

# Nonparametric Prediction in Measurement Error Models

Raymond J. Carroll

Department of Statistics, 3143 TAMU, Texas A&M University, College Station,  
Texas 77843, USA

Aurore Delaigle

Department of Mathematics, University of Bristol, Bristol BS8 1TW, UK and  
Department of Mathematics and Statistics, University of Melbourne, VIC, 3010,  
Australia

Peter Hall

Department of Mathematics and Statistics, University of Melbourne, VIC, 3010,  
Australia and Department of Statistics, University of California at Davis, Davis,  
CA 95616, USA

## Abstract

Predicting the value of a variable  $Y$  corresponding to a future value of an explanatory variable  $X$ , based on a sample of previously observed independent data pairs  $(X_1, Y_1), \dots, (X_n, Y_n)$  distributed like  $(X, Y)$ , is very important in statistics. In the error-free case, where  $X$  is observed accurately, this problem is strongly related to that of standard regression estimation, since prediction of  $Y$  can be achieved via estimation of the regression curve  $E(Y|X)$ . When the observed  $X_i$ s and the future observation of  $X$  are measured with error, prediction is of a quite different nature. Here, if  $T$  denotes the future (contaminated) available version of  $X$ , prediction of  $Y$  can be achieved via estimation of  $E(Y|T)$ . In practice, estimating  $E(Y|T)$  can be quite challenging, as data may be collected under different conditions, making the measurement errors on  $X_i$  and  $X$  non-identically distributed. We take up this problem in the nonparametric setting and introduce estimators which allow a highly adaptive approach to smoothing. Reflecting the complexity of the problem, optimal rates of convergence of estimators can vary from the semiparametric  $n^{-1/2}$  rate to much slower rates that are characteristic of nonparametric problems. Nevertheless, we are able to develop highly adaptive, data-driven methods that achieve very good performance in practice.

**Some Key Words:** Bandwidth, Contamination, Deconvolution, Errors-in-variables, Parametric rates, Regression, Ridge parameter, Smoothing

**Short title:** Errors-in-Variables Prediction

# 1 Introduction

We consider prediction in a problem of nonparametric errors-in-variables regression. In the classical errors-in-variables context, the data consist of a sample of independent and identically distributed observations  $(W_i, Y_i)$ ,  $i = 1, \dots, n$ , generated by the model  $Y_i = g(X_i) + \epsilon_i$ ,  $W_i = X_i + U_i$ , where each  $W_i$  represents a contaminated version of a variable  $X_i$ ,  $X_i$  and  $U_i$  are independent, and, for  $i = 1, \dots, n$ ,  $U_i$  has the distribution with density  $f_U$ , which we indicate by writing  $U_i \sim f_U$ . Nonparametric estimation of  $g$  in this context is a difficult problem, for which optimal estimators converge at notoriously slow rates, see e.g. Fan and Truong (1993). When the interest lies in predicting future values of  $Y$ , however, there is often no need to estimate the function  $g$  explicitly. In particular, if future observations of  $X$  are also measured with an error  $U \sim f_U$ , then it is rarely necessary to address the measurement error, as prediction of  $Y$  can be achieved via estimation of  $E(Y|X + U) = E(Y|W)$  by standard nonparametric regression estimators from the sample  $(W_i, Y_i)$ ,  $i = 1, \dots, n$ . See Carroll et al. (2006) for further discussion of this and related issues.

In empirical applications, however, the above model can be too restrictive, because individuals are not necessarily observed in similar conditions. For example, the data may have been collected from different laboratories (see National Research Council, 1993), and future observations may come from yet another laboratory. In such cases, the data are a sample of independent observations  $(W_i, Y_i)$ ,  $i = 1, \dots, n$ , generated by

$$Y_i = g(X_i) + \epsilon_i, \quad W_i = X_i + U_i, \quad (1.1)$$

where each  $W_i$  represents a contaminated version of a variable  $X_i \sim f_X$ , with error  $U_i \sim f_{U_i}$  and where  $\{X_i, U_i, \epsilon_i\}_{i=1, \dots, n}$  are mutually independent, and future

observations are of the type  $T = X + U^F$ , where  $X$  and  $U^F$  are independent,  $U^F \sim f_{U^F}$  and  $X$  has the same distribution as the  $X_i$ s. As in the setting of the previous paragraph, since future values of  $X$  are of the type  $T = X + U^F$ , nonparametric prediction of  $Y$  can be achieved via nonparametric estimation of  $\mu(t) = E(Y|T = t)$ . Unlike the case of the previous paragraph, however, this cannot be done via standard nonparametric regression estimators. Indeed, given that  $f_{U_1}, \dots, f_{U_n}$  and  $f_{U^F}$  can all be different, the  $W_i$ s are not necessarily identically distributed, nor are they distributed like  $T$ . Despite this difficulty, the major purpose of this paper is to show that it is possible to estimate  $\mu(t)$  nonparametrically. Moreover, the convergence rates of our estimator can be as fast as the parametric  $n^{-1/2}$  rate. While we acknowledge that asymptotic convergence rates do not tell the entire story about the relative performance of estimators, and in particular that multiplicative constants can also be important, our numerical results indicate that our method also does very well in practice.

Heteroscedasticity in the errors arises in many different ways and has been treated by several authors. See, for example, Devanarayana and Stefanski (2002), Kulathinal et al. (2002), Thamerus (2003), Cheng and Riu (2006), Delaigle and Meister (2007), Staudenmayer et al. (2008) and Delaigle and Meister (2008). In some contexts, it is reasonable to assume that there is only a small number of different error densities. In other cases of interest, the error densities could reasonably all come from the same parametric family and differ only through the parameters of their distributions. Indeed, it is commonly assumed that all errors are centered normal random variables. See, for example, Cook and Stefanski (1994), Carroll et al. (1999), Berry et al. (2002), Devanarayana and Stefanski (2002), Kulathinal et al. (2002), Staudenmayer and Ruppert (2004) and Staudenmayer et al. (2008).

The work in this paper was originally motivated by applications where the errors in the sample  $W_1, \dots, W_n$  are of only two types, and the error on future observations is of one of these two types. To fix ideas, suppose the data have been rearranged such that, for  $i = 1, \dots, m$ ,  $U_i \sim f_{U^{(1)}}$  and for  $i = m + 1, \dots, n$ ,  $U_i \sim f_{U^{(2)}}$ , whereas future observations are of the type  $T = X + U^F$  with  $U^F \sim f_{U^{(1)}}$ . Although it is much simpler, this model is important in practical applications (see also Carroll et al. 2006, page 38–39), and we shall discuss it in detail. In Section 5, we apply this two-error model on a dietary data set where the goal is to predict a nutrient intake from a Food Frequency Questionnaire.

There is an extensive literature on estimation of a regression curve from contaminated data sets. A contemporary introduction to this problem is provided by Carroll et al. (2006), and recent contributions include those of Kim and Gleser (2000), Stefanski (2000), Taupin (2001), Linton and Whang (2002), Schennach (2004a,b) and Huang et al. (2006). Nonparametric estimation of a regression curve without contamination is a much older problem, treated in monographs such as those by Wand and Jones (1995) and Simonoff (1996).

An outline of this paper is as follows. In Section 2, when the measurement error densities are known, we describe estimators of the target  $\mu(t)$ . In Section 3, we show how to extend these methods to the case that the measurement error densities are unknown. Section 4 gives the rates of convergence of our estimators, and in particular discusses cases where our estimators achieve parametric rates of convergence. Section 5 gives numerical results, both in simulations and in a nutritional epidemiology context.

## 2 Estimators

### 2.1 Estimators for the general case

Suppose we have a sample of data  $(W_i, Y_i)$ ,  $i = 1, \dots, n$ , generated as at (1.1); that the future observations of  $X$  are of the type  $T = X + U^F$ , where  $X$  and  $U^F$  are independent and  $U^F \sim f_{U^F}$ ; and that the error densities  $f_{U_i}$  and  $f_{U^F}$  are known. The case where these are totally or partially unknown will be discussed in Section 3. We wish to predict  $Y$  nonparametrically, via estimation of

$$\mu(t) = E(Y|T = t) = \int y f_{T,Y}(t, y) dy / f_T(t). \quad (2.1)$$

The task seems challenging, as we need to estimate  $T$ -related quantities from a sample of  $W$ -related quantities. The relationship between  $T$  and each  $W_j$ , however, when expressed in terms of their characteristic functions, is relatively simple. Let  $f_V^{\text{Ft}}$  denote the characteristic function of the distribution of a random variable  $V$ . Then it is easy to check that  $f_T^{\text{Ft}}(t) = f_X^{\text{Ft}}(t)f_{U^F}^{\text{Ft}}(t)$  and  $f_X^{\text{Ft}}(t) = f_{W_j}^{\text{Ft}}(t)/f_{U_j}^{\text{Ft}}(t)$ . Assuming that none of the  $f_{U_j}^{\text{Ft}}(t)$ s vanishes and  $f_T^{\text{Ft}}(t)$  is integrable, it follows that, by the Fourier inversion theorem,

$$f_T(x) = \frac{1}{2\pi} \int e^{-itx} f_T^{\text{Ft}}(t) dt, \quad (2.2)$$

where we can write

$$f_T^{\text{Ft}}(t) = f_{U^F}^{\text{Ft}}(t)n^{-1} \sum_j \{f_{W_j}^{\text{Ft}}(t)/f_{U_j}^{\text{Ft}}(t)\}. \quad (2.3)$$

Based on these considerations, we show below how to estimate  $f_T^{\text{Ft}}$  and  $f_T$ . Then, we construct an estimator of the numerator of (2.1), and finally obtain an estimator of  $\mu$ .

The simplest device to obtain a consistent estimator of  $f_T$  is to replace  $f_{W_j}^{\text{Ft}}(t)$  in (2.3) by  $e^{itW_j}$ , an unbiased estimator. However, the form of a sum of ratios in (2.3) implies that the variance of the resulting estimator of  $f_T$ , which

depends on the behavior of the  $f_{U_j}^{\text{Ft}}$ s in the tails, would be dominated by the variance of the least favorable errors. This simple approach thus does not lead to optimal estimators. Alternatively, to gain more precision, we could rewrite (2.3) by using the ratio of the sums of  $f_{W_j}^{\text{Ft}}$  and  $f_{U_j}^{\text{Ft}}$ . Specifically, first note that  $1/f_{U_j}^{\text{Ft}}(t) = f_{U_j}^{\text{Ft}}(-t)/|f_{U_j}^{\text{Ft}}(t)|^2$ . Use the notation  $\Psi_j(t) = f_{U_j}^{\text{Ft}}(-t)/\sum_k |f_{U_k}^{\text{Ft}}(t)|^2$ . Since  $f_{W_j}^{\text{Ft}}(t) = f_X^{\text{Ft}}(t)f_{U_j}^{\text{Ft}}(t)$ , it follows that  $f_{W_j}^{\text{Ft}}(t)f_{U_j}^{\text{Ft}}(-t) = f_X^{\text{Ft}}(t)|f_{U_j}^{\text{Ft}}(t)|^2$ , which implies that

$$f_T^{\text{Ft}}(t) = f_{U^F}^{\text{Ft}}(t) \sum_j f_{W_j}^{\text{Ft}}(t) f_{U_j}^{\text{Ft}}(-t) / \sum_k |f_{U_k}^{\text{Ft}}(t)|^2 = \sum_j f_{W_j}^{\text{Ft}}(t) \Psi_j(t) f_{U^F}^{\text{Ft}}(t).$$

Now, replacing  $f_{W_j}^{\text{Ft}}(t)$  by its unbiased estimate  $e^{itW_j}$ , we can estimate  $f_T^{\text{Ft}}(t)$  from the data  $(W_1, Y_1), \dots, (W_n, Y_n)$  by

$$\widehat{f}_T^{\text{Ft}}(t) = \sum_{j=1}^n e^{itW_j} \Psi_j(t) f_{U^F}^{\text{Ft}}(t). \quad (2.4)$$

We shall show in Section 4 that this procedure leads to optimal estimators of  $\mu$ .

If  $f_{U^F}^{\text{Ft}}(t) \sum_j \Psi_j(t) \in L_1$  we can obtain an estimator of the denominator of (2.1) by plugging (2.4) into (2.2), which gives  $\widehat{f}_T(x) = \sum_j f_{T,j}(x - W_j)$ , where we employed the notation

$$f_{T,j}(x) = \frac{1}{2\pi} \int e^{-itx} f_{U^F}^{\text{Ft}}(t) \Psi_j(t) dt. \quad (2.5)$$

Using a similar approach, we estimate the numerator of (2.1) by  $\sum_j Y_j f_{T,j}(x - W_j)$ , and we obtain an estimator of  $\mu$  by taking the ratio of these two estimators:

$$\widehat{\mu}(x) = \frac{\sum_j Y_j f_{T,j}(x - W_j)}{\sum_j f_{T,j}(x - W_j)}. \quad (2.6)$$

When  $f_{U^F}^{\text{Ft}}(t) \sum_j \Psi_j(t) \notin L_1$  we need to regularize  $\widehat{f}_T^{\text{Ft}}$  before plugging it into (2.2). This challenge is also encountered in relatively classical deconvolution problems. We use a kernel approach to regularize this problem. Methodology

based on another nonparametric technique, such as splines, orthogonal series or the ridge technique of Hall and Meister (2007), could also be developed. While those methods might be competitive with the kernel approach, the latter benefits from being relatively accessible to asymptotic analysis. Let  $K$  be a kernel function with Fourier transform  $K^{\text{Ft}}$ , and let  $h > 0$  be a bandwidth. Assuming that  $K^{\text{Ft}}(t)f_{U^{\text{F}}}^{\text{Ft}}(t/h)\Psi_j(t/h) \in L_1$ , our regularized estimator of  $f_T$  is  $\tilde{f}_T(x) = h^{-1} \sum_j K_{T,j}\{(x - W_j)/h\}$ , where

$$K_{T,j}(x) = \frac{1}{2\pi} \int e^{-itx} K^{\text{Ft}}(t) f_{U^{\text{F}}}^{\text{Ft}}(t/h) \Psi_j(t/h) dt.$$

Proceeding as above, we define our estimator of  $\mu$  by

$$\tilde{\mu}(x) = \frac{\sum_j Y_j K_{T,j}\left(\frac{x-W_j}{h}\right)}{\sum_j K_{T,j}\left(\frac{x-W_j}{h}\right)}. \quad (2.7)$$

## 2.2 The two-error model

In the two-error model which motivated our work, the estimators of  $\mu$  are particularly simple. To keep the same notation as in the introduction, assume that the first  $m$  observations are contaminated by an error with density  $f_{U^{(1)}}$ , that the last  $n - m$  are contaminated by an error with density  $f_{U^{(2)}}$ , and that the error in the future observation is  $U_F \sim f_{U^{(1)}}$ . Let  $q^{\text{Ft}} = f_{U^{(2)}}^{\text{Ft}}(t)/f_{U^{(1)}}^{\text{Ft}}(t)$ .

If  $1/q^{\text{Ft}} \in L_1$ , we use the estimator at (2.6), which becomes

$$\hat{\mu}(x) = \frac{\sum_{j=1}^m Y_j f_{q,1}(x - W_j) + \sum_{j=m+1}^n Y_j f_{q,2}(x - W_j)}{\sum_{j=1}^m f_{q,1}(x - W_j) + \sum_{j=m+1}^n f_{q,2}(x - W_j)}, \quad (2.8)$$

where we used the notations  $f_{q,1}(x) = (2\pi)^{-1} \int e^{-itx} \{m + (n - m)|q^{\text{Ft}}(t)|^2\}^{-1} dt$  and  $f_{q,2}(x) = (2\pi)^{-1} \int e^{-itx} q^{\text{Ft}}(-t) \{m + (n - m)|q^{\text{Ft}}(t)|^2\}^{-1} dt$ .

If  $1/q^{\text{Ft}} \notin L_1$ , we use the estimator at (2.7), which becomes

$$\tilde{\mu}(x) = \frac{\sum_{j=1}^m Y_j K_{q,1}\left(\frac{x-W_j}{h}\right) + \sum_{j=m+1}^n Y_j K_{q,2}\left(\frac{x-W_j}{h}\right)}{\sum_{j=1}^m K_{q,1}\left(\frac{x-W_j}{h}\right) + \sum_{j=m+1}^n K_{q,2}\left(\frac{x-W_j}{h}\right)}, \quad (2.9)$$

where  $K_{q,1}(x) = (2\pi)^{-1} \int e^{-itx} K^{\text{Ft}}(t) \{m + (n - m)|q^{\text{Ft}}(t/h)|^2\}^{-1} dt$  and  $K_{q,2}(x) = (2\pi)^{-1} \int e^{-itx} K^{\text{Ft}}(t) q^{\text{Ft}}(-t/h) \{m + (n - m)|q^{\text{Ft}}(t/h)|^2\}^{-1} dt$ . Note that in the particular case where we have only observations contaminated by the error density  $f_{U(1)}$ ,  $m = n$  and the estimator at (2.9) is nothing more than the usual Nadaraya-Watson estimator without any contamination, see Wand and Jones (1995, p. 119).

**Remark 2.1.** In the terminology of nonparametric deconvolution, the smoothness of an error (or error density) is usually described in terms of the speed of convergence to zero of its characteristic function in the tails — the faster, the smoother. See Section 4.2 for discussion. In this context, roughly speaking,  $1/q^{\text{Ft}} \in L_1$  implies that  $f_{U(1)}$  is smoother than  $f_{U(2)}$ . For example, if  $f_{U(1)}$  and  $f_{U(2)}$  are normal densities with mean zero and variances  $\sigma_1^2$  and  $\sigma_2^2$ , respectively, then,  $1/q^{\text{Ft}} \in L_1$  if  $\sigma_1^2 > \sigma_2^2$ , and  $1/q^{\text{Ft}} \notin L_1$  if  $\sigma_1^2 \leq \sigma_2^2$ . The condition  $f_{U^{\text{Ft}}}^{\text{Ft}}(t) \sum_j \Psi_j(t) \in L_1$  can be understood in a related manner.

### 2.3 Local polynomial extension

The estimators presented above are an extension of the Nadaraya-Watson estimator, which is nothing more than a local constant estimator appropriate for error-free data. Recently, Delaigle, Fan and Carroll (2008) solved the long-open problem of developing local polynomial estimators for errors-in-variables problems. Using their technique we can give a local polynomial version of our estimator  $\tilde{\mu}$ . More precisely, we define a  $p$ th order local polynomial estimator of  $\mu$  by  $\tilde{\mu}_p(x) = (1, 0, \dots, 0) \widehat{\mathbf{S}}_{\mathbf{n}}^{-1} \widehat{\mathbf{T}}_{\mathbf{n}}$ , where  $\widehat{\mathbf{S}}_{\mathbf{n}} = (\widehat{S}_{n,j+\ell}(x))_{0 \leq j, \ell \leq p}$  and  $\widehat{\mathbf{T}}_{\mathbf{n}} = (\widehat{T}_{n,0}(x), \dots, \widehat{T}_{n,p}(x))^{\text{T}}$  with

$$\widehat{S}_{n,k}(x) = \sum_{j=1}^n K_{U,k,j;h}(W_j - x) \quad \text{and} \quad \widehat{T}_{n,k}(x) = \sum_{j=1}^n Y_j K_{U,k,j;h}(W_j - x),$$

where  $K_{U,k,j;h}(x) = h^{-1}K_{U,k,j}(x/h)$  and

$$K_{U,k,j}(x) = i^{-k} \frac{1}{2\pi} \int e^{-itx} (K^{\text{Ft}})^{(k)}(t) f_{U^{\text{F}}}^{\text{Ft}}(t/h) \Psi_j(t/h) dt.$$

Compared to local constant estimators ( $p = 0$ ), local polynomial estimators for  $p > 0$  have the advantage of being less biased in the presence of boundary points. On the other hand, in practice, increasing the value of  $p$  usually leads to an increase of variability, and using values of  $p$  larger than 1 is rarely useful unless the interest is in estimating derivatives of the curve  $\mu$ . This, however, is usually not the case in the prediction problem.

It is straightforward to extend the local-constant methodology of Delaigle, Hall and Meister (2008) for the case of unknown error distributions to the context of general local polynomial estimators. Properties are similar, too. For example, convergence rates in the case of local linear methods are identical, under regularity conditions discussed towards the end of Section 3.1, to those for the local constant estimators treated in this paper.

### 3 Unknown error densities

There are many examples where it is too restrictive to assume that the error densities are completely known, and in such cases, these densities have to be estimated from the data. For a long time this problem was essentially ignored in the nonparametric literature, where the error distributions were systematically assumed to be known. Recently, however, several authors have shown that this problem can be tackled if every observation is replicated at least once. References include Schennach (2004a,b), Delaigle, Hall and Meister (2008) and Hu and Schennach (2008). In the next section we discuss parametric and nonparametric methods for error density estimation in the two-error model treated in Section 2.2. A procedure for the general model is given in Section 3.2.

### 3.1 Procedures for the two-error model

In the two-population case, if we do not have enough data to estimate the error densities, and if  $m$  is not particularly small, then a consistent estimator of  $\mu$  can be obtained by taking  $n = m$ , i.e. discarding all observations contaminated by  $f_{U(2)}$  and using the standard Nadaraya-Watson estimator. See also Remark 3.1, below. The most interesting setting is undoubtedly that where it is possible to estimate the error densities, as it is in this case that the estimator of  $\mu$  will enjoy the fastest rates of convergence; see Section 4.

As discussed in the introduction, a large literature on measurement errors assumes that the errors are normal. More generally, the errors could belong to some parametric family, not necessarily normal. There, if we have a parametric model  $f_{U(1)}(\cdot | \theta_1)$  (respectively,  $f_{U(2)}(\cdot | \theta_2)$ ) that is identifiable from data on  $U - U'$ , where  $U$  and  $U'$  denote two independent variables from  $f_{U(1)}$  (respectively,  $f_{U(2)}$ ), then  $\theta_1$  and  $\theta_2$  can be estimated from a sample of replicated data, i.e. a sample of the form  $(W_{ij}, Y_i)$ ,  $i = 1, \dots, n$  and  $j = 1, 2$ , generated by the model

$$Y_i = g(X_i) + \epsilon_i, \quad W_{ij} = X_i + U_{ij}, \quad (3.1)$$

where, for  $i = 1, \dots, m$ ,  $U_{ij} \sim f_{U(1)}$ , whereas, for  $i = m + 1, \dots, n$ ,  $U_{ij} \sim f_{U(2)}$ , and the  $U_{ij}$ s are all independent, and independent of each  $X_i$ . For example, if  $\theta_i = \sigma_{U(i)}^2$ ,  $i = 1, 2$ , then we can take  $\hat{\theta}_i$  equal to a weighted average of the within-subject sample variance; see equation (4.3) of Carroll et al. (2006). Once the unknown parameters have been estimated, the resulting estimated characteristic functions  $\hat{f}_{U(1)}^{\text{Ft}}$  and  $\hat{f}_{U(2)}^{\text{Ft}}$  are plugged into the estimators (2.8) and (2.9), to produce our estimators of  $\mu$ .

In some cases it can happen that we have no suitable parametric model for the error densities. If the characteristic functions of the error densities are positive and symmetric, as is the case for many common densities, then

they can be estimated nonparametrically along the lines of Delaigle, Hall and Meister (2008). More precisely,  $f_{U(1)}^{\text{Ft}}$  and  $f_{U(2)}^{\text{Ft}}$ , in (2.8) and (2.9), are estimated by, respectively,  $\widehat{f}_{U(1)}^{\text{Ft}}(t) = |m^{-1} \sum_{j=1}^m \cos\{t(W_{j1} - W_{j2})\}|^{1/2}$  and  $\widehat{f}_{U(2)}^{\text{Ft}}(t) = |(n - m)^{-1} \sum_{j=m+1}^n \cos\{t(W_{j1} - W_{j2})\}|^{1/2}$ . We then replace  $f_{U(1)}^{\text{Ft}}$  and  $f_{U(2)}^{\text{Ft}}$  by  $\widehat{f}_{U(1)}^{\text{Ft}}$  and  $\widehat{f}_{U(2)}^{\text{Ft}}$  in the numerators of  $f_{q,1}$ ,  $f_{q,2}$ ,  $K_{q,1}$  and  $K_{q,2}$ , but in the denominators, to avoid division by zero, we replace  $m|f_{U(1)}^{\text{Ft}}|^2 + (n - m)|f_{U(2)}^{\text{Ft}}|^2$  by  $m|\widehat{f}_{U(1)}^{\text{Ft}}|^2 + (n - m)|\widehat{f}_{U(2)}^{\text{Ft}}|^2 + r$ , with  $r > 0$ ; see Delaigle, Hall and Meister (2008). More general settings are considered by Li and Vuong (1998), Schennach (2004a,b) and Hu and Schennach (2008).

Convergence rates in the unknown error case, and in the setting of classical errors-in-variables problems, have been given by Delaigle, Hall and Meister (2008). The results there state that, if the characteristic function of the unknown error distribution is estimated using a difference-based method, then the convergence rate is the same as in the setting of a known error distribution, provided the density of  $X$  is sufficiently smooth relative to the error density. This is also true in the prediction problem treated in the present paper.

**Remark 3.1.** When  $m$  is large relative to  $n - m$ , and the error densities cannot be well estimated, a classical Nadaraya-Watson estimator of  $\mu$ , based on  $(W_1, Y_1), \dots, (W_m, Y_m)$ , is likely to perform better than our estimator. For example, this could happen if the errors densities had to be estimated nonparametrically from a small number of individuals for which there were replicated observations. In such cases, a conservative approach would be to use the Nadaraya-Watson estimator.

### 3.2 A procedure for the general model

Before we show how to construct a consistent estimator in the general context of the model (1.1), it is important to realize that, whatever approach we take, in order for the function  $\mu$  to be identifiable we need to be able to consistently estimate the error density  $f_{U^F}$  of the future observations. See Section 3.1 for a discussion on how to estimate an error density. We assume that sufficient effort has been made by the experimenters to collect data permitting the construction of a consistent estimator  $\widehat{f}_{U^F}$  of  $f_{U^F}$ . If  $f_{U^F}$  cannot be estimated then the prediction problem is not identifiable.

As in Section 3.1, for simplicity we address the case where there are just two replicated measurements of  $X_i$  for each  $i$ , that is, we have data of the form  $(W_{ij}, Y_i)$ , for  $i = 1, \dots, n$  and  $j = 1, 2$ , generated by the model (3.1), where  $U_{ij} \sim f_{U_i}$  and the  $U_{ij}$ s are all independent, and independent of every  $X_i$  (the case of a larger number of replicates can be treated similarly). Of course, it is not possible to estimate each error density  $f_{U_i}$  from such data. Nevertheless, if the characteristic functions of the error densities are positive and symmetric and if we modify our estimators at (2.6) and (2.7) appropriately, it is possible to construct a consistent estimator of  $\mu$ , as we show below.

Let  $\overline{W}_i = (W_{i1} + W_{i2})/2$ , and note that  $f_X^{\text{Ft}}(t) = f_{\overline{W}_j}^{\text{Ft}}(t)/\{f_{U_j}^{\text{Ft}}(t/2)\}^2 = \Phi(t) \sum_j f_{\overline{W}_j}^{\text{Ft}}(t)$ , where we used the notation  $\Phi(t) = 1/\sum_k \{f_{U_k}^{\text{Ft}}(t/2)\}^2$ . Replacing the unknown  $\Phi(t)^{-1}$  by the estimator  $\widehat{\Phi}(t)^{-1} = \sum_k \exp(it(W_{k,1} - W_{k,2})/2)$ , and proceeding as in Section 2, we obtain the following versions of (2.6) and (2.7):

$$\widehat{\mu}^*(t) = \frac{\sum_{j=1}^n Y_j f_T^*(x - \overline{W}_j)}{\sum_{j=1}^n f_T^*(x - \overline{W}_j)}, \quad \widetilde{\mu}^*(t) = \frac{\sum_{j=1}^n Y_j K_T^*\left(\frac{x - \overline{W}_j}{h}\right)}{\sum_{j=1}^n K_T^*\left(\frac{x - \overline{W}_j}{h}\right)},$$

with  $f_T^*(x) = (2\pi)^{-1} \int e^{-itx} \widehat{f}_{U_F}^{\text{Ft}}(t) / \{\widehat{\Phi}^{-1}(t) + r\}$ , and with

$$K_T^*(x) = (2\pi)^{-1} \int e^{-itx} K^{\text{Ft}}(t) \widehat{f}_{U_F}^{\text{Ft}}(t/h) / \{\widehat{\Phi}^{-1}(t/h) + r\} dt,$$

where, as before,  $r > 0$  is introduced to avoid division by zero.

## 4 Theoretical properties

The properties of the estimator  $\widehat{\mu}$  at (2.6) are clear. In particular, it is easy to check that the numerator and the denominator are both unbiased estimators of  $\mu(t)f_T(t)$  and  $f_T(t)$ , respectively, and that, under conditions similar to those discussed below Theorem 4.1,  $\widehat{\mu}$  converges at the fast parametric  $n^{-1/2}$  rate. Intuitive explanation of why this fast rate can occur in a nonparametric context will be given in Remark 4.2. Properties of the estimator  $\widetilde{\mu}$  are more complicated and, in what follows, we derive them in the general case. Then we obtain more detailed results in the two-error problem. Proofs of the results presented in this section can be found in the supplemental material of this paper, available at [http://www.amstat.org/publications/jasa/supplemental\\_materials](http://www.amstat.org/publications/jasa/supplemental_materials).

### 4.1 Asymptotic results for $\widetilde{\mu}$ in the general case

To study asymptotic properties of our estimator, we assume that:

$$\epsilon_i, \dots, \epsilon_n \text{ have zero means and uniformly bounded variances;} \quad (4.1)$$

$$\begin{aligned} &K \text{ is symmetric, } |K(x)| \leq C_3 (1 + |x|)^{-k-1-C_4} \text{ for an integer } k \geq 2 \text{ and for} \\ &\text{constants } C_3, C_4 > 0, \int u^j K(u) du = 0 \text{ for } 1 \leq j \leq k-1, \sup |K^{\text{Ft}}| < \infty, \\ &K^{\text{Ft}}(0) > 0, \text{ and, for some } C_5 > 0, K^{\text{Ft}}(t) = 0 \text{ for all } |t| > C_5, \end{aligned} \quad (4.2)$$

$$f_X, f_{U_1}, \dots, f_{U_n}, f_{U^F} \text{ and } g \text{ are bounded, and } f_X \text{ and } g \text{ have } k \text{ bounded derivatives;} \quad (4.3)$$

$$\sup_k |f_{U_k}^{\text{Ft}}|^2 \text{ and } |f_{U^F}^{\text{Ft}}| \text{ are bounded and, for all } t, \sum_k |f_{U_k}^{\text{Ft}}(t)| > 0. \quad (4.4)$$

$$h \rightarrow 0 \text{ as } n \rightarrow \infty \text{ and } n/v(h) \rightarrow \infty \text{ as } n \rightarrow \infty, \quad (4.5)$$

where we defined

$$v(h) = nh^{-1} \int |K^{Ft}(t)|^2 |f_{U^F}^{Ft}(t/h)|^2 / \sum_{k=1}^n |f_{U_k}^{Ft}(t/h)|^2 dt. \quad (4.6)$$

Assumptions such as these are fairly standard in the nonparametric regression literature. Condition (4.1) is mild, and the smoothness of the various curves in (4.3) is imposed only to determine the order of the bias of the estimator, which depends of  $k$ . In particular,  $k$  is generally not a tuning parameter, and in empirical examples, where the smoothness of the curves is usually unknown, it is common to set  $k = 2$ . It is well known that, in practice, larger values of  $k$  increase the variability of estimators and usually make them unattractive, see, for example, Marron and Wand (1992). Of course, as in the standard error-free case, our results can be extended to cases where  $f_X$  and  $g$  have discontinuity points, with the obvious modifications to the bias. Condition (4.2) only concerns the kernel (which we can choose) and is satisfied by the kernels used in deconvolution problems. Condition (4.4) is a weaker version of standard conditions usually imposed in deconvolution problems (see e.g Fan, 1991), since, in our case, the characteristic functions of the errors are permitted to vanish. Condition (4.5) is a generalization of the standard condition  $nh \rightarrow \infty$  imposed in the error-free case, but it looks more complicated here because the variance of the estimator is of order  $v(h)/n$  rather than  $1/(nh)$ .

The asymptotic behavior of the estimator is described in the next theorem.

**Theorem 4.1.** *Assume (4.1)–(4.5). Then, for each  $t$  such that  $f_T(t) > 0$ ,*

$$\tilde{\mu}(t) = \mu(t) + O_p\left(\{v(h)/n\}^{1/2} + h^k\right). \quad (4.7)$$

Precise rates of convergence of the estimator depend on the behavior of the ratio  $Q(t) \equiv n|f_{U^F}^{\text{Ft}}(t)|^2 / \sum_k |f_{U_k}^{\text{Ft}}(t)|^2$  in the tails. It is not possible here to consider every possible combination of error types. To get some insight on these results, consider the situation where, for all  $t$ ,  $\sum_k |f_{U_k}^{\text{Ft}}(t)|^2 > n\xi(t)$  where  $\xi$  is a continuous, strictly positive function. Then, if  $Q(t) = o(|t|^{-1})$  as  $t \rightarrow \infty$ , by taking  $h = O(n^{-1/k})$ , the estimator converges at the fast parametric  $n^{-1/2}$  rate. When  $Q(t) = O(1)$  and  $|t|Q(t) \rightarrow \infty$  as  $t \rightarrow \infty$ , the estimator converges at a rate that lies between  $n^{-1/2}$  and the classical nonparametric rate  $n^{-k/(2k+1)}$ , and in other cases it converges more slowly than  $n^{-k/(2k+1)}$ . An intuitive explanation of the occurrence of fast parametric rates will be given in Remark 4.2.

## 4.2 Detailed results in the two-error case

In the two-error problem described in Section 2.2, it is possible to provide a more detailed study of the convergence rates of the estimator  $\tilde{\mu}$ , which, in this case, reduces to (2.9). Precise rates of convergence depend on the values of  $m$  and  $n$ , and on the behavior of  $q^{\text{Ft}}$  in the tails, which itself is dictated by the behaviors of the characteristic functions  $f_{U^{(1)}}^{\text{Ft}}$  and  $f_{U^{(2)}}^{\text{Ft}}$  of the errors. In the measurement error literature it is well known that convergence rates of nonparametric estimators depend heavily on the behavior of the characteristic function of the error in the tails. This tail behavior is usually referred to as the smoothness of the error, and it is standard to divide the error distributions in two quite different categories, called ordinary smooth and supersmooth in the terminology of Fan (1991). The errors  $f_{U^{(1)}}$  and  $f_{U^{(2)}}$  are ordinary smooth of orders  $\beta$  and  $\alpha$ , respectively, if they satisfy, for positive constants  $C_1 < C'_1$  and  $C'_2 < C_2$  and for all  $t$ ,

$$\begin{aligned} C'_2 (1 + |t|)^{-\beta} &\leq |f_{U^{(1)}}^{\text{Ft}}(t)| \leq C_2 (1 + |t|)^{-\beta}, \\ C_1 (1 + |t|)^{-\alpha} &\leq |f_{U^{(2)}}^{\text{Ft}}(t)| \leq C'_1 (1 + |t|)^{-\alpha}. \end{aligned} \tag{4.8}$$

An error density  $f_U$  is supersmooth of order  $\beta > 0$  if it satisfies, for positive constants  $\gamma$ ,  $D_1 < D_2$  and for all  $t$ ,

$$D_1 \exp(-|t|^{-\beta}/\gamma) \leq |f_U^{\text{Ft}}(t)| \leq D_2 \exp(-|t|^{-\beta}/\gamma). \quad (4.9)$$

For simplicity of presentation we give our main results under the assumption that, for constants  $\alpha, \beta, C_1, C_2 > 0$  and all real  $t$ ,

$$C_1 (1 + |t|)^{-\alpha} \leq |f_{U^{(2)}}^{\text{Ft}}(t)|, \quad |f_{U^{(1)}}^{\text{Ft}}(t)| \leq C_2 (1 + |t|)^{-\beta}. \quad (4.10)$$

Obviously, ordinary smooth errors satisfy both inequalities, but supersmooth errors also satisfy the second inequality for any  $\beta > 0$ .

Define  $\delta = \alpha - \beta + \frac{1}{2}$  and, denoting the indicator function by  $1_{\{\cdot\}}$ , let

$$v_1(h) = h^{-2\delta} \cdot 1_{\{\delta > 0\}} + |\log h| \cdot 1_{\{\delta = 0\}} + 1_{\{\delta < 0\}}. \quad (4.11)$$

Assume that, as  $n \rightarrow \infty$ ,

$$h \rightarrow 0, \quad n/v_1(h) \rightarrow \infty \text{ and } mh \rightarrow \infty. \quad (4.12)$$

The asymptotic behavior of  $\tilde{\mu}$  is described in the next theorem.

**Theorem 4.2.** *Assume (4.1)–(4.4) and (4.10)–(4.12). Then, for each  $t$  such that  $f_T(t) > 0$ ,*

$$\tilde{\mu}(t) = \mu(t) + O_P(h^k + \min\{(mh)^{-1/2}, (n-m)^{-1/2}v_1(h)^{1/2}\}). \quad (4.13)$$

This result shows that our estimator  $\tilde{\mu}$  converges at a rate at least as fast as the Nadaraya-Watson estimator of  $\mu$  based on direct data from  $(T, Y)$ , i.e. on the first  $m$  observations  $(W_1, Y_1), \dots, (W_m, Y_m)$ . Indeed, the Nadaraya-Watson estimator is optimized when taking  $h \sim \text{const. } m^{-1/(2k+1)}$ , for which it converges at the rate  $m^{-k/(2k+1)}$ . Of course, the cases where our estimator converges faster than the Nadaraya-Watson estimator depend on the relative sizes of  $m$  and  $n$ , but also on the relative smoothness of the two errors, as we show below.

The most favorable situation is clearly the one where  $\delta < 0$ . This includes the case where  $f_{U(1)}$  and  $f_{U(2)}$  are ordinary smooth of orders  $\beta$  and  $\alpha$ , respectively, with  $\beta > \alpha + \frac{1}{2}$ , but it also includes the case where, simultaneously,  $f_{U(1)}$  is supersmooth of any order, and  $f_{U(2)}$  is ordinary smooth of any order. When  $m = o\{(n - m)^{(2k+1)/2k}\}$ , we obtain the very fast parametric rate  $(n - m)^{-1/2}$ , by taking  $h = O\{(n - m)^{-1/(2k)}\}$ . In particular, when  $m$  and  $n - m$  are of the same order, or  $m = o(n)$ , the estimator converges at the rate  $n^{-1/2}$ . When  $m \neq o\{(n - m)^{(2k+1)/2k}\}$ , i.e. when there are many more data contaminated by  $f_{U(1)}$  than by  $f_{U(2)}$ , then, logically, the estimator converges at the same rate as the Nadaraya-Watson estimator based on the  $m$  first data points. More precisely, when  $h \sim \text{const. } m^{-1/(2k+1)}$  the estimator converges at the rate  $m^{-k/(2k+1)}$ .

The case where  $\delta > 0$  is more involved since, there, the term  $(n - m)^{-1/2} v_1(h)^{1/2}$  in (4.13) is only an upper bound to the contribution of the data  $(W_i, Y_i)$  for  $i = m+1, \dots, n$ , and precise characterization of convergence rates can be obtained only at the expense of more precise characterization of  $q^{\text{Ft}}$ . We shall assume that the errors are ordinary smooth of order  $\beta$  and  $\alpha$ , as defined at (4.8). Under this assumption, the convergence rate of the estimator is exactly of the order given in the theorem. The optimal bandwidth is thus of order  $h \sim \text{const. } m^{-1/(2k+1)}$  when  $(n - m)^{(2k+1)/(2k+2\delta)} = o(m)$ , and the estimator then converges at the rate  $m^{-k/(2k+1)}$ . When  $m = o\{(n - m)^{(2k+1)/(2k+2\delta)}\}$ , the optimal bandwidth is of order  $h \sim \text{const. } (n - m)^{-1/(2k+2\delta)}$  and the estimator converges at rate  $(n - m)^{-k/(2k+2\delta)}$ .

**Remark 4.1.** (*Other types of errors*). Calculation of the rates for  $\tilde{\mu}$  can be extended to cases more general than (4.10). For example, it can be shown that the convergence rates are of order  $\min\{m^{-k/(2k+1)}, (n - m)^{-1/2}\}$  whenever  $t^{1/2}/q^{\text{Ft}}(t) < \text{const.}$  as  $t \rightarrow \infty$ ; they are of order  $\min\{m^{-k/(2k+1)}, B(n)\}$  where  $B(n) \rightarrow 0$  as  $n \rightarrow \infty$  at a speed similar to typical deconvolution rates (see e.g.

Fan and Truong, 1993), when  $1/q^{\text{Ft}}(t) \rightarrow \infty$  as  $t \rightarrow \infty$ ; and it is of an order between  $\min\{m^{-k/(2k+1)}, B(n)\}$  and  $\min\{m^{-k/(2k+1)}, (n-m)^{-1/2}\}$  in all other cases.

**Remark 4.2.** (*On the parametric rates — I*). The very fast parametric rate noted above may appear counter-intuitive. It can be understood from the fact that, since we know that  $f_T = f_X * f_{U^{(1)}}$ , we have some valuable information about the structure of  $f_T$ : we know that it is the convolution of an estimable density  $f_X$  and a known density  $f_{U^{(1)}}$ . If we had a direct sample from  $f_X$ , this would be the so-called Berkson problem, for which it has been established that the rates of convergence are parametric. See Delaigle (2007). Our situation is more complicated since we can estimate  $f_X$  only indirectly, and it is only when  $f_{U^{(1)}}$  is smooth enough that we can obtain fast parametric convergence rates. Unlike our situation, note that, in the Berkson problem, the parametric rate occurs only in the density estimation context, and not in the regression setting.

**Remark 4.3.** (*On the parametric rates — II*). When the covariate  $X$  is observed without error, for past and future observations, instead of applying standard nonparametric estimators of  $E(Y|X)$ , which only converge at the rate  $n^{-k/(2k+1)}$ , it may seem to be a better idea to artificially add random noise  $U \sim f_U$  to the future observed value of  $X$ , with  $f_U$  such that  $f_U^{\text{Ft}} \in L_1$ , and predict  $Y$  by an estimator of  $E(Y|T)$ , where  $T = X + U$ . Indeed, this would correspond to our model, in the situation where  $1/q^{\text{Ft}}(t) = f_U^{\text{Ft}} \in L_1$ , in which case we can use  $\hat{\mu}$ , which converges at a  $n^{-1/2}$  rate. However, it is not clear that, despite the convergence rate, this would lead to better prediction of  $Y$  than the error-free estimator of  $E(Y|X)$ , since  $Y$  is more dispersed around  $E(Y|T)$  than  $Y$  is around  $E(Y|X)$  because  $T$  exhibits larger errors than  $X$ .

### 4.3 Optimal convergence rates

Here we indicate that in the ordinary smooth case, the convergence rates given by Theorem 4.2 are optimal when  $\delta > 0$ . A simpler argument can be used to demonstrate optimality when  $\delta < 0$ , and similar methods can be employed to verify optimality in supersmooth settings. We prove results only in the two-error case, but similar techniques can be used to show that, under regularity conditions, our estimator is also optimal in the general setting of model (1.1).

Let  $f_{U(2)}$  and  $f_{U(1)}$  denote symmetric densities for which

$$\limsup_{|t| \rightarrow \infty} \max_{j=0,1,2} (1 + |t|)^{\alpha+j} \left| \frac{d^j}{dt^j} f_{U(2)}^{\text{Ft}}(t) \right| < \infty, \quad (4.14)$$

$$\liminf_{|t| \rightarrow \infty} \min_{j=0,1,2} (1 + |t|)^{\beta+j} \left| \frac{d^j}{dt^j} f_{U(1)}^{\text{Ft}}(t) \right| > 0. \quad (4.15)$$

Given  $-\infty < a < b < \infty$ ,  $C > 0$  and an integer  $k \geq 1$ , write  $\mathcal{F}(a, b, C, k)$  for the class of densities  $f_{XY}$  of  $(X, Y)$  such that (a)  $\mu = E(Y | T = \cdot)$  and  $f_X$  are both  $k$  times differentiable, with each of these derivatives uniformly bounded, in absolute value, by  $C$ ; (b)  $E(Y^2 | X = x) \leq C$  for all  $x$ ; and (c)  $f_X(x) \geq C^{-1}$  for  $x \in [a, b]$ . Let  $\mathcal{C}$  denote the class of all estimators  $\check{\mu}$  of  $\mu$ .

**Theorem 4.3.** Assume that (4.14) and (4.15) hold, and  $\delta > 0$ . Then, for each real number  $w$ , there exists a constant  $c > 0$  such that

$$\liminf_{n \rightarrow \infty} \inf_{\check{\mu} \in \mathcal{C}} \sup_{f_{XY} \in \mathcal{F}(a, b, C, k)} P \left[ \left| \check{\mu}(w) - \mu(w) \right| > c \min \left\{ (n - m)^{-k/\{2k+2\delta\}}, m^{-k/(2k+1)} \right\} \right] > 0. \quad (4.16)$$

These rates correspond exactly to those in Theorem 4.2. They show that the rate at (4.16) is achieved by the estimator  $\tilde{\mu}$ . Although our upper bound giving this rate was derived only for a particular fixed distribution, that bound is readily established uniformly over a function class for which (4.16) holds.

## 5 Numerical results

We applied our estimators in the particular case where the observations are contaminated by only two types of errors, which is the setting of our empirical example. Note that we have defined two estimators, at (2.8) and (2.9). The first exists only when  $1/q^{\text{Ft}}$  is integrable, and is simpler to calculate. In particular, it requires neither a bandwidth nor a kernel, and our numerical work showed that it systematically outperformed  $\tilde{\mu}$ , which therefore we do not present in the cases where  $\hat{\mu}$  exists. We use the notations  $\hat{\mu}^*$  and  $\tilde{\mu}^*$  for the versions of the estimator  $\hat{\mu}$  and  $\tilde{\mu}$ , respectively, with the error variances estimated from replicated data as in Section 3. We use the notation  $\hat{\mu}_{\text{NW}}$  for the classical Nadaraya-Watson estimator calculated from the data  $(W_i, Y_i)$ ,  $i = 1, \dots, m$ . Note that  $\hat{\mu}_{\text{NW}}$  is exactly equal to  $\tilde{\mu}$  when  $m = n$ .

### 5.1 Simulations

We applied the various estimators introduced above to some simulated examples, corresponding to the following models, where we took the  $\{\epsilon_i\}$  identically distributed as  $\epsilon$  (we use  $\text{Be}(p)$  to denote the Bernoulli( $p$ ) distribution):

(i)  $g(x) = 3x + 20 \exp\{-100(x - 0.5)^2\}/\sqrt{2\pi}$ ,  $X \sim \text{N}(0.5, 1.0/3.92^2)$ ,  $\epsilon \sim \text{N}(0, 0.673)$ ;

(ii)  $Y|X = x \sim \text{Be}\{g(x)\}$ ,  $g(x) = 0.45 \sin(2\pi x) + 0.5$  and  $X \sim \text{U}[0, 1]$ ;

(iii)  $g(x) = \sin(\pi x/2)/\{1 + 2x^2(\text{sgn } x + 1)\}$ ,  $X \sim \text{N}(0, 1)$ ,  $\epsilon \sim \text{N}(0, 0.09)$ .

In each case we took  $U^F \sim f_{U(1)}$ , and  $f_{U(1)}$  and  $f_{U(2)}$  to be either Laplace or normal with zero mean. We generated 200 samples of various sizes, and, for each calculated estimator, say  $\mu^{\text{est}}$ , we computed the integrated squared error,

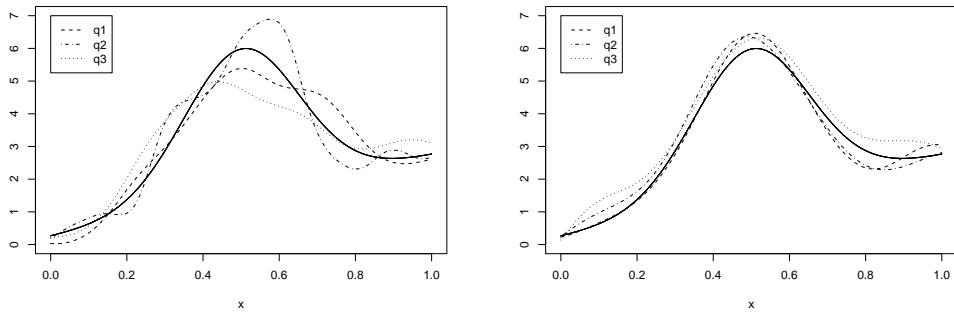


Figure 1: Quartile curves for the estimation of curve (i) when  $f_{U(1)} \sim N$ ,  $f_{U(2)} \sim \text{Laplace}$ ,  $\sigma_{U(1)}^2 = \sigma_{U(2)}^2 = 0.2 \text{var}(X)$ ,  $m = n/2 = 125$ , using  $\hat{\mu}_{\text{NW}}$  (left) or  $\hat{\mu}^*$  (right).

$\text{ISE} = \int (\mu^{\text{est}} - \mu)^2$ . In the graphs, to illustrate the performance of an estimator, we show the estimated curves corresponding to the first (q1), second (q2) and third (q3) quartiles of these calculated ISEs. In each case the target curve is represented by a solid curve. In the tables we provide the average values, denoted by MISE, of the 200 calculated ISEs.

In addition to the bandwidth  $h$ , necessary to calculate  $\tilde{\mu}$  and the classical Nadaraya-Watson estimator  $\hat{\mu}_{\text{NW}}$ , all methods, including  $\hat{\mu}$ , required the choice of a ridge parameter  $\rho$ , used in their denominators to avoid division by a number close to zero. For each method, at points  $x$  where the denominator of the estimator was smaller than  $\rho$ , we replaced it by  $\rho$ . For a given estimator  $\mu^{\text{est}}$ , we selected  $(\rho, h)$  — or  $\rho$  alone for  $\hat{\mu}$  — by minimizing the following cross-validation (CV) criterion:

$$\text{CV} = \sum_{j=1}^m \{Y_j - \mu^{\text{est},(-j)}(W_j)\}^2, \quad (5.1)$$

where the superscript  $(-j)$  meant that the estimator was constructed without using the  $j$ th observation.

Figure 1 and Table 1 compare, for various sample sizes, the results obtained for estimating curve (i) when  $f_{U(1)}$  was smoother than  $f_{U(2)}$ , with either both errors normal, or  $f_{U(1)}$  normal and  $f_{U(2)}$  Laplace. We compare  $\hat{\mu}_{\text{NW}}$ ,  $\hat{\mu}$  and  $\hat{\mu}^*$  (recall

Table 1: MISE for estimation of curve (i) when  $f_{U(1)} \sim \text{Normal (N)}$  and  $f_{U(2)} \sim \text{Laplace (L)}$ , with  $\sigma_{U(1)}^2 = \sigma_{U(2)}^2 = 0.2 \text{var}(X)$ ;  $f_{U(1)}$  and  $f_{U(2)} \sim \text{N}$ , with  $\sigma_{U(1)}^2 = 2\sigma_{U(2)}^2 = 0.2 \text{var}(X)$ ; and  $f_{U(1)}$  and  $f_{U(2)} \sim \text{N}$ , with  $2\sigma_{U(1)}^2 = \sigma_{U(2)}^2 = 0.2 \text{var}(X)$ . Results for  $\hat{\mu}^*$  and  $\tilde{\mu}^*$  are given within parenthesis.

		$f_{U(1)}$ smoother than $f_{U(2)}$		$f_{U(2)}$ smoother than $f_{U(1)}$		
		$f_{U(1)} \sim \text{N}, f_{U(2)} \sim \text{L}$	$f_{U(1)} \sim \text{N}, f_{U(2)} \sim \text{N}$	$f_{U(1)} \sim \text{N}, f_{U(2)} \sim \text{N}$		
$m$	$n$	Method	MISE	MISE	Method	MISE
125	250	$\hat{\mu}_{\text{NW}}$	0.352	0.317	$\hat{\mu}_{\text{NW}}$	0.310
500	1000		0.0859	0.0897		0.132
125	250	$\hat{\mu}$ at (2.8)	0.1290 (0.1363)	0.0897 (0.0985)	$\tilde{\mu}$ at (2.9)	0.294 (0.291)
500	1000		0.0284 (0.0304)	0.0221 (0.0242)		0.120 (0.121)

that the \* version of estimators is used when the error variances are estimated from replicated observations). Our results show that the estimator  $\hat{\mu}$  outperforms  $\hat{\mu}_{\text{NW}}$ . The  $\hat{\mu}^*$  version of  $\hat{\mu}$  worked almost as well as the latter, showing the limited loss incurred by estimating the error variances from replicated data. We also show the results obtained when  $f_{U(2)}$  was smoother than  $f_{U(1)}$  with both errors normal, where we compare  $\hat{\mu}_{\text{NW}}$ ,  $\tilde{\mu}$ , and  $\tilde{\mu}^*$ . Although the new estimator still outperforms the Nadaraya-Watson estimator, here the gain is less impressive, as predicted by the theory.

Figure 2 and Table 2 show the results obtained for estimating curve (ii) when  $f_{U(1)}$  was smoother than  $f_{U(2)}$ , with both errors normal. We compare  $\hat{\mu}_{\text{NW}}$ ,  $\hat{\mu}$  and  $\hat{\mu}^*$  for different combinations of sample sizes. Of course, the situation where we can expect the largest gain by using the new estimator, compared to the classical Nadaraya-Watson estimator, is that where the size,  $m$ , of the sample of data contaminated by  $f_{U(1)}$  is as small as possible, relative to the total sample size,  $n$ . The results, however, indicate that, even when  $m = 5n/6$ , the gain can already be quite significant. In this example,  $\hat{\mu}^*$  performed so well that it even bettered its known error version,  $\hat{\mu}$ , in the majority of cases.

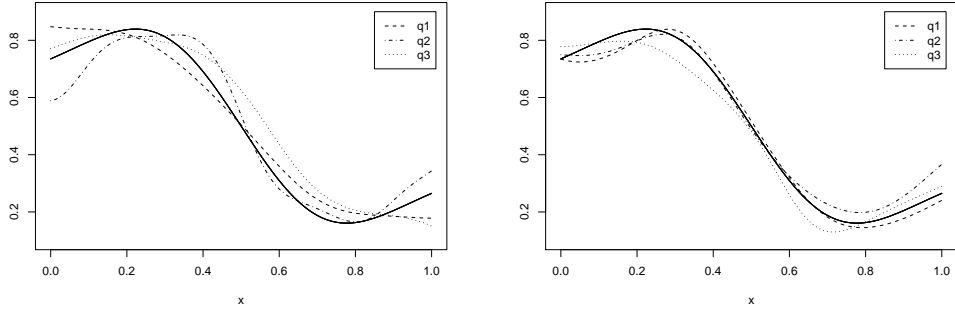


Figure 2: Quartile curves for the estimation of curve (ii) when  $f_{U(1)} \sim N$ ,  $f_{U(2)} \sim N$ ,  $\sigma_{U(1)}^2 = 2\sigma_{U(2)}^2 = 0.2 \text{ var}(X)$  and  $m = n/2 = 250$ , using  $\hat{\mu}_{\text{NW}}$  (left) or  $\hat{\mu}^*$  (right).

Table 2: MISE for estimation of curve (ii) when  $f_{U(1)}$  and  $f_{U(2)} \sim \text{Normal}$  with  $\sigma_{U(1)}^2 = 2\sigma_{U(2)}^2 = 0.2 \text{ var}(X)$ .

Method	$m$	$n$	MISE	$m$	$n$	MISE	$m$	$n$	MISE
$\hat{\mu}_{\text{NW}}$	125	250	0.00891	50	250	0.02118	200	250	0.00496
	250	500	0.00418	100	500	0.00967	400	500	0.00285
$\hat{\mu}$	125	250	0.00312	50	250	0.00499	200	250	0.00373
	250	500	0.00157	100	500	0.00155	400	500	0.00190
$\hat{\mu}^*$	125	250	0.00406	50	250	0.00304	200	250	0.00366
	250	500	0.00156	100	500	0.00151	400	500	0.00186

Finally, Figure 3 compares the results obtained for estimating curve (iii) when both errors are normal with  $\sigma_{U(1)}^2 = 2\sigma_{U(2)}^2 = 0.2 \text{ var}(X)$  and  $m = n/2 = 250$ . We show the quartile curves obtained for  $\hat{\mu}_{\text{NW}}$  and  $\hat{\mu}^*$ . Again, we see the important gain that can be obtained when using the new estimator compared to the classical Nadaraya-Watson estimator, which uses only  $(W_1, Y_1), \dots, (W_m, Y_m)$ .

In summary, our simulations showed that when  $f_{U(1)}$  was smoother than  $f_{U(2)}$ , the new estimator substantially outperformed the Nadaraya-Watson estimator. When  $f_{U(2)}$  was smoother than  $f_{U(1)}$ , the gain from using the new estimator was usually less impressive, unless  $m$  was relatively small, as predicted by the theoretical results in Section 4. The empirical applications we had in mind when developing the new estimators fall into the category where the  $f_{U(1)}$  is smoother than  $f_{U(2)}$ , see Section 5.2.

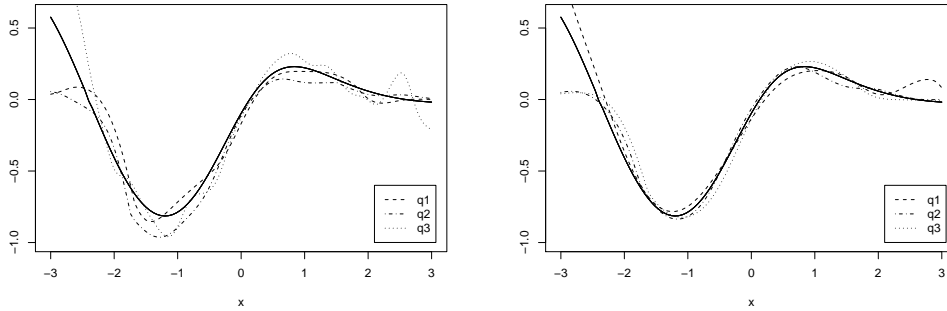


Figure 3: Quartile curves for the estimation of curve (iii) when  $f_{U(1)} \sim N$ ,  $f_{U(2)} \sim N$ ,  $\sigma_{U(1)}^2 = 2\sigma_{U(2)}^2 = 0.2 \text{ var}(X)$ ,  $m = n/2 = 250$ , using  $\hat{\mu}_{\text{NW}}$  (left) or  $\hat{\mu}^*$  (right).

## 5.2 Data Illustration

In part, this paper arises from the following considerations. In nutritional epidemiology, the standard method for correcting for the effects of measurement error in evaluating diet-disease relationships is regression calibration (Carroll, et al., 2006). Using our notation, the method works as follows. Let  $N$  be unobserved true long-term nutrient intake. The goal is to regress a response, say disease status  $D$ , on  $N$ . In the main study, nutrient intakes are measured by a single food frequency questionnaire (FFQ), which is what we call  $W$ . Here  $X$  is the long-term average intake as measured by the FFQ.

Because  $N$  is not observed, most nutritional epidemiology studies take a calibration random sub-sample of the main study population, generally much smaller than the main sample, where they typically measure repeated versions of  $W$ , in an effort to understand the measurement error properties of  $W$  in the sampled population. In addition, in the sub-sample, they observe an unbiased estimate  $Y$  of  $N$ . In regression calibration, instead of regressing  $D$  on the unobserved  $N$ , one regresses  $D$  on  $E(Y|W)$ , where  $E(Y|W)$  is estimated from the observations in the sub-sample. Of course, one way to estimate  $E(Y|W)$  would be to use the classical Nadaraya-Watson estimator of  $E(Y|W)$  based on the direct observations

on  $(W, Y)$ , but our new approach can be used with the averaged replicated data to obtain a more efficient estimator of  $E(Y|W)$ , as we illustrate below, on a calibration sub-study from the American Cancer Society Cancer Prevention Study II Nutrition Survey (ACS, Flagg et al., 2000).

The main study had approximately 185,000 adults, while the calibration sub-study was of size 598. In the calibration sub-study, several variables were measured, including  $Y$ , an average of protein intake from four food records which is taken to be unbiased for usual intake  $N$ , and  $W$ , a log-transformed version of protein intake using a FFQ, which was measured twice with error approximately normal  $N(0, \sigma_U^2)$ . As above,  $X$  is the unobserved long-term average intake as measured by the FFQ. The data we considered were a sample of size  $n = 598$  from  $(W_{i1}, W_{i2}, Y_i)$ , for  $i = 1, \dots, n$ , where, for each  $i$ ,  $W_{i1} = X_i + U_{i1}$  and  $W_{i2} = X_i + U_{i2}$ , with  $U_{i1}$  and  $U_{i2}$  independent and identically distributed as  $N(0, \sigma_U^2)$ . Our target is  $E(Y|W)$ . A point to note here is that we have transformed  $W$  to make the measurement errors normally distributed, but we have not transformed  $Y$ , the idea being that disease risk models are interested in the effects of nutrient intakes and not transformed intakes, see Ferrari, et al. (2004, 2008) for examples of this.

This example is convenient for illustrating the various approaches to regression estimation, since the fact that we have direct data from the quantity of interest allows us to consider three different estimators: the classical Nadaraya-Watson estimator  $\hat{\mu}_{NW}$  of  $E(Y|W)$  based on the independent data  $(W_{i1}, Y_i)$ , for  $i = 1, \dots, n$ , the Nadaraya-Watson estimator  $\hat{\mu}_{NW,dep}$  of  $E(Y|W)$  based on the dependent data  $(W_{i1}, Y_i)$ ,  $(W_{i2}, Y_i)$ , for  $i = 1, \dots, n$ , and our new methodology  $\hat{\mu}^*$ , based on the averaged data  $(\overline{W}_i, Y_i)$ , for  $i = 1, \dots, n$ , where  $\overline{W}_i = (W_{i1} + W_{i2})/2$  and  $\sigma_U^2$  is estimated by half the empirical variance of the differenced replicates. To

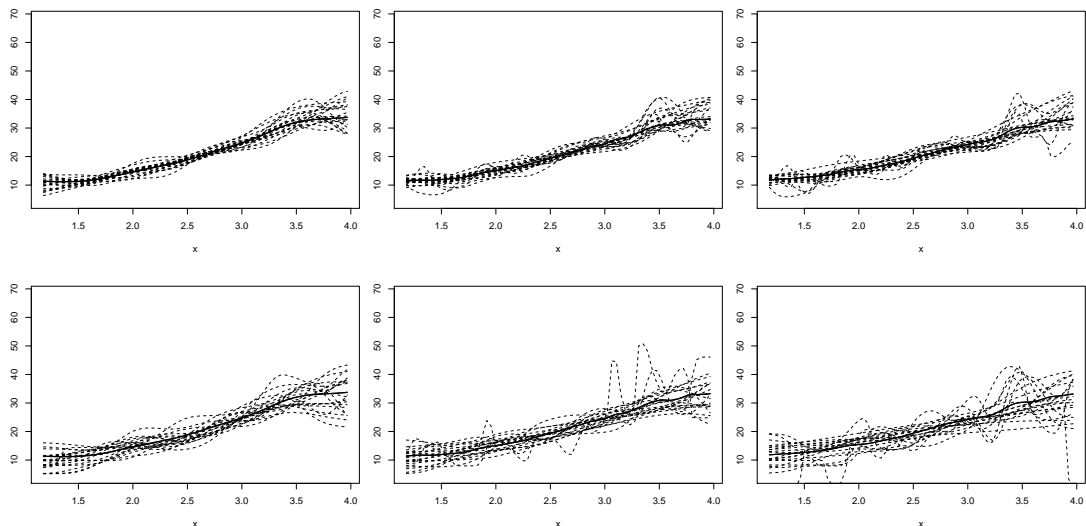


Figure 4: Estimation of  $\mu$  for the American Cancer data, using the new estimator  $\hat{\mu}^*$  (left), the classical NW estimator  $\hat{\mu}_{\text{NW,dep}}$  (middle) with dependent data, or the classical NW estimator  $\hat{\mu}_{\text{NW}}$  (right), for subsamples of size 100 (row 1) or 30 (row 2). The dashed curves show estimators corresponding to 15 subsamples randomly selected among 500 generated subsamples. The thick curves show the respective estimators when using all 598 observations.

Table 3: *Integrated variances (IVAR) for estimation of  $E(Y|W)$  in the empirical example, for various subsample sizes  $n$ .*

Method	$n$	IVAR	$n$	IVAR	$n$	IVAR	$n$	IVAR
$\hat{\mu}_{\text{NW}}$	30	129.933	50	117.645	75	106.527	100	92.440
$\hat{\mu}_{\text{NW,dep}}$	30	124.517	50	112.512	75	99.624	100	75.746
$\hat{\mu}^*$	30	116.875	50	92.934	75	66.149	100	51.018

use the notation of the previous sections, the last approach corresponds to  $m = 0$  (since we do not use the direct data),  $f_{U^{(1)}} \sim N(0, \sigma_U^2)$  and  $f_{U^{(2)}} \sim N(0, \sigma_U^2/2)$ . Clearly,  $f_{U^{(1)}}$  is smoother than  $f_{U^{(2)}}$ , so we can use our estimator  $\hat{\mu}$ .

Of course, here we do not know the true curve  $E(Y|W)$ , so we cannot say which method gives the best estimator. However, the sample size is large, so one way to illustrate the performance of the procedures in a way that is similar to a simulation study is to create a large number (we took 500) of subsamples of smaller size (we took 30, 50, 75 and 100), and examine the variability of

each method, for each subsample size. It is not hard to show that, for our method, the squared bias is of smaller order than the variance, and since we do not know the true target, it thus seems appropriate to focus on variance. In Figure 4 we show the estimated curves for 15 subsamples of size 30 (respectively, 100) randomly selected from the 500 randomly created subsamples of size 30 (respectively, 100). We see that, although all methods indicate the same trend for the relationship between  $W$  and  $Y$ , both versions of the Nadaraya-Watson estimator experience some difficulty, as some of the estimated curves are quite wiggly. To illustrate this further, in Table 3 we show, for each subsample size, the integrated variance of each method on the interval  $[1, 4]$  (calculated via the variance of the 500 replications in each case). The main message is the same: our method is less variable than both Nadaraya-Watson estimators, as expected by the theory.

## 6 Conclusion

We have shown how to predict in errors-in-variables regression, when information from different sources (e.g. different laboratories) is combined, and the errors have different distributions. The methods that we suggest enjoy optimal accuracy, in the sense that the rates of convergence are best possible. However, those rates can vary particularly widely, from root- $n$  rates when the problem is in effect semiparametric, to much slower rates that are characteristic of a genuinely nonparametric problem.

The problem turns out to be quite complex in other ways, too, and has a number of subtle features and apparent contradictions. For example, the results superficially suggest that on occasion it might even be beneficial to artificially add noise to some of the data. However, as explained in Remark 4.3, such a

conclusion is unwarranted because it does not take account of the way in which adding noise would affect the conditioning step.

Our methods can also be applied in settings where the error distributions are not known and are instead estimated, for example from repeated measurements, see section 3. The methodology developed there can be taken further. This, and other practically important variants of the problem, offer interesting avenues for further research.

## References

- Berry, S., Carroll, R.J. and Ruppert, D. (2002). Bayesian Smoothing and Regression Splines for Measurement Error Problems. *J. Amer. Statist. Assoc.* **97**, 160–169.
- Carroll, R.J. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *J. Amer. Statist. Assoc.* **83**, 1184–1186.
- 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: A Modern Perspective*, Second Edition. Chapman and Hall CRC Press.
- Cheng, C-L. and Riu, J. (2006). On estimating linear relationships when both variables are subject to heteroscedastic measurement errors. *Technometrics* **48**, 511–519.
- Cook, J.R. and Stefanski, L.A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *J. Amer. Statist. Assoc.* **89**, 1314–1328.
- Delaigle, A. (2007). Nonparametric density estimation from data with a mixture

- of Berkson and classical errors. *Canad. J. Statist.* **35**, 89–104.
- Delaigle, A., Fan, J. and Carroll, R.J. (2008). Local polynomial estimator for the errors-in-variables problem. *Submitted for publication*.
- Delaigle, A. Hall, P. and Meister, A. (2008). On deconvolution with repeated measurements. *Ann. Statist.* **36**, 665–685.
- Delaigle, A., and Meister, A. (2007). Nonparametric regression estimation in the heteroscedastic errors-in-variables problem. *J. Amer. Statist. Assoc.* **102**, 1416–1426.
- Delaigle, A., and Meister, A. (2008). Density estimation with heteroscedastic error. *Bernoulli*, **14**, 562–579.
- 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. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.* **19**, 1257–1272.
- Fan, J. and Truong, Y.K. (1993). Nonparametric regression with errors in variables. *Ann. Statist.* **21**, 1900–1925.
- Ferrari, P., Kaaks, R., Fahey, M. T., Slimani, N., Day, N. E., Pera, G., Boshuizen, H. C., Roddam, A., Boeing, H., Nagel, G., Thiebaut, A., Orfanos, P., Krogh, P., Braaten, T. and Riboli, E. (2004). Within- and between-cohort variation in measured macronutrient intakes, taking account of measurement errors, in the European Prospective Investigation into Cancer and Nutrition Study. *American Journal of Epidemiology* **160**, 814–822.
- Ferrari, P., Day, N. E., Boshuizen, H. C., Roddam, A., Hoffmann, K., Thiebaut, A., Pera, G., Overvad, K., Lund, E., Trichopoulou, A., Tumino, R.,

- Gullberg, A., Norat, T., Slimani, N., Kaaks, R. and Riboli, E. (2008). The evaluation of the diet/disease relation in the EPIC study: considerations for the calibration and the disease models. *International Journal of Epidemiology*, Advance Access published January 6, 2008.
- Flagg, E.W., Coates, R.J., Calle, E.E., Potischman, N. and Thun, M. (2000). Validation of the American Cancer Society Cancer Prevention Study II Nutrition Survey Cohort Food Frequency Questionnaire. *Epidemiology*, **11**, 462–468.
- Ganase, R.A., Amemiya, Y. and Fuller, w. a. (1983). Prediction when both variables are subject to error, with application to earthquake magnitudes. *J. Amer. Statist. Assoc.* **78**, 761–765.
- Hall, P. and Meister, A. (2007). A ridge-parameter approach to deconvolution. *Ann. Statist.* **35**, 1535–1558.
- Hu, Y. and Schennach, S.M. (2008). Identification and estimation of nonclassical nonlinear errors-in-variables models with continuous distributions. *Econometrica* **76**, 195–216.
- Huang, X.Z., Stefanski, L.A. and Davidian, M. (2006). Latent-model robustness in structural measurement error models. *Biometrika* **93**, 53–64.
- Kim, J. and Gleser, L.J. (2000). SIMEX approaches to measurement error in ROC studies. *Comm. Statist. Theory Meth.* **29**, 2473–2491.
- Kipnis, V., Subar, A.F., Midthune, D., Freedman, L.S., Ballard-Barbash, R., Troiano, R. Bingham, S., Schoeller, D.A., Schatzkin, A. and Carroll, R.J. (2003). The structure of dietary measurement error: results of the OPEN biomarker study. *Amer. J. Epidemiology* **158**, 14–21.
- Kulathinal, S.B., Kuulasmaa, K., and Gasbarra, D. (2002). Estimation of an errors-in-variables regression model when the variances of the measure-

- ment errors vary between the observations. *Statist. Medicine* **21**, 1089–1101.
- Li, T. and Vuong, Q. (1998). Nonparametric estimation of the measurement error model using multiple indicators. *J. Multivariate Anal.* **65**, 139–165.
- Linton, O. and Whang, Y.J. (2002). Nonparametric estimation with aggregated data. *Econometric Theory* **18**, 420–468.
- Marron, J.S. and Wand, M.P. (1992). Exact mean integrated squared error. *Ann. Statist.* **20**, 712–736.
- National Research Council, Committee on Pesticides in the Diets of Infants and Children. (1993). *Pesticides in the Diets of Infants and Children*. National Academies Press.
- Schennach, S.M. (2004a). Estimation of nonlinear models with measurement error. *Econometrica* **72**, 33–75.
- Schennach, S.M. (2004b). Nonparametric regression in the presence of measurement error. *Econometric Theory* **20**, 1046–1093.
- Simonoff, J.S. (1996). *Smoothing Methods in Statistics*. Springer, New York.
- Staudenmayer, J. and Ruppert, D. (2004). Local polynomial regression and simulation-extrapolation. *J. Roy. Statist. Soc., Ser. B* **66**, 17–30
- Staudenmayer, J. and Ruppert, D. and Buonaccorsi, J. (2008). Density estimation in the presence of heteroskedastic measurement error. *J. Amer. Statist. Assoc.*, to appear.
- Stefanski, L. A. (2000). Measurement error models. *J. Amer. Statist. Assoc.* **95**, 1353–1358.
- Stefanski, L. A. and Carroll, R.J. (1990). Deconvoluting kernel density estimators. *Statistics* **21**, 165–184.
- Taupin, M.L. (2001). Semi-parametric estimation in the nonlinear structural

errors-in-variables model. *Ann. Statist.* **29**, 66–93.

Thamerus, M. (2003). Fitting a mixture distribution to a variable subject to heteroscedastic measurement errors. *Comput. Statist.* **18**, 1–17.

Wand, M.P. and Jones, M.C. (1995). *Kernel Smoothing*. Chapman and Hall, London.