Nonparametric Regression Estimation in the Heteroscedastic Errors-in-Variables Problem

Aurore DELAIGLE and Alexander MEISTER

In the classical errors-in-variables problem, the goal is to estimate a regression curve from data in which the explanatory variable is measured with error. In this context, nonparametric methods have been proposed that rely on the assumption that the measurement errors are identically distributed. Although there are many situations in which this assumption is too restrictive, nonparametric estimators in the more realistic setting of heteroscedastic errors have not been studied in the literature. We propose an estimator of the regression function in such a setting and show that it is optimal. We give estimators in cases in which the error distributions are unknown and replicated observations are available. Practical methods, including an adaptive bandwidth selector for the errors-in-variables regression problem, are suggested, and their finite-sample performance is illustrated through simulated and real data examples.

KEY WORDS: Bandwidth selector; Deconvolution; Errors-in-variables; Heteroscedastic contamination; Inverse problem; Regression; Replicated measurement.

1. INTRODUCTION

We consider nonparametric estimation of a regression function m from a sample in which the covariate contains random measurement error. Suppose that the observations are a sample of independent and identically distributed (iid) random vectors $(W_1, Y_1), \ldots, (W_n, Y_n)$ generated by the model

$$Y_j = m(X_j) + \eta_j, \qquad W_j = X_j + U_j, \qquad \mathrm{E}(\eta_j | X_j) = 0, \quad \text{with } X_j \sim f_X \text{ and } U_j \sim f_U, \quad (1)$$

where U_j are the error variables, independent of (X_j, Y_j, η_j), and f_U is known. Under model (1), this estimation problem, referred to as an errors-in-variables problem, has received considerable attention in the literature. (References, not restricted to the nonparametric case, include Fan and Masry 1992; Cook and Stefanski 1994; Stefanski and Cook 1995; Carroll, Maca, and Ruppert 1999; Taupin 2001; Berry, Carroll, and Ruppert 2002; Carroll and Hall 2004; and Liang and Wang 2005; see also Carroll, Ruppert, Stefanski, and Crainiceanu 2006 for an exhaustive review of this problem). In this context, Fan and Truong (1993) proposed a nonparametric estimator of $m(x) = E(Y|X = x)$ that is valid when for all t, $f_U^{\mathrm{ft}}(t) \neq 0$, with g^{ft} denoting the Fourier transform of a function g. Let K be a square-integrable kernel function and $h > 0$ be a smoothing bandwidth parameter. The nonparametric estimator of m is defined by

$$\widehat{m}(x) = (nh)^{-1} \sum_{j=1}^{n} Y_j K_U\left(\frac{x - W_j}{h}\right) \Big/ \widehat{f_n}(x), \qquad (2)$$

where $\widehat{f_n}(x) = (nh)^{-1} \sum_{j=1}^{n} K_U(\frac{x-W_j}{h})$ is the deconvolution kernel density estimator of f_X of Carroll and Hall (1988) and Stefanski and Carroll (1990), with $K_U(x) = (2\pi)^{-1} \times \int e^{-itx} K^{\mathrm{ft}}(t)/f_U^{\mathrm{ft}}(t/h)\, dt$.

In real data applications, there are many examples in which it is not realistic to assume that the errors U_i are homoscedastic. In practice, heteroscedasticity arises as soon as the observations are obtained in nonhomogeneous conditions. For example, the sample might have been obtained by collating data from different laboratories (see, e.g., National Research Council 1993) or from different studies (meta-analysis; see Walter 1997). Groups of individuals (e.g., healthy/unhealthy, smoker/nonsmoker) might be subject to a different contamination process (see Fuller 1987 for an early consideration of this problem), and the measurement process might be subjective and differ among all individuals (see Bennett and Franklin 1954 for an example related to subjective assessment of iron content of substances by different students). (See also Riu and Rius 1995, 1996; Kulathinal, Kuulasmaa, and Gasbarra 2002; Thamerus 2003; and Cheng and Riu 2006.) In such instances, we alter model (1) to

$$Y_j = m(X_j) + \eta_j, \qquad W_j = X_j + U_j, \qquad \mathrm{E}(\eta_j | X_j) = 0, \quad \text{where } X_j \sim f_X \text{ and } U_j \sim f_{U_j}, \quad (3)$$

with U_j independent of (X_j, Y_j, η_j); the error densities f_{U_j} may depend on both the observation number j and the sample size n. Then the estimator (2) cannot be applied, because it uses only one error density f_U in its construction. Despite its numerous applications and the attention that it has received in the parametric literature, the heteroscedastic problem has not yet been considered in the nonparametric literature. In Section 2 we introduce a kernel estimator of the function m that can be applied for heteroscedastic errors when the error distributions are known. We show that the estimator is consistent and achieves optimal convergence rates under smoothness and regularity constraints.

The classical deconvolution methods for nonparametric estimation of the regression function with errors in variables rely on the fact that the error distribution is known (e.g., Fan and Truong 1993). But this assumption is unrealistic in many practical situations. Some recent articles have considered deconvolution with unknown error density when replicates are observed in the homoscedastic error setting (e.g., Horowitz and Markatou 1996; Li and Vuong 1998; Hall and Yao 2003; Delaigle, Hall, and Meister 2007a; Neumann 2007). We show in Section 3 that even in the heteroscedastic case, the regression function can still be estimated consistently if such additional observations are available. We propose consistent estimators of the

Aurore Delaigle is Lecturer, Department of Mathematics, University of Bristol, Bristol BS8 1TW, UK, and Research Fellow, Department of Mathematics and Statistics, University of Melbourne, VIC, 3010, Australia (E-mail: *Aurore.Delaigle@bristol.ac.uk*). Alexander Meister is Postdoctoral Researcher, Graduiertenkolleg 1100, Universität Ulm, D-89081 Ulm, Germany (E-mail: *alexander.meister@uni-ulm.de*). Delaigle's research was supported by a Hellman Fellowship and a Maurice Belz Fellowship. The authors thank the joint editors, the associate editor, and two referees for their valuable comments that helped improve a previous version of the manuscript.

Journal of the American Statistical Association
December 2007, Vol. 102, No. 480, Theory and Methods
DOI 10.1198/016214507000000987

regression function m in the case where the observations are replicated at least once.

We study finite-sample performance of the procedures in Section 4. In practice, the estimator involves selection of a bandwidth parameter from the data. This is extremely complicated in errors-in-variable problems, because classical methods, such as cross-validation, usually cannot be implemented. The method that we propose is based on alternative ideas related to SIMEX and bootstrap procedures and also can be used in the homoscedastic case to improve a first approximate method proposed in the literature. We show in simulated examples that the method works well, illustrate its use on a real data example from a study on coronary heart disease, and compare the results with those from an alternative SIMEX regression estimator.

2. ESTIMATION PROCEDURE

Suppose that we want to estimate a regression curve m from a sample of independent random vectors $(W_1, Y_1), \ldots, (W_n, Y_n)$ generated by the heteroscedastic error model (3). In this context, a natural generalization of (2) leads to the estimator

$$\widehat{m}(x) = (nh)^{-1} \sum_{j=1}^{n} Y_j K_{U_j}\left(\frac{x - W_j}{h}\right) \Big/ \widehat{f}_n(x), \tag{4}$$

where $K_{U_j}(x) = (2\pi)^{-1} \int e^{-itx} K^{\mathrm{ft}}(t) / f_{U_j}^{\mathrm{ft}}(t/h)\, dt$ and $\widehat{f}_n(x) = (nh)^{-1} \sum_{j=1}^{n} K_{U_j}(\frac{x-W_j}{h})$. Careful analysis of the properties of (4), however, shows that its rates of convergence are dictated by the least favorable errors U_j, which makes it unacceptable. The alternative estimator that we propose is defined by

$$\widehat{m}_n(x) = h^{-1} \sum_{j=1}^{n} Y_j K_{U_j}\left(\frac{x - W_j}{h}\right) \Big/ \widehat{f}_n(x), \tag{5}$$

where $\widehat{f}_n(x) = h^{-1} \sum_{j=1}^{n} K_{U_j}(\frac{x-W_j}{h})$ is an estimator of f_X that is valid when the errors are heteroscedastic, with

$$K_{U_j}(x) = (2\pi)^{-1} \int e^{-itx} K^{\mathrm{ft}}(t) \Psi_j(t/h)\, dt,$$

and where $\Psi_j(t) = f_{U_j}^{\mathrm{ft}}(-t) / (\sum_{k=1}^{n} |f_{U_k}^{\mathrm{ft}}(t)|^2)$ generalizes the term $\{n f_U^{\mathrm{ft}}(t)\}^{-1}$ used in the homoscedastic setting. Let $\tau^2(x) = \mathrm{var}(Y|X = x)$. The estimator is well defined under the following conditions:

Condition A.

(A1) There is a j such that $|f_{U_j}^{\mathrm{ft}}(t)| \neq 0$ for all $t \in \mathbb{R}$.

(A2) $K \in L_1(\mathbb{R}) \cap L_2(\mathbb{R})$ is such that K^{ft} is supported on $[-1, 1]$; $|K^{\mathrm{ft}}(t) - 1| \downarrow 0$ as $t \to 0$.

(A3) $f_X(x) \neq 0$.

(A4) $\tau^2 f_X$, f_X, and $m^2 f_X$ are bounded and continuous and $(m f_X)^{\mathrm{ft}}$, $f_X^{\mathrm{ft}} \in L_1(\mathbb{R})$.

These conditions are rather standard in deconvolution and errors-in-variables problems (see, e.g., Fan and Truong 1993 and references therein). The only difference between the homoscedastic and the heteroscedastic cases is condition (A1); whereas under homoscedastic contamination, f_U^{ft} is usually assumed to vanish nowhere, only one of the $f_{U_j}^{\mathrm{ft}}$'s is required to have no 0's for our estimator to be well defined.

Note that our estimator can be applied generally whether or not the errors are heteroscedastic, because it reduces to the classical estimator (2) when the errors are homoscedastic. In particular, all of the results obtained in this article, including the bandwidth selection procedure, can be applied directly to the classical estimator (2).

2.1 Theoretical Properties

First, we study pointwise consistency of the estimator (5) under general conditions. [See Stone 1977 for general consistency in the standard (noncontaminated) nonparametric regression problem.] Although for the estimator to be well defined, only one $f_{U_j}^{\mathrm{ft}}$ must not vanish anywhere, for the estimator to be consistent, we need to ensure that a sufficient number of such functions are nonzero. More precisely, we assume that for almost all t,

$$\begin{aligned} &\sum_{j=1}^{n} |f_{U_j}^{\mathrm{ft}}(t)|^2 \to \infty, \quad \text{as } n \to \infty \qquad \text{and} \\ &\inf_n \sum_{j=1}^{n} |f_{U_j}^{\mathrm{ft}}(t)|^2 \;>\; 0, \qquad \forall t. \end{aligned} \tag{6}$$

In the homoscedastic case (e.g., where $f_{U_j}^{\mathrm{ft}} = f_U^{\mathrm{ft}}$), this condition is satisfied under the usual assumption that $f_U^{\mathrm{ft}}(t)$ does not vanish anywhere. The condition is less restrictive in the heteroscedastic case, because some of the $f_{U_j}^{\mathrm{ft}}(t)$'s are allowed to be equal to 0. We also consider strong consistency of our estimator. There the following more restrictive version of (6) is required: There exist some $\delta > 1$ and $\kappa > 0$ such that

$$\begin{aligned} &0 \leftarrow h \geq c \cdot n^{1-\delta+\kappa} \qquad \text{and} \\ &\int_{|t| \leq 1/h} \left(\sum_{j=1}^{n} \left| f_{U_j}^{\mathrm{ft}}(t) \right|^2 \right)^{-2} dt = O(n^{-\delta}). \end{aligned} \tag{7}$$

Note that it is possible to select the bandwidth h so that (7) is satisfied if for each $\xi > 0$, we have $\inf_n n^{-\alpha} \times \inf_{|t| \leq \xi} \sum_{j=1}^{n} |f_{U_j}^{\mathrm{ft}}(t)|^2 > 0$ for some $\alpha > \delta/2$ [put $\kappa = (\delta - 1)/2$]. We are now ready to establish general pointwise consistency of our estimator.

Theorem 1. Assume Condition A and (6). Then the following results hold:

a. If $h \to 0$ and $\int_{|t| \leq 1/h} (\sum_{j=1}^{n} |f_{U_j}^{\mathrm{ft}}(t)|^2)^{-1}\, dt \to 0$, then

$$\widehat{m}(x) \xrightarrow{P} m(x) \quad \text{as } n \to \infty.$$

b. Also assume that $\|m\|_\infty < \infty$. Under (7), if $E|\eta_j|^l \leq C_l < \infty$ for all integers j and all $0 < l \leq 2\lceil 1/\kappa \rceil + 2$, then

$$\widehat{m}(x) \xrightarrow{\text{a.s.}} m(x) \quad \text{as } n \to \infty.$$

Here we use $\lceil a \rceil$ to denote the smallest integer larger or equal to a.

Next, we derive the pointwise rates of convergence of the estimator $\widehat{m}(x)$ at an arbitrary but fixed $x \in \mathbb{R}$, and show that these are optimal in a minimax sense with respect to any regression estimator in model (3). We consider a weak version of consistency similar to that of Fan and Truong (1993). To

determine the precise behavior of the estimator, we need to quantify the smoothness degree of m and f_X. Toward this end, we fix $x \in \mathbb{R}$ arbitrary and define the following class of functions: $\mathcal{F}_{\beta,C,D} \equiv \{g \in \mathcal{C}^0$ s.t. g and g^{ft} are integrable on $\mathbb{R}$ and $|g^{(\beta)}(y)| \le C, \forall y \in (x-D, x+D)\}$, where $g^{(\beta)}$ denotes the βth derivative of g. We need the following additional assumptions on f_X, m, and K:

Condition B.

(B1) mf_X is continuous and integrable on $\mathbb{R}$, and $mf_X \in \mathcal{F}_{\beta,C_1,D}$ for some $C_1, \beta, D > 0$.
(B2) $f_X \in \mathcal{F}_{\beta,C_1,D}$ with β, D, and C_1 as in (B1).
(B3) $\|m^2 f_X\|_\infty \le C_2$, $\|f_X\|_\infty \le C_3$, and $\|\tau^2 f_X\|_\infty \le C_4$.
(B4) $\int K(y)\,dy = 1$, $\int y^j K(y)\,dy = 0 \ \forall j = 1, \ldots, k$, $\int |y|^k K(y)\,dy < \infty$ for some $k \ge 2\beta$, and K^{ft} is supported on $[-1, 1]$ with β as in (C1).
(B5) $|m(x)| \le C_7$ and $f_X(x) \ge C_8 > 0$.

Note that in (B5), boundedness of m and f_X is needed only at the fixed x, where $m(x)$ is estimated. We also need some conditions on the error distributions. Suppose that there exist α, $C > 0$ and some positive monotonously decreasing functions $\overline{\varphi}_{j,n}(t)$ and $\underline{\varphi}_{j,n}(t)$ such that the following conditions hold:

Condition C.

(C1) $P(|U_j| \le \alpha) \ge C, \forall j, n$.
(C2) $|f_{U_j}^{\text{ft}}(t)| \ge \underline{\varphi}_{j,n}(T), \forall |t| \le T$.
(C3) $\underline{\varphi}_{j,n}(t) \le |f_{U_j}^{\text{ft}}(t)| \le \overline{\varphi}_{j,n}(t), \forall t > T$.
(C4) $|f_{U_j}^{\text{ft}\,\prime}(t)| \le \overline{\varphi}_{j,n}(t), \forall t > T$.
(C5) $\underline{\varphi}_{j,n}(t) \ge c_1 \cdot \overline{\varphi}_{j,n}(c_2 t), \forall t > 0$.

These conditions assume some $T \ge 0$, $c_1 > 0$, and $c_2 \ge 1$ that do not depend on j and n. Note that (C1) gives a regularity condition to prevent the densities f_{U_j} from spreading too heavily and becoming too smooth as j increases; conditions (C2)–(C5) represent a weak version of monotonicity for $|f_{U_j}^{\text{ft}}|$. In particular, the so-called "ordinary smooth densities" f_{U_j}, in the terminology of Fan (1991a,b), satisfy $\underline{\varphi}_{j,n}(t) = C_5|t|^{-\nu}$ and $\overline{\varphi}_{j,n}(t) = C_5'|t|^{-\nu}$, and accordingly, for the supersmooth densities we have $\underline{\varphi}_{j,n}(t) = C_5|t|^{\rho_1}\exp(-c|t|^\gamma)$ and $\overline{\varphi}_{j,n}(t) = C_5'|t|^{\rho_2}\exp(-c|t|^\gamma)$.

Finally, we define the class $\mathcal{F}$ containing all pairs (m, f_X) satisfying Conditions B and C with uniform constants and parameters. Now we are able to derive rate optimality for our estimator (5). In the sequel, "const" denotes an arbitrary positive constant.

Theorem 2. Fix an arbitrary $x \in \mathbb{R}$. Let Conditions B and C hold. Assume that there is a sequence $a_n \uparrow \infty$ such that for some $C_{11} \ge C_{10} > 0$, $\beta > 0$,

$$C_{10} a_n^{1+2\beta} \le \sum_{j=1}^{n} |\overline{\varphi}_{j,n}(a_n)|^2 \le C_{11} a_n^{1+2\beta} \tag{8}$$

is valid for all n. Then,

(a) When putting $h = c_2 a_n^{-1}$ [with c_2 as in (C5)], the estimator $\widehat{m}_n$ satisfies

$$\limsup_{n\to\infty} \sup_{(m,f_X)\in\mathcal{F}} P\left(|\widehat{m}_n(x) - m(x)|^2 > d a_n^{-2\beta}\right) \le \text{const} \cdot d^{-1}, \qquad \forall d > 0.$$

(b) For an arbitrary estimator $\tilde{m}(x) = \tilde{m}_n(x; (W_1, Y_1), \ldots, (W_n, Y_n))$, and for sufficiently large constants C and D in $\mathcal{F} = \mathcal{F}_{\beta,C,D}$, there is some $C_{12} > 0$ such that

$$\liminf_{n\to\infty} \sup_{(m,f_X)\in\mathcal{F}} P\left(|\tilde{m}(x) - m(x)|^2 > C_{12} a_n^{-2\beta}\right) \ge \text{const}.$$

3. THE CASE OF UNKNOWN ERROR DENSITIES

Although in traditional nonparametric deconvolution settings, the error density is usually assumed known, this condition seems unrealistic in many real life situations. When the error distribution is unknown, it is not possible to consistently estimate the functions m and f_X unless extra observations are available. We focus on the problem in which each X_j is repeatedly measured with independent noise, because there is significant potential for obtaining replicated observations on individuals, as reflected in the vast parametric and semiparametric literature, where the error variance is often estimated from replicated observations (see, e.g., Madansky 1959; Carroll, Eltinge, and Ruppert 1993; Stefanski and Bay 1996; Carroll et al. 1999, 2006, and references therein). Recent related work in the econometrics literature includes that of Horowitz and Markatou (1996), Li (2002), Li and Hsiao (2004), and Schennach (2004a,b). Replicated data are also drawing increased interest in the nonparametric literature (see, e.g., Li and Vuong 1998; Linton and Whang 2002; Susko and Nadon 2002; Hall and Yao 2003; Delaigle et al. 2007a; Delaigle, Hall, and Müller 2007b; Neumann 2007).

Replicated data subject to normally distributed heteroscedastic errors have been considered by Devanarayana and Stefanski (2002), who used a parametric SIMEX method to estimate the regression curve. In our context, we treat the problem according to the type of observations available. We distinguish between two cases.

Case I: Groups of Observations. If the observations can be gathered in a small number G [$G = o(n)$] of groups of homoscedastic individuals (see Sec. 1 for some examples), then the methods developed in the homoscedastic case can be extended to the current problem. Assume that we have observations of the form (W_{jk}, I_j, Y_j), $j = 1, \ldots, n$, $k = 1, \ldots, r_j$, where $I_j \in \{1, \ldots, G\}$ indicates the group of the jth observation and

$$W_{j,k} = X_j + U_{j,k} \quad \text{with } U_{j,k} \sim f_{U_j} \equiv f_U^{I_j}, \qquad j \in \{1, \ldots, n\}, k \in \{1, \ldots, r_j\}.$$

Also assume that for each group g, f_U^g is symmetric and $f_U^{g,\text{ft}}$ is nonnegative. Within each group, we use the technique developed by Delaigle et al. (2007a), that is, we estimate $f_U^{g,\text{ft}}(t)$ by $\widehat{f}_U^{g,\text{ft}}(t) = |1/N_g \sum_{(j,k_1,k_2)\in\mathcal{S}_g} \cos\{t(W_{j,k_1} - W_{j,k_2})\}|^{1/2}$, where $\mathcal{S}_g = \{(j, k_1, k_2)$ s.t. $1 \le j \le n$, $1 \le k_1 < k_2 \le r_j$ and $I_j = g\}$ and $N_g = \#\mathcal{S}_g$. Then we define our estimator of m by

$$\tilde{m}_{1,n}(x) = \sum_{j=1}^{n} Y_j \tilde{K}_{U_j}\left(\frac{x - W_j}{h}\right) \Big/ \sum_{j=1}^{n} \tilde{K}_{U_j}\left(\frac{x - W_j}{h}\right), \tag{9}$$

with $\tilde{K}_{U_j}(x) = (2\pi)^{-1}\int e^{-itx} K^{\text{ft}}(t)\tilde{\Psi}_j(t/h)\,dt$ and $\tilde{\Psi}_j(t) = \widehat{f}_{U_j}^{\text{ft}}(t)/(\sum_{k=1}^{n} |\widehat{f}_{U_k}^{\text{ft}}(t)|^2 + \rho)$. Here $\rho > 0$ is a ridge parameter introduced to avoid division by 0, but in Section 4 we show

how to avoid its use in practice. We must assume that sufficient observations are available for each group (e.g., $N_g \geq \text{const} \cdot n$ and $\sum_j (r_j - 1)1_{\{I_j = g\}} \geq \text{const} \cdot n$). In the homoscedastic case, Delaigle et al. (2007a) proved that the convergence rates for this estimator are the same as those for the estimator for the case of known error densities. They were able to derive this result after imposing additional smoothness assumptions on f_X and some appropriate inequalities for the smoothness degrees of the error and f_X. It is possible to extend their results to the context of heteroscedastic errors, but the generalization is nontrivial and requires many technical assumptions on the errors and on the density f_X.

Case II: General Case. In some cases the contamination process is completely subjective and may differ among the observations. This situation is much more complicated, and at first sight, m might seem nonidentifiable unless we have a large number of replications for each individual allowing estimation of each error density f_{U_j}. It is possible to define an estimator of m without having to estimate each component f_{U_j} and without the need for more than two replications per individual. Suppose that we have replicated data in the form $(W_{j,k}, Y_j)$, $j = 1, \ldots, n$, $k = 1, \ldots, r_j$, and

$$W_{j,k} = X_j + U_{j,k}, \qquad U_{j,k} \sim f_{U_{j,k}}, \qquad j \in \{1, \ldots, n\}, k \in \{1, \ldots, r_j\}, r_j \geq 2. \quad (10)$$

Here each observation is replicated at least once, and the replications do not need to have the same error distribution. The only assumption needed for the identification is that at least one of the densities $f_{U_{j,\cdot}}$ is symmetric around 0 for each j; without loss of generality, we assume that $f_{U_{j,r_j}}$ has this property. Let $\varphi_W(t) = n^{-1} \sum_{j=1}^n \sum_{k=1}^{r_j - 1} \exp(it(W_{j,k} + W_{j,r_j})/2)$, $\varphi_U(t) = n^{-1} \sum_{j=1}^n \sum_{k=1}^{r_j - 1} \exp(it(W_{j,k} - W_{j,r_j})/2)$, $\varphi_Y(t) = n^{-1} \sum_{j=1}^n Y_j \sum_{k=1}^{r_j - 1} \exp(it(W_{j,k} + W_{j,r_j})/2)$, and $\phi_U(t) = [|\varphi_U(t)|^2 + \rho|\varphi_U(t)|]/\varphi_U(-t)$. Then our regression estimator is defined by

$$\widehat{m}_{2,n}(x) = \frac{\int \exp(-itx) K^{\text{ft}}(th) \varphi_Y(t)/\phi_U(t)\, dt}{\int \exp(-itx) K^{\text{ft}}(th) \varphi_W(t)/\phi_U(t)\, dt}, \quad (11)$$

where, as before, the ridge parameter ρ is introduced to prevent the denominator from getting too close to 0. In the case, where the errors are symmetric, a simpler version of this estimator takes $\phi_U(t) = \varphi_U(t) + \rho$.

Proposition 1 establishes consistency of the estimator $\widehat{m}_{2,n}(x)$ in the basic setting $r_j \equiv 2$. The basic idea of the proof is to show that the known error version of this estimator, defined by

$$\breve{m}_{2,n}(x) = \frac{\int \exp(-itx) K^{\text{ft}}(th) \varphi_Y(t)/\breve{\phi}_U(t)\, dt}{\int \exp(-itx) K^{\text{ft}}(th) \varphi_W(t)/\breve{\phi}_U(t)\, dt},$$

with φ_W and φ_Y as before and $\breve{\phi}_U(t) = n^{-1} \sum_{j=1}^n f^{\text{ft}}_{U_{j,1}}(t/2) \times f^{\text{ft}}_{U_{j,2}}(t/2)$, is consistent and that the distance between $\widehat{m}_{2,n}$ and $\breve{m}_{2,n}$ tends to 0.

Proposition 1. Assume that Condition A holds for some $x \in \mathbb{R}$ and take h and ρ such that $h \to 0$, $n^{1-\gamma} h \to \infty$, $\inf_{|t| \leq 1/h} |\breve{\phi}_U(t)|^2 \geq c \cdot n^{-\gamma}$, and $\ln \rho / \ln n \to a$ for $c > 0$, $\gamma \in (0, 1)$, $a < -1/2$. Then

$$\widehat{m}_{2,n}(x) \xrightarrow{P} m(x) \quad \text{as } n \to \infty.$$

As for Case I, proving that the rates of convergence of $\widehat{m}_{2,n}$ are the same as those of $\breve{m}_{2,n}$ will require much technical calculations and assumptions. Note that the price to pay for dealing with this very general unknown error scheme is that, because we use averaged observations, some loss of convergence speed may occur for ordinary smooth error densities, compared to the rates derived in Theorem 2. (This is true even for the estimator $\breve{m}_{2,n}(x)$, for which the error densities are known.) One advantage of our estimator is that it is relatively simple, however, it would be interesting to see whether that, in the ordinary smooth error case, there are better ways to estimate the unknown errors.

4. FINITE–SAMPLE PERFORMANCE

4.1 Data-Driven Procedure

The problem of selecting a data-driven bandwidth h is extremely difficult in the errors-in-variables problem (even in the homoscedastic case), because the observations X_k are not available. In particular, even the cross-validation (CV) criterion, which would select

$$h_{\text{CV}} = \arg\min_h \sum_{k=1}^n \big(Y_k - \widehat{m}_n^{-k}(X_k)\big)^2 w(X_k) \quad (12)$$

with w a weight function, where $\widehat{m}_n^{-k}$ denotes the estimator obtained when leaving the kth observation out, cannot be calculated. To the best of our knowledge, the only bandwidth selector proposed in the literature to date is the approximation proposed in the homoscedastic case by Delaigle et al. (2007a), which gives reasonable results when replicated data are available but otherwise tends to select overly large bandwidths. When adapted to the heteroscedastic case, their bandwidth selector $h_{\text{CV},1}$ is the value that minimizes the estimator of the right side of (12) obtained when replacing $e^{-itX_k/h}$ in $\widehat{m}_n^{-k}(X_k)$ by $e^{-itW_k/h} K^{\text{ft}}(t)/f^{\text{ft}}_{U_k}(-t/h)$, which has asymptotically the same expected value. We propose to multiply $h_{\text{CV},1}$ by an estimator of the shrinking factor $c_S = h_{\text{CV}}/h_{\text{CV},1}$ chosen by ideas related to bootstrap methods. Because the shrinking factor in the "original world" is not accessible, we create a "parallel world" that mimics the original world and where everything is calculable; we then replace c_S by $\tilde{c}_S$, the corresponding shrinking factor of the parallel world.

The way in which we mimic the original world is related to SIMEX ideas, because it consists in artificially adding noise to the data, as we describe now. In the original world, the data are of the type (W_j, Y_j), $j = 1, \ldots, n$, where $W_j = X_j + U_j$, $U_j \sim f_{U_j}$ is a contaminated version of the unobservable X_j and the target curve is $m(x) = E(Y|X = x) = n^{-1} \sum_{j=1}^n E(Y_j|X_j = x)$. In the parallel world, we create data of the type (W_j^*, Y_j), $j = 1, \ldots, n$, where Y_j is unchanged from the original world and $W_j^* = W_j + U_j^*$, $U_j^* \sim f_{U_j}$ is a contaminated version of the *observable* W_j (with W_j unchanged from the original world). The target curve is $\tilde{m}(x) = n^{-1} \sum_{j=1}^n E(Y_j|W_j = x)$, and the parallel world versions of h_{CV} and $h_{\text{CV},1}$, denoted by $\tilde{h}_{\text{CV}}$ and

$\tilde{h}_{\mathrm{CV},1}$, are obtained by replacing W_k by W_k^* and X_k by W_k in the respective criteria to minimize.

In most cases, the curves m and $\tilde{m}$ have sufficiently similar properties for the relation between $\tilde{h}_{\mathrm{CV}}$ and $\tilde{h}_{\mathrm{CV},1}$ to mimic the relation between h_{CV} and $h_{\mathrm{CV},1}$. We propose estimating the shrinking factor c_S by $\tilde{c}_S = \tilde{h}_{\mathrm{CV}}/\tilde{h}_{\mathrm{CV},1}$, and our final bandwidth selector is then $\widehat{h}_{\mathrm{CV}} = \tilde{c}_S \cdot h_{\mathrm{CV},1}$. Selected in this way, $\widehat{h}_{\mathrm{CV}}$ tends to 0 at the same speed as h_{CV} (and thus has the appropriate rate of convergence to 0), and only the multiplicative constant term is approximate. The implementation of the method necessitates generation of a parallel sample of $W_1^*, \ldots, W_n^*$ by adding a random error $U_j^* \sim f_{U_j}$ to W_j for $j = 1, \ldots, n$. To avoid the particular effect of the generated sample $W_1^*, \ldots, W_n^*$, we generate, say, B such parallel samples and take the average of the B corresponding calculated $\tilde{c}_S$'s.

In practice, similar to the error-free case, for small sample sizes, the denominator at (5) sometimes can get very close to 0 at some points, causing the estimator there to perform very poorly. Usually this problem arises only at points located near or beyond the smallest and largest observed values. In such cases the problem can be avoided by changing the denominator of $\widehat{m}_n$ into $\max(\widehat{f}_n(x), 0) + \rho_2$, with ρ_2 a small positive value. This approach has been described by Carroll, Delaigle, and Hall (2007) in the homoscedastic case, where it has been proved that for appropriate choices of ρ_2, the regularized estimator has the same asymptotic properties. Similarly, this modification is applied to the denominators used in the CV criteria and their estimators. In practice, we select ρ_2 by CV in the parallel world. To avoid the complication of selecting ρ_2 and $\widehat{h}_{\mathrm{CV}}$ simultaneously, we use the following method:

1. Take a grid of values of $(\rho_2, \tilde{h}_{\mathrm{CV}})$ in the parallel world and choose $\widehat{\rho}_2$ as the first component of the value of $(\rho_2, \tilde{h}_{\mathrm{CV}})$ that minimizes the average of B CVs on that grid.
2. Using the value $\widehat{\rho}_2$ found in step 1, select $\widehat{h}_{\mathrm{CV}}$ by using the procedure described in the previous paragraph.

When calculating the final estimator $\widehat{m}_n$, an adjustment that improves the practical results is to add the value $\widehat{\rho}_2$ only when $\widehat{f}_n$ gets too small. In our simulations, the denominator of $\widehat{m}_n$ that we used was $\max(\widehat{f}_n(x), 0) + \widehat{\rho}_2 \cdot 1\{\widehat{f}_n(x) \le \widehat{\rho}_2, x < q_{.025} \text{ or } x > q_{.975}\}$, where q_p is the pth empirical quantile of W.

In the case of unknown error distributions, we modify the procedure in two ways (we describe only the case where two replications are available per observation and the errors are symmetric): (a) Define the bandwidth $h_{\mathrm{CV},1}$ as the bandwidth that minimizes the approximation of (12) obtained by replacing $e^{-itX_k/h}$ in $\widehat{m}_n^{-k}(X_k)$ by $e^{-it\bar{W}_k/h}/\phi_U(t/h)$ and (b) generate the B samples of the parallel world differently. Here the error densities are unknown, and we generate W_i^* through $W_i^* = \bar{W}_i + \bar{U}_i^*$, where each sample $\bar{U}_1^*, \ldots, \bar{U}_n^*$ is drawn with replacement from $(W_{i1} - W_{i2})/2, \ldots, (W_{n1} - W_{n2})/2$. It would be interesting for future research to see whether a better generation algorithm can be implemented, but our simulations indicate that this simple algorithm already works well.

When the error densities are unknown, we also need to select the ridge parameter ρ (again we restrict ourselves to the symmetric error case), but we can avoid this by adapting the procedure proposed by Delaigle et al. (2007a): Replace $\phi_U(t) = \varphi_U(t) + \rho$ at the denominator of $\widehat{m}_{2,n}$ by $\varphi_U(t)I(t \in A) + \varphi_P(t)I(t \notin A)$, where $A = \{t \text{ s.t. } \varphi_U(t) < c\sqrt{\log n/n}\}$ and $\varphi_P(t)$ is a parametric function defined by $\varphi_P(t) = (1 + A_U t^2)^{-B_U}$, with A_U and B_U chosen so as to match the empirical second and fourth moments of the error with those of ϕ_P, or if A_U and/or B_U is negative, take $B_U = 1$ and A_U to be half the empirical variance of the error, which corresponds to ϕ_P being a Laplace density. The interval A is chosen so as to cut the estimator of φ_U before it becomes erratic. Our simulations indicated that $c = .025$ works well in practice.

To increase the speed of computations, we binned the data in 100 or 200 bins (see, e.g., Wand and Jones 1995) and used a spline approximation of the function φ_U constructed from 400 points. Note that these approximations are used only to select the bandwidth; thus they have a relatively minor impact on our final estimator. The functions K_{U_i} were calculated numerically using the fast Fourier transform (FFT). The code was implemented in C, and the standard routines (i.e., FT, random generation, spline approximation) were taken from Press, Flannery, Teukolsky, and Vetterling (1992). The time needed to run one iteration of the program depends on several factors (e.g., sample size, number of bins, size of grids), but in our simulations, it ranged from 1 minute for small sample sizes ($n \approx 100$) to several minutes for larger sample sizes ($n \approx 1{,}000$).

4.2 An Alternative Estimator

There exist parametric estimators that can be applied in the context of heteroscedastic errors (see, e.g., Devanarayana and Stefanski 2002; Cheng and Riu 2006). As usual, if the parametric assumption is correct, then such estimators perform better than nonparametric estimators and have the advantage of providing intuitive interpretation of the model and easily accessible inference on the estimated curve. On the other hand, nonparametric estimators can be applied in much more general contexts, but interpretation and inference on the estimated model is much more complex. Note that the availability of nonparametric methods is even more important in the error case than in the error-free case, because, in contrast to the latter, it often is impossible to formulate a parametric model for the relation between X and Y by simple inspection of a scatterplot of the data on (W, Y). In Figure 1, for example, we show the (X, Y) values and the (W, Y) values for a sample of size $n = 250$ contaminated by heteroscedastic Laplace errors with variances $\mathrm{var}(U_i) = (1 + i/n) \times .1 \,\mathrm{var}\, X$. The true regression curve, shown in (a), is in fact curve (c) of Section 4.3, which is very hard to guess by looking at the scatterplot of (W, Y).

Devanarayana and Stefanski (2002) proposed a parametric SIMEX regression estimator in the case in which the data are contaminated by heteroscedastic normal errors, and Carroll et al. (1999) proposed a nonparametric SIMEX estimator in the context of homoscedastic normal errors. We generalize their results and compare the numerical performance of our kernel estimator with a nonparametric SIMEX estimator for general heteroscedastic errors.

Note that in our case, the errors are not necessarily normal, and thus we cannot use the refined generation algorithms developed in those two articles. In the known error case, we implemented the procedure in the following way:

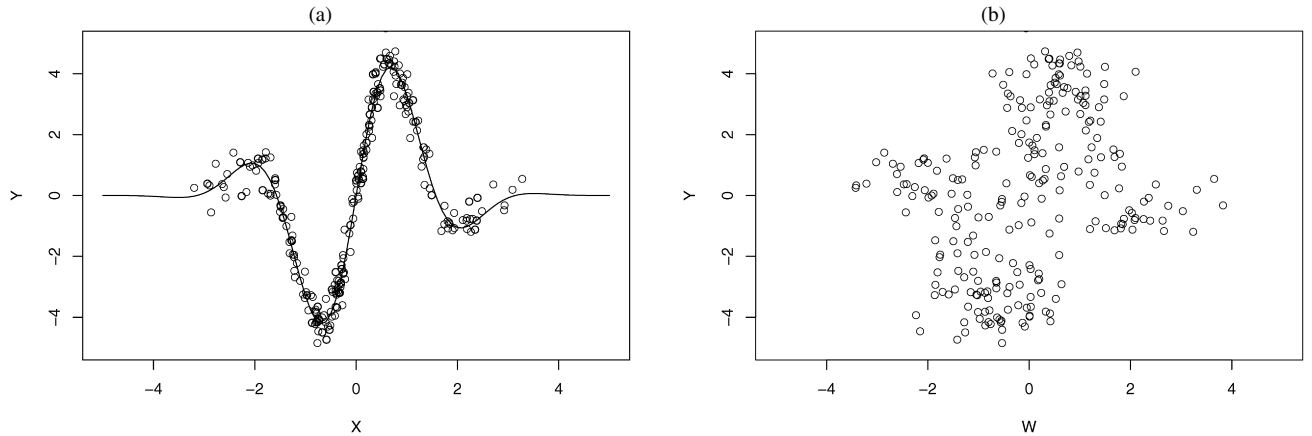

Figure 1. Scatterplot of the unobserved data values (X, Y) (a) and the observed data values (W, Y) (b) contaminated by Laplace errors. The regression curve m is shown in (a).

1. For $b = 1, \ldots, B$ and $\lambda = 1, 2, 3, 4$, generate pseudosamples $\mathcal{W}_{b,\lambda} = (W_{i,b,\lambda})_{1 \le i \le n}$ of size n by $W_{i,b,\lambda} = W_i + \sum_{j=1}^{\lambda} U_{i,b,j,\lambda}$, where $U_{i,b,j,\lambda} \sim U_i$ are independent and independent of W_i.
2. For each sample $(W_{1,b,\lambda}, Y_1), \ldots, (W_{n,b,\lambda}, Y_n)$, calculate the Nadaraya–Watson (NW) estimator $\widehat{m}_{b,\lambda}(x)$ at each x of the grid of interest, using the usual cross-validation bandwidth. Also calculate the NW estimator $\widehat{m}_0(x)$ based on the original data set.
3. Set $\widehat{m}_\lambda = \sum_{b=1}^{B} \widehat{m}_{b,\lambda}/B$ and, at each point x, use a quadratic extrapolant to estimate $m(x)$ by $\widehat{m}_{-1}(x)$.

We used the same procedure in the unknown error case, except that we generated the data as described in the previous section. We took $B = 50$. Although we tried several other ways to generate the samples and select the bandwidth, the method we present here is the one that gave the best results. It would be interesting to see whether a more complex implementation of SIMEX in a context as general as ours could be developed and, if so, whether it could improve the results.

In general, the SIMEX estimator is a consistent estimator not of the curve m, but rather of the curve obtained by extrapolation of the theoretical curves $m_\lambda = n^{-1} \sum_{j=1}^{n} E(Y_j | W_j^\lambda)$, which, if the error variances are not too large, can be a good approximation of m. Thus, in finite samples, if the error variances are not too large, it can provide a good approximation of the target curve and compete with consistent estimators, such as our kernel method. On the other hand, if the approximation of m by the extrapolating curve is too coarse, then the SIMEX method is not as good.

4.3 Simulation Results

We applied our methodology on some simulated examples. We generated samples $(W_{1j}, Y_{1j}), \ldots, (W_{nj}, Y_{nj})$, with $j = 1$ (no replications, known error densities) or $j = 1, 2$ (two replications, unknown error densities), according to model (3). We used one of the following combinations of the distribution of X and the regression function m: (a) $X \sim \mathrm{N}(0, 3)$, $m(x) = x^3 + 100 \cos x$ (unbounded sinusoid); (b) $X \sim \mathrm{N}(0, 1.5)$, $m(x) = \phi_{0,1.5}(4x) + \phi_{1,2}(4x) + \phi_{2,5}(4x)$ (asymmetric); (c) $X \sim \mathrm{N}(0, 1.5)$, $m(x) = 5 \sin(2x) \exp(-16x^2/50)$ (sinusoid); or (d) $X \sim \mathrm{N}(.5, .065)$, $m(x) = 3x + 20(2\pi)^{-1/2} \exp\{-100(x - \frac{1}{2})^2\}$ (mixture of a straight line and an exponential curve), where $\phi_{\mu,\sigma}$ is the density of a $\mathrm{N}(\mu, \sigma^2)$ variable.

In each case, we applied two different error models: (A) $U_1, \ldots, U_{n/2} \sim \mathrm{N}(0, \sigma_1^2)$ and $U_{n/2+1}, \ldots, U_n \sim \mathrm{Laplace}(\sigma_2)$, and (B) f_{U_i} are all normal or all Laplace, with $\mathrm{var}\, U_i = \sigma^2(1 + i/n)$. We took $\eta \sim \mathrm{N}(0, \sigma_\eta^2)$, where $\sigma_\eta^2 = .1 \times \mathrm{var}(|m|)$ and $\mathrm{var}(|m|) = \int_{q_{.01}^m}^{q_{.99}^m} \{|m| - \mathrm{E}(|m|)\}^2 / (q_{.99}^m - q_{.01}^m)$, with $\mathrm{E}(|m|) = \int_{q_{.01}^m}^{q_{.99}^m} |m| / (q_{.99}^m - q_{.01}^m)$ and q_α^m is the αth quantile of $|m|$ rescaled to integrate to 1. We chose σ_η^2 in this way rather than by the usual approach that takes σ_η^2 equal to some percentage of $\sup_x |m(x)|$, because curves (a) and (d) are unbounded.

For each of the foregoing examples, we generated 200 samples and calculated the corresponding 200 estimators and their associated integrated squared error, $ISE = \int_a^b \{\widehat{g}(x) - m(x)\}^2 \, dx$, with $\widehat{g}$ denoting an estimator of m and $[a, b]$ equal to the interval of x values shown on the presented graphs. We summarize the results by showing 3 of the 200 estimated curves, corresponding to the first (q_1), second (q_2), and third (q_3) quartiles of the 200 ISEs. In each graph, the target curve is the darkest, solid curve. We calculated every NW estimator using the standard normal kernel. Our estimator was calculated using the kernel with characteristic function $K^{\mathrm{ft}}(t) = (1 - t^2)^3 \cdot 1\{|t| \le 1\}$, frequently used in contamination problems; the parameters were selected as described in the previous sections, taking $B = 10$ and using the weight function $w(x) = 1\{a \le x \le b\}$, with a (resp. b) the .025th (resp. .975th) empirical quantile of W or W^*.

In some cases, our estimator and the SIMEX estimator gave very similar results. For example, Figure 2 compares these two estimators in the case of curve (a), for samples of size $n = 100$ or $n = 1{,}000$ generated under the error model (B) with unknown Laplace errors and $\sigma^2 = .2 \, \mathrm{var}\, X$. Each observation was replicated once. In this case, even for large sample sizes, the approximation error inherent to the SIMEX method (caused by the replacement of the target curve by the extrapolating curve), is of approximately the same importance as the estimating error made by our estimator, and both methods work well.

Figure 3 shows estimators of curve (c) for samples of size $n = 250$ generated by model (A). We show the results obtained

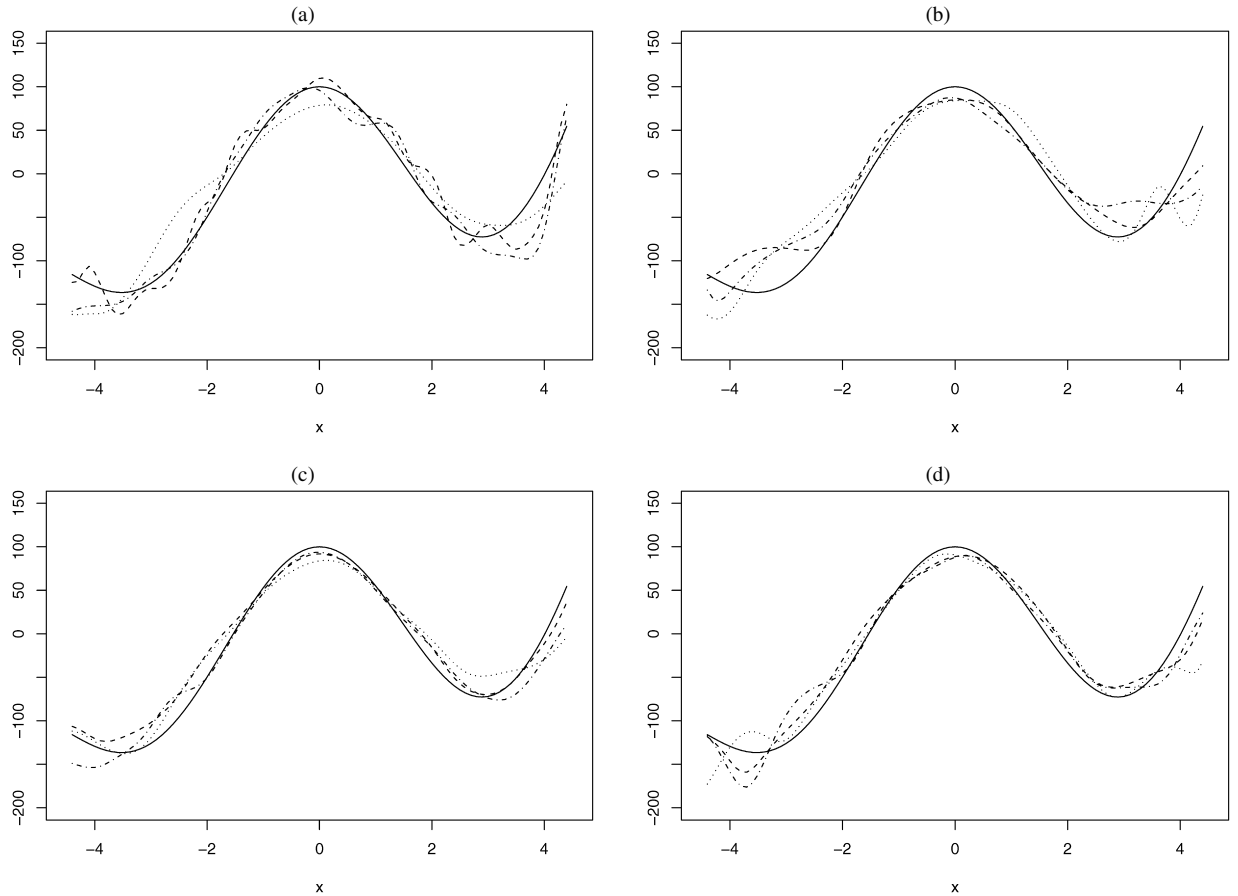

Figure 2. Estimation of curve (a) from samples of size $n = 250$ [(a) and (b)] or $n = 1{,}000$ [(c) and (d)] under the error model (B) when the error densities are unknown Laplace, using the SIMEX method [(a) and (c)] or the kernel method [(b) and (d)] (- - - q_1; ·-· q_2; ····· q_3).

for the SIMEX and our method in the case in which the error densities are unknown and each observation is replicated once, when $\operatorname{var} U_j = .4 \operatorname{var} X$. Here the SIMEX approximation error was larger, and we can see that the SIMEX procedure is attempting to estimate the wrong curve. We added this curve, denoted by m^*, to the graph of the SIMEX results. We can see the interest in using a consistent estimator like ours. To illustrate the effect of estimating the error density, we also show the results obtained with our estimator and with the naive Nadaraya–Watson estimator, when $\operatorname{var} U_j = .2 \operatorname{var} X$, the error densities are known, and no replication is available. Here we halved the error variances, because taking the averages of replicated observations, as was done earlier, also amounts to dividing these variances by 2. We see that in the finite sample, there is a small price to pay for having to estimate the error densities. However, a comparison with the results for the NW estimator (i.e., the estimator that ignores the errors) shows very clearly that the loss incurred by estimating the error densities is considerably smaller than that incurred by ignoring the errors from the analysis.

Finally, in Figure 4, we estimate curve (d) from samples of size $n = 250$ generated by model (B) with Laplace errors. Figure 4(a) shows the results of our estimator in the case where $\sigma^2 = .2 \operatorname{var} X$ and the unknown errors are estimated through the two replications available for each observation; Figure 4(b) shows the results for the naive NW estimator in the case where $\sigma^2 = .1 \operatorname{var} X$. Again, we see that the ignoring the contaminating errors results in a strongly biased estimator.

4.4 Real Data Example

We applied our method to a real data set from the Framingham Study on coronary heart disease described by Carroll et al. (2006). The data consist of measurements of systolic blood pressure (SBP) obtained at two different examinations and the incidence of coronary heart disease (CHD) in 1,615 males on an 8-year follow-up from the first examination. At each examination, the SBP was measured twice and for each individual, we take the average of these two measurements. Our goal is to analyze the relation between SBP and CHD without imposing any distribution to the error made by measuring the SBP and without imposing homoscedasticity. We set Y equal to CHD (0 if no incidence and 1 otherwise) and W_{j1} (resp. W_{j2}) equal to the logarithm of the (average SBP measurement -50) at examination 1 (resp. examination 2) for the jth individual; the transformation is as described by Carroll et al. (2006).

We compare the estimator $\widehat{m}_{2,n}$ with the SIMEX estimator discussed in Section 4.2 and the naive NW estimator (which ignores the error in the data) calculated with a cross-validation bandwidth. In this case, because the regression curve is a probability curve, it is interesting to compare the results with the logistic regression using regression calibration with the replicates, as described by Carroll et al. (2006); we used their linear approximation. The estimated curves are shown at Figure 5.

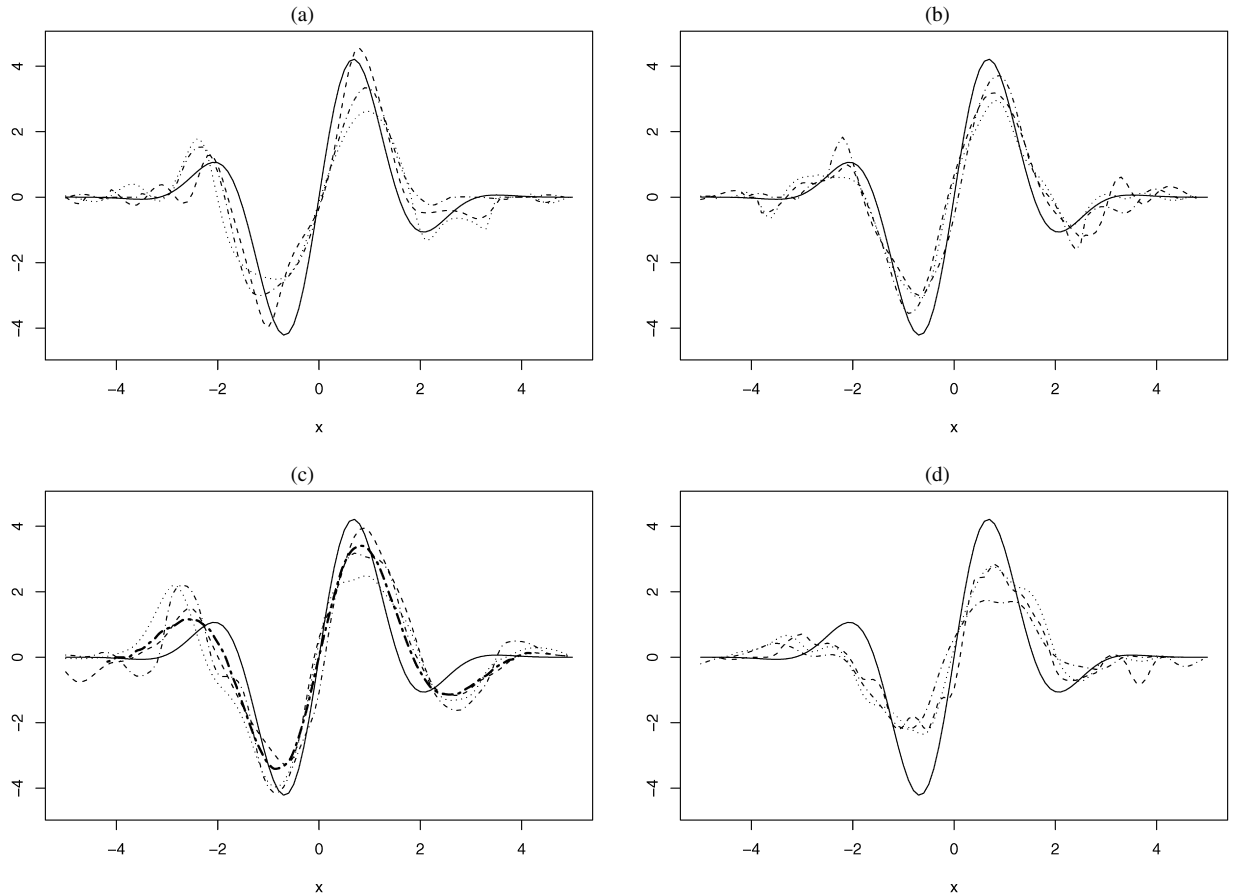

Figure 3. Estimation of curve (c) from samples of size $n = 250$ generated by model (A) when the error densities are unknown, using our method [(a)] or the SIMEX method [(c)], or when the error densities are known, using our method [(b)] or the naive NW estimator [(d)] (- - - q_1; - · - q_2; ····· q_3; - · - · m^*).

The three nonparametric methods are roughly in agreement, in the sense they indicate an increased risk of heart disease with increasing SBP (see also Carroll et al. 2006). As in our simulations, the NW estimator is smoother (probably oversmoothing the data) than the other two nonparametric estimators, both of which indicate a greater increase in the risk in the intervals $[4, 4.6]$ and $[4.8, 5]$. The similarity of these two estimators, which both incorporate the errors but use quite different techniques, gives us some confidence in their validity in this example. The parametric logistic curve closely follows the other three curves on the interval $[4, 4.6]$, but then increases much more rapidly between 4.6 and 5. Of course, we do not know the true curve, and it is impossible to know which estimator is the best. However, nonparametric estimators have the advan-

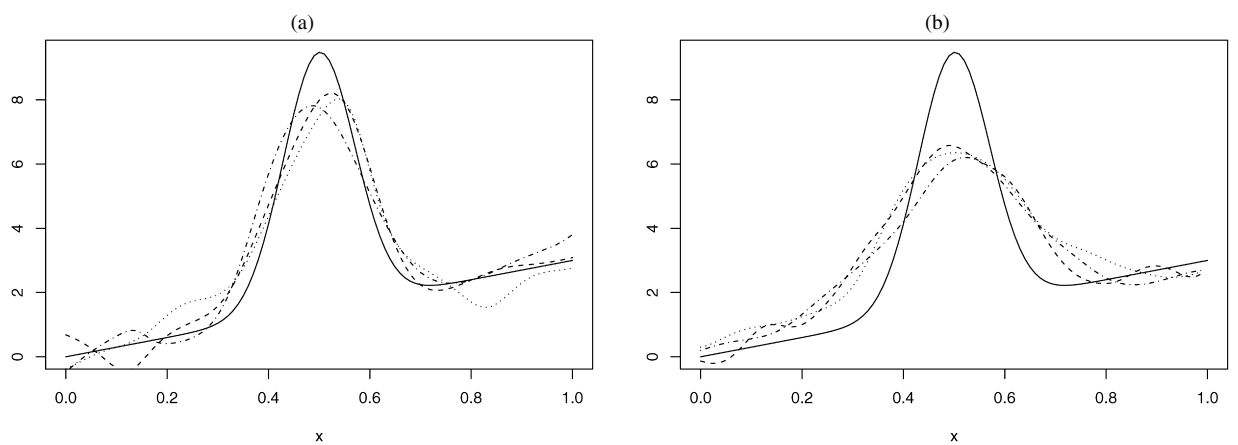

Figure 4. Estimation of curve (d) from samples of size $n = 250$ generated by model (B) with unknown Laplace errors using our method [(a)] or using the naive NW estimator [(b)] (- - - q_1; - · - q_2; ····· q_3).

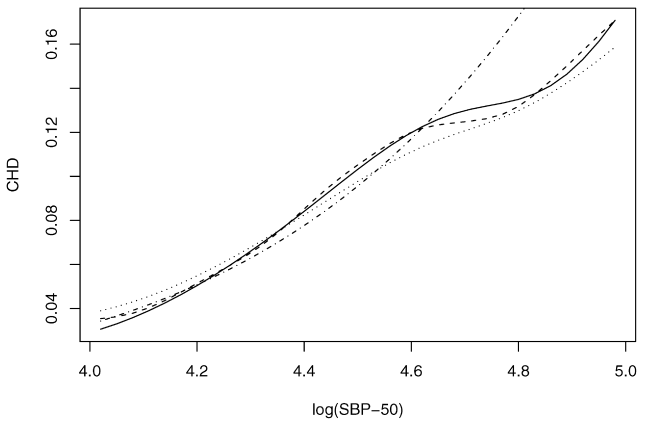

Figure 5. Estimation of the regression function E(CHD| log(SBP − 50)) for the Framingham-study data, using the estimator $\widehat{m}_{2,n}$ (—), the SIMEX procedure of Section 4.2 (- - -), the naive NW estimator (·····) and the logistic calibrated method (-·-·).

tage that they are consistent without the need to specify any model, whereas the logistic estimator can be consistent only if the unknown regression truly follows a logistic model. In this example, the big difference between this and the nonparametric estimators suggests that the logistic model may not be appropriate on the interval [4.6, 5]. It would be interesting for future research to test whether the difference is significant and investigate which of these estimators is most likely to reflect properties of the true regression curve.

APPENDIX: PROOFS

The following lemma, which gives an upper bound for the mean squared error (MSE) for the estimation of mf_X at an arbitrary but fixed x,

$$\mathrm{MSE}_n(f_X(x)m(x)) = \mathrm{E}|\widehat{f}_n(x)\widehat{m}_n(x) - f_X(x)m(x)|^2,$$

is useful for proving Theorems 1 and 2.

Lemma A.1. Let Condition A hold. Then we have for estimator (5),

$$\mathrm{MSE}_n(f_X(x)m(x)) = b_n^2 + v_n,$$

where $b_n^2 = |\int \frac{1}{h}K(\frac{x-y}{h})(mf_X)(y)\,dy - (mf_X)(x)|^2$ and $v_n \le \frac{1}{(2\pi)^2}(\|\tau^2 f_X\|_\infty + \|m^2 f_X\|_\infty)\cdot\int |K^{\mathrm{ft}}(th)|^2[\sum_{k=1}^n |f_{U_k}^{\mathrm{ft}}(t)|^2]^{-1}\,dt$.

Proof. The expression for the square of the bias term $b_n = \mathrm{E}[\widehat{f}_n(x)\widehat{m}_n(x) - f_X(x)m(x)]$ follows easily from $\mathrm{E}(Y_j\exp(itW_j)) = f_{U_j}^{\mathrm{ft}}(t)(mf_X)^{\mathrm{ft}}(t)$. For the variance term $v_n = \mathrm{var}[\widehat{f}_n(x)\widehat{m}_n(x)]$, using the independence between $Y_j|X_j$ and U_j, along with (A4) and Parseval's identity, we obtain

$$\begin{aligned}
v_n &= \sum_{j=1}^n \mathrm{var}\left\{Y_j\frac{1}{2\pi}\int \exp(-itx)K^{\mathrm{ft}}(th)\Psi_j(t)\exp(itW_j)\,dt\right\}\\
&\le \sum_{j=1}^n \mathrm{E}\left\{\mathrm{E}(Y_j^2|X_j)\right.\\
&\qquad\left.\times\left|\frac{1}{2\pi}\int \exp(-itx)K^{\mathrm{ft}}(th)\Psi_j(t)\exp(itW_j)\,dt\right|^2\right\}\\
&\le \frac{1}{2\pi}(\|\tau^2 f_X\|_\infty + \|m^2 f_X\|_\infty)\sum_{j=1}^n\int |K^{\mathrm{ft}}(th)|^2|\Psi_j(t)|^2\,dt\\
&= \frac{1}{2\pi}(\|\tau^2 f_X\|_\infty + \|m^2 f_X\|_\infty)\\
&\qquad\times\int |K^{\mathrm{ft}}(th)|^2\left[\sum_{k=1}^n |f_{U_k}^{\mathrm{ft}}(t)|^2\right]^{-1}dt.
\end{aligned}$$

Note that Lemma A.1 focuses on estimation of $m(x)f_X(x)$. Setting $Y_j = 1$ almost surely and $m \equiv 1$, we immediately get the corresponding result for the estimation of $f_X(x)$ by $\widehat{f}_n(x)$.

Lemma A.2. Assume Condition A and also that $\|m\|_\infty < \infty$. Then, for any integer $\gamma > 0$, if $E|\eta_j|^l \le C_l < \infty\ \forall l \le 2\gamma$, then we have

$$\begin{aligned}
&E|\widehat{f}_n(x)\widehat{m}_n(x) - E\widehat{f}_n(x)\widehat{m}_n(x)|^{2\gamma}\\
&\qquad = O(h^{-\gamma}n^{\gamma})\cdot\left[\int_{|t|\le 1/h}\left(\sum_{k=1}^n |f_{U_k}^{\mathrm{ft}}(t)|^2\right)^{-2}dt\right]^{\gamma}.
\end{aligned}$$

Proof. Let $Z_j(t) = Y_j\exp(itW_j) - E(Y_j\exp(itW_j))$. We have

$$\begin{aligned}
&E|\widehat{f}_n(x)\widehat{m}_n(x) - E\widehat{f}_n(x)\widehat{m}_n(x)|^{2\gamma}\\
&= E\left|\sum_{j=1}^n\frac{1}{2\pi}\int \exp(-itx)K^{\mathrm{ft}}(th)Z_j(t)\Psi_j(t)\,dt\right|^{2\gamma}\\
&\le O(1)\cdot\sum_{j_1}\cdots\sum_{j_{2\gamma}}\int\cdots\int\prod_{k_1=1}^{\gamma}|K^{\mathrm{ft}}(t_{k_1}h)||\Psi_{j_{k_1}}(t_{k_1})|\\
&\qquad\times\prod_{k_2=\gamma+1}^{2\gamma}|K^{\mathrm{ft}}(t_{k_2}h)|\\
&\qquad\times|\Psi_{j_{k_2}}(t_{k_2})|I_{(\#\{j_1,\ldots,j_{2\gamma}\}\le\gamma)}\Big[\sup_k E|Y_k|^{2\gamma}\Big]\,dt_1\cdots dt_{2\gamma}\\
&\le O(1)\cdot\sum_{j_1}\cdots\sum_{j_{2\gamma}}I_{(\#\{j_1,\ldots,j_{2\gamma}\}\le\gamma)}\prod_{l=1}^{2\gamma}\int|K^{\mathrm{ft}}(th)||\Psi_{j_l}(t)|\,dt\\
&\le O(1)h^{-\gamma}\cdot\sum_{j_1}\cdots\sum_{j_{2\gamma}}I_{(\#\{j_1,\ldots,j_{2\gamma}\}\le\gamma)}\\
&\qquad\times\left[\int_{|t|\le 1/h}\left(\sum_k|f_{U_k}(t)|^2\right)^{-2}dt\right]^{\gamma}\\
&\le O(1)n^{\gamma}h^{-\gamma}\cdot\left[\int_{|t|\le 1/h}\left(\sum_k|f_{U_k}(t)|^2\right)^{-2}dt\right]^{\gamma},
\end{aligned}$$

where we used the fact that $E(Z_{j_1}\cdots Z_{j_{2\gamma}}) = 0$ if $\{j_1,\ldots,j_{2\gamma}\}$ contains more than γ different elements because $E(Z_j) = 0$. Furthermore, we applied the Cauchy–Schwarz inequality to the integrals.

Proof of Theorem 1

To prove part a, note that we are able to select h as stated in part a due to condition (6) and the theorem of dominated convergence. The bias term b_n^2 from Lemma A.1 may be written in Fourier representation as

$$b_n^2 = \left|\frac{1}{2\pi}\int\exp(-itx)[K^{\mathrm{ft}}(th) - 1](mf_X)^{\mathrm{ft}}(t)\,dt\right|^2.$$

Then conditions (A2) and (A4) imply that b_n^2 tends to 0 as $h \to 0$. Because $|K^{\mathrm{ft}}|$ is bounded, the variance v_n also converges to 0 as a direct consequence of the conditions of Theorem 1 combined with Lemma A.1. Thus we have convergence of $\widehat{f}_n(x)\widehat{m}_n(x)$ to $f_X(x)m(x)$ and $\widehat{f}_n(x)$ to $f_X(x)$ in probability. By the usual technique of considering subsequences that converge almost surely, we derive that $\widehat{m}_n(x)$

has the weak limit $f_X(x)m(x)/f_X(x) = m(x)$, where condition (A3) is essential.

To prove part b, we consider

$$|\widehat{f}_n(x)\widehat{m}_n(x) - f_X(x)m(x)| \le |\widehat{f}_n(x)\widehat{m}_n(x) - E\widehat{f}_n(x)\widehat{m}_n(x)| + b_n. \quad \text{(A.1)}$$

As shown in part a, the fully deterministic term b_n converges to 0 as $h \to 0$. For the first addend in (A.1), we have, for all $\varepsilon > 0$,

$$\begin{aligned}
&\sum_{n=1}^{\infty} P\big(|\widehat{f}_n(x)\widehat{m}_n(x) - E\widehat{f}_n(x)\widehat{m}_n(x)| > \varepsilon\big) \\
&\quad \le \varepsilon^{-2\gamma} \sum_{n=1}^{\infty} E|\widehat{f}_n(x)\widehat{m}_n(x) - E\widehat{f}_n(x)\widehat{m}_n(x)|^{2\gamma} \\
&\quad \le \varepsilon^{-2\gamma} \sum_{n=1}^{\infty} O(h^{-\gamma})n^{\gamma} \cdot \left[\int_{|t|\le 1/h} \left(\sum_{k=1}^{n} |f_{U_k}^{\mathrm{ft}}(t)|^2\right)^{-2} dt\right]^{\gamma}, \\
&\quad \le \varepsilon^{-2\gamma} \sum_{n} O\big(n^{-(1-\delta+\kappa)\gamma}\big)n^{-\gamma(\delta-1)} \\
&\quad \le O(1) \cdot \sum_{n=1}^{\infty} n^{-\gamma\kappa} < \infty
\end{aligned}$$

if we take $\gamma \ge \lceil 1/\kappa \rceil + 1$; we have used Markov's inequality, Lemma A.2, and (7). Therefore, by the Borel–Cantelli lemma, $|\widehat{f}_n(x)\widehat{m}_n(x) - E\widehat{f}_n(x)\widehat{m}_n(x)| \xrightarrow{\text{a.s.}} 0$, which, combined with (A.1), implies that $\widehat{f}_n(x)\widehat{m}_n(x) \xrightarrow{\text{a.s.}} f_X(x)m(x)$. Using similar arguments, we have that $\widehat{f}_n(x) \xrightarrow{\text{a.s.}} f_X(x)$.

Proof of Theorem 2

To prove part a, by Markov's inequality, we derive the following upper bound:

$$\begin{aligned}
&P\big(|\widehat{m}_n(x) - m(x)|^2 > da_n^{-2\beta}\big) \\
&\quad \le P\big(|\widehat{m}_n(x)\widehat{f}_n(x) - m(x)f_X(x)|^2 > (dC_8^2/16)a_n^{-2\beta}\big) \\
&\qquad + P\big(|\widehat{f}_n(x) - f_X(x)|^2 > [dC_8^2/(16C_7^2)]a_n^{-2\beta}\big) \\
&\qquad + P\big(|\widehat{f}_n(x) - f_X(x)|^2 > C_8^2/4\big) \\
&\quad \le \text{const} \cdot d^{-1}a_n^{2\beta} \\
&\qquad \times \big\{E|\widehat{m}_n(x)\widehat{f}_n(x) - m(x)f_X(x)|^2 + E|\widehat{f}_n(x) - f_X(x)|^2\big\}.
\end{aligned}$$

Applying Lemma A.1, (C2),(C3), (C5), and (8), we can verify part a.

To prove part b, we start by introducing the specific densities $f(x) = [1 - \cos(x)]/(\pi x^2)$ with the Fourier transform $f^{\mathrm{ft}}(t) = (1 - |t|) \cdot \chi_{[-1,1]}(t)$, the supersmooth Cauchy density $s_X(x) = 1/[\pi(1 + x^2)]$, $s_Y(x) = (1/2)\exp(-|x|)$, $\Delta_Y(x) = (1/4)\,\mathrm{sign}(x)\exp(-|x|)$, and $\Delta_X(x) = a_n^{-\beta}\cos(2a_nx)f(a_nx)$. As competing bivariate densities for (X_j, Y_j), we give

$$f_{(X,Y),\theta}(x, y) = s_X(x)s_Y(y) + \theta\,\text{const}\,\Delta_X(x)\Delta_Y(y),$$

where $\theta \in \{0, 1\}$ and a_n as in (8). We obtain $f_{X,\theta}(x) = \int f_{(X,Y),\theta}(x, y)\,dy = s_X(x)$ and

$$\begin{aligned}
m_\theta(x)f_{X,\theta}(x) &= \int yf_{(X,Y),\theta}(x, y)\,dy \\
&= \theta\,\text{const}\,a_n^{-\beta}f(a_nx)\cos(2a_nx).
\end{aligned}$$

From there, we can verify the smoothness assumptions on mf_X and f_X, that is, (B1) and (B2).

Let

$$h_{j,\theta}(w, y) = \int f_{(X,Y),\theta}(x, y)f_{U_j}(w - x)\,dx$$

denote the density of the observation (W_j, Y_j). Putting $x = 0$ for simplicity, we have, for any estimator $\tilde{m}(x) \equiv \tilde{m}_n(x)$,

$$\begin{aligned}
&2 \sup_{(m,f_X)\in\mathcal{F}} P\big(|\tilde{m}(0) - m(0)|^2 \ge Da_n^{-2\beta}\big) \\
&\quad \ge P_{(m_0,f_{X,0})}\big(|\tilde{m}(0) - m_0(0)|^2 \ge Da_n^{-2\beta}\big) \\
&\qquad + P_{(m_1,f_{X,1})}\big(|\tilde{m}(0) - m_1(0)|^2 \ge Da_n^{-2\beta}\big) \\
&\quad \ge \int \cdots \int \min\big\{h_{1,0}(w_1, y_1)\cdots h_{n,0}(w_n, y_n), \\
&\qquad h_{1,1}(w_1, y_1)\cdots h_{n,1}(w_n, y_n)|\big\}\,dw_1\,dy_1\cdots dw_n\,dy_n, \qquad \text{(A.2)}
\end{aligned}$$

when setting $D = \inf_n\{a_n^{2\beta}|m_0(0) - m_1(0)|^2/4\}$; it follows from there that $\{|\tilde{m}(0) - m_0(0)|^2 < Da_n^{-2\beta}\}$ and $\{|\tilde{m}(0) - m_1(0)|^2 < Da_n^{-2\beta}\}$ are disjoint. Considering $m_0(0)$ and $m_1(0)$, we note that $D > 0$. Thus it remains to be shown that (A.2) is bounded away from 0. This corresponds to

$$\prod_{j=1}^{n} \int\int [h_{j,0}(w, y)h_{j,1}(w, y)]^{1/2}\,dw\,dy \ge \text{const} > 0, \qquad \text{(A.3)}$$

where LeCam's inequality (see, e.g., Devroye 1987, p. 7) has been used.

Equivalent to (A.3), we write

$$\sum_{j=1}^{n} \left|\ln\left(\int\int [h_{j,0}(w, y)h_{j,1}(w, y)]^{1/2}\,dw\,dy\right)\right| \le \text{const}.$$

We assume that n is sufficiently large. From the definition of $f_{(X,Y),\theta}$, we derive $h_{j,\theta}(w, y) \ge \text{const}\exp(-|y|)g_j(w)$, where $g_j(w) = \int s_X(x)f_{U_j}(w - x)\,dx$ is a density. This gives us a positive lower bound on all integrals occurring in (A.3). Combining this with the inequality $|\ln\xi| \le |(\xi - 1)/\xi|$ for all $\xi \in (0, 1]$, we obtain the inequality

$$\sum_{j=1}^{n} \left(1 - \int\int [h_{j,0}(w, y)h_{j,1}(w, y)]^{1/2}\,dw\,dy\right) = O(1), \qquad \text{(A.4)}$$

which implies (A.3). Condition (A.4) follows from

$$\sum_{j=1}^{n} \chi^2(h_{j,0}, h_{j,1}) = O(1), \qquad \text{(A.5)}$$

which is therefore sufficient for (A.3), where $\chi^2(f, g) = \int (f - g)^2/f\,dx$ denotes the chi-squared distance, according to the notation of Fan (1991b).

To verify (A.5), we consider, for $\theta \in \{0, 1\}$,

$$\begin{aligned}
h_{j,0}(w, y) &= s_Y(y)\int s_X(x)f_{U_j}(w - x)\,dx \\
&\ge s_Y(y)\int_{|x|\le d} s_X(w - x)f_{U_j}(x)\,dx \\
&\ge s_Y(y)\frac{1}{2(1 + 2w^2 + 2d^2)}\int_{|x|\le d} f_{U_j}(x)\,dx \\
&\ge c\exp(-|y|) \cdot \frac{1}{1 + w^2},
\end{aligned}$$

due to (C1), when selecting d sufficiently large and c sufficiently small. By Parseval's identity and the equality $f^{\mathrm{ft}\prime}(t) = i(\bullet f(\bullet))^{\mathrm{ft}}(t)$, we see

that (A.5) is satisfied if

$$\sum_{j=1}^{n} \int\int \left| \int \Delta_Y(y)\Delta_X(x) f_{U_j}(w-x)\,dx \right|^2 (1+w^2)\exp(|y|)\,dw\,dy$$

$$\le O(1)\cdot \sum_{j=1}^{n} \int \exp(-|y|)\,dy$$

$$\times \frac{1}{2\pi} \int \Big[\big|a_n^{-\beta-1} f^{\mathrm{ft}}((t\pm 2a_n)/a_n) f_{U_j}^{\mathrm{ft}}(t)\big|^2$$

$$+ \big|a_n^{-\beta-2} f^{\mathrm{ft}\prime}((t\pm 2a_n)/a_n) f_{U_j}^{\mathrm{ft}}(t)\big|^2$$

$$+ \big|a_n^{-\beta-1} f^{\mathrm{ft}}((t\pm 2a_n)/a_n) f_{U_j}^{\mathrm{ft}\,\prime}(t)\big|^2 \Big]\,dt$$

$$\le \sum_{j=1}^{n} O(a_n^{-2\beta-2}) \cdot \int_{|t|\in[a_n,3a_n]} \left(|f_{U_j}^{\mathrm{ft}}(t)|^2 + |f_{U_j}^{\mathrm{ft}\,\prime}(t)|^2 \right) dt$$

$$\le O(a_n^{-2\beta-1}) \sum_{j=1}^{n} (\overline{\varphi}_{j,n}(a_n))^2 \le \text{const}, \qquad \text{(A.6)}$$

where we have used (C3). The inequalities (A.6), and thus (A.2), follow from (8).

[Received November 2006. Revised July 2007.]

REFERENCES

Bennett, C. A., and Franklin, N. L. (1954), *Statistical Analysis in Chemistry and the Chemical Industry*, New York: Wiley.

Berry, S. M., Carroll, R. J., and Ruppert, D. (2002), "Bayesian Smoothing and Regression Splines for Measurement Error Problems," *Journal of the American Statistical Association*, 97, 160–169.

Carroll, R. J., and Hall, P. (1988), "Optimal Rates of Convergence for Deconvolving a Density," *Journal of the American Statistical Association*, 83, 1184–1186.

——— (2004), "Low-Order Approximations in Deconvolution and Regression With Errors in Variables," *Journal of the Royal Statistical Society*, Ser. B, 66, 31–46.

Carroll, R. J., Delaigle, A., and Hall, P. (2007), "Nonparametric Regression Estimation From Data Contaminated by a Mixture of Berkson and Classical Errors," *Journal of the Royal Statistical Society*, Ser. B, to appear.

Carroll, R. J., Eltinge, J. L., and Ruppert, D. (1993), "Robust Linear Regression in Replicated Measurement Error Models," *Statistics and Probability Letters*, 16, 169–175.

Carroll, R. J., Maca, J. D., and Ruppert, D. (1999), "Nonparametric Regression in the Presence of Measurement Error," *Biometrika*, 86, 541–554.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006), *Measurement Error in Nonlinear Models* (2nd ed.), Boca Raton, FL: Chapman & Hall CRC.

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," *Journal of the American Statistical Association*, 89, 1314–1328.

Delaigle, A., Hall, P., and Meister, A. (2007a), "On Deconvolution With Repeated Measurements," *The Annals of Statistics*, to appear.

Delaigle, A., Hall, P., and Müller, H. G. (2007b), "Accelerated Convergence for Nonparametric Regression With Coarsened Predictors," *The Annals of Statistics*, to appear.

Devanarayana, V., and Stefanski, L. A. (2002), "Empirical Simulation Extrapolation for Measurement Error Models With Replicate Measurements," *Statistics and Probability Letters*, 59, 219–225.

Devroye, L. (1987), *A Course in Density Estimation*, Boston: Birkhäuser.

Fan, J. (1991a), "Global Behaviour of Deconvolution Kernel Estimates," *Statistica Sinica*, 1, 541–551.

——— (1991b), "On the Optimal Rates of Convergence for Nonparametric Deconvolution Problems," *The Annals of Statistics*, 19, 1257–1272.

Fan, J., and Masry, E. (1992), "Multivariate Regression Estimation With Errors in Variables: Asymptotic Normality for Mixing Processes," *Journal of Multivariate Analysis*, 43, 237–271.

Fan, J., and Truong, Y. K. (1993), "Nonparametric Regression With Errors in Variables," *The Annals of Statistics*, 21, 1900–1925.

Fuller, W. A. (1987), *Measurement Error Models*, New York: Wiley.

Hall, P., and Yao, Q. (2003), "Inference in Components of Variance Models With Low Replication," *The Annals of Statistics*, 31, 414–441.

Horowitz, J. L., and Markatou, M. (1996), "Semiparametric Estimation of Regression Models for Panel Data," *Review of Economic Studies*, 63, 145–168.

Kulathinal, S. B., Kuulasmaa, K., and Gasbarra, D. (2002), "Estimation of an Errors-in-Variables Regression Model When the Variances of the Measurement Errors Vary Between the Observations," *Statistics in Medicine*, 21, 1089–1101.

Li, T. (2002), "Robust and Consistent Estimation of Nonlinear Errors-in-Variables Models," *Journal of Econometrics*, 110, 1–26.

Li, T., and Hsiao, C. (2004), "Robust Estimation of Generalised Linear Models With Measurement Errors," *Journal of Econometrics*, 118, 51–65.

Li, T., and Vuong, Q. (1998), "Nonparametric Estimation of the Measurement Error Model Using Multiple Indicators," *Journal of Multivariate Analysis*, 65, 139–165.

Liang, H., and Wang, N. (2005), "Large-Sample Theory in a Semiparametric Partially Linear Errors-in-Variables Model," *Statistica Sinica*, 15, 99–117.

Linton, O., and Whang, Y.-J. (2002), "Nonparametric Estimation With Aggregated Data," *Econometric Theory*, 18, 420–468.

Madansky, A. (1959), "The Fitting of Straight Lines When Both Variables Are Subject to Error," *Journal of the American Statistical Association*, 54, 173–205.

National Research Council, Committee on Pesticides in the Diets of Infants and Children (1993), *Pesticides in the Diets of Infants and Children*, Washington, DC: National Academies Press.

Neumann, M. H. (2007), "Deconvolution From Panel Data With Unknown Error Distribution," *Journal of Multivariate Analysis*, to appear.

Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T. (1992), *Numerical Recipes in C, the Art of Scientific Computing* (2nd ed.), New York: Cambridge University Press.

Riu, J., and Rius, F. X. (1995), "Univariate Regression Models With Errors in Both Axes," *Journal of Chemometrics*, 9, 343–362.

——— (1996), "Assessing the Accuracy of Analytical Methods Using Linear Regression Both Axes," *Analytical Chemistry*, 68, 1851–1857.

Schennach, S. M. (2004a), "Estimation of Nonlinear Models With Measurement Error," *Econometrica*, 72, 33–75.

——— (2004b), "Nonparametric Regression in the Presence of Measurement Error," *Econometric Theory*, 20, 1046–1093.

Stefanski, L., 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, J. R. (1995), "Simulation-Extrapolation: The Measurement Error Jackknife," *Journal of the American Statistical Association*, 90, 1247–1256.

Stone, C. J. (1977), "Consistent Nonparametric Regression," *The Annals Statistics*, 5, 595–645.

Susko, E., and Nadon, R. (2002), "Estimation of a Residual Distribution With Small Numbers of Repeated Measurements," *Canadian Journal of Statistics*, 30, 383–400.

Taupin, M. L. (2001), "Semi-Parametric Estimation in the Nonlinear Structural Errors-in-Variables Model," *The Annals of Statistics*, 29, 66–93.

Thamerus, M. (2003), "Fitting a Mixture Distribution to a Variable Subject to Heteroscedastic Measurement Errors," *Computational Statistics*, 18, 1–17.

Walter, S. D. (1997), "Variation in Baseline Risk as an Explanation of Heterogeneity in Meta-Analysis," *Statistics in Medicine*, 16, 2883–2900.

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