Statistica Sinica **21** (2011), 1093-1113

NONPARAMETRIC REGRESSION FUNCTION ESTIMATION FOR ERRORS-IN-VARIABLES MODELS WITH VALIDATION DATA

Lilun Du, Changliang Zou and Zhaojun Wang

Nankai University

Abstract: This paper develops an estimation approach for nonparametric regression analysis with measurement error in covariates, assuming the availability of independent validation data on them, in addition to primary data on the response variable and surrogate covariates. Without specifying any error model structure between the surrogate and true covariates, we propose an estimator that integrates local linear regression and Fourier transformation methods. Under mild conditions, the consistency of the proposed estimator is established and the convergence rate is also obtained. Numerical examples show that it performs well in applications.

Key words and phrases: Asymptotic normality, local linear regression, measurement error, trigonometric series.

1. Introduction

In the last two decades, the errors-in-variables (EV) model has drawn much attention. An increasing number of applications of the linear and non-linear EV models have been seen in recent years due to their simple forms and wide applicabilities. Comprehensive reviews can be found in Fuller (1987), Carroll et al. (2006) and the references therein. In practice, the relationship between the measured (surrogate) variables and the true variables can be rather complicated compared to the classical additive error structure. In such cases, obtaining correct statistical analyses becomes challenging. One solution is to use validation data to capture the underlying relationship between the true and surrogate variables. For instance, in the measurement of heart muscle damage caused by a myocardial infection, the peak cardiac enzyme level in the bloodstream is a variable easily obtained, though it cannot accurately assess the damage to the heart muscle. Instead, the arterioscintograph, an invasive and expensive procedure, can be employed to produce a more accurate measure of the heart muscle in a small subset of subjects(cf., Wittes, Lakatos and Probstfield (1989)). Here, for heart damage, peak cardiac enzyme level in the bloodstream is used as a surrogate variable, with exact data measured by arterioscintograph used as a validation variable for a small subset of subjects.

Inference based on surrogate data and a validation sample has been the object of much attention. Carroll and Stefanski (1990), Carroll and Wand (1991), Pepe and Fleming (1991), Pepe (1992), Carroll, Knickerbocker, and Wang (1995), Lee and Sepanski (1995), Sepanski and Lee (1995), Wang (1999), Wang and Rao (2002), and Stute, Xue and Zhu (2007) developed suitable methods for different models. Most focus on specifying some parametric relationship between covariates and response. While parametric methods are useful in certain applications, questions always arise about the adequacy of parametric model assumptions and about potential impact of model mis-specifications on statistical analyses. In comparison, it is well known that nonparametric regression provides a useful tool, especially when the relationship is complicated. This motivates us to consider nonparametric regression when the covariate is measured erroneously and some validation data are available to relate the surrogate and true variables.

To be specific, we assume that an independent validation dataset $\mathcal{V} = \{(t_j, \tilde{t}_j)\}_{j=N+1}^{N+n}$ is available, in addition to the primary (surrogate) dataset $\mathcal{S} = \{(Y_i, \tilde{t}_i)\}_{i=1}^{N}$ which is generated by the nonparametric model

$$Y = m(t) + \varepsilon, \tag{1.1}$$

where Y is a scalar response variable, t is a univariate explanatory variable, $\tilde{t}$ is the surrogate variable of t, and ε is a random error with $E[\varepsilon|t] = 0$ and $E[\varepsilon^2] < \infty$. Given t_i's, the errors ε_i's are assumed to be independent and identically distributed. Our objective is to estimate the unknown regression function $m(t)$ in (1.1) with the datasets $\mathcal{V}$ and $\mathcal{S}$. Obviously, standard methodologies based on regression calibration developed for parametric inferences, such as Carroll and Wand (1991), Sepanski and Lee (1995), and Stute, Xue and Zhu (2007) are not applicable in this situation. Recently, Wang (2006) developed an estimation approach for nonparametric regression analysis with surrogate data and validation sampling, but which cannot be applied to our problem.

In this paper, we propose a nonparametric estimator $\widehat{m}(z)$ that integrates the local linear regression and Fourier transformation methods. This estimation approach consists of two major steps: An estimator combining two local linear kernel estimators (LLKE) based on $\mathcal{V}$ and $\mathcal{S}$, respectively, is first proposed to calibrate the function $E[Y|U(\tilde{t}) = z]$, in which $U(\tilde{t}) = E[t|\tilde{t}]$. Although $E[Y|U(\tilde{t}) = z]$ is not our objective function, the relationship between $m(z) = E[Y|t = z]$ and $E[Y|U(\tilde{t}) = z]$ can be derived through the trigonometric series approach suggested by Delaigle, Hall, and Qiu (2006). Under mild conditions, the consistency of $\widehat{m}(z)$ is established and the convergence rate derived.

In the next section, we describe our estimation approach and state the main asymptotic results. In Section 3, we investigate the finite sample properties of

our proposed approach. A data example is used to demonstrate the method in Section 4. Section 5 concludes this paper by suggesting some future research issues. The proofs are given in the Appendix.

2. Methodology

Recall model (1.1) and the assumptions below it. We rewrite (1.1) as

$$Y = m(U(\tilde{t}) + \eta) + \varepsilon, \tag{2.1}$$

$$t = U(\tilde{t}) + \eta, \tag{2.2}$$

where $U(\tilde{t})$, ε, and η are independent. With $U(\tilde{t})$ as a new variable, this can be regarded as a Berkson EV model (Berkson (1950); Carroll et al. (2006)). This enables us to apply a recent nonparametric technique developed by Delaigle, Hall, and Qiu (2006), as long as we have an estimate of the distribution of η and $E[Y|U(\tilde{t}) = z]$, on which we elaborate next.

2.1. Estimating $E[Y|U(\tilde{t}) = z]$

Represent (2.1) as

$$Y = M(U(\tilde{t})) + \xi, \tag{2.3}$$

where $M(z) \equiv E[Y|U(\tilde{t}) = z]$, $E[\xi|\tilde{t}] = 0$, and $E[\xi^2] < \infty$. Generally, M is not equal to m and $\xi \neq \varepsilon + \eta$. Based on the validation set $\mathcal{V}$, we can estimate $U(\tilde{t})$ by means of local linear fit, that is

$$\widehat{U}(\tilde{t}) = \frac{\sum_{j=N+1}^{N+n} t_j L\left((\tilde{t}_j - \tilde{t})/b_n\right)\{S_{n2} - (\tilde{t}_j - \tilde{t})S_{n1}\}}{\sum_{j=N+1}^{N+n} L\left((\tilde{t}_j - \tilde{t})/b_n\right)\{S_{n2} - (\tilde{t}_j - \tilde{t})S_{n1}\}},$$

where

$$S_{n\gamma} = \frac{1}{nb_n}\sum_{j=N+1}^{N+n} L\left(\frac{\tilde{t}_j - \tilde{t}}{b_n}\right)(\tilde{t}_j - \tilde{t})^{\gamma}, \quad \gamma = 0, 1, 2,$$

$L(\cdot)$ is a symmetric density kernel function, and b_n is a bandwidth. Here we choose the LLKE rather than the Nadaraya-Watson estimator since the former possesses superior boundary behavior (cf., Fan (1992)).

So far, with the datasets $\mathcal{V}$ and $\mathcal{S}$, we can estimate the function $M(\cdot)$ by plugging $\widehat{U}(\tilde{t})$ into the LLKE of $\{(Y_i, U(\tilde{t}_i))\}_{i=1}^N$, that is,

$$\widehat{M}(z) = \frac{\sum_{i=1}^N Y_i K\left((\widehat{U}(\tilde{t}_i) - z)/h_N\right)\{S_{N2} - (\widehat{U}(\tilde{t}_i) - z)S_{N1}\}}{\sum_{i=1}^N K\left((\widehat{U}(\tilde{t}_i) - z)/h_N\right)\{S_{N2} - (\widehat{U}(\tilde{t}_i) - z)S_{N1}\}}, \tag{2.4}$$

where

$$S_{N\gamma} = \frac{1}{Nh_N} \sum_{i=1}^{N} K\left(\frac{\widehat{U}(\tilde{t}_i) - z}{h_N}\right) (\widehat{U}(\tilde{t}_i) - z)^{\gamma}, \quad \gamma = 0, 1, 2, 3,$$

$K(\cdot)$ is a symmetric density kernel function, and h_N is a bandwidth.

Remark 1. The preceding algorithms require the LLKE of $U(\tilde{t})$ regressed on $\tilde{t}$ in the validation data. As discussed by Carroll and Wand (1991), Sepanski and Carroll (1993), Sepanski, Knickerbocker and Carroll (1994), and others, the range of $\tilde{t}$ in the validation data is usually smaller than that in primary data observed, which could affect the LLKE $\widehat{U}(\tilde{t})$ studied above; if used blindly, this would lead to an extrapolation. The LLKE $\widehat{U}(\tilde{t})$ used here is calculated on a compact set $\Theta = [\tilde{t}_{V\min}, \tilde{t}_{V\max}]$ interior to the support of $\tilde{t}$, where $\tilde{t}_{V\min} = \min\{\tilde{t}_j\}_{j=N+1}^{N+n}$ and $\tilde{t}_{V\max} = \max\{\tilde{t}_j\}_{j=N+1}^{N+n}$. Sums in the primary data are taken only for those $\tilde{t} \in \Theta$. While this truncation causes a certain loss in efficiency, it is counterbalanced by a gain in robustness.

2.2. Estimating $m(z)$

In what follows, we assume that the densities of t, $\tilde{t}$, and η, denoted as f_t, $f_{\tilde{t}}$ and f_η, respectively, are compactly supported and bounded away from zero. Without loss of generality, we suppose their support intervals have been rescaled to be within $\Omega = [-\pi, \pi]$. In addition, to assure that m is identifiable, we suppose that the the support of f_t is contained within the range of $U(\tilde{t})$. After obtaining $\widehat{M}(z)$ by (2.4), the Fourier transformation method introduced by Delaigle, Hall, and Qiu (2006) can be used to find the relationship between $m(z)$ and $M(z)$.

On Ω we write the trigonometric series for $m(z)$ as

$$m(z) = m_0 + \sum_{l=1}^{\infty} \{m_{1l} \cos(lz) + m_{2l} \sin(lz)\}, \tag{2.5}$$

where

$$m_0 = \frac{1}{2\pi} \int_\Omega m(t)dt, \quad m_{1l} = \frac{1}{\pi} \int_\Omega m(t) \cos(lt)dt, \quad m_{2l} = \frac{1}{\pi} \int_\Omega m(t) \sin(lt)dt.$$

Analogously, $M(U)$ can be written as

$$M(U) = M_0 + \sum_{l=1}^{\infty} \{M_{1l} \cos(lU) + M_{2l} \sin(lU)\}, \tag{2.6}$$

where the constants M_0, M_{1l}, and M_{2l}, $l = 1, 2, \ldots$, are the Fourier coefficients determined by M. Furthermore, the coefficients m_{1l}, m_{2l} are uniquely determined

by M_{l1} and M_{2l}, if U and η are independent, which is implied by the Berkson model (2.1)−(2.2). Hence, simple calculations yield

$$\begin{pmatrix} m_{1l} \\ m_{2l} \end{pmatrix} = \frac{1}{\alpha_{1l}^2 + \alpha_{2l}^2} \begin{pmatrix} \alpha_{1l} & -\alpha_{2l} \\ \alpha_{2l} & \alpha_{1l} \end{pmatrix} \begin{pmatrix} M_{1l} \\ M_{2l} \end{pmatrix}, \tag{2.7}$$

where $\alpha_{1l} = E\{\cos(l\eta)\}$ and $\alpha_{2l} = E\{\sin(l\eta)\}$, provided $\alpha_{1l}^2 + \alpha_{2l}^2 \neq 0$ for $l \geq 1$.

In the present problem, the distribution of η is not explicitly available. However, with the help of validation data, the empirical distribution of the $\widehat{\eta}_j$'s can be used instead, where

$$\widehat{\eta}_j = t_j - \widehat{U}(\tilde{t}_j), \quad j = N+1, \ldots, N+n.$$

This type of estimation has been proposed and studied by Akritas and Van Keilegom (2001). Correspondingly, α_{1l} and α_{2l}, $l = 1, 2, \ldots$, can be estimated by

$$\widehat{\alpha}_{1l} = \frac{1}{n} \sum_{j=N+1}^{N+n} \cos(l\widehat{\eta}_j), \quad \widehat{\alpha}_{2l} = \frac{1}{n} \sum_{j=N+1}^{N+n} \sin(l\widehat{\eta}_j). \tag{2.8}$$

Thus, the estimated coefficients of m_{1l}, m_{2l} in (2.7) can be represented as

$$\begin{pmatrix} \widehat{m}_{1l} \\ \widehat{m}_{2l} \end{pmatrix} = \frac{1}{\widehat{\alpha}_{1l}^2 + \widehat{\alpha}_{2l}^2} \begin{pmatrix} \widehat{\alpha}_{1l} & -\widehat{\alpha}_{2l} \\ \widehat{\alpha}_{2l} & \widehat{\alpha}_{1l} \end{pmatrix} \begin{pmatrix} \widehat{M}_{1l} \\ \widehat{M}_{2l} \end{pmatrix}, \tag{2.9}$$

where

$$\widehat{M}_{1l} = \frac{1}{\pi} \int_{\mathcal{H}} \widehat{M}(t) \cos(lt) dt, \quad \widehat{M}_{2l} = \frac{1}{\pi} \int_{\mathcal{H}} \widehat{M}(t) \sin(lt) dt,$$
$$\widehat{m}_0 = \widehat{M}_0 = \frac{1}{2\pi} \int_{\mathcal{H}} \widehat{M}(t) dt,$$

$\widehat{M}(\cdot)$ is defined in (2.4), and $\mathcal{H} \subseteq \Omega$ contains the support of M. Combining (2.5) and (2.9), our estimator is

$$\widehat{m}(z) = \widehat{m}_0 + \sum_{l=1}^{q} [\widehat{m}_{1l} \cos(lz) + \widehat{m}_{2l} \sin(lz)], \tag{2.10}$$

where q denotes the number of Fourier coefficients that are included in the estimator and can be deemed as another smoothing (regularization) parameter besides b_n and h_N.

Remark 2. Based on (2.1)−(2.2), we can see that inference on the relationship between $m(z)$ and $M(z)$ essentially lies in nonparametric regression with

additive EV structure. Nonparametric methods for inference in the settings of the classical EV model, that is $\tilde{t} = t + \delta$, include kernel approaches (e.g., Fan and Truong (1993); Delaigle, Fan and Carroll (2009)) and techniques based on simulation and extrapolation (Cook and Stefanski (1994); Carroll, Maca, and Ruppert (1999); Staudenmayer and Ruppert (2004)). These approaches may be also used in the present circumstance with modifications, although the Berkson model is particularly appropriate by (2.2). This problem is a subject for future research.

2.3. Main results

In this subsection, we study the asymptotic behavior of the proposed estimator. Some conditions are needed. Let $F_{t|\tilde{t}}$ denotes the conditional distribution function of t given $\tilde{t}$.

(C1) $m(\cdot)$ has an s-order Lipschitz continuous second derivative for some real $s > 0$.

(C2) The functions $U(\cdot)$ and $M(\cdot)$ are bounded and twice continuously differentiable.

(C3) The density function of $U(\tilde{t})$, denoted as f_U, is Lipschitz continuous and bounded away from 0.

(C4) The kernel functions $K(w)$ and $L(w)$ are bounded and symmetric probability density functions and satisfy $\int K(w)w^2dw < \infty$, $\int L(w)w^2dw < \infty$. In addition, $K(w)$ is twice differentiable and satisfies $\int K'^2(w)w^2dw < \infty$ and $\int |K''(w)|dw < \infty$.

(C5) N, n, b_n, h_N satisfy $\gamma_n/h_N \to 0$, $nb_n \to \infty$, and $Nh_N \to \infty$, where $\gamma_n = (1/\sqrt{nb_n} + b_n^2)\log^{1/2}(1/b_n)$.

(C6) $\lim N/n = \lambda \in [0, \infty)$.

(C7) The density function $f_{\tilde{t}}$ is three times continuously differentiable.

(C8) $F'_{t|\tilde{t}}$ is continuous in $(t, \tilde{t})$ and $\sup_{t,\tilde{t}} |t^2 F'_{t|\tilde{t}}| < \infty$; the same holds for all other partial derivatives of $F_{t|\tilde{t}}$ with respect to t and $\tilde{t}$ up to order two.

Remark 3. Conditions (C1)−(C4) are standard in nonparametric regression, as is the bandwidth condition, (C5) (C6) is standard in the study of validation sampling, and is reasonable in practice. (C7) and (C8) are mild conditions used in the proof of Theorem 2, that guarantees the consistency of the moment estimators (2.8).

Due to the relationship between $m(z)$ and $M(z)$, the performance of estimator $\widehat{M}(z)$ plays a role in the convergence rate of $\widehat{m}(z)$. Moreover, $\widehat{M}(z)$ is

rather complicated and its properties cannot be derived by using the standard technique of LLKE, since it involves the two error terms, ξ and η. We begin by studying the asymptotic properties of $\widehat{M}(z)$.

Theorem 1. *If Conditions* (C1)−(C6) *hold, then*

$$\sqrt{Nh_N}[\widehat{M}(z) - M(z) - B(z)] \xrightarrow{\mathcal{L}} N(0, V(z)), \tag{2.11}$$

where

$$\begin{aligned} V(z) &= \left[\mathrm{Var}\,(\xi) + \lambda \mathrm{Var}\,(\eta) M^{'2}(z)\right] \frac{\int K^2(w)dw}{f_U(z)}, \\ B(z) &= \frac{1}{2} M^{''}(z) h_N^2 \int K(w) w^2 dw + O(b_n^2). \end{aligned} \tag{2.12}$$

Furthermore, if U is reversible,

$$B(z) = \frac{1}{2} M^{''}(z) h_N^2 \int K(w) w^2 dw - \frac{1}{2} M^{'}(z) U^{''}(U^{-1}(z)) b_n^2 \int L(w) w^2 dw,$$

where U^{-1} denotes the inverse function of U.

Remark 4. If $\lambda = 0$, which means $n >> N$, the asymptotic property of $\widehat{M}(z)$ is the same as that of the standard LLKE (Fan (1992)). In reality, λ is usually larger than 1, because the covariates $\{t_j\}_{j=N+1}^{N+n}$ are expensive to obtain. However, $\widehat{M}(z)$ still achieves the optimal convergence rate of the standard LLKE by choosing appropriate h_N and b_n.

The next theorem establishes the convergence rate of the estimator $\widehat{m}(z)$.

Theorem 2. *Suppose Conditions* (C1)−(C8) *hold. If $c_1 j^{-a_1} \le |\alpha_{kj}| \le c_2 j^{-a_1}$ and $|M_{kj}| \le c_3 j^{-a_2}$ for some positive constants c_1, c_2, c_3, a_1 and a_2, we have*

$$\int_\Omega E[\widehat{m}(t) - m(t)]^2 dt = O(N^{-1} q^{2b+1} + h_N^4 q^{2b+1} + b_n^4 q^{2b+1} + q^{-2-s}),$$

where $b = \max(a_1, 2a_1 - a_2)$, $s > 0$ is defined in Condition (C1).

Remark 5. The convergence rate of $\widehat{M}(z)$ presented in Theorem 1 has the mean square error of $\widehat{M}_{kl} - M_{kl}$ of $O_p(N^{-1} + h_N^4 + b_n^4)$. Similarly, the mean square error of $\widehat{\alpha}_{kl} - \alpha_{kl}$ is of order $O_p(n^{-1} + b_n^4)$, as shown in the proof of Theorem 2. Combining the convergence rates of $\widehat{M}_{kl} - M_{kl}$ and $\widehat{\alpha}_{kl} - \alpha_{kl}$ results in the first three terms of mean integrated square error of $\widehat{m}(z)$ presented in Theorem 2. Condition (C5) and Theorem 2 motivate us to choose h_N and b_n in the region

of $[N^{-1/3}, N^{-1/4}]$ as then the mean integrated squared error of $\widehat{m}(t)$ will mainly depend on N and q. In particular, undersmoothing at the stage of estimating $\widehat{M}$ decreases the estimation bias at the expense of variance to some extent, but further oversmoothing at the stage of Fourier transformation can compensate for the increase of variance.

2.4. Smoothing parameters selection

As is the case with any nonparametric regression procedure, an important choice to be made is the amount of local averaging performed to obtain the regression estimate. For the local polynomial regression estimator, bandwidth selection rules were considered in Ruppert, Sheather and Wand (1995) and Fan and Gijbels (1996), among others. Delaigle, Hall, and Qiu (2006) proposed an automatic way of choosing the smoothing parameters q and h_N as a combination of an existing plug-in bandwidth selector for LLKE and the cross-validation (CV) rule for trigonometric series. Since our estimator $\widehat{m}(z)$ involves three regularization parameters h_N, b_n, and q, we present modification of the leave-one-out CV selection criterion.

First, we use $\widehat{b}_n = b_{\mathrm{C}} n^{-1/20}$, where b_{C} is the delete-one CV bandwidth selector for the validation sample, that is the minimizer of

$$\mathrm{CV}(b_n) = \frac{1}{n} \sum_{j=N+1}^{N+n} (t_j - \widehat{U}^{(-j)}(\tilde{t}_j; b_n))^2,$$

where $\widehat{U}^{(-j)}(\cdot; b_n)$ denotes the leave-one-out version of $\widehat{U}$ using b_n. Here the multiplier $n^{-1/20}$ is recommended partially due to Condition (C5) and the order of the cross-validation selector that has an optimal $O_p(n^{-1/5})$ rate (Härdle, Hall, and Marron (1988)). This bandwidth undersmoothes $\widehat{U}$, which is appropriate from the explanation provided in Remark 5.

After obtaining $\widehat{b}_n$ and defining the corresponding $\widehat{U}$ with the selected bandwidth $\widehat{b}_n$, a cross-validation criterion for selecting h_N and q chooses

$$(\widehat{h}_N, \widehat{q}) = \arg\min_{h_N, q} \frac{1}{N} \sum_{i=1}^{N} [Y_i - \widehat{m}^{(-i)}(\widehat{U}(\tilde{t}_i); h_N, q)]^2,$$

where $\widehat{m}^{(-i)}(\cdot; h_N, q)$ denotes the version of $\widehat{m}$ that is constructed on omitting $(\tilde{t}_i, Y_i)$ from the surrogate data using h_N and q terms in the trigonometric series. In this criterion, $\widehat{U}(\tilde{t}_i)$ is used as an empirical approximation of t_i in the sense of ignoring the error item η_i, because in applications the true regressor t_i is not observable. Note that, since q is an integer, this two-dimensional CV criterion does not require extensive computation effort. Furthermore, based on our numerical results, it suffices to choose q from 1 to 5 in accordance with the empirical

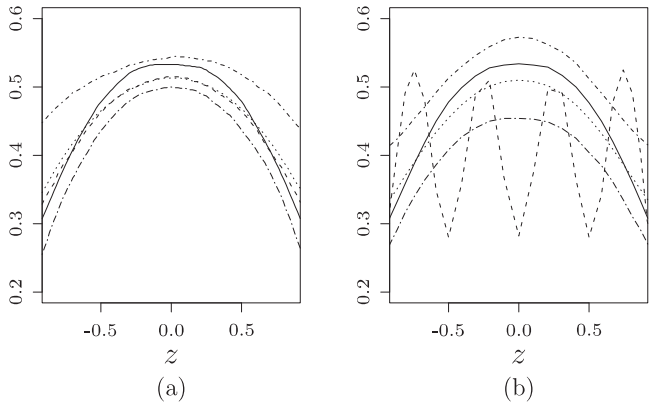

Figure 1. The estimated curves of $M(z)$ for the regression function (I) through 1,000 repetitions when (a): $(n, N) = (30, 120)$, U is given by (i); (b): $(n, N) = (30, 120)$, U is given by (iii). The solid, dashed, dotted, and two dashed-dotted curves represent $M(z)$, the median of $\widehat{m}_2(z)$, the median of $\widehat{M}(z)$, and the quartiles of $\widehat{M}(z)$, respectively.

findings in Delaigle, Hall, and Qiu (2006). For instance, when $(n, N) = (60, 120)$, it requires less than 3 seconds to complete the curve fitting on 201 grid points using a Pentium-M 2.4MHz CPU.

3. Numerical Performance Assessment

We conducted a simulation study to evaluate numerical properties of the proposed estimator using the smoothing parameters selection method described in Section 2.4. Our goal is to show the effectiveness and robustness of the proposed estimator $\widehat{m}(z)$, and thus we only chose certain representative examples for illustration.

We considered two regression functions $m(\cdot)$ taken from the examples of Delaigle, Hall, and Qiu (2006):

(I) $m(z) = (1 - z^2)^2 I_{\{z\in[-1,1]\}}$,

(II) $m(z) = (1 - z^2)^2 \exp(2z) I_{\{z\in[-1,1]\}}$,

where $I_{\{\cdot\}}$ is the indicator function. Three cases for U were considered: (i) $t = \tilde{t}+\delta$, the classical Berkson model; (ii) $t = \tilde{t}^2-1/4+\delta$; (iii) $t = \cos(2\pi\tilde{t})+\delta$, where δ is independent of $\tilde{t}$ and follows the uniform distribution U$[-1, 1]$. The $\tilde{t}$'s were generated from two distributions: (1) $\tilde{t} \sim N(0, 0.5^2)$; (2) $\tilde{t} \sim \exp(1)-1$, where the latter denotes the centered standard exponential distribution. Throughout this section the ε's are assumed to follow the normal distribution $N(0, 0.25^2)$. For each of several choices of m, U and the distribution of $\tilde{t}$, 1,000 simulated datasets were

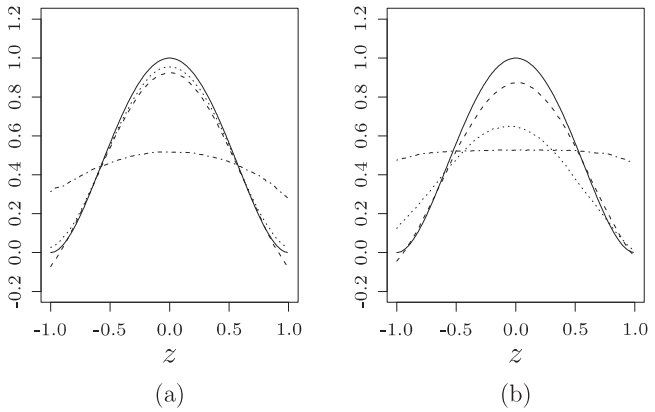

Figure 2. Median Curves of 1,000 estimators of regression function (I) with $(n, N) = (60, 120)$ when (a): U is given by (i), $f_{\tilde{t}}$ is given by (1); (b): U is given by (ii), $f_{\tilde{t}}$ is given by (1). The solid, dashed, dotted, and dashed-dotted curves represent $m(z)$, $\widehat{m}(z)$, $\widehat{m}_1(z)$, and $\widehat{m}_2(z)$, respectively.

generated for each sample size combination of $(n, N) = (30, 120), (60, 200)$. The kernel functions K and L used in (2.4) were both chosen to be the Epanechnikov kernel function $0.75(1 - x^2)I(-1 \leq x \leq 1)$, that has certain optimal properties (cf., Fan and Gijbels (1996)).

It is challenging to compare the proposed method with alternative methods, since there is no obvious comparable method in the literature. Here, we consider the method of Delaigle, Hall, and Qiu (2006) where a simple Berkson error structure is assumed (denoted as $\widehat{m}_1$). Another alternative is the naive LLKE based on the dataset $\{(Y_i, \tilde{t}_i)\}_{i=1}^N$ (denoted as $\widehat{m}_2$), although it is not a consistent estimate due to ignoring of the measurement error. For $\widehat{m}_1$, the method proposed in Delaigle, Hall, and Qiu (2006) for determining the parameters is considered. For $\widehat{m}_2$, the leave-one-out CV approach is used for choosing bandwidth h_N.

First, as the performance of our final estimator $\widehat{m}(z)$ relies heavily on the estimator $\widehat{M}(z)$, it is interesting to see how well $\widehat{M}(z)$ works. Figure 1 shows the regression function $M(z)$, the curves of the median and quartiles of 1,000 estimates $\widehat{M}(z)$ and the median of $\widehat{m}_2(z)$ with two different settings for the regression function (I). In both settings, the first $f_{\tilde{t}}$ is considered. Since the closed form of true function $M(z)$ is not available, we used simulation to approximate it through 100,000 repetitions. In the first setting where U is considered as the additive Berkson error model, $\widehat{m}_2(z)$ certainly looks a consistent estimator in Figure 1(a). However, this is not the case for the second setting in which the nonlinear error structure (III) is considered. Here, $\widehat{m}_2(z)$ is far from the true $M(z)$, as we would expect, but $\widehat{M}(z)$ still performs reasonably well as guaranteed by Theorem 1.

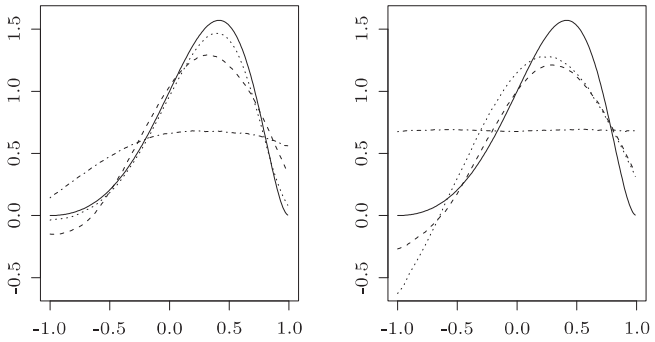

Figure 3. Median Curves of 1,000 estimators of regression function (II) with $(n, N) = (60, 120)$ when (a): U given by (i), $f_{\tilde{t}}$ given by (2); (b): U given by (ii), $f_{\tilde{t}}$ given by (2). The solid, dashed, dotted, and dashed-dotted curves represent $m(z)$, $\widehat{m}(z)$, $\widehat{m}_1(z)$ and $\widehat{m}_2(z)$ respectively.

Table 1. The estimated MISE comparison for the estimators $\widehat{m}(z)$, $\widehat{m}_1(z)$ and $\widehat{m}_2(z)$.

(n, N)	U	$f_{\tilde{t}}$	$\widehat{m}(z)$		$\widehat{m}_1(z)$ regression function m		$\widehat{m}_2(z)$	
			(I)	(II)	(I)	(II)	(I)	(II)
$(30, 120)$	(i)	(1)	1.73e-2	7.32e-2	1.05e-2	2.54e-2	1.14e-1	2.65e-1
	(i)	(2)	1.53e-2	6.95e-2	1.10e-2	2.35e-2	1.11e-1	2.74e-1
	(ii)	(1)	3.69e-2	8.95e-2	9.80e-2	1.47e-1	1.33e-1	3.67e-1
	(ii)	(2)	3.40e-2	8.19e-2	9.66e-2	1.39e-1	1.28e-1	3.68e-1
$(60, 120)$	(i)	(1)	1.25e-2	6.22e-2	7.49e-3	1.96e-2	1.01e-1	2.60e-1
	(i)	(2)	1.07e-2	5.80e-2	7.18e-3	1.75e-2	9.85e-2	2.48e-1
	(ii)	(1)	2.55e-2	7.77e-2	9.64e-2	1.39e-1	1.23e-1	3.45e-1
	(ii)	(2)	2.22e-2	7.49e-2	1.03e-1	1.58e-1	1.16e-1	3.31e-1

Figures 2 and 3 show the regression function curve $m(z)$, the curves of the medians of 1,000 estimates $\widehat{m}(z)$, $\widehat{m}_1(z)$ and $\widehat{m}_2(z)$ under different settings of U and $f_{\tilde{t}}$ for $(n, N) = (60, 120)$, in the two examples (I) and (II) respectively. From these two figures, we see that $\widehat{m}(z)$ can capture the patterns of the true curves as well as $\widehat{m}_1(z)$ does, although $\widehat{m}(z)$ tends to have larger bias at boundaries and peaks due to its explicit dependence on the size of the validation dataset. Taking sample size and noise level into account, the proposed estimator $\widehat{m}(z)$ and smoothing parameter selection method appear to perform very well for the test functions considered in this study, while $\widehat{m}_2(z)$ fails to produce the correct curves, as we expected.

Table 1 summarizes the results shown in Figures 2−3. The estimated mean

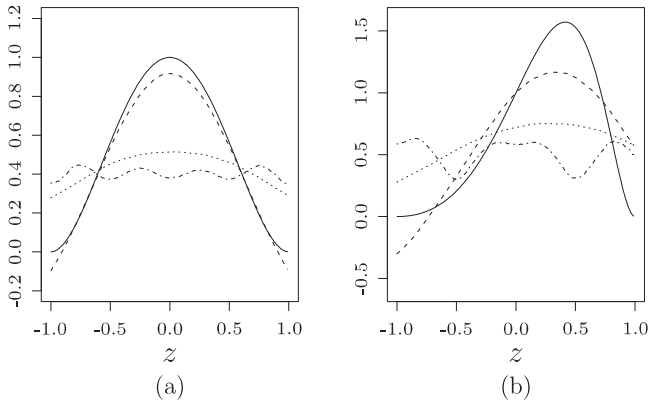

Figure 4. Median Curves of 1,000 estimators with $(n, N) = (60, 120)$ when (a): m is given by (I), U is given by (iii), $f_{\tilde{t}}$ is given by (2); (b): m is given by (II), U is given by (iii), $f_{\tilde{t}}$ is given by (2). The solid, dashed, dotted, and dashed-dotted curves represent $m(z)$, $\widehat{m}(z)$, $\widehat{m}_1(z)$, and $\widehat{m}_2(z)$, respectively.

integrated squared errors (MISE), evaluated on a grid of 201 equidistant values of z in [-1,1], are presented. We can see that $\widehat{m}_1(z)$ had certain advantage over $\widehat{m}(z)$ when the function (i) is considered because its assumed error model is correct in such a case. $\widehat{m}(z)$ had much smaller MISE than $\widehat{m}_1(z)$ when the nonlinear Berkson model (ii) were used. In these cases, the performance of $\widehat{m}(z)$ improved considerably as the sample sizes increased in terms of MISE, whereas $\widehat{m}_1(z)$ was hardly affected by those changes. It is worth pointing out that the superiority of $\widehat{m}(z)$ to $\widehat{m}_1(z)$ may become more significant as the difference between the U and the simple linear model gets more prominent. Figure 4 shows the median curve comparison when U is chosen as (iii) with $(n, N) = (60, 120)$. In both examples, $\widehat{m}_1(z)$ fails to capture the main profile of the underlying regression function. Note that in all the cases listed in Table 1, the $\widehat{m}_2(z)$ are totally incorrect which indicates that the proposed method is necessary.

4. A Data Example

In this section, we apply our approach to a dataset of enzyme reaction speeds collected in 1974. The reaction speed (Y) is calculated by the particle number of radioactive matter obtained by reaction per minute in basal density (t). The objective of this analysis is to relate Y to the basal density. There are two ways to measure basal density: a simple chemical method can be used, but with measurement errors; Another approach is to use a precision machine tool and an expensive procedure to produce a more accurate measure for a small subset of subjects enrolled in the study. The basal density obtained by the chemical

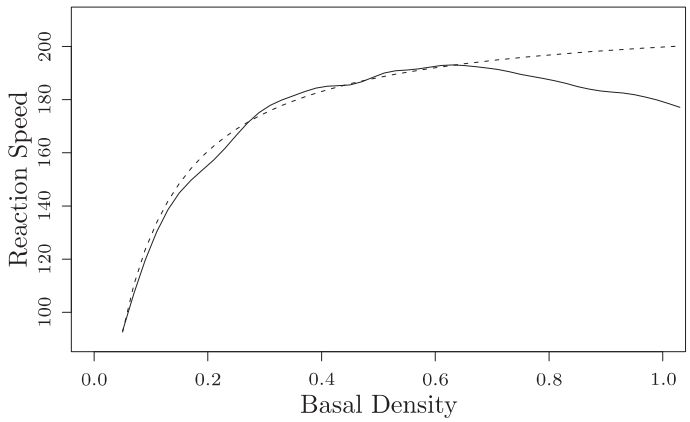

Figure 5. Our estimator $\widehat{m}(z)$ and the parametric model (4.1) fit of the regression function m, represented by the solid and dotted curves, respectively.

method is used as the surrogate variable $\tilde{t}$, and the corresponding exact measure for a small subset of subjects is used as the validation variable t.

This dataset has been analyzed by Stute, Xue and Zhu (2007), in which a nonparametric fit is applied to relate t and $\tilde{t}$, and the following non-linear model is considered as the underlying regression function m:

$$m(t, \boldsymbol{\beta}) = \frac{\beta_1 t}{\beta_2 + t}. \tag{4.1}$$

The dataset consists of $n = 10$ validation observations and $N = 30$ surrogate observations. Using the smoothing parameters selection method given in Section 2.4 results in three parameters $b_n = 0.453$, $h_N = 0.144$, and $q = 3$. Figure 5 displays the corresponding estimated curve ($\widehat{m}$), along with the curve produced by the parametric model (4.1) with $\beta_1 = 212.7$ and $\beta_2 = 0.06484$ obtained from Stute, Xue and Zhu (2007). It is readily seen that the two curves have similar patterns, although the estimator we proposed looks coarser, as the nonparametric smoother absorbs considerably more degrees of freedom than does the parametric approach. In addition, $\widehat{m}(z)$ differs considerably from the parametric fit at the right boundary. This is not surprising because of the lack of data in the interval $t \in [-0.8, 1.0]$. Thus again, it should be emphasized that compared with parametric methods, to use the proposed method requires relatively large sample size, especially at the boundary. We think this has become a less significant limitation with advances in various technologies. New instruments can capture more information, and a large amount of data become available in modern statistical analysis.

5. Discussion

In the foregoing investigation, we have assumed without comment that in the validation dataset only the true and surrogate covariables are observed. In fact, in certain applications (e.g., Chen (2002)), observations on the response variable are also available. In such situations, several naive nonparametric function estimators are obtainable. One is to directly use the matched $\{Y_j, t_j\}_{j=N+1}^{N+n}$ which is apparently inefficient because the size of the validation dataset is usually relatively small. An alternative estimator uses the proposed approach given in Section 2, except in replacing the surrogate dataset with an extended sample by combining $\{Y_j, \tilde{t}_j\}_{j=N+1}^{N+n}$ and $\{Y_i, \tilde{t}_i\}_{i=1}^{N}$. However, this estimator ignores the information provided by $\{Y_j, t_j\}_{j=N+1}^{N+n}$ and hence may not be efficient either. How to construct an estimator able to fully incorporate the information given by data in this case warrants further research.

Another question is how to estimate nonparametric curve in a high dimensional space when multi-covariates are contaminated with non-additive errors with the use of a validation data. As we know, the Fourier transformation is not directly applicable in this situation. To avoid the "curse of dimensionality", several dimension reduction models, such as the additive model, the single-index model, the partially linear model, and the varying coefficient model, have been studied and widely applied in applications. It is of interest to incorporate these dimension reduction methods into our proposed methodology to solve high-dimensional problems with validation data. Specially, the case in which only a univariate variable is measured with error in a multi-covariates model, say, $y = m(t, \mathbf{x}) + \varepsilon$, where t is measured with error but $\mathbf{x}$ is measured exactly, is a common situation. How to deal with this kind of problem is a topic for future study.

Acknowledgement

The authors thank the Editor, an associate editor, and two referees for many constructive comments and suggestions which greatly improved the quality of the paper. This research was supported by the NSF of Tianjin Grant 07JCYBJC04300, the NNSF of China Grants 10771107, 10711120448, 11071128 and 11001138.

Appendix: Proofs of Theorems

Unless otherwise stated, subscript i runs from 1 to N, j from $N+1$ to $N+n$, and denote surrogate and validation data, respectively. Let $C_1, C_2, \ldots$ be generic positive constants, not depending on M, h_N, b_n, n or N.

As Theorem 1 and its proof play an important role on establishing the other results, we detail the steps of its proof. First, we state the following necessary lemma. In what follows, we assume the function U is reversible; otherwise, the terms involving U^{-1} can be expressed as $O(1)$ and the corresponding convergence rate can be achieved.

Lemma 1. *If Conditions* C1−C6 *hold, then*

(i) $S_{N0} \xrightarrow{P} f_U(z)$;
(ii) $S_{N1}^2 = o_p(S_{N0}S_{N2}), \quad S_{N1}S_{N3} = o_p(S_{N2}^2)$;
(iii) $S_{N1}\phi_{N1} = o_p(S_{N2}\phi_{N0})$, *where*

$$\phi_{Nk} = \frac{1}{Nh_N}\sum_i K\left(\frac{\widehat{U}(\tilde{t}_i)-z}{h_N}\right)(\widehat{U}(\tilde{t}_i)-z)^k\xi_i, \quad k=0,1.$$

Proof. (i) Using a second-order Taylor expansion, we have

$$\begin{aligned} S_{N0} &:= \widehat{f}_U(z) \\ &= \frac{1}{Nh_N}\sum_i K\left(\frac{U(\tilde{t}_i)-z}{h_N}\right) + \frac{1}{Nh_N}\sum_i K'\left(\frac{U(\tilde{t}_i)-z}{h_N}\right)\frac{\left(\widehat{U}(\tilde{t}_i)-U(\tilde{t}_i)\right)}{h_N} \\ &\quad + \frac{1}{2Nh_N}\sum_i K''\left(\frac{U(\tilde{t}_i)-z}{h_N}\right)\frac{\left(\widehat{U}(\tilde{t}_i)-U(\tilde{t}_i)\right)^2}{h_N^2} + D_4 \\ &:= D_1 + D_2 + D_3 + D_4. \end{aligned}$$

Standard density theory (Parzen (1962)) leads to $D_1 = f_U(z) + O_p(h_N^2)$. Next, we focus on D_2 and D_3.

First, according to the asymptotic properties of LLKE (Fan and Gijbels (1996)), we have

$$\begin{aligned} \left(\widehat{U}(\tilde{t}_i)-U(\tilde{t}_i)\right) &= \left\{\frac{1}{nb_nf_{\tilde{t}}(\tilde{t}_i)}\sum_j\left(t_j-U(\tilde{t}_j)\right)L\left(\frac{\tilde{t}_j-\tilde{t}_i}{b_n}\right)\right. \\ &\quad \left. + \frac{1}{2nb_nf_{\tilde{t}}(\tilde{t}_i)}\sum_j\left[U''(\tilde{t}_i)\left(\tilde{t}_j-\tilde{t}_i\right)^2\right]L\left(\frac{\tilde{t}_j-\tilde{t}_i}{b_n}\right)\right\}(1+o_p(1)) \\ &:= (U_V+U_B)(1+o_p(1)). \end{aligned} \tag{A.1}$$

Following Condition (C5) and (A.1), the uniform convergence rate of the local linear smoother (e.g., Carroll et al. (1997)) reveals that

$$\sup_{\tilde{t}\in\Omega}\left|\widehat{U}(\tilde{t})-U(\tilde{t})\right| = O_p(\gamma_n). \tag{A.2}$$

Moreover, for any function $g(x) \in \{K(x), K^{'}(x), K^{''}(x)\}$, by taking the expectation one can show that

$$\frac{1}{Nh_N}\sum_i \left| g\left(\frac{U(\tilde{t}_i)-z}{h_N}\right)\right| = O_p(1). \tag{A.3}$$

By using (A.2) and (A.3), we can verify that

$$\begin{aligned}|D_2| &\le \frac{1}{Nh_N^2}\sum_i \left|K^{'}\left(\frac{U(\tilde{t}_i)-z}{h_N}\right)\right| \times \left|\widehat{U}(\tilde{t}_i)-U(\tilde{t}_i)\right| \\ &\le \sup_{\tilde{t}\in\Omega}\left|\widehat{U}(\tilde{t})-U(\tilde{t})\right| \times \frac{1}{Nh_N^2}\sum_i \left|K^{'}\left(\frac{U(\tilde{t}_i)-z}{h_N}\right)\right| = O_p\left(\frac{\gamma_n}{h_N}\right), \end{aligned} \tag{A.4}$$

$$\begin{aligned}|D_3| &\le \frac{1}{2Nh_N^3}\sum_i \left|K^{''}\left(\frac{U(\tilde{t}_i)-z}{h_N}\right)\right| \left|\widehat{U}(\tilde{t}_i)-U(\tilde{t}_i)\right|^2 \\ &\le \sup_{\tilde{t}\in\Omega}\left|\widehat{U}(\tilde{t})-U(\tilde{t})\right|^2 \times \frac{1}{2Nh_N^3}\sum_i \left|K^{''}\left(\frac{U(\tilde{t}_i)-z}{h_N}\right)\right| = O_p\left(\frac{\gamma_n^2}{h_N^2}\right). \end{aligned} \tag{A.5}$$

Similarly, we can derive $D_4 = o_p(D_3)$. Combining this with (A.4)−(A.5), we can complete the proof.

(ii) By using similar arguments but tedious algebra in (i), and Conditions (C4) and (C5), we get $S_{N1} = O_p(h_N^2+\gamma_n), S_{N2} = O_p(h_N^2+\gamma_n h_N), S_{N3} = O_p(h_N^4+\gamma_n h_N^2)$, which directly lead to the result (ii).

(iii) The proof follows from similar but tedious calculations, and hence is omitted.

Proof of Theorem 1. By Lemma 1, collecting the leading terms, $\widehat{M}(z)-M(z)$ can be written as

$$\begin{aligned}\widehat{M}(z)-M(z) &= \left\{\frac{(Nh_N)^{-1}\sum_i K\left((\widehat{U}(\tilde{t}_i)-z)/h_N\right)\left[Y_i - M(\widehat{U}(\tilde{t}_i))\right]}{f_U(z)}\right. \\ &\quad \left. + \frac{(M^{''}(z)/2Nh_N)\sum_i K\left((\widehat{U}(\tilde{t}_i)-z)/h_N\right)\left[\widehat{U}(\tilde{t}_i)-z\right]^2}{f_U(z)}\right\}(1+o_p(1)) \\ &:= \frac{M_V+M_B}{f_U(z)}(1+o_p(1)),\end{aligned}$$

where M_B and M_V can be rewritten as

$$M_B = \left\{\frac{1}{2Nh_N}\sum_i K\Big(\frac{\widehat{U}(\tilde{t}_i)-z}{h_N}\Big)M^{''}(z)\left[U(\tilde{t}_i)-z\right]^2\right.$$

$$
\begin{aligned}
&+\frac{1}{Nh_N}\sum_i K\Big(\frac{\widehat{U}(\tilde{t}_i)-z}{h_N}\Big)M^{''}(z)\left[\hat{U}(\tilde{t}_i)-U(\tilde{t}_i)\right]^2\Bigg\}(1+o_p(1))\\
&:=\{M_{B1}+M_{B2}\}(1+o_p(1)),\\
M_V \;=\; &\frac{1}{Nh_N}\sum_i K\left(\frac{\widehat{U}(\tilde{t}_i)-z}{h_N}\right)\left[Y_i-M(U(\tilde{t}_i))\right]\\
&-\frac{1}{Nh_N}\sum_i K\left(\frac{\widehat{U}(\tilde{t}_i)-z}{h_N}\right)\left[M(\widehat{U}(\tilde{t}_i))-M(U(\tilde{t}_i))\right]\\
&:=M_{V1}-M_{V2}.
\end{aligned}
$$

For M_{B1}, a Taylor expansion of the kernel function yields

$$
\begin{aligned}
M_{B1} \;=\; &\frac{M^{''}(z)}{2Nh_N}\sum_i K\left(\frac{U(\tilde{t}_i)-z}{h_N}\right)\left(U(\tilde{t}_i)-z\right)^2\\
&+\frac{M^{''}(z)}{2Nh_N^2}\sum_i K^{'}\left(\frac{U(\tilde{t}_i)-z}{h_N}\right)\left(\widehat{U}(\tilde{t}_i)-U(\tilde{t}_i)\right)\left(U(\tilde{t}_i)-z\right)^2(1+o_p(1))\\
&:=M_{B11}+M_{B12}\,(1+o_p(1))\,.
\end{aligned}
$$

Simple calculations yield that

$$
\begin{aligned}
M_{B11} &= \frac{1}{2h_N}E\left[K\left(\frac{U(\tilde{t})-z}{h_N}\right)M^{''}(z)(U(\tilde{t})-z)^2\right](1+o_p(1))\\
&= \frac{1}{2}\int K(w)w^2dwM^{''}(z)f_U(z)h_N^2+o_p(h_N^2),
\end{aligned}
$$

which is the same as the bias term of LLKE. As for M_{B12}, using (A.2) and (A.3),

$$
\begin{aligned}
M_{B12} &\le \sup_{\tilde{t}\in\Omega}\left|\widehat{U}(\tilde{t})-U(\tilde{t})\right|\times\left|M^{''}(z)\right|\frac{1}{2Nh_N^2}\sum_i\left|K^{'}\left(\frac{U(\tilde{t}_i)-z}{h_N}\right)\right|\left(U(\tilde{t}_i)-z\right)^2\\
&= O_p\left(\gamma_n h_N\right).
\end{aligned}
$$

Hence, we conclude that

$$
M_{B1}=\frac{1}{2}\int K(w)w^2dwM^{''}(z)f_U(z)h_N^2+o_p(h_N^2). \tag{A.6}
$$

Similarly, we get

$$
M_{B2}=O_p(\gamma_n^2). \tag{A.7}
$$

Using (A.6)−(A.7) with Condition (C5) leads to

$$
M_B=\frac{1}{2}\int K(w)w^2dwM^{''}(z)f_U(z)h_N^2+o_p(h_N^2). \tag{A.8}
$$

Now we turn to M_V. First, by using a Taylor expansion and (A.2), we can show that

$$M_{V1} = \Delta_1 + O_p(\frac{1}{\sqrt{Nh_N}} \times \frac{\gamma_n}{h_N}), \tag{A.9}$$

where we write $\Delta_1 = \frac{1}{Nh_N}\sum_i K\left(\frac{U(\tilde{t}_i)-z}{h_N}\right)\xi_i$. A simple calculation yields

$$\mathrm{Var}\,(M_{V1}) = \frac{\mathrm{Var}\,(\xi) f_U(z)}{Nh_N}\int K^2(w)dw(1+o(1)). \tag{A.10}$$

Finally, we focus on M_{V2}, the main term in the variance:

$$\begin{aligned} M_{V2} =& \left\{\frac{1}{Nh_N}\sum_i K\left(\frac{U(\tilde{t}_i)-z}{h_N}\right) M^{'}(U(\tilde{t}_i))\left(\widehat{U}(\tilde{t}_i)-U(\tilde{t}_i)\right)\right. \\ &\left. +\frac{1}{Nh_N^2}\sum_i K^{'}\left(\frac{U(\tilde{t}_i)-z}{h_N}\right) M^{'}(U(\tilde{t}_i))\left(\widehat{U}(\tilde{t}_i)-U(\tilde{t}_i)\right)^2\right\}(1+o_p(1)) \\ :=& \{M_{V21}+M_{V22}\}(1+o_p(1)), \end{aligned}$$

where we can show that $M_{V22} = O_p\left(\gamma_n^2/h_N\right)$, while

$$\begin{aligned} M_{V21} =& \left\{\frac{1}{Nh_N}\sum_i K\left(\frac{U(\tilde{t}_i)-z}{h_N}\right) M^{'}(U(\tilde{t}_i))(U_V+U_B)\right\}(1+o_p(1)) \\ =& \left\{\frac{1}{nh_N}\sum_j \eta_j K\left(\frac{U(\tilde{t}_j)-z}{h_N}\right) M^{'}(U(\tilde{t}_j))\right. \\ &\left. +\frac{1}{2Nh_N}\sum_i K\left(\frac{U(\tilde{t}_i)-z}{h_N}\right) M^{'}(U(\tilde{t}_i))U^{''}(\tilde{t}_i)b_n^2\int L(w)w^2dw\right\}(1+o_p(1)) \\ :=& \left\{\Delta_2 + \frac{1}{2}M^{'}(z)U^{''}[U^{-1}(z)]f_U(z)b_n^2\int L(w)w^2dw\right\}(1+o_p(1)). \end{aligned} \tag{A.11}$$

Note that

$$E[\Delta_2]=0, \quad E[\Delta_2^2] = \frac{\mathrm{Var}\,(\eta)}{nh_N}M^{'2}(z)f_U(z)\int K^2(w)dw. \tag{A.12}$$

Recalling the independence between validation data and surrogate data, and combining (A.8)−(A.12), the Central Limit Theorem leads to the result.

Proof of Theorem 2. Based on the proof of Theorem 1, we have

$$\widehat{M}(z) - M(z) = (\Delta_1(z)+\Delta_2(z)+B(z))(1+o_p(1)), \tag{A.13}$$

where Δ_1, Δ_2, and B are defined in (A.9), (A.11) and (2.12), respectively. By using (A.13), it can be seen that

$$|\,\mathrm{Cov}\,[\widehat{M}(z_1), \widehat{M}(z_2)]|$$

$$\leq \left| \frac{\mathrm{Var}\,(\eta)}{n^2 h_N^2} \sum_j K\left(\frac{z_1 - U(\tilde{t}_j)}{h_N}\right) K\left(\frac{z_2 - U(\tilde{t}_j)}{h_N}\right) [M'(U(\tilde{t}_j))]^2 \right|$$
$$+ \left| \frac{\mathrm{Var}\,(\xi)}{N^2 h_N^2} \sum_i K\left(\frac{z_1 - U(\tilde{t}_i)}{h_N}\right) K\left(\frac{z_2 - U(\tilde{t}_i)}{h_N}\right) \right| + o(h_N^4 + b_n^4 + (Nh_N)^{-1})$$
$$\leq \frac{C_1}{Nh_N} K * K\left(\frac{z_2 - z_1}{h_N}\right) + o(h_N^4 + b_n^4 + (Nh_N)^{-1}). \tag{A.14}$$

Now (A.13), (A.14) and the definitions of $\widehat{M}_0$ and $\widehat{M}_{kl}$ imply that

$$\begin{aligned} E(\widehat{M}_0 - M_0)^2 &\leq C_2 N^{-1} + C_3 h_N^4 + C_4 b_n^4, \\ E(\widehat{M}_{kl} - M_{kl})^2 &\leq C_2 N^{-1} + C_3 h_N^4 + C_4 b_n^4, \end{aligned} \tag{A.15}$$

uniformly in k and l. Hence, we have

$$\begin{aligned} &\int_\Omega E(\widehat{m}(t) - m(t))^2 dt \\ &\quad = \int_\Omega E\bigg((\widehat{m}_0 - m_0) + \sum_{l=1}^{q} [(\widehat{m}_{1l} - m_{1l}) \cos(lt) + (\widehat{m}_{2l} - m_{2l}) \sin(lt)] \bigg)^2 dt \\ &\qquad + \int_\Omega \bigg(\sum_{l=q+1}^{\infty} [m_{1l} \cos(lt) + m_{2l} \sin(lt)] \bigg)^2 dt \\ &\quad = 2\pi E(\widehat{m}_0 - m_0)^2 + \pi \sum_{l=1}^{q} \sum_{k=1}^{2} E(\widehat{m}_{kl} - m_{kl})^2 + R_q. \end{aligned}$$

Under Conditions (C7) and (C8), by some modifications of the proof of Theorem 1 in Akritas and Van Keilegom (2001), we can show

$$E(\widehat{\alpha}_{kj} - \alpha_{kj})^2 \leq C_5 n^{-1} + C_6 b_n^4,$$

uniformly in k and l. This property, the conditions given in Theorem 2, and (A.15) yield that

$$\sum_{l=1}^{q} \sum_{k=1}^{2} E(\widehat{m}_{kl} - m_{kl})^2 \leq (C_7 N^{-1} + C_8 h_N^4 + C_9 b_n^4) \sum_{l=1}^{q} l^{2b}.$$

Note that by the general theory of trigonometric series we have

$$\sum_{l=q+1}^{\infty} [m_{1l} \cos(lt) + m_{2l} \sin(lt)] \leq C_{10} q^{-2-s}$$

uniformly in t. Therefore, R_q is of the order q^{-2-s}. Taking these results together we can complete the proof.

References

Akritas, M. G. and Van Keilegom, I. (2001). Non-parametric estimation of the residual distribution. *Scand. J. Statist.* **28**, 549-568.

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

Carroll, R. J., Fan, J., Gijbels, I. and Wand, M. P. (1997). Generalized partially linear single-index models. *J. Amer. Statist. Assoc.* **92** 477-489.

Carroll, R. J., Knickerbocker, R. K. and Wang, C. Y. (1995). Dimension reduction in a semiparametric regression model with errors in covariates. *Ann. Statist.* **23**, 161-181.

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. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. Second edition. Chapman and Hall, London.

Carroll, R. J. and Stefanski, L. A. (1990). Approximate quasi-likelihood estimation in models with surrogate predictors. *J. Amer. Statist. Assoc.* **85**, 652-663.

Carroll, R. J. and Wand, M. P. (1991). Semiparametric estimation in logistic measurement error models. *J. Roy. Statist. Soc. Ser. B* **53**, 573-585.

Chen, Y. (2002). Cox regression in cohort studies with validation sampling. *J. Roy. Statist. Soc. Ser. B* **64**, 51-62.

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., Fan, J. and Carroll, R. J. (2009). A design-adaptive local polynomial estimator for the errors-in-variables problem. *J. Amer. Statist. Assoc.* **104**, 348-359.

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

Fan, J. (1992). Design-adaptive nonparametric regression. *J. Amer. Statist. Assoc.* **87**. 998-1004.

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

Fan, J. and Truong, Y. K. (1993). Nonparametric regression with errors in variables. *Ann. Statist.* **21**, 1990-1925.

Fuller, W. A. (1987). *Measurement Errors Models*. John Wiley, New York.

Härdle, W., Hall, P. and Marron, J. S. (1988). How far are automatically chosen regression smoothing parameter from theire optimum? (with discussion). *J. Amer. Statist. Assoc.* **83**, 86-99.

Lee, L. F. and Sepanski, J. (1995). Estimation of linear and nonlinear errors-in-variables models using validation data. *J. Amer. Statist. Assoc.* **90**, 130-140.

Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.* **33**, 1065-1076.

Pepe, M. S. (1992). Inference using surrogate outcome data and a validation sample. *Biometrika* **79**, 355-365.

Pepe, M. S. and Fleming, T. R. (1991). A general nonparametric method for dealing with errors in missing or surrogate covaraite data. *J. Amer. Statist. Assoc.* **86**, 108-113.

Ruppert, D., Sheather, S. J. and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *J. Amer. Statist. Assoc.* **90**, 1257-1270.

Sepanski, J. H. and Carroll, R. J. (1993). Semiparametric quasilikelihood and variance function estimation in measurement error models. *J. Econom.* **58**, 223-256.

Sepanski, J. H., Knickerbocker, R. K. and Carroll, R. J. (1994). A semiparametric correction for attenuation. *J. Amer. Statist. Assoc.* **89**, 1366-1373.

Sepanski, J. H. and Lee, L. F. (1995). Semiparametric estimation of nonlinear errors-in-variables models with validation study. *J. Nonparametr. Stat.* **4**, 365-394.

Staudenmayer, J. H. and Ruppert, D. (2004). Local polynomial regression and simulation-extrapolation. *J. Roy. Statist. Soc. Ser. B.* **66**, 17-30.

Stute, W., Xue, L. and Zhu, L. (2007). Empirical likelihood inference in nonlinear errors-in-covariables models with validation data. *J. Amer. Statist. Assoc.* **102**, 332-346.

Wang, Q. (1999). Estimation of partial linear errors-in-variables models with validation data. *J. Multivariate Anal.* **69**, 30-64.

Wang, Q. (2006). Nonparametric regression function estimation with surrogate data and validation sampling. *J. Multivariate Anal.* **97**, 1142-1161.

Wang, Q. and Rao, J. N. K. (2002). Empirical likelihood-based inference in linear errors-in-covariables models with validation data. *Biometrika* **89**, 345-358.

Wittes, J., Lakatos, E. and Probstfield, J. (1989). Surrogate endpoints in clinical trials: Cardiovascular diseases. *Statist. Medicine* **8**, 4150-425.

LPMC and Department of Statistics, School of Mathematical Sciences, Nankai University, Tianjin, 300071, China.

E-mail: feilen45@yahoo.com.cn

LPMC and Department of Statistics, School of Mathematical Sciences, Nankai University, Tianjin, 300071, China.

E-mail: chlzou@yahoo.com.cn

LPMC and Department of Statistics, School of Mathematical Sciences, Nankai University, Tianjin, 300071, China.

E-mail: zjwang@nankai.edu.cn

(Received February 2009; accepted January 2010)