Accelerated convergence for nonparametric regression with coarsened predictors

Aurore Delaigle[*,**,1], Peter Hall[**,2] and Hans-Georg Müller[***,3]

[*]Department of Mathematics, University of Bristol, BS8 1TW, UK

[**]Department of Mathematics and Statistics, University of Melbourne, Parkville, VIC, 3010, Australia

[***]Department of Statistics, University of California, Davis, CA 95616 USA

Short title: Accelerated Convergence

February 2, 2007

Abstract

We consider nonparametric estimation of a regression function for a situation where precisely measured predictors are used to estimate the regression curve for coarsened, i.e., less precise or contaminated predictors. Specifically, while one has available a sample $(W_1, Y_1), \ldots, (W_n, Y_n)$ of independent and identically distributed data, representing observations with precisely measured predictors, where $\mathrm{E}(Y_i|W_i) = g(W_i)$, instead of the smooth regression function g, the target of interest is another smooth regression function m that pertains to predictors X_i that are noisy versions the W_i. Our target is then the regression function $m(x) = E(Y|X = x)$, where X is a contaminated version of W, i.e., $X = W + \delta$. It is assumed that either the density of the errors is known, or replicated data are available resembling, but not necessarily the same as, the variables X. In either case, and under suitable conditions, we obtain $\sqrt{n}$-rates of convergence of the proposed estimator and its derivatives, and establish a functional limit theorem. Weak convergence to a Gaussian limit process implies pointwise and uniform confidence intervals and $\sqrt{n}$-consistent estimators of extrema and zeros of m. It is shown that these results are preserved under more general models in which X is determined by an explanatory variable. Finite sample performance is investigated in simulations and illustrated by a real data example.

Key words and phrases: confidence bands, errors-in-variables, estimation of extremes, functional limit theorem, smoothing, uniform convergence, weak convergence

[1]Supported by a Hellman Fellowship and by a Belgian American and Educational Foundation Post-Doctoral Fellowship, held at Department of Statistics, University of California, Davis.

[2]Supported in part by an Australian Research Council Fellowship.

[3]Supported in part by National Science Foundation grants DMS02-04869, DMS03-54448 and DMS05-05537.

1 Introduction

1.1. Motivation and models. In this paper, we consider nonparametric estimation of a regression function in the framework of a novel errors-in-variables problem. In the classical errors-in-variables problem, the interest is to estimate a regression function m, where

$$Y = m(G) + \epsilon,$$

and a sample $(F_1, Y_1), \ldots, (F_n, Y_n)$ of independent and identically distributed (i.i.d.) data is available, with $F_i = G_i + \delta_i$, where G and δ are independent random variables and the distribution of δ is known. References include Fan et al. (1991), Fan and Masry (1992), Fan and Truong (1993), Stefanski and Cook (1995), Carroll et al. (1995), Carroll et al. (1999), Taupin (2001), Devanarayan and Stefanski (2002), Ioannides and Matzner-Lober (2002), Linton and Whang (2002), and Carroll and Hall (2004).

The situation we consider here is different: we assume that an i.i.d. sample $(W_1, Y_1), \ldots, (W_n, Y_n)$ is observed, where

$$Y_i = g(W_i) + \varepsilon_i \quad \text{for} \quad 1 \leq i \leq n, \tag{1.1}$$

with independent errors ε_i with mean zero and finite variance. Instead of estimating the regression function $g(w) = E(Y|W = w)$ generating the observations, the goal is to estimate the target regression function $m(x) = E(Y|X = x)$, which differs from g, as X is a contaminated (coarsened) version of W.

Specifically, $X \sim f_X$ and $X = W + \delta$, where $\delta \sim f_\delta$ represents a random distortion, and W and δ are independent random variables. We refer to X as coarsened predictor of Y. In section 1.3 we shall note that the model for X can be generalized, without altering the main properties of our methods, to the situation where X is a proxy for a variable T related to W, provided we have additional data to infer the relationship between T and X.

The motivating idea is that the sample $(W_1, Y_1), \ldots, (W_n, Y_n)$, where one has precise predictors, is hard to obtain, and therefore future values of Y will be predicted from easier-to-obtain contaminated observations X of W. This type of problem arises in situations where it is expensive or involved to measure W accurately, so that, in routine applications, only the contaminated and less precise predictors X are available. At the same time, a training set is available containing more precise predictors. For example, if we have a sample of repeated contaminated observations of the predictor for several individuals, the averaged observations $W_i = \bar{X}_{i\cdot}$ will provide relatively accurate measurements of the predictor.

The problem we address is how to use the information in the training sample, with its accurate measurements, to predict a future response Y from a future contaminated predictor X. One of our central findings is that this coarsening of the predictor has the consequence of accelerating the convergence of the proposed estimator of m from the usual nonparametric rate, strictly slower than $\sqrt{n}$, to a parametric $\sqrt{n}$-rate, even if the target regression function is known only to be smooth and does not follow any particular parametric model.

In the setting of (1.1), m is generally not identifiable unless we know f_δ. The latter assumption is commonly made in errors-in-variable problems. See, for example, Stefanski and Carroll (1990) and Fan (1991). However, if we have additional data directly on δ, or if the data at (1.1) are replicated, then we can identify $m(x)$ without knowing f_δ. In either of these settings we might have a parametric model for f_δ, or we might wish to treat inference about f_δ from a nonparametric viewpoint. In order to show that estimation of m is a semiparametric problem, even if f_δ is not known and we treat it nonparametrically, we shall consider a more general, relatively "uninformative" type of replication, where we observe only

$$U_{ij} = V_i + \delta_{ij} \quad \text{for} \quad 1 \le j \le n_i \quad \text{and} \quad 1 \le i \le N\,. \tag{1.2}$$

Here, $V_1, \ldots, V_N$ are arbitrary random variables, $\delta_{11}, \ldots, \delta_{Nn_N}$ are mutually independent, the δ_{ij} are all distributed as δ, and it is assumed that each $n_i \ge 2$. Our results demonstrate that it is possible to attain $\sqrt{n}$-consistency without making joint assumptions about the data at (1.1) and (1.2). In particular, it is not necessary to suppose that the U_{ij} are independent of the (W_i, δ_i) or that the V_i are independent of the δ_{ij}. A direct application of the model in (1.2) is where U_{ij} are replicated measurements of X_i, and $V_i = W_i$.

1.2. Estimators. First we express m as a ratio, where each component can be estimated separately. Since $m(x) = \mathrm{E}(Y|X = x) = \mathrm{E}(g(W)|X = x)$ then

$$m(x) = \frac{\int g(w) f_{X|W}(x|w) f_W(w)\,dw}{f_X(x)} = \frac{\int g(w) f_\delta(x-w) f_W(w)\,dw}{\int f_\delta(x-w) f_W(w)\,dw} = \frac{\varphi(x)}{\psi(x)}, \tag{1.3}$$

where we define $\psi(x) = \int f_\delta(x-w) f_W(w)\,dw = \mathrm{E}(f_\delta(x-W))$ and

$$\varphi(x) = \int g(w) f_\delta(x-w) f_W(w)\,dw = \mathrm{E}(g(W) f_\delta(x-W)) = \mathrm{E}(Y f_\delta(x-W)).$$

If the data (W_i, Y_i) are generated by the model (1.1), and f_δ is assumed known, then the representations above motivate the estimators

$$\hat\varphi(x) \;=\; n^{-1} \sum_{i=1}^{n} Y_i f_\delta(x - W_i),$$

$$\hat\psi(x) \;=\; n^{-1}\sum_{i=1}^n f_\delta(x - W_i)$$

of $\varphi(x)$ and $\psi(x)$, respectively, leading to the estimators

$$\hat m(x) = \frac{\sum_{i=1}^n Y_i f_\delta(x - W_i)}{\sum_{i=1}^n f_\delta(x - W_i)} = \frac{\hat\varphi(x)}{\hat\psi(x)} \tag{1.4}$$

of $m(x)$. An attractive feature of $\hat m$ is that it does not require a smoothing parameter.

When additional data following (1.2) are available, we propose a Fourier-inversion approach to estimating φ and ψ, as follows. Assume that δ has a symmetric distribution, with positive characteristic function $f_\delta^{\rm ft}$:

$$f_\delta^{\rm ft}(t) = \Re f_\delta^{\rm ft}(t) > 0 \quad \text{for all real } t, \tag{1.5}$$

where the superscript ft denotes Fourier transform, and the Fourier transform of a function f is given by $f^{\rm ft}(t) = \int f(x)\, e^{itx}\, dx$. The real part of $f^{\rm ft}$ is denoted by $\Re f^{\rm ft}$. (Our methods can be generalized to the case of asymmetric error distributions, using techniques borrowed from Li and Vuong (1998).) Our estimator of $f_\delta^{\rm ft}$ is

$$\hat f_\delta^{\rm ft}(t) = \left| \frac{1}{M} \sum_{j=1}^N \sum\sum_{1 \le k_1 < k_2 \le n_j} \exp[it\,(U_{jk_1} - U_{jk_2})] \right|^{1/2}, \tag{1.6}$$

where $M = \frac12 \sum_{j=1}^N n_j(n_j - 1)$. (Here and below, $\hat f^{\rm ft}$ denotes an estimator of the Fourier transform of f, not the Fourier transform of an estimator $\hat f$ of f.)

Writing f_W for the density of W, estimators of the Fourier transforms, $f_W^{\rm ft}$ and $(f_W g)^{\rm ft}$, of f_W and $f_W g$ are respectively given by

$$\hat f_W^{\rm ft}(t) = \frac{1}{n} \sum_{j=1}^n \exp(itW_j)\,, \quad \widehat{(f_W g)}^{\rm ft}(t) = \frac{1}{n} \sum_{j=1}^n Y_j\, \exp(itW_j)\,. \tag{1.7}$$

Estimators of ψ and φ, based on Fourier inversion, are then obtained as

$$\tilde\psi(x) = \frac{1}{2\pi} \int_{|t| \le \tau_n} \hat f_W^{\rm ft}(t)\, \hat f_\delta^{\rm ft}(t)\, e^{-itx}\, dt\,, \quad \tilde\varphi(x) = \frac{1}{2\pi} \int_{|t| \le \tau_n} \widehat{(f_W g)}^{\rm ft}(t)\, \hat f_\delta^{\rm ft}(t)\, e^{-itx}\, dt\,, \tag{1.8}$$

where τ_n is a smoothing parameter. Our estimator of m is $\widetilde m = \Re\tilde\varphi / \Re\tilde\psi$.

1.3. Generalizations. The main features of our approach also apply to the more general case where $X = p(T \mid \theta) + \delta$, i.e. where $W = p(T \mid \theta)$ for a r.v. T, and (T, Y) rather than (W, Y) is observed in a subset of the available data. Here $p(\,\cdot \mid \theta)$ is a parametric model, determined by the finite parameter θ.

In this setting, we would ideally take $W_i = p(T_i \mid \theta)$. However, in most cases, we have to settle instead for $\widehat W_i = p(T_i \mid \hat\theta)$, where $\hat\theta$ is a $\sqrt{n}$-consistent estimator of θ,

computed by least-squares from data (T_i', X_i'), with the same distribution as (T, X), and related by $X_i' = p(T_i' \mid \theta) + \delta_i'$, for $1 \le i \le r$, say. The most important special case is that where p is linear: $p(t \mid \theta) = \theta^{(1)} + \theta^{(2)}\, t$, with $\theta = (\theta^{(1)}, \theta^{(2)})$ denoting a vector of length two.

In this model, the variable X typically represents a proxy for the variable T, where T often is not available in applications, because it is too costly to measure it, for example. In some applications, however, we are able to observe (T_i, X_i, Y_i) for $1 \le i \le n$ in a "training set", where $r = n$. We then propose to use the estimator $\hat{m}$, rather than a more conventional nonparametric regression based on (X_i, Y_i), since it is more accurate, as we will demonstrate. In some cases the training set (T_i', X_i') might be genuinely different from (T_i, X_i). For example (T_i', X_i') might represent external data.

In the case where $X = p(T \mid \theta) + \delta$ the estimators $\hat{m}$ and $\widetilde{m}$ differ only in that we replace W_i by $\widehat{W}_i$ at each appearance. Under appropriate regularity conditions the main properties of $\hat{m}$ and $\widetilde{m}$, and in particular their $\sqrt{n}$-consistency (provided $n = O(r)$), do not change. This point will be discussed in section 2.

2 Asymptotic results

2.1. Case where f_δ is assumed known. Here we discuss asymptotic properties of the estimator defined at (1.4). A central result is the weak convergence of a suitably scaled estimator process, with $\sqrt{n}$-scaling, to a Gaussian limit process in the location argument x. This result (Theorem 1 below) implies, among other matters, pointwise and uniform limits, local and simultaneous confidence bands, and convergence of estimated extrema locations.

We assume throughout section 2 that the distribution corresponding to f_δ is absolutely continuous, and in section 2.1 that f_δ has a bounded derivative. In section 2.1 it is not necessary to suppose that the densities f_W or $f_{W,Y}$ exist, although it is convenient to use the notation f_W and $f_{W,Y}$ when introducing the quantities needed to state and derive our results. However, the differential elements $f_W(w)\, dw$ and $f_{W,Y}(w, y)\, dw\, dy$ may be interpreted as $F_W(dw)$ and $F_{W,Y}(dw, dy)$, respectively; the distributions need not be absolutely continuous.

Given an integer $\nu \ge 0$, and assuming all quantities are well defined, let φ and ψ be as in Section 1 and define

$$\begin{aligned} h(x, w) &= f_\delta^{(\nu)}(x - w), \\ \alpha(x) &= \varphi^{(\nu)}(x) = \iint y f_{W,Y}(w, y) h(x, w)\, dw\, dy, \end{aligned}$$

$$\beta(x) \quad = \quad \psi^{(\nu)}(x) = \int h(x,w) f_W(w)\, dw.$$

Below, the notation D denotes a compact set on which we shall estimate φ and ψ. The following conditions, indexed by $\nu = 0$ or 1, will be used to prove our results. For $\nu = 1$ they can be relaxed to an assertion about the modulus of continuity for the corresponding quantity when $\nu = 0$; we impose the more stringent condition only for simplicity and brevity.

($\mathrm{A}_{\nu,1}$) [boundedness of $f_\delta^{(\nu)}$] $\sup_{x,y\in\mathbb{R}} |h(x,y)| < \infty$;

($\mathrm{A}_{\nu,2}$) [smoothness of $f_\delta^{(\nu)}$] $h(x,w)$ is an integrable function which is uniformly Lipschitz continuous in x, i.e., $\sup_w |h(x_1,w) - h(x_2,w)| \le L|x_1 - x_2|$, for a constant $L > 0$;

($\mathrm{A}_{\nu,3}$) [boundedness of β^{-1}] $\inf_{x\in D} |\beta(x)| = c_\beta > 0$;

(A_4) [finiteness of moments] $\int |y| f_Y(y)\, dy < \infty$ and $\int y^2 f_Y(y)\, dy < \infty$.

Note in particular that Conditions ($\mathrm{A}_{\nu,1}$) and (A_4) guarantee that all the quantities defined above exist, and α and β satisfy $\sup_{x\in D} |\alpha(x)| < \infty$ and $\sup_{x\in D} |\beta(x)| < \infty$.

Let $\Rightarrow$ denote weak convergence in $\mathcal{C}(D)$ and define

$$\begin{aligned}
\mu(x_1,x_2) &= \int y f_{W,Y}(w,y) f_\delta(x_1 - w) f_\delta(x_2 - w)\, dw, \\
\varphi_1(x_1,x_2) &= \int y^2 f_{W,Y}(w,y) f_\delta(x_1 - w) f_\delta(x_2 - w)\, dw, \\
\psi_1(x_1,x_2) &= \int f_W(w) f_\delta(x_1 - w) f_\delta(x_2 - w)\, dw.
\end{aligned}$$

Our main result is a functional limit theorem for the proposed estimator. (All proofs are deferred to Section 5.)

Theorem 1. *Under Conditions ($A_{\nu,1}$), ($A_{\nu,2}$) for $\nu = 0, 1$, ($A_{0,3}$) and (A_4), we have that, for the process $Z_n(x) = \sqrt{n}\,(\widehat{m}(x) - m(x))$, $Z_n \Rightarrow Z$ on $\mathcal{C}(D)$, where Z is a Gaussian process with zero mean and covariance*

$$\begin{aligned}
\mathrm{cov}\{Z(x_1), Z(x_2)\} =& \varphi_1(x_1,x_2)/\{\psi(x_1)\psi(x_2)\} + \psi_1(x_1,x_2)\varphi(x_1)\varphi(x_2)/\{\psi^2(x_1)\psi^2(x_2)\} \\
&- \mu(x_1,x_2)\{\varphi(x_1)\psi(x_2) + \varphi(x_2)\psi(x_1)\}/\{\psi^2(x_1)\psi^2(x_2)\},
\end{aligned}$$

for $x_1, x_2 \in D$.

The correlation structure for estimates at points $x_1 \neq x_2$ is seen not to vanish asymptotically, in contrast to the well-known behavior of local smoothing estimators where estimates at different points become asymptotically uncorrelated as

bandwidths and windows converge to zero. Define $\widehat{\mu}(x_1, x_2) = n^{-1} \sum_{i=1}^n Y_i f_\delta(x_1 - W_i) f_\delta(x_2 - W_i)$, $\hat{\psi}_1(x_1, x_2) = n^{-1} \sum_{i=1}^n f_\delta(x_1 - W_i) f_\delta(x_2 - W_i)$ and $\hat{\varphi}_1(x_1, x_2) = n^{-1} \sum_{i=1}^n Y_i^2 f_\delta(x_1 - W_i) f_\delta(x_2 - W_i)$. Particular consequences of Theorem 1 include the properties $\sup_{x \in D} \sqrt{n}\,(\widehat{m}(x) - m(x)) \xrightarrow{D} \sup_{x \in D} Z(x)$ and $\sqrt{n}\,(\widehat{m}(x) - m(x)) \xrightarrow{D} N(0, V(x))$ as $n \to \infty$, where $V(x) = \operatorname{cov}\big(Z(x), Z(x)\big)$ is estimated uniformly and $\sqrt{n}$-consistently by $\hat{V}(x) = \hat{\varphi}_1(x,x)\hat{\psi}^{-2}(x) + \hat{\varphi}^2(x)\hat{\psi}_1(x,x)\hat{\psi}^{-4}(x) - 2\hat{\varphi}(x)\widehat{\mu}(x,x)\hat{\psi}^{-3}(x)$, in the sense that $\sup_{x \in D} |\hat{V}(x) - V(x)| = O_P(n^{-1/2})$. It follows that an asymptotic $(1-\alpha)$-level confidence interval for $m(x)$ has endpoints $\hat{m}(x) \pm \hat{V}(x)^{1/2} \Phi^{-1}(1 - \alpha/2), \hat{m}(x)$, where Φ denotes the standard normal distribution function.

Semiparametric efficiency of $\hat{m}$ can be established, in regular cases where $f_\delta(x - w)$ is monotone in x for w in the support of W, by considering the following simpler problem. Suppose we observe independent and identically distributed pairs $(R_1, S_1), \ldots, (R_n, S_n)$, where $R_i \geq 0$ and $S_i = \rho(R_i) + \epsilon_i$, with ρ a smooth function and ϵ_i independent of R_i and distributed as $\mathrm{N}(0, \sigma^2)$. Consider the problem of estimating $(\theta_1, \theta_2) = (E(R), E\{R\rho(R)\})$ from these data. The estimator $(\hat{\theta}_1, \hat{\theta}_2) = (n^{-1} \sum_{i=1}^n R_i, n^{-1} \sum_{i=1}^n R_i S_i)$ is asymptotically normally distributed and semiparametric efficient in this problem, and thus $\hat{\theta}_2/\hat{\theta}_1$ is semiparametric efficient for θ_2/θ_1. (The proof follows via Examples 3.2.1 and 3.3.4, and Propositions 3.3.1 and A.5.2, of Bickel et al. (1993).) We may identify $m(x)$ and $\hat{m}(x)$ with θ_2/θ_1 and $\hat{\theta}_2/\hat{\theta}_1$, respectively, by taking $R_i = f_\delta(x - W_i)$ and $\rho(r) = g\{f_{\delta x}^{-1}(r)\}$, where $f_{\delta x}(w) = f_\delta(x - w)$.

Under additional regularity conditions, Theorem 1 continues to hold, although with an altered covariance structure for the limiting process Z, in the more general setting described in section 1.3. There we observe T_i, in the setting of an unknown parameter θ, rather than $W_i = p(T_i \mid \theta)$, and W_i is replaced by $\widehat{W}_i = p(T_i \mid \hat{\theta})$ in the definition of $\hat{m}$. If the model $p(\cdot \mid \theta)$ is linear then the only additional assumptions needed are two bounded derivatives of f_δ, and $E(T^2) < \infty$, where T denotes a generic T_i. See section 5.3 for an outline proof.

2.2. Case where f_δ is estimated from replicated data. The conditions imposed below (see particularly (2.2)) imply that the distributions of W and δ are absolutely continuous, and in particular that the respective densities f_W and f_δ are square-integrable. We shall assume that,

$$\max_{i \geq 1} n_i < \infty\,, \quad n = o(N)\,, \quad n^{1/(2(\lambda + \lambda_\delta - 1))} \ll \tau_n \ll N^{1/(2(1+\lambda_\delta))}\,, \tag{2.1}$$

where $a_n \ll b_n$, for positive sequences a_n and b_n, means that $a_n/b_n \to 0$ as $n \to \infty$; and that, for constants $\lambda, \lambda_\delta > 0$ satisfying

$$\lambda > \lambda_\delta + 1 \quad \text{and} \quad \lambda_\delta > 1, \tag{2.2}$$

we have

$$\begin{aligned} \bigl|(f_W g)^{\rm ft}(t)\bigr| + \bigl|f_W^{\rm ft}(t)\bigr| \le \text{const.}\, |t|^{-\lambda} \quad \text{for all } t, \\ f_\delta^{\rm ft}(t) > 0 \quad \text{for all } t\,, \quad \bigl|f_\delta^{\rm ft}(t)\bigr| \asymp |t|^{-\lambda_\delta} \quad \text{as } t \to \infty. \end{aligned} \tag{2.3}$$

The second part of (2.1) asks that there be an order of magnitude more values of U_{ij}, at (1.2), than there are pairs (W_i, Y_i), at (1.1). Conditions (2.2) and (2.3) ask that f_δ be sufficiently smooth, with its Fourier transform decaying in the standard polynomial way, and that f_W and $f_W g$ be sufficiently smooth relative to f_δ. The second part of (2.1), and (2.2), imply that it is always possible to choose the smoothing parameter τ_n such as to satisfy the third part of (2.1).

Theorem 2. *If the function g is uniformly bounded, if the errors ϵ_i at (1.1) have zero mean and finite variance, and if (2.1)–(2.3) hold, then, uniformly in x,*

$$\tilde\psi(x) = \hat\psi(x) + o_p\bigl(n^{-1/2}\bigr), \quad \tilde\varphi(x) = \hat\varphi(x) + o_p\bigl(n^{-1/2}\bigr). \tag{2.4}$$

Let $\mathcal{I}$ denote an interval for which $\inf_{x\in\mathcal{I}}\, \psi(x) > 0$. Result (2.4) implies that, under the additional conditions imposed for Theorem 1, the estimator $\widetilde m = \tilde\varphi/\tilde\psi$, which is an alternative to $\hat m = \hat\varphi/\hat\psi$ discussed in Section 2.1, satisfies $\widetilde m(x) = \hat m(x) + o_p(n^{-1/2})$ uniformly in $x \in \mathcal{I}$. Therefore $\widetilde m$ inherits the weak convergence and semiparametric-efficiency properties of $\widetilde m$ on $\mathcal{I}$. Theorem 2 holds, under more restrictive assumptions, in the more general setting of section 1.3; see section 5.3.

3 Simulations

We implemented our estimator $\hat m(x)$ of $m(x)$ on samples of (W, Y) generated from models of two types:

(1) $g(w) = [3w + 20\,(2\pi)^{-1/2} \exp(-200(w - 1/2)^2)]1_{[0,1]}(w)$, $W \sim U[0, 1]$, $\epsilon \sim N(0, \sigma_\epsilon^2)$ and $\delta \sim N(0, \sigma_\delta^2)$ or $\delta \sim U[-1/2, 1/2]$;

(2) $Y|W = w \sim$ Bernoulli$(g(w))$, with $g(w) = \exp(6w)/[1 + \exp(6w)]$, $W \sim U[-0.5, 0.5]$, $\delta \sim N(0, \sigma_\delta^2)$ or with $g(w) = 0.45 \sin(a\pi w) + 0.5$, $a = 2$ or 4, $W \sim U[0, 1]$, $\delta \sim N(0, \sigma_\delta^2)$ or $\delta \sim U[-1/2, 1/2]$.

The last example was used by Hobert and Wand (2000). In each case, we considered several sample sizes ($n = 50, 100$ and 250) and the parameters $\text{var}(\delta)$ and $\text{var}(\epsilon)$ were chosen such that the noise-to-signal ratios $NS_\delta = \text{var}(\delta)/\,\text{var}(W)$ and $NS_\epsilon = \text{var}(\epsilon)/\|g\|_\infty$ equal 10%, 25% or 50%. We considered the situation where the values of X are available as well, which allowed us to compare our estimators with the $n^{-2/5}$-consistent Nadaraya-Watson estimator $\hat m_N$ of $m(x)$, based on observations of (X, Y). In all cases, our estimators based on (W, Y) performed much better than $\hat m_N$, which was biased and much more variable. These findings continued to hold

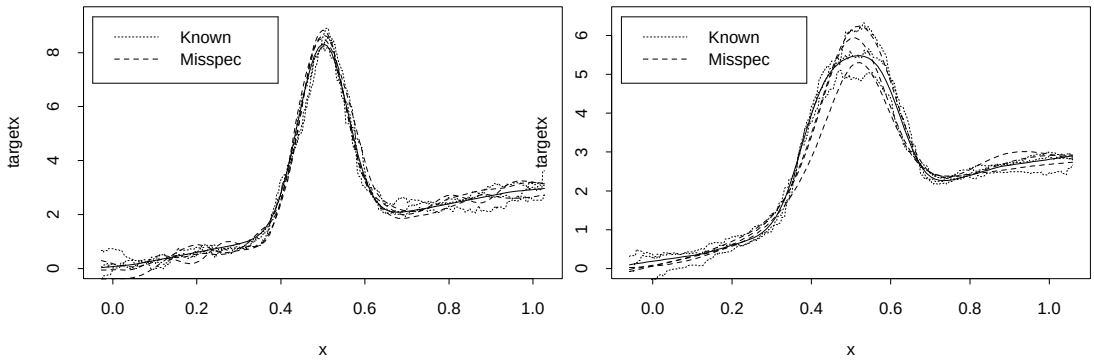

Figure 3.1: The estimator $\hat{m}$ with the error f_δ known (uniform) or misspecified (Gaussian) for case (2), with $NS_\delta = 0.1$ (left panel) or $NS_\delta = 0.25$ (right panel), with $NS_\epsilon = 0.1$ and $n = 250$. The solid curve is the target curve m.

in the setting of section 1.3, where the sample available was (T_i, X_i, Y_i), $i = 1, \ldots, n$ and the error variance was unknown and estimated by the empirical variance of the sample $X_i - \hat{W}_i$, $i = 1, \ldots, n$. More details are available from the first author's website.

The typical behavior of our estimator is illustrated in Figure 3.1, where we compare, for case (1) with uniform δ, $NS_\delta = 0.1$, $NS_\delta = 0.25$ and $n = 250$, the results of 1000 replications of the estimators $\hat{m}$ with the correct error density f_δ and $\hat{m}$ with f_δ misspecified (here we used Gaussian error instead of the uniform error). In both cases, the estimates shown correspond to the first, fifth and ninth deciles of the ordered 1000 values of the Integrated Squared Error $\int (\hat{m}(x) - m(x))^2 \, dx$. We see that for small NS_δ, the estimator is quite robust to error misspecification, but, without any surprise, the quality deteriorates as the ratio increases. Note however that the results remain quite good for $NS_\delta = 0.25$.

4 Real data illustration

We illustrate the proposed estimator in the setting of Section 1.3 on a real data example. The data set was collected during a South African study on heart disease and was used by Hastie et al. (2001). The data are available at http://www-stat.stanford.edu/ElemStatLearn. During the study, several variables were measured on males in a heart-disease high-risk region of the Western Cape, including low density lipoprotein cholesterol (LDL) and total cholesterol (CHOL) as predictors, and coronary heart disease history (CHD) as response, coded as 0 = non-incidence of CHD, 1 = incidence of CHD. LDL is much more difficult to measure than CHOL, which motivates the use of CHOL as a proxy for LDL (Carroll et al., 1995). After

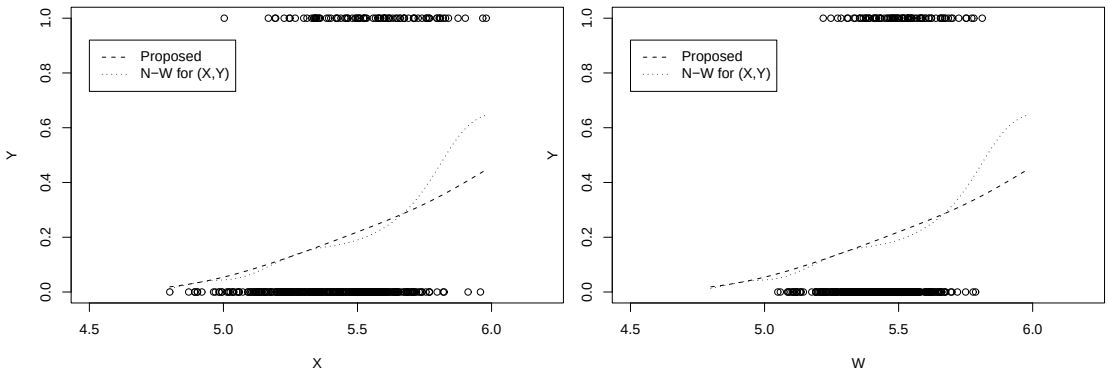

Figure 4.1: The proposed estimator $\hat{m}$ and the Nadaraya-Watson estimator $\hat{m}_N$ based on the observations of (X, Y), and a scatter plot of the 446 observed values of (X, Y) (left panel) or the 446 observed values of $(\hat{W}, Y)$ (right panel), for the coronary heart disease data.

deleting several outliers, the relationship between LDL and CHOL can be reasonably well modelled as $\log(\text{CHOL}) = \theta^{(1)} + \theta^{(2)} \log(\text{LDL}) + \delta$ with δ a random variable of zero mean, see Carroll et al. (1995) who use the same model for a similar data set. Checking for outliers, we deleted the observations corresponding to the smallest (respectively two largest) value(s) of CHOL, the smallest three (resp. largest two) values of LDL, and the eight points of $(\log(\text{CHOL}), \log(\text{LDL}))$ the furthest away from the least squares line.

We set $Y = \text{CHD}$, $X = \log(\text{CHOL})$ and $\hat{W} = \hat{\theta}^{(1)} + \hat{\theta}^{(2)} \log(\text{LDL})$, where $\hat{\theta}^{(1)} = 4.8890$ and $\hat{\theta}^{(2)} = 0.3663$ are the least squares estimators of $\theta^{(1)}$ and $\theta^{(2)}$. Our goal is to estimate $m(x) = \text{E}(Y|X = x)$, the conditional expectation of incidence of coronary heart disease given the (transformed) total cholesterol level, using the sample of $n = 446$ observations.

We compare the proposed estimator $\hat{m}(x)$ with the Nadaraya-Watson estimator $\hat{m}_N$. The data suggest that it is reasonable to assume that the errors $\delta_i = X_i - W_i$ are normal, where the variance can be estimated from the differences $X_i - \hat{W}_i$. In Figure 4.1, we overlay the proposed estimator $\hat{m}$ and the Nadaraya-Watson estimator $\hat{m}_N$ calculated with an appropriate data-driven cross-validation bandwidth. The graphs suggest that the probability of coronary heart disease increases with the cholesterol level. The increase is highly nonlinear, and there are clear differences between the classical Nadaraya-Watson estimator and the proposed estimator. The Nadaraya-Watson estimator exhibits additional fluctuations, especially in the right tail, thus giving a less stable appearance.

5 Proofs

5.1. Outline proof of Theorem 1. Define the auxiliary quantities

$$\begin{aligned}
\tilde{Z}_n(x) &= \sqrt{n}\left[\frac{\hat{\varphi}(x)-\varphi(x)}{\psi(x)} - \frac{(\hat{\psi}(x)-\psi(x))\varphi(x)}{\psi^2(x)}\right],\\
\hat{\alpha}(x) &= n^{-1}\sum_{i=1}^{n} Y_i h(x,W_i),\\
\alpha_1(x_1,x_2) &= \iint y^2 h(x_1,w)h(x_2,w)f_{W,Y}(w,y)\,dw\,dy,\\
\hat{\beta}(x) &= n^{-1}\sum_{i=1}^{n} h(x,W_i),\\
\beta_1(x_1,x_2) &= \int h(x_1,w)h(x_2,w)f_W(w)\,dw.
\end{aligned}$$

The next two results will be useful to prove the theorem. Their proof is given at the end of this section.

Lemma 1. *Let ν be a positive integer and $x \in D$. Under Conditions $(A_{\nu,1})$, $(A_{\nu,2})$ and (A_4),*

$$\sqrt{n}\,(\hat{\alpha}(x)-\alpha(x)) \Rightarrow Z_\alpha(x), \quad \sqrt{n}\,(\hat{\beta}(x)-\beta(x)) \Rightarrow Z_\beta(x),$$

where Z_α, Z_β are Gaussian processes characterized by the moments $\mathrm{E}(Z_\alpha(x)) = \mathrm{E}(Z_\beta(x)) = 0$, and $\mathrm{cov}(Z_\alpha(x_1), Z_\alpha(x_2)) = \alpha_1(x_1,x_2) - \alpha(x_1)\alpha(x_2)$, $\mathrm{cov}(Z_\beta(x_1), Z_\beta(x_2)) = \beta_1(x_1,x_2) - \beta(x_1)\beta(x_2)$, for all $x_1, x_2 \in D$.

Lemma 2. *Let $x_1,\ldots,x_k \in D$. Under Conditions $(A_{0,1})$ and (A_4), for all $\boldsymbol{t} = (t_1,\ldots,t_k)' \in I\!R^k$, we have $\sum_{j=1}^k t_j \tilde{Z}_n(x_j) \xrightarrow{D} N(0, \boldsymbol{t}'\Sigma\boldsymbol{t})$, where*

$$(\Sigma)_{jl} = \frac{\varphi_1(x_j,x_l)}{\psi(x_j)\psi(x_l)} + \frac{\varphi(x_j)\varphi(x_l)\psi_1(x_j,x_l)}{\psi^2(x_j)\psi^2(x_l)} - \frac{\varphi(x_l)\mu(x_j,x_l)}{\psi(x_j)\psi^2(x_l)} - \frac{\varphi(x_j)\mu(x_j,x_l)}{\psi(x_l)\psi^2(x_j)}.$$

Put $Z_n(x) = \sqrt{n}\,(\widehat{m}(x) - m(x)) = X_n(x) + Y_n(x)$, where $\psi(x)X_n = \sqrt{n}\,(\hat{\varphi}(x) - \varphi(x))$ and $\hat{\psi}(x)\psi(x)Y_n(x) = -\sqrt{n}\,(\hat{\psi}(x)-\psi(x))\hat{\varphi}(x)$. It suffices to prove (a) convergence of the finite dimensional limit distribution of Z_n, and (b) tightness of Z_n. To establish (a), note that

$$Z_n(x) = \tilde{Z}_n(x) - \sqrt{n}\,\frac{\hat{\psi}(x)-\psi(x)}{\psi(x)}\left[\frac{\hat{\varphi}(x)}{\hat{\psi}(x)} - \frac{\varphi(x)}{\psi(x)}\right]. \tag{5.1}$$

Now

$$\sup_{x\in D}\left|\frac{\hat{\varphi}(x)}{\hat{\psi}(x)} - \frac{\varphi(x)}{\psi(x)}\right| \le \frac{\sup_{x\in D}|\hat{\varphi}(x)-\varphi(x)|}{\inf_{x\in D}|\hat{\psi}(x)|} + \frac{\sup_{x\in D}|\varphi(x)|\cdot\sup_{x\in D}|\psi(x)-\hat{\psi}(x)|}{\inf_{x\in D}|\psi(x)\hat{\psi}(x)|},$$

where $\inf_{x\in D}|\psi_n(x)| \xrightarrow{P} \inf_{x\in D}|\psi(x)| > 0$, which, combined with Lemma 1, proves that the last term of (5.1) tends to zero as n tends to infinity, and thus $Z_n(x)$

has the same finite-dimensional limit distribution as $\tilde{Z}_n(x)$. From Lemma 2 and the Cramér-Wold device, this limit distribution is the same as that claimed for Z in Theorem 1. To prove (b), note that, by the proof of Lemma 1, the sequences $\sqrt{n}\,(\hat{\varphi}(x)-\varphi(x))$ and $\sqrt{n}\,(\hat{\psi}(x)-\psi(x))$ are tight. The sequence $\hat{\varphi}(x)/\hat{\psi}(x)$ is tight if we show that for given $\epsilon,\eta\geq 0$ and sufficiently small δ and large n,

$$P\Big(\sup_{|x-y|\leq\delta}|\hat{\varphi}(x)/\hat{\psi}(x)-\hat{\varphi}(y)/\hat{\psi}(y)|\geq\epsilon\Big)\leq\eta. \tag{5.2}$$

Now, defining $\xi(x)=\iint y f_{W,Y}(w,y)f'_\delta(x-w)\,dw\,dy$, $\zeta(x)=\int f_W(w)f'_\delta(x-w)\,dw$, $\hat{\xi}(x)=n^{-1}\sum_{i=1}^n Y_i f'_\delta(x-W_i)$ and $\hat{\zeta}(x)=n^{-1}\sum_{i=1}^n f'_\delta(x-W_i)$, let $\hat{T}(x)=[\hat{\xi}(x)\hat{\psi}(x)-\hat{\varphi}(x)\hat{\zeta}(x)]/\hat{\psi}^2(x)$. By the mean value theorem, the left-hand side of (5.2) is bounded by $P\big(\sup_{x\in D}|\hat{T}(x)|\geq\epsilon/\delta\big)$ and (5.2) follows if we note that

$$\begin{aligned}\sup_{x\in D}|\hat{T}(x)|\leq&\sup_{x\in D}|\hat{\xi}(x)-\xi(x)|/|\hat{\psi}(x)|+\sup_{x\in D}|\xi(x)|/|\hat{\psi}(x)|+\big[\sup_{x\in D}|\hat{\varphi}(x)-\varphi(x)|/|\hat{\psi}(x)|^2\\&+\sup_{x\in D}|\varphi(x)|/|\hat{\psi}(x)|^2\big]\cdot\big[\sup_{x\in D}|\hat{\zeta}(x)-\zeta(x)|/|\hat{\psi}(x)|^2+\sup_{x\in D}|\zeta(x)|/|\hat{\psi}(x)|^2\big]\end{aligned}$$

which tends to zero as n tends to infinity. Property (b) follows.

Proof of Lemma 1. We prove the result for α; the proof for β is analogous. Let $x_1,\ldots,x_k\in D$, $\hat{\boldsymbol{\alpha}}=(\hat{\alpha}(x_1),\ldots,\hat{\alpha}(x_k))'$, $\boldsymbol{\alpha}=(\alpha(x_1),\ldots,\alpha(x_k))'$ and $\boldsymbol{Z}_\alpha\sim N_k(0,\Sigma_\alpha)$, where $(\Sigma_\alpha)_{ij}=\alpha_1(x_i,x_j)-\alpha(x_i)\alpha(x_j)$. Applying the central limit theorem to the i.i.d. sequence $T_1,\ldots,T_n$, with $T_i=\sum_{j=1}^k t_j Y_i h(x_j,W_i)$, it is not hard to prove that, for all $\boldsymbol{t}=(t_1,\ldots,t_k)'\in\mathbb{R}^k$, $\sqrt{n}\,\boldsymbol{t}'(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha})\xrightarrow{D}\boldsymbol{t}'\boldsymbol{Z}_\alpha$. From the Cramér-Wold device, we deduce that $\sqrt{n}(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha})\xrightarrow{D}\boldsymbol{Z}_\alpha$, which implies weak convergence of the finite-dimensional distributions. Using uniform Lipschitz continuity of h in the first coordinate, one can show that $\mathrm{E}(\sqrt{n}[\hat{\alpha}(x_1)-\alpha(x_1)-\hat{\alpha}(x_2)+\alpha(x_2)])^2\leq c|x_1-x_2|^2$, which implies tightness of $\sqrt{n}(\hat{\alpha}-\alpha)$.

Proof of Lemma 2. Since $\mathrm{E}\,\hat{\varphi}(x)=\varphi(x)$ and $\mathrm{E}\,\hat{\psi}(x)=\psi(x)$, we have that $\mathrm{E}(\tilde{Z}_n(x))=0$. The result follows from the central limit theorem if we note that $\sqrt{n}\sum_{j=1}^k t_j\tilde{Z}_n(x_j)$ may be written as $\sum_{i=1}^n T_i$, where $T_i=\sum_{j=1}^k t_j f_\delta(x_j-W_i)[Y_i/\psi(x_j)-\varphi(x_j)/\psi^2(x_j)]$.

5.2. Outline proof of Theorem 2. We shall derive the second result at (2.4); a proof of the first result there is similar. Define the functions $a=(f_W g)^{\mathrm{ft}}$, $\hat{a}=\widehat{(f_W g)}^{\mathrm{ft}}$, $b=(f_\delta^{\mathrm{ft}})^2$, $\hat{b}=(\hat{f}_\delta^{\mathrm{ft}})^2$, $c=f_\delta^{\mathrm{ft}}$, $\Delta_a=\hat{a}-a$ and $\Delta_b=\hat{b}-b$. Let $\mathcal{T}$ denote the interval $[-\tau_n,\tau_n]$, write $\tilde{\mathcal{T}}$ for the complement in $\mathbb{R}$ of $\mathcal{T}$, and put $u_x(t)=e^{-itx}$. Then, uniformly in x,

$$\begin{aligned}2\pi\,\tilde{\varphi}(x)&=\int_{\mathcal{T}}\hat{a}\,|\hat{b}|^{1/2}\,u_x=\int_{\mathcal{T}}(a+\Delta_a)\,|b|^{1/2}\left(1+b^{-1}\Delta_b\right)^{1/2}u_x\\&=\int_{\mathcal{T}}(a+\Delta_a)\,c\,u_x+O_p\Big[\int_{\mathcal{T}}|a/c|\left(E\Delta_b^2\right)^{1/2}+\int_{\mathcal{T}}c^{-1}\left(E\Delta_a^2\,E\Delta_b^2\right)^{1/2}\Big].\end{aligned}$$

Using the fact that $\hat{a}$ equals a sum of n independent and identically distributed random variables, and $\hat{b}$ is expressed in a form similar to a U-statistic, it can be shown that $E[\Delta_a(t)^2] = O(n^{-1})$ and $E[\Delta_b(t)^2] = O(N^{-1})$, uniformly in t. Moreover, (2.2) and (2.3) imply that $\int_{\mathcal{T}} |a/c| = O(1)$, $\int_{\mathcal{T}} c^{-1} = O(\tau_n^{\lambda_\delta+1})$, $\int_{\tilde{\mathcal{T}}} acu_x = O(\tau_n^{1-\lambda-\lambda_\delta})$ and $\int_{\tilde{\mathcal{T}}} \Delta_a c u_x = O_p(n^{-1/2}\,\tau_n^{1-\lambda_\delta})$, the latter two results holding uniformly in x. Therefore, uniformly in x,

$$\begin{aligned} 2\pi\,\tilde{\varphi}(x) &= \int (a+\Delta_a)\,c\,u_x + O_p\Big(N^{-1/2} + n^{-1/2}N^{-1/2}\,\tau_n^{\lambda_\delta+1} + \tau_n^{1-\lambda-\lambda_\delta} + n^{-1/2}\,\tau_n^{1-\lambda_\delta}\Big) \\ &= \int \hat{a}\,c\,u_x + o_p\big(n^{-1/2}\big). \end{aligned} \tag{5.3}$$

Since $\hat{\varphi}(x) = (2\pi)^{-1}\int \hat{a}cu_x$ then the second part of (2.4) follows from (5.3).

5.3. Case where $X = p(T \mid \theta) + \delta$. This generalisation, in which (T, Y) rather than (W, Y) is observed, was introduced in section 1.3. There we noted that the unknown parameter θ could be estimated by least-squares from data (T_i', X_i'), for $1 \le i \le r$, on (T, X). In the case of a linear model, $p(t \mid \theta) = \theta^{(1)} + \theta^{(2)}t$, and our estimator of $W_i = \theta^{(1)} + \theta^{(2)}T_i$ is $\widehat{W}_i = \hat{\theta}^{(1)} + \hat{\theta}^{(2)}T_i$. We shall treat this particular case below; other models for p can be addressed similarly.

Let $\hat{m}^*$, $\hat{\varphi}^*$, $\hat{\psi}^*$ and $\tilde{\psi}^*$ denote the versions of $\hat{m}$, $\hat{\varphi}$, $\hat{\psi}$ and $\tilde{\psi}$, respectively, obtained on replacing W_i by $\widehat{W}_i$ throughout. It will be assumed that $n = O(r)$. In this case the least-squares estimators $\hat{\theta}^{(1)}$ and $\hat{\theta}^{(2)}$ are $\sqrt{n}$-consistent.

First we consider the setting where f_δ is known. Provided f_δ has two bounded derivatives, we may write:

$$\begin{aligned} \hat{\varphi}^*(x) = n^{-1}\sum_{i=1}^n Y_i\,f_\delta(x - \widehat{W}_i) = \hat{\varphi}(x) &- \big(\hat{\theta}^{(1)} - \theta^{(1)}\big)\,E\{Y\,f_\delta'(x-W)\} \\ &- \big(\hat{\theta}^{(2)} - \theta^{(2)}\big)\,E\{TY\,f_\delta'(x-W)\} + o_p\big(n^{-1/2}\big), \end{aligned} \tag{5.4}$$

$$\begin{aligned} \hat{\psi}^*(x) = n^{-1}\sum_{i=1}^n f_\delta(x - \widehat{W}_i) = \hat{\psi}(x) &- \big(\hat{\theta}^{(1)} - \theta^{(1)}\big)\,E\{f_\delta'(x-W)\} \\ &- \big(\hat{\theta}^{(2)} - \theta^{(2)}\big)\,E\{T\,f_\delta'(x-W)\} + o_p\big(n^{-1/2}\big). \end{aligned} \tag{5.5}$$

Here, $\hat{\varphi}$ and $\hat{\psi}$ are the original estimators of φ and ψ, given in section 1.2 for the case where W_i is directly observed; $W = \theta^{(1)} + \theta^{(2)}T$; and the remainder terms $o_p(n^{-1/2})$ are uniform in x, provided the conditions of Theorem 2 hold and, in addition, $E(T^2) < \infty$.

It follows from (5.4) and (5.5) that $\hat{\varphi}$ and $\hat{\psi}$ are $\sqrt{n}$-consistent for ϕ and ψ, respectively, and $\hat{m}^* = \hat{\varphi}^*/\hat{\psi}^*$ is $\sqrt{n}$-consistent for m. A version of Theorem 1 is readily obtained in this setting, using (5.4) and (5.5). Unless $r/n \to \infty$, the

covariance structure of the limiting Gaussian process depends on whether the data (T_i', X_i'), from which $\hat{\theta}^{(1)}$ and $\theta^{(2)}$ are computed, are independent of the data (W_i, Y_i) used to calculate $\hat{\varphi}$ and $\hat{\psi}$, or whether $(T_i', X_i') = (T_i, X_i)$ and the triples (T_i, X_i, Y_i) are observed.

The case where f_δ is not known, and is consistently estimated from replicated data as discussed in section 1.2, is similar although more complex. Our estimator $\hat{f}_\delta^{\rm ft}$, given at (1.6), does not alter since it does not use the data W_i. On the other hand, the estimators $\hat{f}_W^{\rm ft}$ and $\widehat{(f_W g)}^{\rm ft}$, given at (1.6) and (1.7), are replaced by

$$\hat{f}_W^{\rm ft\,*}(t) = n^{-1} \sum_{j=1}^n \exp(it\widehat{W}_j)\,, \quad \widehat{(f_W g)}^{\rm ft\,*} = n^{-1} \sum_{j=1}^n Y_j \exp(it\widehat{W}_j)\,.$$

Substituting the latter for $\hat{f}_W^{\rm ft}$ and $\widehat{(f_W g)}^{\rm ft}$, respectively, in (1.8); Taylor-expanding $\exp(-it\widehat{W}_j)$ as $\exp(itW_j)\,\{1 + it\,(\widehat{W}_j - W_j) + \ldots\}$; and taking the smoothing parameter τ_n in (1.8) to be of order $n^{(1/2)-2\eta}$, for some $\eta > 0$ (so that, under moment conditions on W_j, $\tau_n \sup_{j \le n} |\widehat{W}_j - W_j| = O_p(n^{-\eta})$); we may deduce that (2.4) continues to hold if $\hat{\varphi}$ and $\hat{\psi}$ there are replaced by $\hat{\varphi}^*$ and $\hat{\psi}^*$, provided more restrictive assumptions than those given in Theorem 2 are imposed.

Acknowledgements

We are grateful to Chris Klaassen for helpful discussion, and to the Editor, the Associate Editor and two reviewers for valuable comments and suggestions which led to considerable improvement of the paper.

References

Bickel, Peter J., Klaassen, C.A.J., Ritov, Y. and Wellner, J.A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* Johns Hopkins University Press, Baltimore, MD.

Carroll, R.J. and Hall, P. (2004). Low-order approximations in deconvolution and regression with errors in variables. *J. Roy. Statist. Soc.* Ser. B, **66**, 31–46.

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. and Stefanski, L. (1995). *Measurement error in nonlinear models.* Chapman and Hall, London.

Devanarayan, V. and Stefanski, L.A. (2002). Empirical simulation extrapolation for measurement error models with replicate measurements. *Statist. Probab. Lett.*, **59**, 219-225.

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

Fan, J. and Masry, E. (1992). Multivariate regression estimation with errors-in-variables: asymptotic normality for mixing processes. *J. Multivariate Anal.*, **43**, 237–271.

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

Fan, J., Truong, Y.K. and Wang, Y. (1991). Nonparametric function estimation involving errors-in-variables. In: *Nonparametric Functional Estimation and Related Topics*, Ed. G. Roussas, pp. 613–627. Kluwer, Dordrecht.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning.* Springer-Verlag, New-York.

Hobert, J. P. and Wand, M. P. (2000). Automatic generalized nonparametric regression via maximum likelihood. *Technical Report*, Department of Biostatistics, Harvard School of Public Health.

Ioannides, D.A. and Matzner-Lober, E. (2002). Nonparametric estimation of the conditional mode with errors-in-variables: Strong consistency for mixing processes. *J. Nonparam. Statist.*, **14**, 341–352.

Li, T. and Vuong, Q. (1998). Nonparametric estimation of the measurement error model using multiple indicators. *J. Multivar. Anal.*, **65**, 139–165.

Linton, O. and Whang, Y.J. (2002). Nonparametric estimation with aggregated data. *Econometric Theory*, **18**, 420–468.

Stefanski, L. and Carroll, R.J. (1990). Deconvoluting kernel density estimators. *Statistics*, **2**, 169–184.

Stefanski, L.A. and Cook, J.R. (1995). Simulation-extrapolation: The measurement error jackknife *J. Amer. Statist. Assoc.*, **90**, 1247–1256.

Taupin, M.L. (2001). Semi-parametric estimation in the nonlinear structural errors-in-variables model. *Ann. Statist.*, **29**, 66–93.