Electronic Journal of Statistics
Vol. 13 (2019) 3985–4014
ISSN: 1935-7524
https://doi.org/10.1214/19-EJS1612

Efficient estimators for expectations in nonlinear parametric regression models with responses missing at random

Guorong Dai

Department of Statistics, Texas A&M University, College Station, TX, USA
e-mail: rondai@stat.tamu.edu

and

Ursula U. Müller

Department of Statistics, Texas A&M University, College Station, TX, USA
Department of Mathematics, University of Hamburg, Hamburg, Germany
e-mail: uschi@stat.tamu.edu

Abstract: We consider nonlinear regression models that are solely defined by a parametric model for the regression function. The responses are assumed to be missing at random, with the missingness depending on multiple covariates. We propose estimators for expectations of a known function of response and covariates. Our estimator is a nonparametric estimator corrected for the regression function. We show that it is asymptotically efficient in the Hájek and Le Cam sense. Simulations and an example using real data confirm the optimality of our approach.

MSC 2010 subject classifications: Primary 62J02, 62F12; secondary 62G05.
Keywords and phrases: Nonlinear regression, conditional mean model, efficiency, imputation, multivariate covariates.

Received September 2018.

Contents

1. Introduction

In this article we study efficient estimation of expectations in a nonlinear regression model that is defined solely by the conditional constraint

$$E(Y|X) = r_\vartheta(X), \quad \vartheta \in \Theta \subset \mathbb{R}^p, \tag{1.1}$$

and therefore also known as the *conditional mean model*. Here the regression function r_ϑ is assumed to be known up to a parameter vector ϑ and X is a d-dimensional random vector. The nonlinear regression model is an important model for applications; see, for example, the books by Bates and Watts (1998 [1]) and Seber and Wild (1989 [18]).

In the literature it is quite common to introduce a third variable $\varepsilon = Y - r_\vartheta(X)$, especially if the covariates X and errors ε can be assumed to be independent; see, for example, Wang and Rao (2001 [22]), who study linear regression with missing responses. We do not make the independence assumption: in many situations, especially in applications in econometrics, model (1.1), which we consider here, is more suitable because of its flexibility.

We are interested in the scenario when responses Y are possibly missing and work with an indicator variable Z that is 1 if a response Y is observed and 0 if it is missing. Our sample consists of independent copies (X_i, Z_iY_i, Z_i), $i = 1, ..., n$, of a base observation (X, ZY, Z). The indicator Z tells us if a zero response is a numerical zero or a missing value. More specifically, we assume that the responses are *missing at random* (MAR), i.e. the probability that Y is missing depends only on the covariate vector X that is always observed,

$$P(Z = 1|X, Y) = P(Z = 1|X) = \pi(X).$$

The MAR assumption is common in applications; see, for example, the book by Little and Rubin [10]. It in particular implies that Z and Y are conditionally independent given X.

Our goal is to efficiently estimate expectations $E\{h(X, Y)\}$ of the joint distribution in model (1.1), where h is some known square-integrable function. This is a quite general problem: we basically estimate the entire joint distribution of the vector (X, Y). In the literature usually only estimation of the mean response $E(Y)$ is considered; see, for example, Matloff (1981 [11]), Cheng (1994 [3]), Wang and Rao (2001, 2002 [22, 23]) and, for further references, Müller (2009 [12]). Other examples of such expectations are moments of Y or X, mixed moments, and probabilities involving X and Y such as $P(X < Y)$. Estimation of $E\{h(X, Y)\}$ is also considered in Müller (2009 [12]) in the more restrictive nonlinear regression model with independent covariates and errors. That article

exploits the independence assumption by writing $E\{h(X,Y)\}$ as a convolution integral, which can be estimated in a relatively straightforward way. Since the distribution of the errors in our model depends on the covariates, our approach is quite different.

An obvious approach to estimate $E\{h(X,Y)\}$ in the missing data model with MAR responses is to use the Horvitz-Thompson family of estimators

$$\frac{1}{n}\sum_{i=1}^{n}\frac{Z_i}{\widehat{\pi}(X_i)}h(X_i,Y_i),$$

where $\widehat{\pi}(\cdot)$ is an estimator of the probability $\pi(\cdot)$ from above. Hirano, Imbens and Ridder (2003 [8]) prove its root-n asymptotic normality in a binary treatment model when $\pi(x)$ is estimated by the series logit method. We will consider this estimator in Section 4 and compare it with our method.

As in Müller, Schick and Wefelmeyer (2006 [15]), who discuss estimation of expectations $E\{h(X,Y)\}$ in a simple linear regression model, we use a nonparametric estimator $\widehat{H}_{np}$ and improve it by adding a correction term $\widehat{\Gamma}$ that takes the nonlinear structure into account,

$$\widehat{H} = \widehat{H}_{np} - \widehat{\Gamma}, \tag{1.2}$$

with $\widehat{\Gamma}$ defined in equation (2.1) in Section 2. Our nonparametric estimator $\widehat{H}_{np}$ for the first part of (1.2) is a partially imputed estimator,

$$\widehat{H}_{np} = \frac{1}{n}\sum_{i=1}^{n}\{Z_i h(X_i,Y_i) + (1-Z_i)\widehat{\chi}(X_i)\}, \tag{1.3}$$

where $\widehat{\chi}(x)$ is the Nadaraya-Watson estimator of $\chi(x) = E\{h(X,Y)|X=x\}$, similar to that used in Cheng (1994 [3]) (see Section 2.1 for details). Alternatively one could, as in Cheng and Wei (1986 [4]) and Cheng (1990 [2]), use a full imputation approach, which also replaces observed cases with estimators. In the nonparametric model full imputation and partial imputation are asymptotically equivalent (see Cheng, 1994 [3]), which is intuitively clear since the model contains no structural information. For this article we prefer partial imputation, for reasons of speed and simplicity.

We will show that the estimator proposed in this paper is efficient in the sense of Hájek and Le Cam. The efficiency results imply asymptotic normality, which is useful for constructing approximative confidence intervals for expectations $E\{h(X,Y)\}$ of known square-integrable functions $h(X,Y)$. To the best of our knowledge, our estimator is the first efficient estimator for $E\{h(X,Y)\}$ in the parametric MAR multiple regression model (1.1). Müller et al. (2006 [15]) propose an efficient estimator for univariate linear regression, but does not provide technical details. The results of this paper also apply to the usual model with no missing data, i.e. when all indicators equal one and $\pi(\cdot)\equiv 1$, so this is covered as a special case.

This paper is organized as follows. In the next section we provide a complete and detailed derivation of the stochastic expansion of the nonparametric estimator and of the correction term. Section 3 characterizes efficient estimators of functionals of the joint distribution and gives the efficient influence function for estimating $E\{h(X,Y)\}$ in our model. The efficiency of our estimator is established by showing that the expansion in Section 2 matches the efficient influence function in Section 3. In Section 4 we explain how our estimator can be implemented and compare it with other methods in various scenarios, using computer simulations. The results are positive throughout and confirm the theoretically proved optimality of our approach. In Section 5 we illustrate our approach by means of a real data set. Some technical details can be found in the Appendix.

2. Expansion of the estimator

Our estimator $\widehat{H} = \widehat{H}_{np} - \widehat{\Gamma}$ from (1.2) consists of the nonparametric estimator $\widehat{H}_{np}$ from equation (1.3) and a correction term $\widehat{\Gamma}$, which has the form

$$\widehat{\Gamma} = \frac{1}{n}\sum_{i=1}^{n} Z_i \widehat{g}(X_i)\widehat{\varepsilon}_i. \tag{2.1}$$

Here $\widehat{g}(x)$ is a consistent estimator of

$$g(x) = \frac{\rho_h(x)}{\pi(x)\sigma^2(x)},$$

uniformly in x on the support $\mathcal{I}$ of X, with $\rho_h(x) = E\{h(X,Y)\varepsilon | X = x\}$ and $\sigma^2(x) = E(\varepsilon^2|X=x)$. The term $\widehat{\Gamma}$ incorporates the parametric regression structure and is suggested by the canonical gradient, which characterizes the influence function of an efficient estimator (see Section 3).

To estimate $g(x)$ we can, for example, use a combination of Nadaraya-Watson estimators introduced by Nadaraya and Watson (1964 [16, 24]). The residuals $\widehat{\varepsilon}_i = Y_i - r_{\widehat{\vartheta}}(X_i)$ are based on an efficient estimator $\widehat{\vartheta}$ of ϑ; see Müller and Van Keilegom (2012 [13]) for an approach using estimating equations, and also for an overview of related efficient methods.

All estimators in $\widehat{\Gamma}$, including $\widehat{\vartheta}$, are complete case estimators since only observations with $Z = 1$ are used; see Müller and Schick (2017 [14]) who show that in the model with MAR responses complete case analysis is efficient for estimating characteristics of the *conditional* distribution of Y given X. The estimator $\widehat{H}_{np}$ from (1.3), on the contrary, is an imputation estimator. Hence our estimator (1.2) is a combination of imputation and complete case analysis.

In the usual model with no missing data, the partially imputed estimator $\widehat{H}_{np}$ for $E\{h(X,Y)\}$ reduces to the empirical estimator. However, it is not efficient unless we enhance it by correcting for the unknown parametric regression function using $\widehat{\Gamma}$ with all $Z_i = 1$ $(i = 1,\dots,n)$ and $\pi \equiv 1$, i.e. the efficient estimator becomes

$$\widehat{H} = \frac{1}{n}\sum_{i=1}^{n} h(X_i,Y_i) + \frac{1}{n}\sum_{i=1}^{n} \frac{\widehat{\rho}_h(x)}{\widehat{\sigma}^2(x)}\widehat{\varepsilon}_i.$$

In the following, for convenience of notation, we will always use the lower case letter c to represent a generic constant. The norm brackets $\|\cdot\|$ refer to the Euclidean norm of a vector.

We will assume throughout that $\pi(x) > 0$, for all x in the support $\mathcal{I}$ of X, to exclude the extreme case that no response is observed, that $h(X, Y)$ is square-integrable and that $E(\varepsilon^2)$ is positive and finite. The covariate vector X and the regression function need to satisfy the following conditions.

Assumption (X). *The d-dimensional random vector X has a compact support $\mathcal{I}$ and a density f that is bounded and bounded away from zero on $\mathcal{I}$.*

Assumption (R). *The regression function $\tau \mapsto r_\tau(x)$ is differentiable at $\tau = \vartheta$ with a p-dimensional square-integrable gradient $\dot{r}_\vartheta(x)$ that satisfies*

$$\sup_{x \in \mathcal{I}} \|\dot{r}_\tau(x) - \dot{r}_\vartheta(x)\| \le L \|\tau - \vartheta\| \quad \textit{for some constant } L \in \mathbb{R}.$$

To construct an efficient estimator of $E\{h(X, Y)\}$, an efficient estimator of ϑ, say $\widehat{\vartheta}$, is needed. Efficient estimation of ϑ in models defined by conditional constraints is discussed in Müller and Van Keilegom (2012 [13]). They show that an efficient estimator $\widehat{\vartheta}$ is characterized by the following expansion.

Assumption (T). *The estimator $\widehat{\vartheta}$ of ϑ satisfies*

$$\widehat{\vartheta} - \vartheta = \frac{1}{n} I^{-1} \sum_{i=1}^{n} Z_i \dot{r}_\vartheta(X_i) \sigma^{-2}(X_i) \varepsilon_i + o_p(n^{-1/2}),$$

with $I = E\{Z \dot{r}_\vartheta(X) \dot{r}_\vartheta(X)^\top \sigma^{-2}(X)\}$, which is assumed to be invertible.

An example of an efficient estimator is provided by Müller and Van Keilegom (2012 [13]), who propose using

$$\widehat{\vartheta} = \arg\min_\theta \Big\| \sum_{i=1}^{n} Z_i \dot{r}_\theta(X_i) \widehat{\sigma}^{-2}(X_i) \{Y_i - r_\theta(X_i)\} \Big\|,$$

where $\widehat{\sigma}^2(x)$ is a consistent estimator of $\sigma^2(x)$ uniformly in $x \in \mathcal{I}$, for example, the Nadaraya-Watson estimator. Under the conditions of their Theorem 2.1, the solution $\widehat{\vartheta}$ of the above estimating equation satisfies Assumption (T).

In the next two subsections we will expand the partially imputed estimator $\widehat{H}_{np}$ of equation (1.3) and derive the expansion of the correction term $\widehat{\Gamma}$ introduced in (2.1). Combining the two parts gives the expansion of $\widehat{H} = \widehat{H}_{np} - \widehat{\Gamma}$, which is stated in Corollary 2.1 at the end of this section.

2.1. Expansion of the nonparametric estimator

Consider the nonparametric partial imputation estimator introduced in (1.3), which imputes only the incomplete cases, as in Cheng (1994 [3]). We propose estimating the conditional expectation $\chi(x) = E\{h(X, Y)|X = x\}$ by

the Nadaraya-Watson estimator

$$\widehat{\chi}(x) = \frac{\sum_{j=1}^{n} K_b(X_j, x) Z_j h(X_j, Y_j)}{\sum_{j=1}^{n} K_b(X_j, x) Z_j}$$

with $K_b(u, x) = b^{-d} K(b^{-1}(u - x), x)$, where $K(\cdot, x)$ is a kernel function with integrated boundary correction, i.e. the form of the kernel is different for interior and boundary points x. The letter $b = b_n$ denotes a bandwidth sequence which tends to zero as n increases. By using boundary kernels we assume that we know the support of $\mathcal{I}$, which is a rather strong assumption. For practical applications we recommend estimating $\mathcal{I}$ using extreme values; see also Remark 2.1.

To derive the expansion of the partially imputed estimator (1.3) we stipulate the following assumption on the kernel K.

Assumption (K). *The kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a function that satisfies the following properties.*

(i) The kernel K is bounded and

$$\int_{\mathbb{R}^d} \left| s_1^{q_1} ... s_d^{q_d} K(s, x) \right| ds < \infty$$

for $x \in \mathcal{I}$ and any non-negative integers $q_1, q_2, ..., q_d$ satisfying $q_1 + ... + q_d = d + 1$, where $s_1, ..., s_d$ are the components of s.

(ii) Denote the region $\mathcal{S}_b(x) = \{b^{-1}(y - x) : y \in \mathcal{I}\}$. Then

$$\int_{\mathcal{S}_b(x)} K(s, x) ds = 1 \text{ and } \int_{\mathcal{S}_b(x)} s_1^{l_1} ... s_d^{l_d} K(s, x) ds = 0$$

for $x \in \mathcal{I}$ and non-negative integers $l_1, l_2, ..., l_d$ satisfying $0 < l_1 + ... + l_d < d + 1$.

(iii) $K(s, x)$ is differentiable with respect to s. For some constants η, ζ and $\nu > 1$, $\|\partial K(s, x)/\partial s\| \leq \eta$, and $\|\partial K(s, x)/\partial s\| \leq \eta \|s\|^{-\nu}$ for any s satisfying $\|s\| \geq \zeta$.

Note that assumption (K) is necessary because we consider a scenario with a covariate *vector* X. This is in contrast to Cheng (1994 [3]), who considers the nonparametric model with *univariate* covariates X. Cheng uses Theorem 1 of [5] to derive the expansion of his version of $\widehat{H}_{np}$. The theorem requires a non-negative kernel function, so it cannot be applied to our multivariate scenario, which requires using higher order kernels. For the construction of such kernels we can use results from Simonoff (1996 [19]); see Remark 2.1 at the end of this subsection for details.

For the ease of derivation we will further assume that $\pi(x)$ and $\sigma^2(x)$ are bounded away from zero on $\mathcal{I}$. In the second conclusion of Lemma 2.1 below we will show that $\sum_{j=1}^{n} K_b(X_j, x) Z_j / n$ converges to $\pi(x) f(x)$ in probability, uniformly in x. It follows that

$$\left\{ \inf_{x \in \mathcal{I}} \left| \frac{1}{n} \sum_{j=1}^{n} Z_j K_b(X_j, x) \right| \right\}^{-1} < \infty \tag{2.2}$$

with probability tending to one. Hence we can assume, without loss of generality, that the denominator in the Nadaraya-Watson estimator $\widehat{\chi}(\cdot)$ is bounded away from zero on $\mathcal{I}$. Finally we need the following two conditions.

Assumption (B)**.** *The bandwidth $b = b_n$ satisfies $nb^{2d}(\log n)^{-2} \to \infty$ and $nb^{2(d+1)} \to 0$ as $n \to \infty$.*

Assumption (D)**.** *The functions χ, π and f are $d+1$ times continuously differentiable on $\mathcal{I}$.*

Lemma 2.1 below facilitates the derivation of the asymptotic linearity of $\widehat{H}_{np}$ in Theorem 2.1.

Lemma 2.1. *Let $(X_1, V_1), \ldots, (X_n, V_n)$ be i.i.d. copies of a base observation (X, V), where X satisfies Assumption (X) and V is a q-dimensional random vector. For some function $g(x, v) : \mathbb{R}^d \times \mathbb{R}^q \to \mathbb{R}$, set $m(x) = E\{g(X, V)|X = x\}$. Suppose further that the distribution of (X, V) has a joint density and that Assumptions (K), (B) and (D) are satisfied.*

1. *If $m(x)$ is $d+1$ times continuously differentiable on $\mathcal{I}$, then*

$$\sup_{x \in \mathcal{I}} \left| E\{g(X, V)K_b(X, x)\} - f(x)m(x) \right| = o_p(n^{-1/2}).$$

2. *Further, if $E\{g^2(X, V)\}$ is finite, then*

$$\sup_{x \in \mathcal{I}} \left| \frac{1}{n} \sum_{i=1}^{n} g(X_i, V_i)K_b(X_i, x) - f(x)m(x) \right| = o_p(n^{-1/4}).$$

The vector X is our covariate vector from Assumption (X), whereas V is an arbitrary random vector, so the results can be used at various points of the proof. We will, for example, use Lemma 2.1 with Z in place of $g(X, V)$. The proof is given in the Appendix.

Theorem 2.1. *Suppose Assumptions (X), (K), (B) and (D) are satisfied. Then the nonparametric estimator $\widehat{H}_{np}$ given in (1.3) has the expansion*

$$\widehat{H}_{np} = \frac{1}{n} \sum_{i=1}^{n} \left[\chi(X_i) + \frac{Z_i}{\pi(X_i)} \{h(X_i, Y_i) - \chi(X_i)\} \right] + o_p(n^{-1/2}).$$

Proof. To prove this theorem we write $\widehat{H}_{np} = A + B_1$, where

$$A = \frac{1}{n} \sum_{i=1}^{n} [\chi(X_i) + Z_i\{h(X_i, Y_i) - \chi(X_i)\}], \tag{2.3}$$

$$B_1 = \frac{1}{n} \sum_{i=1}^{n} (1 - Z_i)\{\widehat{\chi}(X_i) - \chi(X_i)\}.$$

We will show that B_1 and the term B_3 given below are asymptotically equivalent. This will be established using Assumptions (X), (K), (B) and (D) and Lemma

2.1. Then we show that the leading term of $\widehat{H}_{np}$ stated in the theorem is an approximation of $A + B_3$.

We first introduce

$$B_2 = \frac{1}{n}\sum_{i=1}^{n}(1 - Z_i)\{\phi(X_i) - \widetilde{\phi}(X_i)\}$$

with

$$\phi(X_i) = \frac{n^{-1}\sum_{j=1}^{n} Z_j K_b(X_j, X_i) h(X_j, Y_j)}{\pi(X_i) f(X_i)},$$
$$\widetilde{\phi}(X_i) = \frac{n^{-1}\sum_{j=1}^{n} Z_j K_b(X_j, X_i) \chi(X_i)}{\pi(X_i) f(X_i)}.$$

Then we define B_3 as the conditional expectation of B_2 given the completely observed cases "$\mathcal{B}$", i.e.

$$B_3 = E(B_2|\mathcal{B}).$$

Formally $\mathcal{B}$ stands for the subset $\{(X_j, Y_j, Z_j), j = 1, \ldots, n : Z_j = 1\}$. A fairly straightforward but lengthy calculation yields the asymptotic equivalence of B_1 and B_3,

$$n^{1/2}|B_1 - B_3| = o_p(1). \tag{2.4}$$

The detailed proof of (2.4) is provided in the Appendix. It remains to examine B_3 more closely. Assume that (X_p, Y_p, Z_p) does not belong to the set of complete observations $\mathcal{B}$. We have

$$\begin{aligned}
B_3 &= E(B_2|\mathcal{B}) \\
&= E\Big[\frac{1}{n}\sum_{i=1}^{n}(1 - Z_i)\frac{n^{-1}\sum_{j=1}^{n} Z_j K_b(X_j, X_i)\{h(X_j, Y_j) - \chi(X_i)\}}{\pi(X_i) f(X_i)}\Big|\mathcal{B}\Big] \\
&= E\Big[\{1 - \pi(X_p)\}\frac{n^{-1}\sum_{j=1}^{n} Z_j K_b(X_j, X_p)\{h(X_j, Y_j) - \chi(X_p)\}}{\pi(X_p) f(X_p)}\Big|\mathcal{B}\Big] \\
&= \frac{1}{n}\sum_{j=1}^{n} Z_j E\Big[\{1 - \pi(X_p)\}\frac{K_b(X_j, X_p)\{h(X_j, Y_j) - \chi(X_p)\}}{\pi(X_p) f(X_p)}\Big|\mathcal{B}\Big] \\
&= \frac{1}{n}\sum_{j=1}^{n} Z_j h(X_j, Y_j) E\Big[\frac{\{1 - \pi(X_p)\}K_b(X_j, X_p)}{\pi(X_p) f(X_p)}\Big|X_j\Big] \\
&\qquad -\frac{1}{n}\sum_{j=1}^{n} Z_j E\Big[\frac{\{1 - \pi(X_p)\}K_b(X_j, X_p)\chi(X_p)}{\pi(X_p) f(X_p)}\Big|X_j\Big] \\
&= \frac{1}{n}\sum_{j=1}^{n} Z_j h(X_j, Y_j)\frac{1 - \pi(X_j)}{\pi(X_j)} - \frac{1}{n}\sum_{j=1}^{n} Z_j \chi(X_j)\frac{1 - \pi(X_j)}{\pi(X_j)} + o_P(n^{-1/2}) \\
&= \frac{1}{n}\sum_{j=1}^{n} Z_j\{h(X_j, Y_j) - \chi(X_j)\}\frac{1 - \pi(X_j)}{\pi(X_j)} + o_P(n^{-1/2}).
\end{aligned}$$

The last but one step follows from the first conclusion in Lemma 2.1. This combined with (2.3) and (2.4) gives the expansion provided in the theorem:

$$\begin{aligned}\widehat{H}_{np} &= A + B_1 \\ &= A + B_3 + o_p(n^{-1/2}) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[\chi(X_i) + \frac{Z_i}{\pi(X_i)}\{h(X_i, Y_i) - \chi(X_i)\}\right] + o_p(n^{-1/2}).\end{aligned}$$

This completes the proof. □

Remark 2.1. *To specify a kernel that satisfies Assumption (K), we can extend the construction of second order boundary kernels in Section 3.3.1 of Simonoff (1996 [19]) to higher order boundary kernels. Consider, for example, the case* $d = 2$ *and* $X = (X_1, X_2)^\top$, $s = (s_1, s_2)^\top$ *and* $x = (x_1, x_2)^\top$. *Based on four different univariate bounded functions* $L_i(\cdot)$, $i = 1, \ldots, 4$, *which satisfy* $\int |s_1|^3 |L_i(s_1)| ds_1 < \infty$, $|\partial L_i(s_1)/\partial s_1| < \eta$, *and* $|\partial L_i(s_1)/\partial s_1| < \eta |s_1|^\nu$ *for any* s_1 *satisfying* $|s_1| > \zeta$, *with some constants* η, ζ *and* $\nu > 1$, *we first calculate second order kernels*

$$T_1(s_1, x_1) = \frac{\ell_2^{(1)}(x_1) L_1(s_1) - \ell_1^{(1)}(x_1) L_2(s_1)}{\ell_2^{(1)}(x_1)\ell_1^{(0)}(x_1) - \ell_1^{(1)}(x_1)\ell_2^{(0)}(x_1)},$$

$$T_2(s_1, x_1) = \frac{\ell_4^{(1)}(x_1) L_3(s_1) - \ell_3^{(1)}(x_1) L_4(s_1)}{\ell_4^{(1)}(x_1)\ell_3^{(0)}(x_1) - \ell_3^{(1)}(x_1)\ell_4^{(0)}(x_1)},$$

with $\ell_i^{(j)}(x_1) = \int_{\mathcal{S}_{b,1}(x_1)} s_1^j L_i(s_1) ds_1$ *and* $S_{b,1}(x_1) = \{b^{-1}(y_1 - x_1) : y_1 \in \mathcal{I}_1\}$, *where* $\mathcal{I}_1$ *denotes the support of* X_1. *Then the linear combination of* $T_1(s_1, x_1)$ *and* $T_2(s_1, x_1)$,

$$K_1(s_1, x_1) = \frac{t_2^{(2)}(x_1) T_1(s_1, x_1) - t_1^{(2)}(x_1) T_2(s_1, x_1)}{t_2^{(2)}(x_1) - t_1^{(2)}(x_1)},$$

with $t_i^{(j)}(x_1) = \int_{\mathcal{S}_{b,1}(x_1)} s_1^j T_i(s_1, x_1) ds_1$, *is a univariate third order boundary kernel for* X_1. *A third order boundary kernel* $K_2(s_2, x_2)$ *for* X_2 *can be constructed analogously. By taking the product we obtain the desired bivariate third order boundary kernel* $K(s, x) = K_1(s_1, x_1) K_2(s_2, x_2)$ *for* X.

The construction of general multivariate higher order boundary kernels is done analogously. For $j = 2, 3 \ldots, d$, *we first calculate univariate* j*-th order boundary kernels* T_1 *and* T_2, *and then a univariate* $(j+1)$*-th order boundary kernel as the linear combination given above. The product of* j *such univariate* $(j+1)$*-th order boundary kernels yields a multivariate* $(j+1)$*-th order boundary kernel* K.

A multivariate $(d+1)$*-th order boundary kernel constructed in this way will satisfy Assumption (K). If the boundary of the support* $\mathcal{I}$ *is unknown, it can be estimated using extreme values, i.e.* $(min_{1 \le i \le n}\{X_{i1}\}, \ldots, min_{1 \le i \le n}\{X_{id}\})^\top$ *and* $(max_{1 \le i \le n}\{X_{i1}\}, \ldots, max_{1 \le i \le n}\{X_{id}\})^\top$.

2.2. Expansion of the correction term

To expand the additive correction

$$\widehat{\Gamma} = \frac{1}{n}\sum_{i=1}^{n} Z_i \widehat{g}(X_i)\widehat{\varepsilon}_i,$$

the following assumption is required:

Assumption (S). *The function $\rho_h(X) = E\{h(X,Y)\varepsilon|X\}$ is square-integrable.*

Under Assumption (R) on the regression function and Assumption (S) we expand the nonlinear correction $\widehat{\Gamma}$ in the next theorem. Remember that $g(x) = \rho_h(x)/\{\pi(x)\sigma^2(x)\}$.

Theorem 2.2. *Suppose that Assumptions (X), (R), (T) and (S) hold and that $\widehat{g}(x)$ is a consistent estimator of $g(x)$, uniformly in $x \in \mathcal{I}$. Then the nonlinear correction $\widehat{\Gamma} = \sum_{i=1}^{n} Z_i \widehat{g}(X_i)\widehat{\varepsilon}_i/n$ has the expansion*

$$\widehat{\Gamma} = \frac{1}{n}\sum_{i=1}^{n} \frac{Z_i\varepsilon_i}{\sigma^2(X_i)}\Big\{\frac{\rho_h(X_i)}{\pi(X_i)} - \dot{r}_\vartheta(X_i)^\top I^{-1}\Delta\Big\} + o_p(n^{-1/2})$$

with $\Delta = E\{Z\dot{r}_\vartheta(X)g(X)\} = E\{\dot{r}_\vartheta(X)h(X,Y)\sigma^{-2}(X)\varepsilon\}$.

Proof. Consider $\Gamma = \sum_{i=1}^{n} Z_i g(X_i)\widehat{\varepsilon}_i/n$. We begin with an auxiliary result. A first order Taylor expansion, using Assumption (R), yields

$$\begin{aligned}
&\sum_{i=1}^{n}[g(X_i)\{r_\tau(X_i) - r_\vartheta(X_i) - \dot{r}_\vartheta(X_i)^\top(\tau-\vartheta)\}]^2 \\
&\qquad = \sum_{i=1}^{n} g^2(X_i)\Big[\int_0^1 \{\dot{r}_{\vartheta+u(\tau-\vartheta)}(X_i) - \dot{r}_\vartheta(X_i)\}^\top(\tau-\vartheta)\,du\Big]^2 \\
&\qquad \le \|\tau-\vartheta\|^2 \sum_{i=1}^{n} g^2(X_i)\int_0^1 \|\dot{r}_{\vartheta+u(\tau-\vartheta)}(X_i) - \dot{r}_\vartheta(X_i)\|^2\,du \\
&\qquad \le \|\tau-\vartheta\|^4 \sum_{i=1}^{n} g^2(X_i)L^2.
\end{aligned}$$

This combined with the square integrability of $\rho_h(X)$, Assumption (S), guarantees for any constant c that

$$\sup_{\|\tau-\vartheta\|\le cn^{-1/2}} \sum_{i=1}^{n}\{g(X_i)[r_\tau(X_i) - r_\vartheta(X_i) - \dot{r}_\vartheta(X_i)(\tau-\vartheta)]\}^2 = o_p(1). \quad (2.5)$$

We now approximate Γ by $\sum_{i=1}^{n} Z_i g(X_i)\varepsilon_i^*/n$, where $\varepsilon_i^* = \varepsilon_i - \dot{r}_\vartheta(X_i)^\top(\widehat{\vartheta}-\vartheta)$. We have

$$\Big|\frac{1}{n}\sum_{i=1}^{n} Z_i g(X_i)(\widehat{\varepsilon}_i - \varepsilon_i^*)\Big| \;\le\; \frac{1}{n}\sum_{i=1}^{n} Z_i|g(X_i)(\widehat{\varepsilon}_i - \varepsilon_i^*)|$$

$$\leq \frac{1}{n}\Big\{n\sum_{i=1}^{n} Z_i g^2(X_i)(\widehat{\varepsilon}_i - \varepsilon_i^*)^2\Big\}^{1/2}$$
$$= n^{-1/2}\Big\{\sum_{i=1}^{n} Z_i g^2(X_i)(\widehat{\varepsilon}_i - \varepsilon_i^*)^2\Big\}^{1/2}. \quad (2.6)$$

The second relation uses the Cauchy-Schwarz inequality. Now apply (2.5) to obtain

$$\sum_{i=1}^{n} Z_i g^2(X_i)(\widehat{\varepsilon}_i - \varepsilon_i^*)^2 = \sum_{i=1}^{n} Z_i g^2(X_i)[\widehat{\varepsilon}_i - \{\varepsilon_i - \dot{r}_\vartheta(X_i)^\top(\widehat{\vartheta} - \vartheta)\}]^2$$
$$\leq \sum_{i=1}^{n} g^2(X_i)\{r_{\widehat{\vartheta}}(X_i) - r_\vartheta(X_i) - \dot{r}_\vartheta(X_i)^\top(\widehat{\vartheta} - \vartheta)\}^2 = o_p(1). \quad (2.7)$$

This combined with (2.6) gives

$$\frac{1}{n}\sum_{i=1}^{n} Z_i g(X_i)\widehat{\varepsilon}_i$$
$$= \frac{1}{n}\sum_{i=1}^{n} Z_i g(X_i)\varepsilon_i^* + o_p(n^{-1/2})$$
$$= \frac{1}{n}\sum_{i=1}^{n} Z_i g(X_i)\varepsilon_i - \frac{1}{n}\sum_{i=1}^{n} Z_i g(X_i)\dot{r}_\vartheta^\top(X_i)(\widehat{\vartheta} - \vartheta) + o_p(n^{-1/2})$$
$$= \frac{1}{n}\sum_{i=1}^{n} Z_i g(X_i)\varepsilon_i - E\{Zg(X)\dot{r}_\vartheta(X)^\top\}(\widehat{\vartheta} - \vartheta) + o_p(n^{-1/2})$$
$$= \frac{1}{n}\sum_{i=1}^{n} Z_i g(X_i)\varepsilon_i - \frac{1}{n}E\{Zg(X)\dot{r}_\vartheta(X)^\top\}I^{-1}\sum_{i=1}^{n} Z_i\sigma^{-2}(X_i)\dot{r}_\vartheta(X_i)\varepsilon_i$$
$$+ o_p(n^{-1/2})$$
$$= \frac{1}{n}\sum_{i=1}^{n} \frac{Z_i\varepsilon_i}{\sigma^2(X_i)}\Big\{\frac{\rho_h(X_i)}{\pi(X_i)} - \dot{r}_\vartheta(X_i)^\top I^{-1}\Delta\Big\} + o_p(n^{-1/2}). \quad (2.8)$$

Here we use the law of large numbers in the third step and the fourth equation uses the asymptotic linearity of $\widehat{\vartheta}$ stated in Assumption (T). This gives the influence function of the correction term, which involves the unknown quantity $g(x) = \rho_h(x)/\{\pi(x)\sigma^2(x)\}$. Replacing $g(x)$ by a uniformly consistent estimator $\widehat{g}(x)$ does not change the asymptotic expansion because

$$\Big|n^{-1}\sum_{i=1}^{n} Z_i\{g(X_i) - \widehat{g}(X_i)\}\widehat{\varepsilon}_i\Big| = o_p(n^{-1/2}). \quad (2.9)$$

We verify (2.9) in the Appendix. This completes the proof. □

A common estimator of $g(x)$ is $\widehat{g}(x) = \widehat{\rho}_h(x)/\{\widehat{\sigma}^2(x)\widehat{\pi}(x)\}$ with $\widehat{\rho}_h(x)$, $\widehat{\sigma}^2(x)$ and $\widehat{\pi}(x)$ being Nadaraya-Watson estimators of $\rho_h(x)$, $\sigma^2(x)$ and $\pi(x)$, respectively. The Nadaraya-Watson estimator is uniformly consistent when X has a compact support. In Section 4 we will use this estimator for our simulation study, and also show more details.

We conclude the section with the final expansion of our estimator $\widehat{H} = \widehat{H}_{np} - \widehat{\Gamma}$. The result follows directly from the statements in Theorems 2.1 and 2.2 on $\widehat{H}_{np}$ and $\widehat{\Gamma}$. We therefore formulate the result as a corollary.

Corollary 2.1. *Write $\Delta = E\{\dot{r}_\vartheta(X)h(X,Y)\sigma^{-2}(X)\varepsilon\}$ as in Theorem 2.2 and let the assumptions of that theorem be satisfied. Suppose that Assumptions (K), (B), (X) and (D) from Section 2.1 hold true. Then the estimator $\widehat{H} = \widehat{H}_{np} - \widehat{\Gamma}$ from equation (1.2) has the expansion*

$$\begin{aligned} &n^{1/2}[\widehat{H} - E\{h(X,Y)\}] \\ &= n^{-1/2}\sum_{i=1}^{n}\Big[\chi(X_i) - E\{h(X,Y)\} + \frac{Z_i}{\pi(X_i)}\{h(X_i,Y_i) - \chi(X_i)\} \\ &\qquad - \frac{Z_i\varepsilon_i}{\sigma^2(X_i)}\Big\{\frac{\rho_h(X_i)}{\pi(X_i)} - \dot{r}_\vartheta(X_i)^\top I^{-1}\Delta\Big\}\Big] + o_p(n^{-1/2}). \end{aligned}$$

3. Efficiency

In this section we calculate the canonical gradient of $E\{h(X,Y)\}$, which characterizes the influence function of an efficient estimator of that expectation. The efficiency of our estimator will be established by showing that the canonical gradient equals the influence function obtained in Section 2. We will use results from Müller et al. (2006 [15]) and Müller (2009 [12]) about the canonical gradient, and also from Schick (1993 [17]) about the tangent space in nonlinear regression.

Essential for the derivation of canonical gradients is the notion of tangent space: a canonical gradient is characterized as an orthogonal projection of a gradient onto the tangent space, which is the closed linear span of the set of all perturbations of the joint distribution $P(dx, dy, dz)$ within the model. The distribution P depends on the marginal distribution $G(dx)$ of X, the conditional probability $\pi(x)$ of $Z = 1$ given $X = x$ and the conditional distribution $Q(x, dy)$ of Y given $X = x$. Müller et al. (2006 [15]), who also considers regression models with MAR responses, were the first to describe the tangent space for general differentiable functionals $\kappa(G, Q, \pi)$ in this model. They write the joint distribution in the form

$$P(dx, dy, dz) = G(dx)B_{\pi(x)}\{zQ(x, dy) + (1 - z)\delta_0(dy)\},$$

where $B_p = p\delta_1 + (1 - p)\delta_0$ denote the Bernoulli distribution with parameter p and δ_t the Dirac measure at t. To specify the tangent space we assume that G,

Q and π are Hellinger differentiable:

$$\begin{aligned} G_{nu}(dx) &\doteq G(dx)\{1 + n^{-1/2}u(x)\}, \\ Q_{nv}(x, dy) &\doteq Q(x, dy)\{1 + n^{-1/2}v(x, y)\}, \\ B_{\pi_{nw}(x)}(dz) &\doteq B_{\pi(x)}(dz)[1 + n^{-1/2}\{z - \pi(x)w(x)\}], \end{aligned} \tag{3.1}$$

where $\doteq$ means ignoring $o_p(n^{-1/2})$ items. Since the perturbed distributions are probability distributions, the Hellinger derivative u belongs to

$$L_{2,0}(G) = \Big\{u \in L_2(G) : \int udG = 0\Big\},$$

v belongs to

$$V_0 = \Big\{v \in L_2(M) : \int v(x, y)Q(x, dy) = 0\Big\},$$

with $M(dx, dy) = Q(x, dy)G(dx)$, and w belongs to

$$W = \big\{w \in L_2(G_\pi) : G_\pi(dx) = \pi(x)\{1 - \pi(x)\}G(dx)\big\}.$$

The tangent space is the orthogonal sum

$$\{u(X) : u \in U\} \oplus \{Zv(X, Y) : v \in V\} \oplus \{\{Z - \pi(X)\}w(X) : w \in W\}.$$

As in Müller et al. (2006 [15]), we have no structural assumptions on G and π. This means that we have no further restrictions on the perturbations u and w and can therefore take $U = L_{2,0}(G)$ and $W = L_2(G_\pi)$. We must, however, take the regression structure into account, i.e. the space V is the subset of V_0 to which v is now restricted. In the following we assume that the subspaces U, V and W are closed and linear.

The canonical gradient g_* is an element of the tangent space and has the form

$$g_*(X, ZY, Z) = u_*(X) + Zv_*(X, Y) + \{Z - \pi(X)\}w_*(X), \tag{3.2}$$

where $u_*(X)$, $Zv_*(X)$ and $\{Z - \pi(X)\}w_*(X)$ are projections of the gradient (that characterizes the differentiable functional) onto the three orthogonal subspaces of the tangent space.

For full details of the results we have just summarized, see pages 352–355 in Müller et al. (2006 [15]). There they provide a detailed characterization of efficient estimators in the model with MAR responses and then specialize them to four specific models for the conditional distribution Q. In the current paper we have $Q(x, dy) = f\{y - r_\vartheta(x)|x\}dy$, where $f(\cdot|x)$ denotes the conditional density of the (conditional mean zero) error distribution given $X = x$. In order to find V we introduce perturbations of the parameter ϑ and the conditional error density. The exact form of V is only relevant for the derivation of v_* and is therefore located in the Appendix; see Section A.4 for details. The derivation of u_* and w_* is given in the proof of Theorem 3.1 below. In that theorem we provide the explicit representation of the canonical gradient of $E\{h(X, Y)\}$. The efficiency of our estimator is formulated subsequently in Corollary 3.1

Theorem 3.1. *Let the vector Δ and the matrix I be defined as in Section 2, i.e. $\Delta = E\{\dot{r}_\vartheta(X)h(X,Y)\sigma^{-2}(X)\varepsilon\}$ and $I = E\{Z\dot{r}_\vartheta(X)\dot{r}_\vartheta(X)^\top\sigma^{-2}(X)\}$. Suppose Assumptions (R), (S) and the Hellinger differentiability assumption (3.1) are satisfied. Further assume that the conditional density of ε given x, $f(\cdot|x)$, has a finite Fisher information I and that I is invertible. Then the canonical gradient of the functional $E\{h(X,Y)\}$ is*

$$g_*(X,ZY,Z) = \chi(X) - E\{h(X,Y)\} + \frac{Z}{\pi(X)}\{h(X,Y) - \chi(X)\} - \frac{Z\varepsilon}{\sigma^2(X)}\Big\{\frac{\rho_h(X)}{\pi(X)} - \dot{r}_\vartheta(X)^\top I^{-1}\Delta\Big\}.$$

Proof. Müller et al. (2006 [15]) and Müller (2009 [12]) show that the canonical gradient $g_*(X,ZY,Z)$ from (3.2), now specifically for the functional $E\{h(X,Y)\}$, is determined by

$$\begin{aligned} &E\{u_*(X)u(X)\} + E\{Zv_*(X,Y)v(X,Y)\} \\ &\quad + E[\{Z-\pi(X)\}^2 w_*(X)w(X)] = E[h(X,Y)\{u(X)+v(X,Y)\}] \end{aligned} \tag{3.3}$$

for all $u \in U$, $v \in V$ and $w \in W$. Here we use the fact that the canonical gradient of $E\{h(X,Y)\}$ is a projection of a gradient of $E\{h(X,Y)\}$ onto the tangent space. To determine the specific form of g_* we set $u = 0$ and $v = 0$ in (3.3), which gives

$$w_* = 0. \tag{3.4}$$

Then, setting $v = 0$ in (3.3) yields that $u_*(X)$ is the projection of $h(X,Y)$ onto U:

$$u_*(X) = E\{h(X,Y)|X\} - E\{h(X,Y)\} = \chi(X) - E\{h(X,Y)\}. \tag{3.5}$$

In order to find v_* we must take the parametric model structure into account, i.e. the special form of the subset $V \subset V_0$. The derivation of V and v_* is quite elaborate and therefore given in Section A.4 of the Appendix where we prove

$$v_*(X,Y) = \frac{\varepsilon}{\sigma^2(X)}\dot{r}_\vartheta(X)^\top I^{-1}\Delta + \frac{1}{\pi(X)}\Big\{h(X,Y) - \chi(X) - \frac{\varepsilon\rho_h(X)}{\sigma^2(X)}\Big\}. \tag{3.6}$$

Combining equations (3.2), (3.4), (3.5) and (3.6) yields the canonical gradient of $E\{h(X,Y)\}$ given in the theorem. □

It follows from Corollary 2.1 that our estimator is asymptotically linear, and from Theorem 3.1 that the influence function given in Corollary 2.1 equals the canonical gradient g_* from Theorem 3.1. Hence our estimator is efficient in the sense of the Hájek and Le Cam theory, which implies asymptotic normality. We formulate this as a corollary.

Corollary 3.1. *Let the assumptions of Corollary 2.1 and Theorem 3.1 be satisfied. Then the estimator $\widehat{H} = \widehat{H}_{np} - \widehat{\Gamma}$ introduced in equation (1.2) is asymptotically efficient. In particular it is asymptotically normally distributed with variance $E\{g_*(X,ZY,Z)^2\}$, with $g_*(X,ZY,Z)$ specified in Theorem 3.1 above.*

4. Simulations

4.1. Linear and nonlinear regression with one covariate

To illustrate the results of the previous sections, we conduct a simulation study comparing various estimators for $E(Y)$, $E(Y^2)$, $E(XY)$ and $E\{X\exp(XY)\}$; see Tables 1-4. In each case we consider two regression functions, $r_\vartheta(x) = \vartheta x$ and $r_\vartheta(x) = \cos(\vartheta x)$ with $\vartheta = 2$, and two variance functions, namely a linear variance function $\sigma^2(x) = 0.6 - 0.5x$ and a parabolic variance function $\sigma^2(x) = (x - 0.4)^2 + 0.1$. The covariate X is generated from a uniform distribution on $[-1, 1]$ and the error variable η in $\varepsilon = \sigma(X)\eta$ from a standard normal distribution. In all scenarios we use the logistic distribution function $\pi(x) = 1/\{1 + \exp(-x)\}$ for the conditional probability, so that about half of the simulated responses are missing. In this section we use, for simplicity, only ordinary kernels instead of the boundary kernels discussed in Remark 2.1.

To evaluate the performance of our asymptotically optimal estimator when sample sizes are small we simulate the mean squared errors (MSE) of $\widehat{H}_\vartheta$ and $\widehat{H}_{\widehat{\vartheta}}$. Here $\widehat{H}_\vartheta$ denotes the version of our estimator $\widehat{H}$ from (1.2) that uses the true values of $\sigma^2(x)$, $\pi(x)$ and ϑ in the correction term, whereas $\widehat{H}_{\widehat{\vartheta}}$ uses estimators for those quantities. For the calculation of

$$\widehat{\vartheta} = \arg\min_\theta \Big| \sum_{i=1}^n Z_i \widehat{\sigma}^{-2}(X_i) \dot{r}_\theta(X_i)\{Y_i - r_\theta(X_i)\} \Big|$$

we use a consistent nonparametric estimator for $\sigma^2(x)$, namely

$$\widehat{\sigma}^2(x) = \frac{\sum_{i=1}^n Z_i K_{b_1}(x - X_i)\{Y_i - r_{\widehat{\vartheta}_0}(X_i)\}^2}{\sum_{i=1}^n Z_i K_{b_1}(x - X_i)},$$

where $K_{b_1}(\cdot)$ is a Gaussian kernel with bandwidth b_1 and $\widehat{\vartheta}_0$ is the ordinary least squares estimator (or some other consistent estimator of ϑ). In the model with a linear regression function $\widehat{\vartheta}$ and $\widehat{\vartheta}_0$ have a closed form, while for the cosine function we use the `nls` function in `R` to obtain them. Our nonparametric estimator for $\pi(x)$ is

$$\widehat{\pi}(x) = \frac{\sum_{i=1}^n Z_i K_{b_2}(x - X_i)}{\sum_{i=1}^n K_{b_2}(x - X_i)}, \tag{4.1}$$

where $K_{b_2}(\cdot)$ is a Gaussian kernel with bandwidth b_2; $\widehat{\rho}_h(x)$ is a plug-in estimator for $\rho_h(x) = E\{h(X, Y)\varepsilon | X = x\}$. For our choices of h it will involve the estimators $\widehat{\sigma}^2(x)$ and $\widehat{\vartheta}$ just described; see below for more details. The nonparametric part $\widehat{H}_{np}$ of our estimator $\widehat{H}$ is the partially imputed estimator (1.3). It is based on a Nadaraya-Watson estimator for the conditional expectation $\chi(x) = E\{h(X, Y)|X = x\}$, with a Gaussian kernel $K_{b_3}(\cdot)$ with bandwidth b_3.

We also compare $\widehat{H}_\vartheta$ and $\widehat{H}_{\widehat{\vartheta}}$ with the simple Horvitz-Thompson type estimator $S = n^{-1}\sum_{i=1}^n \{Z_i h(X_i, Y_i)/\pi(X_i)\}$ (based on the true $\pi(x)$), and the

nonparametric estimator $\widehat{H}_{np}$ without the nonlinear correction. For each setting simulations with sample sizes $n = 50, 100$ and 200 are conducted based on $5,000$ repetitions. The `nls` routine does not always converge for the cosine regression function. We therefore list the MSEs of $\widehat{H}_{\widehat{\vartheta}}$ only for sample sizes $n = 100$ and $n = 200$ for that scenario.

For estimators involving kernel estimation we use leave-one-out cross validation to select the bandwidth. For example, to obtain the bandwidth b_1 of $\widehat{\sigma}^2(x)$, we first calculate, for each complete observation (X_j, Y_j),

$$\widetilde{\sigma}^2_{jb} = \sum_{\substack{i=1 \\ i\neq j}}^{n} Z_i K_b(X_j - X_i)\{Y_i - r_{\widehat{\vartheta}_0}(X_i)\}^2 \Big/ \sum_{\substack{i=1 \\ i\neq j}}^{n} Z_i K_b(X_j - X_i),$$

for bandwidths b from a candidate set $\mathcal{G}$. Then b_1 is obtained as

$$b_1 = \arg\min_{b\in\mathcal{G}} \sum_{i=1}^{n} Z_i[\widetilde{\sigma}^2_{ib} - \{Y_i - r_{\widehat{\vartheta}_0}(X_i)\}^2]^2.$$

For the case $h(x, y) = y$, for example, we used the set $\mathcal{G} = \{0.1, 0.2, ...0.5\}$ for b_1 and also for b_2. The bandwidth b_3 for the nonparametric part $\widehat{H}_{np}$ has the form $b_3 = an^{-2/5}$, which is indicated to have optimal convergence rate by Cheng (1994 [3]). We chose $a = 0.5, 0.6, \ldots, 0.9$ to determine b_3.

TABLE 1
Simulated MSEs of estimators of $E(Y)$

		$r_\vartheta(X) = \vartheta X \quad (\vartheta = 2)$				$r_\vartheta(X) = \cos(\vartheta X) \quad (\vartheta = 2)$			
$\sigma^2(X)$	n	$\widehat{H}_\vartheta$	$\widehat{H}_{\widehat{\vartheta}}$	$\widehat{H}_{np}$	S	$\widehat{H}_\vartheta$	$\widehat{H}_{\widehat{\vartheta}}$	$\widehat{H}_{np}$	S
(a)	50	0.0341	0.0319	0.0634	0.0941	0.0102	–	0.0390	0.0426
	100	0.0152	0.0144	0.0291	0.0443	0.0037	0.0071	0.0179	0.0215
	200	0.0070	0.0067	0.0144	0.0229	0.0015	0.0031	0.0085	0.0104
(b)	50	0.0353	0.0323	0.0655	0.0968	0.0112	–	0.0411	0.0451
	100	0.0157	0.0146	0.0308	0.0462	0.0041	0.0082	0.0195	0.0233
	200	0.0072	0.0068	0.0151	0.0236	0.0016	0.0035	0.0092	0.0112

The entries in both the left and the right panels are the simulated mean squared errors of estimators of the mean response. The first two columns of each panel show the MSEs of the two versions $\widehat{H}_\vartheta$ and $\widehat{H}_{\widehat{\vartheta}}$ of the efficient estimator. The third and fourth columns list the results for the nonparametric estimator $\widehat{H}_{np}$ (no correction) and the simple estimator S. The variance functions are (a) $\sigma^2(X) = 0.6 - 0.5X$ and (b) $\sigma^2(X) = (X - 0.4)^2 + 0.1$.

The simulated mean squared errors for estimating the mean response are given in Table 1. In this case $\rho_h(x) = E\{h(X, Y)\varepsilon|X = x\} = E\{Y\varepsilon|X = x\} = \sigma^2(x)$. In each row of Table 1 the efficient estimator outperforms the nonparametric estimator without the nonlinear correction, while the simple estimator is inferior to any of its competitors. In the linear regression model the two versions $\widehat{H}_\vartheta$ and $\widehat{H}_{\widehat{\vartheta}}$ of the efficient estimator differ slightly, in contrast to the cosine regression model, where the difference is quite large. This is because the estimator of the slope in the linear regression model is better than that of the frequency

parameter in the model with the cosine regression function. The MSEs for different sample sizes confirm the root-n convergence rate of the efficient estimator, as stated in Corollary 2.1.

TABLE 2
Simulated MSEs of estimators of $E(Y^2)$

		$r_\vartheta(X)=\vartheta X \quad (\vartheta=2)$				$r_\vartheta(X)=\cos(\vartheta X) \quad (\vartheta=2)$			
$\sigma^2(X)$	n	$\widehat{H}_\vartheta$	$\widehat{H}_{\widehat{\vartheta}}$	$\widehat{H}_{np}$	S	$\widehat{H}_\vartheta$	$\widehat{H}_{\widehat{\vartheta}}$	$\widehat{H}_{np}$	S
	50	0.1235	0.1725	0.2755	0.4206	0.0626	–	0.1065	0.1267
(a)	100	0.0545	0.0818	0.1394	0.2099	0.0285	0.0275	0.0493	0.0630
	200	0.0247	0.0381	0.0660	0.1029	0.0134	0.0133	0.0234	0.0307
	50	0.1973	0.2484	0.4135	0.6012	0.0924	–	0.1207	0.1520
(b)	100	0.0891	0.1180	0.2130	0.3060	0.0456	0.0448	0.0592	0.0786
	200	0.0402	0.0544	0.1010	0.1475	0.0215	0.0214	0.0281	0.0375

The entries are mean squared errors as in Table 1, now with $h(x,y)=y^2$. The variance functions are again (a) $\sigma^2(X)=0.6-0.5X$ and (b) $\sigma^2(X)=(X-0.4)^2+0.1$.

Table 2 displays the simulation results for the same scenario as in Table 1, but now the second moment of the response is estimated. The efficient estimator again outperforms both the nonparametric estimator and the simple estimator. In our scenario with normal errors we have $\rho_h(x)=2r_\vartheta(x)\sigma^2(x)$ with $r_\vartheta(x)=\vartheta x$ and $r_\vartheta(x)=\cos(\vartheta x)$. In both regression models it makes little difference whether true values or estimators are used.

TABLE 3
Simulated MSEs of estimators of $E(XY)$

		$r_\vartheta(X)=\vartheta X \quad (\vartheta=2)$				$r_\vartheta(X)=\cos(\vartheta X) \quad (\vartheta=2)$			
$\sigma^2(X)$	n	$\widehat{H}_\vartheta$	$\widehat{H}_{\widehat{\vartheta}}$	$\widehat{H}_{np}$	S	$\widehat{H}_\vartheta$	$\widehat{H}_{\widehat{\vartheta}}$	$\widehat{H}_{np}$	S
	50	0.0120	0.0147	0.0215	0.0402	0.0034	–	0.0136	0.0131
(a)	100	0.0050	0.0068	0.0105	0.0203	0.0013	0.0009	0.0065	0.0068
	200	0.0022	0.0030	0.0048	0.0099	0.0005	0.0004	0.0031	0.0032
	50	0.0131	0.0159	0.0268	0.0456	0.0044	–	0.0186	0.0187
(b)	100	0.0055	0.0073	0.0134	0.0231	0.0017	0.0011	0.0093	0.0096
	200	0.0023	0.0032	0.0062	0.0113	0.0007	0.0004	0.0044	0.0046

We consider the same scenario as in Tables 1 and 2, now with $h(x,y)=xy$.

The MSEs for estimating $E(XY)$ are given in Table 3. In both regression models $\rho_h(x)=x\sigma^2(x)$. The first two columns of the left panel (linear regression) indicate that estimating $\sigma^2(x)$, $\pi(x)$ and ϑ increases the MSE slightly. The MSEs in the corresponding columns in the right panel (cosine regression) appear to be similar. The results in Table 3 again confirm the superiority of the efficient estimator as well as the convergence rate.

The results for $E\{X\exp(XY)\}$ are listed in Table 4. Straightforward calculations yield $\rho_h(x)=\sigma^2(x)x^2\exp[\{\vartheta+\sigma^2(x)/2\}x^2]$. The efficient estimator clearly outperforms the competing estimators. As in the previous tables we see that the two estimators $\widehat{H}_\vartheta$ and $\widehat{H}_{\widehat{\vartheta}}$ based on true values and on estimates perform similarly.

TABLE 4
Simulated MSEs of estimators of $E\{X\exp(XY)\}$

		$r_\vartheta(X)=\vartheta X$ $(\vartheta=2)$				$r_\vartheta(X)=\cos(\vartheta X)$ $(\vartheta=2)$			
$\sigma^2(X)$	n	$\widehat{H}_\vartheta$	$\widehat{H}_{\widehat{\vartheta}}$	$\widehat{H}_{np}$	S	$\widehat{H}_\vartheta$	$\widehat{H}_{\widehat{\vartheta}}$	$\widehat{H}_{np}$	S
(a)	50	0.5273	0.4491	0.7517	1.2958	0.0286	–	0.0415	0.0727
	100	0.2442	0.2164	0.3693	0.6207	0.0136	0.0144	0.0210	0.0350
	200	0.1289	0.1197	0.2161	0.3420	0.0072	0.0075	0.0122	0.0191
(b)	50	2.2743	1.9389	3.0689	5.0159	0.1148	–	0.1566	0.2547
	100	0.9657	0.8976	1.3701	2.2641	0.0491	0.0504	0.0706	0.1166
	200	0.4589	0.4624	0.7310	1.4083	0.0238	0.0264	0.0383	0.0710

In this table $h(x,y)=x\exp(xy)$; the scenario is the same as in Tables 1-3.

The influence function of the efficient estimator in Corollary 2.1 contains a non-negligible part that comes from the difference $n^{1/2}(\widehat{\vartheta}-\vartheta)$. This part is missing if we replace $\widehat{\vartheta}$ by ϑ, which explains why in some cases, e.g. the upper left panel in Table 4, $\widehat{H}_{\widehat{\vartheta}}$ outperforms $\widehat{H}_\vartheta$. However, estimating $\sigma^2(x)$ and $\pi(x)$ adds uncertainty, especially if n is not very large, so that in other cases, for example in the right panel in Table 4, the MSE of $\widehat{H}_{\widehat{\vartheta}}$ is larger than that of $\widehat{H}_\vartheta$.

4.2. Linear regression with two covariates

Finally we consider a bivariate covariate vector $X=(X_1,X_2)^\top$ and a linear regression function $r_\vartheta(x)=\vartheta_1x_1+\vartheta_2x_2$ with $\vartheta_1=1$ and $\vartheta_2=2$. We modify the scenario of the previous section as follows: the variance function $\sigma^2(x)=\sigma^2(x_1,x_2)$ is set to be $2.1-0.5(x_1+x_2)$ or $(x_1+x_2-0.8)^2+0.1$, and $\pi(x)=1/[1+\exp\{-(x_1+x_2)\}]$. In order to generate correlated covariates X_1,X_2 we first sample auxiliary random variables W,X_1' and X_2' independently: W is generated from a uniform distribution on $[-0.5,0.5]$, and X_1' and X_2' are generated from a uniform distribution on $[-1,1]$. Then we take $X_1=X_1'+W$ and $X_2=X_2'+W$. Our final estimator is based on kernel estimators. For example, $\widehat{\sigma}^2(x)$ now involves a product of two Gaussian-based kernels of order 4 (Wand and Schucany, 1990 [21]), i.e. $K(x)=(3-x^2)\Phi(x)/2$, where $\Phi(\cdot)$ is the standard Gaussian density function, both using the same bandwidth, to estimate the unknown conditional expectations. Table 5 shows the simulated mean squared errors of estimators of the mean response in the bivariate regression model. In this case $\rho_h(x)=E\{h(X,Y)\varepsilon|X=x\}=\sigma^2(x)$. Again our efficient estimator outperforms the competing estimators and confirms our theoretical results. The efficient estimator that uses estimates $\widehat{\sigma}^2(x)$, $\widehat{\pi}(x)$ and $\widehat{\vartheta}$ is better than the estimator $\widehat{H}_\vartheta$, which uses the true values.

5. An example

In this section we apply our method to a data set of 2139 HIV positive patients from a clinical trial (Hammer et al., 1996 [6]). The data are freely accessible in the `R` package `speff2trial`.

TABLE 5
Simulated MSEs of estimators of $E(Y)$

$\sigma^2(X)$	n	$\hat{H}_{\vartheta}$	$\hat{H}_{\hat{\vartheta}}$	$\hat{H}_{np}$	S
$2.1 - 0.5(X_1 + X_2)$	50	0.1081	0.0963	0.2022	0.2991
	100	0.0514	0.0444	0.1013	0.1529
	200	0.0247	0.0214	0.0507	0.0759
$(X_1 + X_2 - 0.8)^2 + 0.1$	50	0.1263	0.1054	0.1876	0.3338
	100	0.0617	0.0514	0.1020	0.1738
	200	0.0297	0.0248	0.0550	0.0861

The entries are simulated mean squared errors of estimators of the mean response, here for the scenario with the bivariate linear regression function $r_{\vartheta}(X) = \vartheta_1 X_1 + \vartheta_2 X_2$ ($\vartheta_1 = 1$, $\vartheta_2 = 2$) described in Section 4.2.

In the trial patients were randomly assigned to four antiretroviral therapies: (i) zidovudine (ZDV) monotherapy, (ii) ZDV + didanosine (DDI), (iii) ZDV + zalcitabine, and (iv) DDI monotherapy. We want to compare the ZDV monotherapy (i) with the alternative group of therapies (ii)-(iv), and estimate the mean number of CD4 cells in both groups, i.e. the number of white blood cells that fight the infection. An increasing CD4 count indicates that the HIV treatment is more effective.

We are interested in the difference between the mean CD4 counts (Y) in the monotherapy group and the mean CD4 counts in the alternative therapy group at 96±5 weeks post therapy. There are six covariates: $X^{(1)}$, age; $X^{(2)}$, weight; $X^{(3)}$, CD4 counts at baseline; $X^{(4)}$, CD4 counts at 20±5 weeks; $X^{(5)}$, CD8 (immune cells) counts at baseline; $X^{(6)}$, CD8 counts at 20±5 weeks. Because of deaths and dropouts, 39% of the responses in the monotherapy group and 37% of the responses of the combined therapy group are missing, while all covariates are observed for all patients. Let Z again denote the missingness indicator (which is 1 if Y is observed and 0 if it is missing). As indicated by Hu et al. (2010 [9]) and Tang et al. (2018 [20]), who consider the same data set, it is reasonable to asume that the conditional expectation of the response given the covariates can be modelled using linear regression, and that the response is missing at random. The variable selection results in Tang et al. (2018 [20]) suggest that only $X^{(3)}$, $X^{(4)}$ and $X^{(6)}$ actually affect Y. We therefore assume $E(Y|X) = \vartheta^\top X$, with a covariate vector $X = (1, X^{(3)}, X^{(4)}, X^{(6)})^\top$ and a regression parameter $\vartheta \in \mathbb{R}^4$.

We apply our method to the two groups of data separately and construct the efficient estimator for the mean response $\mu^{(0)}$ in the monotherapy group and the mean response $\mu^{(1)}$ in the combined therapy group. Then we calculate the difference between the means, $\mu^{(1)} - \mu^{(0)}$. For the construction of the efficient estimator see Section 4.1.

For comparison we also consider the three estimators for the mean difference $\mu^{(1)} - \mu^{(0)}$ in Section 7 of Hu et al. (2010 [9]): inverse probability weighting estimation (IPW), augmented inverse probability weighting estimation (AIPW), and semiparametric dimension reduction estimation (SDR). Besides the linear regression model between Y and X, Hu et al. additionally assume a parametric logistic model for the probablity of missingness, i.e. $\text{logit}\{\pi(X)\} = \gamma^\top X$ for some

parameter γ (which is technically a different statistical model). For the term $\pi(X)$ in the nonlinear correction term of our efficient estimator, we therefore use the nonparametric estimator (4.1) and a parametric estimator for the logistic model, both based on $(X^{(3)}, X^{(4)}, X^{(6)})^\top$.

TABLE 6
Estimates of the mean difference $\mu^{(1)} - \mu^{(0)}$.

	Point estimator	Standard error	95% confidence interval
IPW	58.19	10.33	[37.94, 78.44]
AIPW	61.91	8.83	[44.60, 79.22]
SDR	62.42	9.02	[44.74, 80.10]
EENP	63.75	9.07	[45.98, 81.52]
EEP	63.40	9.08	[45.60, 81.20]

IPW, inverse probability weighting estimation; AIPW, augmented inverse probability weighting estimation; SDR, semiparametric dimension reduction estimation; EENP, efficient estimator with the nonparametric estimator for the probability of missingness; EEP, a version of the efficient estimator with the logistic model for the probability of missingness.

The point estimators, standard errors and 95% confidence intervals of various methods are given in Table 6. The results of the IPW, AIPW and SDR are taken from Hu et al. (2010 [9]) for comparison. The standard errors of the two versions of the efficient estimator (EENP and EEP in Table 6) are obtained using the bootstrap based on 500 repetitions. The point estimators, standard errors and confidence intervals of our method are close to those of the AIPW and SDR, which both attain an efficiency bound if $E(Y|X)$ and $\pi(X)$ are correctly specified, as discussed in Section 3 of Hu et al. (2010 [9]). However, our method is efficient without specifying an auxiliary parametric model for $\pi(X)$. From Table 6 we can see that the results of the EENP and the EEP are very close.

Appendix A

A.1. Proof of Lemma 2.1

Let f_2 denote the joint density of (X, V) and $f(\cdot|x)$ the conditional density of V given $X = x$. For the proof of part (1) we write $\mu(x) = E\{g(X, V)K_b(X, x)\}$ and, using substitution, obtain

$$\begin{aligned}
\mu(x) &= \int_{\mathcal{I}} \int_{\mathbb{R}^q} g(u, v) b^{-d} K\Big(b^{-1}(u - x), x\Big) f_2(u, v) dv du \\
&= \int_{\mathcal{S}_b(x)} \int_{\mathbb{R}^q} g(bs + x, v) K(s, x) f_2(bs + x, v) dv ds \\
&= \int_{\mathcal{S}_b(x)} K(s, x) f(bs + x) \int_{\mathbb{R}^q} g(bs + x, v) f(v|bs + x) dv ds \\
&= \int_{\mathcal{S}_b(x)} K(s, x) f(bs + x) m(bs + x) ds.
\end{aligned}$$

A Taylor expansion gives that for $s = (s_1, \ldots, s_d)^\top \in \mathcal{S}_b(x)$,

$$f(bs+x)m(bs+x) = \sum_{|\alpha|\le d} \frac{D^\alpha\{f(x)m(x)\}}{\alpha!}(bs)^\alpha + \sum_{|\alpha|=d+1} R_\alpha(x,s)(bs)^\alpha,$$

where $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d$, $|\alpha| = \sum_{i=1}^d \alpha_i$, $\alpha! = \prod_{i=1}^d \alpha_i!$, $(bs)^\alpha = b^{|\alpha|}s^\alpha = b^{|\alpha|}\prod_{i=1}^d s_i^{\alpha_i}$ and

$$D^\alpha\{f(x)m(x)\} = \frac{\partial^{|\alpha|}\{f(x)m(x)\}}{\partial x_1^{\alpha_1}\ldots\partial x_d^{\alpha_d}}.$$

For example, if $x = (x_1, x_2)^\top$ and $s = (s_1, s_2)^\top$ are two-dimensional vectors, we have

$$\begin{aligned}&\sum_{|\alpha|=2} \frac{D^\alpha\{f(x)m(x)\}}{\alpha!}(bs)^\alpha \\ &= \frac{\partial^2 f(x)m(x)}{\partial x_1^2}\frac{(bs_1)^2}{2} + \frac{\partial^2 f(x)m(x)}{\partial x_2^2}\frac{(bs_2)^2}{2} + \frac{\partial^2 f(x)m(x)}{\partial x_1\partial x_2}b^2 s_1 s_2.\end{aligned}$$

Since $f(x)$ and $m(x)$ are $d+1$ times continuously differentiable on $\mathcal{I}$, the term in the remainder is

$$R_\alpha(x,s) = \frac{|\alpha|}{\alpha!}\int_0^1 (1-t)^{|\alpha|-1} D^\alpha\{f(x+tbs)m(x+tbs)\}dt.$$

When $|\alpha| = d+1$, it follows that

$$\big|R_\alpha(x,s)\big| \le \frac{1}{\alpha!}\sup_{|\beta|=d+1}\sup_{w\in\mathcal{I}}\big|D^\beta\{f(w)m(w)\}\big| \le c \tag{A.1}$$

with $\beta \in \mathbb{R}^d$, because $f(x)$ and $m(x)$ are $d+1$ times continuously differentiable on $\mathcal{I}$. By Assumption (K) (ii) we have

$$\mu(x) = f(x)m(x) + b^{d+1}\int_{\mathcal{S}_b(x)} K(s,x)\sum_{|\alpha|=d+1} R_\alpha(x,s)s^\alpha ds,$$

which implies that

$$\begin{aligned}&\sup_{x\in\mathcal{I}}|\mu(x) - f(x)m(x)| \\ &\le \sup_{x\in\mathcal{I}}\Big\{b^{d+1}\int_{\mathcal{S}_b(x)}\big|K(s,x)\big|\sum_{|\alpha|=d+1}\Big(\big|R_\alpha(x,s)\big||s^\alpha|\Big)ds\Big\} \\ &\le cb^{d+1}\sup_{x\in\mathcal{I}}\Big\{\sum_{|\alpha|=d+1}\int\big|K(s,x)s^\alpha\big|ds\Big\} = O_p(b^{d+1}),\end{aligned}$$

where the second step holds true because of (A.1) and the last step because of Assumption (K) (i). Therefore, by Assumption (B), we have

$$\sup_{x\in\mathcal{I}}|\mu(x) - f(x)m(x)| = o_p(n^{-1/2}). \tag{A.2}$$

We now prove part (2). Analogously as in the derivation of Theorem 2 in Hansen (2008 [7]), where Y in the original proof is replaced by $g(X,V)$, we have

$$\sup_{x\in\mathcal{I}}\Big|\frac{1}{n}\sum_{i=1}^{n} g(X_i,V_i)K_b(X_i,x)-\mu(x)\Big| = O_p\Big(\Big(\frac{\log n}{nb^d}\Big)^{1/2}\Big) = o_p(n^{-1/4}). \quad \text{(A.3)}$$

The assumptions of that theorem are satisfied:

1. Assumptions 1 and 3 in Hansen (2008 [7]) hold true by Assumption (K)(i) and (K)(iii), respectively;
2. we have independent observations, so conditions (2), (4), (7) and (10) in [7] are not needed, and (11) in that article simplifies to $\theta = 1$;
3. condition (5) in [7] is satisfied by Assumption (X), and inspecting the proofs of Theorems 1 and 2 in [7] reveals that condition (3) and (6) in that article can be replaced by the assumption that $g(X,Y)$ is square integrable for independent data;
4. equation (12) in [7] is met by Assumption (B); equation (13) in that article is satisfied since the support $\mathcal{I}$ in (A.3) is compact.

Combining (A.2) and (A.3) gives the desired statement,

$$\sup_{x\in\mathcal{I}}\Big|\frac{1}{n}\sum_{i=1}^{n} g(X_i,V_i)K_b(X_i,x)-f(x)m(x)\Big| = o_p(n^{-1/4}).$$

□

A.2. Proof of equation (2.4) (Theorem 2.1)

We can write B_1 in the form

$$B_1 = \frac{1}{n}\sum_{i=1}^{n}(1-Z_i)\frac{\sum_{j=1}^{n} Z_jK_b(X_j,X_i)\{h(X_j,Y_j)-\chi(X_i)\}}{\sum_{j=1}^{n} Z_jK_b(X_j,X_i)}.$$

By the second conclusion of Lemma 2.1 we have

$$J_1 = \sup_{x\in\mathcal{I}}\Big|\frac{1}{n}\sum_{j=1}^{n} Z_jK_b(X_j,x)\{h(X_j,Y_j)-\chi(x)\}\Big| = o_p(n^{-1/4}),$$

$$J_2 = \sup_{x\in\mathcal{I}}\Big|\frac{1}{n}\sum_{j=1}^{n} Z_jK_b(X_j,x)-\pi(x)f(x)\Big| = o_p(n^{-1/4}),$$

and therefore

$$|B_2-B_1| \le \frac{1}{n}\sum_{i=1}^{n}(1-Z_i)\frac{|n^{-1}\sum_{j=1}^{n} Z_jK_b(X_j,X_i)\{h(X_j,Y_j)-\chi(X_i)\}|}{\pi(X_i)f(X_i)|n^{-1}\sum_{j=1}^{n} Z_jK_b(X_j,X_i)|} \times\Big|\frac{1}{n}\sum_{j=1}^{n} Z_jK_b(X_j,X_i)-\pi(X_i)f(X_i)\Big|$$

$$\le J_1 J_2 \frac{1}{n}\sum_{i=1}^{n} \frac{1-Z_i}{\pi(X_i)f(X_i)|n^{-1}\sum_{j=1}^{n} Z_j K_b(X_j,X_i)|}$$
$$\le o_p(n^{-1/2})\Big\{\inf_{x\in\mathcal{I}} \pi(x)f(x)\Big\}^{-1}\Big\{\inf_{x\in\mathcal{I}}\Big|\frac{1}{n}\sum_{j=1}^{n} Z_j K_b(X_j,x)\Big|\Big\}^{-1},$$

where the two items are both bounded away from zero by Assumption (X) and (2.2). This shows

$$|B_2 - B_1| = o_p(n^{-1/2}).$$

In the following we assume that (X_p, Y_p, Z_p) and (X_q, Y_q, Z_q) are two different observations which do not belong to the set of complete observations $\mathcal{B}$. Consider

$$\begin{aligned} B_3^2 &= E^2(B_2|\mathcal{B}) \\ &= E^2\big[(1-Z_p)\{\phi(X_p) - \widetilde{\phi}(X_p)\}|\mathcal{B}\big] \\ &= E\big[(1-Z_p)\{\phi(X_p) - \widetilde{\phi}(X_p)\}|\mathcal{B}\big]\, E\big[(1-Z_q)\{\phi(X_q) - \widetilde{\phi}(X_q)\}|\mathcal{B}\big] \\ &= E\big[(1-Z_p)\{\phi(X_p) - \widetilde{\phi}(X_p)\}(1-Z_q)\{\phi(X_q) - \widetilde{\phi}(X_q)\}|\mathcal{B}\big], \end{aligned}$$

where the last equality holds because $\phi(X_p)$ and $\phi(X_q)$ are conditionally independent given $\mathcal{B}$. Then

$$E(B_3^2) = E\big[(1-Z_1)(1-Z_2)\{\phi(X_1) - \widetilde{\phi}(X_1)\}\{\phi(X_2) - \widetilde{\phi}(X_2)\}\big].$$

This combined with

$$\begin{aligned} E(B_2^2) &= \frac{1}{n^2} E\Big(\Big[\sum_{i=1}^{n}(1-Z_i)\{\phi(X_i) - \widetilde{\phi}(X_i)\}\Big]^2\Big) \\ &= \frac{1}{n} E\big[(1-Z)\{\phi(X) - \widetilde{\phi}(X)\}^2\big] \\ &\quad + \frac{n-1}{n} E\big[(1-Z_1)(1-Z_2)\{\phi(X_1) - \widetilde{\phi}(X_1)\}\{\phi(X_2) - \widetilde{\phi}(X_2)\}\big] \end{aligned}$$

yields

$$E(B_2^2) - E(B_3^2) = n^{-1} E\big[(1-Z)\{\phi(X) - \widetilde{\phi}(X)\}^2\big] - n^{-1} E(B_3^2). \tag{A.4}$$

Further we obtain

$$\begin{aligned} &E[\{\phi(X) - \widetilde{\phi}(X)\}^2] \\ &= E\Big(\Big[\frac{1}{n}\frac{\sum_{j=1}^{n} Z_j K_b(X_j,X)\{h(X_j,Y_j) - \chi(X)\}}{\pi(X)f(X)}\Big]^2\Big) \\ &\le \frac{c}{n^2} E\Big(\Big[\sum_{j=1}^{n} Z_j K_b(X_j,X)\{h(X_j,Y_j) - \chi(X)\}\Big]^2\Big) \\ &= \frac{c}{n^2} E\Big[\sum_{j=1}^{n} Z_j K_b^2(X_j,X)\{h(X_j,Y_j) - \chi(X)\}^2\Big] \end{aligned}$$

$$+\frac{c}{n^2}E\Big[\sum_{i\neq j}^{n} Z_iZ_jK_b(X_i,X)K_b(X_j,X)\{h(X_i,Y_i)-\chi(X)\}\{h(X_j,Y_j)-\chi(X)\}\Big]$$
$$= T_1 + T_2$$

with

$$T_1 = c/n\, E[Z_1K_b^2(X_1,X)\{h(X_1,Y_1)-\chi(X)\}^2]$$
$$T_2 = c(n-1)/n \times E[Z_1Z_2K_b(X_1,X)K_b(X_2,X)\{h(X_1,Y_1)-\chi(X)\}\{h(X_2,Y_2)-\chi(X)\}].$$

For T_1 we have

$$\begin{aligned} T_1 &\le c/n\, E[\{h(X_1,Y_1)-\chi(X)\}^2] \\ &= c/n\,[E\{h^2(X_1,Y_1)\}+E\{\chi^2(X)\}-2E\{h(X_1,Y_1)\chi(X)\}] \\ &= c/n\,[E\{h^2(X_1,Y_1)\}-E^2\{h(X_1,Y_1)\}+E\{\chi^2(X)\}-E^2\{\chi(X)\}] \\ &= c/n\,\{\mathrm{Var}[h(X,Y)]+\mathrm{Var}[\chi(X)]\}\to 0 \quad (n\to\infty). \end{aligned}$$

In the third step we used

$$E\{h(X_1,Y_1)\chi(X)\} = E\{h(X_1,Y_1)\}E\{\chi(X)\} = E^2\{h(X_1,Y_1)\} = E^2\{\chi(X)\},$$

and in the last statement that the variances are finite by assumption.

The second term T_2 computes to

$$\begin{aligned} T_2 &= c(n-1)/nE(E[Z_1Z_2K_b(X_1,X)K_b(X_2,X)\{h(X_1,Y_1)-\chi(X)\} \\ &\qquad \times\{h(X_2,Y_2)-\chi(X)\}|X]) \\ &= c(n-1)/n\, E(E^2[Z_1K_b(X_1,X)\{h(X_1,Y_1)-\chi(X)\}|X]) \\ &\le c\sup_{x\in\mathcal{I}} E^2[K_b(X_1,x)\{h(X_1,Y_1)-\chi(x)\}] \\ &= c\big(\sup_{x\in\mathcal{I}}\big|E[K_b(X_1,x)\{h(X_1,Y_1)-\chi(x)\}]\big|\big)^2\to 0 \quad (n\to\infty). \end{aligned}$$

The last step follows from the first conclusion of Lemma 2.1. Hence we have

$$E[\{\phi(X)-\widetilde{\phi}(X)\}^2] = T_1+T_2\to 0 \quad (n\to\infty).$$

This combined with (A.4) yields

$$\begin{aligned} nE\{(B_2-B_3)^2\} &= n\{E(B_2^2)-2E(B_2B_3)+E(B_3^2)\} \\ &= n\{E(B_2^2)-2E\{E(B_2B_3|\mathcal{B})\}+E(B_3^2)\} \\ &= n\{E(B_2^2)-2E\{B_3E(B_2|\mathcal{B})\}+E(B_3^2)\} \\ &= n\{E(B_2^2)-E(B_3^2)\} \\ &= E[(1-Z)\{\phi(X)-\widetilde{\phi}(X)\}^2]-E(B_3^2) \\ &\le E[\{\phi(X)-\widetilde{\phi}(X)\}^2]\longrightarrow 0. \end{aligned}$$

Now use $E(B_2-B_3)=0$ and Chebyshev's inequality to obtain $n^{1/2}|B_2-B_3| = o_p(1)$. This and $n^{1/2}|B_1-B_2| = o_p(1)$ finally give

$$n^{1/2}|B_1-B_3| = o_p(1).$$

□

A.3. Proof of equation (2.9) (Theorem 2.2)

We will use similar arguments as in the first part of the proof of the theorem, in particular we write again $\varepsilon_i^* = \varepsilon_i - \dot{r}_\vartheta(X_i)^\top(\widehat{\vartheta} - \vartheta)$. Using this notation, equation (2.9) becomes

$$\begin{aligned}&\Big|n^{-1}\sum_{i=1}^n Z_i\{g(X_i) - \widehat{g}(X_i)\}\widehat{\varepsilon}_i\Big| \\ &= \Big|n^{-1}\sum_{i=1}^n Z_i\{g(X_i) - \widehat{g}(X_i)\}\{(\widehat{\varepsilon}_i - \varepsilon_i^*) - \dot{r}_\vartheta(X_i)^\top(\widehat{\vartheta} - \vartheta) + \varepsilon_i\}\Big| = o_p(n^{-1/2}).\end{aligned}$$

We treat the three parts separately. As in the proof of (2.6), we obtain for the first part

$$\begin{aligned}&\Big|n^{-1}\sum_{i=1}^n Z_i\{g(X_i) - \widehat{g}(X_i)\}(\widehat{\varepsilon}_i - \varepsilon_i^*)\Big| \\ &\qquad\le n^{-1}\sum_{i=1}^n \big|\{g(X_i) - \widehat{g}(X_i)\}(\widehat{\varepsilon}_i - \varepsilon_i^*)\big| \\ &\qquad\le n^{-1}\Big\{n\sum_{i=1}^n \{g(X_i) - \widehat{g}(X_i)\}^2(\widehat{\varepsilon}_i - \varepsilon_i^*)^2\Big\}^{1/2} \\ &\qquad= n^{-1/2}\sup_{x\in\mathcal{I}}\{g(x) - \widehat{g}(x)\}^2\Big\{\sum_{i=1}^n(\widehat{\varepsilon}_i - \varepsilon_i^*)^2\Big\}^{1/2} \\ &\qquad= o_p(n^{-1/2}).\end{aligned} \tag{A.5}$$

In the last step we use the arguments following (2.7) with $g(\cdot) \equiv 1$, and the fact that $\widehat{g}(x)$ is a consistent estimator of $g(x)$. The second part computes to

$$\begin{aligned}&\Big|n^{-1}\sum_{i=1}^n Z_i\{g(X_i) - \widehat{g}(X_i)\}\dot{r}_\vartheta(X_i)^\top(\widehat{\vartheta} - \vartheta)\Big| \\ &\qquad\le n^{-1}\|\widehat{\vartheta} - \vartheta\|\sum_{i=1}^n \big|g(X_i) - \widehat{g}(X_i)\big|\|\dot{r}_\vartheta(X_i)\| \\ &\qquad\le n^{-1}\sup_{x\in\mathcal{I}}\big|g(x) - \widehat{g}(x)\big|\|\widehat{\vartheta} - \vartheta\|\sum_{i=1}^n \|\dot{r}_\vartheta(X_i)\| \\ &\qquad= o_p(n^{-1/2})\end{aligned} \tag{A.6}$$

where the last step uses Assumptions (R) and (T) as well as the uniform consistency of $\widehat{g}(x)$.

Finally we show

$$n^{-1}\sum_{i=1}^n Z_i\{\widehat{g}(X_i) - g(X_i)\}\varepsilon_i = o_p(n^{-1/2}). \tag{A.7}$$

In equation (2.8) in the first part of proof of Theorem 2.2 we have seen that $n^{-1}\sum_{i=1}^{n} Z_i g(X_i)\varepsilon_i$ is part of the approximation and therefore has the order $O_p(n^{-1/2})$. The term on the left-hand side of (A.7) is approximately conditionally centered (given X_i). Since $\widehat{g}(x)-g(x)$ is asympotically negligible, we obtain the desired order $o_p(n^{-1/2})$.

Combining (A.5), (A.6) and (A.7) gives the desired statement

$$\Big|n^{-1}\sum_{i=1}^{n} Z_i\big\{g(X_i)-\widehat{g}(X_i)\big\}\widehat{\varepsilon}_i\Big| = o_p(n^{-1/2}).$$

To prove that the term in (A.7) is exactly conditionally centered, we propose using leave-one-out estimators $\widetilde{\rho}(X_i)$, $\widetilde{\sigma}^2(X_i)$ and $\widetilde{\pi}(X_i)$ to estimate $\widehat{g}(X_i)$ ($i=1,\dots,n$), i.e.

$$\widehat{g}(X_i) = \frac{\widetilde{\rho}(X_i)}{\widetilde{\sigma}^2(X_i)\widetilde{\pi}(X_i)}.$$

Choose, for example,

$$\widetilde{\sigma}^2(X_i) = \frac{\sum_{j=1,j\neq i}^{n} Z_j K_b(X_i-X_j)\{Y_j - r_{\widetilde{\vartheta}_i}(X_j)\}^2}{\sum_{j=1,j\neq i}^{n} Z_j K_b(X_i-X_j)},$$

where $\widetilde{\vartheta}_i$ is some consistent estimator of ϑ that does not use (X_i,Y_i) if that pair is observed. The other two leave-one-out estimator are defined similarly. Thanks to this construction $\widehat{g}(X_i)$ is independent of Y_i and Z_i conditional on X_i, and we obtain (suppressing the subscript i)

$$\begin{aligned} &E\big\{Z(\widehat{g}(X)-g(X))\varepsilon\big\} = E\big\{Z\widehat{g}(X)\varepsilon\big\} \\ &= E\Big\{E\{Z\widehat{g}(X)\varepsilon|X\}\Big\} = E\Big\{\pi(X)E\{\widehat{g}(X)|X\}E\{\varepsilon|X\}\Big\} = 0. \end{aligned}$$

□

A.4. Proof of equation (3.6) (Theorem 3.1)

To specify the tangent space V concerning the conditional distribution we introduce perturbations s and t of the two parameters $f(\cdot|x)$ and ϑ. Write $F(\cdot|x)$ for the conditional distribution function of $f(\cdot|x)$ and assume that $f(\cdot|x)$ has finite Fisher information for location, $E\ell^2(\varepsilon|x)<\infty$, where $\ell(\cdot|x) = -f'(\cdot|x)/f(\cdot|x)$ is the score function. The perturbed conditional distribution is

$$Q_{nv}(x,dy) = Q_{nsa}(x,dy) = f_{ns}\{y - r_{\vartheta_{na}(x)}|x\}dy$$

with $\vartheta_{na} = \vartheta + n^{-1/2}a$, $a\in\mathbb{R}^d$, $f_{ns}(y|x) = f(y|x)\{1+n^{-1/2}s(x,y)\}$ and $s\in S$, where

$$S = \Big\{s\in L_2(F) : \int s(x,y)f(y|x)dy = 0, \int ys(x,y)f(y|x)dy = 0\Big\}.$$

Here S is determined by two constraints: the perturbed error conditional density $f_{ns}(\cdot|x)$ must integrate to one, $\int f_{ns}(y|x)dy = 1$, and must be centered at zero, $\int y f_{ns}(y|x)dy = 0$. As in Schick (1993 [17]), Section 3, we have

$$\begin{aligned}
&f_{ns}\{y - r_{\vartheta_{na}}(x)|x\} \\
&= f\{y - r_{\vartheta_{na}}(x)|x\}[1 + n^{-1/2}s\{x, y - r_{\vartheta_{na}}(x)\}] \\
&\doteq [f\{y - r_\vartheta(x)|x\} - n^{-1/2}f'\{y - r_\vartheta(x)|x\}\dot{r}_\vartheta(x)^\top a][1 + n^{-1/2}s\{x, y - r_\vartheta(x)\}] \\
&\doteq f\{y - r_\vartheta(x)|x\}\Big(1 + n^{-1/2}\Big[s\{x, y - r_\vartheta(x)\} - \frac{f'\{y - r_\vartheta(x)|x\}}{f\{y - r_\vartheta(x)|x\}}\dot{r}_\vartheta(x)^\top a\Big]\Big) \\
&= f\{y - r_\vartheta(x)|x\}(1 + n^{-1/2}[s\{x, y - r_\vartheta(x)\} + \ell\{y - r_\vartheta(x)|x\}\dot{r}_\vartheta(x)^\top a]).
\end{aligned}$$

Therefore the subspace V of V_0 is

$$V = \big\{s\{x, y - r_\vartheta(x)\} + \ell\{y - r_\vartheta(x)|x\}\dot{r}_\vartheta(x)^\top a : s \in S, a \in \mathbb{R}^d\big\}.$$

Setting $\widetilde{V} = \{v(X, Y) : v \in V\}$ and writing $v \in \widetilde{V}$ as a sum of three terms, we obtain

$$\begin{aligned}
v(X, Y) &= s(X, \varepsilon) + \ell(\varepsilon|X)\dot{r}_\vartheta(X)^\top a \\
&= s(X, \varepsilon) + \Big\{\ell(\varepsilon|X) - \frac{\varepsilon}{\sigma^2(X)}\Big\}\dot{r}_\vartheta(X)^\top a + \frac{\varepsilon}{\sigma^2(X)}\dot{r}_\vartheta(X)^\top a.
\end{aligned}$$

The third term is obviously an element of

$$V_1 = \{\sigma^{-2}(X)\dot{r}_\vartheta(X)^\top a\varepsilon : a \in \mathbb{R}^d\}.$$

It is easy to check that the first two terms (and their sum) belong to

$$V_2 = \{t(X, Y) : t \in S\}$$

and that V_1 and V_2 are orthogonal. Hence we can write $\widetilde{V}$ as an orthogonal sum, $\widetilde{V} = V_1 \oplus V_2$. To specify v_* in the canonical gradient formula (3.2), we use this presentation and write

$$\begin{aligned}
v_*(X, Y) &= \sigma^{-2}(X)\dot{r}_\vartheta(X)^\top a_*\varepsilon + t_*(X, Y), \\
v(X, Y) &= \sigma^{-2}(X)\dot{r}_\vartheta(X)^\top a\varepsilon + t(X, Y),
\end{aligned} \tag{A.8}$$

where a_*, $a \in \mathbb{R}^d$ and $t_*, t \in S$. Setting $u = 0$ and $w = 0$ in equation (3.3), we obtain

$$\begin{aligned}
&E[Z\{\sigma^{-2}(X)\dot{r}_\vartheta(X)^\top a_*\varepsilon + t_*\}\{\sigma^{-2}(X)\dot{r}_\vartheta(X)^\top a\varepsilon + t\}] \\
&\quad = E[h(X, Y)\{\sigma^{-2}(X)\dot{r}_\vartheta(X)^\top a\varepsilon + t\}].
\end{aligned} \tag{A.9}$$

Set $t = 0$ in (A.9) and use $E\{Z\sigma^{-2}(X)\dot{r}_\vartheta(X)^\top a\varepsilon t_*\} = 0$, which holds since $t_* \in S$. Then (A.9) becomes

$$E\{Z\sigma^{-4}(X)\dot{r}_\vartheta(X)^\top a_*\dot{r}_\vartheta(X)^\top a\varepsilon^2\} = E\{h(X, Y)\sigma^{-2}(X)\dot{r}_\vartheta(X)^\top a\varepsilon\},$$

and, since the equation must be satisfied for arbitrary vectors a,

$$a_*^\top E\{Z\sigma^{-4}(X)\varepsilon^2\dot{r}_\vartheta(X)\dot{r}_\vartheta(X)^\top\} = E\{h(X,Y)\sigma^{-2}(X)\varepsilon\dot{r}_\vartheta(X)^\top\}.$$

The term on the left-hand side computes to

$$\begin{aligned} a_*^\top E\{Z\sigma^{-4}(X)\varepsilon^2\dot{r}_\vartheta(X)\dot{r}_\vartheta(X)^\top\} &= a_*^\top E[E\{Z\sigma^{-4}(X)\varepsilon^2\dot{r}_\vartheta(X)\dot{r}_\vartheta(X)^\top|X\}] \\ &= a_*^\top E\{Z\sigma^{-2}(X)\dot{r}_\vartheta(X)\dot{r}_\vartheta(X)^\top\}, \end{aligned}$$

and, assuming $E\{Z\sigma^{-2}(X)\dot{r}_\vartheta(X)\dot{r}_\vartheta(X)^\top\}$ is invertible, we obtain

$$a_* = [E\{Z\sigma^{-2}(X)\dot{r}_\vartheta(X)\dot{r}_\vartheta(X)^\top\}]^{-1}E\{h(X,Y)\sigma^{-2}(X)\varepsilon\dot{r}_\vartheta(X)\} = I^{-1}\Delta,$$

with I and Δ as in Theorem 2.2 and Corollary 2.1. Now set $a = 0$ in (A.9) and use

$$\begin{aligned} E\{Z\sigma^{-2}(X)\dot{r}_\vartheta(X)^\top a_*\varepsilon t\} &= E[E\{Z\sigma^{-2}(X)\dot{r}_\vartheta(X)^\top a_*\varepsilon t|X\}] \\ &= E\{\sigma^{-2}(X)\dot{r}_\vartheta(X)^\top a_*\pi(X)E(\varepsilon t|X)\} = 0 \end{aligned}$$

to obtain

$$E(Zt_*t) = E\{h(X,Y)t\}.$$

Writing this as an iterated expectation,

$$E\{E(Zt_*t|X)\} = E\{\pi(X)E(t_*t|X)\} = E[E\{h(X,Y)t|X\}], \tag{A.10}$$

we see that $h(X,Y)/\pi(X)$ is a candidate for $t_*(X,Y)$. Since t_* must be in S we choose a suitably modified version, namely

$$\begin{aligned} t_*(X,Y) &= \frac{1}{\pi(X)}[h(X,Y) - E\{h(X,Y)|X\} - E\{h(X,Y)\varepsilon|X\}\sigma^{-2}(X)\varepsilon] \\ &= \frac{1}{\pi(X)}\Big\{h(X,Y) - \chi(X) - \frac{\varepsilon\rho_h(X)}{\sigma^2(X)}\Big\} \end{aligned} \tag{A.11}$$

with $\rho_h(x) = E\{h(X,Y)\varepsilon|X = x\}$. Now plug a_* and t_* into the formula for v_* in (A.8) to obtain formula (3.6).

To verify (A.11) formally, we show that t_* satisfies characterization (A.10) and that t_* is in $V_2 = \{t(X,Y) : t \in S\}$. To prove the first part we consider $t = t(X,Y) \in V_2$, that is, by definition of S, $E(t|X) = 0$ and $E(t\varepsilon|X) = 0$. Then

$$\begin{aligned} &E(Zt_*t|X) \\ &= E\big(\big[h(X,Y) - E\{h(X,Y)|X\} - E\{h(X,Y)\varepsilon|X\}\sigma^{-2}(X)\varepsilon\big]t|X\big) \\ &= E\{h(X,Y)t\} - E\{h(X,Y)|X\}E(t|X) - E\{h(X,Y)\varepsilon|X\}\sigma^{-2}(X)E(t\varepsilon|X) \\ &= E\{h(X,Y)t\}, \end{aligned}$$

which shows that t_* satisfies (A.10). The second part, $t_* = t_*(X, Y) \in V_2$, follows from

$$E(t_*|X) = \frac{1}{\pi(X)}[E\{h(X,Y)|X\} - E\{h(X,Y)|X\} \\ - E\{h(X,Y)\varepsilon|X\}\sigma^{-2}(X)E(\varepsilon|X)] = 0$$

and

$$E(t_*\varepsilon|X) = \frac{1}{\pi(X)}\big[E\{h(X,Y)\varepsilon|X\} - E\{h(X,Y)|X\}E(\varepsilon|X) \\ - E\{h(X,Y)\varepsilon|X\}\sigma^{-2}(X)E(\varepsilon^2|X)\big] \\ = \frac{1}{\pi(X)}\big[E\{h(X,Y)\varepsilon|X\} - E\{h(X,Y)\varepsilon|X\}\big] = 0.$$

□

Acknowledgments

We thank Raymond Carroll for reading an early draft of this article and for his constructive and detailed remarks. We also thank three referees for their valuable suggestions and comments, which helped us improve the paper.

References

[1] Bates, D.M. and Watts, D.G. (1998). *Nonlinear Regression Analysis and Its Applications*, Wiley, New York. MR1060528

[2] Cheng, P.E. (1990). Applications of kernel regression estimation: A Survey. *Comm. Statist. Theory Methods*, **19**, 4103–4134. MR1103005

[3] Cheng, P.E. (1994). Nonparametric estimation of mean functionals with data missing at random. *J. Am. Statist. Assoc.*, **89**, 81–87.

[4] Cheng, P.E. and Wei, L.J. (1986). Nonparametric inference under ignorable missing data process and treatment assignment. International Statistical Symposium, Taipei, **1**, 97–112.

[5] Devroye, L.P. and Wagner, T.J. (1980). Distribution-free consistency results in nonparametric discrimination and regression function estimations. *Ann. Statist.*, **8**, 231–239. MR0560725

[6] Hammer, S.M., Katzenstein, D.A., Hughes, M.D., Gundacker, H., Schooley, R.T., Haubrich, R.H., Henry, W.K., Lederman, M.M., Phair, J.P., Niu, M., Hirsch, M.S. and Merigan, T.C. (1996). A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *New England J. of Med.*, **335(15)**, 1081–1090.

[7] Hansen, B.E. (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory*, **24**, 726–748. MR2409261

[8] Hirano, K., Imbens, G.W. and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, **714**, 1161–1189. MR1995826

[9] Hu, Z., Follmann, D.A., and Qin, J. (2010). Semiparametric dimension reduction estimation for mean response with missing data. *Biometrika*, **97(2)**, 305–319. MR2650740

[10] Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. Second edition. Wiley-Interscience, New York. MR1925014

[11] Matloff, N.S. (1981). Use of regression functions for improved estimation of means. *Biometrika*, **68**, 685–689. MR0637788

[12] Müller, U.U. (2009). Estimating linear functionals in nonlinear regression with responses missing at random. *Ann. Statist.*, **37**, 2245–2277. MR2543691

[13] Müller, U.U. and Van Keilegom, I. (2012). Efficient parameter estimation in regression with missing responses. *Electron. J. Stat.*, **6**, 1200–1219. MR2988444

[14] Müller, U.U. and Schick, A. (2017). Efficiency transfer for regression models with responses missing at random. *Bernoulli*, **23**, 2693–2719. MR3648042

[15] Müller, U.U., Schick, A. and Wefelmeyer, W. (2006). Imputing responses that are not missing. *Probability, Statistics and Modelling in Public Health*, Symposium in Honor of Marvin Zelen (M. Nikulin, D. Commenges and C. Huber, eds.), 350–363, Springer. MR2230741

[16] Nadaraya, E.A. (1964). On estimating regression. *Theory Probab. Appl.*, **9**, 141–142.

[17] Schick, A. (1993). On efficient estimation in regression models. *Ann. Statist.*, **21**, 1486–1521. Correction and addendum: **23** (1995), 1862–1863. MR1241276 MR1370311

[18] Seber, G.A.F. and Wild, C.J. (1989). *Nonlinear Regression*. J. Wiley & Sons. New York. MR0986070

[19] Simonoff, J.S. (1996). *Smoothing Methods in Statistics*. Springer. MR1391963

[20] Tang, M.L., Tang, N.S., Zhao, P.Y., and Zhu, H. (2018). Efficient robust estimation for linear models with missing response at random. *Scand. J. Stat.*, **45(2)**, 366–381. MR3803594

[21] Wand, M.P. and Schucany, W.R. (1990). Gaussian-based kernel. *Canad. J. Statist.*, **18**, 197–204. MR1079592

[22] Wang, Q. and Rao, J.N.K. (2001). Empirical likelihood for linear regression models under imputation for missing responses. *Canad. J. Statist.*, **29**, 597–608. MR1888507

[23] Wang, Q. and Rao, J.N.K. (2002). Empirical likelihood-based inference under imputation for missing response data. *Ann. Statist.*, **30**, 896–924. MR1922545

[24] Watson, G.S. (1964). Smooth regression analysis. *Sankhya, Ser. A*, **26**, 359–372. MR0185765