Estimation in semiparametric models with missing data

Chen, Songxi

December 2012

Estimation in semiparametric models with missing data*

Song Xi Chen
Guanghua School of Management and Center for Statistical Science
Peking University
Department of Statistics, Iowa State University

Ingrid Van Keilegom
Institut de Statistique, Biostatistique et Sciences Actuarielles
Université catholique de Louvain

October 20, 2012

Abstract

This paper considers the problem of parameter estimation in a general class of semiparametric models when observations are subject to missingness at random. The semiparametric models allow for estimating functions that are non-smooth with respect to the parameter. We propose a nonparametric imputation method for the missing values, which then leads to imputed estimating equations for the finite dimensional parameter of interest. The asymptotic normality of the parameter estimator is proved in a general setting, and is investigated in detail for a number of specific semiparametric models. Finally, we study the small sample performance of the proposed estimator via simulations.

Key words: Copulas; imputation; kernel smoothing; missing at random; nuisance function; partially linear model; semiparametric model; single index model.

*Chen acknowledges financial support from a National Science Foundation grant SES-0518904 and Center for Statistical Science, Peking University. Van Keilegom acknowledges financial support from IAP research network P6/03 of the Belgian Government (Belgian Science Policy), and from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement No. 203650. Both authors thank two referees for constructive comments and suggestions which have improved the presentation of the paper.

1 Introduction

Semiparametric models encompass a large class of statistical models. They have the advantage of being more interpretable and parsimonious than nonparametric models, and at the same time they are less restrictive than purely parametric models. Let (X, Y) and (X_i, Y_i), $i = 1, \ldots, n$, be independent and identically distributed random vectors. We denote Θ for a finite dimensional parameter set (a compact subset of $\mathbb{R}^p$) and $\mathcal{H}$ for a set of infinite dimensional functions depending on X and/or Y. The functions in $\mathcal{H}$ are allowed to depend on θ. Suppose that $g(X, Y, \theta, h)$ is an estimating function which is known up to the finite dimensional parameter $\theta \in \Theta$ and the infinite dimensional nuisance function $h \in \mathcal{H}$, and which satisfies

$$G(\theta, h) := E\{g(X, Y, \theta, h)\} = 0 \tag{1.1}$$

at $\theta = \theta_0$ and $h = h_0$, which are respectively the true parameter value and the true infinite dimensional nuisance function.

Model (1.1) includes as special cases many well known semiparametric models. For instance, by adding a nonparametric functional component $h(\cdot)$ to the classical linear regression model, we get the following partially linear model :

$$Y = \theta^T X_1 + h(X_2) + \varepsilon, \tag{1.2}$$

where $X = (X_1, X_2)$ is a covariate vector, Y is univariate response, and the error ε satisfies some identifiability constraint, like $E(\varepsilon|X) = 0$ or $\text{med}(\varepsilon|X) = 0$. Here h is a nuisance function that summarizes the nonparametric covariate effect due to a group of predictors X_2. In the context of the generalized linear model (McCullagh and Nelder, 1983), if the known link function is replaced by an unknown nonparametric link function h, we arrive at the single-index regression model $Y = h(\theta^T X) + \varepsilon$. Other semiparametric models that are special cases of model (1.1) include copula models, semiparametric transformation models, Cox models, among many others.

Missing values are commonly encountered in statistical applications. In survey sampling e.g., there are typically non-responses of respondants to some survey questions. In biological applications, part of the data vector is often incompletely collected. The presence of missing values means that the entire sample $\{(X_i, Y_i)\}_{i=1}^n$ is not available. Without loss of generality, we assume that Y_i is subject to missingness, whereas X_i is always available. Note that this convention implies that if the vector (X_i, Y_i) follows

a regression model, the vector Y_i possibly contains some covariates, and the vector X_i possibly contains the (or a) response.

There are basically two streams of inference methods for missing values. The first one is the imputation approach. The celebrated multiple imputation method of Little and Rubin (2002) is a popular representation of this approach. The idea of the second approach is based on inverse weighting by the missing propensity function proposed by James Robins and colleagues, see for instance Robins, Rotnitzky and Zhao (1994). The implementation of both approaches usually requires a parametric model for the missing propensity function or the missing at random mechanism (Rubin, 1976).

The aim of this paper is to provide a general estimator of the finite dimensional parameter θ in the presence of the nuisance function h and of missing values. To make the estimation fully respective to the underlying missing values mechanism without assuming a parametric model, we impute for each missing Y_i multiple copies from a kernel estimator of the conditional distribution of Y_i given X_i, under the assumption of missingness at random. This nonparametric imputation method can be viewed as a nonparametric counter part of the multiple imputation approach of Little and Rubin (2002). With the imputed missing values and a separate estimator for the nonparametric function h, the estimator of θ is obtained by solving an estimating equation based on (1.1). The consistency and asymptotic normality of the estimator are established under a set of mild conditions.

We end this section by mentioning some related papers on parametric and semiparametric models with missing data. Recent contributions have been made e.g. by Wang, Wang, Gutierrez and Carroll (1998), Wang, Linton and Härdle (2004), Müller, Schick and Wefelmeyer (2006), Chen, Zeng and Ibrahim (2007), Wang and Sun (2007), Liang (2008), Müller (2009), Wang (2009) and Wang, Shen, He and Wang (2010). All these contributions are however limited to specific (often quite narrow) classes of models, whereas we aim in this paper at developing a general approach, applicable not only to regression models (with missing responses and/or covariates), but also to any other semiparametric model with missing data. We also refer to Chen, Hong and Tarozzi (2008), who study semiparametric efficiency bounds and efficient estimation of parameters defined through general moment restrictions with missing data. Their method relies however on auxiliary data containing information about the distribution of the missing variables conditional on proxy variables that are observed in both the primary and the auxiliary database, when such distribution is common to the two data sets.

The paper is organized as follows. The nonparametric imputation and the estimation

framework are introduced in Section 2. Section 3 reports the general asymptotic result regarding the consistency and the asymptotic normality of the proposed estimator. The general result is illustrated and applied to a set of popular semiparametric models in Section 4. In Section 5 we study the small sample performance of the proposed estimator via simulations. All the technical details are provided in the Appendix.

2 General method

Let X be a d_x-dimensional vector that is always observable, and let Y be a d_y-dimensional vector that is subject to missingness. Define $\Delta = 1$ if Y is observed, and $\Delta = 0$ if Y is missing. We assume that Y is missing at random, i.e. Δ and Y are conditionally independent given X :

$$P(\Delta = 1|X, Y) = P(\Delta = 1|X) =: p(X).$$

Note that using Y to denote the missing vector does not mean that we work under a regression model and that Y is the response in that model. Specificially Y can represent a set of covariates in a regression problem. Hence, our framework includes the case where the covariates are subject to missingness.

In the absence of missing values, the semiparametric model is defined by an r-dimensional real valued estimation function $g(X, Y, \theta, h)$, where θ is a finite dimensional parameter taking values in a compact $\Theta \subset I\!R^p$, and h is an unknown function taking values in a functional space $\mathcal{H}$ (an infinite dimensional parameter set of functions) and is depending on X and/or Y. The functions in $\mathcal{H}$ are allowed to depend on θ too (but we will often suppress this dependence when no confusion is possible). Let θ_0 and h_0 be the true unknown finite and infinite dimensional parameters. We often omit the arguments of the function h for notational convenience, i.e. $(\theta, h) \equiv (\theta, h_\theta)$, $(\theta, h_0) \equiv (\theta, h_{0\theta})$ and $(\theta_0, h_0) \equiv (\theta_0, h_{0\theta_0})$. The estimating function g is known up to θ and h. Suppose that $r \geq p$, meaning that the number of estimating functions may be larger than the dimension of θ, so we allow for an over-identified set of equations, popular in e.g. econometrics. Moreover, by allowing the function g to be a non-smooth function of its arguments, our general model also includes e.g. quantile regression models or change-point models.

Let $G(\theta, h) = E[g(X, Y, \theta, h)]$, which is a non-random vector-valued function $G : \Theta \times \mathcal{H} \to I\!R^r$, such that $G(\theta, h_0) = 0$ for $\theta = \theta_0$. If all data (X_i, Y_i), $i = 1, \ldots, n$ were observed, the parameter θ could be estimated by minimizing a weighted norm of

$n^{-1}\sum_{i=1}^{n} g(X_i, Y_i, \theta, \widehat{h}_\theta)$, where $\widehat{h}_\theta$ is an appropriate estimator of h_θ. See Chen, Linton and Van Keilegom (2003) for more details.

The issue of interest here is the estimation of θ in the presence of missing values. Let $(X_i, \Delta_i Y_i, \Delta_i)$, $i = 1, \ldots, n$, be i.i.d. random vectors having the same distribution as $(X, \Delta Y, \Delta)$. We use a nonparametric approach to impute the missing values via a nonparametric kernel estimator of $F(y|x) = P(Y \leq y|X = x)$, the conditional distribution of Y given $X = x$. The kernel estimator of $F(y|x)$ is

$$\widehat{F}(y|x) = \sum_{j=1}^{n} \frac{\Delta_j K_a(X_j - x) I(Y_j \leq y)}{\sum_{l=1}^{n} \Delta_l K_a(X_l - x)}, \tag{2.1}$$

based on the portion of the sample without missing data, where K is a d_x-dimensional kernel function, $a = a_n$ is a bandwidth sequence and $K_a(\cdot) = K(\cdot/a_n)/a_n^{d_x}$.

For each missing Y_i, we generate κ (conditionally) independent $Y_{i1}^*, \ldots, Y_{i\kappa}^*$ from $\widehat{F}(\cdot|X_i)$ as imputed values for the missing Y_i. Define now the imputed estimating function :

$$G_n(\theta, h) = n^{-1} \sum_{i=1}^{n} \Big\{ \Delta_i g(X_i, Y_i, \theta, h) + (1 - \Delta_i) \frac{1}{\kappa} \sum_{l=1}^{\kappa} g(X_i, Y_{il}^*, \theta, h) \Big\}.$$

Note that the value of κ controls the variance of the imputed component. Theoretically speaking, we will let κ tend to infinity. Our numerical experience shows that the choice $\kappa = 50$ is sufficient when the dimension is not too large. For larger dimension, κ should be chosen larger. We note that $\frac{1}{\kappa}\sum_{l=1}^{\kappa} g(X_i, Y_{il}^*, \theta, h)$ approximates $\int g(X_i, y, \theta, h) d\hat{F}(y|X_i)$. If the integral can be computed directly, then explicit imputation can be avoided. If a parametric model $F(y|x;\theta)$ is available for the conditional distribution, where θ is a finite dimensional parameter and $\hat{\theta}$ is its maximum likelihood estimator, then we can use $F(y|x;\hat{\theta})$ instead of the nonparametric estimator $\hat{F}(y|x)$ to generate the imputed Y_i^*. We would like to emphasize here that our general model is not necessarily a regression model, and hence in general we cannot impute missing Y's by using conditional mean imputation. And even if we would have a regression structure, the conditional mean imputation approach would still not be applicable in general, since Y does not necessarily represent the response variable in that model. See also Wang and Chen (2009), where a similar imputation approach has been used in the context of parametric estimating equations with missing values.

From the imputed estimating function $G_n(\theta, h)$ and for a given estimator $\widehat{h}_\theta$ of h_θ, depending on the particular model at hand, we define the estimator of θ by :

$$\widehat{\theta} = \text{argmin}_{\theta \in \Theta} \, \|G_n(\theta, \widehat{h}_\theta)\|_W, \tag{2.2}$$

where $\|A\|_W = (tr(A^T W A))^{1/2}$ for any r-dimensional vector A and for some fixed symmetric $r \times r$ positive definite matrix W (and where tr stands for the trace of a matrix). Note that when $r = p$ (so the number of equations equals the number of parameters to be estimated) and when the function g is smooth in θ, the system of equations $G_n(\theta, \widehat{h}_\theta) = 0$ has a solution (namely $\widehat{\theta}$ defined in (2.2)) under certain regularity conditions. In other situations (e.g. in the case of quantile regression or in an overidentified case), there is no vector θ that solves this equation, in which case we have to use the (more general) definition given in (2.2).

3 Main result

Below, we state the asymptotic normality of the estimator $\widehat{\theta}$ and we also give the formula of its asymptotic variance. The conditions under which this result is valid are given in the Appendix, and they will be checked in detail in Section 4 for a number of specific semiparametric models.

Theorem 3.1 *Assume that conditions (A1)-(A5), (B1)-(B5) and (C1)-(C3) hold. Then,*

$$\widehat{\theta} - \theta_0 = n^{-1} \sum_{i=1}^{n} (\Lambda^T W \Lambda)^{-1} \Lambda^T W k(X_i, \Delta_i Y_i, \Delta_i) + o_P(n^{-1/2}),$$

and

$$n^{1/2}(\widehat{\theta} - \theta_0) \xrightarrow{d} N(0, \Omega),$$

where $\Omega = (\Lambda^T W \Lambda)^{-1} \Lambda^T W \, Var\{k(X, \Delta Y, \Delta)\} W \Lambda (\Lambda^T W \Lambda)^{-1}$,

$$k(x, \delta y, \delta) = \frac{\delta}{p(x)} g(x, y, \theta_0, h_0) + \Big(1 - \frac{\delta}{p(x)}\Big) E[g(x, Y, \theta_0, h_0) | X = x] + \xi(x, \delta y, \delta),$$

the function ξ *is defined in condition (A4), and* $\Lambda = \Lambda(\theta_0)$, *with*

$$\Lambda(\theta) = \frac{d}{d\theta} G(\theta, h_0) = \lim_{\tau \to 0} \frac{1}{\tau} \Big[G(\theta + \tau, h_{0,\theta+\tau}) - G(\theta, h_{0\theta}) \Big].$$

The proof of this result can be found in the Appendix.

Remark 3.2

(i) Instead of using an imputation approach to take care of the missing values, we could also estimate θ by minimizing

$$\Big\| n^{-1} \sum_{i=1}^{n} \frac{\Delta_i g(X_i, Y_i, \theta, \widehat{h})}{p(X_i, \widehat{\beta})} \Big\|_W$$

with respect to θ, where $p(X_i) = p(X_i, \beta)$ follows e.g. a logistic model, and β can be estimated by

$$\widehat{\beta} = \text{argmax}_\beta \sum_{i=1}^{n} \big\{\Delta_i \log p(X_i, \beta) + (1 - \Delta_i) \log(1 - p(X_i, \beta))\big\}.$$

See Robins, Rotnitzky and Zhao (1994), among many other papers, for more details on this estimation procedure based on the inverse weighting by the missing propensity function.

(ii) Based on Theorem 3.1 the efficiency of the proposed estimator can be studied, and the optimal choice of the weight matrix W can be obtained. We do not elaborate on this in this paper, and refer e.g. to Section 6 in Ai and Chen (2003) for more details.

(iii) In the above i.i.d. representation of $\widehat{\theta} - \theta_0$, the function ξ comes from the estimation of h_0 by $\widehat{h}$. Also, note that if there are no missing data, then $\delta = 1$ and $p(\cdot) \equiv 1$, so $k(x, \delta y, \delta) = g(x, y, \theta_0, h_0) + \xi(x, y, 1)$ in that case.

(iv) Note that when $r = p$ (i.e. there are as many equations as there are parameters to be estimated), then Ω reduces to

$$\Omega = \Lambda^{-1} \text{Var}\{k(X, \Delta Y, \Delta)\} (\Lambda^T)^{-1},$$

provided Λ is of full rank.

4 Examples

4.1 Partially linear regression model

The first example we consider is that of a partial linear mean regression model :

$$Y = \theta^T X_1 + h(X_2) + \varepsilon, \tag{4.1}$$

where $E(\varepsilon|X) = 0$, $X = (X_1^T, X_2)^T$ is $(d+1)$-dimensional, Y is one-dimensional and for identifiability reasons we let $E(h(X_2)) = 0$ (the linear part $\theta^T X_1$ contains an intercept). We suppose that the response Y is missing at random. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. coming from model (4.1). Define $g(x, y, \theta, h) = (y - \theta^T x_1 - h(x_2))x_1$, and let

$$\widehat{h}_\theta(x_2) = \sum_{i=1}^{n} W_{ni}(x_2, b_n) \Big\{\Delta_i [Y_i - \theta^T X_{1i}] + (1 - \Delta_i)[\widehat{m}(X_i) - \theta^T X_{1i}]\Big\},$$

where

$$W_{ni}(x_2, b_n) = \frac{L_b(X_{2i} - x_2)}{\sum_{j=1}^n L_b(X_{2j} - x_2)},$$

L is a univariate kernel function, $b = b_n \to 0$ is a bandwidth sequence, $L_b(\cdot) = L(\cdot/b)/b$, and $\widehat{m}(x) = \int y d\widehat{F}(y|x)$, with $\widehat{F}(y|x)$ given in (2.1). Finally, let $\widehat{\theta} = \text{argmin}_\theta \|G_n(\theta, \widehat{h}_\theta)\|$, where $\|\cdot\|$ is the Euclidean norm. Instead of working with the above estimator of $h_\theta(x_2)$, we could also work with a weighted average of the non-missing observations only.

We now verify conditions (A1)-(A5) and (B1)-(B5). Let $h_{0\theta}(x_2) = h_0(x_2) - (\theta - \theta_0)^T E(X_1|X_2 = x_2) = E(Y|X_2 = x_2) - \theta^T E(X_1|X_2 = x_2)$, and let $\|h\|_{\mathcal{H}} = \sup_{\theta, x_2} |h_\theta(x_2)|$ for any h. First, note that

$$\begin{aligned}\widehat{h}_\theta(x_2) - h_{0\theta}(x_2) &= \big\{\widehat{h}_\theta(x_2) - E[\widehat{h}_\theta(x_2)|\mathbf{X}]\big\} + \big\{E[\widehat{h}_\theta(x_2)|\mathbf{X}] - h_{0\theta}(x_2)\big\} \\ &= (T_1 + T_2)(x_2),\end{aligned}$$

where $\mathbf{X} = (X_1, \ldots, X_n)$. Denoting $\widetilde{Y}_i = \Delta_i Y_i + (1 - \Delta_i)\widehat{m}(X_i)$, we have that $E[\widetilde{Y}_i|\mathbf{X}] = p(X_i)m(X_i) + (1 - p(X_i))(m(X_i) + O_P(a_n^q)) = m(X_i) + o_P(n^{-1/4})$ uniformly in i, provided $na_n^{4q} \to 0$, and where $m(x) = E(Y|X = x)$ and q is the order of the kernel k. Hence,

$$\begin{aligned}T_1(x_2) &= \sum_{i=1}^n W_{ni}(x_2, b_n)\Delta_i[Y_i - m(X_i)] \\ &\quad + \sum_{i=1}^n W_{ni}(x_2, b_n)(1 - \Delta_i)[\widehat{m}(X_i) - m(X_i)] + o_P(n^{-1/4}) \\ &= O_P((nb_n)^{-1/2}(\log n)^{1/2}) + O_P((na_n^{d+1})^{-1/2}(\log n)^{1/2}) + o_P(n^{-1/4}) \\ &= o_P(n^{-1/4}),\end{aligned}$$

provided $nb_n^2(\log n)^{-2} \to \infty$ and $na_n^{2(d+1)}(\log n)^{-2} \to \infty$. Next, consider

$$\begin{aligned}T_2(x_2) &= \sum_{i=1}^n W_{ni}(x_2, b_n)[m(X_i) - \theta^T X_{1i}] - h_{0\theta}(x_2) + o_P(n^{-1/4}) \\ &= \sum_{i=1}^n W_{ni}(x_2, b_n)[E(Y|X_i) - \theta^T X_{1i} - E(Y|X_{2i}) + \theta^T E(X_1|X_{2i})] \\ &\quad + O_P(b_n^2) + o_P(n^{-1/4}) \\ &= O_P((nb_n)^{-1/2}(\log n)^{1/2}) + o_P(n^{-1/4}) = o_P(n^{-1/4}),\end{aligned}$$

uniformly in θ, provided $nb_n^8 \to 0$. Hence, (A1) is verified. Next, for (A2), define $\mathcal{H} = C_M^1(R_{X_2})$, where the space $C_M^1(R_{X_2})$ is defined in Remark A.1, and where R_{X_2} is the

(compact) support of X_2. Using a similar derivation as for verifying condition (A1), we can show that $\sup_{\|\theta-\theta_0\|=o(1)} \sup_{x_2} |\widehat{h}'_\theta(x_2) - h'_{0\theta}(x_2)| = o_P(1)$, provided $nb_n^3(\log n)^{-1} \to \infty$, $na_n^{d+1}b_n^2(\log n)^{-1} \to \infty$ and $a_n^q b_n^{-1} \to 0$. It then follows that $P(\widehat{h}_\theta \in C_M^1(R_{X_2})) \to 1$. Moreover, the second part of (A2) is valid for $s = 1$ by Remark A.1. For condition (A3) note that for any θ and h, $\Gamma(\theta, h_0)[h - h_0] = -E\{(h_\theta(X_2) - h_{0\theta}(X_2))X_1\}$. Hence,

$$\begin{aligned}
&\Big\|\Gamma(\theta, h_0)[\widehat{h} - h_0] - \Gamma(\theta_0, h_0)[\widehat{h} - h_0]\Big\| \\
&= \Big\|(\theta - \theta_0)^T E\Big\{\Big[\sum_{i=1}^n W_{ni}(X_2, b_n)X_{1i} - E(X_1|X_2)\Big]X_1\Big\}\Big\| = o_P(\|\theta - \theta_0\|).
\end{aligned}$$

Next, consider (A4). Using the above derivations, write

$$\begin{aligned}
&\widehat{h}_{\theta_0}(x_2) - h_0(x_2) \\
&= \sum_{i=1}^n W_{ni}(x_2, b_n)\Big[\Delta_i\{Y_i - m(X_i)\} + (1 - \Delta_i)\{\widehat{m}(X_i) - m(X_i)\}\Big] + o_P(n^{-1/2}),
\end{aligned}$$

provided $na_n^{2q} \to 0$ and $nb_n^4 \to 0$. Replacing $W_{ni}(x_2, b_n)$ by $(nb_n)^{-1}K(\frac{X_{2i}-x_2}{b_n})/f_{X_2}(x_2)$, and noting that

$$b_n^{-1}E\Big\{\frac{X_1}{f_{X_2}(X_2)}K\Big(\frac{X_{2i} - X_2}{b_n}\Big)\Big\} = E(X_1|X_2 = X_{2i}) + O(b_n^2),$$

uniformly in i, we obtain

$$\begin{aligned}
&\Gamma(\theta_0, h_0)[\widehat{h} - h_0] = -E\{(\widehat{h}_{\theta_0}(X_2) - h_0(X_2))X_1\} \\
&= -n^{-1}\sum_{i=1}^n E(X_1|X_{2i})\Big[\Delta_i\{Y_i - m(X_i)\} + (1 - \Delta_i)\{\widehat{m}(X_i) - m(X_i)\}\Big] + o_P(n^{-1/2}).
\end{aligned}$$

It can be easily seen that

$$\begin{aligned}
&n^{-1}\sum_{i=1}^n E(X_1|X_{2i})(1 - \Delta_i)\{\widehat{m}(X_i) - m(X_i)\} \\
&= n^{-1}\sum_{i=1}^n \frac{1 - p(X_i)}{p(X_i)}E(X_1|X_{2i})\Delta_i\{Y_i - m(X_i)\} + o_P(n^{-1/2}),
\end{aligned}$$

and hence

$$\xi(X_i, \Delta Y_i, \Delta_i) = -E(X_1|X_{2i})\frac{\Delta_i}{p(X_i)}\{Y_i - m(X_i)\}.$$

For (A5) consider

$$G_n(\theta,h)-\widetilde{G}_n(\theta,h)=n^{-1}\sum_{i=1}^{n}(1-\Delta_i)\Big[\frac{1}{\kappa}\sum_{l=1}^{\kappa}Y_{il}^{*}-E(Y|X=X_i)\Big]X_{1i}$$
$$=n^{-1}\sum_{i=1}^{n}(1-\Delta_i)\Big(\widehat{m}(X_i)-m(X_i)\Big)X_{1i}+o_P(1)=o_P(1),$$

as κ and n tend to infinity. Since this does not depend on θ, the second part of (A5) is obvious.

We now turn to the B-conditions. Condition (B1) is an identifiability condition, whereas (B2) holds true provided $E|X_{1j}|<\infty$ for $j=1,\ldots,d$. Next, for (B3) it is easily seen that

$$\Lambda=-E\Big[(X_1-E(X_1|X_2))^TX_1\Big]=-\mathrm{Var}[X_1|X_2],$$

which we assume to be of full rank. Condition (B4) is automatically fulfilled since $G(\theta,h)$ is linear in h. Finally, condition (B5) holds true for $s=1$.

It now follows from Theorem 3.1 that $n^{1/2}(\widehat{\theta}-\theta_0)$ is asymptotically normally distributed with asymptotic variance given by

$$\Omega=\Lambda^{-1}\mathrm{Var}\Big[\frac{\Delta}{p(X)}\{Y-\theta_0^TX_1-h_0(X_2)\}\{X_1-E(X_1|X_2)\}\Big](\Lambda^T)^{-1},$$

since $E[g(x,Y,\theta_0,h_0)|X=x]=0$. Note that if all data would be observed, the matrix Ω equals the variance-covariance matrix given in Robinson (1988), who considers the special case where ε is independent of X.

4.2 Single index regression model

We now consider a single index regression model :

$$Y=h(\theta^TX)+\varepsilon, \tag{4.2}$$

where X is d-dimensional, Y is one-dimensional, $E(\varepsilon|X)=0$ and $\theta\in\Theta=\{\theta\in I\!R^d:\|\theta\|=1\}$ for identifiability reasons. See Powell, Stock and Stoker (1989), Ichimura and Lee (1991), Ichimura (1993), Härdle, Hall and Ichimura (1993), among many others for important results on the estimation and inference for this model when all the data are completely observed. We assume here that the response Y is missing at random. Suppose that the density of θ_0^TX is bounded away from zero on its support (which is supposed

to be compact). Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a sample of i.i.d. data drawn from model (4.2). Define $g(x, y, \theta, h, h') = (y - h(\theta^T x))h'(\theta^T x)x$, and let

$$\widehat{h}_\theta(u) = \sum_{j=1}^{n} W_{nj}(u) Y_j$$

be a kernel estimator of $h_{0\theta}(u) = E(Y|\theta^T X = u)$, where

$$W_{nj}(u) = \frac{\Delta_j L_b(\theta^T X_j - u)}{\sum_{\ell=1}^{n} \Delta_\ell L_b(\theta^T X_\ell - u)},$$

L is a kernel function and b is an appropriate bandwidth. Finally, let $\widehat{\theta}$ be a solution of the equation $G_n(\theta, \widehat{h}_\theta, \widehat{h}'_\theta) = 0$.

We restrict attention here to the calculation of the matrix Λ and the function ξ, which determine the asymptotic variance of $\widehat{\theta}$. The verification of conditions (A1)-(A5) and (B1)-(B5) can be done by adapting the arguments used in the previous example to the context of single index models. Details are omitted. Note that

$$\Lambda = \frac{d}{d\theta} E\Big[(Y - h_{0\theta}(\theta^T X))h'_{0\theta}(\theta^T X)X\Big]\Big|_{\theta=\theta_0}.$$

It can be easily seen that Λ can also be written as

$$\Lambda = -E\Big[h_0'^2(\theta_0^T X)\{X - \mu_X(\theta_0^T X)\}\{X - \mu_X(\theta_0^T X)\}^T\Big],$$

where $\mu_X(u) = E(X|\theta_0^T X = u)$. Also note that

$$\begin{aligned}\Gamma(\theta_0, h_0)[h - h_0, h' - h'_0] = &-E\Big[\{h_{\theta_0}(\theta_0^T X) - h_0(\theta_0^T X)\}h'_0(\theta_0^T X)X\Big] \\ &+E\Big[\{Y - h_0(\theta_0^T X)\}\{h'_{\theta_0}(\theta_0^T X) - h'_0(\theta_0^T X)\}X\Big].\end{aligned}$$

Since the latter term equals zero, we have that

$$\begin{aligned}&\Gamma(\theta_0, h_0)[\widehat{h} - h_0, \widehat{h}' - h'_0] \\ &= -n^{-1}\sum_{i=1}^{n} \frac{\Delta_i}{p(X_i)}\{Y_i - h_0(\theta_0^T X_i)\}h'_0(\theta_0^T X_i)E(X|\theta_0^T X = \theta_0^T X_i) + o_P(n^{-1/2}).\end{aligned}$$

It now follows that the asymptotic variance of $\widehat{\theta}$ equals

$$\Omega = \Lambda^{-1}\mathrm{Var}\Big[\frac{\Delta}{p(X)}\{Y - h_0(\theta_0^T X)\}h'_0(\theta_0^T X)\{X - E(X|\theta_0^T X)\}\Big](\Lambda^T)^{-1}.$$

It is easily seen that if all data would be observed and the model is homoscedastic, the matrix Ω reduces to the asymptotic variance formula given in Härdle, Hall and Ichimura (1993) (see their section 2.5).

4.3 Copula model

In the third example we consider a copula model for two random variables X and Y, with X being always observed and Y being missing at random. We suppose that the copula belongs to a parametric family $\{C_\theta : \theta \in \Theta\}$ where Θ is a subset of $\mathbb{R}^p$. Hence, for any x, y,

$$F(x,y) = P(X \le x, Y \le y) = C_\theta(F_X(x), F_Y(y)),$$

where $F_X(x) = P(X \le x)$ and $F_Y(y) = P(Y \le y)$ are the marginals of X and Y, which are completely unspecified and will be estimated nonparametrically. We assume that F_X and F_Y are continuous, so that θ_0 is unique. Let (X_i, Y_i), $i = 1, \ldots, n$ be i.i.d. with common distribution F. Define $\widehat{F}_X(x) = (n+1)^{-1} \sum_{i=1}^n I(X_i \le x)$ and

$$\widehat{F}_Y(y) = (n+1)^{-1} \sum_{i=1}^n \Big\{ \Delta_i I(Y_i \le y) + (1 - \Delta_i) \widehat{F}(y|X_i) \Big\},$$

where $\widehat{F}(y|x)$ is defined in (2.1). Note that in a second step the estimator $\widehat{F}_Y(y)$ could be improved, by replacing $\widehat{F}(y|x)$ by $\widehat{F}_\theta(y|x) = C^2_\theta(\widehat{F}_X(x), \widehat{F}_Y(y))$, where C^2_θ denotes the derivative of C_θ with respect to the second argument (similar notations are used to denote higher order derivatives). In what follows we restrict attention to the estimator defined in the first step. Now, define

$$g(x, y, \theta, F_X, F_Y) = \frac{C^{12'}_\theta(F_X(x), F_Y(y))}{C^{12}_\theta(F_X(x), F_Y(y))}$$

(the prime in $C^{12'}_\theta$ stands for the derivative with respect to θ), i.e. $g(x, y, \theta, F_X, F_Y)$ is the derivative of the log-likelihood function under the copula model, and define $\widehat{\theta} = \text{argmin}_{\theta \in \Theta} \| G_n(\theta, \widehat{F}_X, \widehat{F}_Y) \|$.

We now calculate the function ξ defined in condition (A4), from which the formula of the asymptotic variance of $\widehat{\theta}$ can be easily obtained. We omit the verification of the conditions of Theorem 3.1, but more details can be obtained from the authors. Let $d_\theta(u, v) = C^{12'}_\theta(u, v)/C^{12}_\theta(u, v)$. Then, straightforward calculations show that

$$\begin{aligned}\Gamma(\theta, F_X, F_Y)[\widehat{F}_X - F_X, \widehat{F}_Y - F_Y] = E\Big[& \frac{\partial}{\partial u} d_\theta(u, F_Y(Y))|_{u=F_X(X)} (\widehat{F}_X(X) - F_X(X)) \\ & + \frac{\partial}{\partial v} d_\theta(F_X(X), v)|_{v=F_Y(Y)} (\widehat{F}_Y(Y) - F_Y(Y)) \Big].\end{aligned}$$

Next, note that

$$\widehat{F}_Y(y) - F_Y(y) = n^{-1}\sum_{i=1}^{n}\Big[\Delta_i I(Y_i \le y) + (1-\Delta_i)F(y|X_i) - F_Y(y)\Big]$$
$$+n^{-1}\sum_{i=1}^{n}(1-\Delta_i)\Big[\widehat{F}(y|X_i) - F(y|X_i)\Big].$$

For the second term above it can be shown that

$$n^{-1}\sum_{i=1}^{n}(1-\Delta_i)E\Big[\frac{\partial}{\partial v}d_\theta(F_X(X),v)|_{v=F_Y(Y)}(\widehat{F}(Y|X_i) - F(Y|X_i))\Big]$$
$$= n^{-1}\sum_{i=1}^{n}\frac{1-p(X_i)}{p(X_i)}E\Big[\frac{\partial}{\partial v}d_\theta(F_X(X),v)|_{v=F_Y(Y)}(I(Y_i \le Y) - F(Y|X_i))\Big] + o_P(n^{-1/2}).$$

It now follows that

$$\Gamma(\theta_0, F_X, F_Y)[\widehat{F}_X - F_X, \widehat{F}_Y - F_Y]$$
$$= n^{-1}\sum_{i=1}^{n}E\Big[\frac{\partial}{\partial u}d_{\theta_0}(u, F_Y(Y))|_{u=F_X(X)}\Big\{I(X_i \le X) - F_X(X)\Big\}$$
$$+\frac{\partial}{\partial v}d_{\theta_0}(F_X(X),v)|_{v=F_Y(Y)}\Big\{\Delta_i I(Y_i \le Y) + (1-\Delta_i)F(Y|X_i) - F_Y(Y)$$
$$+\frac{1-p(X_i)}{p(X_i)}(I(Y_i \le Y) - F(Y|X_i))\Big\}\Big] + o_P(n^{-1/2}).$$

The formula of the asymptotic variance can now be easily obtained from Theorem 3.1.

5 Simulation study

In this section we present simulation results for a single index mean regression model. Consider

$$Y = h(\theta^T X) + \varepsilon,$$

where $X = (X_1, X_2, X_3)^T$, $X_j \sim$ Unif$[0,1]$ $(j = 1,2,3)$, (X_1, X_2, X_3) are mutually independent, $\varepsilon \sim N(0, 0.5^2)$, $\theta \in \Theta = \{\theta \in I\!R^3 : \|\theta\| = 1\}$, and $h(u) = \exp(u)$. The probability that the response Y is missing depends on the covariate X via a logistic model :

$$P(\Delta = 1|X,Y) = \frac{\exp(\beta^T X)}{1+\exp(\beta^T X)}.$$

Note that Θ can be regarded as compact by reparametrizing the unit circle via a polar transformation and using the angles of the polar transformation as the new parameter to

replace θ, i.e. we write $\theta(\alpha) = (\sin(\alpha_1)\sin(\alpha_2), \sin(\alpha_1)\cos(\alpha_2), \cos(\alpha_1))^T$ for some α_1, α_2. Note that $\|\theta\| = 1$ for any value of α_1 and α_2. In the simulations, we work with $\alpha_1 = \pi/3$ and $\alpha_2 = \pi/6$, and we take $\beta = (1, 1, 0)^T$ in Table 1, and $\beta = (0.5, 0.5, 0)^T$ in Table 2.

The estimating function for α_1 and α_2 is given by

$$g(x, y, \alpha, h)$$
$$= \Big\{y - h(\theta(\alpha)^T x)\Big\} h'(\theta(\alpha)^T x) \begin{pmatrix} \cos(\alpha_1)\sin(\alpha_2)x_1 + \cos(\alpha_1)\cos(\alpha_2)x_2 - \sin(\alpha_1)x_3 \\ \sin(\alpha_1)\cos(\alpha_2)x_1 - \sin(\alpha_1)\sin(\alpha_2)x_2 \end{pmatrix}.$$

Next, define $\widehat{h}(u) = \sum_{j=1}^n W_{nj}(u) Y_j$, where

$$W_{nj}(u) = \frac{\Delta_j K_b(u - \theta(\alpha)^T X_j)}{\sum_{i=1}^n \Delta_i K_b(u - \theta(\alpha)^T X_i)}.$$

The imputed estimating equation is

$$G_n(\alpha, h) = n^{-1} \sum_{i=1}^n \Big[\Delta_i g(X_i, Y_i, \alpha, h) + (1 - \Delta_i) \frac{1}{\kappa} \sum_{l=1}^{\kappa} g(X_i, Y_{il}^*, \alpha, h)\Big]. \qquad (5.1)$$

We now define $\widehat{\alpha}$ as the solution of the equation $G_n(\alpha, \widehat{h}) = 0$ with respect to all vectors $\alpha \in [0, 2\pi]^2$. Throughout the simulation k was set to be 50.

Two bandwidths need to be selected. The first one is the bandwidth a in the kernel estimator of the conditional distribution function $F(y|x)$ defined in (2.1), which is used for the imputation procedure. The second one is in the bandwidth b in the kernel estimator of the function h. Because the cross-validation method for selecting bandwidths is complicated when imputation is involved, and because it can only produce one combination of bandwidths, we prefer to use a sequence of bandwidths for a and b to study the influence of the bandwidths on the estimator. In our simulation, we considered for a and b all values between 0.02 and 0.30 with step size 0.02. We so obtain 225 combinations of the bandwidths used for estimating α. We report in Tables 1 and 2 the results corresponding to the 9 pairs of bandwidths a and b that lead to the smallest MSEs.

Table 1 summarizes the bias, standard deviation and MSE of the estimators of α_1 and α_2 based on the selected a and b. The MSE is the sum of the MSEs of $\widehat{\alpha}_1$ and $\widehat{\alpha}_2$. Each entry is based on 500 simulations. The table shows that the influence of the bandwidth is rather small. We also notice that the bias is reasonably close to 0 and as the sample size increases, the standard deviation and MSE of $\widehat{\alpha}_1$ and $\widehat{\alpha}_2$ decrease. In Table 2 the percentage of missing values is 38% compared to 27% in Table 1. As can be expected, the standard deviation and the MSE of $\widehat{\alpha}_1$ and $\widehat{\alpha}_2$ are larger compared to Table 1.

Appendix

Below we list the assumptions that are needed for the main result in Section 3. The following notations are needed. We equip the space $\mathcal{H}$ with a semi-norm $\|\cdot\|_{\mathcal{H}}$, defined by $\|h\|_{\mathcal{H}} = \sup_{\theta\in\Theta} \|h_\theta\|_S$ for any $h \in \mathcal{H}$, i.e. $\|\cdot\|_{\mathcal{H}}$ is a sup-norm with respect to the θ-argument and a semi-norm $\|\cdot\|_S$ with respect to all the other arguments. Also, $N(\lambda, \mathcal{H}, \|\cdot\|_{\mathcal{H}})$ is the covering number of the class $\mathcal{H}$ with respect to the norm $\|\cdot\|_{\mathcal{H}}$, i.e. the minimal number of balls of $\|\cdot\|_{\mathcal{H}}$-radius λ needed to cover $\mathcal{H}$. We use the notation $\Lambda(\theta) = \frac{d}{d\theta}G(\theta, h_0)$ to denote the complete derivative of $G(\theta, h_0)$ with respect to θ, i.e.

$$\Lambda(\theta) = \frac{d}{d\theta}G(\theta, h_0) = \lim_{\tau\to 0} \frac{1}{\tau}\Big[G(\theta+\tau, h_{0,\theta+\tau}) - G(\theta, h_{0\theta})\Big], \tag{A.1}$$

and the notation $\Gamma(\theta, h_0)[h - h_0]$ to denote the functional derivative of $G(\theta, h_0)$ in the direction $[h - h_0]$, i.e.

$$\Gamma(\theta, h_0)[h - h_0] = \lim_{\tau\to 0} \frac{1}{\tau}\Big[G(\theta, h_0 + \tau(h - h_0)) - G(\theta, h_0)\Big].$$

We also need to introduce the function

$$\widetilde{G}_n(\theta, h) = n^{-1}\sum_{i=1}^{n}[\Delta_i g(X_i, Y_i, \theta, h) + (1-\Delta_i)E\{g(X_i, Y, \theta, h)|X = X_i\}].$$

Finally, $\|\cdot\|$ denotes the Euclidean norm.

Assumptions on the estimator $\widehat{h}$

(A1) $\sup_{\theta\in\Theta} \|\widehat{h}_\theta - h_{0\theta}\|_S = o_P(1)$, and $\sup_{\|\theta-\theta_0\|\le\delta_n} \|\widehat{h}_\theta - h_{0\theta}\|_S = o_P(n^{-1/4})$, where $\delta_n \to 0$, and where $h_{0\theta}$ is such that $h_{0\theta_0} \equiv h_0$.

(A2) $P(\widehat{h}_\theta \in \mathcal{H}) \to 1$ as n tends to infinity, uniformly over all θ with $\|\theta - \theta_0\| = o(1)$, and $\int_0^\infty \sqrt{\log N(\lambda^{1/s}, \mathcal{H}, \|\cdot\|_{\mathcal{H}})}d\lambda < \infty$, where $0 < s \le 1$ is defined in condition (B5).

(A3) For θ in a neighborhood of θ_0, $\|\Gamma(\theta, h_0)[\widehat{h} - h_0] - \Gamma(\theta_0, h_0)[\widehat{h} - h_0]\|_W = o_P(\|\theta - \theta_0\|)$.

(A4) $\Gamma(\theta_0, h_0)[\widehat{h} - h_0] = n^{-1}\sum_{i=1}^n \xi(X_i, \Delta_i Y_i, \Delta_i) + o_P(n^{-1/2})$, where the function $\xi = (\xi_1, \ldots, \xi_r)$ satisfies $E[\xi_j(X, \Delta Y, \Delta)] = 0$ and $E[\xi_j^2(X, \Delta Y, \Delta)] < \infty$ for $j = 1, \ldots, r$.

(A5) $\sup_{\theta\in\Theta} \|G_n(\theta, \widehat{h}) - \widetilde{G}_n(\theta, \widehat{h})\|_W = o_P(1)$, and for any $\delta_n \to 0$,

$$\sup_{\|\theta-\theta_0\|\le\delta_n} \|G_n(\theta, \widehat{h}) - \widetilde{G}_n(\theta, \widehat{h}) - G_n(\theta_0, h_0) + \widetilde{G}_n(\theta_0, h_0)\|_W = o_P(n^{-1/2}).$$

Assumptions on the function G

(B1) For all $\delta > 0$, there exists $\epsilon > 0$ such that $\inf_{\|\theta-\theta_0\|>\delta} \|G(\theta, h_0)\|_W \geq \epsilon > 0$.

(B2) Uniformly for all $\theta \in \Theta$, $G(\theta, h)$ is continuous in h with respect to the $\|\cdot\|_{\mathcal{H}}$-norm at $h = h_0$.

(B3) The matrix $\Lambda(\theta)$ exists for θ in a neighborhood of θ_0, and is continuous at $\theta = \theta_0$. Moreover, $\Lambda \equiv \Lambda(\theta_0)$ is of full rank.

(B4) For all θ in a neighborhood of θ_0, $\Gamma(\theta, h_0)[h - h_0]$ exists in all directions $[h - h_0]$, and for some $0 < c < \infty$, $\|G(\theta, h) - G(\theta, h_0) - \Gamma(\theta, h_0)[h - h_0]\|_W \leq c\|h - h_0\|^2_{\mathcal{H}}$.

(B5) For each $j = 1, \ldots, r$,

$$E\Big[\sup_{(\theta',h'):\|\theta'-\theta\|\leq\delta,\|h'-h\|_{\mathcal{H}}\leq\delta} |g_j(X, Y, \theta', h') - g_j(X, Y, \theta, h)|^2\Big] \leq K\delta^{2s},$$

for all $(\theta, h) \in \Theta \times \mathcal{H}$ and $\delta > 0$, and for some constants $0 < K < \infty$ and $0 < s \leq 1$.

Regularity assumptions

(C1) The kernel K satisfies $K(u_1, \ldots, u_{d_x}) = \prod_{j=1}^{d_x} k(u_j)$, where k is a qth-order ($q \geq 2$) univariate probability density function supported on $[-1, 1]$, and k is bounded, symmetric and Lipschitz continuous. The bandwidth a_n satisfies $na_n^{d_x} \to \infty$ and $na_n^{2q} \to 0$. Moreover, $\kappa \to \infty$.

(C2) For all $x_0 \in \mathcal{X}$ and $\theta \in \Theta$, the function $x \to E[g^\ell(x_0, Y, \theta, h_0)|X = x]$ is uniformly continuous in x for $\ell = 1$ and 2, and $E[g^2(X, Y, \theta, h_0)] < \infty$.

(C3) For all $x_0 \in \mathcal{X}$ and $\theta \in \Theta$, the functions $x \to p(x)$, $f_X(x)$, $m_g(x_0, x, \theta, h_0)$ and $m_{g^2}(x_0, x, \theta, h_0)$ are q times continuously differentiable with respect to the components of x on the interior of their support. Here, f_X is the probability density function of X and $m_{g^\ell}(x_0, x, \theta, h) = E[g^l(x_0, Y, \theta, h)|X = x]$. Moreover, $\inf_{x\in\mathcal{X}} p(x) > 0$.

Remark A.1 Suppose the functions in $\mathcal{H}$ have a compact support R of dimension $d \leq d_x + d_y$. In order to check condition (A2), define for any vector $a = (a_1, \ldots, a_d)$ of d integers, the differential operator

$$D^a = \frac{\partial^{|a|}}{\partial u_1^{a_1} \ldots \partial u_d^{a_d}},$$

where $|a| = \sum_{i=1}^{d} a_i$. For any smooth function $h : R \to I\!R$ and some $\alpha > 0$, let $\underline{\alpha}$ be the largest integer smaller than α, and

$$\|h\|_{\infty,\alpha} = \max_{|a|\leq\underline{\alpha}} \sup_u |D^a h(u)| + \max_{|a|=\underline{\alpha}} \sup_{u\neq u'} \frac{|D^a h(u) - D^a h(u')|}{\|u-u'\|^{\alpha-\underline{\alpha}}}.$$

Further, let $C_M^\alpha(R)$ be the set of all continuous functions $h : R \to I\!R$ with $\|h\|_{\infty,\alpha} \leq M$. If $\mathcal{H} \subset C_M^\alpha(R)$ with $\|\cdot\|_{\mathcal{H}} = \|\cdot\|_\infty$, then $\log N(\lambda, \mathcal{H}, \|\cdot\|_\infty) \leq K(M/\lambda)^{d/\alpha}$, where K is a constant depending only on d, α and the Lebesgue measure of the domain R. Hence,

$$\int_0^\infty \sqrt{\log N(\lambda^{1/s}, \mathcal{H}, \|\cdot\|_\infty)} d\lambda < \infty \quad \text{if} \quad \alpha > \frac{d}{2s}$$

(see Theorem 2.7.1 in Van der Vaart and Wellner (1996)).

Proof of Theorem 3.1. The proof is based on Theorems 1 and 2 in Chen, Linton and Van Keilegom (2003) (CLV hereafter). In these theorems high-level conditions are given under which the estimator $\widehat{\theta}$ is, respectively, weakly consistent and asymptotically normal. We start with verifying the conditions of Theorem 1 in CLV. Condition (1.1) holds by definition of $\widehat{\theta}$, while the second, third and fourth condition are guaranteed by assumptions (B1), (B2) and (A1). Finally, condition (1.5) can be treated in a similar way as condition (2.5) of Theorem 2 of CLV, which we check below. Next, we verify conditions (2.1)–(2.6) of Theorem 2 in CLV. Condition (2.1) is, as for condition (1.1), valid by construction of the estimator $\widehat{\theta}$. For (2.2) and the first part of (2.3), use assumptions (B3) and (B4). For the second part of (2.3), it follows from the proof in CLV that it suffices to assume (A3). Next, for condition (2.4), we use assumptions (A1) and (A2). For (2.5), note that

$$\begin{aligned}
&\|G_n(\theta,h) - G(\theta,h) - G_n(\theta_0,h_0) + G(\theta_0,h_0)\|_W \\
&\leq \|G_n(\theta,h) - \widetilde{G}_n(\theta,h) - G_n(\theta_0,h_0) + \widetilde{G}_n(\theta_0,h_0)\|_W \\
&\quad + \|\widetilde{G}_n(\theta,h) - G(\theta,h) - \widetilde{G}_n(\theta_0,h_0) + G(\theta_0,h_0)\|_W.
\end{aligned}$$

For the first term above, note that it follows from the proof of Theorem 2 in CLV that it suffices to take $h = \widehat{h}$. Hence, assumption (A5) gives the required rate. For the second term we verify the conditions of Theorem 3 in CLV. For condition (3.2) note that for $j = 1, \ldots, r$ and for fixed $(\theta, h) \in \Theta \times \mathcal{H}$,

$$\begin{aligned}
&E\Big[\sup{}^*\big|\Delta g_j(X,Y,\theta',h') - \Delta g_j(X,Y,\theta,h) - (1-\Delta)E\{g_j(X,Y,\theta',h')|X\} \\
&\qquad + (1-\Delta)E\{g_j(X,Y,\theta,h)|X\}\big|\Big] \\
&\leq 2E\Big[\sup{}^*\big|g_j(X,Y,\theta',h') - g_j(X,Y,\theta,h)\big|\Big],
\end{aligned}$$

where sup* is the supremum over all $\|\theta' - \theta\| \leq \delta_n$ and $\|h' - h\|_{\mathcal{H}} \leq \delta_n$, with $\delta_n \to 0$. Hence, condition (3.2) follows from assumption (B5), whereas condition (3.3) is given in (A2). Finally, condition (2.6) is valid by combining assumption (A4), Proposition A.1 and the central limit theorem. □

Proposition A.1 *Assume that conditions (A1)-(A5), (B1)-(B5) and (C1)-(C3) hold. Then,*

$$G_n(\theta_0, h_0) = n^{-1} \sum_{i=1}^{n} \Big\{ \frac{\Delta_i}{p(X_i)} g(X_i, Y_i, \theta_0, h_0) + \Big(1 - \frac{\Delta_i}{p(X_i)}\Big) E[g(X_i, Y, \theta_0, h_0)|X = X_i] \Big\} + o_P(n^{-1/2}).$$

Proof. First we consider

$$\begin{aligned} & G_n(\theta_0, h_0) - \widetilde{G}_n(\theta_0, h_0) \\ &= n^{-1} \sum_{i=1}^{n} (1 - \Delta_i) \Big\{ \frac{1}{\kappa} \sum_{l=1}^{\kappa} g(X_i, Y_{il}^*, \theta_0, h_0) - \widetilde{m}_{ga}(X_i, \theta_0, h_0) \Big\} \\ &\quad + n^{-1} \sum_{i=1}^{n} (1 - \Delta_i) \Big\{ \widetilde{m}_{ga}(X_i, \theta_0, h_0) - m_g(X_i, \theta_0, h_0) \Big\} \\ &:= V_{n1}(\theta_0, h_0) + V_{n2}(\theta_0, h_0), \end{aligned}$$

where $m_g(x, \theta_0, h_0) = E[g(x, Y, \theta_0, h_0)|X = x]$ and

$$\widetilde{m}_{ga}(x, \theta_0, h_0) = \sum_{j=1}^{n} \frac{\Delta_j K_a(X_j - x)}{\sum_{l=1}^{n} \Delta_l K_a(X_l - x)} g(x, Y_j, \theta_0, h_0)$$

is the conditional mean imputation based on the kernel estimator of the conditional distribution. We will first show that $V_{n1}(\theta_0, h_0) = o_P(n^{-1/2})$, which means that we can just substitute $\kappa^{-1} \sum_{l=1}^{\kappa} g(X_i, Y_{il}^*, \theta_0, h_0)$ by the conditional mean imputation $\widetilde{m}_{ga}(X_i, \theta_0, h_0)$, which would simplify the theoretical analysis. However, the proposed imputation is attractive in practical implementations as it separates the imputation and analysis steps, as proposed by Little and Rubin (2002). Let $\chi_{nc} = \{(X_j, Y_j, \Delta_j = 1) : j = 1, \ldots, n\}$ be the complete part of the sample with no missing values. Given X_i with $\Delta_i = 0$, write $\widehat{m}_{g\kappa}(X_i, \theta_0, h_0) = \kappa^{-1} \sum_{l=1}^{\kappa} g(X_i, Y_{il}^*, \theta_0, h_0)$. From the way we impute Y_{il}^*,

$$E\{\widehat{m}_{g\kappa}(X_i, \theta_0, h_0)|\chi_{nc}, X_i, \Delta_i = 0\} = \widetilde{m}_{ga}(X_i, \theta_0, h_0). \tag{A.2}$$

Hence, $E\{V_{n1}(\theta_0, h_0)\} = 0$. We next calculate the variance of $V_{n1}(\theta_0, h_0)$. Note that

$$\mathrm{Var}\{V_{n1}(\theta_0, h_0)\} = n^{-2}\sum_{i,j}^{n}\mathrm{Cov}\Big\{(1-\Delta_i)[\widehat{m}_{g\kappa}(X_i,\theta_0,h_0) - \widetilde{m}_{ga}(X_i,\theta_0,h_0)],$$
$$(1-\Delta_j)[\widehat{m}_{g\kappa}(X_j,\theta_0,h_0) - \widetilde{m}_{ga}(X_j,\theta_0,h_0)]\Big\}.$$

If $i \neq j$, then conditioning on χ_{nc}, $(X_i, \Delta_i = 0)$ and $(X_j, \Delta_j = 0)$,

$$\mathrm{Cov}\Big[\{\widehat{m}_{g\kappa}(X_i,\theta_0,h_0), \widehat{m}_{g\kappa}(X_j,\theta_0,h_0)\}|\chi_{nc}, (X_i,\Delta_i=0), (X_j,\Delta_j=0)\Big] = 0.$$

This together with (A.2) implies that

$$\begin{aligned}
&\mathrm{Var}\{V_{n1}(\theta_0,h_0)\} \qquad\qquad (A.3)\\
&= n^{-2}\sum_{i=1}^{n}\mathrm{Var}\Big\{(1-\Delta_i)[\widehat{m}_{g\kappa}(X_i,\theta_0,h_0) - \widetilde{m}_{ga}(X_i,\theta_0,h_0)]\Big\}\\
&= n^{-1}E\Big[\mathrm{Var}\Big\{\widehat{m}_{g\kappa}(X_i,\theta_0,h_0) - \widetilde{m}_{ga}(X_i,\theta_0,h_0)|\chi_{nc},(X_i,\Delta_i=0)\Big\}\Big]\\
&= (n\kappa)^{-1}E\Big[\frac{\sum_{l=1}^{n}\Delta_l K_a(X_l-X_i)(gg^T)(X_i,Y_l,\theta_0,h_0)}{\sum_{l=1}^{n}\Delta_l K_a(X_l-X_i)} - (\widetilde{m}_{ga}\widetilde{m}_{ga}^T)(X_i,\theta_0,h_0)\Big]\\
&= O\{(n\kappa)^{-1}\}, \qquad\qquad (A.4)
\end{aligned}$$

provided (C2) and (C3) hold true. Hence, $V_{n1}(\theta_0, h_0) = o_P(n^{-1/2})$. Next, we consider the term $V_{n2}(\theta_0, h_0)$. Defining $w_j(x, a) = \Delta_j K_a(X_j - x)/\{\sum_{l=1}^{n}\Delta_l K_a(X_l - x)\}$, we have,

$$\begin{aligned}
V_{n2}(\theta_0,h_0) &= n^{-1}\sum_{i=1}^{n}(1-\Delta_i)\sum_{k=1}^{n}w_k(X_i,a)\Big\{g(X_i,Y_k,\theta_0,h_0) - m_g(X_i,\theta_0,h_0)\\
&\qquad -g(X_k,Y_k,\theta_0,h_0) + m_g(X_k,\theta_0,h_0)\Big\}\\
&\quad + n^{-1}\sum_{i=1}^{n}(1-\Delta_i)\sum_{k=1}^{n}w_k(X_i,a)\Big\{g(X_k,Y_k,\theta_0,h_0) - m_g(X_k,\theta_0,h_0)\Big\}\\
&= V_{n21}(\theta_0,h_0) + V_{n22}(\theta_0,h_0).
\end{aligned}$$

First, we will show that $V_{n21}(\theta_0, h_0) = o_P(n^{-1/2})$. Note that $E[V_{n21}(\theta_0, h_0)] = 0$ and using the notation

$$\gamma_{nj}(x_1,x_2) = (1-p(x_1))(1-p(x_2))\mathrm{Cov}[g(x_1,Y_j,\theta_0,h_0) - g(X_j,Y_j,\theta_0,h_0),$$
$$g(x_2,Y_j,\theta_0,h_0) - g(X_j,Y_j,\theta_0,h_0)\,|X_j],$$

we have,

$$\begin{aligned}\mathrm{Var}[V_{n21}(\theta_0,h_0)] &= E\Big\{\mathrm{Var}[V_{n21}(\theta_0,h_0)|X_1,\ldots,X_n]\Big\}\\ &= n^{-2}\sum_{i,j,k=1}^{n} E\Big\{w_j(X_i,a)w_j(X_k,a)\gamma_{nj}(X_i,X_k)\Big\}\\ &= n^{-2}\sum_{i,j,k=1}^{n} E\Big\{w_j(X_i,a)w_j(X_k,a)\gamma_{nj}(X_j,X_j)\Big\} + O(n^{-1}a_n^q)\\ &= O(n^{-1}a_n^q),\end{aligned}$$

since $\gamma_{nj}(X_j,X_j)=0$. Hence, $V_{n21}(\theta_0,h_0)=o_P(n^{-1/2})$. Next, note that

$$V_{n22}(\theta_0,h_0) = n^{-1}\sum_{i=1}^{n}\Delta_i\{g(X_i,Y_i,\theta_0,h_0)-m_g(X_i,\theta_0,h_0)\}\frac{1-p(X_i)}{p(X_i)}+o_P(n^{-1/2}),$$

where the rate of the remainder term can be shown in a similar way as for $V_{n21}(\theta_0,h_0)$. This shows that

$$\begin{aligned}&G_n(\theta_0,h_0)\\ &= \widetilde{G}_n(\theta_0,h_0)+\Big[G_n(\theta_0,h_0)-\widetilde{G}_n(\theta_0,h_0)\Big]\\ &= n^{-1}\sum_{j=1}^{n}[\Delta_j g(X_j,Y_j,\theta_0,h_0)+(1-\Delta_j)m_g(X_j,\theta_0,h_0)]\\ &\quad +n^{-1}\sum_{j=1}^{n}\Delta_j[g(X_j,Y_j,\theta_0,h_0)-m_g(X_j,\theta_0,h_0)]\frac{1-p(X_j)}{p(X_j)}+o_P(n^{-1/2})\\ &= n^{-1}\sum_{j=1}^{n}\Big[\frac{\Delta_j}{p(X_j)}g(X_j,Y_j,\theta_0,h_0)+\Big(1-\frac{\Delta_j}{p(X_j)}\Big)m_g(X_j,\theta_0,h_0)\Big]+o_P(n^{-1/2}).\end{aligned}$$

□

References

Ai, C. and Chen, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, **71**, 1795-1843.

Chen, X., Hong, H. and Tarozzi, A. (2008). Semiparametric efficiency in GMM models with auxiliary data. *Ann. Statist.*, **36**, 808-843.

Chen, X., Linton, O.B. and Van Keilegom, I. (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica*, **71**, 1591-1608.

Chen, Q., Zeng, D. and Ibrahim, J.G. (2007). Sieve maximum likelihood estimation for regression models with covariates missing at random. *J. Amer. Statist. Assoc.*, **102**, 1309-1317.

Härdle, W., Hall, P. and Ichimura, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.*, **21**, 157-178.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *J. Econometrics*, **58**, 71–120.

Ichimura, H. and Lee, L.-F. (1991). Semiparametric least squares estimation of multiple index models: single equation estimation. In: Barnett, W.A., Powell, J. and Tauchen, G. (Eds.), *Nonparametric and Semiparametric Methods in Statistics and Econometrics*. Cambridge University Press, Cambridge. (Chapter 1).

Liang, H. (2008). Generalized partially linear models with missing covariates. *J. Multiv. Anal.*, **99**, 880-895.

Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, Wiley.

McCullagh, P. and Nelder, J.A. (1983). *Generalized Linear Models*. Chapman & Hall, London.

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

Müller, U.U., Schick, A. and Wefelmeyer, W. (2006). Imputing responses that are not missing. *Probability, Statistics and Modelling in Public Health* (M. Nikulin, D. Commenges and C. Huber, eds.), 350-363, Springer.

Powell, J.L., Stock, J.M. and Stoker, T.M. (1989). Semiparametric estimation of index coefficients. *Econometrica*, **57**, 1403–1430.

Robins, J.M., Rotnitzky, A. and Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.*, **89**, 846-866.

Robinson, P.M. (1988). Root-N-consistent semiparametric regression. *Econometrica*, **56**, 931-954.

Rubin, D.B. (1976). Inference and missing values (with discussion). *Biometrika*, **63**, 481-592.

Van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag, New York.

Wang, C.Y., Wang, S., Gutierrez, R.G. and Carroll, R.J. (1998). Local linear regression

for generalized linear models with missing data. *Ann. Statist.*, **26**, 1028-1050.

Wang, D. and Chen, S.X. (2009). Empirical likelihood for estimating equations with missing values. *Ann. Statist.*, **37**, 490-517.

Wang, Q.-H. (2009). Statistical estimation in partial linear models with covariate data missing at random. *Ann. Inst. Statist. Math.*, **61**, 47-84.

Wang, Q., Linton, O. and Härdle, W. (2004). Semiparametric regression analysis with missing response at random. *J. Amer. Statist. Assoc.*, **99**, 334-345.

Wang, Q. and Sun, Z. (2007). Estimation in partially linear models with missing responses at random. *J. Multiv. Anal.*, **98**, 1470-1493.

Wang, Y., Shen, J., He, S. and Wang, Q. (2010). Estimation of single index model with missing response at random. *J. Statist. Plann. Inference*, **140**, 1671-1690.

		Bandwidths		$\widehat{\alpha}_1$		$\widehat{\alpha}_2$		
n	κ	a	b	Bias	Std	Bias	Std	MSE
50	50	0.14	0.30	-0.0161	0.114	0.0318	0.137	0.0331
		0.14	0.28	-0.0203	0.119	0.0234	0.142	0.0351
		0.12	0.22	-0.0131	0.123	0.0131	0.141	0.0355
		0.18	0.30	-0.0205	0.111	0.0255	0.149	0.0356
		0.14	0.20	-0.0078	0.128	0.0171	0.140	0.0363
		0.08	0.26	-0.0001	0.126	0.0121	0.144	0.0366
		0.12	0.26	-0.0004	0.124	0.0180	0.145	0.0367
		0.02	0.28	-0.0167	0.133	0.0202	0.136	0.0369
		0.16	0.24	-0.0103	0.127	0.0088	0.146	0.0375
100	50	0.14	0.24	-0.0135	0.083	0.0192	0.091	0.0156
		0.12	0.26	-0.0080	0.081	0.0206	0.093	0.0158
		0.14	0.28	-0.0068	0.078	0.0284	0.094	0.0159
		0.12	0.16	-0.0062	0.082	0.0108	0.096	0.0160
		0.16	0.30	-0.0116	0.078	0.0193	0.097	0.0161
		0.08	0.24	-0.0171	0.081	0.0210	0.095	0.0164
		0.10	0.20	-0.0069	0.084	0.0069	0.098	0.0167
		0.10	0.30	-0.0238	0.081	0.0347	0.093	0.0169
		0.14	0.20	-0.0109	0.083	0.0151	0.099	0.0170

Table 1: *Bias, standard deviation and MSE of* $\widehat{\alpha}_1$ *and* $\widehat{\alpha}_2$ *when* $\beta = (1, 1, 0)^T$ *(leading to 27% of missing values).*

		Bandwidths		$\widehat{\alpha}_1$		$\widehat{\alpha}_2$		
n	κ	a	b	Bias	Std	Bias	Std	MSE
50	50	0.02	0.22	-0.0125	0.147	0.0120	0.150	0.0442
		0.18	0.30	-0.0076	0.132	0.0175	0.164	0.0447
		0.12	0.26	-0.0009	0.144	0.0138	0.160	0.0463
		0.14	0.22	-0.0224	0.142	0.0123	0.161	0.0469
		0.18	0.26	-0.0078	0.147	0.0178	0.162	0.0481
		0.02	0.26	-0.0199	0.151	0.0267	0.157	0.0486
		0.14	0.28	-0.0069	0.140	0.0173	0.170	0.0486
		0.12	0.24	-0.0084	0.144	0.0102	0.167	0.0487
		0.18	0.28	-0.0034	0.146	0.0159	0.166	0.0490
100	50	0.12	0.30	-0.0166	0.092	0.0225	0.103	0.0199
		0.14	0.18	-0.0068	0.092	0.0119	0.107	0.0202
		0.10	0.28	-0.0092	0.091	0.0119	0.110	0.0207
		0.20	0.30	-0.0046	0.089	0.0304	0.108	0.0207
		0.14	0.24	-0.0079	0.094	0.0125	0.110	0.0211
		0.14	0.28	-0.0111	0.093	0.0276	0.108	0.0211
		0.12	0.16	-0.0028	0.096	0.0126	0.109	0.0211
		0.14	0.22	-0.0097	0.094	0.0136	0.112	0.0216
		0.20	0.22	-0.0038	0.094	0.0122	0.112	0.0216

Table 2: *Bias, standard deviation and MSE of* $\widehat{\alpha}_1$ *and* $\widehat{\alpha}_2$ *when* $\beta = (0.5, 0.5, 0)^T$ *(leading to 38% of missing values).*