Statistical Theory and Related Fields

ISSN: 2475-4269 (Print) 2475-4277 (Online) Journal homepage: https://www.tandfonline.com/loi/tstf20

Pseudo likelihood and dimension reduction for data with nonignorable nonresponse

Ji Chen, Bingying Xie & Jun Shao

To cite this article: Ji Chen, Bingying Xie & Jun Shao (2018) Pseudo likelihood and dimension reduction for data with nonignorable nonresponse, Statistical Theory and Related Fields, 2:2, 196-205, DOI: 10.1080/24754269.2018.1516101

To link to this article: https://doi.org/10.1080/24754269.2018.1516101

Published online: 01 Sep 2018.

STATISTICAL THEORY AND RELATED FIELDS
2018, VOL. 2, NO. 2, 196–205
https://doi.org/10.1080/24754269.2018.1516101

Pseudo likelihood and dimension reduction for data with nonignorable nonresponse

Ji Chen[a], Bingying Xie[b] and Jun Shao[a,b]

[a]School of Statistics, East China Normal University, Shanghai, China; [b]Department of Statistics, University of Wisconsin-Madison, Madison, WI, United states

ABSTRACT

Tang et al. (2003. Analysis of multivariate missing data with nonignorable nonresponse. *Biometrika*, *90*(4), 747–764) and Zhao & Shao (2015. Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data. *Journal of the American Statistical Association*, *110*(512), 1577–1590) proposed a pseudo likelihood approach to estimate unknown parameters in a parametric density of a response Y conditioned on a vector of covariate X, where Y is subjected to nonignorable nonersponse, X is always observed, and the propensity of whether or not Y is observed conditioned on Y and X is completely unspecified. To identify parameters, Zhao & Shao (2015. Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data. *Journal of the American Statistical Association*, *110*(512), 1577–1590) assumed that X can be decomposed into U and Z, where Z can be excluded from the propensity but is related with Y even conditioned on U. The pseudo likelihood involves the estimation of the joint density of U and Z. When this density is estimated nonparametrically, in this paper we apply sufficient dimension reduction to reduce the dimension of U for efficient estimation. Consistency and asymptotic normality of the proposed estimators are established. Simulation results are presented to study the finite sample performance of the proposed estimators.

ARTICLE HISTORY
Received 23 April 2018
Accepted 22 August 2018

KEYWORDS
Dimension reduction; kernel estimation; nonignorable nonresponse; nonresponse instrument; pseudo likelihood

1. Introduction

Missing data or nonresponse is common in various statistical applications such as sample surveys and biomedical studies. Let Y be a univariate response variable subject to nonresponse, X be a vector of covariates that is always observed, and R be the indicator of whether Y is observed. When the propensity $P(R = 1 \mid Y, X) = P(R = 1 \mid X)$, missing data are ignorable and there is a rich literature on methodology of handling nonresponse (Little & Rubin, 2014). In many applications, however, Y cannot be excluded from the propensity $P(R = 1 \mid Y, X)$ and missing data are nonignorable. With nonignorable nonresponse, it is challenging to estimate unknown characteristics in the conditional distribution of Y given X or the unconditional distribution of Y.

Throughout we use $p(\cdot \mid \cdot)$ or $p(\cdot)$ as a generic notation for the conditional or unconditional probability density with respect to an appropriate measure. When nonresponse is nonignorable and both $p(Y \mid X)$ and $P(R = 1 \mid Y, X)$ are nonparametric, $p(Y \mid X)$ is not identificable (Robins & Ritov, 1997). When both $p(Y \mid X)$ and $P(R = 1 \mid Y, X)$ have parametric forms, maximum likelihood methods have been developed (Baker & Laird, 1988; Greenlees, Reece, & Zieschang, 1982). Since parametric methods are sensitive to model violations, efforts have been made under semiparametric models. Qin, Leung, and Shao (2002) and Wang, Shao, and Kim (2014) imposed a parametric model on $P(R = 1 \mid Y, X)$ but allowed $p(Y \mid X)$ to be nonparametric. Assuming $P(R = 1 \mid Y, X) = P(R = 1 \mid Y)$, i.e., the entire covariate vector X can be excluded from the propensity, Tang, Little, and Raghunathan (2003) proposed a pseudo likelihood method in which $P(R = 1 \mid Y)$ is nonparametric and $p(Y \mid X)$ is parametric:

$$p(Y \mid X) = p(Y \mid X; \theta), \tag{1}$$

where θ is an unknown parameter vector and $p(y \mid x; \theta)$ is a conditional density which is known when θ is known. Zhao and Shao (2015) extended the pseudo likelihood method to the case where part of X can be excluded from the propensity, i.e.,

$$P(R = 1 \mid Y, X) = P(R = 1 \mid Y, U), \tag{2}$$

where $X = (U, Z)$ and the covariate Z, termed as instrumental variable, cannot be excluded from $p(Y \mid X; \theta)$ in (1). Here is a brief description of what has been done under (1)–(2). Under (1)–(2) and the Bayes formula,

$$\begin{aligned} p(Z \mid Y, U, R = 1) &= p(Z \mid Y, U) \\ &= \frac{p(Y \mid U, Z; \theta) p(U, Z)}{\int p(Y \mid U, z; \theta) p(U, z)\, \mathrm{d}z}. \end{aligned} \tag{3}$$

Let (y_i, x_i, r_i), $i = 1, \ldots, n$, be n independent and identically distributed observations from (Y, X, R). Based

CONTACT Ji Chen jaechen3@126.com School of Statistics, East China Normal University, 500 Dongchuan Road, 200241 Shanghai, China

on (3), if $p(U,Z)$ is known, then we can estimate θ by maximising the following likelihood:

$$\prod_{i:r_i=1} \frac{p(y_i \mid u_i, z_i; \theta) p(u_i, z_i)}{\int p(y_i \mid u_i, z; \theta) p(u_i, z)\, \mathrm{d}z} \propto \prod_{i:r_i=1} \frac{p(y_i \mid u_i, z_i; \theta)}{\int p(y_i \mid u_i, z; \theta) p(u_i, z)\, \mathrm{d}z}. \tag{4}$$

Usually $p(U,Z)$ is unknown. Substituting $p(u_i, z_i)$ in (4) by its estimate results in a pseudo likelihood. For example, Zhao and Shao (2015) assumed a parametric model $p(U \mid Z; \eta)$ for $p(U \mid Z)$ and replaced $\int p(y_i \mid u_i, z; \theta) p(u_i, z)\, \mathrm{d}z$ in (4) by $\int p(y_i \mid u_i, z; \theta)\, p(u_i \mid z; \widehat{\eta})\, \mathrm{d}\widehat{F}(z)$, where $\widehat{\eta}$ is an estimator of η based on (u_i, z_i), $i = 1, \ldots, n$, and $\widehat{F}$ is the empirical distribution of z_i, $i = 1, \ldots, n$. To avoid model misspecification on $p(U \mid Z)$, Zhao and Shao (2015) also suggested a nonparametric kernel estimator $\widehat{p}(u, z)$ to replace $p(u, z)$ in (4).

However, kernel estimation is unstable when the dimension of $X = (U, Z)$ is not small. The purpose of our work is to propose an alternative way to handle $\int p(y_i \mid u_i, z; \theta) p(u_i, z)\, \mathrm{d}z$ in (4), which adopts dimension reduction techniques to improve the resulting pseudo likelihood estimator of θ. Our main idea is described in the next section, along with the proposed pseudo likelihood and estimator of θ. Under typical conditions for kernel estimation, our proposed pseudo likelihood estimator is asymptotically normal with convergence rate $n^{-1/2}$. We also perform some simulations to examine the finite sample properties of our proposed estimator.

2. Methodology and theory

As pointed out in the previous section, the main contribution of this paper is to estimate $\int p(Y \mid U, z; \theta)\, p(z, U)\, \mathrm{d}z$ in the denominator of (4) in a more efficient way, especially when the covariate U is of high dimension. Often times, the dimension of the instrument variable Z is small, while the covariate U contains a lot of variables (demographic variables for instance), and its dimension p is not small. A straightforward idea is to split $p(U,Z)$ as $p(U,Z) = p(Z \mid U) p(U)$, instead of $p(U,Z) = p(U \mid Z) p(Z)$ as in Zhao and Shao (2015). Since $p(U)$ does not involve θ, the likelihood in (4) is equivalent to

$$\prod_{i:r_i=1} \frac{p(y_i \mid u_i, z_i; \theta)}{\int p(y_i \mid u_i, z; \theta) p(z \mid u_i)\, \mathrm{d}z}, \tag{5}$$

and our main task is to estimate the denominator in (5), i.e., the integral

$$\delta(y, u; \theta) = \int p(y \mid u, z; \theta) p(z \mid u)\, \mathrm{d}z = E\left\{p(y \mid u, Z; \theta) \mid U = u\right\} \tag{6}$$

for arbitrarily given θ, y and u. Since the real form of $p(Z \mid U)$ is not our main concern, for robustness, we adopt a nonparametric kernel regression, called Nadaraya-Watson (NW) estimation (Nadaraya, 1964; Watson, 1964), to estimate the conditional expectation in (6). The way we split $p(U,Z)$, along with the nonparametric estimation, actually frees ourselves from parameterising or modelling the relationship between Z and U.

Before building our estimators, we introduce a generic notation K_h for a kernel with an appropriate dimension and bandwidth h, i.e., K_h appeared in different places may be different. In what follows K_h is chosen to be a product kernel of dimension s and order $m \geq 2$ in the sense that $K_h(x) = h^{-s} \prod_{j=1}^{s} \kappa(x_j/h)$, where x_j is the jth component of the s-dimensional x and $\kappa(\cdot)$ is a bounded and Lipschitz continuous univariate kernel having a compact support and satisfying $\int \kappa(t)\, \mathrm{d}t = 1$, $\int t^m \kappa(t)\, \mathrm{d}t$ is finite and nonzero, and $\int t^l \kappa(t)\, \mathrm{d}t = 0$ for all $0 < l < m$. The NW estimator of $\delta(y, u; \theta)$ in (6) is

$$\widetilde{\delta}(y, u; \theta) = \frac{\sum_{j=1}^{n} K_h(u_j - u) p(y \mid u, z_j; \theta)}{\sum_{j=1}^{n} K_h(u_j - u)}. \tag{7}$$

Substituting this estimator into (5) leads to a maximum pseudo likelihood estimator of θ,

$$\widetilde{\theta} = \arg\max_{\theta \in \Theta} \prod_{i:r_i=1} \frac{p(y_i \mid u_i, z_i; \theta)}{\widetilde{\delta}(y_i, u_i; \theta)}. \tag{8}$$

Although p, the dimension of U, does not exert a direct influence on the convergence rate of $\widetilde{\theta}$ as shown in Theorem 2.1, it is well known in the literature that kernel estimators do not perform well when p is very large. For example, their convergence requires a very large sample size n.

Fortunately, there is a well-developed nonparametric method called Sufficient Dimension Reduction (SDR) (Cook & Weisberg, 1991; Li & Wang, 2007; Li, 1991; Ma & Zhu, 2012; Xia, Tong, Li, & Zhu, 2002), which helps to reduce the dimension of predictors by finding a $p \times d$ matrix B with the smallest possible $d \leq p$ such that $Z \perp\!\!\!\perp U \mid B^T U$, meaning that Z is only related to U via $B^T U$, where B^T is the transpose of B. It is common that $d < p$ and we are able to improve the estimation in (7) and (8) by applying NW estimation directly to Z and $\widehat{B}^T U$ with $\widehat{B}$ being an SDR estimator of B. Starting from sliced inverse regression (SIR) (Li, 1991), research has been done in the literature to develop SDR estimators of B, including sliced average variance estimation (SAVE) (Cook & Weisberg, 1991), directional regression (DR) (Li & Wang, 2007), (conditional) minimum average variance estimation (MAVE) (Xia et al., 2002), semiparametric approach to dimension reduction (Ma & Zhu, 2012) and etc. We adopt SIR method developed by Li (1991) to estimate B in our simulation studies in

Section 3 because it is easy to implement and works out well in practice.

When $Z \perp\!\!\!\perp U \mid B^T U$, $\delta(y,u;\theta)$ in (6) is equal to $E\{p(y \mid u, Z;\theta) \mid B^T U = B^T u\}$, which can be estimated by a new kernel estimator

$$\widehat{\delta}(y,\widehat{B}^T u;\theta) = \frac{\sum_{j=1}^n K_h(\widehat{B}^T u_j - \widehat{B}^T u) p(y \mid u, z_j;\theta)}{\sum_{j=1}^n K_h(B^T u_j - \widehat{B}^T u)}.$$

Then, a new maximum pseudo likelihood estimator of θ,

$$\widehat{\theta} = \arg\max_{\theta\in\Theta} \prod_{i:r_i=1} \frac{p(y_i \mid u_i, z_i;\theta)}{\widehat{\delta}(y_i,\widehat{B}^T u_i;\theta)}. \tag{9}$$

In the rest of this section we establish asymptotic consistency and normality of our proposed estimator $\widetilde{\theta}$ in (8) and $\widehat{\theta}$ in (9). We first introduce some notation. Let $w_i = (y_i, z_i, u_i, r_i)$ and $\nabla_\theta(\cdot)$ be the first derivative with respect to θ. For the estimator without SDR, denote

$$\gamma(y,u;\theta) = \begin{pmatrix} \gamma_1(y,u;\theta) \\ \gamma_2(y,u;\theta) \end{pmatrix} = \begin{pmatrix} E\left\{p(y \mid Z,u;\theta) \mid U = u\right\} p(u) \\ E\left\{\nabla_\theta p(y \mid Z,u;\theta) \mid U = u\right\} p(u) \end{pmatrix}, \tag{10}$$

and the kernel estimator of γ is denoted by

$$\widetilde{\gamma}(y,u;\theta) = \begin{pmatrix} \widetilde{\gamma}_1(y,u;\theta) \\ \widetilde{\gamma}_2(y,u;\theta) \end{pmatrix} = \begin{pmatrix} n^{-1}\sum_{j=1}^n p(y \mid z_j,u;\theta) K_h(u-u_j) \\ n^{-1}\sum_{j=1}^n \nabla_\theta p(y \mid z_j,u;\theta) K_h(u-u_j) \end{pmatrix}. \tag{11}$$

Therefore, maximising the pseudo likelihood in (8) is the same as maximising the pseudo log-likelihood

$$\begin{aligned} l(\theta,\widetilde{\gamma}_1) &= \frac{1}{n}\sum_{i=1}^n H(w_i;\theta,\widetilde{\gamma}_1), \\ H(w_i;\theta,\widetilde{\gamma}_1) &= r_i\Big\{\log(p(y_i \mid z_i,u_i;\theta)) - \log(\widetilde{\gamma}_1(y_i,u_i;\theta))\Big\}. \end{aligned} \tag{12}$$

Differentiating $l(\theta,\widetilde{\gamma}_1)$ with respect to θ, we obtain the score function

$$\begin{aligned} S(\theta,\widetilde{\gamma}) &= \sum_{i=1}^n g(w_i;\theta,\widetilde{\gamma}), \\ g(w_i;\theta,\widetilde{\gamma}) &= r_i\left\{\frac{\nabla_\theta p(y_i \mid z_i,u_i;\theta)}{p(y_i \mid z_i,u_i;\theta)} - \frac{\widetilde{\gamma}_2(y_i,u_i;\theta)}{\widetilde{\gamma}_1(y_i,u_i;\theta)}\right\}. \end{aligned} \tag{13}$$

Thus, $\widetilde{\theta}$ can also be obtained by solving $S(\theta,\widetilde{\gamma}) = 0$.

We use the same notation γ as in (10) for the estimator with SDR,

$$\gamma(y,u;\theta,B) = \begin{pmatrix} \gamma_1(y,u;\theta,B) \\ \gamma_2(y,u;\theta,B) \end{pmatrix} = \begin{pmatrix} E\left\{p(y \mid Z,u;\theta) \mid B^T U = B^T u\right\} p(B^T u) \\ E\left\{\nabla_\theta p(y \mid Z,u;\theta) \mid B^T U = B^T u\right\} p(B^T u) \end{pmatrix}. \tag{14}$$

Its kernel estimator is

$$\widehat{\gamma}(y,u;\theta,\widehat{B}) = \begin{pmatrix} \widehat{\gamma}_1(y,u;\theta,\widehat{B}) \\ \widehat{\gamma}_2(y,u;\theta,\widehat{B}) \end{pmatrix} = \begin{pmatrix} n^{-1}\sum_{j=1}^n p(y \mid z_j,u;\theta) K_h(\widehat{B}^T u - \widehat{B}^T u_j) \\ n^{-1}\sum_{j=1}^n \nabla_\theta p(y \mid z_j,u;\theta) K_h(\widehat{B}^T u - \widehat{B}^T u_j) \end{pmatrix}. \tag{15}$$

The pseudo log-likelihood and score function of the $\widehat{\theta}$ can be obtained by replacing $\widetilde{\gamma}$ in (12) and (13) by $\widehat{\gamma}$.

As the structure of two etimators are alike, we only provide the sufficient conditions for $\widetilde{\theta}$ in Assumptions 2.1–2.2 where γ is defined as in (10). For $\widehat{\theta}$, however, Assumptions 2.1–2.2 need to be modified by replacing (10) by (14), (11) by (15) and p by d. In addition, in order for $\widehat{\theta}$ to converge with rate $n^{-1/2}$, $\widehat{B}$ should also possess some asymptotic properties as indicated in Assumption 2.3 with γ defined in (14). In fact, SDR estimator $\widehat{B}$ and kernel estimators $\widetilde{\gamma}$ and $\widehat{\gamma}$ all have good asymptotic performances (Bierens, 1987; Li, 1991; Nadaraya, 1964; Watson, 1964).

Throughout, θ_0 denotes the true but unknown value of θ.

Assumption 2.1: *There exists a constant $c > 0$ such that $\inf_{y,u} p(y,u) \geq c$. The function $\gamma(y,u,\theta_0)$ has bounded mth derivative with respect to u. There exists a $q > 1$ such that $n^{1-1/q}h^p/\log n \to \infty$ as $n \to \infty$ and $E\{(p^2(y \mid Z,u;\theta_0) + \left\|\nabla_\theta p(y \mid Z,u)\right\|^2)^q \mid U = u\}$ is bounded.*

Let $||\cdot||_\infty$ be the sup-norm. Under Assumption 2.1, $\|\widetilde{\gamma} - E(\widetilde{\gamma})\|_\infty = O_p((\log n/nh^p)^{1/2})$ and $\|E(\widetilde{\gamma}) - \gamma\|_\infty = O_p(h^m)$, which means that the estimator $\widetilde{\gamma}$ converges uniformly to γ at a certain rate, i.e., $\|\widetilde{\gamma} - \gamma\|_\infty = O_p((\log n/nh^p)^{1/2} + h^m)$ (Hansen, 2008; Newey & McFadden, 1994).

Assumption 2.2: (i) *There exists $\epsilon_1 > 0$ such that*

$$\lim_{n\to\infty} \sup_{\theta, \|\bar{\gamma}_1 - \gamma_1\|_\infty < \epsilon_1} \left| \frac{1}{n}\sum_{i=1}^n H(w_i;\theta,\bar{\gamma}_1) - E\{H(w_i;\theta,\bar{\gamma}_1)\} \right| = 0,$$

where $H(w_i;\theta,\gamma_1)$ is defined in (12).

(ii) $E\{\|\nabla_\theta p(Y \mid Z, U; \theta_0)\|\} < \infty$, $E\{\|(\gamma_2(Y, U; \theta_0)/\gamma_1(Y, U; \theta_0))p(Y \mid Z, U; \theta_0)\|\} < \infty$, *for* $\|\nu\| < \epsilon_2$ *with small enough* $\epsilon_2 > 0$, $E\{\|(\gamma_2(Y, U + \nu; \theta_0)/\gamma_1(Y, U + \nu; \theta_0))\|^2\} < \infty$, *and* $E_{Z \mid U}\{\|\nabla_\theta p(y \mid Z, u + \nu; \theta_0)\|^2 \mid U = u\}$ *and* $E_{Z \mid U}\{p^2(y \mid Z, u + \nu; \theta_0) \mid U = u\}$ *are bounded as functions of y and u.*

(iii) *For small enough* $\|\widetilde{\gamma} - \gamma\|_\infty$, $g(w; \theta, \widetilde{\gamma})$ *is continuously differentiable in* θ *on a neighbourhood of* θ_0, *where* $g(w; \theta, \widetilde{\gamma})$ *is defined in* (13). *There exists a* $b(w)$ *with* $E\{b(W)\} < \infty$ *such that* $\|\nabla_\theta g(w; \theta, \widetilde{\gamma}) - \nabla_\theta g(w; \theta_0, \gamma)\| \le b(w)(\|\widetilde{\gamma} - \gamma\|_\infty^\epsilon + \|\theta - \theta_0\|^\epsilon)$ *for an* $\epsilon > 0$. $E\{\nabla_\theta g(W; \theta_0, \gamma_0)\}$ *exists and is nonsingular.*

(iv) $\sqrt{n}(\log n/nh^p) \to 0$ *and* $\sqrt{n}h^{2m} \to 0$ *as* $n \to \infty$.

Assumptions 2.1–2.2 together guarantee that $\widetilde{\theta}$ is consistent for θ_0. As to $\widehat{\theta}$, there is an extra step of SDR estimation requiring Assumption 2.3.

Assumption 2.3: *Let* $\Omega = \{(y, z, u, \bar{B}) : y \in \mathbb{R}, z \in \mathbb{R}^{p^*}, u \in \mathbb{R}^p, \|\bar{B} - B\| \le cn^{-1/n}$ *for some* $c > 0\}$, *where* p^* *is the dimension of z, and* γ *is as defined in* (14).

(i) *Uniformly in* Ω, *The mth derivatives of* $\gamma_1(y, u; \theta, \bar{B})$ *and* $\gamma_2(y, u; \theta, \bar{B})$ *on* $\bar{B}$ *are locally Lipschitz-continuous as functions of* $\bar{B}^T u$. $E\{p^2(y \mid Z, u; \theta) \mid \bar{B}^T U = \bar{B}^T u\}$ *and each entry in the matrices* $E\{\nabla_\theta p(y \mid Z, u; \theta)\nabla_\theta^T p(y \mid Z, u; \theta) \mid \bar{B}^T U = \bar{B}^T u\}$ *are locally Lipschitz-continuous and bounded from above as a functions of* $\bar{B}^T u$.

(ii) $\widehat{B} \to B$ *in probability as* $n \to \infty$, *and* $\sqrt{n}(\widehat{B} - B) = \sum_{i=1}^n \pi(z_i, u_i)/\sqrt{n} + o_p(1)$, *where* $E\{\pi(z_i, u_i)\} = 0$.

(iii) *The bandwidth* $h = O(n^{-\tau})$ *for* $1/(4m) < \tau < 1/(2d)$.

Note that B is a matrix except when $d = 1$, and when it comes to the calculation of derivatives or norms, we are treating B as vec(B), which denotes the vector formed by concatenating the columns of B. For simplicity, we are still using B to represent vec(B). Similarly, we denote vec($\widehat{B}$) as $\widehat{B}$.

Theorem 2.1: (i) *If Assumptions* 2.1–2.2 *hold. Then, as* $n \to \infty$, $\widetilde{\theta} \to \theta_0$ *in probability and*

$$\sqrt{n}(\widetilde{\theta} - \theta_0) \xrightarrow{d} N(0, G^{-1}\Omega G^{-T}),$$

where $\xrightarrow{d}$ *means convergence in distribution, and*

$$\begin{aligned} G &= E\left\{\nabla_\theta g(W; \theta_0, \gamma_0)\right\}, \\ G^{-T} &= (G^{-1})^T, \\ \Omega &= \mathrm{Var}\left\{g(W; \theta_0, \gamma) - A(Z, U; \theta_0)\right\}, \\ A(Z, U; \theta_0) &= E\left\{g(W; \theta_0, \gamma) \mid Z, U\right\}. \end{aligned}$$

(ii) *If Assumptions* 2.1–2.2 *hold with* (10), (11) *and p replaced by* (14), (15) *and d, respectively, and if Assumption* 2.3 *also holds. Then, as* $n \to \infty$, $\widehat{\theta} \to \theta_0$ *in probability and*

$$\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow{d} N(0, G^{-1}\Omega_B G^{-T}),$$

where

$$\begin{aligned} \Omega_B &= \mathrm{Var}\{g(W; \theta_0, \gamma) - A(Z, U; \theta_0, B) \\ &\quad - \rho(W; \theta_0, B)\}, \\ A(Z, U; \theta_0, B) &= E_{U \mid B^T U}\{A(Z, U; \theta_0)\}, \\ \rho(W; \theta_0, B) &= E\{R\nabla_B[\gamma_2(Y, U; \theta_0, B)/ \\ &\quad \gamma_1(Y, U; \theta_0, B)]\}\pi(Z, U). \end{aligned}$$

It follows from Theorem 2.1 that the asymptotic variance of $\widehat{\theta}$ may or may not be smaller than that of $\widetilde{\theta}$, partly due to the variability of SDR estimation. Owing to SDR, the dimension of kernel estimation is reduced from p to d, which is the main advantage of $\widehat{\theta}$ over $\widetilde{\theta}$, although they have the same convergence rate. A problem with a not so small p is the selection of bandwidth h for kernel estimation. Specifically, if $h = O(n^{-\tau})$ with a $\tau > 0$, then from Assumptions 2.1 and 2.2(iv), τ must be between $1/(4m)$ and $\min\{1/(2p), (1 - 1/q)/p\}$ for $\widetilde{\theta}$, where m is the order of kernel. When p is not so small, the upper bound $\min\{1/(2p), (1 - 1/q)/p\}$ is pretty small, and we may need to increase the kernel order m in order to find such τ satisfying the constraints. If $m > 2$, the kernel takes negative values, which may reduce the stability of kernel estimation. For example, if $q = 2$ in Assumption 2.1, then we must have $1/(4m) < \tau < 1/(2p)$ and we cannot use $m = 2$ when $p > 3$; if $p = 8$, then we must use a kernel of order $m = 5$. On the other hand, if we reduce U to $\widehat{B}^T U$ with dimension d, the upper bound increases to $\min\{1/(2d), (1 - 1/q)/d\}$, allowing enough flexibility in choosing τ to be larger than the lower bound $1/(4m)$, which results in a low-order kernel. If $q = 2$, then we may use $m = 2$ when$d \le 3$.

Although both $\widetilde{\theta}$ and $\widehat{\theta}$ have convergence rate $n^{-1/2}$, $\widehat{\theta}$ may still have an edge over $\widetilde{\theta}$ in finite sample performance, because of smaller dimension and order used in the kernel estimation. This is supported by the simulation results in Section 3.

3. Simulation studies

We study the finite-sample performance of the proposed pseudo likelihood estimators in two simulation studies. Four estimators are compared: the estimator based on full data assuming no missing data ($\widehat{\theta}_{\text{full}}$), the estimator based on complete case analysis ($\widehat{\theta}_{\text{cc}}$), the maximum pseudo likelihood estimator without dimensional reduction ($\widetilde{\theta}$ defined in (8)), and the maximum pseudo likelihood estimator with

Table 1. Simulation results of experiment 1.

		Estimator				Standard Deviation			
		$\widehat{\theta}_{\text{full}}$	$\widehat{\theta}_{\text{cc}}$	$\widetilde{\theta}$	$\widehat{\theta}$	$\widehat{\theta}_{\text{full}}$	$\widehat{\theta}_{\text{cc}}$	$\widetilde{\theta}$	$\widehat{\theta}$
$d=1$	$\beta_0=1$	0.9935	1.8226	1.5812	0.9523	0.1526	0.2002	0.2429	0.2567
	$\beta_1=1$	1.0017	0.8448	0.8462	1.0133	0.0842	0.0950	0.1072	0.1042
	$\beta_2=1$	1.0005	0.8417	0.8285	1.0031	0.0819	0.0946	0.1094	0.1066
	$\beta_3=1$	1.0046	0.8449	0.8331	1.0065	0.0850	0.0957	0.1062	0.1086
	$\beta_4=1$	1.0016	0.8448	0.8403	1.0161	0.0818	0.0942	0.1031	0.1060
	$\beta_5=-1$	−1.0008	−0.9620	−0.9230	−0.9979	0.0416	0.0469	0.0612	0.0530
	$\sigma^2=2$	1.9725	1.7240	1.9107	1.9991	0.1415	0.1442	0.1854	0.2265
$d=2$	$\beta_0=1$	0.9995	1.8152	1.3759	0.9347	0.1654	0.2637	0.2453	0.2402
	$\beta_1=1$	0.9959	0.7565	0.9055	1.0319	0.0910	0.1231	0.1164	0.1226
	$\beta_2=1$	0.9960	0.7596	0.9031	1.0323	0.0906	0.1240	0.1186	0.1271
	$\beta_3=1$	1.0001	0.7629	0.9090	1.0340	0.0907	0.1240	0.1189	0.1283
	$\beta_4=1$	1.0020	0.7630	0.9079	1.0362	0.0914	0.1201	0.1163	0.1206
	$\beta_5=-1$	−0.9983	−0.8816	−0.9408	−1.0077	0.0436	0.0588	0.0623	0.0647
	$\sigma^2=2$	1.9750	1.7556	1.9051	2.0024	0.1429	0.1479	0.1884	0.2170

dimensional reduction ($\widehat{\theta}$ defined in (9)). The NW estimations are computed using a Gaussian kernel $K(u) = \exp(-u^2/2)/(2\pi)^{1/2}$. The bandwidths are selected by rule of thumb proposed by Silverman (1986), i.e., $h = (4/3)\widehat{\sigma}n^{-1/5}$, where $\widehat{\sigma}$ is the estimated standard deviation of U. For SDR, we apply SIR in Li (1991) with the number of slices = 5. All results are based on 1000 simulation replications and sample size $n = 400$.

3.1. Experiment 1

In the first experiment, $U = (U_1, U_2, U_3, U_4, U_5, U_6)$, where $U_j \sim N(1,1)$ and $\text{Cov}(U_j, U_{j'}) = 0$. Hence, the dimension p equals 6. To evaluate the performance of SDR, two models of $p(Z \mid U)$ are studied: $Z \mid U \sim N(U_1 + U_2 + U_3 + U_4 + U_5 + U_6 - 2, 1)$ and $Z \mid U \sim N((U_1 + U_2)^2/3 + (U_3 + U_4)^2/3, 1)$, where d equals 1 and 2, respectively. The response Y is generated as $Y \mid X \sim N(\beta_0 + \beta_1 U_1 + \beta_2 U_2 + \beta_3 U_3 + \beta_4 U_4 + \beta_5 Z, \sigma^2)$, where $\theta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \sigma^2) = (1, 1, 1, 1, 1, 1, -1, 2)$. The propensity $P(R = 1 \mid Y, U) = (1 + \exp\{5 - (Y + U_1 + U_2 + U_3 + U_4 + U_5 + U_6)\})^{-1}$, resulting in unconditional response rates 73% and 74% in two cases, respectively. Table 1 reports the estimators and their standard deviations.

The simulation results can be summarised as follows. First, $\widehat{\theta}$ is nearly unbiased for θ no matter d equals 1 or 2. Second, the bias of $\widetilde{\theta}$ may not be negligible and, in some cases, the bias is comparable to $\widehat{\theta}_{\text{cc}}$ that is biased in theory. As we discussed in Section 2, compared with $\widehat{\theta}$, the assumptions are harder to satisfy for $\widetilde{\theta}$. Third, the standard deviations of $\widetilde{\theta}$ and $\widehat{\theta}$ are quite close and, in some cases, $\widetilde{\theta}$ has slightly smaller standard deviation than that of $\widehat{\theta}$.

3.2. Experiment 2

In the second experiment, we consider a binary outcome Y. Covariate U is generated the same as that in the first experiment, $Z \mid U \sim N(\log(U_1 + U_2 + U_3 + U_4)^2 - 1, 1)$, or $Z \mid U \sim N(\log(U_1 + U_2)^2 + \log(U_3 + U_4)^2 - 1, 1)$. The binary Y is generated according to $P(Y = 1 \mid X) = (1 + \exp\{-(\beta_0 + \beta_1 U_1 + \beta_2 U_2 + \beta_3 U_3 + \beta_4 U_4 + \beta_5 Z)\})^{-1}$, where $\theta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5) = (-1, 1, 1, 1, 1, -1)$. The propensity $P(R = 1 \mid Y, U) = (1 + \exp\{-\frac{1}{3}Y(U_1 + U_2 + U_3 + U_4 + U_5 + U_6 - 2)\})^{-1}$, resulting unconditional response rates 71% or 72%. The simulation results are reported in Table 2. The conclusions are quite similar to those for experiment 1, but the bias of $\widetilde{\theta}$ is more serious: the bias of $\widetilde{\theta}$ are even larger than the bias of $\widehat{\theta}_{\text{cc}}$. Meanwhile, $\widehat{\theta}$ is still nearly unbiased.

Table 2. Simulation results of experiment 2.

		Estimator				Standard Deviation			
		$\widehat{\theta}_{\text{full}}$	$\widehat{\theta}_{\text{cc}}$	$\widetilde{\theta}$	$\widehat{\theta}$	$\widehat{\theta}_{\text{full}}$	$\widehat{\theta}_{\text{cc}}$	$\widetilde{\theta}$	$\widehat{\theta}$
$d=1$	$\beta_0=-1$	−1.0369	−0.9579	−3.7354	−0.9329	0.3671	0.5210	0.5141	0.5301
	$\beta_1=1$	1.0307	1.1445	1.8894	0.9512	0.1858	0.2649	0.2733	0.2544
	$\beta_2=1$	1.0245	1.1253	1.8723	0.9524	0.1911	0.2626	0.2849	0.2657
	$\beta_3=1$	1.0236	1.1281	1.8988	0.9612	0.1883	0.2711	0.2782	0.2648
	$\beta_4=1$	1.0283	1.1376	1.9018	0.9577	0.1854	0.2623	0.2735	0.2613
	$\beta_5=-1$	−1.0265	−1.0499	−1.1685	−0.9859	0.1498	0.2181	0.2105	0.2120
$d=2$	$\beta_0=-1$	−1.0307	−0.9436	−4.6084	−1.0587	0.4623	0.6637	0.6068	0.6450
	$\beta_1=1$	1.0271	1.1336	1.9316	1.0554	0.2154	0.3046	0.3262	0.3091
	$\beta_2=1$	1.0318	1.1401	1.9586	1.0627	0.2085	0.3064	0.3353	0.3147
	$\beta_3=1$	1.0227	1.1270	2.0029	1.0553	0.2043	0.2973	0.3217	0.2928
	$\beta_4=1$	1.0268	1.1367	1.9532	1.0458	0.2140	0.3070	0.3324	0.3011
	$\beta_5=-1$	−1.0275	−1.0568	−1.2408	−1.0294	0.1495	0.2176	0.2011	0.2104

4. Proofs

Proof of Theorem 2.1

We first prove part (i). Here, γ is defined as in (10) and $\widetilde{\gamma}$ is defined as in (11). When $\theta = \theta_0$, denote $\gamma(y, u; \theta_0)$ as $\gamma_0(y, u) = (\gamma_{01}(y, u), \gamma_{02}(y, u))^T$. Note that, in Assumption 2.1, we assume that there exists a constant $c > 0$ such that $\inf p(y, u) = \inf \gamma_{01}(y, u) \geq c$. Moreover, the univariate kernel $\kappa(\cdot)$ is bounded, hence, there exists a constant $b > 0$ such that $\inf \widetilde{\gamma}_1(y, u; \theta_0) \geq b$.

For the consistency of $\widetilde{\theta}$, we only need to verify the assumptions of Theorem 2(a) in Zhao and Shao (2015). Note that Assumptions 2.1 and 2.2(i), (iv) guarantee that $\widetilde{\gamma}$ is consistent. Then,

$$\begin{aligned}
&|\widetilde{\gamma}_1(y_i, u_i; \theta) - \gamma_1(y_i, u_i; \theta)| \\
&\quad \leq \left\| \widetilde{\gamma}_1(y_i, u_i; \theta) - \gamma_1(y_i, u_i; \theta) \right\|_\infty \\
&\quad = o_p(1).
\end{aligned}$$

Hence,

$$\begin{aligned}
&|H(w_i, \theta, \widetilde{\gamma}_1) - H(w_i; \theta, \gamma_1)| \\
&\quad = |-r_i \left\{ \log(\widetilde{\gamma}_1(y_i, u_i; \theta)) - \log(\gamma_1(y_i, u_i; \theta)) \right\}| \\
&\quad = \left| -r_i \left\{ \log \frac{\widetilde{\gamma}_1(y_i, u_i; \theta)}{\gamma_1(y_i, u_i; \theta)} \right\} \right| \\
&\quad \leq \left| \log \frac{\widetilde{\gamma}_1(y_i, u_i; \theta)}{\gamma_1(y_i, u_i; \theta)} \right| \\
&\quad = \left| \log \frac{\gamma_1(y_i, u_i; \theta) + o_p(1)}{\gamma_1(y_i, u_i; \theta)} \right|.
\end{aligned}$$

Since $\inf \gamma_1(y_i, u_i; \theta) \geq c$, $|H(w_i, \theta, \widetilde{\gamma}_1) - H(w_i; \theta, \gamma_1)| \to 0$ in probability as $n \to \infty$. Therefore,

$$E\{H(w_i, \theta, \widetilde{\gamma}_1) - H(w_i; \theta, \gamma_1)\} \to 0 \quad \text{inprobability.}$$

Following Theorem 2 in Zhao and Shao (2015), with Assumption 2.2(i), $\widetilde{\theta} \to \theta_0$ in probability as $n \to \infty$, where θ_0 is the true value of θ.

For the asymptotic normality of $\widetilde{\theta}$, we only need to verify the assumptions of Theorem 8.11 in Newey and McFadden (1994).

Step 1. Let $w_i = (y_i, z_i, u_i, r_i)$, and

$$G(w_i, \gamma) = -\frac{r_i}{\gamma_{01}(w_i)} \left[\gamma_2(w_i) - \frac{\gamma_{02}(w_i)}{\gamma_{01}(w_i)} \gamma_1(w_i) \right].$$

We would like to prove

$$\begin{aligned}
&\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[g(w_i; \theta_0, \widetilde{\gamma}) - g(w_i; \theta_0, \gamma_0) - G(w_i, \widetilde{\gamma} - \gamma_0) \right] \\
&= o_p(1).
\end{aligned}$$

Note that

$$\begin{aligned}
&\sqrt{n} E \left\| \left[g(w_i; \theta_0, \widetilde{\gamma}) - g(w_i; \theta_0, \gamma_0) - G(w_i, \widetilde{\gamma} - \gamma_0) \right] \right\| \\
&\quad \leq \sqrt{n} E \left\{ r_i \frac{1}{\widetilde{\gamma}_1(w_i) \gamma_{01}(w_i)} \left[1 + \left\| \frac{\gamma_{02}(w_i)}{\gamma_{01}(w_i)} \right\| \right] \right. \\
&\quad \left. \|\widetilde{\gamma} - \gamma_0\|^2 \right\} \\
&\quad \leq \sqrt{n} E \left\{ r_i \frac{1}{\widetilde{\gamma}_1(w_i) \gamma_{01}(w_i)} \frac{\gamma_{01}(w_i)}{c} \frac{\widetilde{\gamma}_1(w_i)}{b} \right. \\
&\quad \left. \left[1 + \left\| \frac{\gamma_{02}(w_i)}{\gamma_{01}(w_i)} \right\| \right] \|\widetilde{\gamma} - \gamma_0\|^2 \right\} \\
&\quad \leq \frac{1}{bc} E \left\{ r_i \left[1 + \left\| \frac{\gamma_{02}(w_i)}{\gamma_{01}(w_i)} \right\| \right] \right\} \\
&\quad \left[\sqrt{n} \sup \|\widetilde{\gamma} - \gamma_0\|^2 \right] = o_p(1), \qquad (16)
\end{aligned}$$

which follows from the Assumptions 2.1 and 2.2(ii).

Step 2. Let

$$\begin{aligned}
&\sqrt{n} \left[\frac{1}{n} \sum_{i=1}^{n} G(w_i, \widetilde{\gamma} - \gamma_0) - \int G(w_i, \widetilde{\gamma} - \gamma_0) \, \mathrm{d}w_i \right] \\
&\quad = \sqrt{n} \left[\frac{1}{n} \sum_{i=1}^{n} G(w_i, \widetilde{\gamma} - \bar{\gamma}) \right. \\
&\quad \left. - \int G(w_i, \widetilde{\gamma} - \bar{\gamma}) \, \mathrm{d}F(w_i) \right] \\
&\qquad + \sqrt{n} \left[\frac{1}{n} \sum_{i=1}^{n} G(w_i, \bar{\gamma} - \gamma_0) \right. \\
&\quad \left. - \int G(w_i, \bar{\gamma} - \gamma_0) \, \mathrm{d}F(w_i) \right] \\
&\quad = S_{n1} + S_{n2}.
\end{aligned}$$

We would like to prove $S_{n1} = o_p(1)$ and $S_{n2} = o_p(1)$. For S_{n1}, V-statistics method is applied. We use the result of Lemma 8.4 by Newey and McFadden (1994) directly. In our case, $m_n(w_i, w_j) = -(r_i/\gamma_{01}(w_i))[\nabla_\theta p(y_i \mid z_j, u_i) - (\gamma_{02}(w_i)/\gamma_{01}(w_i)) p(y_i \mid z_j, u_i)] K_h(u_i - u_j)$. Then

$$\begin{aligned}
&E \|m_n(w_i, w_i)\| \\
&\quad \leq E \left\| \frac{r_i}{\gamma_{01}(w_i)} \left[\nabla_\theta p(y_i \mid z_i, u_i) \right. \right. \\
&\quad \left. \left. - \frac{\gamma_{02}(w_i)}{\gamma_{01}(w_i)} p(y_i \mid z_i, u_i) \right] K_h(0) \right\| \\
&\quad \leq h^{-p} K(0) E \left\{ \frac{r_i}{\gamma_{01}} \frac{\gamma_{01}}{c} \left[\|\nabla_\theta p(y_i \mid x_i, u_i)\| \right. \right. \\
&\quad \left. \left. + \left\| \frac{\gamma_{02}(w_i)}{\gamma_{01}(w_i)} \right\| p(y_i \mid z_i, u_i) \right] \right\} \\
&\quad = c^{-1} h^{-p} K(0) E \left\{ r_i \left[\|\nabla_\theta p(y_i \mid x_i, u_i)\| \right. \right. \\
&\quad \left. \left. + \left\| \frac{\gamma_{02}(w_i)}{\gamma_{01}(w_i)} \right\| p(y_i \mid z_i, u_i) \right] \right\}.
\end{aligned}$$

Let $t \in \Omega^k, u_i \in \mathbb{R}^p, y_i \in \mathbb{R}, 0 < \bar{h} \le h$, then

$$\begin{aligned}
&E\left[\left\|m_n(w_i, w_j)\right\|^2\right] \\
&\quad \le E\left\{\left\|\frac{r_i}{\gamma_{01}(w_i)}\left[\nabla_\theta p(y_i \mid z_j, u_i)\right.\right.\right. \\
&\quad \left.\left.\left. -\frac{\gamma_{02}(w_i)}{\gamma_{01}(w_i)} p(y_i \mid z_j, u_j)\right] K_h(u_i - u_j)\right\|^2\right\} \\
&\quad \le R_{n1} + R_{n2},
\end{aligned}$$

where

$$\begin{aligned}
R_{n1} &= E\left\{\left\|\frac{r_i}{\gamma_{01}(w_i)} \nabla_\theta p(y_i \mid z_j, u_i) K_h(u_i - u_j)\right\|^2\right\} \\
&\le E\left\{\frac{r_i^2}{\gamma_{01}(w_i)^2} \frac{\gamma_{01}(w_i)^2}{c^2} \left\|\nabla_\theta p(y_i \mid z_i, u_j)\right\|^2\right. \\
&\quad \left. K_h^2(u_i - u_j)\right\} \\
&= c^{-2} h^{-p} E[r_i^2] \int K^2(t)\, \mathrm{d}t \\
&\quad \times \sup_{t, y_i, u_i, \bar{h}} \left\{E_{Z_j \mid U_j}\left[\left\|\nabla_\theta p(y_i \mid Z_j, u_i)\right\|^2\right.\right. \\
&\quad \left.\left. \mid U_j = u_i + v\right] p(u_i + \bar{h}t)\right\}
\end{aligned}$$

and

$$\begin{aligned}
R_{n2} &= E\left\{\left\|\frac{r_i}{\gamma_{01}(w_i)} p(y_i \mid z_j, u_i) \frac{\gamma_{02}(w_i)}{\gamma_{01}(w_i)} K_h(u_i - u_j)\right\|^2\right\} \\
&\le E\left\{\frac{r_i^2}{\gamma_{01}(w_i)^2} \frac{\gamma_{01}(w_i)^2}{c^2} \left\|p(y_i \mid z_j, u_i) \frac{\gamma_{02}(w_i)}{\gamma_{01}(w_i)}\right\|^2\right. \\
&\quad \left. K_h^2(u_i - u_j)\right\} \\
&= c^{-2} h^{-p} E\left\{r_i^2 \left\|\frac{\gamma_{02}(w_i)}{\gamma_{01}(w_i)}\right\|^2\right\} \int K^2(t)\, \mathrm{d}t \\
&\quad \times \sup_{t, y_i, u_i, \bar{h}} \left\{E_{Z_j \mid U_j}\left[\left\|p(y_i \mid Z_j, u_i)\right\|^2 \mid U_j = u_i + v\right]\right. \\
&\quad \left. \times\, p(u_i + \bar{h}t)\right\}
\end{aligned}$$

Then, following Assumption 2.2(ii) and (iv), we obtain that

$$\begin{aligned}
\sqrt{n} E \left\|m_n(w_i, w_i)\right\| / n &= o_p(1), \\
\sqrt{n} E^{1/2}[\left\|m_n(w_i, w_j)\right\|^2]/n &= o_p(1).
\end{aligned}$$

Hence, $S_{n1} = o_p(1)$.

As to S_{n2}, by Chebychev's Inequality, since $E[S_{n2}] = 0$,

$$\begin{aligned}
&P\left(\left\|\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^n G(w_i, \bar{\gamma} - \gamma_0)\right.\right. \\
&\quad \left.\left. - \int G(w_i, \bar{\gamma} - \gamma_0)\, \mathrm{d}F(w_i)\right]\right\| > \epsilon\right) \\
&= P\left(\left\|\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^n G(w_i, \bar{\gamma} - \gamma_0)\right.\right. \\
&\quad \left.\left. - \int G(w_i, \bar{\gamma} - \gamma_0)\, \mathrm{d}w_i\right]\right\|^2 > \epsilon^2\right) \\
&\le n^2(n-1) \left\|E\left\{G(w_i, \bar{\gamma} - \gamma_0)\right.\right. \\
&\quad \left.\left. - \int G(w_i, \bar{\gamma} - \gamma_0)\, \mathrm{d}F(\omega_i)\right\}\right\|^2 \Big/ (n^2\epsilon^2) \\
&\quad + n^2 E\left\{\left\|G(w_i, \bar{\gamma} - \gamma_0)\right.\right. \\
&\quad \left.\left. - \int G(w_i, \bar{\gamma} - \gamma_0)\, \mathrm{d}F(w_i)\right\|^2\right\} \Big/ (n^2\epsilon^2) \\
&= \epsilon^2 E\left\{\left\|G(w_i, \bar{\gamma} - \gamma_0)\right.\right. \\
&\quad \left.\left. - \int G(w_i, \bar{\gamma} - \gamma_0)\, \mathrm{d}F(w_i)\right\|^2\right\} \\
&\le \epsilon^2 E\{\|G(w_i, \bar{\gamma} - \gamma_0)\|^2\} \\
&= \epsilon^2 E\left\{\frac{r_i^2}{\gamma_{01}^2(w_i)} \left\|\left[1, -\frac{\gamma_{02}(w_i)}{\gamma_{01}(w_i)}\right]\right.\right. \\
&\quad \left.\left. [\bar{\gamma}_2 - \gamma_{02}, \bar{\gamma}_1 - \gamma_{01}]\right\|^2\right\} \\
&\le \epsilon^2 E\left\{r_i^2\left[1 + \left\|\frac{\gamma_{02}(w_i)}{\gamma_{01}(w_i)}\right\|^2\right]\right\} \frac{\sup \|\widehat{\gamma} - \gamma_{01}\|^2}{c^2} \\
&= o_p(1),
\end{aligned}$$

which follows the Assumptions 2.1 and 2.2(ii).

Step 3. Note that

$$\begin{aligned}
\int G(w, \gamma)\, \mathrm{d}F(w) &= \iiint -\frac{rp(r, y, u)}{\gamma_{01}(w)} \\
&\quad \times \left[1, -\frac{\gamma_{02}(w)}{\gamma_{01}(w)}\right] \begin{bmatrix}\gamma_2(w) \\ \gamma_1(w)\end{bmatrix} \mathrm{d}r\, \mathrm{d}y\, \mathrm{d}u.
\end{aligned}$$

Let $v(r, y, u) = -(rp(r, y, u)/\gamma_{01}(w))[1, -(\gamma_{02}(w)/\gamma_{01}(w))]$,

$$A(z, u) = \iint v\left(r, y, u\begin{bmatrix}\nabla_\theta p(y \mid z, u) \\ p(y \mid z, u)\end{bmatrix}\right) \mathrm{d}r\, \mathrm{d}y,$$

and $\delta(z, u) = A(z, u) - E[A(Z, U)] = A(z, u)$. Then,

$$\begin{aligned}
&\sqrt{n} \int G(w, \widehat{\gamma} - \gamma_0)\, \mathrm{d}F(w) \\
&\quad = \int G(w, \widehat{\gamma})\, \mathrm{d}F(w) \\
&\quad = \sqrt{n} \frac{1}{n} \sum_{j=1}^n \iiint v(r, y, u) \\
&\quad \left[\nabla_\theta p(y \mid z_j, u); \quad p(y \mid z_j, u)\right] \\
&\quad K_h(u_j - u)\, \mathrm{d}r\, \mathrm{d}y\, \mathrm{d}u.
\end{aligned}$$

We would like to prove

$$\sqrt{n}\left(\int G(w,\widehat{\gamma}-\gamma_0)\,\mathrm{d}F(w)-\frac{1}{n}\sum_{j=1}^{n}\iint v(r,y,u_j)=\right.$$
$$\left.\times\begin{bmatrix}\nabla_\theta p(y\mid z_j,u_j)\\ p(y\mid z_j,u_j)\end{bmatrix}\mathrm{d}r\,\mathrm{d}y\right)$$
$$=\sqrt{n}\frac{1}{n}\sum_{j=1}^{n}\iint\left(\int v(r,y,u)\begin{bmatrix}\nabla_\theta p(y\mid z_j,u)\\ p(y\mid z_j,u)\end{bmatrix}\right.$$
$$\left.\times K_h(u_j-u)\,\mathrm{d}u-v(r,y,u_j)\begin{bmatrix}\nabla_\theta p(y\mid z_j,u_j)\\ p(y\mid z_j,u_j)\end{bmatrix}\right)\mathrm{d}r\,\mathrm{d}y$$
$$=\sqrt{n}\frac{1}{n}\sum_{j=1}^{n}D(z_j,u_j)$$
$$=o_p(1).$$

By Chebychev's Inequality, we only need to prove

$$P\left(\left\|\sqrt{n}\frac{1}{n}\sum_{j=1}^{n}D(z_j,u_j)\right\|^2>\epsilon^2\right)$$
$$\le n\left[n(n-1)\,\|E[D(Z,U]\|^2+nE(\|D(Z,U)\|^2)\right]/$$
$$\times(n^2\epsilon^2)\to 0.$$

So we only need to prove $\sqrt{n}\left\|E[D(Z_j,U_j)]\right\|\to 0$ and $E[\left\|D(Z_j,U_j)\right\|^2]\to 0$. Let

$$E_{Z_j\mid U_j=u_j}[p(y\mid Z_j,u)\mid u_j]p(u_j)=\gamma_{01}(y,u,u_j),$$
$$E_{Z_j\mid U_j=u_j}[\nabla_\theta p(y\mid Z_j,u)\mid u_j]p(u_j)=\gamma_{02}(y,u,u_j).$$

Then,

$$\sqrt{n}\left\|E[D(Z_j,U_j)]\right\|$$
$$=\sqrt{n}\left\|\iiint E\left\{v(r,y,u)\right.\right.$$
$$\left.\times\begin{bmatrix}\nabla_\theta p(y\mid z_j,u)\\ p(y\mid z_j,u)\end{bmatrix}K_h(u_j-u)\right\}\mathrm{d}r\,\mathrm{d}y\,\mathrm{d}u$$
$$\left.-\iint E\left\{v(r,y,u_j)\begin{bmatrix}\nabla_\theta p(y\mid z_j,u_j)\\ p(y\mid z_j,u_j)\end{bmatrix}\right\}\mathrm{d}r\,\mathrm{d}y\right\|$$
$$=\sqrt{n}\left\|\iiiint v(r,y,u)\begin{bmatrix}\gamma_{02}(y,u,u+\bar{h}t)\\ \gamma_{01}(y,u,u+\bar{h}t)\end{bmatrix}\right.$$
$$\times K(t)\,\mathrm{d}t\,\mathrm{d}r\,\mathrm{d}y\,\mathrm{d}u$$
$$\left.-\iiint v(r,y,u_j)\begin{bmatrix}\gamma_{02}(y,u_j,u_j)\\ \gamma_{01}(y,u_j,u_j)\end{bmatrix}\mathrm{d}r\,\mathrm{d}y\,\mathrm{d}u_j\right\|$$
$$\le\sqrt{n}\iiiint\|v(r,y,u)\|\left\|\begin{bmatrix}\gamma_{02}(y,u,u+\bar{h}t)\\ \gamma_{01}(y,u,u+\bar{h}t)\end{bmatrix}\right.$$
$$\left.-\begin{bmatrix}\gamma_{02}(y,u,u)\\ \gamma_{01}(y,u,u)\end{bmatrix}\right\|K(t)\,\mathrm{d}t\,\mathrm{d}r\,\mathrm{d}y\,\mathrm{d}u$$
$$\le\sqrt{n}h^m\iiint\|v(r,y,u)\|\,\mathrm{d}r\,\mathrm{d}y\,\mathrm{d}u$$
$$\times\sup_{t,y_i,u_i,\bar{h}}\left\|\frac{\partial^m\gamma_0(y,u,u_j)}{\partial u_j}\bigg|_{u_j=u+\bar{h}t}\right\|\int K(t)\,\mathrm{d}t$$
$$\le c^{-1}\sqrt{n}h^mE\left\{r\left[1+\left\|\frac{\gamma_{02}(w)}{\gamma_{01}(w)}\right\|\right]\right\}$$
$$\times\sup_{t,y_i,u_i,\bar{h}}\left\|\frac{\partial^m\gamma_0(y,u,u_j)}{\partial u_j}\bigg|_{u_j=u+\bar{h}t}\right\|\int K(t)\,\mathrm{d}t,$$

Therefore, by Assumption 2.2(ii) and (iv), $\sqrt{n}\left\|E[D(Z_j,U_j)]\right\|\to 0$. And

$$E[\left\|D(Z_j,U_j)\right\|^2]$$
$$=E_{Z_j,U_j}\left\{\left\|\iint\left(\int v(r,y,u)\begin{bmatrix}\nabla_\theta p(y\mid z_j,u)\\ p(y\mid z_j,u)\end{bmatrix}\right.\right.$$
$$\left.K_h(U_j-u)\,\mathrm{d}u-v(r,y,U_j)\begin{bmatrix}\nabla_\theta p(y\mid Z_j,U_j)\\ p(y\mid Z_j,U_j)\end{bmatrix}\right)$$
$$\left.\mathrm{d}r\,\mathrm{d}y\right\|^2\right\}$$
$$\le E_{Z_j,U_j}\left\{\left\|\iiint v(r,y,u)\begin{bmatrix}\nabla_\theta p(y\mid z_j,u)\\ p(y\mid z_j,u)\end{bmatrix}\right.\right.$$
$$\left.K_h(U_j-u)\,\mathrm{d}u\,\mathrm{d}r\,\mathrm{d}y\right\|^2\right\}$$
$$+E_{Z_j,U_j}\left\{\left\|\iint v(r,y,U_j)\begin{bmatrix}\nabla_\theta p(y\mid Z_j,U_j)\\ p(y\mid Z_j,U_j)\end{bmatrix}\right.\right.$$
$$\left.\mathrm{d}r\,\mathrm{d}y\right\|^2\right\}$$
$$=Q_{n1}+Q_{n2}.$$

Note that,

$$Q_{n1}\le E_{Z_j,U_j}\left\{\int\left\|\iint v(r,y,U_j+\bar{h}t)\right.\right.$$
$$\left.\left.\times\begin{bmatrix}\nabla_\theta p(y\mid z_j,U_j+\bar{h}t)\\ p(y\mid z_j,U_j+\bar{h}t)\end{bmatrix}\mathrm{d}r\,\mathrm{d}y\right\|^2K(t)^2\,\mathrm{d}t\right\}$$
$$\le CK^2(0)\sup_{\|\nu\|<\nu_0}\iiint\|v(r,y,u_j+\nu)\|^2\,\mathrm{d}u_j\,\mathrm{d}r\,\mathrm{d}y$$
$$\times\sup_{y,u_j,t,\bar{h}}\left\{E_{Z_j\mid U_j}\left[\left\|\begin{bmatrix}\nabla_\theta p(y\mid z_j,u_j+\bar{h}t)\\ p(y\mid z_j,u_j+\bar{h}t)\end{bmatrix}\right\|^2\right.\right.$$
$$\left.\left.\mid U_j=u_j\right]p(u_j)\right\}$$
$$\le CK^2(0)c^{-2}\sup_{\|\nu\|<\nu_0}\iiint r^2$$
$$\left[1+\left\|\frac{\gamma_{02}(y,u+\nu)}{\gamma_{01}(y,u+\nu)}\right\|\right]^2$$
$$p^2(r,y,u+\nu)\,\mathrm{d}r\,\mathrm{d}y\,\mathrm{d}u$$

$$\times \sup_{y,u_j,t,\bar{h}} \left\{ E_{Z_j \mid U_j} \left[\left\| \begin{bmatrix} \nabla_\theta p(y \mid z_j, u_j + \bar{h}t) \\ p(y \mid z_j, u_j + \bar{h}t) \end{bmatrix} \right\|^2 \mid U_j = u_j \right] p(u_j) \right\} < \infty$$

for some constant C, and

$$Q_{n2} \le \iiint \|v(r, y, u_j)\|^2 \, \mathrm{d}u_j \, \mathrm{d}r \, \mathrm{d}y \sup_{y,u_j} \left\{ E_{Z_j \mid U_j} \left[\left\| \begin{bmatrix} \nabla_\theta p(y \mid Z_j, U_j) \\ p(y \mid Z_j, U_j) \end{bmatrix} \right\|^2 \mid U_j = u_j \right] p(u_j) \right\}$$

$$\le c^{-2} \sup_{\|v\| < v_0} \iiint r^2 \left[1 + \left\| \frac{\gamma_{02}(y, u + v)}{\gamma_{01}(y, u + v)} \right\| \right]^2 p^2(r, y, u + v) \, \mathrm{d}r \, \mathrm{d}y \, \mathrm{d}u$$

$$\times \sup_{y,u_j,t,\bar{h}} \left\{ E_{Z_j \mid U_j} \left[\left\| \begin{bmatrix} \nabla_\theta p(y \mid z_j, u_j + \bar{h}t) \\ p(y \mid z_j, u_j + \bar{h}t) \end{bmatrix} \right\|^2 \mid U_j = u_j \right] p(u_j) \right\} < \infty.$$

Since when n is large enought, and $t \in \Omega_K$ is bounded,

$$\left\| v(r, y, u_j + \bar{h}t) \begin{bmatrix} \nabla_\theta p(y \mid z_j, u_j + \bar{h}t) \\ p(y \mid z_j, u_j + \bar{h}t) \end{bmatrix} \right\| \le \left\| v(r, y, u_j + v) \begin{bmatrix} \nabla_\theta p(y \mid z_j, u_j + v) \\ p(y \mid z_j, u_j + v) \end{bmatrix} \right\| \quad a.s. \quad (r, y, u_j).$$

Then by conditional dominated convergence theorem,

$$\int v(r, y, u) \begin{bmatrix} \nabla_\theta p(y \mid z_j, u) \\ p(y \mid z_j, u) \end{bmatrix} K_h(u_j - u) \, \mathrm{d}u - v(r, y, u_j) \times \begin{bmatrix} \nabla_\theta p(y \mid z_j, u_j) \\ p(y \mid z_j, u_j) \end{bmatrix}$$

$$= \int v(r, y, u_j + \bar{h}t) \begin{bmatrix} \nabla_\theta p(y \mid z_j, u_j + \bar{h}t) \\ p(y \mid z_j, u_j + \bar{h}t) \end{bmatrix} K(t) \, \mathrm{d}t - v(r, y, u_j) \begin{bmatrix} \nabla_\theta p(y \mid z_j, u_j) \\ p(y \mid z_j, u_j) \end{bmatrix}$$

$$\to 0 \quad a.s. \quad (r, y, u_j).$$

By Dominated Convergence Theorem, $E[\|D(Z_j, U_j)\|^2] \to 0$.

Then, we combine Step 1–3, based on Theorems 8.11 and 8.12 in Newey and McFadden (1994), with Assumption 2.2(iii), $\sum_{i=1}^n g(w_i; \theta_0, \widetilde{\gamma})/\sqrt{n} \to_d N(0, \Omega)$ and $\sqrt{n}(\widetilde{\theta} - \theta_0) \to_d N(0, G^{-1}\Omega G^{-T})$.

Second, we prove part (ii). Here, γ is defined as in (10) and $\widehat{\gamma}$ is defined as in (15). We consider the normality of $g(w_i; \theta_0, \widehat{\gamma}, \widehat{B})$. When $\widehat{B}$ is replaced by the true SDR direction B, the proof is exactly the same as in part (i). Then we could easily get that $\sum_{i=1}^n g(w_i; \widehat{\gamma}, \theta_0, B)/\sqrt{n} \to_d N(0, \Omega_d)$. Now we only need to consider

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (g(w_i; \theta_0, \widehat{\gamma}, \widehat{B}) - g(w_i; \theta_0, \widehat{\gamma}, B))$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^n -r_i \left(\frac{\widehat{\gamma}_2(w_i, \widehat{B})}{\widehat{\gamma}_1(w_i, \widehat{B})} - \frac{\widehat{\gamma}_2(w_i, B)}{\widehat{\gamma}_1(w_i, B)} \right)$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^n -r_i \left(\frac{\widehat{\gamma}_2(w_i, \widehat{B})}{\widehat{\gamma}_1(w_i, \widehat{B})} - \frac{\widehat{\gamma}_2(w_i, B)}{\widehat{\gamma}_1(w_i, B)} - \frac{\gamma_2(w_i, \widehat{B})}{\gamma_1(w_i, \widehat{B})} + \frac{\gamma_2(w_i, B)}{\gamma_1(w_i, B)} \right)$$

$$+ \frac{1}{\sqrt{n}} \sum_{i=1}^n -r_i \left(\frac{\gamma_2(w_i, \widehat{B})}{\gamma_1(w_i, \widehat{B})} - \frac{\gamma_2(w_i, B)}{\gamma_1(w_i, B)} \right).$$

Based on Lemma 3 in Ma and Zhu (2012), with Assumption 2.3(i) and (iii),

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n -r_i \left(\frac{\widehat{\gamma}_2(w_i, \widehat{B})}{\widehat{\gamma}_1(w_i, \widehat{B})} - \frac{\widehat{\gamma}_2(w_i, B)}{\widehat{\gamma}_1(w_i, B)} - \frac{\gamma_2(w_i, \widehat{B})}{\gamma_1(w_i, \widehat{B})} + \frac{\gamma_2(w_i, B)}{\gamma_1(w_i, B)} \right)$$

$$= O_p(h^m + n^{-1/2} h^{-(q+1)} \log n) \to 0.$$

And, with Assumption 2.3(ii), we can easily prove the normality of

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n -r_i \left(\frac{\gamma_2(w_i, \widehat{B})}{\gamma_1(w_i, \widehat{B})} - \frac{\gamma_2(w_i, B)}{\gamma_1(w_i, B)} \right).$$

Then we finish the proof of part (ii).

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by Division of Mathematical Sciences [1612873] and the Chinese Ministry of Education 111 Project [B14019].

Notes on contributors

Ji Chen is a PhD candidate in East China Normal University.

Bingying Xie is a statistician at Roche in Shanghai, China.

Jun Shao is a professor in University of Wisconsin-Madison.

References

Baker, S. G., & Laird, N. M. (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical association*, *83*(401), 62–69.

Bierens, H. J. (1987). Kernel estimators of regression functions. In *Advances in econometrics: Fifth world congress* (Vol. 1, pp. 99–144).

Cook, R. D., & Weisberg, S. (1991). Comment. *Journal of the American Statistical Association*, *86*(414), 328–332.

Greenlees, J. S., Reece, W. S., & Zieschang, K. D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association*, *77*(378), 251–261.

Hansen, B. E. (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory*, *24*(3), 726–748.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, *86*(414), 316–327.

Li, B., & Wang, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association*, *102*(479), 997–1008.

Little, R. J., & Rubin, D. B. (2014). *Statistical analysis with missing data* (Vol. 333). New York: John Wiley & Sons.

Ma, Y., & Zhu, L. (2012). A semiparametric approach to dimension reduction. *Journal of the American Statistical Association*, *107*(497), 168–179.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications*, *9*(1), 141–142.

Newey, W. K., & McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics*, *4*, 2111–2245.

Qin, J., Leung, D., & Shao, J. (2002). Estimation with survey data under nonignorable nonresponse or informative sampling. *Journal of the American Statistical Association*, *97*(457), 193–200.

Robins, J., & Ritov, Y. (1997). Toward a curse of dimensionality appropriate (coda) asymptotic theory for semi-parametric models. *Statistics in Medicine*, *16*(3), 285–319.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis* (Vol. 26). New York: CRC press.

Tang, G., Little, R. J., & Raghunathan, T. E. (2003). Analysis of multivariate missing data with nonignorable nonresponse. *Biometrika*, *90*(4), 747–764.

Wang, S., Shao, J., & Kim, J. K. (2014). An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica*, *24*(3), 1097–1116.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, *26*(4), 359–372.

Xia, Y., Tong, H., Li, W., & Zhu, L.-X. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(3), 363–410.

Zhao, J., & Shao, J. (2015). Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data. *Journal of the American Statistical Association*, *110*(512), 1577–1590.