Biometrika (2019), **106**, 4, *pp.* 875–888
Printed in Great Britain

doi: 10.1093/biomet/asz048
Advance Access publication 24 September 2019

Causal inference with confounders missing not at random

By S. YANG

Department of Statistics, North Carolina State University, 2311 Stinson Drive, Raleigh, North Carolina 27695, U.S.A.

syang24@ncsu.edu

L. WANG

Department of Statistical Sciences, University of Toronto, 100 St. George Street, Toronto, Ontario M5S 3G3, Canada

linbo.wang@utoronto.ca

and P. DING

Department of Statistics, University of California, 367 Evans Hall, Berkeley, California 94720, U.S.A.

pengdingpku@berkeley.edu

Summary

It is important to draw causal inference from observational studies, but this becomes challenging if the confounders have missing values. Generally, causal effects are not identifiable if the confounders are missing not at random. In this article we propose a novel framework for nonparametric identification of causal effects with confounders subject to an outcome-independent missingness, which means that the missing data mechanism is independent of the outcome, given the treatment and possibly missing confounders. We then propose a nonparametric two-stage least squares estimator and a parametric estimator for causal effects.

Some key words: Completeness; Identifiability; Ill-posed inverse problem; Integral equation; Outcome-independent missingness; Two-stage least squares estimator.

1. Introduction

Causal inference plays an important role in biomedical studies and social sciences. If all the confounders of the treatment-outcome relationship are observed, one can use standard techniques, such as propensity score matching, subclassification and weighting, to adjust for confounding (e.g., Rosenbaum & Rubin, 1983; Imbens & Rubin, 2015).

Much less work has been done on the case where confounders have missing values. Rosenbaum & Rubin (1984) and D'Agostino Jr & Rubin (2000) developed a generalized propensity score approach. Under a modified unconfoundedness assumption, they showed that adjusting for the missing pattern and the observed values of confounders removes all confounding bias, and hence the causal effects are identifiable. Moreover, the balancing property of the propensity score carries over to the generalized propensity score. Standard propensity score methods can therefore be used to estimate the causal effects. However, the modified unconfoundedness assumption implies that units may have different confounders depending on the missing pattern, which is often difficult to justify scientifically. An alternative approach assumes that the confounders are missing at random (Rubin, 1976). Under this assumption, both the full data distribution and the causal effects are identifiable, and multiple imputation can be used to obtain estimates of the causal effects (Rubin, 1987; Qu & Lipkovich, 2009; Crowe et al., 2010; Mitra & Reiter, 2011; Seaman & White, 2014). In practice, however, the missing pattern often depends on the missing values themselves, a scenario commonly known as missing not at random (Rubin, 1976). Multiple-imputation methods may fail to provide valid inference in this scenario. See Mattei (2009) for a comparison of various methods and Lu & Ashmead (2018) for a sensitivity analysis.

Causal inference with confounders missing not at random is challenging because neither the full data distribution nor the causal effects are identifiable without further assumptions. We consider a novel setting in which the confounders are subject to an outcome-independent missingness; that is, the missing data mechanism is independent of the outcome, given the treatment and possibly missing confounders. This outcome-independent missingness is plausible if the outcome happens after the covariate measurements and missing data indicators. To identify the causal effects in this setting, we formulate the identification problem as solving an integral equation, and show that the identification of the full data distribution is equivalent to the existence of a unique solution to an inverse problem. This new perspective allows us to establish a general condition for identifiability of the causal effects. Our condition generalizes existing results for discrete covariates and outcome (Ding & Geng, 2014). Motivated by the identification result, we develop a nonparametric two-stage least squares estimator by solving the sample analogue of the integral equation. To avoid the curse of dimensionality, we further develop parametric likelihood-based methods.

2. Set-up and assumptions

2.1. *Potential outcomes, causal effects and unconfoundedness*

We use potential outcomes to define causal effects (Neyman, 1923; Rubin, 1974). Suppose that the binary treatment is $A \in \{0, 1\}$, with 0 and 1 being the labels for the control and active treatments, respectively. Each level of treatment a corresponds to a possibly multi-dimensional potential outcome $Y(a)$, representing the outcome had the subject, possibly contrary to the fact, been given treatment a. The observed outcome is $Y = Y(A) = AY(1) + (1 - A)Y(0)$. Let $X = (X_1, \ldots, X_p)$ be a vector of p-dimensional pre-treatment covariates. We assume that a sample of size n consists of independent and identically distributed draws from the distribution of $\{A, X, Y(0), Y(1)\}$. The covariate-specific causal effect is $\tau(X) = E\{Y(1) - Y(0) \mid X\}$, and the average causal effect is $\tau = E\{Y(1) - Y(0)\} = E\{\tau(X)\}$. We focus on τ; a similar discussion applies to the average causal effect on the treated, $\tau_{\text{ATT}} = E\{Y(1) - Y(0) \mid A = 1\} = E\{\tau(X) \mid A = 1\}$. The following assumptions are standard in causal inference with observational studies (Rosenbaum & Rubin, 1983).

Assumption 1. We have that $\{Y(0), Y(1)\} \perp\!\!\!\perp A \mid X$.

Assumption 2. There exist constants c_1 and c_2 such that $0 < c_1 \leqslant e(X) \leqslant c_2 < 1$ almost surely, where $e(X) = \text{pr}(A = 1 \mid X)$ is the propensity score.

Under Assumptions 1 and 2, $\tau = E\{E(Y \mid A = 1, X) - E(Y \mid A = 0, X)\}$ is identifiable from the joint distribution of the observed data (A, X, Y). Rosenbaum & Rubin (1983) showed that $\{Y(0), Y(1)\} \perp\!\!\!\perp A \mid e(X)$, so adjusting for the propensity score removes all confounding. We can estimate τ through propensity score matching, subclassification or weighting.

2.2. *Confounders with missing values and the generalized propensity score*

We consider the case where X contains missing values. Let $R = (R_1, \dots, R_p)$ be the vector of missing indicators such that $R_j = 1$ if the jth component X_j is observed and 0 if X_j is missing. Let $\mathcal{R}$ be a subset of all possible values of R. We use 1_p to denote the p-vector of 1s and 0_p the p-vector of 0s. The missingness pattern $R = r \in \mathcal{R}$ partitions the covariates X into X_r and $X_{\bar{r}}$, the observed and missing parts of X, respectively. Using the standard notation, $X_R = X_{\text{obs}}$ and $X_{\bar{R}} = X_{\text{mis}}$ are the realized observed and missing covariates, respectively. For example, if $R_1 = 1$ and $R_j = 0$ for $j = 2, \dots, p$, then $X_R = X_1$ and $X_{\bar{R}} = (X_2, \dots, X_p)$. Assume that the full data are independent and identically distributed draws from $\{A, X, Y(0), Y(1), R\}$, and so the observed data are independent and identically distributed draws from (A, R, X_R, Y). Rosenbaum & Rubin (1984) introduced the following modified unconfoundedness assumption.

Assumption 3. We have that $\{Y(0), Y(1)\} \perp\!\!\!\perp A \mid (X_R, R)$.

Under Assumption 3, the generalized propensity score $e(X_R, R) = \text{pr}(A = 1 \mid X_R, R)$ plays the same role as the usual propensity score $e(X) = \text{pr}(A = 1 \mid X)$ in the settings without missing covariates. Rosenbaum & Rubin (1984) showed that adjusting for $e(X_R, R)$ balances (X_R, R) and removes all confounding on average. Their approach has the advantage of requiring no assumptions on the missing data mechanism of X for the identification of causal effects. However, Assumption 3 implies that a pre-treatment covariate can be a confounder when it is observed, but is not a confounder when it is missing; this is often hard to justify scientifically. Moreover, if the covariate measurement occurs after the treatment assignment, then R is a post-treatment variable affected by A. In this case, even if A is completely randomized, Assumption 3 is unlikely to hold when conditioning on the post-treatment variable R (Frangakis & Rubin, 2002).

2.3. *Missing data mechanisms of the confounders*

Without Assumption 3, identification of causal effects relies on alternative assumptions on the missing data mechanism. We now describe existing approaches under different missingness mechanisms of the confounders, the first of which is missing completely at random (Rubin, 1976).

Assumption 4 (Missing completely at random). We have that $R \perp\!\!\!\perp (A, X, Y)$.

Assumption 4 requires that the missingness of confounders be independent of all variables (A, X, Y). It implies $\tau = E\{\tau(X) \mid R = 1_p\}$ and thus justifies the complete-case analysis that uses only the units with fully observed confounders. This complete-case analysis is, however, inefficient as it discards all units with missing confounders. Moreover, confounders are rarely missing completely at random.

The second missingness mechanism is missing at random (Rubin, 1976).

Assumption 5 (Missing at random). We have that $R \perp\!\!\!\perp X \mid (A, Y)$.

Under Assumption 5, conditioning on the treatment and outcome, the missing mechanism of confounders is independent of the missing values themselves. Assumption 5 implies $f(A, X, Y) = f(A, Y)f(X \mid A, Y, R = 1_p)$, and therefore the joint distribution $f(A, X, Y)$ and its functionals, including τ, are all identifiable. Rubin (1976) showed that the missing data mechanism can be ignored in the likelihood-based and Bayesian inferences under Assumption 5. In this case, multiple imputation is a popular tool for causal inference (e.g., Qu & Lipkovich, 2009; Crowe et al., 2010; Mitra & Reiter, 2011; Seaman & White, 2014).

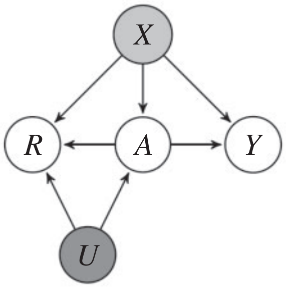


Fig. 1. A direct acyclic graph illustrating Assumptions 1 and 6; white nodes represent observed variables, the light grey node represents the variable with missing values, and the dark node represents an unmeasured variable U.

However, imputing the missing confounders based on $f(X_{\bar{R}} \mid X_R, A, Y) \propto f(X)f(A \mid X)f(Y \mid A, X)$ involves an outcome model in general (U.S. Department of Education, 2017), which is contrary to the suggestion of Rubin (2007) that the outcome should not be used in the design of an observational study. More importantly, missing at random is not plausible if the missing pattern depends on the missing values themselves. Instead, we consider the following missing data mechanism.

Assumption 6 (Outcome-independent missingness). We have that $R \perp\!\!\!\perp Y \mid (A, X)$.

Assumption 6 is plausible for prospective observational studies with covariates measured long before the outcome takes place (e.g., Hsu & Small, 2013; Hanna-Attisha et al., 2016). Figure 1 is a special causal diagram (Pearl, 1995) illustrating Assumptions 1 and 6. Graphically, A and Y have no common parents except for X, encoding Assumption 1, and R and Y have no common parents except A and X, encoding Assumption 6. Our framework allows for unmeasured common causes of R and A, as well as the dependence of R on the missing confounders $X_{\bar{R}}$. Moreover, it allows R to be a post-treatment variable affected by A. We give more graphical illustrations of Assumption 6 in the Supplementary Material.

We also make the following assumption to rule out degeneracy of the missing data mechanism.

Assumption 7. We have that $\text{pr}(R = 1_p \mid A, X, Y) > c_3 > 0$ almost surely for some constant c_3.

3. Nonparametric identification

3.1. *Identification strategy*

Assume that the distribution of (A, X, Y, R) is absolutely continuous with respect to some measure, with $f(A, X, Y, R)$ being the density or probability mass function. Under Assumptions 1 and 2, the key is to identify the joint distribution of $f(A, X, Y)$ because τ is its functional. The following identity relates the full data distribution to the observed data distribution:

$$f(A, X, Y, R = 1_p) = f(A, X, Y)\,\text{pr}(R = 1_p \mid A, X, Y). \quad (1)$$

The left-hand side of (1) is identifiable under Assumption 7. Therefore, the identification of $f(A, X, Y)$ relies on the identification of $\text{pr}(R = 1_p \mid A, X, Y)$. We now discuss how to identify $\text{pr}(R = 1_p \mid A, X, Y) = \text{pr}(R = 1_p \mid A, X)$ under Assumption 6.

3.2. *Integral equation representation*

Under Assumption 6, let

$$\xi_{ra}(X) = \frac{\mathrm{pr}(R = r \mid A = a, X, Y)}{\mathrm{pr}(R = 1_p \mid A = a, X, Y)} = \frac{\mathrm{pr}(R = r \mid A = a, X)}{\mathrm{pr}(R = 1_p \mid A = a, X)} \quad (a = 0, 1; r \in \mathcal{R}).$$

It then suffices to identify $\xi_{ra}(X)$, because it determines the missing data mechanism via

$$\mathrm{pr}(R = r \mid A = a, X, Y) = \frac{\mathrm{pr}(R = r \mid A = a, X, Y)}{\sum_{r' \in \mathcal{R}} \mathrm{pr}(R = r' \mid A = a, X, Y)} = \frac{\xi_{ra}(X)}{\sum_{r' \in \mathcal{R}} \xi_{r'a}(X)}. \tag{2}$$

The following theorem shows that $\xi_{ra}(X)$ is a key term connecting the observed data distribution $f(A, X_r, Y, R = r)$ and the complete-case distribution $f(A, X, Y, R = 1_p)$. Throughout the paper, $\nu(\cdot)$ denotes a generic measure, such as the Lebesgue measure for a continuous variable or the counting measure for a discrete variable.

Theorem 1. *Under Assumption 6, for any r and a, the following integral equation holds:*

$$f(A = a, X_r, Y, R = r) = \int \xi_{ra}(X) f(A = a, X, Y, R = 1_p)\, \mathrm{d}\nu(X_{\bar{r}}). \tag{3}$$

Proof. The result follows because the observed data distribution is the complete-data distribution averaged over the missing data:

$$\begin{aligned} f(A = a, X_r, Y, R = r) &= \int f(A = a, X, Y, R = r)\, \mathrm{d}\nu(X_{\bar{r}}) \\ &= \int \frac{\mathrm{pr}(R = r \mid A = a, X, Y)}{\mathrm{pr}(R = 1_p \mid A = a, X, Y)} f(A = a, X, Y, R = 1_p)\, \mathrm{d}\nu(X_{\bar{r}}) \\ &= \int \xi_{ra}(X) f(A = a, X, Y, R = 1_p)\, \mathrm{d}\nu(X_{\bar{r}}). \end{aligned}$$ □

Theorem 1 is the basis of our identification analysis. In (3), $f(A = a, X_r, Y, R = r)$ and $f(A = a, X, Y, R = 1_p)$ are identifiable from the observed data. We have thus turned the identification of $\xi_{ra}(X)$ into the problem of solving for $\xi_{ra}(X)$ from (3). This requires additional technical assumptions, given below.

3.3. *Bounded completeness and identification of the joint distribution*

To motivate our identification conditions, we first consider the case of discrete X and Y, so that (3) becomes a linear system. To solve for $\xi_{ra}(X)$ from (3), we need the linear system to be nondegenerate.

Proposition 1. *Under Assumption 6, suppose that X and Y are discrete, with $X_j \in \{x_{j1}, \ldots, x_{jJ_j}\}$ for $j = 1, \ldots, p$ and $Y \in \{y_1, \ldots, y_K\}$. Let $q = J_1 \times \cdots \times J_p$, and let Θ_a be a $K \times q$ matrix with the kth row being $f(X, y_k, R = 1_p, A = a)$ evaluated at all possible values of X. The distribution of (A, X, Y, R) is identifiable if* $\mathrm{Rank}(\Theta_a) = q$ *for $a = 0, 1$.*

We relegate the proof to the Supplementary Material. For the special case of a binary X and a discrete Y, the rank condition in Proposition 1 is equivalent to $X \not\perp\!\!\!\perp Y \mid (A = a, R = 1)$ for $a = 0$ and 1, which is testable based on the observed data (Ding & Geng, 2014). For general cases, we need to extend the rank condition that ensures the unique existence of $\xi_{ra}(X)$. We use the notion of bounded completeness for general X and Y, which is related to the concept of a complete statistic (Lehmann & Scheffé, 1950; Newey & Powell, 2003). Below, we say that a function $g(x)$ is bounded in $\mathcal{L}_1$-metric if $\sup_x |g(x)| \leqslant c$ for some $0 < c < \infty$.

DEFINITION 1. *A function $f(X, Y)$ is bounded complete in Y if $\int g(X) f(X, Y)\, \mathrm{d}\nu(X) = 0$ implies $g(X) = 0$ almost surely for any measurable function $g(X)$ bounded in $\mathcal{L}_1$-metric.*

D'Haultfoeuille (2011) gave sufficient conditions for bounded completeness. Bounded completeness has also appeared in other identification analyses, such as nonparametric instrumental variable regression models (Darolles et al., 2011) and measurement error models (An & Hu, 2012).

We invoke the following assumption motivated by Theorem 1 and Definition 1.

Assumption 8. The joint distribution $f(A = a, X, Y, R = 1_p)$ is bounded complete in Y for $a = 0, 1$.

Remark 1. When X and Y are discrete with finite supports, Assumption 8 is equivalent to the rank condition in Proposition 1. For continuous X and Y, Assumption 8 requires that the dimension of Y be at least as large as the dimension of X in general. Moreover, Assumption 8 implies Assumption 2. We give more details for these results in the Supplementary Material.

Under Assumption 7, Assumption 8 is sufficient to ensure the existence and uniqueness of $\xi_{ra}(X)$ from (3). We state the result in the following theorem.

THEOREM 2. *Under Assumptions 6–8, the distribution of (A, X, Y, R) is identifiable.*

Proof. Suppose that $\xi_{ra}^{(1)}(X)$ and $\xi_{ra}^{(2)}(X)$ are two solutions to (3):

$$f(A = a, X_r, Y, R = r) = \int \xi_{ra}^{(k)}(X) f(A = a, X, Y, R = 1_p)\, \mathrm{d}\nu(X_{\bar{r}}) \quad (k = 1, 2),$$

implying that $\int \{\xi_{ra}^{(1)}(X) - \xi_{ra}^{(2)}(X)\} f(A = a, X, Y, R = 1_p)\, \mathrm{d}\nu(X_{\bar{r}}) = 0$. Integrating this identity with respect to X_r gives

$$\int \{\xi_{ra}^{(1)}(X) - \xi_{ra}^{(2)}(X)\} f(A = a, X, Y, R = 1_p)\, \mathrm{d}\nu(X) = 0.$$

Assumption 7 implies that $\xi_{ra}(X)$ is bounded in $\mathcal{L}_1$-metric, which further implies that $\xi_{ra}^{(1)}(X) - \xi_{ra}^{(2)}(X)$ is bounded in $\mathcal{L}_1$-metric. Under Assumption 8, Definition 1 implies that $\xi_{ra}^{(1)}(X) - \xi_{ra}^{(2)}(X) = 0$ almost surely. Therefore, (3) has a unique solution $\xi_{ra}(X)$. Based on the definition of $\xi_{ra}(X)$, we can identify $\mathrm{pr}(R = 1_p \mid A, X, Y)$ by (2). Finally, we identify $f(A, X, Y)$ through (1) as $f(A, X, Y) = f(A, X, Y, R = 1_p)/\mathrm{pr}(R = 1_p \mid A, X, Y)$. □

If the distribution of (A, X, Y) is identifiable, we can use a standard argument to show that τ and τ_{ATT} are identifiable under Assumption 1. In the next subsection we give explicit identification formulas for τ and τ_{ATT}, which form the basis for constructing the nonparametric estimator.

3.4. *Nonparametric identification formulas for average causal effects*

Under Assumptions 1 and 6–8, we can identify τ and τ_{ATT} in two steps. First,

$$\tau(X) = E(Y \mid A = 1, X) - E(Y \mid A = 0, X) \tag{4}$$

$$= E(Y \mid A = 1, X, R = 1_p) - E(Y \mid A = 0, X, R = 1_p), \tag{5}$$

where (4) follows from Assumption 1 and (5) follows from Assumption 6. Therefore, we can identify $\tau(X)$ using a complete-case analysis based on (5).

Second, under Assumptions 6–8, Theorem 2 shows that the distribution of (A, X, Y, R) is identifiable, which implies that the marginal distribution of X, $f(X)$, and the conditional distribution of X, $f(X \mid A = 1)$, are also identifiable. Therefore, both $\tau = E\{\tau(X)\}$ and $\tau_{\text{ATT}} = E\{\tau(X) \mid A = 1\}$ are identifiable. The following theorem summarizes these results and gives the explicit formulas.

THEOREM 3. *Under Assumptions 1 and 6–8, the average causal effect τ is identified by*

$$\tau = \sum_{a=0}^{1} \int \tau(X) \frac{f(A = a, X, R = 1_p)}{\text{pr}(R = 1_p \mid A = a, X)} \, \text{d}\nu(X), \tag{6}$$

and the average treatment effect on the treated, τ_{ATT}, is identified by

$$\tau_{\text{ATT}} = \int \tau(X) \frac{f(X, R = 1_p \mid A = 1)}{\text{pr}(R = 1_p \mid A = 1, X)} \, \text{d}\nu(X), \tag{7}$$

where $\tau(X)$ is identified by (5), $\text{pr}(A = a, R = 1_p)$ and $f(A = a, X, R = 1_p)$ depend only on the observed data, and $\text{pr}(R = 1_p \mid A = a, X)$ can be identified from (2) and (3) for $a = 0, 1$.

Proof. First, we can identify the conditional distribution of X given $A = a$ by

$$f(X \mid A = a) = \frac{f(X, R = 1_p \mid A = a)}{\text{pr}(R = 1_p \mid A = a, X)} \quad (a = 0, 1).$$

Averaging $\tau(X)$ over $f(X \mid A = 1)$ yields the identification formula (7).

Second, we can identify the marginal distribution of X by

$$f(X) = \sum_{a=0}^{1} f(A = a, X) = \sum_{a=0}^{1} \frac{f(A = a, X, R = 1_p)}{\text{pr}(R = 1_p \mid A = a, X)}.$$

Averaging $\tau(X)$ over the above distribution gives the identification formula (6). □

4. ESTIMATION OF THE AVERAGE CAUSAL EFFECT

4.1. *Nonparametric two-stage least squares estimator*

Theorem 3 gives the nonparametric identification formulae at the population level. Based on (6), we propose a nonparametric two-stage least squares estimator of τ with finite samples $(A_i, R_i, X_{R_i}, Y_i)_{i=1}^{n}$. Estimation of τ_{ATT} is similar in spirit and hence omitted. We can use standard nonparametric or machine learning methods to estimate $\tau(X)$, $\text{pr}(A = a, R = 1_p)$ and

$f(X \mid A=a, R=1_p)$; let $\hat{\tau}(X)$, $\hat{\text{pr}}(A=a, R=1_p)$ and $\hat{f}(X \mid A=a, R=1_p)$ denote the respective estimators. Therefore, the key is to estimate $\text{pr}(R=1_p \mid A=a, X)$ or, equivalently, $\xi_{ra}(X)$ based on (3).

In the first stage, we obtain $\hat{f}(X_r, Y, R=r \mid A=a)$ and $\hat{f}(X, Y, R=1_p \mid A=a)$ as the nonparametric sample analogues of $f(X_r, Y, R=r \mid A=a)$ and $f(X, Y, R=1_p \mid A=a)$. Substituting these estimates into (3) leads to

$$\hat{f}(X_r, Y, R=r \mid A=a) = \int \xi_{ra}(X)\hat{f}(X, Y, R=1_p \mid A=a)\, \mathrm{d}\nu(X_{\bar{r}}), \tag{8}$$

which is a Fredholm integral equation of the first kind. Solving (8) presents several challenges. First, although Theorem 2 states that the population equation (3) has a unique solution, the sample equation (8) may not have a unique solution. Second, $\xi_{ra}(X)$ is an infinite-dimensional parameter, and its estimation often relies on some approximation. Third, solving for $\xi_{ra}(X)$ from (8) is an ill-conditioned problem, in the sense that even a slight perturbation of $\hat{f}(X_r, Y, R=r \mid A=a)$ and $\hat{f}(X, Y, R=1_p \mid A=a)$ can lead to a large variation in the solution for $\xi_{ra}(X)$. As a result, replacing $f(X_r, Y, R=r \mid A=a)$ and $f(X, Y, R=1_p \mid A=a)$ in (3) by their consistent estimators does not necessarily yield a consistent estimator of $\xi_{ra}(X)$ (Darolles et al., 2011).

To deal with these issues, we use a series approximation (Kress et al., 1999; Newey & Powell, 2003) in the second stage. Let the set $\mathcal{H}_J = \{h^j(X) = \exp(-X^{\mathrm{T}}X)X^{\lambda_j} : j=1,\ldots,J\}$ form a Hermite polynomial basis, where $X^{\lambda_j} = X_1^{\lambda_{j1}} \cdots X_p^{\lambda_{jp}}$ with $\lambda_j = (\lambda_{j1}, \ldots, \lambda_{jp})$ and $|\lambda_j| = \sum_{l=1}^{p} \lambda_{jl}$ increasing in j. Let $\tilde{X} = \Sigma^{-1/2}(X-\mu)$ be a standardized version of X, where μ and Σ are a constant vector and matrix. We approximate $\xi_{ra}(X)$ by $\xi_{ra}(X) \approx \sum_{j=1}^{J} \beta_{ra}^j h^j(\tilde{X})$. Thus, for each missing pattern $R=r$, we approximate (3) by

$$\begin{aligned} f(X_r, Y, R=r \mid A=a) &\approx \sum_{j=1}^{J} \beta_{ra}^j \int h^j(\tilde{X}) f(X, Y, R=1_p \mid A=a)\, \mathrm{d}\nu(X_{\bar{r}}) \\ &= \sum_{j=1}^{J} \beta_{ra}^j H_{ra}^j(X_r, Y) f(X_r, Y, R=1_p \mid A=a), \end{aligned} \tag{9}$$

where the conditional expectation $H_{ra}^j(X_r, Y) = E\{h^j(\tilde{X}) \mid A=a, X_r, Y, R=1_p\}$ is over the distribution $f(X_{\bar{r}} \mid A=a, X_r, Y, R=1_p)$.

We need the empirical versions of $H_{ra}^j(X_r, Y)$ and $f(X_r, Y, R=1_p \mid A=a)$ for estimation. First, for unit i, let $\hat{H}_{ra,i}^j = \hat{E}\{h^j(\tilde{X}) \mid A_i=a, X_{r,i}, Y_i, R_i=1_p\}$ be a nonparametric estimator of the conditional expectation. Second, we obtain $\hat{f}(X_r, Y, R=1_p \mid A=a)$, a nonparametric estimator of $f(X_r, Y, R=1_p \mid A=a)$. Although we obtain these estimators based on the complete cases, we still need to partition the confounders into $(X_r, X_{\bar{r}})$ based on the missing pattern $R=r$. Because the sample version of the approximation (9) is linear, we can estimate the β_{ra}^j by minimizing the residual sum of squares

$$\sum_{i=1}^{n} I(R_i=r)\left\{\hat{f}(X_{r,i}, Y_i, R_i=r \mid A_i=a) - \sum_{j=1}^{J} \beta_{ra}^j \hat{H}_{ra,i}^j \hat{f}(X_{r,i}, Y_i, R_i=1_p \mid A_i=a)\right\}^2. \tag{10}$$

To ensure the estimates from (10) are well-behaved asymptotically, we need a large number of observations for each pattern $r \in \mathcal{R}$. To solve the ill-conditioned problem, we restrict the parameter space of $\xi_{ra}(X)$ to a compact space, which effectively regularizes the problem, making it well-posed. Given the approximation of $\xi_{ra}(X)$, we require the vector of coefficients β_{ra}, the concatenation of $(\beta_{ra}^1, \ldots, \beta_{ra}^J)$, to satisfy $\beta_{ra}^{\mathrm{T}} \Lambda \beta_{ra} \leqslant B$, where Λ is a positive-definite $J \times J$ matrix and B is a positive constant. Therefore, we propose to estimate β_{ra} by minimizing (10) subject to the constraint $\beta_{ra}^{\mathrm{T}} \Lambda \beta_{ra} \leqslant B$. More details of the regularization are presented in the Supplementary Material.

We then estimate $\xi_{ra}(X)$ and the probability $\mathrm{pr}(R = 1_p \mid A = a, X)$ by

$$\hat{\xi}_{ra}(X) = \sum_{j=1}^{J} \hat{\beta}_{ra}^{j} h^{j}(\tilde{X}), \quad \hat{\mathrm{pr}}(R = 1_p \mid A = a, X) = \left\{ 1 + \sum_{r \neq 1_p} \hat{\xi}_{ra}(X) \right\}^{-1}$$

and finally estimate τ by

$$\hat{\tau} = \sum_{a=0}^{1} \hat{\mathrm{pr}}(A = a, R = 1_p) \int \hat{\tau}(X) \frac{\hat{f}(X \mid A = a, R = 1_p)}{\hat{\mathrm{pr}}(R = 1_p \mid A = a, X)} \, \mathrm{d}\nu(X). \tag{11}$$

We now comment on some subtle technical issues in implementing the above estimator. First, we standardize the confounders by $\tilde{X} = \Sigma^{-1}(X - \mu)$ for numerical stability. We choose μ and Σ to be the mean and covariance matrix of confounders for the complete cases. This choice is innocuous because $\mathcal{H}_J$ remains the same for other values of μ and Σ. Second, we use the importance sampling technique to approximate the integral in (11), because it is difficult to directly sample from the nonparametric density estimators. Third, we use the bootstrap to construct confidence intervals. Newey (1997) proposed a relatively simple variance estimation approach that treats the nonparametric estimators as if they were parametric given the fixed tuning parameters. For all bootstrap samples we use the same tuning parameters, such as the smoothing parameter in the smoothing splines and the bandwidth in the kernel density estimator. In the Supplementary Material, we give more technical details and illustrate the procedure with an example involving a scalar confounder.

4.2. *Parametric estimation: likelihood-based and Bayesian inferences*

The nonparametric estimator above suffers from the curse of dimensionality. We propose a parametric approach for moderate- or high-dimensional covariates. Let $Z_i = (A_i, X_i, Y_i, R_i)$ be the complete data and $Z_{R,i} = (A_i, R_i, X_{R,i}, Y_i)$ the observed data for unit i. The complete-data likelihood is $L(\theta \mid Z_1, \ldots, Z_n) = \prod_{i=1}^{n} f(Z_i; \theta)$, where $\theta = (\alpha, \beta_0, \beta_1, \eta_0, \eta_1, \lambda)$ and

$$f(Z_i; \theta) = \mathrm{pr}(R_i \mid A_i, X_i; \eta_{A_i}) f(Y_i \mid A_i, X_i; \beta_{A_i}) \, \mathrm{pr}(A_i \mid X_i; \alpha) f(X_i; \lambda). \tag{12}$$

The observed-data likelihood is

$$L(\theta \mid Z_{R,1}, \ldots, Z_{R,n}) = \prod_{i=1}^{n} \left\{ \sum_{r \in \mathcal{R}} I(R_i = r) \int f(Z_i; \theta) \, \mathrm{d}\nu(X_{\bar{r},i}) \right\}.$$

Under Assumptions 6–8 as in Theorem 2, θ is identifiable if the parametric models in (12) are not overparameterized. The bounded completeness condition holds for many commonly used models, such as generalized linear models and a location family of absolutely continuous distributions with compact support; see Blundell et al. (2007), Hu & Shiu (2018) and the Supplementary Material for additional examples. Moreover, parametric assumptions can further help to identify the model parameters even without the bounded completeness assumption. We illustrate this later.

We first discuss likelihood-based inference. Let $\tau(X_i;\theta) = E(Y_i \mid A_i = 1, X_i; \beta_1) - E(Y_i \mid A_i = 1, X_i; \beta_0)$ be the covariate-specific average causal effect, and let

$$\hat{\tau}(\theta) = n^{-1}\sum_{i=1}^{n} \tau(X_i;\theta), \quad \tau = \tau(\theta) = E\{\tau(X_i;\theta)\} = E\{\hat{\tau}(\theta)\}.$$

We first obtain the maximum likelihood estimate $\hat{\theta}$ and then estimate τ by $\tau(\hat{\theta})$. The formula $\tau(\theta)$ involves integrating over the distribution of the confounders. To avoid this complexity, we use $\hat{\tau}(\hat{\theta})$ to estimate τ. The bootstrap can be used to construct confidence intervals.

Next, we discuss Bayesian inference. Suppose that we can simulate the posterior distributions of the missing confounders and the parameter θ. These further induce posterior distributions of $\hat{\tau}(\theta)$ and $\tau = \tau(\theta)$. Technically, the posterior distribution of $\hat{\tau}(\theta)$ is different from that of τ. The former depends on the observed confounder values, but the latter does not. See Ding & Li (2018) for more discussion.

We give more computational details in the Supplementary Material, including a fractional imputation algorithm (Yang & Kim, 2016) and a Bayesian procedure for a parametric model. In future work we will develop multiple-imputation methods under Assumptions 6–8. From (12), we need to use both treatment and outcome models in the imputation step as in the full Bayesian procedure.

5. Simulation

5.1. *Design of the simulation*

We use simulation to compare our estimators with existing ones. First, we consider the unadjusted estimator, which is the simple difference-in-means of the outcomes between the treated and control groups. We use it to quantify the degree of confounding. Second, we consider the generalized propensity score weighting estimator, with the generalized propensity scores estimated separately by a logistic regression for each missing pattern (Rosenbaum & Rubin, 1984). Third, we consider three multiple-imputation estimators. The first uses the outcome in the imputation model, but the second does not (Mitra & Reiter, 2011); the third estimator uses the missingness pattern in the propensity score model (Qu & Lipkovich, 2009).

We evaluate the finite-sample performance of these estimators with the missingness of confounders satisfying Assumption 6. In the first setting, in § 5.2, one confounder has missing values and we investigate the performance of the proposed nonparametric estimator and the sensitivity to the choice of tuning parameters. In the second setting, in § 5.3, multiple confounders have missing values and we investigate the performance of the proposed parametric estimator. In each setting, we choose the sample size to be $n = 400$, 800 and 1600, and we generate 2000 Monte Carlo samples for each sample size. For the multiple-imputation estimators, we generate 100 imputed datasets. For all estimators, we use the bootstrap with 500 replicates to estimate the variances.

5.2. *One confounder subject to missingness*

The confounders $X_i = (X_{1i}, X_{2i})$ follow $X_{1i} \sim N(1,1)$ and $X_{2i} \sim \text{Ber}(0.5)$. The potential outcomes follow $Y_i(0) = 0.5 + 2X_{1i} + X_{2i} + \epsilon_i(0)$ and $Y_i(1) = 3X_{1i} + 2X_{2i} + \epsilon_i(1)$, where

Table 1. *Simulation results: bias* $(\times 10^{-2})$ *and variance* $(\times 10^{-3})$ *of the point estimator of* τ, *variance estimate* $(\times 10^{-3})$, *and coverage* (%) *of* 95% *confidence intervals*

Method	Bias	Var	VE	Cvg	Bias	Var	VE	Cvg	Bias	Var	VE	Cvg
(a) Comparing the nonparametric estimator with existing estimators												
	$n = 400$				$n = 800$				$n = 1600$			
Unadj	−127.5	77.4	73.7	0.3	−127.4	38.0	37.5	0.0	−127.2	17.5	18.6	0.0
GPSW	−55.1	42.4	44.2	22.2	−54.9	20.9	20.7	5.8	−54.4	9.5	9.9	0.4
MI1	41.5	35.4	36.7	40.6	41.0	15.5	17.2	9.5	40.8	7.6	8.3	0.5
MI2	−10.8	60.0	63.8	91.4	−9.2	28.8	30.8	91.4	−9.1	13.7	14.9	86.6
MIMP	29.3	73.5	71.5	83.7	28.5	33.7	32.6	65.0	28.3	14.9	16.0	30.8
NonPara	1.2	19.4	18.8	95.1	0.9	9.6	8.1	95.2	0.8	3.9	3.8	94.9
(b) Comparing the parametric estimator with existing estimators												
	$n = 400$				$n = 800$				$n = 1600$			
Unadj	32.2	85.2	85.8	81.5	32.2	44.3	42.9	65.8	31.9	20.3	21.6	43.1
GPSW	8.4	174.6	246.1	97.2	8.8	84.2	94.2	94.9	8.3	40.0	44.0	92.4
MI1	7.7	180.5	238.0	96.1	7.1	93.5	106.4	95.2	6.9	47.5	54.8	93.4
MI2	3.0	162.1	209.9	97.3	3.1	84.2	94.1	95.8	2.6	42.8	49.1	94.6
MIMP	12.9	177.0	239.2	95.7	12.2	93.9	107.5	93.8	12.1	47.4	55.0	91.8
Para	1.6	95.4	95.4	95.3	0.4	48.3	48.0	95.0	0.0	23.0	24.2	95.4

Var, variance of the point estimator of τ; VE, variance estimate; Cvg, coverage of 95% confidence intervals; Unadj, the unadjusted estimator; GPSW, the generalized propensity score weighting estimator; NonPara, the proposed nonparametric estimator; Para, the proposed parametric estimator; for the multiple-imputation estimators, MI1 uses the outcome in the imputation, MI2 does not use the outcome in the imputation, and MIMP is the multiple-imputation missingness pattern method of Qu & Lipkovich (2009).

$\epsilon_i(0) \sim N(0,1)$ and $\epsilon_i(1) \sim N(0,1)$. The average causal effect τ is 1. The treatment indicator A_i follows $\text{Ber}(\pi_i)$, where $\text{logit}(\pi_i) = 1.25 - 0.5X_{1i} - 0.5X_{2i}$. The missing indicator of X_{1i}, R_{1i}, follows $\text{Ber}(p_i)$, where $\text{logit}(p_i) = -2 + 2X_{1i} + A_i(1.5 + X_{2i})$. The average response rate is about 67%. Other variables do not have missing values.

For the proposed nonparametric estimator, we estimate $\hat{\tau}(X)$ using cubic splines with five knots and estimate the density functions using kernel-based estimators with the Gaussian kernel. We use ten-fold crossvalidation to choose the smoothing parameters in the smoothing spline estimator and the bandwidths in the kernel-based estimators. For $\hat{\xi}_{ra}(X)$, we choose $J = 5$ Hermite polynomial basis functions and $B = 50$ as the bound for regularization.

Table 1(a) compares the nonparametric estimator with the existing estimators. The unadjusted estimator, the propensity score weighting estimator and multiple-imputation estimators are biased. As a result, the coverage rates of the confidence intervals for these methods are quite poor. Our proposed method has negligible biases and good coverages, with variances decreasing with the sample size.

To assess the sensitivity of the nonparametric estimator to the choice of the tuning parameters J and B, we specify a 4×3 design with $(J, B) \in \{(3, 50), (3, 100), (5, 50), (5, 100)\}$ and $n \in \{400, 800, 1600\}$. Table 2 shows the mean squared errors. For each (J, B), the mean squared error decreases with the sample size. The mean squared error decreases with J, is relatively insensitive to the choice of B, and remains small across all cases.

5.3. *Multiple confounders subject to missingness*

Let $X_i = (X_{1i}, \ldots, X_{6i})$. We generate X_{1i} and X_{2i} from $N(1, 1)$, X_{3i} and X_{4i} from $\{\text{Ber}(0.5) - 0.5\}/0.5$, $X_{5i} = X_{1i} + X_{2i} + X_{3i} + X_{4i} + \epsilon_{5i}$ with $\epsilon_{5i} \sim N(0, 1)$, and X_{6i} from $\text{Ber}(p_{6i})$ with $\text{logit}(p_{6i}) = -X_{5i}$. The potential outcomes follow $Y_i(0) = (1, X_i^{\mathrm{T}})\beta_0 + \epsilon_i(0)$ and $Y_i(1) =$

Table 2. *Simulation results for different tuning parameters: mean squared errors* ($\times 10^{-3}$) *of the proposed estimator of* τ *for different choices of* (J, B) *based on* 2000 *Monte Carlo samples*

(J, B)	$n = 400$	$n = 800$	$n = 1600$
$(3, 50)$	26.8	13.9	8.3
$(3, 100)$	27.0	14.1	8.7
$(5, 50)$	19.5	9.7	4.1
$(5, 100)$	21.3	10.2	4.5

$(1, X_i^{\mathrm{T}})\beta_1 + \epsilon_i(1)$, where $\beta_0 = (-1.5, 1, -1, 1, -1, 1, 1)^{\mathrm{T}}$, $\beta_1 = (0, -1, 1, -1, 1, -1, -1)^{\mathrm{T}}$, $\epsilon_i(0) \sim N(0, 1)$ and $\epsilon_i(1) \sim N(0, 1)$. The average treatment effect is $\tau = -0.5$. The treatment indicator A_i follows $\mathrm{Ber}(\pi_i)$, where $\mathrm{logit}(\pi_i) = (1, X_i^{\mathrm{T}})\alpha$ and $\alpha = 0.5 \times (2, 1, 1, 1, 1, -2, -2)^{\mathrm{T}}$. Covariates X_{5i} and X_{6i} have missing values, but the other variables do not. The missingness pattern for X_{5i} and X_{6i}, $R_i = (R_{5i}, R_{6i}) \in \{(11), (10), (01), (00)\}$, follows a multinomial distribution with parameters $(p_{11,i}, p_{10,i}, p_{01,i}, p_{00,i})$ where

$$\mathrm{logit}(p_{11,i}) = [1 + 3\exp\{(1, A_i, X_i^{\mathrm{T}})\eta\}]^{-1}, \quad \mathrm{logit}(p_{kl,i}) = [\exp\{-(1, A_i, X_i^{\mathrm{T}})\eta\} + 3]^{-1}$$

for $kl \in \{10, 01, 00\}$, with $\eta = 0.25 \times (-4, 1, 1, 1, 1, 1, -1, -1)^{\mathrm{T}}$. The average percentages of these missingness patterns are about 49%, 17%, 17% and 17%, respectively.

Table 1(b) compares the parametric maximum likelihood estimator with the existing estimators. The unadjusted estimator has large biases due to confounding. The multiple-imputation estimators have large biases, although the coverages of confidence intervals seem good due to the overestimation of variances. In contrast, our estimator has negligible biases and good coverages.

6. Application

6.1. *The causal effect of smoking on blood lead level*

We use a dataset from the 2015–2016 U.S. National Health and Nutrition Examination Survey to estimate the causal effect of smoking on blood lead level (Hsu & Small, 2013). The dataset includes 2949 adults, consisting of 1102 smokers, denoted by $A = 1$, and 1847 nonsmokers, denoted by $A = 0$. All subjects were at least 15 years old and had no tobacco use besides cigarette smoking in the previous five days. The outcome Y is the lead level in blood, ranging from 0.05 to 23.51 μg/dl. The confounders X include the income-to-poverty level ratio, age and gender. The income-to-poverty level ratio has missing values, but the other variables do not. The missingness of income-to-poverty level is likely to be not at random because subjects with high incomes may be less likely to disclose their income information (Davern et al., 2005). It is plausible that Assumption 6 holds, i.e., that this missingness is unrelated to the blood lead level after controlling for income information. The missing rate of income-to-poverty level is 14.0% for smokers and 15.2% for nonsmokers. We apply the proposed procedure to obtain estimates separately for groups stratified by age and gender, and then average over the empirical distribution of age and gender.

Table 3(a) shows the results. Note the substantial differences in point estimates between our estimator and the competitors, illustrating the impact of the missing data assumption on causal inference in the presence of missing confounders. In contrast to the existing estimators, our estimator is better able to handle the confounders missing not at random. Based on the nonparametric estimator, smoking increases blood lead level by 0.20 μg/dl on average.

Table 3. *Results from the analysis of datasets: point estimate, standard error by the bootstrap, and* 95% *confidence interval*

	Est	SE	95% CI		Est	SE	95% CI
(a) The causal effect of smoking on blood lead level in § 6.1							
Unadj	0.44	0.05	(0.35, 0.54)	MI1	0.34	0.05	(0.25, 0.44)
PSW	0.12	0.05	(0.02, 0.22)	MI2	0.35	0.05	(0.25, 0.44)
NonPara	0.20	0.07	(0.05, 0.36)	MIMP	0.35	0.05	(0.25, 0.44)
(b) The causal effect of education on general health satisfaction in § 6.2							
Unadj	−0.57	0.034	(−0.64, −0.51)	MI1	−0.24	0.057	(−0.36, −0.13)
GPSW	−0.25	0.054	(−0.36, −0.14)	MI2	−0.26	0.057	(−0.38, −0.15)
Para	−0.32	0.051	(−0.41, −0.21)	MIMP	−0.23	0.057	(−0.34, −0.11)

Est, point estimate; SE, standard error; CI, confidence interval; Unadj, the unadjusted estimator; GPSW, the generalized propensity score weighting estimator; NonPara, the proposed nonparametric estimator; Para, the proposed parametric estimator; for the multiple-imputation estimators, MI1 uses the outcome in the imputation, MI2 does not use the outcome in the imputation, and MIMP is the multiple-imputation missingness pattern method of Qu & Lipkovich (2009).

6.2. *The causal effect of education on general health satisfaction*

We use a dataset from the 2015–2016 U.S. National Health and Nutrition Examination Survey to estimate the average causal effect of education on general health satisfaction. The dataset includes 4845 subjects. Among them, 76% have at least high school education, denoted by $A = 1$, and 24% do not, denoted by $A = 0$. The outcome Y is the general health satisfaction score, which ranges from 1 to 5, with lower values indicating greater satisfaction. The observed outcomes have mean 2.88 and standard deviation 0.96. The confounders X include age, gender, race, marital status, income-to-poverty level ratio, and an indicator of ever having risk of prediabetes. The income-to-poverty level and prediabetes risk variables have missing values, whereas the other variables do not. The missingness of the income-to-poverty level ratio and the prediabetes risk variable is likely to be related to the missing values themselves. It is plausible that this missingness is unrelated to the outcome value conditioning on the treatment and confounders.

Table 3(b) reports the results. Although qualitatively all estimators show that education is beneficial in improving general health satisfaction, differences can be observed in the point estimates of our estimator and the competitors. This illustrates the impact of the missing data assumption on causal inference with missing confounders. Based on the parametric estimator, education improves general health satisfaction by 0.32 on average.

Acknowledgement

Yang was supported in part by Oak Ridge Associated Universities, the U.S. National Science Foundation and National Institutes of Health. Wang was supported in part by the Natural Sciences and Engineering Research Council of Canada. Ding was supported in part by the U.S. National Science Foundation and the Institute of Education Sciences. The authors thank Professor Eric Tchetgen Tchetgen for valuable discussions, Professor Xiaohong Chen for useful references and the associate editor and two reviewers for helpful comments.

Supplementary material

Supplementary material available at *Biometrika* online includes additional proofs, further discussions on the nonparametric and parametric estimators, and additional simulations.

References

An, Y. & Hu, Y. (2012). Well-posedness of measurement error models for self-reported data. *J. Economet.* **168**, 259–69.

Blundell, R., Chen, X. & Kristensen, D. (2007). Semi-nonparametric IV estimation of shape-invariant Engel curves. *Econometrica* **75**, 1613–69.

Crowe, B. J., Lipkovich, I. A. & Wang, O. (2010). Comparison of several imputation methods for missing baseline data in propensity scores analysis of binary outcome. *Pharm. Statist.* **9**, 269–79.

D'Agostino Jr, R. B. & Rubin, D. B. (2000). Estimating and using propensity scores with partially missing data. *J. Am. Statist. Assoc.* **95**, 749–59.

Darolles, S., Fan, Y., Florens, J.-P. & Renault, E. (2011). Nonparametric instrumental regression. *Econometrica* **79**, 1541–65.

Davern, M., Rodin, H., Beebe, T. J. & Call, K. T. (2005). The effect of income question design in health surveys on family income, poverty and eligibility estimates. *Health Serv. Res.* **40**, 1534–52.

D'Haultfoeuille, X. (2011). On the completeness condition in nonparametric instrumental problems. *Economet. Theory* **27**, 460–71.

Ding, P. & Geng, Z. (2014). Identifiability of subgroup causal effects in randomized experiments with nonignorable missing covariates. *Statist. Med.* **33**, 1121–33.

Ding, P. & Li, F. (2018). Causal inference: A missing data perspective. *Statist. Sci.* **33**, 214–37.

Frangakis, C. E. & Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58**, 21–9.

Hanna-Attisha, M., LaChance, J., Sadler, R. C. & Champney Schnepp, A. (2016). Elevated blood lead levels in children associated with the flint drinking water crisis: A spatial analysis of risk and public health response. *Am. J. Public Health* **106**, 283–90.

Hsu, J. Y. & Small, D. S. (2013). Calibrating sensitivity analyses to observed covariates in observational studies. *Biometrics* **69**, 803–11.

Hu, Y. & Shiu, J.-L. (2018). Nonparametric identification using instrumental variables: Sufficient conditions for completeness. *Economet. Theory* **34**, 659–93.

Imbens, G. W. & Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge: Cambridge University Press.

Kress, R., Maz'ya, V. & Kozlov, V. (1999). *Linear Integral Equations*. New York: Springer, 2nd ed.

Lehmann, E. L. & Scheffé, H. (1950). Completeness, similar regions, and unbiased estimation: Part I. *Sankhyā* **10**, 305–40.

Lu, B. & Ashmead, R. (2018). Propensity score matching analysis for causal effects with MNAR covariates. *Statist. Sinica* **28**, 2005–25.

Mattei, A. (2009). Estimating and using propensity score in presence of missing background data: An application to assess the impact of childbearing on wellbeing. *Statist. Meth. Appl.* **18**, 257–73.

Mitra, R. & Reiter, J. P. (2011). Estimating propensity scores with missing covariate data using general location mixture models. *Statist. Med.* **30**, 627–41.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *J. Economet.* **79**, 147–68.

Newey, W. K. & Powell, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica* **71**, 1565–78.

Neyman, J. (1923). Sur les applications de la thar des probabilities aux experiences Agaricales: Essay de principle. *Statist. Sci.* **5**, 465–72. English translation of excerpts by D. Dabrowska and T. Speed.

Pearl, J. (1995). Causal diagrams for empirical research (with Discussion). *Biometrika* **82**, 669–88.

Qu, Y. & Lipkovich, I. (2009). Propensity score estimation with missing values using a multiple imputation missingness pattern (MIMP) approach. *Statist. Med.* **28**, 1402–14.

Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

Rosenbaum, P. R. & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *J. Am. Statist. Assoc.* **79**, 516–24.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66**, 688–701.

Rubin, D. B. (1976). Inference and missing data (with Discussion). *Biometrika* **63**, 581–92.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statist. Med.* **26**, 20–36.

Seaman, S. & White, I. (2014). Inverse probability weighting with missing predictors of treatment assignment or missingness. *Commun. Statist.* A **43**, 3499–515.

U.S. Department of Education (2017). *What Works Clearinghouse: Standards Handbook, Version 4.0*. Washington, DC: Institute of Education Sciences.

Yang, S. & Kim, J. K. (2016). Fractional imputation in survey sampling: A comparative review. *Statist. Sci.* **31**, 415–32.

[*Received on* 28 *April* 2018. *Editorial decision on* 27 *February* 2019]

Biometrika (2019), **xx**, x, *pp*. 1–14
Printed in Great Britain

Supplementary material for Causal inference with confounders missing not at random

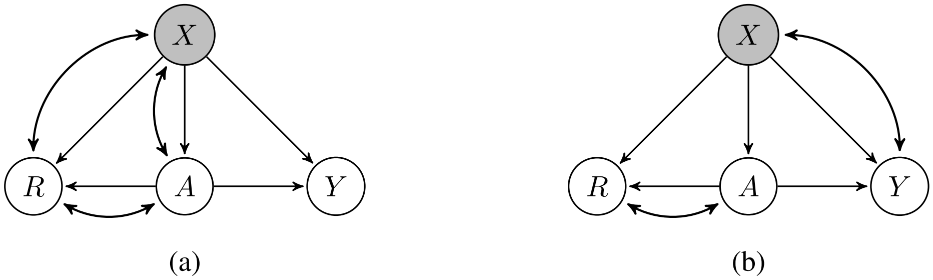
BY S. YANG
Department of Statistics, North Carolina State University, North Carolina 27695, U.S.A.
syang24@ncsu.edu

L. WANG
Department of Statistical Sciences, University of Toronto, Toronto, Ontario M5S 3G3, Canada.
linbo.wang@utoronto.ca

AND P. DING
Department of Statistics, University of California, Berkeley, California 94720, U.S.A.
pengdingpku@berkeley.edu

S1. CAUSAL DIAGRAMS ILLUSTRATING ASSUMPTIONS 1 AND 6

Figure 1 gives a causal diagram to illustrate Assumptions 1 and 6. Figure S1 gives two more diagrams.



(a)

(b)

Fig. S1. White nodes represent observed variables, the light grey nodes represent the variables with missing values, and the double arrows represent unmeasured common causes.

S2. PROOFS

S2·1. *Proof of Proposition* 1

We prove the result for $p = 2$. Proofs for general p are similar and hence omitted. For discrete covariates with $R = (0, 0)$, (3) reduces to

$$f\{A = a, Y, R = (0,0)\} = \sum_{i=1}^{J_1} \sum_{j=1}^{J_2} \frac{\mathrm{pr}\{R = (0,0) \mid X_{1i}, X_{2j}, A = a\}}{\mathrm{pr}\{R = (1,1) \mid X_{1i}, X_{2j}, A = a\}} \times f\{A = a, X_{1i}, X_{2j}, Y, R = (1,1)\}, \quad (a = 0, 1). \qquad \text{(S1)}$$

In a matrix form, (S1) becomes

$$\begin{pmatrix} f\{A=a,y_1,R=(0,0)\} \\ \vdots \\ f\{A=a,y_K,R=(0,0)\} \end{pmatrix}_{K\times 1} = \Theta_a \begin{pmatrix} \xi_{(0,0)a}(x_{11},x_{21}) \\ \vdots \\ \xi_{(0,0)a}(x_{1J_1},x_{2J_2}) \end{pmatrix}_{(J_1J_2)\times 1}, \tag{S2}$$

where

$$\Theta_a = \begin{pmatrix} f\{A=a,x_{11},x_{21},y_1,R=(1,1)\} & \cdots & f\{A=a,x_{1J_1},x_{2J_2},y_1,R=(1,1)\} \\ \vdots & \ddots & \vdots \\ f\{A=a,x_{11},x_{21},y_K,R=(1,1)\} & \cdots & f\{A=a,x_{1J_1},x_{2J_2},y_K,R=(1,1)\} \end{pmatrix}_{K\times(J_1J_2)},$$

and

$$\xi_{(0,0)a}(x_{1i},x_{2j}) = \frac{\mathrm{pr}\{R=(0,0)\mid A=a,x_{1i},x_{2j}\}}{\mathrm{pr}\{R=(1,1)\mid A=a,x_{1i},x_{2j}\}}.$$

In the linear system (S2), the vector on the left hand side and the coefficients in Θ_a on the right hand side are identifiable because they depend only on the observed data. The linear system for the $\xi_{(0,0)a}(X_1,X_2)$'s has a unique solution if and only if Θ_a has a full column rank J_1J_2. Similarly, for $R=(1,0)$,

$$\begin{pmatrix} f\{A=a,X_1,y_1,R=(1,0)\} \\ \vdots \\ f\{A=a,X_1,y_K,R=(1,0)\} \end{pmatrix}_{K\times 1} = \Theta_{X_1a} \begin{pmatrix} \xi_{(1,0)a}(X_1,x_{21}) \\ \vdots \\ \xi_{(1,0)a}(X_1,x_{2J_2}) \end{pmatrix}_{J_2\times 1}, \quad (a=0,1), \tag{S3}$$

where

$$\Theta_{X_1a} = \begin{pmatrix} f\{A=a,X_1,x_{21},y_1,R=(1,1)\} & \cdots & f\{A=a,X_1,x_{2J_2},y_1,R=(1,1)\} \\ \vdots & \ddots & \vdots \\ f\{A=a,X_1,x_{21},y_K,R=(1,1)\} & \cdots & f\{A=a,X_1,x_{2J_2},y_K,R=(1,1)\} \end{pmatrix}_{K\times J_2}.$$

The linear system (S3) has a unique solution for the $\xi_{(1,0)a}(X_1,X_2)$'s if and only if Θ_{X_1a} has a column rank J_2, which is guaranteed if Θ_a has a full column rank J_1J_2 . For $R=(0,1)$,

$$\begin{pmatrix} f\{A=a,X_2,y_1,R=(0,1)\} \\ \vdots \\ f\{A=a,X_2,y_K,R=(0,1)\} \end{pmatrix}_{K\times 1} = \Theta_{x_2a} \begin{pmatrix} \xi_{(0,1)a}(x_{11},X_2) \\ \vdots \\ \xi_{(0,1)a}(x_{1J_1},X_2) \end{pmatrix}_{J_1\times 1}, \quad (a=0,1), \tag{S4}$$

where

$$\Theta_{X_2a} = \begin{pmatrix} f\{A=a,x_{11},X_2,y_1,R=(1,1)\} & \cdots & f\{A=a,x_{1J_1},X_2,y_1,R=(1,1)\} \\ \vdots & \ddots & \vdots \\ f\{A=a,x_{11},X_2,y_K,R=(1,1)\} & \cdots & f\{A=a,x_{1J_1},X_2,y_K,R=(1,1)\} \end{pmatrix}_{K\times J_1}.$$

The linear system (S4) has a unique solution for the $\xi_{(0,1)a}(X_1,X_2)$'s if and only if Θ_{X_2a} has a column rank J_1, which is guaranteed if Θ_a has a full column rank J_1J_2. Therefore, $\xi_{ra}(X_1,X_2)$ is identifiable if and only if Θ_a has a full column rank J_1J_2.

It follows that

$$\mathrm{pr}(R = r \mid A = a, X_1, X_2) = \frac{\xi_{ra}(X_1, X_2)}{\sum_{r' \in \mathcal{R}} \xi_{r'a}(X_1, X_2)}$$

is identifiable. It then follows that

$$f(A = a, X, Y) = \frac{f(A = a, X, Y, R = 1_p)}{\mathrm{pr}(R = 1_p \mid A = a, X_1, X_2)}$$

is identifiable. Therefore, the joint distribution of (A, X, Y, R), $f(A = a, X, Y)\mathrm{pr}(R = r \mid A = a, X)$, is identifiable. This completes the proof.

S2·2. *Proof of Remark* 1

We first prove that when X and Y are discrete with finite supports, Assumption 8 is equivalent to the rank condition in Proposition 1.

PROPOSITION S1. *Suppose that X and Y are discrete, and that $X_j \in \{x_{j1}, \ldots x_{jJ_j}\}$ for $j = 1, \ldots, p$ and $Y \in \{y_1, \ldots, y_K\}$. The bounded completeness in Y of $f(A = a, X, Y, R = 1_p)$ is equivalent to the condition that Θ_a is of full column rank, for $a = 0, 1$.*

*Proof of Proposition S*1. Suppose that $\int g(X) f(A = a, X, Y, R = 1_p)\mathrm{d}\nu(X) = 0$ for all $Y = y_1, \ldots, y_K$. For discrete X, the integral equation (3) reduces to

$$\Theta_a \begin{pmatrix} g(x_{11}, \ldots, x_{p1}) \\ \vdots \\ g(x_{1J_1}, \ldots, x_{pJ_p}) \end{pmatrix}_{(J_1 \times \cdots \times J_p) \times 1} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}_{K \times 1}. \tag{S5}$$

If Θ_a is of full column rank, then the solution to the linear system (S5) is zero, that is, $g(X) = 0$, which indicates that $f(A = a, X, Y, R = 1_p)$ is bounded complete in Y.

On the other hand, suppose $f(A = a, X, Y, R = 1_p)$ is bounded complete in Y. Therefore, $\int g(X) f(A = a, X, Y, R = 1_p)\mathrm{d}\nu(X) = 0$ for all $Y = y_1, \ldots, y_K$ implies $g(X) = 0$. In this case, the only solution to (S5) is

$$\begin{pmatrix} g(x_{11}, \ldots, x_{p1}) \\ \vdots \\ g(x_{1J_1}, \ldots, x_{pJ_p}) \end{pmatrix}_{(J_1 \times \cdots \times J_p) \times 1} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}_{(J_1 \times \cdots \times J_p) \times 1}.$$

Therefore, Θ_a is of full column rank. This completes the proof. □

We then prove that Assumption 8 implies Assumption 2.

PROPOSITION S2. *Assumption* 8 *implies Assumption* 2.

*Proof of Proposition S*2. For the discrete X and Y, suppose that there exists x^* with $\mathrm{pr}(X = x^*) > 0$, such that $e(x^*) = \mathrm{pr}(A = 1 \mid X = x^*) = 0$. Then,

$$f(A = 1, X = x^*, Y, R = 1_p) = e(x^*) f(X = x^*, Y)\mathrm{pr}(R = 1_p \mid X = x^*, Y, A = 1) = 0,$$

which indicates that one column in Θ_1 is zero. Therefore, Θ_1 is not of full column rank, violating the bounded completeness condition.

For the continuous X and Y, suppose that there exists a subset $\mathcal{X}^*$ with $\mathrm{pr}(x^* \in \mathcal{X}^*) > 0$, such that $e(x^*) = \mathrm{pr}(A = 1 \mid X = x^*) = 0$ for any $x^* \in \mathcal{X}^*$. Following the same derivation as for the discrete case, we have, for any $x^* \in \mathcal{X}^*$, that $f(A = 1, X = x^*, Y, R = 1_p) = 0$. Then,

$f(A = 1, X, Y, R = 1_p)$ is not bounded complete in Y. To see this, suppose $\int g(X)f(A = 1, X, Y, R = 1_p)\mathrm{d}\nu(X) = 0$ for any Y, we can let $g(X)$ be zero outside of $\mathcal{X}^*$ but non-zero inside of $\mathcal{X}^*$, violating the bounded completeness condition. □

S3. MORE DETAILS FOR THE NONPARAMETRIC ESTIMATION OF τ

S3·1. *Regularization of series estimators*

We consider a scalar Y and a p-vector of X with one component subject to missingness. Although we can use other regularization techniques to solve the ill-conditioned inverse problem such as Tikhonov's regularization (Darolles et al., 2011) and a penalized sieve minimum distance criterion (Chen & Pouzo, 2015), we follow Newey & Powell (2003) to restrict $\xi_{ra}(X)$ and its estimator $\hat{\xi}_{ra}(X)$ to belong to a compact space. Because the inverse of integration restricted to a compact space is continuous, this regularization turns the problem to be well-posed.

We now describe the compact space and its norm. Recall that p is the dimension of X. For any function $g(X)$, denote

$$D^{\lambda}g(X) = \frac{\partial^{\lambda_1}}{\partial x_1^{\lambda_1}} \cdots \frac{\partial^{\lambda_p}}{\partial x_p^{\lambda_p}} g(X),$$

and $|\lambda| = \sum_{l=1}^{p} \lambda_l$ gives the order of the derivative. In particular, the zero order derivative is the function itself; that is, $D^0 g(X) = g(X)$. For $H(X) = \{h_1(X), \ldots, h_J(X)\}$, we define $D^{\lambda}H(X) = \{D^{\lambda}h_1(X), \ldots, D^{\lambda}h_J(X)\}^{\mathrm{T}}$. Let $\lambda = (\lambda_1, \ldots, \lambda_p)^{\mathrm{T}}$ be a p-vector with non-negative integers as components. For $m > 0$, $m_0, \delta_0 > p/2$, and $p/2 < \delta < \delta_0$, consider the following functional space

$$\mathcal{G}_{m,m_0,\delta_0,B} = \left\{ g(X) : \sum_{|\lambda| \le m+m_0} \int \{D^{\lambda}g(\tilde{X})\}^2 (1 + \tilde{X}^{\mathrm{T}}\tilde{X})^{\delta_0} \mathrm{d}x \le B \right\}, \tag{S6}$$

where $\tilde{X} = \Sigma^{-1/2}(X - \mu)$ is a linear transformation of X . Consider the norm

$$||g||_{\mathcal{G}} = \max_{|\lambda| \le m} \sup_{X} |D^{\lambda}g(\tilde{X})|(1 + \tilde{X}^{\mathrm{T}}\tilde{X})^{\delta}.$$

Gallant & Nychka (1987) showed that the closure of $\mathcal{G}_{m,m_0,\delta_0,B}$ with respect to the norm $||g||_{\mathcal{G}}$ is compact.

*Assumption S*1 (*Regularization of the parameter space*). Assume that $\xi_{ra}(X)$ and its estimator $\hat{\xi}_{ra}(X)$ belong to $\mathcal{G}_{m,m_0,\delta_0,B}$ in (S6), for any r and a.

*Remark S*1. The regularization is not restrictive for the following reasons. First, by the definition of $\mathcal{G}_{m,m_0,\delta_0,B}$, the bound B requires the functions of $\mathcal{G}_{m,m_0,\delta_0,B}$ to be smooth to a certain degree and the tail areas of these functions to be small. In most applications, we would expect that the functions $\xi_{ra}(X)$ to be smooth and mainly concern with the functional forms of $\xi_{ra}(X)$ over some compact region that is large enough to cover the region where observations are measured.

Given the Hermite approximation of $\xi_{ra}(X)$, the regularization in Assumption S1 becomes

$$\beta_{ra}^{\mathrm{T}} \left[\sum_{|\lambda| \le m+m_0} \int \{D^{\lambda}H(\tilde{X})\}\{D^{\lambda}H(\tilde{X})\}^{\mathrm{T}} (1 + \tilde{X}^{\mathrm{T}}\tilde{X})^{\delta_0} \mathrm{d}\nu(X) \right] \beta_{ra} \le B, \tag{S7}$$

where $\beta_{ra} = (\beta_{ra}^1, \ldots, \beta_{ra}^J)^{\mathrm{T}}$. Therefore, we choose the positive definite matrix Λ in the constraint for regularization in §4·1 to be

$$\Lambda = \sum_{|\lambda| \leq m+m_0} \int \{D^\lambda H(\tilde{X})\}\{D^\lambda H(\tilde{X})\}^{\mathrm{T}}(1+\tilde{X}^{\mathrm{T}}\tilde{X})^{\delta_0}\mathrm{d}\nu(X).$$

The proposed estimator of ξ_{ra} is $\hat{\xi}_{ra}(X) = \sum_{j=1}^J \hat{\beta}_{ra}^j h^j(\tilde{X})$, where $\hat{\beta}_{ra}$ minimizes (10) with the constraint $\beta_{ra}^{\mathrm{T}}\Lambda\beta_{ra} \leq B$.

S3·2. *The computational algorithm in* §4·1 *and an example*

We summarize the computation algorithm for τ as follows.

*Step S*1. Obtain nonparametric estimators of $\tau(X)$, $f(X \mid A = a, R = 1_p)$, $f(X_r, Y \mid A = a, R = r)$, for all r and a. Specifically, we use

$$\hat{\tau}(X) = \hat{E}(Y \mid A = 1, X, R = 1_p) - \hat{E}(Y \mid A = 0, X, R = 1_p), \tag{S8}$$

where $\hat{E}(Y \mid A = a, X, R = 1_p)$ is a smoothing spline estimator of $E(Y \mid A = a, R = 1_p)$, for $a = 0, 1$. Also let $\hat{f}(X \mid A = a, R = 1_p)$ and $\hat{f}(X_r, Y \mid A = a, R = r)$ be the kernel density estimators of $f(X \mid A = a, R = 1_p)$ and $f(X_r, Y \mid A = a, R = r)$, respectively.

*Step S*2. Obtain a series estimator of $\xi_{ra}(X)$ using the Hermite polynomials, $\hat{\xi}_{ra}(X) \approx \sum_{j=1}^J \hat{\beta}_{ra}^j h^j(\tilde{X})$, where $(\hat{\beta}_{ra}^1, \ldots, \hat{\beta}_{ra}^J)^{\mathrm{T}}$ minimizes (10) with the constraint $\beta_{ra}^{\mathrm{T}}\Lambda\beta_{ra} \leq B$.

*Step S*3. Estimate the probabilities $\mathrm{pr}(R = 1_p \mid A = a, X)$ by

$$\widehat{\mathrm{pr}}(R = 1_p \mid A = a, X) = \left\{1 + \sum_{r \neq 1_p} \hat{\xi}_{ra}(X)\right\}^{-1}.$$

*Step S*4. Estimate τ by (11) using a numerical approximation.

For illustration of the proposed computational algorithm, we provide an example with a scalar X, which is subject to the outcome-independent missingness. In this case, $R \in \mathcal{R} = \{0, 1\}$.

*Example S*1. In Step S1, obtain a nonparametric estimator of $\tau(X)$ as

$$\hat{\tau}(X) = \hat{E}(Y \mid A = 1, X, R = 1) - \hat{E}(Y \mid A = 0, X, R = 1),$$

where $\hat{E}(Y \mid A = a, X, R = 1)$ is a smoothing spline estimator of $E(Y \mid A = a, X, R = 1)$, for $a = 0, 1$. Also let

$$\hat{f}(X \mid A = a, R = 1), \quad \hat{f}(Y \mid A = a, R = 0), \quad \hat{f}(Y, R = 0 \mid A = a), \quad \hat{f}(Y, R = 1 \mid A = a)$$

be the kernel density estimators of

$$f(X \mid A = a, R = 1), \quad f(Y \mid A = a, R = 0), \quad f(Y, R = 0 \mid A = a), \quad f(Y, R = 1 \mid A = a).$$

In Step S2, (8) becomes

$$\hat{f}(Y, R = 0 \mid A = a) = \int \xi_{0a}(X)\hat{f}(X \mid A = a, Y, R = 1)\mathrm{d}\nu(X) \times \hat{f}(Y, R = 1 \mid A = a).$$

Let $\hat{E}\{h^j(\tilde{X}) \mid y, A = a, R = 1\}$ be a nonparametric estimator of $E\{h^j(\tilde{X}) \mid y, A = a, R = 1\}$. For unit i, evaluate this nonparametric estimator at Y_i, we have $\hat{H}_{0a,i}^j = \hat{E}\{h^j(\tilde{X}) \mid A_i =$

$a, Y_i, R_i = 1\}$. We obtain a series estimator of $\xi_{0a}(X)$ using the Hermite polynomials, $\hat{\xi}_{0a}(X) \approx \sum_{j=1}^{J} \hat{\beta}_{0a}^{j} h^{j}(\tilde{X})$, where the $\hat{\beta}_{0a}^{j}$'s minimize the objective function

$$\sum_{i=1}^{n} I(R_i = 0) \left\{ \hat{f}(Y_i, R_i = 0 \mid A_i = a) - \sum_{j=1}^{J} \beta_{0a}^{j} \hat{H}_{0a,i}^{j} \hat{f}(Y_i, R_i = 1 \mid A_i = a) \right\}^2, \quad \text{(S9)}$$

subject to the constraint $\beta_{0a}^{\mathrm{T}} \Lambda \beta_{0a} \leq B$.

In Step S3, estimate the probability $\mathrm{pr}(R = 1 \mid A = a, X)$ by $\widehat{\mathrm{pr}}(R = 1 \mid A = a, X) = \{1 + \hat{\xi}_{0a}(X)\}^{-1}$.

In Step S4, obtain the estimator of τ by using a numerical approximation of

$$\sum_{a=0}^{1} \widehat{\mathrm{pr}}(A = a, R = 1) \int \hat{\tau}(X) \frac{\hat{f}(X \mid A = a, R = 1)}{\widehat{\mathrm{pr}}(R = 1 \mid A = a, X)} \mathrm{d}\nu(X). \quad \text{(S10)}$$

S3·3. *Choice of tuning parameters*

The proposed estimator depends on several tuning parameters: the number of the Hermite polynomial functions J, the bound B for regularization, and tuning parameters in the kernel-based estimators. On the one hand, J and B should be large enough to ensure that the series estimator approximates the true underlying function well. On the other hand, J and B should not be too large to control the variance of our estimator. Chen & Pouzo (2012) and Chen & Christensen (2015) investigated the general requirements for these tuning parameters in terms of the growing rate with the sample size in the penalized sieve minimum distance estimation. In practice, we suggest using data-driven methods, such as cross-validation, to choose these parameters, and conducting sensitivity analysis varying the tuning parameters.

S4. ASYMPTOTIC RESULTS FOR THE NONPARAMETRIC ESTIMATION

We study the consistency of the proposed estimator $\hat{\tau}$ of τ. The literature has established comprehensive consistency results for nonparametric estimators and series estimators. For completeness of our theory, in §S4·1 and §S4·2, we establish the consistency of the nonparametric estimators in Step S1 and the series estimator of $\xi_{ra}(X)$ in Step S2, which serve building blocks for deriving the consistency result for $\hat{\tau}$ in §S4·3.

S4·1. *The consistency of the nonparametric estimators in Step S*1

We assume that the kernel functions and the bandwidth h_n satisfy the following regularity conditions:

*Assumption S*2. (i)$\int_{\mathcal{R}^p} K(s)\mathrm{d}s = 1$; (ii) $||K||_{\infty} = \sup_{x \in \mathcal{R}^p} |K(x)| = \kappa < \infty$; (iii) $K(\cdot)$ is right continuous; (iv) $\int_{\mathcal{R}^p} \Psi_K(x)\mathrm{d}x < \infty$, where $\Psi_K(x) = \sup_{||y|| \geq ||x||} |K(y)|$, for $x \in \mathcal{R}^p$; and (v) the kernel function is regular and satisfies the following uniform entropy condition. Let $\mathcal{K}$ be the class of functions indexed by x,

$$\mathcal{K} = \left\{ K\left(\frac{x - \cdot}{h^{1/p}} \right) : h > 0, x \in \mathcal{R}^p \right\}.$$

Suppose $\mathcal{B}$ is a Borel set in $\mathcal{R}^p$, and Q is some probability measure on $(\mathcal{R}^p, \mathcal{B})$. Define d_Q to be the $L_2(Q)$-metric, and $N(\epsilon, \mathcal{K}, d_Q)$ the minimal number of balls $\{g : d_Q(g, g') < \epsilon\}$ of d_Q-radius ϵ needed to cover $\mathcal{K}$. Let $N(\epsilon, \mathcal{K}) = \sup_Q N(\kappa\epsilon, \mathcal{K}, d_Q)$, where the supremum is taken over all probability measures Q. For some $C > 0$ and $\nu > 0$, $N(\epsilon, \mathcal{K}) \leq C\epsilon^{-\nu}$ for any $0 < \epsilon < 1$.

van der Vaart & Wellner (1996) provides sufficient conditions for Assumption S2 (v).

*Assumption S*3. h_n decreases to zero, h_n/h_{2n} is bounded, $\log(1/h_n)/\log\log n \to \infty$ and $nh_n/\log n \to \infty$, as $n \to \infty$.

LEMMA S1 (CONSISTENCY OF KERNEL DENSITY ESTIMATORS). *Let $\hat{f}(X \mid A = a, R = 1_p)$ be the kernel density estimator of $f(X \mid A = a, R = 1_p)$, where the kernel function satisfies Assumption S*2*, and the bandwidth h_n satisfies Assumption S*3*. Suppose that the true density function $f(X \mid A = a, R = 1_p)$ is bounded and uniformly continuous in X, then*

$$\lim_{n\to\infty} \left\| \hat{f}(X \mid A = a, R = 1_p) - f(X \mid A = a, R = 1_p) \right\|_\infty = 0 \tag{S11}$$

almost surely.

The Nadaraya–Watson estimators of $E(Y \mid A = a, X, R = 1_p)$ is

$$\hat{E}(Y \mid A = a, X, R = 1_p) = \sum_a Y_i K\left(\frac{X - X_i}{h_n^{1/p}}\right) \Big/ \sum_a K\left(\frac{X - X_i}{h_n^{1/p}}\right), \tag{S12}$$

where $\sum_a$ represents the summation over units $\{i : A_i = a, R_i = 1_p\}$. We focus on the Nadaraya–Watson estimator, but we can also consider other nonparametric estimators, such as local polynomial estimator.

Let I be a compact subset of $\mathcal{R}^p$. For any function $\psi : \mathcal{R}^p \to \mathcal{R}$, define

$$||\psi||_I = \sup_{X\in I} |\psi(X)|. \tag{S13}$$

Also, denote $I^\epsilon = \{X \in \mathcal{R}^p : \max_{1\leq i\leq p} |X_i| \leq \epsilon\}$.

LEMMA S2 (CONSISTENCY OF KERNEL-BASED ESTIMATORS FOR CONDITIONAL MEANS). *Suppose that the kernel function $K(\cdot)$ in (S*12*) satisfies Assumption S*2 *with support contained in $[-1/2, 1/2]^p$, and the bandwidth h_n satisfies Assumption S*3*. Suppose that there exists an $\epsilon > 0$ such that $f(X \mid A = a, R = 1_p) = \int_{-\infty}^{\infty} f(X, Y \mid A = a, R = 1_p)\mathrm{d}Y$ is continuous and strictly positive on I^ϵ, and that $f(X, Y \mid A = a, R = 1_p)$ is continuous in X for almost every $Y \in \mathcal{R}$. Suppose further that there exists an $M > 0$ such that for $X \in I^\epsilon$, $|Y| \leq M$ almost surely. Then, for $a = 0, 1$,*

$$\lim_{n\to\infty} \left\| \hat{E}(Y \mid A = a, X, R = 1_p) - E(Y \mid A = a, X, R = 1_p) \right\|_I = 0 \tag{S14}$$

almost surely.

A large literature has developed consistency of kernel-based estimators. The proofs of Lemmas S1 and S2 are similar to those given by Deheuvels (2000) and Giné & Guillou (2002), and therefore are omitted. The smoothing spline estimator is asymptotically equivalent to a kernel-based estimator that employs the so-called spline kernel (Silverman, 1984). Both spline kernels and Gaussian kernels satisfy Assumption S2 (van der Vaart & Wellner, 1996). Therefore, by Lemmas S1 and S2, the nonparametric estimators in Step S1 are consistent.

S4·2. *The consistency of the series estimator of $\xi_{ra}(X)$ in Step S*2

For any r and a, $\xi_{ra}(X)$ satisfies the conditional moment restriction

$$E\{f(X_r, Y, R = r \mid A = a) - \xi_{ra}(X)f(X, Y, R = 1_p \mid A = a) \mid A = a, X_r, Y, R = r\} = 0.$$

We define a generalized residuals with the function of interest $h(X)$ as

$$\rho_{ra}(X, Y; h) = f(X_r, Y, R = r \mid A = a) - h(X)f(X, Y, R = 1_p \mid A = a),$$

the conditional mean function of $\rho_{ra}(X, Y; h)$ given $(A = a, X_r, Y, R = r)$ as

$$m_{ra}(X_r, Y; h) = E\{\rho_{ra}(X, Y; h) \mid A = a, X_r, Y, R = r\},$$

and the series least square estimator of the conditional mean function as

$$\begin{aligned}\hat{m}_{ra}(X_r, Y; h) = {} & \hat{f}(X_r, Y, R = r \mid A = a) \\ & -\hat{E}\{h(X) \mid A = a, X_r, Y, R = r\}\hat{f}(X, Y, R = 1_p \mid A = a).\end{aligned}$$

Following these definitions, $m_{ra}(X_r, Y; \xi_{ra}) = 0$ for any r and a. Let the project of ξ_{ra} onto $\mathcal{H}_J$ be $\prod_{\mathcal{H}_J} \xi_{ra}(\cdot) = \sum_{j=1}^{J} \beta_{ra}^j h^j(\cdot)$ such that $|| \prod_{\mathcal{H}_J} \xi_{ra} - \xi_{ra}||_\infty = o(1)$.

To avoid technicality, we assume the following regularity conditions.

Assumption S4. (i) $E\{||m_{ra}(X_r, Y; \prod_{\mathcal{H}_J} \xi_{ra})||_{\mathcal{G}}^2\} = o(1)$; (ii) $n^{-1} \sum_{i=1}^n ||m_{ra}(X_{r,i}, Y_i; \prod_{\mathcal{H}_J} \xi_{ra})||_{\mathcal{G}}^2 \leq c_0 E||m_{ra}(X_r, Y; \prod_{\mathcal{H}_J} \xi_{ra})||_{\mathcal{G}}^2 + o_p(1)$ and a finite constant $c_0 > 0$; (iii) $n^{-1} \sum_{i=1}^n ||\hat{m}_{ra}(X_{r,i}, Y_i; h)||_{\mathcal{G}}^2 \geq c_1 E||m_{ra}(X_r, Y; h)||_{\mathcal{G}}^2 - o_p(1)$ uniformly for h over $\mathcal{H}_J$ and a finite constant $c_1 > 0$.

Assumption S4 (i) holds if $E\{||m_{ra}(X_r, Y; h)||_{\mathcal{G}}^2\}$ is continuous at $h = \xi_{ra}$ under $|| \cdot ||_\infty$. Assumption S4 (ii) and (iii) are sample criteria to regularize the asymptotic behavior of the series estimator of $m_{ra}(X_{r,i}, Y_i; h)$. Chen & Pouzo (2012) provided sufficient conditions for Assumption S4.

LEMMA S3 (CONSISTENCY OF $\hat{\xi}_{ra}$). *Under Assumptions* 1, 7, 6 *and Assumption S*4*, the series estimator* $\hat{\xi}_{ra}(X) = \sum_{j=1}^{J} \hat{\beta}_{ra}^j h^j(\tilde{X})$ *is consistent for* $\xi_{ra}(X)$ *in the sense that* $||\hat{\xi}_{ra} - \xi_{ra}||_\infty = o_p(1)$ *as* $J \to \infty$ *and* $n \to \infty$.

Chen & Pouzo (2012) provided a proof for Lemma S3 in the context of estimation of nonparametric conditional moment models. Our proof for Lemma S3 is similar, and therefore omitted.

S4·3. *The consistency of the proposal estimator of τ in Step S*4

Let $||X||$ be the Euclidean norm for X. Denote $I_K = \{X : ||X|| > K\}$ for a constant K, and I_K^c to be the complement set of I_K.

THEOREM S1 (CONSISTENCY OF $\hat{\tau}$). *Suppose that the assumptions in Theorem* 3 *and Lemmas S*1*–S*3 *hold. Suppose further that for some* $B > 0$*,* $\hat{\tau}(X)$ *and* $\tau(X)$ *are uniformly bounded for* $X \in I_B$*, and that*

$$\int_{I_K^c} \frac{f(X \mid A = a, R = 1_p)}{\mathrm{pr}(R = 1_p \mid A = a, X)} \mathrm{d}\nu(X) \to 0, \qquad \text{(S15)}$$

as $K \to \infty$*. Then, the nonparametric estimator* $\hat{\tau}$ *resulting from (*11*) is consistent for* τ.

The proposed estimator $\hat{\tau}$ is a linear functional of $\hat{\tau}(\cdot)$, $\hat{f}(\cdot \mid A = a, R = 1_p)$, and $\hat{\xi}_{ra}(\cdot)$. A large literature has established the root-n asymptotic normality and the consistent variance estimation for plug-in series estimators of functionals; see, for example, Newey (1997), Shen (1997), Chen & Shen (1998), Li & Racine (2007), Chen (2007), Chen & Pouzo (2009), Chen & Pouzo (2012), and Chen & Liao (2014). Alternatively, Chen & Pouzo (2015) provided Wald and quasi-likelihood ratio inference results for the general models in Chen & Pouzo (2012), including series two stage least squares as an example. A relatively simple approach is to treat the nonparametric estimators as if they were parametric given the fixed tuning parameters, so that there is only a finite number of parameters. From this point of view, we can use standard approaches for variance estimation under parametric models. This approach is asymptotically valid for nonparametric series regression; see, for example, Newey (1997). In the light of treating the nonparametric estimators as if they were parametric, one might expect the nonparametric bootstrap to work for our estimator. For all bootstrap samples, we use the same tuning parameters, such as the smoothing parameter in the smoothing splines and the bandwidth in the kernel density estimator. In our simulation study, inference based on the above bootstrap is promising. However, it is a difficult task (if it is possible) to prove that the bootstrap is consistent which is beyond the scope of this article. Recent work has shown that it does work for some nonparametric instrumental variable series estimators (Horowitz, 2007).

*Proof of Theorem S*1. By Lemmas S1 and S2,

$$\lim_{n\to\infty}\left|\left|\frac{\hat{f}(X \mid A=a, R=1_p)}{\widehat{\mathrm{pr}}(R=1_p \mid A=a, X)} - \frac{f(X \mid A=a, R=1_p)}{\mathrm{pr}(R=1_p \mid A=a, X)}\right|\right|_{\infty} = 0 \tag{S16}$$

almost surely. Since $\hat{\tau}(X)$ and $\tau(X)$ are uniformly bounded in I_K for $K > B$, together with (S15) and (S16), for any ϵ, there exists $K_2 > 0$, such that for any $K > K_2$,

$$\lim_{n\to\infty}\mathrm{pr}\left[\left|\int_{I_K^c}\hat{\tau}(X)\left\{\frac{\hat{f}(X \mid A=a, R=1_p)}{\widehat{\mathrm{pr}}(R=1_p \mid A=a, X)} - \frac{f(X \mid A=a, R=1_p)}{\mathrm{pr}(R=1_p \mid A=a, X)}\right\}\mathrm{d}\nu(X)\right| > \frac{\epsilon}{4}\right] < \frac{\epsilon}{4}, \tag{S17}$$

and

$$\lim_{n\to\infty}\left|\int_{I_K^c}\{\hat{\tau}(X) - \tau(X)\}\frac{f(X \mid A=a, R=1_p)}{\mathrm{pr}(R=1_p \mid A=a, X)}\mathrm{d}\nu(X)\right| < \frac{\epsilon}{4}. \tag{S18}$$

By Theorem S3, for any K,

$$\lim_{n\to\infty}\left|\left|\hat{\tau}(X, R=1_p)\frac{\hat{f}(X \mid A=a, R=1_p)}{\widehat{\mathrm{pr}}(R=1_p \mid A=a, X)} - \tau(X, R=1_p)\frac{f(X \mid A=a, R=1_p)}{\mathrm{pr}(R=1_p \mid A=a, X)}\right|\right|_{I_K} = 0 \tag{S19}$$

almost surely, where $||\cdot||_I$ is defined in (S13). Therefore, for any ϵ, by (S19), we choose K_1 such that for any $K > K_1$,

$$\lim_{n\to\infty}\mathrm{pr}\left\{\left|\int_{I_{K_1}}\hat{\tau}(X, R=1_p)\frac{\hat{f}(X \mid A=a, R=1_p)}{\widehat{\mathrm{pr}}(R=1_p \mid A=a, X)}\mathrm{d}\nu(X) - \int_{I_{K_1}}\tau(X, R=1_p)\frac{f(X \mid A=a, R=1_p)}{\mathrm{pr}(R=1_p \mid A=a, X)}\mathrm{d}\nu(X)\right| > \frac{\epsilon}{2}\right\} < \frac{\epsilon}{2}. \tag{S20}$$

Combing (S17), (S18) and (S20), for any $\epsilon > 0$, we choose $K > \max(K_1, K_2)$,

$$
\begin{aligned}
&\lim_{n\to\infty} \mathrm{pr}(|\hat{\tau} - \tau| > \epsilon) \\
&= \lim_{n\to\infty} \mathrm{pr}\left\{\left| \int \hat{\tau}(X) \frac{\hat{f}(X \mid A = a, R = 1_p)}{\widehat{\mathrm{pr}}(R = 1_p \mid A = a, X)} \mathrm{d}\nu(X) - \int \tau(X) \frac{f(X \mid A = a, R = 1_p)}{\mathrm{pr}(R = 1_p \mid A = a, X)} \mathrm{d}\nu(X) \right| > \epsilon \right\} \\
&\leq \lim_{n\to\infty} \mathrm{pr}\left\{\left| \int_{I_K} \hat{\tau}(X) \frac{\hat{f}(X \mid A = a, R = 1_p)}{\widehat{\mathrm{pr}}(R = 1_p \mid A = a, X)} \mathrm{d}\nu(X) - \int_{I_K} \tau(X) \frac{f(X \mid A = a, R = 1_p)}{\mathrm{pr}(R = 1_p \mid A = a, X)} \mathrm{d}\nu(X) \right| > \frac{\epsilon}{2} \right\} \\
&+ \lim_{n\to\infty} \mathrm{pr}\left[\left| \int_{I_K^c} \hat{\tau}(X) \left\{ \frac{\hat{f}(X \mid A = a, R = 1_p)}{\widehat{\mathrm{pr}}(R = 1_p \mid A = a, X)} - \frac{f(X \mid A = a, R = 1_p)}{\mathrm{pr}(R = 1_p \mid A = a, X)} \right\} \mathrm{d}\nu(X) \right| > \frac{\epsilon}{4} \right] \\
&+ \lim_{n\to\infty} \mathrm{pr}\left[\left| \int_{I_K^c} \{\hat{\tau}(X) - \tau(X)\} \frac{f(X \mid A = a, R = 1_p)}{\mathrm{pr}(R = 1_p \mid A = a, X)} \mathrm{d}\nu(X) \right| > \frac{\epsilon}{4} \right] < \epsilon,
\end{aligned}
$$

that is, $\hat{\tau}$ is consistent for τ. □

S5. MORE DETAILS FOR THE PARAMETRIC ESTIMATION OF τ

S5·1. *Bounded completeness and an example*

Bounded completeness is weaker than completeness. We say that a function $f(X, Y)$ is complete in Y if $\int g(X) f(X, Y) \mathrm{d}\nu(X) = 0$ implies $g(X) = 0$ almost surely for any squared integrable function $g(X)$. For illustration, we give sufficient conditions for completeness of distribution functions in an exponential family, which implies bounded completeness.

LEMMA S4. *The distribution $f(X, Y) = \psi(X) h(Y) \exp\{\lambda(Y)^{\mathrm{T}} \eta(X)\}$ is complete in Y if (i) $\psi(X) > 0$, (ii) the support of $\lambda(Y)$ is an open set, and (iii) the mapping $X \mapsto \eta(X)$ is one-to-one.*

Lemma S4 is the same as Theorem 2.2 in Newey & Powell (2003). We give an example below.

PROPOSITION S3. *For scalar X and Y, the Gaussian model*

$$
f(X, Y) = f(Y \mid X) f(X) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{ -\frac{(Y - \beta_0 - \beta_1 X)^2}{2\sigma^2} \right\} f(X), \tag{S21}
$$

is complete in Y.

*Proof of Proposition S*3. Using the notation in Lemma S4, (S21) can be expressed as $f(X, Y) = \psi(X) \exp\{\lambda(Y) \eta(X)\}$ with $\psi(X) = (2\pi\sigma^2)^{-1/2} f(X)$, $\lambda(Y) = \sigma^{-2} \beta_1 Y$ and $\eta(X) = X$. Therefore, (S21) satisfies the conditions for $\lambda(Y)$ and $\eta(X)$, and it is complete in Y. □

S5·2. *Likelihood-based inference: a fractional imputation approach*

Let $S(\theta; Z_i) = \partial \log f(Z_i; \theta) / \partial\theta$ be the complete-data score for unit i. The maximum likelihood estimator $\hat{\theta}$ is a solution of the conditional score equation (Kim & Shao, 2013)

$$
n^{-1} \sum_{i=1}^{n} \sum_{r \in \mathcal{R}} I(R_i = r) E\{S(\theta; Z_i) \mid Z_{r,i}, R_i = r; \theta\} = 0, \tag{S22}
$$

where the conditional expectation is with respect to

$$f(X_{\bar{r},i} \mid Z_{r,i}, R_i = r; \theta) = \frac{f(A_i, X_i, Y_i, R_i = r; \theta)}{\int f(A_i, X_i, Y_i, R_i = r; \theta) \mathrm{d}\nu(X_{\bar{r},i})}. \tag{S23}$$

The EM algorithm is a standard tool for solving (S22). However, it has several drawbacks. First, the computation of the conditional expectation in (S22) can be difficult due to the possibly high-dimensional integration. Second, the conditional distribution (S23) may not have an explicit form. We can use the fractional imputation (Yang & Kim, 2016) to overcome the computation difficulties. The fractional imputation uses importance sampling to avoid analytical calculation for evaluating the conditional expectation.

In fractional imputation, we approximate the conditional expectation in (S22) by

$$\sum_{r \in \mathcal{R}} I(R_i = r) E\{\tau(Z_i; \theta) \mid Z_{r,i}, R_i = r; \theta\} \approx \sum_{j=1}^{M} \omega_{ij}^{*} \tau(Z_{ij}^{*}; \theta), \tag{S24}$$

where $\{Z_{ij}^{*} = (A_i, X_{R_i,i}, X_{\bar{R}_i,i}^{*(j)}, Y_i, R_i) : j = 1, \ldots, M\}$ are the fractional observations and the ω_{ij}^{*}'s are the fractional weights that satisfy $\omega_{ij}^{*} \geq 0$ and $\sum_{j=1}^{M} \omega_{ij}^{*} = 1$. The approximation in (S24) is due to the Monte Carlo error, which becomes more accurate as M increases. Approximately, we can solve $\hat{\theta}$ from

$$n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{M} \omega_{ij}^{*} S(\theta; Z_{ij}^{*}) = 0. \tag{S25}$$

Computationally, we iteratively generate weighted fractional observations satisfying (S24) and solve the conditional score equation (S25). This often converges to $\hat{\theta}$ as $M \to \infty$.

The key is to construct (S24) using importance sampling. For each missingness pattern $R_i = r$ and the missing value $X_{\bar{r},i}$, we first generate $X_{\bar{r},i}^{*(1)}, \ldots, X_{\bar{r},i}^{*(M)}$ from a proposal distribution $h(X_{\bar{r},i} \mid Z_{r,i})$ for some $h(\cdot)$ that is easy to simulate. We then compute

$$\omega_{ij}^{*} \propto \frac{f(X_{\bar{r},i}^{*(j)} \mid Z_{r,i}; \hat{\theta})}{h(X_{\bar{r},i}^{*(j)} \mid Z_{r,i})} \propto \frac{f(Z_{ij}^{*}; \hat{\theta})}{h(X_{\bar{r},i}^{*(j)} \mid Z_{r,i})},$$

subject to $\sum_{j=1}^{M} \omega_{ij}^{*} = 1$, as the fractional weight for Z_{ij}^{*}.

As a by product, we can also use

$$\tilde{\tau}(\hat{\theta}) = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{M} \hat{\omega}_{ij} \tau(\hat{Z}_{ij}; \hat{\theta})$$

as an estimator for τ, where the $\hat{\omega}_{ij}$'s are the weights for the fractional observations $\hat{Z}_{ij}$'s at the maximum likelihood estimator $\hat{\theta}$. Clearly, $\tilde{\tau}(\hat{\theta})$ is an approximation to

$$\tilde{\tau}(\theta) = n^{-1} \sum_{i=1}^{n} \sum_{r \in \mathcal{R}} I(R_i = r) E\{\tau(X_i; \theta) \mid Z_{r,i}, R_i = r; \theta\},$$

which satisfies $E\{\tilde{\tau}(\theta)\} = \tau$.

S5·3. *Bayesian approach: an example with a scalar X*

Let R be the missing indicator the scalar X. Suppose

$$\begin{aligned} \mathrm{pr}(R=0 \mid A=a, X, Y;\eta) = \mathrm{pr}(R=0 \mid A=a, X;\eta) &= \{1+\exp(\eta_{a0}+\eta_{a1}X)\}^{-1}, \\ f(Y \mid A=a, X;\beta) &= (2\pi\sigma_a^2)^{-1/2}\exp\{-(Y-\beta_{a0}-\beta_{a1}X)^2/(2\sigma_a^2)\}, \\ \mathrm{pr}(A=1 \mid X;\alpha) &= \mathrm{logit}(\alpha_0+\alpha_1 X), \\ f(X;\lambda) &= (2\pi\sigma_x^2)^{-1/2}\exp\{-(X-\mu_x)^2/(2\sigma_x^2)\}, \end{aligned}$$

where $\eta=(\eta_{00},\eta_{01},\eta_{10},\eta_{11})$, $\beta=(\beta_{00},\beta_{01},\sigma_0^2,\beta_{10},\beta_{11},\sigma_1^2)$, $\alpha=(\alpha_0,\alpha_1)$, and $\lambda=(\mu_x,\sigma_x^2)$. The parametric $\theta=(\alpha,\beta,\eta,\lambda)$ has prior $\pi(\theta)$. The complete-data likelihood is $L(\theta \mid Z_1,\ldots,Z_n)=\prod_{i=1}^n f(Z_i;\theta)$, where

$$\begin{aligned} f(Z_i;\theta) = &\left[\frac{\exp(\eta_{10}+\eta_{11}X_i)^{R_i}}{1+\exp(\eta_{10}+\eta_{11}X_i)}\frac{1}{(2\pi\sigma_1^2)^{1/2}}\exp\left\{-\frac{(Y_i-\beta_{10}-\beta_{11}X_i)^2}{2\sigma_1^2}\right\}\right]^{A_i} \\ &\times\left[\frac{\exp(\eta_{00}+\eta_{01}X_i)^{R_i}}{1+\exp(\eta_{00}+\eta_{01}X_i)}\frac{1}{(2\pi\sigma_0^2)^{1/2}}\exp\left\{-\frac{(Y_i-\beta_{00}-\beta_{01}X_i)^2}{2\sigma_0^2}\right\}\right]^{1-A_i} \\ &\times\frac{\exp(\alpha_0+\alpha_1X_i)^{A_i}}{1+\exp(\alpha_0+\alpha_1X_i)}\times\frac{1}{(2\pi\sigma_x^2)^{1/2}}\exp\left\{-\frac{(X_i-\mu_x)^2}{2\sigma_x^2}\right\}. \end{aligned} \tag{S26}$$

By Lemma S4, it is easy to verify that $f(A=a,X,Y,R=1)$ is bounded complete in Y. By Theorem 2, θ is identifiable.

In the Bayesian estimation, we first simulate the posterior distribution of the Z_i's and θ. Given the parameter value $\theta^*=(\alpha^*,\beta^*,\eta^*,\lambda^*)$, we generate

$$\begin{aligned} X_i^* &\sim f(X_i \mid A_i,Y_i,R_i;\theta^*) \\ &\propto\left[\frac{\exp(\eta_{10}^*+\eta_{11}^*X_i)^{R_i}}{1+\exp(\eta_{10}^*+\eta_{11}^*X_i)}\frac{1}{(2\pi\sigma_1^{*2})^{1/2}}\exp\left\{-\frac{(Y_i-\beta_{10}^*-\beta_{11}^*X_i)^2}{2\sigma_1^{*2}}\right\}\right]^{A_i} \\ &\times\left[\frac{\exp(\eta_{00}^*+\eta_{01}^*X_i)^{R_i}}{1+\exp(\eta_{00}^*+\eta_{01}^*X_i)}\frac{1}{(2\pi\sigma_0^{*2})^{1/2}}\exp\left\{-\frac{(Y_i-\beta_{00}^*-\beta_{01}^*X_i)^2}{2\sigma_0^{*2}}\right\}\right]^{1-A_i} \\ &\times\frac{\exp(\alpha_0^*+\alpha_1^*X_i)^{A_i}}{1+\exp(\alpha_0^*+\alpha_1^*X_i)}\frac{1}{(2\pi\sigma_x^{*2})^{1/2}}\exp\left\{-\frac{(X_i-\mu_x^*)^2}{2\sigma_x^{*2}}\right\} \end{aligned}$$

for units with $R_i=0$. For units with $R_i=1$, let $X_i^*=X_i$. Given the imputed values X_i^*, we have the complete data Z_i^*, and then generate $\theta^*\sim f(\theta \mid Z_1^*,\ldots,Z_n^*)\propto L(\theta \mid Z_1^*,\ldots,Z_n^*)\pi(\theta)$. Both steps may involve the Markov chain Monte Carlo.

Given $(\theta^*,X_1^*,\ldots,X_n^*)$, we calculate $\hat{\tau}(\theta^*)=n^{-1}\sum_{i=1}^n\tau(X_i^*;\theta^*)$ as a posterior draw of $\hat{\tau}(\theta)$. This gives the posterior distribution of the average causal effect conditioning on the covariate values.

REFERENCES

CHEN, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics* **6**, 5549–5632.

CHEN, X. & CHRISTENSEN, T. M. (2015). Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. *Journal of Econometrics* **188**, 447–465.

CHEN, X. & LIAO, Z. (2014). Sieve M inference on irregular parameters. *Journal of Econometrics* **182**, 70–86.

CHEN, X. & POUZO, D. (2009). Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals. *Journal of Econometrics* **152**, 46–60.

CHEN, X. & POUZO, D. (2012). Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica* **80**, 277–321.

CHEN, X. & POUZO, D. (2015). Sieve Wald and QLR inferences on semi/nonparametric conditional moment models. *Econometrica* **83**, 1013–1079.

CHEN, X. & SHEN, X. (1998). Sieve extremum estimates for weakly dependent data. *Econometrica* **66**, 289–314.

DAROLLES, S., FAN, Y., FLORENS, J.-P. & RENAULT, E. (2011). Nonparametric instrumental regression. *Econometrica* **79**, 1541–1565.

DEHEUVELS, P. (2000). Uniform limit laws for kernel density estimators on possibly unbounded intervals. In *Recent Advances in Reliability Theory: Methodology, Practice and Inference*. Birkhauser, Basel: Springer, pp. 477–492.

GALLANT, A. R. & NYCHKA, D. W. (1987). Semi-nonparametric maximum likelihood estimation. *Econometrica* **55**, 363–390.

GINÉ, E. & GUILLOU, A. (2002). Rates of strong uniform consistency for multivariate kernel density estimators. In *Annales de l'IHP Probabilités et Statistiques*, vol. 38.

HOROWITZ, J. L. (2007). Asymptotic normality of a nonparametric instrumental variables estimator. *International Economic Review* **48**, 1329–1349.

KIM, J. K. & SHAO, J. (2013). *Statistical Methods for Handling Incomplete Data*. New York: Chapman and Hall/CRC.

LI, Q. & RACINE, J. S. (2007). *Nonparametric Econometrics: Theory and Practice*. Princeton University Press.

NEWEY, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics* **79**, 147–168.

NEWEY, W. K. & POWELL, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica* **71**, 1565–1578.

SHEN, X. (1997). On methods of sieves and penalization. *Ann. Statist.* **25**, 2555–2591.

SILVERMAN, B. W. (1984). Spline smoothing: the equivalent variable kernel method. *Ann. Statist.* **12**, 898–916.

VAN DER VAART, A. W. & WELLNER, J. A. (1996). *Weak Convergence and Emprical Processes: With Applications to Statistics*. New York: Springer.

YANG, S. & KIM, J. K. (2016). Fractional imputation in survey sampling: A comparative review. *Statistical Science* **31**, 415–432.