Discussion Paper Series

RIEB

Kobe University

DP2020-06

Semiparametric Bayesian Instrumental Variables Estimation for Nonignorable Missing Instruments

Ryo KATO
Takahiro HOSHINO

February 5, 2020

Research Institute for Economics and Business Administration
Kobe University
2-1 Rokkodai, Nada, Kobe 657-8501 JAPAN

Semiparametric Bayesian Instrumental Variables Estimation for Nonignorable Missing Instruments

Ryo Kato*

Research Institute for Economics and Business Administration, Kobe University, 2-1 Rokkodai-cho, Nada-ku, Kobe, Japan

and

Takahiro Hoshino†

Department of Economics, Keio University, 2-15-45 Mita, Minato-ku, Tokyo, Japan / RIKEN Center for Advanced Intelligence Project, 1-4-1 Nihonbashi, Chuo-ku, Tokyo, Japan

SUMMARY.This paper considers the case where instrumental variable (IV) are available to infer the effect of interested variable to the outcome (or the causal effect), but some components of IV are missing with the missing mechanism of not missing at random (NMAR). Although NMAR requires the analysis to prespecify the missing mechanism, it is unknown for us and what is worse, it is generally not identified. We use the IV distribution of original population as an auxiliary information, and show that missing mechanism can be represented as identifiable nonparametric generalized additive model. We also introduce MCMC algorithm that impute the missing values and simultaneously estimate parameters of interested.

KEY WORDS: Instrumental variable; missing not at random; auxiliary information.

1. Introduction

When it is not feasible to conduct randomized controlled trials (RCT) or quasi-randomized experiments, the IV approach can be a very useful tool to infer the causal effect if it is possible to find sufficient IVs. Since they can properly eliminate the confoundings caused by unobserved factors, IV models are developed and applied in many empirical economics researches, where unobserved confoundings are ubiquitous. Therefore, introducing sufficient

* *email:* kato ryo@rieb.kobe-u.ac.jp

† *email:* bayesian@jasmine.ocn.ne.jp

IVs can in itself be great invention, and many researchers are trying to find them.

Despite the desperate efforts, IVs tend to be missing. For example, information of twin are often used as an instrument, but it is only observed for sub-samples. Aaslund and Gronquist (2010) used twin birth as an instrument to survey the effect of family size on the quality of children. In this case, a twin birth can be observed for families with twins. Of course, the complete case analysis seems to result in biased results, and as an alternative approach, restricting the sample to families with more than two children lose the efficiency in sample size. Variables on children, such as child BMI, are also frequently employed as instruments. However, since child information comes from different sources than the endogenous parents' information (e.g. BMI), there is a tendency for the former to be missing. Real data analysis section is an example wherein instruments are missing for many observations since they are sourced from other surveys.

In addition to economics, missing IV is a common problem in other fields. Mendelian randomization uses genotype information as an instrumental variable to infer the causal effect of a biomarker to a disease. Since the appropriate genetic variant is independent of the confounders of the intermediate phenotype-outcome association and can affect the outcome only through the causal intermediate phenotype as long as it is related to the intermediate phenotype, it has recently been applied in economics as well as in biostatistics. In general, as genetic variants explain only a small portion of the endogenous population, Mendelian randomization requires large sample sizes (Smith, 2006) to satisfy enough causal associations. However, Mendelian randomization datasets are often missing (Palmer et al., 2012) and a large enough sample size cannot be guaranteed.

These example of missing IV shows us that complete case analysis, namely, only those samples where all the instruments are observed, are used in the analysis, thus resulting in biased results and wrong decision-making. In this paper, we develop a semiparametric method to impute the missing portion of IV and simultaneously infer the causal effect. As the most general case, we consider the case with not missing at random (NMAR). Therefore, missingness of instruments remains associated with the missing instruments even after controlling for other observed variables. In the NMAR case, the interested regression model cannot be identified without an additional assumption (Little and Rubin, 2002) . An example of such an assumption is strong parametric assumption on the regression and missing mechanism (Kott and Chang, 2010) . However, Miao et al. (2016) showed that probit specification

on the missing mechanism can identify normal and normal-mixture models while logit specification can less identify them.

Although there exists a lot of literature on IV, the model for missing instrument is scant (Kennedy and Small, 2017) except for (Burgess et al., 2011; Mogstad and Wiswall, 2012; Chaudhuri and Guilkey, 2016). Burgess et al. (2011) considered missing instruments for Mendelian randomization, and proposed a Bayesian multiple imputation method. However, they considered the case of missing at random. Mogstad and Wiswall (2012) proposed an IV estimator for partially missing instrument, but they also assume missing at random. Chaudhuri and Guilkey (2016) developed semiparametric efficient GMM method for missing IV, assuming IV missing at random. As another related work Ertefaie et al. (2017) considered the case of NMAR with observed confonders, but they assume IVs are completely observed. Therefore, there exist no literature which considered missing IV with NMAR.

We take an assumption that the IV distribution of the original population is available as an auxiliary information. In many cases, the population-level information is available from other data sources. Government statistics is such an example and some researches utilize this auxiliary information to estimate individual-level causality. Imbens and Lancaster (1994) and Hellerstein and Imbens (1999) incorporated population-level information as moment conditions to infer individual-level models using the generalized method of moments (GMM). Another instance where population-level information is used is the empirical likelihood estimation (Qin, 2000; Chaudhuri et al., 2008). Such approaches are also applied to the missing data issues. Nevo (2003) proposed the propensity score weighting method using the moment conditions obtained from auxiliary population-level information. Igari and Hoshino (2018) introduced the Bayesian method with population-level information that dealt with repeated durations under unobserved missing indicators.

Although prior works incorporating population-level information to deal with missing variables use moment conditions, our proposed method uses probability distribution of population as auxiliary information since the moment conditions have less information than original distribution. Under the condition that the original population distribution of the missing IV is known, followed by the theorem in Hirano et al. (2001), we show that the missing mechanism is nonparametrically identified with generalized additive model, and the substantive IV regression models are also identified. Figure 1 illustrates the model considered in this paper. In general, since fully nonparametric missing mechanism are not identified, parametric missing mechanism are

frequently assumed (Kott and Chang, 2010). However, misspecification of missing mechanisms results in severely biased estimates (Kim and Yu, 2011). Kim and Yu (2011) developed a semiparametric missing mechanism approach which incorporates nonparametric specifications on observed variables but not on unobserved variables. However, their method, as well as other prior works, cannot identify the nonparametric part of unobserved variables. On the other hand, we assume the availability of the information of the original population distribution of missing IV so that our proposed method can specify the fully nonparametric missing mechanisms on observed and unobserved variables. Furthermore, our missing mechanism can incorporate cross terms of observed and unobserved variables, which cannot be identified by the existing methods.

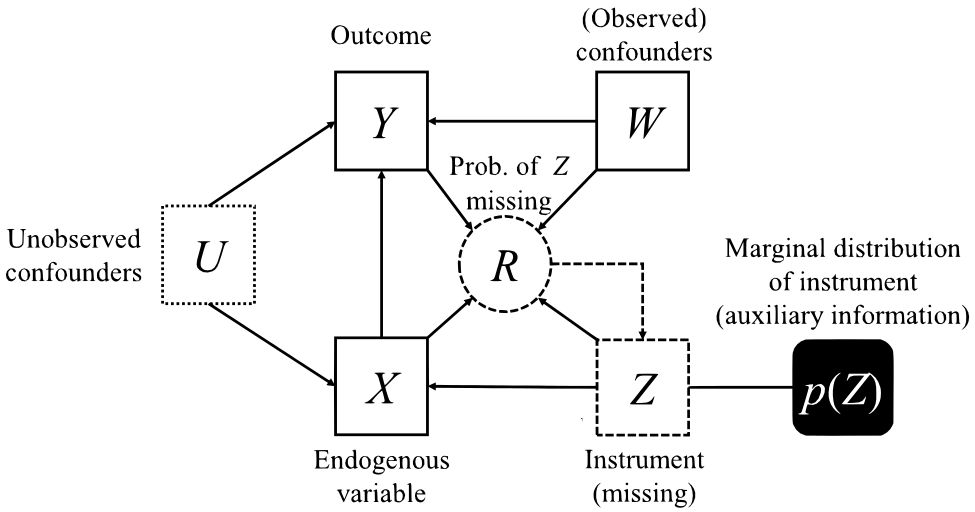

Fig.1 The model considered in this paper

2. The challenges of handling missing IV

2.1 *Setup and notations*

Before discussing the methodologies for handling datasets with missing IVs, we introduce the notations and setup.

IV approach aims to overcome the unmeasured confounders problem in order to make causal inference. We consider potential outcome approach, then define the causal effect. Let Y be the outcome variable, X be a treatment variable or endogenous variable. Let $Y_i^{(X^*)}$ denote the potential outcome that would be observed for unit i if the individual were to have the treatment level of X^*. For each unit, only one possible realization of $Y_i^{obs} = Y_i$ and $X_i^{obs} = X$ is observed. In this potential outcome approach, although each

unit has potential outocomes corresponding to the possible level of X, only one potential outcome $Y_i = Y_i^{(X^*)}$ can be observed. Let Z_i be a instrumental variable for unit i.

Let (Y_i, X_i, Z_i), $(i = 1, ..., N)$ be an i.i.d. sample. We consider IV regression model with one structural equation or "second-stage" and one reduced form equation or "first-stage" in the following form

$$\begin{cases} Y_i = M(X_i, \beta_0) + \epsilon_{1,i} \\ X_i = f(Z_i; \boldsymbol{\delta}) + \epsilon_{2,i} \end{cases}. \tag{1}$$

where $M(X_i, \beta_0)$ has only a finite-dimensional unknown parameter β_0, but $f(Z_i)$ are an unknown functions. We consider the situation where $\epsilon_{1,i}$ and $\epsilon_{2,i}$ are correlated, which causes the endogeneity problem. We assume Z_i to be independent of $\epsilon_{2,i}$ and $E(\epsilon_{1,i} Z_i) = 0$. We call

We consider the case where X and β are scalar, that is $M(X_i, \beta_0) = \beta_0 X_i$. In this case, our parameter of interest is the causal effect of increasing X by one unit: $\beta_0 = Y_i^{(X^*+1)} - Y_i^{(X^*)}$.

Although some recent literature consider the case that the function M is unknown and they estimate M nonparametrically, we assume $M(X_i, \beta_0) = \beta_0 X_i$ because nonparametric estimation of M require us to very strong identifiability assumptions (Newey and Powell, 2003; Hall and Horowitz, 2005; Darolles et al., 2011; Liao and Jiang, 2011; Kato, 2013).

We consider the missingness of instrumental variables Z. We assume that Z for some observations are missing and the other variables, Y and X are completely observed. We denote the missing indicator as R_i which takes 1 when the corresponding element of the Z_i is observed and 0 otherwise.

We assume the missing mechanism to be nonignorable (NMAR). We consider the case with the missing probability as depending on all the other observed variables $p(R = 1|Y, X, Z)$ (See also Figure 1).

2.2 *Missing instruments and complete case analysis*

The simplest and the most applied way to deal with missing IVs are thought to be complete case analysis. We consider the results obtained from complete case analysis using estimating equation. The estimating equation of complete case analysis is

$$E\left[I(R = 1)\epsilon_1\left\{Y - f(Z_i; \boldsymbol{\delta}_0)\beta\right\}\right]$$

where $\boldsymbol{\delta}_0$ is the true value of $\boldsymbol{\delta}$ on the first stage, and $I(R = 1) = 1$ for complete case and $I(R = 1) = 0$, otherwise.

If the probability of missing depends only on Z (NMAR), that is, $p(R=1|Z)$, then the estimates of β obtained from complete case analysis results has consistency, as shown from the following estimating equation since $E(\epsilon_1|Z)$ is assumed to be 0.

$$\begin{aligned}
& E\left[I(R=1|Z)\,\epsilon_1\left\{Y-f(Z_i;\boldsymbol{\delta}_0)\,\beta\right\}\right] \\
= \; & E_Z\left(E_Y\left[I(R=1|Z)\,\epsilon_1\left\{Y-f(Z_i;\boldsymbol{\delta}_0)\,\beta\right\}|Z\right]\right) \\
= \; & \int_Z p(R=1|Z)\,\epsilon_1\left[\int_Y \left\{Y-f(Z_i;\boldsymbol{\delta}_0)\,\beta\right\} f(Y|Z)\,dY\right] f(Z)\,dZ \\
= \; & \int_Z p(R=1|Z)\,\epsilon_1\left[\int_{\epsilon_1} \epsilon_1 f(\epsilon_1|Z)\,d\epsilon_1\right] f(Z)\,dZ \\
= \; & 0
\end{aligned}$$

On the other hand, for example, if the probability of missing only depends on X (MAR), that is, $p(R=1|X)$,

$$\begin{aligned}
& E\left[I(R=1|X)\,\epsilon_1\left\{Y-f(Z_i;\boldsymbol{\delta}_0)\,\beta\right\}\right] \\
= \; & E_X\left(E_Y\left[I(R=1|X)\,\epsilon_1\left\{Y-f(Z_i;\boldsymbol{\delta}_0)\,\beta\right\}|X\right]\right) \\
= \; & \int_X p(R=1|X)\,\epsilon_1\left[\int_Y \left\{Y-f(Z_i;\boldsymbol{\delta}_0)\,\beta\right\} f(Y|X)\,dY\right] f(X)\,dX \\
= \; & \int_z p(R=1|Z)\,\epsilon_1\left[\int_{\epsilon_1} \epsilon_1 f(\epsilon_1|X)\,d\epsilon_1\right] f(Z)\,dZ \\
\neq \; & 0
\end{aligned}$$

shows complete case analysis results in inconsistent result since $E(\epsilon_1|X) \neq 0$.

In the setting, since $E\left[\delta(R=1|Y,X,Z)\,\epsilon_1\left\{Y-f(z;\boldsymbol{\delta}_0)\,\beta\right\}\right] \neq 0$, the complete case analysis results in biased estimates. Therefore, we must specify the missing mechanism to obtain the consistent estimates of the interested parameters.

2.3 *Missing instruments and existing imputation methods*

As stated above, no method has proposed to impute missing instruments when the pattern of missingness is NMAR. We review some imputation methods which can just imputed missing components, though they are not expected to obtain consistent estimates.

2.3.1 Multiple imputation by chained equation Multiple imputation by chained equation (MICE) (van Buuren, 2007) is one of the most applied imputation method since researchers can avoid the difficulty in specifying the

distributional assumptions. MICE specifies a multivariate distribution by a sequence of univariate regressions for each missing variable. Moreover, it can easily implement using existing several software packages, such as the "mice" package in R, "proc mi" with the FCS option in SAS, and "mi impute" in STATA.

However, Liu et al. (2014) showed that the imputations with MICE do not guarantee the asymptotic distributions to be consistent with the existing Bayesian joint model multiple imputation estimator when the family of conditional models and their joint distributions are not "compatible". The violation of compatible is ubiquitous and MICE is shown to result in severely biased estimated in many cases using simulation studies (Kato and Hoshino, 2019). What is worse, since it do not specify the pattern of missingness which is required to NMAR, MICE result in biased in the case. Then the situation considered here are not appropriate for MICE. In fact, the simulation study conducted in this paper shows considerable biased results when missing IV is imputed using MICE.

2.3.2 Imputation by machine learning method Recently, machine leaning based method has been applied to impute the missing values. Especially, Stekhoven and Buhlmann (2012) proposed the missForest algorithm, in which the missing values are imputed by the predictors from the random forest. Since missForest predicts missing values based on random forest, it accommodate interactions and nonlinearities. Therefore, we do not have to specify a particular regression model for imputation. missForest is often implemented to the real world dataset since they are reported to provide lower imputation errors than the FCS method (Waljee et al., 2013; Liao et al., 2014).

However, it is reported that missForest approach results in biased estimates in some cases when the number of covariates are small (Kato and Hoshino, 2019). Moreover, even if the algorithm correctly estimates the mean, it always underestimates the variance of the estimates and results in poor CIs (Shah et al., 2014) since they do not rely on the probabilistic model. Therefore, it is not suitable for the case applying IV method where the results of statistical significance (or hypothesis testing) are crucial (such as economics, medical, or epidemiological research). In fact, the simulation study conducted in this paper shows considerable biased results when missing IV is imputed using missForest.

3. Semiparametric Bayes model for instrumental variable with nonignorable missing

In this section, we proposed semiparametric Bayesian model for instrumental variable with nonignorable missing that overcome the shortcomings of existing imputation methods.

3.1 *Identification*

We are interested in recovering the joint distribution of (Y, X, Z) or possibly the conditional distribution $(Y, X|Z)$ with missing problem of Z. Since

$$p(Y, X, Z) = \frac{f(Y, X, Z|R = 1)\, p(R = 1)}{p(R = 1|Y, X, Z)}$$

and $f(Y, X, Z|R = 1)$ and $p(R = 1)$ can be directly estimated, identification of $p(Y, X, Z)$ depends on the identification of $p(R = 1|Y, X, Z)$. Because the missing probability of Z depends on Z itself: $p(R = 1|Y, X, Z)$, this is NMAR case.

Generally, the identification of missing mechanism is difficult under NMAR (Little and Rubin, 2002). However, by considering following assumptions, we can identify the model.

Theorem 1:
Assume that the following conditions hold.
$(A1)$ *the support of* $f(y, x, z)$ *coincides with* $f(y, x, z|R = 1)$;
$(A2)$ $f(y, x, z)$ *and* $f(y, x, z|R = 1)$ *are square integrable;*
$(A3)$ *the marginal distribution of* Z, $p(Z)$ *is known.*
Then the missing mechanism is identified through the form of

$$p(R_i = 0|Y = y, X = x, Z = z) = h(k_0 + k_1(y, x) + k_2(z)) \qquad (2)$$

where h *is a known function which is differentiable, strictly increasing with* $\lim_{a\to-\infty} g(a) = 0$ *and* $\lim_{a\to\infty} g(a) = 1$, *and* $k_1(\cdot), k_2(\cdot), k_3(\cdot)$ *are unique set of functions subject to normalization* $k_1(0) = k_2(0) = k_3(0) = 0$.

Proof:
The proof for theorem 1 is very similar to that of Hirano et al. (2001). By regarding the set (Y, X) as time-variant completely observed variables, and Z as time-variant incompletely observed variable as in the variable definition of Hirano et al. (2001), we can straightforwardly obtain the desired result.

We can also include measured confounders (exogenous variables) in the second stage regression model. We denote W_i as a k-dimensional vector of

measured confounders which contains Z_i. We consider following system of equations

$$\begin{cases} Y_i = M\left(X_i, \beta_0\right) + g\left(W_i; \gamma\right) + \epsilon_{1,i} \\ X_i = f\left(Z_i; \delta\right) + \epsilon_{2,i} \end{cases}. \tag{3}$$

where $g(\cdot)$ is an unknown functions. With this specification, we can identify the missing mechanism with some assumptions.

Theorem 2

Assume that the following conditions hold.

$(A1')$ *the support of* $f(y, x, z|w)$ *coincides with* $f(y, x, z|w, R = 1)$;

$(A2')$ $f(y, x, z|w)$ *and* $f(y, x, z, w|R = 1)$ *are square integrable almost surely with respect to* w;

$(A3')$ *the marginal distribution of* Z, $p(Z)$ *is known.*

Then the missing mechanism is identified through the form of

$$p\left(R_i = 0|Y = y, X = x, Z = z, W = w\right) = h(k_0'\left(w\right) + k_1'\left(y, x, w\right) + k_2'\left(z, w\right)) \tag{4}$$

where h *is a known function which is differentiable, strictly increasing with* $\lim_{a\to-\infty} g(a) = 0$ *and* $\lim_{a\to\infty} g(a) = 1$, *and* $k_1(\cdot), k_2(\cdot), k_3(\cdot)$ *are unique set of functions subject to normalization* $k_1'(0) = k_2'(0) = k_3'(0) = 0$.

Proof:

The proof for theorem 2 is the same as that of Hirano et al. (2001). By regarding the set (Y, X) as time-variant completely observed variables, Z as time-variant incompletely observed variable, and W as time-invariant covariates as in the variable definition of Hirano et al. (2001), we can straightforwardly obtain the desired result.

The advantages of this missing mechanism specification is: (i) we can specify nonparametric forms on the observed variables and missing variables (IVs), and (ii) we can consider cross term effect of observed variables and missing variables. Neither of these formulations are identified by the semiparametric models proposed by Kott and Chang (2010) or Kim and Yu (2011) (Because the method of Kott and Chang (2010) is developed for nonignorable nonresponse and Kim and Yu (2011) is for mean functionals, they are not applied to IV regression).

3.2 *Semiparametric formulation*

We formulate the conditional distribution of the missing IV z given all the other variables as follows.

$$
\begin{aligned}
p\left(Z|Y,X,W,R=0\right) &= \frac{p\left(R=0|Y,X,W,Z\right)p\left(Z|Y,X,W\right)}{\int_z p\left(R=0,Z|Y,X,W\right)dZ} \\
&\propto p\left(R=0|Y,X,W,Z\right)p\left(Y,X|W,Z\right)p\left(Z|W\right)p\left(Z\right)
\end{aligned}
$$

where $p\left(R=0|Y,X,W,Z\right)$ is the missing mechanism, $p\left(Y,X|W,Z\right)$ is the substantive IV model (the structural equation and the reduced-form equation), $p\left(Z|W\right)$ is conditional instrumental variables distribution, and $p\left(Z\right)$ is known and true marginal distribution of Z.

As stated in the identification issue, we can specify nonparametric missing mechanism for $p\left(R=0|Y,X,W,Z\right)$ using generalized additive model in equation (4). Needless to say, parametric probit or logistic regression can be prespecified to $p\left(R=0|Y,X,W,Z\right)$.

3.2.1 Substantive model and conditional IV distribution We formulate the substantive model $p\left(Y,X|Z,W\right)=p\left(Y|X,W\right)p\left(X|Z\right)$ in the semiparametric form. As stated, $p\left(X|Z\right)$ represents the first stage: $X_i=f\left(Z_i;\delta\right)+\epsilon_{2,i}$, and $p\left(Y|X,W\right)$ represents the second stage: $Y_i=M\left(X_i,\beta_0\right)+g\left(W_i;\gamma\right)+\epsilon_{1,i}$ in equation (3), respectively. Therefore, we consider $p\left(Y,X|Z,W\right)$ to be semiparametric IV regression model. Moreover, we specify the conditional IV distribution $p\left(Z|W\right)$ to be nonparametric.

We use DPM representation to represent semiparametric IV regression model. Fortunately, the DPM model can be estimated with a relatively simpler MCMC algorithm by applying blocked Gibbs sampling (Ishwaran and James, 2011). Therefore, DPM modeling has been often employed to avoid distributional assumptions on the parameters within the Bayesian framework (e.g. Hirano, 2002; Xue et al., 2007; Conley et al., 2008; Shahbaba and Neal, 2009). The theoretical properties of DPM were investigated in Shen et al. (2013).

Dirichlet process prior is assumed for a random distribution G, denoted as $G\sim DP\left(\eta,G_0\right)$ is expressed as follows:

$$
G=\sum_{l=1}^{\infty}\pi_l\delta_{\theta_l},\ \theta_l\sim G_0
$$

where η is the concentration parameter and G_0 is the base distribution. δ_θ is a point mass at θ and

$$
\pi_l=\xi_l\prod\nolimits_{h<l}\left(1-\xi_h\right)
$$

where $\xi_l \sim Be(1, \alpha)$.

By applying DPM, the resulting multivariate regression function of y_i, x_i, and z_i is represented by following mixture model;

$$\begin{aligned} p(y_i, x_i, z_i|w_i) &= p(y_i, x_i|z_i, w_i)\, p(z_i|w_i) \\ &= \sum_{l=1}^{\infty} \pi_l N\left(\widetilde{\boldsymbol{y}}_i|\widetilde{\boldsymbol{X}}_i\widetilde{\beta}_l, \boldsymbol{\Sigma}_l\right) N(z_i|\boldsymbol{\Gamma}_l w_i, \boldsymbol{\Phi}_l) \end{aligned}$$

where

$$\widetilde{\boldsymbol{y}}_i = \begin{pmatrix} y_i \\ x_i \end{pmatrix}, \widetilde{\boldsymbol{X}}_i = \begin{pmatrix} x_i & w_i & \mathbf{0} \\ 0 & \mathbf{0} & z_i \end{pmatrix}, \widetilde{\boldsymbol{\beta}}_l = \begin{pmatrix} \beta \\ \boldsymbol{\gamma}_l \\ \boldsymbol{\delta}_l \end{pmatrix},$$
$$\widetilde{\boldsymbol{\epsilon}}_l = \begin{pmatrix} \epsilon_{1,l} \\ \epsilon_{2,l} \end{pmatrix}, \widetilde{\boldsymbol{\epsilon}}_l \sim N(\mathbf{0}, \boldsymbol{\Sigma}_l).$$

4. MCMC algorithm

Blocked Gibbs sampler (Ishwaran and James, 2011) is applied to the posterior computation of the DPM parameters. It is proved that the case with finite number of classes l can be used to approximate the inference that is based on infinite classes with satisfactory accuracy when the maximum number of classes L is large enough. Therefore, blocked Gibbs sampler considers the case for truncation of the number of classes (e.g. $L = 20$), hence the simpler posterior computation.

We obtain the detailed posterior computation using the MCMC estimation as follows.

1. Conditional for K_i $(i = 1, ..., N)$

   Let K_i be the indicator denoting where case i belongs, and $K_i = l$ if case i belongs to class l. To assign samples to each class, generate K_i by $\sum_{l=1}^{L} \pi_{li}\delta_l(\cdot)$, where π_{li} is

   $$\pi_{li} = \frac{\pi_l N\left(\widetilde{\boldsymbol{y}}_i|\widetilde{\boldsymbol{X}}_i\widetilde{\boldsymbol{\beta}}_l, \boldsymbol{\Sigma}_l\right) N(z_i|\boldsymbol{\Gamma}_l \boldsymbol{w}_i, \boldsymbol{\Phi}_l)}{\sum_{l=1}^{L} \pi_l N\left(\widetilde{\boldsymbol{y}}_i|\widetilde{\boldsymbol{X}}_i\widetilde{\boldsymbol{\beta}}_l, \boldsymbol{\Sigma}_l\right) N(z_i|\boldsymbol{\Gamma}_l \boldsymbol{w}_i, \boldsymbol{\Phi}_l)},$$

   with $\pi_l = \xi_l \prod_{h<l}(1 - \xi_h)$.

2. Conditional for $\alpha(l = 1, ..., L-1)$

   Simulate α_l from the following normal distribution.

$$\alpha_l \sim N\left(\frac{\sum_{i:K_i \geq l} W_{il}^* + \mu_v}{N_l + 1}, \frac{1}{N_l + 1}\right),$$

3. Update $\boldsymbol{\Gamma}_l, \boldsymbol{\Phi}_l$

   Draw $\boldsymbol{\Gamma}_l$ and $\boldsymbol{\Phi}_l$ from the following multivariate normal and inverted Wishart distribution.

$$\begin{aligned}
\boldsymbol{\Gamma}_l|rest &\sim N\left(vec(\widehat{\boldsymbol{\Gamma}}), \boldsymbol{\Phi}_l \otimes \left(\boldsymbol{W}_l^T \boldsymbol{W}_l\right)^{-1}\right), \\
\boldsymbol{\Phi}_l|rest &\sim IW\left(f_0 + N, \boldsymbol{G}_0^{-1} + (\boldsymbol{Z}_l - \boldsymbol{\Gamma}_l \boldsymbol{W}_l)^T (\boldsymbol{Z}_l - \boldsymbol{\Gamma}_l \boldsymbol{W}_l)\right),
\end{aligned}$$

   where $\widehat{\boldsymbol{\Gamma}} = \left(\boldsymbol{W}_l^T \boldsymbol{W}_l\right)^{-1} \boldsymbol{W}_l \boldsymbol{Z}_l$, $\boldsymbol{W}_l = \left(\boldsymbol{w}_1^T, ..., \boldsymbol{w}_N^T\right)^T$, $\boldsymbol{Z}_l = \left(z_1^T, ..., z_N^T\right)^T$, and $\boldsymbol{W}_l$ and $\boldsymbol{Z}_l$ denote the subset of $\boldsymbol{W}_l$ and $\boldsymbol{Z}_l$ whose case i belong to class l. f_0 and $\boldsymbol{G}_0^{-1}$ denotes the parameter of the prior distribution of $\boldsymbol{\Gamma}_l$; $\boldsymbol{\Gamma}_l \sim IW(f_0, \boldsymbol{G}_0^{-1})$.

4. Update the missing mechanism

   If we assume parametric probit or logistic model to the missing mechanism, we can use Gibbs sampling proposed by Albert and Chib (1993) for probit and M-H algorithm for the logistic, given the pseudo-complete dataset. If we assume generalized additive model to the missing mechanism, we can directly employ Bayesian GAM estimation proposed by Klein et al. (2015) given the pseudo-complete dataset.

5. Update the missing components

   Draw the missing component of $\boldsymbol{z}_i$ from a density proportional to $p\left(z_i|y_i, x_i, \boldsymbol{w}_i, \boldsymbol{r}_i = 0\right)$. Since it is difficult to draw the missing z_i, we employ the Metropolis-Hastings algorithm and use $p\left(z_i|\boldsymbol{w}_i\right)$ as a proposal density in order to draw a candidate of z_i, z_i^c. We accept the candidates with the following probability:

$$\min\left[1, \frac{p\left(r_i = 1|y_i, x_i, \boldsymbol{w}_i, z_i^*\right) p\left(y_i, x_i|z_i^*, \boldsymbol{w}_i\right) p(z_i^*|\boldsymbol{w}_i)}{p\left(r_i = 1|y_i, x_i, \boldsymbol{w}_i, z_i^{l-1}\right) p\left(y_i, x_i|z_i^{l-1}, \boldsymbol{w}_i\right) p(z_i^{l-1}|\boldsymbol{w}_i)}\right]$$

6. Conditional for $\boldsymbol{\delta}_l$

The standardized reduced model is used to draw. Substituting in for x in the second stage regression, the followings are obtained

$$\begin{pmatrix} x_i \\ y_i - \boldsymbol{w}_i\boldsymbol{\gamma}_l \end{pmatrix} = \begin{pmatrix} z_i \\ \beta z_i \end{pmatrix} \boldsymbol{\delta}_l + \upsilon_i$$

where $Var(\upsilon_i) = \begin{bmatrix} 1 & 0 \\ \beta & 1 \end{bmatrix} \boldsymbol{\Sigma}_l \begin{bmatrix} 1 & \beta \\ 0 & 1 \end{bmatrix}$. Then we draw $\boldsymbol{\delta}_l$ using Bayesian regression with unit normal errors.

7. Conditional for β and $\boldsymbol{\gamma}_l$

   The structural model parameters β and $\boldsymbol{\gamma}$ can be simulated by applying Bayesian linear regression draw with $N(0,1)$ error terms

$$\frac{y_i - E\left[\epsilon_{1,i}|\epsilon_{2,i}\right]}{\sigma_{\epsilon_{1,i}|\epsilon_{2,i}}} = \frac{x_i\beta}{\sigma_{\epsilon_{1,i}|\epsilon_{2,i}}} + \frac{\boldsymbol{w}_i\boldsymbol{\gamma}}{\sigma_{\epsilon_{1,i}|\epsilon_{2,i}}} + u_i, \quad u_i \sim N(0,1)$$

   where

$$\sigma_{\epsilon_{1,i}|\epsilon_{2,i}} = \sigma_{11,\omega_i} - \frac{\sigma^2_{12,\omega_i}}{\sigma^2_{11,\omega_i}}$$

   and

$$E\left[\epsilon_{1,i}|\epsilon_{2,i}\right] = \frac{\sigma_{12,\omega_i}}{\sigma_{22,\omega_i}}\epsilon_{2,i}$$

   When we draw β, we consider linear regression model with standard normal error

$$y_i^{\boldsymbol{w}} = \frac{x_i\beta}{\sigma_{\epsilon_{1,i}|\epsilon_{2,i}}} + u_i, \quad u_i \sim N(0,1)$$

   where $y_i^{\boldsymbol{w}} = \left(y_i - E\left[\epsilon_{1,i}|\epsilon_{2,i}\right] - \boldsymbol{w}_i\boldsymbol{\gamma}_l\right)/\sigma_{\epsilon_{1,i}|\epsilon_{2,i}}$.

   When we draw $\boldsymbol{\gamma}_l$, we consider linear regression model with standard normal error

$$y_i^{\boldsymbol{x}} = \frac{\boldsymbol{w}_i\boldsymbol{\gamma}_l}{\sigma_{\epsilon_{1,i}|\epsilon_{2,i}}} + u_i, \quad u_i \sim N(0,1)$$

   where $y_i^{\boldsymbol{x}} = \left(y_i - E\left[\epsilon_{1,i}|\epsilon_{2,i}\right] - x_i\beta\right)/\sigma_{\epsilon_{1,i}|\epsilon_{2,i}}$.

5. Simulation study

We conduct two simulation studies in order to illustrate the performance of the proposed method when some components of instruments are missing with NMAR and the original population information is available. Simulation 1 considers the case where the original population distribution of IV is log-normally distributed with a linear reduced-form equation and additive disturbance that follows the Gaussian distribution. The missing mechanism contains the cross-term of IV and observed variables, which may not be identified by existing methods that do not have population-level information. Simulation 2 considers the case where the original population distribution of IV is log-normally distributed with the mixture of two classes' ($l = 2$) reduced-form equation. The missing mechanism also contains the cross-term of IV and observed variables. Through the simulation study, 30% of the incomplete covariates are set to be missing. We generate the missing values based on $p\left(r_i = 0|y_i, x_i, w_i, z_i\right)$ (see Appendix for detailed missing probability setting).

In the two simulation studies, we evaluate the finite sample property of the estimators. We consider $1,000$ replications of the dataset and confirm if the true value of endogenous variable coefficient $\beta = 1$ can be recovered. Throughout the simulations, the sample size for each dataset is set to be $N = 400$. We calculate standard performance measures, such as empirical mean, standard deviation, the coverage of nominal 95% confidence (or credible) intervals (CI) of the estimate, and the deviation (MSE) from the true value. Although it could be sometimes inappropriate, we take more practical perspective to compare the Bayesian credible intervals with the classical confidence intervals in Conley et al. (2008).

We evaluate the proposed method by comparing it with other data imputation methods, i.e., MICE-GMM and "pseudo-proposed method" (Ps-Proposed). As described, the MICE approach specifies a multivariate covariate distribution by a sequence of univariate regressions for each missing variable. In our simulation settings, where the conditional models (set of univariate regressions) and their joint distributions are incompatible, MICE estimates do not guarantee consistency. After creating multiple datasets from MICE, the IV estimates are obtained by GMM (generalized methods of moments). We call this competitor MICE-GMM. We also employed missForest imputation followed by IV estimation with GMM (MF-GMM). "Pseudo-proposed method" (Ps-Proposed) is the alternative which does not use the original population information, but where the missing instruments are imputed from $p\left(z_i|y_i, x_i, w_i, r_i = 0\right)$. Ps-Proposed might not have identi-

fication. We also compared this method with IV estimates from other classical methods based on complete case analysis by GMM (CC-GMM), and OLS estimates which ignore the existence of endogeneity (CC-OLS). As for Bayesian methods, burn-in of 5,000 iterations followed by 10,000 iterations were used for the posterior inference. MICE-GMM and CC-GMM assumes an ordinary linear regression model for the reduced-form equation, that is, $x_i = z_i\boldsymbol{\delta} + \epsilon_{2,i}$. MICE-GMM and GMM-CC used the moment condition of $E(\epsilon_{1,i}z_i) = 0$, therefore GMM based methods do not prespecify the functional form of the reduced-form equation. We used inverse variance of the moments as a weighting matrix, namely, the optimal weighting matrix. We calculated the intervals based on large sample approximations for the classical estimators.

This section presents the results of simulation study 1. The detailed simulation settings and the results of simulation 2 are provided in the Appendix.

Table 1 shows the results of simulation 1, including the empirical mean, standard deviation, the coverage of nominal 95% confidence intervals (CI) of the estimate, and the mean squared error (MSE) from the true value of β. Figure 2 shows the box plot of the estimates for β obtained by $1,000$ replications.

Table 1 Results of Simulation 1

	proposed				Ps-Proposed				MICE-GMM			
TRUE	Mean	(s.d.)	Cov	MSE	Mean	(s.d.)	Cov	MSE	Mean	(s.d.)	Cov	MSE
β = 1	1.023	(0.043)	94.2%	0.0021	2.192	(0.547)	32.3%	2.4652	1.190	(0.166)	89.8%	0.0692
MSE ratio		1.000				1164.345				32.674		

	MF-GMM				CC-GMM				CC-OLS			
TRUE	Mean	(s.d.)	Cov	MSE	Mean	(s.d.)	Cov	MSE	Mean	(s.d.)	Cov	MSE
β = 1	1.196	(0.168)	84.4%	0.079	0.859	(2.303)	99.6%	1.895	0.866	(0.031)	1.1%	0.019
MSE ratio		37.518				895.250				8.968		

Empirical mean, standard deviation, coverage of nominal 95% CIs, and mean squared error of the estimates, and MSE Ratio of the proposed basis from 1000 simulations are described. Ps-Proposed, pseudo-proposed method which does not use population-level information; MICE-GMM, IV estimates by generalized method of moments based on multiple imputation by chained equation; MF-GMM, IV estimates by generalized method of moments based on missForest imputation; CC-GMM, IV estimates by generalized method of moments based on complete case; CC-OLS, ordinary least squares regression based on complete case

As can be seen from the table and figure, all estimation methods except the proposed seem to be biased. Since Ps-Proposed do not have the information of the original distribution of IV, the missing mechanism cannot be identified and the obtained estimates are upwardly biased. The coverage of 95% credible intervals is also poor at 32.3%. As a result, the quasi-proposed

yields an MSE over 1,000 times larger compared with the proposed. The results indicate that population-level information is very useful to NMAR IV regression case. Another imputation approach is MICE-GMM, which also shows biased results. The biased results seem to be caused by incompatibility. Therefore, the standard deviation and the MSE are much larger than the proposed method.

The complete case analysis-based methods also show poor results. CC-GMM estimates have large standard deviation and range in very wide region. The MSE is about 900 times larger than the proposed method. The complete case analysis results as biased are indicated in Section 2.2. The coverage of 95% confidence intervals of CC-OLS is only 1.1%. This is because CC-OLS ignores the existence of IV and missing values.

As can be seen from comparison with the results from the complete data, as well as the competing methods, the proposed semiparametric approach shows very good estimates.

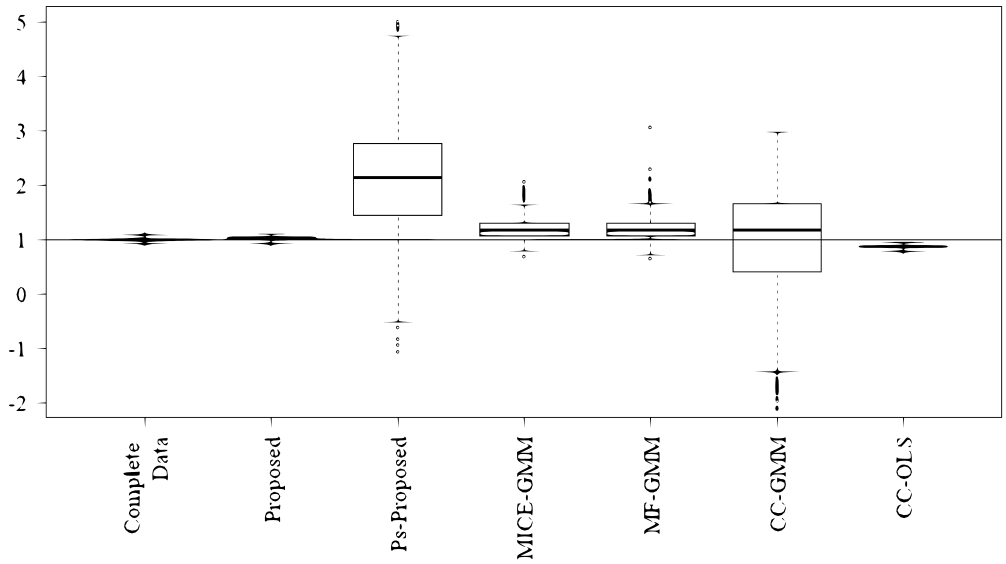

Fig.2 The solid horizontal line is the "true" coefficient value $\beta = 1$. The boxes span the range from the 25th to the 75th percentiles, and the whiskers extend to an area no more than 1.5 times the range from the 25th to the 75th percentiles from the box. The circles above and below the whiskers represent outliers. Complete, complete case analysis; Ps-Proposed, pseudo-proposed method which does not use population-level information; MICE-GMM, IV estimates by generalized method of moments based on multiple imputation by chained equation; MF-GMM, IV estimates by generalized method of moments based on missForest imputation;

CC-GMM, IV estimates by generalized method of moments based on complete case; CC-OLS, ordinary least squares regression based on complete case

6. Real data analysis

We applied our methods to an example from Acemoglu et al. (2001), which consider the effect of institutions on economic performance. Acemoglu et al. (2001) surveyed the causal relationship between "institution" (such as, more secure property rights and less distortionary policies) and "economic performance" proxied by GDP per capita. It is indicated that better institution and GDP per capita has positive relationship, but the causal relationship had not been proven. To solve the endogeneity problem, Acemoglu et al. (2001) employed European settlers mortality as an instrument. They argue that countries with higher mortality rate are fixed to be extractive institutions.

We used the same dataset as Acemoglu et al. (2001) and complete case analysis results in $N = 64$. However, if instruments are available, 111 samples are ready for analysis given they completely observed other variables. The outcome in this dataset is natural logarithm of GDP ($logGDP$), and the endogenous variable is Average Protection Against Ex-propriation Risk ($APER$). Available exogenous variables are latitude and country dummy (Asia, Africa, and other countries). We use the natural logarithm of annualized European settlers' mortality per thousand mean strength ($logSM$) as an instrument, of which 36.9% are missing. Since Acemoglu et al. (2001) constructed the IV sourced from and estimated by other research, i.e., Curtin (1989) and Curtin (1998), it tends to be missing. The F-statistic of reduced-form equation is 2.25, indicating weak instruments; hence, using more samples is effective in obtaining reliable results.

As population-level information, we use the information obtained from Albouy (2012), which improved the data for $logSM$. Based on AIC fitted to Albouy (2012)'s $logSM$, we use log-normal distribution with mean parameter 1.50 and standard deviation parameter 0.28 (AL-ln), and normal distribution with mean 4.65 and standard deviation 1.24 (AL-n). Another candidate is the complete case of Acemoglu et al. (2001), which has log-normal distribution with mean parameter 1.47 and standard deviation parameter 0.35 (AC-ln). Figure 3 describes the density of AL-ln, AL-n, and AC-ln.

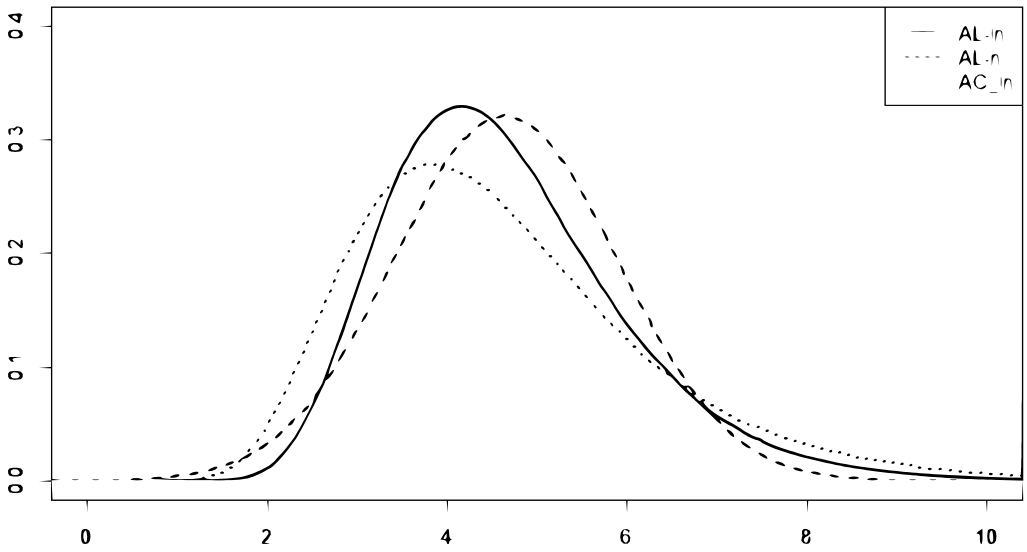

Fig.3 Probability density function of assumed population distribution of logSM based on AL-ln, AL-n, and AC-ln.

Burn-in of 25,000 iterations followed by 50,000 iterations were used for the posterior inference.

We also compared the results obtained from the pseudo-proposed method with do not consider the original population distribution (Ps-Proposed), MICE-GMM, FM-GMM, CC-GMM.

The results are described in Table 2. Three estimates based on the proposed method (AL-ln, AL-n, and AC-ln) imputed the missing values, and none of the three have 0 for their credible intervals. However, they employed different distributional information and all the posteriors are different. Pseudo-proposed methods, which do not have identification, yield uninterpretable results. MICE-GMM and CC-GMM show significant coefficients, but the means are smaller than the proposed methods (AL-ln, AL-n, and AC-ln). MF-GMM shows the smallest coefficient with statistical significance. As discussed before, the estimates from missForest result in smaller standard deviation, hence the obtained s.d. seems to abnormally small.

Given that the original population distributions are true, the causal effects of $APER$ to GDP is much larger than the prior surveys.

Table 2 Estimates results of Acemoglu et al. (2001) dataset

	coef.	(s.d.)	upper-CI	lower-CI
AL-ln	1.548	(0.542)	2.611	0.485
AL-n	1.304	(0.559)	2.401	0.208
AC-ln	1.072	(0.495)	2.043	0.101
Ps-Proposed	1.495	(5.651)	12.571	-9.580
MICE-GMM	1.121	(0.452)	2.008	0.234
MF-GMM	0.862	(0.176)	1.206	0.518
CC-GMM	1.100	(0.460)	2.002	0.198

7. Discussion

We have proposed semiparametric Bayes IV method which incorporates the missing instruments with NMAR using the original population information of IV. In this paper, we assume that the missing IV is one-dimensional. Within the field of economics, our method can deal with several datasets, since almost all the data have only one IV. However, if we consider application to other fields such as biometrics, the extension to multiple instruments might be required. As described, for example, Mendelian randomizations use multiple instruments, and IV used for them can be missing since they are genetic information. We should consider another identification condition for multiple IVs with NMAR. The condition might vary by assumptions, for instance, if joint distribution of original population is available, or if just the original marginal distribution of each IV is available.

Acknowledgements

This work was supported by the JSPS Grant-in-Aid for Research Activity Start-Up (18H05697) and the JSPS KAKENHI (18H03209; 17K18598).

References

Aaslund, O. and Gronquist, H. (2010). Family size and child outcomes: is there really no trade-off. *Labour Economics* **17**, 130–139.

Acemoglu, D., Johnson, S. and Robinson, J. A. (2001). The colonial origins of comparative development: An empirical investigation. *American Economic Review* **91**, 1369–1401.

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association* **88**, 669–679.

Albouy, D. Y. (2012). The colonial origins of comparative development: an empirical investigation: comment. *American Economic Review* **102**, 3059–3076.

Burgess, S., Seaman, S., Lawlor, D. A., Casas, J. P. and Thompson, S. G. (2011). Missing data methods in mendelian randomization studies with multiple instruments. *American Journal of Epidemiology* **174**, 1069–1076.

Chaudhuri, S. and Guilkey, D. K. (2016). Gmm with multiple missing variables. *Journal of Applied Econometrics* **31**, 678–706.

Chaudhuri, S., Handcock, M. S. and Rendall, M. S. (2008). Generalized linear models incorporating population level information: an empirical-likelihood-based approach. *Journal of the Royal Statistical Society: Series B* **70**, 311–328.

Conley, T. G., Hansen, C. B., McCulloch, R. E. and Rossi, P. E. (2008). A semi-parametric bayesian approach to the instrumental variable problem. *Journal of Econometrics* **144**, 276–305.

Curtin, P. D. (1989). *Death by migration: Europe's encounter with the tropical world in the nineteenth century*. Cambridge University Press.

Curtin, P. D. (1998). *Disease and empire: The health of European Troops in the Conquest of Africa*. Cambridge University Press.

Darolles, S., Fan, Y., Florens, J. P. and Renault, E. (2011). Nonparametric instrumental regression. *Econometrica* **79**, 1541–1565.

Ertefaie, A., Flory, J. H., S, H. and D., S. (2017). Instrumental variable methods for continuous outcomes that accommodate nonignorable missing baseline values. *American Journal of Epidemiology* **185**, 1233–1239.

Hall, P. and Horowitz, J. L. (2005). Nonparametric methods for inference in the presence of instrumental variables. *Annals of Statistics* **33**, 2904–2929.

Hellerstein, J. K. and Imbens, G. W. (1999). Imposing moment restrictions from auxiliary data by weighting. *Review of Economics and Statistics* **81**, 1–14.

Hirano, K. (2002). Semiparametric Bayesian inference in autoregressive panel data models. *Econometrica* **70**, 781–799.

Hirano, K., Imbens, G. W., Ridder, G. and Rubin, D. B. (2001). Combining panel data sets with attrition and refreshment samples. *Econometrica* **69**, 1645–1659.

Igari, R. and Hoshino, T. (2018). A bayesian data combination approach for repeated durations under unobserved missing indicators: Application

to interpurchase-timing in marketing. *Computational Statistics and Data Analysis* **126**, 150–166.

Imbens, G. W. and Lancaster, T. (1994). Combining micro and macro data in microeconometric models. *Review of Economic Studies* **61**, 655–680.

Ishwaran, H. and James, L. F. (2011). Gibbs sampling methods for stick-breaking priors. *Journal of American Statistical Association* **96**, 161–173.

Kato, K. (2013). Quasi-bayesian analysis of nonparametric instrumental variables models. *Annals of Statistics* **41**, 2359–2390.

Kato, R. and Hoshino, T. (2019). Semiparametric bayesian multiple imputation for regression models with missing mixed continuous-discrete covariates. *Annals of the Institute of Statistical Mathematics, in press* .

Kennedy, E. H. and Small, D. S. (2017). Paradoxes in instrumental variable studies with missing data and one-sided noncompliance. *arXiv* page 1705.00506.

Kim, J. K. and Yu, C. L. (2011). A semiparametric estimation of mean functionals with nonignorable missing data. *Journal of the American Statistical Association* **106**, 157–165.

Klein, N., Kneib, T. and Lang, S. (2015). Bayesian generalized additive models for location, scale, and shape for zero-inflated and overdispersed count data. *Journal of the American Statistical Association* **110**, 405–419.

Kott, P. S. and Chang, T. (2010). Using calibration weighting to adjust for nonignorable unit nonresponse. *Journal of the American Statistical Association* **105**, 1265–1275.

Liao, S. G., Lin, Y., Kang, D. D., Chandra, D., Bon, J., Kaminski, N., Sciurba, F. C. and Tseng, G. C. (2014). Missing value imputation in high-dimensional phenomic data: imputable or not, and how? *BMC Bioinformatics* **15**, 346.

Liao, Y. and Jiang, W. (2011). Posterior consistency of nonparametric conditional moment restricted models. *Annals of Statistics* **39**, 3003–3031.

Little, R. J. and Rubin, D. B. (2002). Bayes and multiple imputation. *Statistical analysis with missing data* pages 200–220.

Liu, J., Gelman, A., Hill, J., Su, Y. S. and Kropko, J. (2014). On the stationary distribution of iterative imputations. *Biometrika* **101**, 155–173.

Miao, W., Ding, P. and Geng, Z. (2016). Identifiability of normal and normal mixture models with nonignorable missing data. *Journal of the American Statistical Association* **111**, 1673–1683.

Mogstad, M. and Wiswall, M. (2012). Instrumental variables estimation with partially missing instruments. *Economics Letters* **114**, 186–189.

Nevo, A. (2003). Using weights to adjust for sample selection when auxiliary information is available. *Journal of Business and Economic Statistics* **21**, 43–52.

Newey, W. K. and Powell, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica* **71**, 1565–1578.

Palmer, T. M., Lawlor, D. A., Harbord, R. M., Sheehan, N. A., Tobias, J. H., Timpson, N. J., Smith, G. D. and Sterne, J. A. (2012). Using multiple genetic variants as instrumental variables for modifiable risk factors. *Statistical methods in medical research* **21**, 223–242.

Qin, J. (2000). Combining parametric and empirical likelihoods. *Biometrika* **87**, 484–490.

Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O. and Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using mice: a caliber study. *American Journal of Epidemiology* **179**, 764–774.

Shahbaba, B. and Neal, R. (2009). Nonlinear models using dirichlet process mixtures. *Journal of Machine Learning Research* **10**, 1829–1850.

Shen, W., Tokdar, S. T. and Ghosal, S. (2013). Adaptive Bayesian multivariate density estimation with dirichlet mixtures. *Biometrika* **100**, 623–640.

Smith, G. D. (2006). Randomised by (your) god: robust inference from an observational study design. *Epidemiology and Community Health* **60**, 382–388.

Stekhoven, D. J. and Buhlmann, P. (2012). Missforest–non-parametric missing value imutation for mixed-type data. *Bioinformatics* **28**, 112–118.

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research* **16**, 219–242.

Waljee, A. K., Mukherjee, A., Singal, A. G., Zhang, Y., Warren, J., Balis, U., Marrero, J., Zhu, J. and Higgins, P. D. (2013). Comparison of imputation methods for missing laboratory data in medicine. *BMJ open* **3**, e002847.

Xue, Y., Liao, X., Carin, L. and Krishnapuram, B. (2007). Multi-task learning for classification with dirichlet process priors. *Journal of Machine Learning Research* **8**, 35–63.

Appendix A
Detailed simulation design

A1. Setup for simulation study 1

In this study, we assume 1-dimensional endogenous regressor x_i, 4-dimensional exogenous regressor $\boldsymbol{w}_i$, whose first elements are set to be 1 (i.e. intercept), and 1-dimensional instruments $\boldsymbol{z}_i$, whose first elements are set to be 1 and are independent of $\epsilon_{2,i}$. The true relationship of the full model is as follows.

$$\begin{cases} y_i = x_i\beta + \boldsymbol{w}_i\boldsymbol{\gamma} + \epsilon_{1,i} \\ x_i = \boldsymbol{z}_i\boldsymbol{\delta} + \epsilon_{2,i} \end{cases}, \boldsymbol{\epsilon}_i \equiv \begin{pmatrix} \epsilon_{1,i} \\ \epsilon_{2,i} \end{pmatrix} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$$

We set $\boldsymbol{\delta} = (1,1,0,1,1)^t$, $\boldsymbol{\gamma} = (1,1,-1.1)^t$, $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}$, and the interested parameter $\beta = 1$. As for $\boldsymbol{z}_i$, we first generate $\boldsymbol{z}^*_{-1,i}$ from log-normal distribution with mean parameter 0 and variance parameter 0.5. Then, $\boldsymbol{w}_{-1,i} \sim MVN\left(\boldsymbol{z}^*_{-1,i}\begin{pmatrix} 0.2 \\ -0.2 \\ 0.2 \end{pmatrix}, \begin{pmatrix} 1 & 0.3 & 0.2 \\ 0.3 & 1 & 0.3 \\ 0.2 & 0.3 & 1 \end{pmatrix}\right)$, where $\boldsymbol{w}_{-1,i}$ denote the components of $\boldsymbol{w}_i$ with the first component 1 removed, and set $\boldsymbol{z}_i$ to be $\boldsymbol{z}_i = \left(1, \boldsymbol{z}^*_{-1,i}, \boldsymbol{w}_i\right)$.

After creating complete dataset, we set some components of IV to be missing. We firstly calculated

$$\lambda_i = U_i - \Phi^{-1}\left(\eta_1 y_i + \eta_2 x_i + \boldsymbol{\eta}_3\boldsymbol{w}_i + \eta_4 z_i + \eta_5 y_i z_i + \eta_6 x_i z_i + \boldsymbol{\eta}_7\boldsymbol{w}_i z_i\right)$$

where $U_i \sim uniform(0,1)$, and corresponding i of z_i which has the top 30% of λ_i are converted to be missing. η are set to be $(\eta_1, \eta_2, ..., \eta_7) = (0.25, 0.25, \mathbf{0}, 0.25, 0, -0.5, \mathbf{0})$. Since the missing probability of IV depends on IV, this is the NMAR case. We assume parametric probit model with cross terms.

The results are provided in the main article.

A2. Setup for simulation study 2

Simulation study 2 consider the finite dimensional mixture of regression model to the reduced-form equation. In this study, we assume 1-dimensional endogenous regressor x_i, 4-dimensional exogenous regressor $\boldsymbol{w}_i$, whose first elements are set to be 1 (i.e. intercept), and 1-dimensional instruments $\boldsymbol{z}_i$, whose first elements are set to be 1 and are independent of $\epsilon_{2,i}$. The true relationship of the model is as follows.

$$\begin{cases} y_i = x_i\beta + \boldsymbol{w}_i\boldsymbol{\gamma} + \epsilon_{1,i} \\ x_i = \sum_{l=1}^{2} \mathbf{1}_{i\in l}\boldsymbol{z}_i\boldsymbol{\delta}_l + \epsilon_{2,i} \end{cases}, \boldsymbol{\epsilon}_i \equiv \begin{pmatrix} \epsilon_{1,i} \\ \epsilon_{2,i} \end{pmatrix} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_l)$$

where $\boldsymbol{c}$ is taken so that the mean of $\boldsymbol{\epsilon}_i$ is zero. For weak instrument case, we set $(\pi_1, \pi_2) = (0.5, 0.5)$, $\boldsymbol{\delta}_1 = (-1, 2, 2, 0, 0)^t$, $\boldsymbol{\delta}_2 = (-1, -1, 0, 0, 0)^t$, $\boldsymbol{\gamma} = (1, 1, -1.1)^t$, $(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2) = \left(\begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}, \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \right)$, and the interested parameter $\beta = 1$. As for $\boldsymbol{z}_i$, we first generate $\boldsymbol{z}^*_{-1,i}$ from log-normal distribution with mean parameter 0 and variance parameter 0.5. Then, $\boldsymbol{w}_{-1,i} \sim MVN\left(\boldsymbol{z}^*_{-1,i} \begin{pmatrix} 0.2 \\ -0.2 \\ 0.2 \end{pmatrix}, \begin{pmatrix} 1 & 0.3 & 0.2 \\ 0.3 & 1 & 0.3 \\ 0.2 & 0.3 & 1 \end{pmatrix} \right)$, where $\boldsymbol{w}_{-1,i}$ denote the components of $\boldsymbol{w}_i$ with the first component 1 removed, and set $\boldsymbol{z}_i$ to be $\boldsymbol{z}_i = \left(1, \boldsymbol{z}^*_{-1,i}, \boldsymbol{w}_i\right)$.

The data missing process and the hyperparameters are the same as simulation 1.

A2.1 *Results of Simulation study 2*

Table A1 shows the results of simulation 2, including the empirical mean, standard deviation, the coverage of nominal 95% confidence intervals (CI) of the estimate, and the mean squared error (MSE) from the true value of β. Figure 5-1 shows the box plot of the estimates for β obtained by $1,000$ replications.

As can be seen from the table, this simulation study considers the case where MICE works relatively well. However, it shows biased results and the coverage is poor (75.3%). The MSE is about 5.0 times larger than the proposed. MF-GMM shows the largest MSE, and the coverage is very poor. Ps-Proposed, which does not have the information of the original distribution of IV, contains missing mechanism that cannot be identified and the obtained estimates are slightly upwardly biased. It should be noted that standard deviation obtained from Ps-Proposed is much lower than others, but the coverage is poor (72.6%). As a result, the MSE of Ps-Proposed is 4.1 times that of the proposed. The proposed obtained the smallest MSE with smaller s.d. and better coverage. Comparing the results of the proposed with Ps-Proposed and MICE-GMM, original population information combined with semiparametric IV model specification seems very useful.

The complete case analysis-based methods also shows poor results. CC-GMM estimates have upwardly biased results and the MSE is about 7.1 times

that of the proposed method. The coverage of 95% confidence intervals of CC-OLS is only 3.3%, since OLS-CC ignores the presence of endogeneity.

Table A1 Results of Simulation 2

	proposed				Ps-Proposed				MICE-GMM			
TRUE	Mean	(s.d.)	Cov	MSE	Mean	(s.d.)	Cov	MSE	Mean	(s.d.)	Cov	MSE
$\beta = 1$	1.017	(0.047)	91.8%	0.004	1.086	(0.067)	72.6%	0.018	1.094	(0.102)	75.3%	0.021
MSE ratio		1.000				4.101				4.998		

	MF-GMM				CC-GMM				CC-OLS			
TRUE	Mean	(s.d.)	Cov	MSE	Mean	(s.d.)	Cov	MSE	Mean	(s.d.)	Cov	MSE
$\beta = 1$	1.146	(0.114)	61.4%	0.042	1.111	(0.118)	79.7%	0.030	1.111	(0.028)	3.3%	0.0133
MSE ratio		9.738				7.083				3.108		

Empirical mean, standard deviation, coverage of nominal 95% CIs, and mean squared error of the estimates, and MSE Ratio of the proposed basis from 1000 simulations are described. Ps-Proposed, pseudo-proposed method which does not use population-level information; MICE-GMM, IV estimates by generalized method of moments based on multiple imputation by chained equation; MF-GMM, IV estimates by generalized method of moments based on missForest imputation; CC-GMM, IV estimates by generalized method of moments based on complete case; CC-OLS, ordinary least squares regression based on complete case