Discussion Paper Series

RIEB

Kobe University

DP2018-15

Semiparametric Bayes Multiple Imputation for Regression Models with Missing Mixed Continuous-Discrete Covariates

Ryo KATO
Takahiro HOSHINO

May 15, 2018

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

Semiparametric Bayes multiple imputation for regression models with missing mixed continuous-discrete covariates

Ryo Kato*

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

and

Takahiro Hoshino†

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. Issues regarding missing data are critical in observational and experimental research, as they induce loss of information and biased result. Recently, for datasets with mixed continuous and discrete variables in various study areas, multiple imputation by chained equation (MICE) has been more widely used, although MICE may yield severely biased estimates. We propose a new semiparametric Bayes multiple imputation approach that can deal with continuous and discrete variables. This enables us to overcome the shortcomings of multiple imputation by MICE; they must satisfy strong conditions (known as compatibility) to guarantee that obtained estimators are consistent. Our exhaustive simulation studies show that the coverage probability of 95 % interval calculated using MICE can be less than 1 %, while the MSE of the proposed one can be less than one-fiftieth. We also applied our method to the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, and the results are consistent with those of the previous research works that used panel data other than ADNI database, whereas the existing methods such as MICE, resulted in entirely inconsistent results.

KEY WORDS: Full conditional specification; Missing data; Multiple imputation; Probit stick-breaking process mixture; Semiparametric Bayes model.

* *email:* kato.ryo@keio.jp

† *email:* bayesian@jasmine.ocn.ne.jp

1. Introduction

Issues regarding missing data are critical in observational and experimental research, as they induce loss of information and biased result, and at times, lead to wrong decisions. The National Research Council (2010) published a report including recommendations on treating missing data in medical science research. According to the recommendation, researchers should employ as many confounders as possible in order to obtain valid estimates from analysis such as logistic regressions or Cox proportional hazards model. However, when they employ more covariates, the number of observations with at least one missing component increase. Also, if a researcher is interested in using a regression model containing missing components in covariates, a complete case analysis results in biased estimates even when the missing mechanism is missing at random (MAR) (Ibrahim et al., 2005).

In this case, if conditional distributions of incompletely observed covariates, given completely observed covariates, are correctly specified, we can obtain consistent estimators using the expectation-maximization (EM) algorithm or Bayesian estimation with the Markov chain Monte Carlo (MCMC) method. However, it is usually difficult to specify such a distribution because both incompletely and completely observed covariates generally have large dimensions, and the distributional form is not expressed by well-known distributional families owing to the mixed-scale variables.

For datasets with mixed continuous and discrete variables in various study areas, multiple imputation by chained equation (MICE), in which missing variables are iteratively imputed based on full conditional specification (FCS), has been cited numerous times by researchers from several fields including medical statistics (van Buuren, 2007,White et al., 2011,Paton et al., 2014). This is because the researchers, especially the imputers, are not required to construct an explicit joint multivariate model with mixed-scale variables (continuous, categorical, ordinal, and so on). More specifically, the MICE-FCS approach specifies a multivariate imputation model using a sequence of seemingly "appropriate" univariate regression models

corresponding to the types of missing variables; namely, one only needs to assign a univariate linear regression with a normally distributed error term for an incomplete continuous variable, a logistic regression for an incomplete binary variable, an ordered logistic regression for an incomplete ordinal variable, and so on. Moreover, researchers can easily implement MICE-FCS using several existing statistical software packages, such as the mice package in R and S-plus, proc mi with the FCS option in SAS, and mi impute in STATA.

In spite of the widespread use of MICE-FCS, recent studies showed that it leads to severely biased estimates in various setups. Liu et al. (2014) proved that using MICE-FCS does not guarantee that the asymptotic distribution is equivalent with the existing Bayesian MI estimator when the families of the conditional models are "incompatible" (see Section 4 in Liu et al. (2014)). In fact, simulation studies by Bartlett et al. (2015) showed that MICE yields biased estimates when treating incompatible conditional models. Unfortunately, violation of the compatibility assumption is not uncommon (the example of the violation of the compatibility assumption is provided in section 2.1). Therefore, although MICE-FCS is simple and convenient to use, it can provide statistically valid estimates in very limited cases.

1.1 *Motivating example*

We briefly introduce a motivating example of a real world dataset in which it is very hard to properly impute the missing components using FCS approach. The data used in this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org.

Jack et al. (2010) used data from the ANDI to study baseline predictors that contribute

to the progression of AD. Figure 1 shows the distributions of amyloid $\beta 1-42$ ($ABETA_{1-42}$), tau (total tau protein), and p-tau (phosphorylated tau protein; P-tau181p) of MCI at baseline subjects, and Jack et al. (2010) show that they are associated with time to conversion to AD. Since the participants were not forced to contribute to the CSF measurement, around 50% of the data for $ABETA_{1-42}$, tau, and p-tau are missing. Jack et al. (2010) employed the Cox proportional hazards model, in which the covariates include $ABETA_{1-42}$, tau, and p-tau. The analysis was restricted to 218 samples whose $ABETA_{1-42}$ were available, nevertheless the dataset contains approximately 400 subjects. If the researchers try to address the missing components using the MI method, EM algorithm, or Bayesian estimation with the MCMC technique, they must correctly specify the complex joint distribution of the covariates.

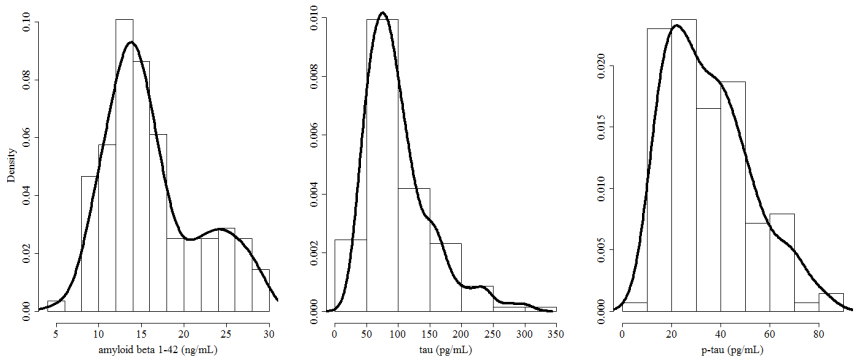

Fig.1. Histograms of observed $ABETA_{1-42}$ (Amyloid beta 1-42), tau (total tau protein), and p-tau (phosphorylated tau protein). The bold lines represent the kernel densities of the data.

Figure 1 shows that they are not normally distributed and seem to be skewed, following a fat-tailed distribution. Accordingly, specifying the covariate distribution seems to be very difficult in such cases. Bartlett et al. (2015) employed the SMC-FCS approach to impute the missing covariates and estimated the Cox regression. They added information pertaining to the family history of Alzheimer's patients, namely, whether the subject's father and mother had AD or not. However, their results seem to be biased owing to the violation of the model compatibility assumption required be the FCS approaches; several covariates do not follow

normal distributions, as seen from Figure 1. Also, more than two binary missing covariates are employed. As shown in Section 2, the FCS approaches results in biased estimates when missing covariates include two or more binary variables because of the model incompatibility. Therefore, it is evident that the existing FCS approaches should not be applied to these kinds of datasets, which are often found in the real world.

1.2 *New contribution*

In this paper, we propose a new flexible semiparametric Bayesian framework for MI, which is capable of treating mixed-scale incomplete variables. The model formulation is different from that seen in the existing literature in two ways.

First, we express the full model as the product of the covariate distribution (conditional distribution of incompletely observed covariates given completely observed covariates) and the substantive model (the regression model researchers are interested in). We assume the parametric model to the substantive model since the researchers conducting applied research are generally concerned with the parameters of the functions in the substantive model, which should be built upon the existing theories or previous literature in the field of study. Examples of the parametric substantive model are the Cox regression and the logistic regression in epidemiological and clinical research. On the other hand, with regard to the covariate distribution, we specify a joint distribution of the missing variables using the probit stick-breaking process mixture (PSBPM) model proposed by Chung and Dunson (2009), whose model specification is based on the Dirichlet process mixture (DPM) model. Ibrahim et al. (2005) also pointed out that one of the caveats of treating missing covariates lies in specifying the parametric model of the covariate distribution. However, it is nearly impossible to correctly prespecify the covariate distribution based on existing theories or some inferences, because the relationships of the missing variable and the complete variables are often "multivariate-to-multivariate", they can be non-linear relationships, or they may be non-normally distributed. Therefore, we employ the nonparametric Bayesian specification; specifically, we use PSBPM

modeling instead of DPM since the stick-breaking weights can vary depending on the predictors. Since our approach do not rely on FCS approach, we do not have to consider the compatibility assumption holding. Murray and Reiter (2016) proposed fully nonparametric multiple imputation method using DPM model with local dependence. However, they do not consider the existence of the substantive model, hence their main scope of the inference is the means or the variances of the imputed variables, and it cannot estimate the interested parameters of the substantive model.

Second, we express mixed-scale variables through the transformation of the latent continuous variables for probit modeling. This underlying continuous variables approach is used in the context of the DPM model, as in Kottas et al. (2005) for ordinal variables; in Canale and Dunson (2011) for count variables; and in Kim and Ratchford (2013) for ordinal variables. This approach enables us to deal in a straightforward manner with many types of variables in the joint covariate distribution without specifying the complicated conditional joint distribution of mixed-scale variables.

1.3 *Organization*

The rest of the paper is organized as follows. In the next section, we propose and formulate a semiparametric Bayesian multiple imputation (SB-MI) algorithm that can overcome the drawbacks of the existing methods. In Section 3, we describe the model specification, imputation procedure and posterior computation of the proposed model in detail. The simulation studies illustrating the performance of the proposed method compared with the MICE-FCS, SMC-FCS, and missForest approaches are presented in Section 4. In Section 5, we apply our proposed method to the real dataset described in the motivating example in Section 1.1. Section 6 concludes after providing a short discussion. The detailed descriptions of the simulation design appear in the Appendix A. We are preparing an R package to implement our algorithm. In advance, computer code in Matlab are provided, and may result in minor variations, all of which will be resolved and covered by the R package.

2. Model setup

In this paper, we consider a dataset consisting of N $(i = 1, ..., N)$ cases, where the interest of the researchers lies in a model with outcomes $\boldsymbol{y} : j \times 1$, completely observed covariates $\boldsymbol{v} : p \times 1$, and incompletely observed covariates $\boldsymbol{w} : q \times 1$. Let $\boldsymbol{a}^{obs}$ and $\boldsymbol{a}^{mis}$ be the observed and missing subsets of variable $\boldsymbol{a} \in (\boldsymbol{y}, \boldsymbol{v}, \boldsymbol{w})$ in the dataset, respectively. Let $\boldsymbol{r}$ be the vector of observation indicators whose element equals 1 if the corresponding element of the dataset is observed and 0 otherwise. Throughout this paper, we consider that the data are MAR, that is, $p(\boldsymbol{r}|\boldsymbol{a}^{obs}, \boldsymbol{a}^{mis}) = p(\boldsymbol{r}|\boldsymbol{a}^{obs})$. Additionally, we assume that all the observations are independent and identically distributed.

Let $\boldsymbol{\vartheta}_s$ be the parameter vectors of the substantive model $p(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{v}, \boldsymbol{\vartheta}_s)$. Note that the researcher's prime target of inference lies in $\boldsymbol{\vartheta}_s$, even in the context of missing data analysis. We propose a SB-MI algorithm expressed by the following imputation model as the product of two submodels:

$$p(\boldsymbol{y}, \boldsymbol{w}|\boldsymbol{v}) \propto p(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{v}, \boldsymbol{\vartheta}_s)p(\boldsymbol{w}|\boldsymbol{v}, \boldsymbol{\vartheta}_m) \tag{1}$$

where $p(\boldsymbol{w}|\boldsymbol{v}, \boldsymbol{\vartheta}_m)$ represents the covariate distribution with parameters of lower interest $\boldsymbol{\vartheta}_m$, and $p(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{v}, \boldsymbol{\vartheta}_s)$ represents the substantive model with parameters of higher interest $\boldsymbol{\vartheta}_s$. We assume the parametric model for the substantive model $p(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{v}, \boldsymbol{\vartheta}_s)$ since the researchers' concern generally lies in the parameters of the substantive model $\boldsymbol{\vartheta}_s$. Additionally, the substantive model should be built upon the existing knowledge corresponding to the purpose of the study. Accordingly, the researcher may assume a linear regression with an interaction term, a Cox regression, or a logistic regression for $p(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{v}, \boldsymbol{\vartheta}_s)$. While we assume a parametric structure for the substantive model, we do consider Bayesian nonparametric form rather than a parametric form for the covariate distribution $p(\boldsymbol{w}|\boldsymbol{v}, \boldsymbol{\vartheta}_m)$, because researchers generally have no interest in $\boldsymbol{\vartheta}_m$ and parametric modeling of a large number of covariates can result in misspecification bias (Chib, 2007). Moreover, we express mixed-scale variables

through the transformation of latent continuous variables for probit modeling in order to deal with many types of continuous and discrete variables in the joint covariate distribution in a straightforward manner. This transformation enables us to avoid considering the compatibility assumption, which is required in MICE-FCS or SMC-FCS.

2.1 *Existing method*

In this situation, $\boldsymbol{\vartheta}_s$ can be estimated by the existing MI methodology, EM algorithm, or Bayesian MCMC estimation. For example, MI uses the MCMC approach, and researchers iteratively draw the parameter of the joint model $\boldsymbol{\psi}$ from $p(\boldsymbol{\psi}|\boldsymbol{a})$, and then draw $\boldsymbol{a}^{mis}$ from $p(\boldsymbol{a}^{mis}|\boldsymbol{a}^{obs}, \boldsymbol{\psi})$ D times. However, as is the case with the maximum likelihood estimation of the EM algorithm or the Bayesian MCMC estimation, it is difficult to correctly specify the joint distribution of all the variables that have missing elements $p(\boldsymbol{a}^{mis}|\boldsymbol{a}^{obs}, \boldsymbol{\psi})$, especially when both continuous and discrete variables are missing.

The MICE-FCS method has become a more widely used methodology as researchers seek to avoid the difficulty in specifying the conditional joint distribution $p(\boldsymbol{a}^{mis}|\boldsymbol{a}^{obs}, \boldsymbol{\psi})$. The MICE-FCS approach specifies a multivariate covariate distribution by a sequence of univariate regressions for each missing variable. More specifically, MICE-FCS iterates drawing $\boldsymbol{\psi}_j$ from $p(\boldsymbol{\psi}_j|\boldsymbol{a})$ and $\boldsymbol{a}^{mis}$ from $p(\boldsymbol{a}_j^{mis}|\boldsymbol{a}_{-j}^{mis}, \boldsymbol{a}^{obs}, \boldsymbol{\psi}_j)$ for each $\boldsymbol{a}_j^{mis}$, where $\boldsymbol{a}_{-j}$, denoting the components of $\boldsymbol{a}$ with $\boldsymbol{a}_j$ removed. Because of the simplicity of its covariate distribution specification, MICE is popularly used to deal with missing data.

In spite of the widespread use of MICE-FCS due to its convenience, it was recently proved that the asymptotic distribution drawn using MICE-FCS is not equivalent to the existing Bayesian simulation in several settings. Liu et al. (2014) showed that the MICE-FCS algorithm does not guarantee that the asymptotic distributions are consistent with the existing Bayesian joint model MI estimator when the family of conditional models and their joint distributions are incompatible. According to Liu et al. (2014), the compatibility assumption is satisfied when a parameter set of conditional models $f_j(a_j^{mis}|\boldsymbol{a}_{-j}^{mis}, \boldsymbol{a}^{obs}, \boldsymbol{\psi}_j)$ is

represented by surjective mapping of a collection of the joint model $p(\boldsymbol{a}^{mis}, \boldsymbol{a}^{obs}|\boldsymbol{\psi})$ parameter $\boldsymbol{\psi}$, that is, $g_j(\boldsymbol{\psi}) = \boldsymbol{\psi}_j$ and hence $p(a_j^{mis}|\boldsymbol{a}_{-j}^{mis}, \boldsymbol{a}^{obs}, \boldsymbol{\psi}) = f_j(a_j^{mis}|\boldsymbol{a}_{-j}^{mis}, \boldsymbol{a}^{obs}, \boldsymbol{\psi}_j)$; otherwise they are said to be incompatible. Put simply, compatibility purports that the parameters of the conditional distribution can be expressed by the parameters of the joint distribution of the model. Liu et al. (2014) also showed that the MICE-FCS algorithm generates a consistent estimator using Rubin's rules, but the variance of the parameters cannot be applied to Rubin's rules if the family of the conditional models is semicompatible as a special case of incompatibility. On the other hand, if the model is compatible, MICE-FCS is asymptotically equivalent to the existing Bayesian simulation; hence, one can apply Rubin's rule to calculate the mean and variance of the parameters of interest.

In what kinds of cases this compatibility assumption holds? If the all variables in the datasets are consist only of continuous variables that follow an i.i.d. multivariate normal distribution and the conditional models are linear regressions with normally distributed error terms, the conditionals and joint model are compatible, and the estimators of MICE-FCS are applicable to Rubin's rules. If one variable is binary variable and the rest are continuous, one can also apply Rubin's rules to the substantive model described in the form of a linear regression with normally distributed error terms. However, when the researcher is interested in binary outcome modeling with a logistic regression, wherein there exist binary covariates in the datasets, and even if all the other covariates are continuous, the conditionals and substantive model (logistic model specification) are incompatible, and the MICE-FCS estimators are not equivalent to those corresponding to Gibbs sampling. In epidemiological and clinical research, researchers often assume non-linear models such as the Cox proportional hazards model, regression models with quadratic terms, or regression models with interaction terms. Yet, unfortunately, these are examples of model incompatibility. In addition, the conditionals of MICE-FCS are under the immediate control of the researcher, and hence, the joint distribution is only implicitly known and may not exist (van Buuren, 2012). Therefore, although

MICE-FCS is simple and convenient, the estimators are valid in a very limited number of cases only.

Bartlett et al. (2015) recently developed SMC-FCS in order to relax the compatibility assumption, which assigns the imputer compatibility of the joint distribution of covariates only and not all the variables. However, it has hardly solved the problem of the compatibility assumption holding because the number of covariates is generally larger than the outcome variable. For example, if the missing covariates contain two or more binary variables, a parameter set of conditional models cannot be represented by onto mapping of a collection of joint model parameters. Hence, the model compatibility assumption among covariates is violated, and the estimates from the SMC-FCS do not guarantee the consistency. Additionally, if some covariates do not follow a normal or Bernoulli distribution (say, they follow a log-normal distribution or a mixture of normal distributions), the compatibility assumption of joint distribution among covariates required in the SMC-FCS method cannot be generally satisfied, as seen in the motivating example in Section 1.1.

3. Semiparametric Bayes multiple imputation

3.1 *Semiparametric model formulation*

As stated in Equation (1), we propose a SB-MI algorithm, expressed by imputaion model $p(\boldsymbol{y}, \boldsymbol{w}|\boldsymbol{v})$ as the product of two submodels $p(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{v}, \boldsymbol{\vartheta}_s)p(\boldsymbol{w}|\boldsymbol{v}, \boldsymbol{\vartheta}_m)$. Assuming the independence of the priors $p(\boldsymbol{\vartheta}_m)$ and $p(\boldsymbol{\vartheta}_s)$, the posteriors are

$$p(\boldsymbol{\vartheta}_m, \boldsymbol{\vartheta}_s|y, \boldsymbol{v}, \boldsymbol{w}) \propto p(\boldsymbol{\vartheta}_m)p(\boldsymbol{\vartheta}_s)p(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{v}, \boldsymbol{\vartheta}_s)p(\boldsymbol{w}|\boldsymbol{v}, \boldsymbol{\vartheta}_m)$$

$\boldsymbol{\vartheta}_m$ and $\boldsymbol{\vartheta}_s$ can be drawn from $p(\boldsymbol{\vartheta}_m)p(\boldsymbol{w}|\boldsymbol{v}, \boldsymbol{\vartheta}_m)$ and $p(\boldsymbol{\vartheta}_s)p(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{v}, \boldsymbol{\vartheta}_s)$, respectively. Given these parameters, the missing values are drawn from the density proportional to $p(\boldsymbol{y}, \boldsymbol{w}|\boldsymbol{v})$.

The specification of the substantive model $p(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{v}, \boldsymbol{\vartheta}_s)$ varies by the purpose of the analysis and the properties of the outcome $\boldsymbol{y}$. One may specify the linear regression to the

continuous outcome or the logistic regression to the discrete outcome. Besides, one must employ specified model forms on $p(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{v}, \boldsymbol{\vartheta}_s)$, such as the Cox proportional hazards model or quadratic models. Our proposed methodology, in any case, can properly estimate $\boldsymbol{\vartheta}_s$ even when MICE-FCS or SMC-FCS cannot because of model incompatibility.

3.2 *Covariate distribution*

On the other hand, we have to specify the complicated covariate distribution $p(\boldsymbol{w}|\boldsymbol{v}, \boldsymbol{\vartheta}_m)$. Usually, $\boldsymbol{w}$ will include continuous and discrete variables. In order to deal with mixed-scale covariates, we employ a transformation of the latent continuous variables for probit modeling. Thus, we rewrite $\boldsymbol{w}$ as $\boldsymbol{w} = (\boldsymbol{w}_c, \boldsymbol{w}_d, \boldsymbol{w}_n)$ where $\boldsymbol{w}_c$ denotes the continuous variable component, $\boldsymbol{w}_d$ denotes the ordered variable component with $Q_d(= 1, ..., q_d, ..., Q_d)$ orders, and $\boldsymbol{w}_n$ denotes the nominal variable component with $Q_n(= 0, ..., q_n, ..., Q_n)$ choices. To deal with discrete variables simply, we introduce continuous latent variables $\boldsymbol{u}_d$ and $\boldsymbol{u}_n$ where

$$w_d = \ q_d \quad \text{if } \zeta_{q_d-1} < u_d \leq \zeta_{q_d}$$

$$w_n = \begin{cases} 0 & \text{if } \max(u_n) < 0 \\ q_n & \text{if } \max(u_n) = u_{nq_n} > 0 \end{cases}.$$

We assume the following structure on the covariate distribution.

$$\boldsymbol{w}_i^* = f(\boldsymbol{v}_i) + \boldsymbol{\eta}_i$$

with $\boldsymbol{w}_i = g(\boldsymbol{w}_i^*)$, where $\boldsymbol{w}^* = (\boldsymbol{w}_c, \boldsymbol{u}_d, \boldsymbol{u}_n)$ and $g(\cdot)$ represents the function converting the latent continuous variables $\boldsymbol{w}_i^*$ to $\boldsymbol{w}_i$. This enables us to deal with many types of continuous and discrete variables in the covariate distribution in a straightforward manner.

More concretely, we employ DPM modeling to represent the covariate distribution. DPM modeling is frequently utilized in applied statistical modeling when researchers intend to

avoid making assumptions about parameter distribution within the Bayesian framework. For example, Hirano (2002) developed autoregressive models with individual effects where the disturbances are not restricted to a parametric class. Rodriuez et al. (2009) used DPM to develop a Bayesian semiparametric approach for functional data analysis. Kunihama et al. (2016) developed a nonparametric Bayes model with DPM to incorporate sample survey weights. The theoretical properties of DPM were investigated by Shen et al. (2013).

According to Sethuraman (1994), the Dirichlet process as a prior for a random distribution G can be represented by the stick-breaking process. Let $\xi_1, \xi_2, ...$ be an independent draw from a beta distribution $Be(1, \gamma)$. If G follows the Dirichlet process prior with concentration parameter γ and base distribution G_0, that is, $G \sim DP(\gamma, G_0)$, G can be represented as

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

where $\kappa_l = \xi_l \prod_{h<l} (1 - \xi_h)$ and δ_θ is a point mass at θ (refer to Walker et al. (1999) for a detailed description of DPM).

Although DPM is used to flexibly express a variety of parameters or distributions, they are greatly restricted because probability weight κ_l is a constant (Dunson et al., 2007). If the stick-breaking weights π_l are constant and independent of predictor $\boldsymbol{x}_i$, as in DPM and other nonparametric Bayesian models, the mean regression structure is reduced to a linear one, namely, $\sum_{l=1}^{\infty} \pi_l \boldsymbol{\beta}_l^T \boldsymbol{x}_i = \overline{\boldsymbol{\beta}}^T \boldsymbol{x}_i$, where $\overline{\boldsymbol{\beta}} = \left(\sum_{l=1}^{\infty} \pi_l \boldsymbol{\beta}_l^T\right)$. Therefore, in this paper, we apply the PSBPM model proposed by Chung and Dunson (2009) since the algorithm allows for greater flexibility through predictor-dependent stick-breaking weights $\pi_l(\boldsymbol{x}_i)$. In addition, PSBPM results in a conjugate structure, and hence simpler posterior calculation. The statistical properties of PSBPM are described in Pati et al. (2013). For example, Hoshino (2013) proposed a semiparametric Bayesian model for causal inference where the parameters of no interest for researchers are modeled using PSBPM.

We apply PSBPM modeling to the covariate distribution specification. The resulting regression function of $\boldsymbol{w}^*$ on $\boldsymbol{v}$ can be represented as

$$f(\boldsymbol{w}_i^*|\boldsymbol{v}_i) = \sum_{l=1}^{\infty} \pi_l(\boldsymbol{v}) N(\boldsymbol{\Gamma}_l \boldsymbol{v}_i, \boldsymbol{\Phi}_l)$$

with the probability weights

$$\pi_l(\boldsymbol{v}) = \Phi(\boldsymbol{\xi}_l(\boldsymbol{v})) \prod\nolimits_{h<l} \{1 - \Phi(\boldsymbol{\xi}_h(\boldsymbol{v}))\}.$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. To make the probability weights $\pi_l(\boldsymbol{v})$ vary with covariates $\boldsymbol{v}$, we let $\boldsymbol{\xi}_l(\boldsymbol{v}) = \alpha_l + f_l(\boldsymbol{v})$, $\alpha_l \sim N(\mu_\alpha, 1)$, $\mu_\alpha \sim N(\mu_{\alpha_0}, \sigma^2_{\alpha_0})$ and we introduce the following regression function as in Chung and Dunson (2009):

$$f_l(\boldsymbol{v}) = -\sum_{k=1}^{q} \psi_{lk} \left| v_k - \Omega_{lk} \right|$$

with $\psi_{lk} \sim N(\mu_{\psi_k}, \sigma^2_{\psi_k}) \mathbf{1}_{R^+}, \Omega_{lk} \sim \sum_{m=1}^{M_k} \frac{1}{M_k} \delta_{\Omega^*_{km}}(\Omega_{lk})$, where $N(\mu_{\boldsymbol{U}}, \sigma^2_{\boldsymbol{U}}) \mathbf{1}_{\boldsymbol{U}}$ denotes a normal distribution with mean $\mu_{\boldsymbol{U}}$ and variance $\sigma^2_{\boldsymbol{U}}$ truncated to the set $\boldsymbol{U}$ and Ω^*_{km} are discrete points over a reasonable range of the k-th covariate v_k.

3.3 *Imputation and analysis procedure*

Data imputation and analysis procedures are as follows.

1. Assign each case i to any class l of the Dirichlet mixture.

2. Generate $\boldsymbol{\vartheta}_m^{(i)}$ from the posterior distribution based on the likelihood of $p(\boldsymbol{w}|\boldsymbol{v}, \boldsymbol{\vartheta}_m)$ calculated on the complete case dataset or pseudo-complete dataset.

3. If $\boldsymbol{w}$ is missing, generate the missing component of $\boldsymbol{w}$ proportional to $p(\boldsymbol{y}, \boldsymbol{w}|\boldsymbol{v})$.

4. If $\boldsymbol{y}$ is missing, generate the missing component of $\boldsymbol{y}$ from $p(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{v}, \boldsymbol{\vartheta}_s^{(i-1)})$. If case i

has missingness both on $\boldsymbol{w}_i$ and $\boldsymbol{y}_i$, generate $\boldsymbol{y}_i$ from $p(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{v}, \boldsymbol{\vartheta}_s^{(i-1)})$, where missing $\boldsymbol{w}_i$ is imputed in step 3.

5. Given the pseudo-complete dataset, generate $\boldsymbol{\vartheta}_s^{(i)}$ from the posterior distribution based on the likelihood of $p(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{v}, \boldsymbol{\vartheta}_s)$.

3.4 *Posterior computation of the proposed model*

Let $\boldsymbol{K}_i$ be the indicator denoting where case i belongs, and $\boldsymbol{K}_i = l$ if case i belongs to class l. Recall that $\boldsymbol{\vartheta}_m$ and $\boldsymbol{\vartheta}_s$ are the parameter vectors for the covariate distribution $p(\boldsymbol{w}|\boldsymbol{v})$ and the substantive model $p(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{v})$, respectively. This yields the following hierarchical representation of the finite-dimensional PSBPM model:

$$
\begin{aligned}
\boldsymbol{y}_i|\boldsymbol{w}_i, \boldsymbol{v}_i, \boldsymbol{\vartheta}_s &\sim p\left(\boldsymbol{y}_i|\boldsymbol{w}_i, \boldsymbol{v}_i, \boldsymbol{\vartheta}_s\right), \\
\boldsymbol{w}_i|\boldsymbol{v}_i, \boldsymbol{\vartheta}_m, \boldsymbol{K}_i &\sim p\left(\boldsymbol{w}_i|\boldsymbol{v}_i, \boldsymbol{\vartheta}_m, \boldsymbol{K}_i\right), \\
\boldsymbol{K}|\boldsymbol{\phi} &\sim \sum_{l=1}^{\infty} \pi_l(\boldsymbol{v}_i|\boldsymbol{\phi}_l)\delta_l\left(\cdot\right) \ \ (i = 1, ..., N), \\
\boldsymbol{\phi}^* &\sim p\left(\boldsymbol{\phi}^*|\boldsymbol{\tau}_{\boldsymbol{\phi}}\right), \\
\boldsymbol{\vartheta}_s \sim p\left(\boldsymbol{\vartheta}_s|\boldsymbol{\tau}_{\vartheta_s}\right), \boldsymbol{\vartheta}_m &\sim p\left(\boldsymbol{\vartheta}_m|\boldsymbol{\tau}_{\vartheta_m}\right), \\
\boldsymbol{\tau} &\sim p(\boldsymbol{\tau}),
\end{aligned}
$$

where $\boldsymbol{\phi}^* = \left(\alpha_l, \phi_{11}, ..., \phi_{Lq}, \Omega_{11}, ..., \Omega_{Lq}\right)$ and $\boldsymbol{\tau} = (\boldsymbol{\tau}_{\boldsymbol{\phi}}^T, \boldsymbol{\tau}_{\vartheta_s}^T, \boldsymbol{\tau}_{\vartheta_m}^T)^T$.

The blocked Gibbs sampler (Ishwaran and James, 2011) is applied to the posterior computation of the PSBPM parameters $\boldsymbol{\vartheta}_m$. The blocked Gibbs sampler is very similar to the Gibbs sampler except for the assignment of samples to each class. Since we employ PSBPM modeling, we can directly apply the estimation algorithm of Chung and Dunson (2009) for the simulation of $\boldsymbol{\vartheta}_m$. Each case i is assigned to one of L class in the blocked Gibbs sampling, where L denotes the maximum number or truncation of classes. As stated above, $\boldsymbol{w}$ may include continuous and discrete variables. We employ the transformation of the latent continuous variables $\boldsymbol{w}^*$ for probit modeling through a function g such that $\boldsymbol{w} = g(\boldsymbol{w}^*)$. Given

the draw of $\boldsymbol{\vartheta}_m$, the missing components of $\boldsymbol{y}$ and $\boldsymbol{w}$ are imputed, and then, the substantive model parameter $\boldsymbol{\vartheta}_s$ is simulated. We obtain the detailed posterior computation using the MCMC estimation as follows.

1. Update K_i $(i = 1, ..., N)$

   To assign samples to each class, generate K_i by $\sum_{l=1}^{L} \pi_{li} \delta_l(\cdot)$, where π_{li} is

$$\pi_{li} = \frac{\pi_l(\boldsymbol{v}_i) N(\boldsymbol{\Gamma}_l \boldsymbol{v}_i, \boldsymbol{\Phi}_l)}{\sum_{l=1}^{L} \pi_l(\boldsymbol{v}_i) N(\boldsymbol{\Gamma}_l \boldsymbol{v}_i, \boldsymbol{\Phi}_l)}$$

   with $\pi_l(\boldsymbol{v}_i) = \Phi(\boldsymbol{\xi}_l(\boldsymbol{v}_i)) \prod_{h<l} \{1 - \Phi(\boldsymbol{\xi}_h(\boldsymbol{v}_i))\}$.

2. Update Z_{il}^*

   We introduce latent variable Z_{il}^* where $Z_{il} = \mathbf{1}(Z_{il}^* > 0)$ and

$$Z_{il}^* \sim \begin{cases} N\left(\alpha_l - \sum_{k=1}^{q} \psi_{lk} \left|v_{ik} - \Omega_{lk}\right|, 1\right) \mathbf{1}_{\mathbb{R}^+} & \text{for } l = K_i \\ N\left(\alpha_l - \sum_{k=1}^{q} \psi_{lk} \left|v_{ik} - \Omega_{lk}\right|, 1\right) \mathbf{1}_{\mathbb{R}^-} & \text{for } l < K_i. \end{cases}$$

3. Update $\alpha_l (l = 1, ..., L-1)$

   Draw α_l from the following normal distribution.

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

   where $N_l = \sum_{i=1}^{N} 1(K_i \geq l)$ and $W_{il}^* = Z_{il}^* + \sum_{k=1}^{q} \psi_{lk} \left|v_{ik} - \Omega_{lk}\right|$.

4. Update $\psi_{lk} (l = 1, ..., L-1,\ k = 1, ..., q)$

Draw ψ_{lk} from the following left-truncated normal distribution.

$$\psi_{lk} \sim N\left(\frac{\sigma^2_{\psi_k}\mu_{\psi_k} + \sum_{i:K_i \geq l} |v_{ik} - \Omega_{lk}| \, U^*_{il}}{\sigma^2_{\psi_k} + \sum_{i:K_i \geq l} |v_{ik} - \Omega_{lk}|^2}, \frac{1}{\sigma^2_{\psi_k} + \sum_{i:K_i \geq l} |v_{ik} - \Omega_{lk}|^2}\right) \mathbf{1}_{\mathbb{R}^+},$$

where $U^*_{il} = \alpha_l - Z^*_{il} - \sum_{s=1, s\neq k}^{q} \psi_{ls} \, |v_{is} - \Omega_{ls}|$.

5. Update $\Omega_{lk}(l = 1, ..., L-1, \; k = 1, ..., q)$

   Draw Ω_{lk} from the following probability.

$$\Pr\left(\Omega_{lk} = \Omega^*_{km}\right) = \frac{\frac{1}{M_k} \prod_{i:K_i \geq l} N\left(Z^*_{il}; \alpha_l - \sum_{s=1, s\neq k}^{q} \psi_{ls} \, |v_{is} - \Omega_{ls}| - \psi_{lk} \, |v_{ik} - \Omega^*_{km}|, 1\right)}{\sum_{m=1}^{M_k} \frac{1}{M_k} \prod_{i:K_i \geq l} N\left(Z^*_{il}; \alpha_l - \sum_{s=1, s\neq k}^{q} \psi_{ls} \, |v_{is} - \Omega_{ls}| - \psi_{lk} \, |v_{ik} - \Omega^*_{km}|, 1\right)}.$$

6. Update $\boldsymbol{\vartheta}_m (= \boldsymbol{\Gamma}_l, \boldsymbol{\Phi}_l)$

   Draw $\boldsymbol{\vartheta}_m (= \boldsymbol{\Gamma}_l, \boldsymbol{\Phi}_l)$ from the following multivariate normal and inverted Wishart distribution.

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

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

7. Update μ_α

Draw μ_α from the following normal distribution

$$\mu_\alpha \sim N\left((L-1+\sigma_{\alpha_0}^{-2})^{-1}\left[\sum_{l=1}^{L-1}\alpha_l + \boldsymbol{\sigma}_{\alpha_0}^{-2}\mu_{\alpha_0}\right], (L-1+\sigma_{\alpha_0}^{-2})^{-1}\right).$$

8. Update the missing components

   Draw the missing component of $\boldsymbol{w}$ from a density proportional to $p(\boldsymbol{y}, \boldsymbol{w}|\boldsymbol{v})$. Since it is difficult to draw the missing $\boldsymbol{w}$, we employ the Metropolis-Hastings algorithm and use $p(\boldsymbol{w}|\boldsymbol{v}, \boldsymbol{\vartheta}_m)$ as a proposal density in order to draw a candidate of $\boldsymbol{w}_i$, $\boldsymbol{w}_i^c$. Note that the candidates are obtained after the transformation $\boldsymbol{w} = g(\boldsymbol{w}^*)$. We accept the candidates with the following probability:

$$\min\left(\frac{p(\boldsymbol{y}|\boldsymbol{w}^c, \boldsymbol{v}, \boldsymbol{\vartheta}_s)}{p(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{v}, \boldsymbol{\vartheta}_s)}, 1\right). \tag{2}$$

   If any component of $\boldsymbol{y}$ is missing, we draw $\boldsymbol{y}^{mis}$ from

$$p(\boldsymbol{y}^{mis}|\boldsymbol{w}, \boldsymbol{v}, \boldsymbol{\vartheta}_s)$$

9. Update $\boldsymbol{\vartheta}_s$

   Draw $\boldsymbol{\vartheta}_s$ from the density proportional to

$$p(\boldsymbol{\vartheta}_s)p(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{v}, \boldsymbol{\vartheta}_s)$$

For example, if we are interested in inferring the binary logistic regression coefficients, the acceptance probability in Equation (2) can be written as

$$\min\left(\frac{\left\{\exp(\boldsymbol{\Gamma}^T\boldsymbol{x}_i^c)/\left[1+\exp(\boldsymbol{\Gamma}^T\boldsymbol{x}_i^c)\right]\right\}^{y_i}\left\{1/\left[1+\exp(\boldsymbol{\Gamma}^T\boldsymbol{x}_i^c)\right]\right\}^{1-y_i}}{\left\{\exp(\boldsymbol{\Gamma}^T\boldsymbol{x}_i)/\left[1+\exp(\boldsymbol{\Gamma}^T\boldsymbol{x}_i)\right]\right\}^{y_i}\left\{1/\left[1+\exp(\boldsymbol{\Gamma}^T\boldsymbol{x}_i)\right]\right\}^{1-y_i}}, 1\right)$$

where $\boldsymbol{\Gamma}$ is a vector of coefficients, $\boldsymbol{x}_i = \left(\boldsymbol{w}_i^T, \boldsymbol{v}_i^T\right)^T$, and $\boldsymbol{x}_i^c$ is the vector $\boldsymbol{x}_i$ whose missing components are replaced by the candidate value.

4. Simulation study

We conduct the following four simulation studies in order to illustrate the performance of the proposed method when MICE-FCS cannot draw from a Bayesian joint model: (i) linear regression with a quadratic term, (ii) linear regression with an interaction term, (iii) proportional hazards model with a binary covariate, and (iv) logistic regression with a binary covariate. Through the simulation study, we consider the case $N = 400$, and 30% of the incomplete covariates are set to be missing. We generate the missing values based on $p(\boldsymbol{r}|\boldsymbol{y}, \boldsymbol{w}, \boldsymbol{v}) = p(\boldsymbol{r}|\boldsymbol{y})$. We compare the following imputation methods with SB-MI: MICE-FCS, SMC-FCS, and missForest.

Stekhoven and Buhlmann (2012) proposed the missForest algorithm, which imputes missing values from the random forest predictors and they are reported to provide lower imputation errors than the FCS method (Liao et al., 2014,Waljee et al., 2013). The detailed simulation design, results, and detailed discussion appear in Appendix A.

This section summarizes the results of simulation study (iv). We specify the substantive model of the logistic function as follows:

$$logit(y = 1) = \Gamma_0 + \Gamma_1 w_1 + \Gamma_2 w_2 + \Gamma_3 v$$

with $\Gamma_0 = 1$, $\Gamma_1 = 2$, $\Gamma_2 = -2$, and $\Gamma_3 = 3$. In this simulation, we consider three scenarios where the missing covariates follow (a) a multivariate normal distribution, (b) a multivariate log-normal distribution, and (c) a multivariate normal mixture distribution. We consider a case where one of the incompletely observed covariates w_1 is binary, where $w_{i,1} = 1$ if the latent variable, which is simulated based on above three process, $w_{i,1}^* > 0$ and $w_{i,1} = 0$ if the latent variable $w_{i,1}^* \leq 0$. The detailed data generating process is described in Appendix A.

Table 1. Simulation results of case (iv) - Logistic regression with binary covariates

	Complete Case				MICE-FCS				SMC-FCS				missForest				SB-MI			
TRUE	Mean	(s.d.)	Cov	MSE	Mean	(s.d.)	Cov	MSE	Mean	(s.d.)	Cov	MSE	Mean	(s.d.)	Cov	MSE	Mean	(s.d.)	Cov	MSE
(a) Normal																				
$\Gamma_0 = 1$	1.023	(0.286)	0.950	0.086	0.767	(0.257)	0.830	0.313	0.721	(0.260)	0.793	0.243	0.600	(0.240)	0.582	0.235	1.027	(0.223)	0.966	0.053
$\Gamma_1 = 2$	2.058	(0.517)	0.955	0.283	1.929	(0.485)	0.959	0.286	2.180	(0.502)	0.943	0.300	2.529	(0.449)	0.694	0.664	1.965	(0.414)	0.945	0.066
$\Gamma_2 = -2$	-2.081	(0.391)	0.946	0.165	-1.944	(0.355)	0.960	0.212	-2.161	(0.368)	0.944	0.169	-2.452	(0.349)	0.681	0.413	-1.989	(0.294)	0.941	0.056
$\Gamma_3 = 3$	3.127	(0.500)	0.952	0.265	3.118	(0.412)	0.979	0.254	3.388	(0.443)	0.911	0.365	3.542	(0.410)	0.718	0.530	2.985	(0.353)	0.935	0.078
MSE ratio			0.750				1.000				1.011				1.728				0.237	
(b) Log-normal																				
$\Gamma_0 = 1$	1.026	(0.312)	0.950	0.098	0.554	(0.302)	0.673	0.492	0.873	(0.309)	0.917	0.259	1.070	(0.299)	0.930	0.105	1.024	(0.259)	0.964	0.062
$\Gamma_1 = 2$	2.055	(0.401)	0.954	0.167	1.748	(0.385)	0.885	0.350	2.195	(0.397)	0.960	0.293	2.224	(0.376)	0.890	0.218	1.980	(0.341)	0.961	0.065
$\Gamma_2 = -2$	-2.056	(0.297)	0.952	0.091	-1.706	(0.276)	0.774	0.275	-2.190	(0.287)	0.928	0.297	-2.268	(0.276)	0.844	0.156	-2.006	(0.228)	0.962	0.066
$\Gamma_3 = 3$	3.090	(0.386)	0.947	0.157	2.794	(0.353)	0.896	0.264	3.356	(0.370)	0.913	0.414	3.194	(0.345)	0.909	0.167	3.019	(0.302)	0.940	0.081
MSE ratio			0.371				1.000				0.914				0.467				0.198	
(c) Mixture of normals																				
$\Gamma_0 = 1$	1.035	(0.487)	0.951	0.254	0.805	(0.324)	0.884	0.281	1.021	(0.471)	0.942	0.251	0.681	(0.442)	0.863	0.326	1.023	(0.235)	0.943	0.046
$\Gamma_1 = 2$	2.143	(0.886)	0.935	0.915	2.143	(0.582)	0.962	0.395	1.983	(0.851)	0.943	0.794	2.675	(0.826)	0.857	1.319	1.978	(0.439)	0.955	0.082
$\Gamma_2 = -2$	-2.105	(0.404)	0.947	0.192	-2.125	(0.396)	0.969	0.266	-2.137	(0.380)	0.912	0.181	-2.384	(0.357)	0.793	0.321	-1.989	(0.299)	0.956	0.069
$\Gamma_3 = 3$	3.151	(0.515)	0.950	0.312	3.379	(0.451)	0.934	0.329	3.388	(0.465)	0.890	0.398	3.393	(0.431)	0.852	0.393	2.997	(0.354)	0.947	0.067
MSE ratio			1.318				1.000				1.279				1.859				0.208	

Empirical mean, standard deviation, coverage of nominal 95% CIs, and mean squared error of the estimates, andsum of MSE Ratio of MICE-FCS basis from 1000 simulations are described. CC, complete case analysis; MICE-FCS, multiple imputation by chained equation - fully conditionalspecication; SMC-FCS, substantive model compatible - fully conditional specication; SB-MI, semiparametric Bayesian multiple imputation (Proposed)

Table 1 describes the results of the simulation, including the empirical mean, standard deviation, the coverage of nominal 95% confidence intervals (CIs) of the estimate, and the mean squared error (MSE) from the true value of $\mathbf{\Gamma}$. The last row for each scenario shows each sum of the MSE ratio for MICE-FCS. With Scenario (a), namely, the normally distributed covariates, SB-MI gives the most accurate estimates, with an empirical CI coverage of approximately 0.95. However, MICE-FCS, SMC-FCS, and missForest result in biased estimates, and their CI coverages are also considerably poor for all Γs. With Scenario (b), namely, the log-normally distributed missing covariates, MICE-FCS and SMC-FCS continue to be biased with incorrect empirical CI coverage. missForest provides relatively good estimates in terms of MSE, but, once again, the CIs are poor. SB-MI gives relatively correct estimates, with the CI coverage closer to 0.95 compared with MICE-FCS, SMC-FCS, and missForest. In Scenario (c), namely, the mixture of normally distributed covariates, MICE-FCS, SMC-FCS, and missForest once again result in biased estimates, and end with poor empirical CI coverages for some Γs. On the other hand, MSE of SB-MI provides the best result among the three, and the CI coverages are very close to 0.95 for all Γs.

Figure 2 compares the results of MSE and the coverage of nominal 95% CIs. We observe

that the proposed method gives the smallest MSE and the best coverage.

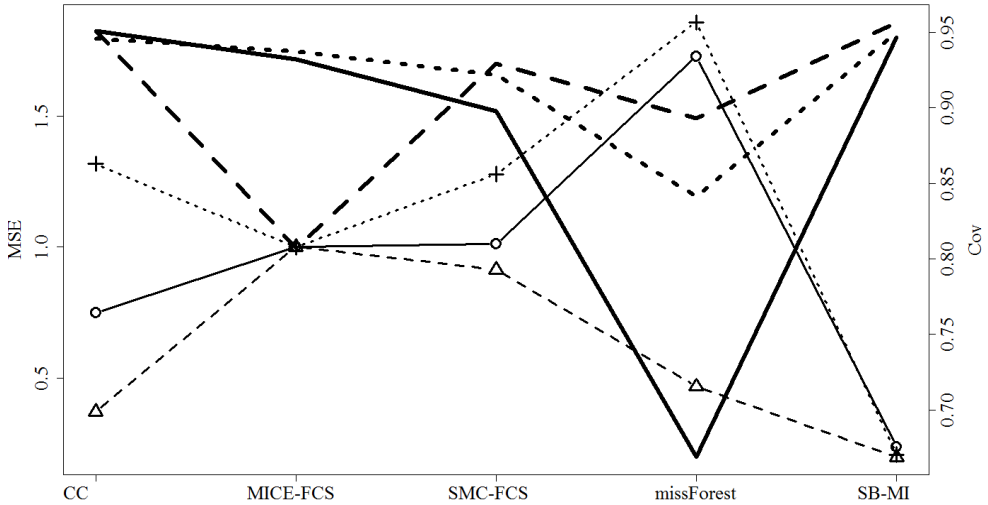

Fig.2. The thinner lines with markers and the thicker lines correspond to the left vertical axis (MSE) and the right vertical axis (Cov), respectively. The solid, dashed, and dotted lines represent the results of the following scenarios, respectively: (a) multivariate normal distribution, (b) multivariate log-normal distribution, and (c) multivariate normal mixture distribution. CC: complete case analysis, MICE-FCS: multiple imputation by chained equation-fully conditional specification, SMC-FCS: substantive model compatible-fully conditional specification, SB-MI: semiparametric Bayes-multiple imputation (proposed method).

For the other three simulations, the summarized main results are as follows. First, in terms of the MSE, the proposed method gives estimates equivalent to those found with SMC-FCS when the latter gives consistent estimates, but MICE-FCS results in biased estimates because of the violation of the model compatibility assumption. The coverage of nominal 95% CIs for the proposed method is very close to that of SMC-FCS. Note that missForest shows unbiased estimates in some situations, but produces underestimated standard deviations, and hence, poor CIs. This indicates that missForest is occasionally good at inferring unbiased estimates,

but it should not be applied in fields such as medical or epidemiologyical research, where the results of statistical significance (or hypothesis testing) are crucial.

Second, the proposed method shows smaller MSEs when the model compatibility assumptions of MICE-FCS and SMC-FCS are not satisfied. Our simulation study includes the missing covariate that follows a log-normal or mixture of normal distribution. Although MICE-FCS and SMC-FCS result in larger MSEs in these simulation settings, our proposed method gives considerably smaller MSEs. Even in these situations, the coverage of the CIs is better compared to that under the imputation methods. This indicates that the proposed method can deal with more complicated covariate distributions that the researcher cannot prespecify. Therefore, these results suggest that SB-MI approach is very practical for treating missing datasets in the real world.

5. Real data analysis

In this section, we apply our proposed algorithm to the real dataset with missing components. The data used in this implementation are sourced from the ADNI dataset described in the motivating example in Section 1.1. The substantive model in this example is the Cox proportional hazards model, which helps us study the time to conversion to AD. The samples comprise 382 observations of MCI in baseline subjects who had at least one follow-up after the first diagnosis. Of these subjects, 167 participants converted to AD during the data period. The dataset contains missing covariates $ABETA_{1-42}$, the square of $ABETA_{1-42}$, tau, p-tau, the dummy variable of whether or not the subject's mother had AD, and the dummy variable of whether or not the subject's father had AD. The dataset also contains the completely observed dummy variable APOE4, which equals 1 if the subject has the APOE4 gene, and 0 otherwise. Jack et al. (2010) found evidence that $ABETA_{1-42}$ is positively associated, p-tau is positively associated after controlling the effect of tau, and APOE4 is positively associated with the hazard of converting to AD. Bartlett et al. (2015) showed that contrary

to expectations, "mother had AD" and "father had AD" are negatively associated with the hazard of converting to AD. We should note that $ABETA_{1-42}$ of 190 observations, tau of 193 observations, p-tau of 189 observations, "mother had AD" of 77 observations, and "father had AD" of 93 observations are missing.

We employ a gamma process prior to the cumulative baseline hazard proposed by Kalbfleisch (1978), namely, $H_0 \sim GP(c_0 H^*, c_0)$. We specify the hyperparameters to be $c_0 = 0.01$, and H^* follows an exponential distribution with parameter λ. Note that Kalbfleisch (1978) and Sinha et al. (2003) showed that the estimates from the gamma process prior to the cumulative baseline hazard are equivalent to the non-Bayesian estimates based on Cox's partial likelihood (Cox, 1975) when $c_0 \to 0$ (see Chen et al. (2006) for a detailed description of the MCMC method for a Bayesian Cox regression).

Table 2 shows the results of the coefficients estimated using the Cox proportional hazards model. A positive coefficient indicates that the variable is associated with hazard of the subject converting to AD. We compare these results with those obtained using the other imputation methods, namely, MICE-FCS, SMC-FCS, and missForest. Note that the results of MICE-FCS and SMC-FCS can be biased because the model compatibility assumption is violated, as noted in Section 2. The estimated results of our proposed method (SB-MI) are different from those of MICE-FCS, SMC-FCS, missForest, as well as the complete case analysis in some ways. The results of SB-MI suggest that the effect of increasing $ABETA_{1-42}$ to the hazard of conversion is non-linear as in Bartlett et al. (2015), whereas the coefficient of $ABETA_{1-42}$ is not statistically significant for the estimates of the complete case samples, MICE-FCS, SMC-FCS, and missForest at the 5% level. The coefficients of $ABETA_{1-42}$ and $(ABETA_{1-42})^2$ from MICE-FCS are statistically significant at the 10% and 5% level, respectively, but they show signs opposite to those of SB-MI. In addition, the estimated coefficient of p-tau from SB-MI is much larger and closer to that of the complete case analysis compared to the other methods. Like the other existing methods, the hazard of the presence

of parents who had AD is not statistically significant. The presence of the APOE4 gene, which is suspected to be associated with the development of AD, is positively associated with the hazard of converting to AD for SB-MI. This relationship cannot be found when we use complete case samples only. In addition, compared with the complete case analysis, several coefficients from SB-MI are statistically significant because they avoid the restrictions posed by the complete case sample. We conduct a logistic regression and find that data availability of each biomarker is not related to time to conversion to AD at the 1% level, which suggests that the missing of the data do not depend on the outcome, and assuming unbiased results from the complete case analysis is reasonable. However, the analysis based on the restricted sample fails to detect statistically significant variables related to the hazard of converting to AD.

Table 2. Estimated results of Cox proportional hazard model

	CC (n=139)			MICE-FCS (n=382)			SMC-FCS (n=382)			missForest (n=382)			SB-MI (n=382)		
	Coef.	(s.d.)		Coef.	(s.d.)		Coef.	(s.d.)		Coef.	(s.d.)		Coef.	(s.d.)	
$ABETA_{1\text{-}42}$ (ng/mL)	0.228	(0.178)		-0.048	(0.029)	*	0.203	(0.162)		-0.012	(0.128)		0.191	(0.092)	**
$(ABETA_{1\text{-}42})^2$ (ng^2/mL^2)	-0.009	(0.005)	*	0.001	(0.000)	**	-0.008	(0.005)	*	-0.002	(0.004)		-0.009	(0.005)	*
tau (pg/mL)	-0.028	(0.032)		-0.065	(0.020)	***	-0.025	(0.027)		-0.023	(0.024)		-0.013	(0.011)	
p-tau (pg/mL)	0.282	(0.105)	***	0.241	(0.067)	**	0.131	(0.076)	*	0.163	(0.087)	*	0.250	(0.094)	***
Mother had AD	-0.259	(0.293)		-0.187	(0.187)		-0.125	(0.194)		-0.130	(0.188)		-0.098	(0.214)	
Father had AD	-0.194	(0.485)		0.028	(0.241)		-0.126	(0.285)		-0.202	(0.301)		-0.053	(0.343)	
APOE4 positive	0.001	(0.289)		0.642	(0.176)	***	0.333	(0.195)	*	0.455	(0.177)	**	0.455	(0.178)	***

*, **, and *** represent statistical significance at 10%, 5%, and 1% level, respectively. CC, complete case analysis; MICE-FCS, multiple imputation by chained equation - fully conditionalspecication; SMC-FCS, substantive model compatible - fully conditional specication; SB-MI, semiparametric Bayesian multiple imputation (proposed)

On the other hand, the results from SB-MI are consistent with the findings of previous studies, such as Hansson et al. (2006) and Okello et al. (2009), in terms of their signs. These studies employed separate longitudinal datasets and not the ADNI dataset. In conclusion, our method is very practical when applied to a real dataset, which often contains non-normally distributed and mixed-scale missing variables.

6. Discussion and conclusions

In this study, we proposed a SB-MI approach for regression models with missing mixed continuous and discrete covariates, in which the substantive model of the researcher's interest

is a parametric formulation, and the covariate distributions are nonparametric formulations employing PSBPM modeling.

If the covariate model can be correctly specified, the EM algorithm or Bayesian MCMC approach can estimate unbiased results. However, prespecifying a covariate distribution is generally impossible, especially in cases where the missing variables are continuous and discrete. The FCS approach including MICE and SMC has been widely used because researchers are not required to specify the covariate distribution among mixed-scale variables. However, these methods yield severely biased estimates if the compatibility assumption of the model is violated.

On the other hand, the SB-MI framework, which is proposed in this paper, is capable of easily dealing with incompletely observed mixed-scale variables in the covariate distribution without using FCS. Therefore, we do not have to consider whether the compatibility assumption holds, and thus, we can assume a non-linear regression, such as a linear regression with quadratic terms, Cox proportional hazards model, or logistic regression model on the substantive model even when the variables include discrete and non-normal continuous variables. The simulation studies show that the proposed method gives the best estimator in terms of MSE in cases where MICE-FCS and SMC-FCS result in biased estimates due to the violation of the model compatibility assumption. Furthermore, the proposed model is more robust when the distributions of the missing variables are non-normal. Since it is usual that some variables do not follow the normal distribution, and hence, the compatibility assumption is not satisfied, the results suggest that the SB-MI approach is very practical. Although missForest, which accommodates the random forest approach, can sometimes give estimates closer to the "true" value compared to the proposed and existing methods, it underestimates the variance of the estimates, resulting in poor CIs. Consequently, it should not be applied to fields where the results of statistical significance (or hypothesis testing) are concern to researchers. The results of the real data analysis show that our proposed method can pro-

vide new insights that cannot be obtained from existing and/or and statistically improper methods.

Further study is needed to improve the efficiency of the SB-MI algorithm. Since our inference is based on the MCMC algorithm, the computation time required to obtain valid estimates is higher than that of the existing imputation method. In addition, our model can be extended to missing not at random (MNAR) by adding a submodel of the missing mechanism to our semiparametric specification using PSBPM modeling. However, it is very difficult to correctly specify the missing mechanism even if we assume a nonparametric formulation. Consequently, we did not consider the missing mechanism to be MNAR.

Acknowledgements

Funding: This work was supported by JSPS KAKENHI Grant Number JP26285151.

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu <http://adni.loni.usc.edu>). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

References

Bartlett, J. W., Seaman, S. R., White, I. R. and Carpenter, J. R. (2015). Multiple imputation of covariates by fully conditional specification: accommodating the substantive model. *Stat. Methods. Med. Res.* **24**, 462–487.

Canale, A. and Dunson, D. B. (2011). Bayesian kernel mixtures for counts. *J. Am. Statist. Ass.* **106**, 1528–1539.

Chen, M., Ibrahim, J. G. and Shao, Q. (2006). Posterior propriety and computation for the Cox regression model with application to missing covariates. *Biometrika* **93**, 791–807.

Chib, S. (2007). Analysis of treatment response data without the joint distribution of potential outcomes. *J. Econometrics* **140**, 401–412.

Chung, Y. and Dunson, D. B. (2009). Nonparametric Bayes conditional distribution modeling with variable selection. *J. Am. Statist. Ass.* **104**, 1646–1660.

Cox, D. R. (1975). Partial likelihood. *Biometrika* **62**, 269–276.

Dunson, D. B., Pillai, N. and Park, J. H. (2007). Bayesian density regression. *J. Roy. Stat. Soc. B Met.* **69**, 163–183.

Hansson, O., Zetterberg, H., Buchhave, P., Londos, E., Blennow, K. and Minthon, L. (2006). Association between csf biomarkers and incipient alzheimer's disease in patients with mild cognitive impairment: a follow-up study. *Lancet Neurol.* **5**, 228–234.

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

Hoshino, T. (2013). Semiparametric Bayesian estimation for marginal parametric potential outcome modeling: Application to causal inference. *J. Am. Statist. Ass.* **108**, 1189–1204.

Ibrahim, J. G., Chen, M. H., Lipsitz, S. R. and Herring, A. H. (2005). Missing-data methods for generalized linear models: A comparative review. *J. Am. Statist. Ass.* **100**, 332–346.

Ishwaran, H. and James, L. F. (2011). Gibbs sampling methods for stick-breaking priors. *J. Am. Statist. Ass.* **96**, 161–173.

Jack, C. R., Wiste, H. J., Vemuri, P., Weigand, S. D., Senjem, M. L. and Zeng, G. (2010). Brain beta-amyloid measures and magnetic resonance imaging atrophy both predict time-to-progression from mild cognitive impairment to alzheimerĄfs disease. *Brain* **133**, 3336–3348.

Kalbfleisch, J. D. (1978). Non-parametric Bayesian analysis of survival time data. *J. Roy. Stat. Soc. B Met.* **40**, 214–221.

Kim, J. S. and Ratchford, B. T. (2013). A Bayesian multivariate probit for ordinal data with semiparametric random-effects. *Comput. Stat. Data An.* **64**, 192–208.

Kottas, A., Muller, P. and Quintana, F. (2005). Nonparametric Bayesian modeling for multivariate ordinal data. *J. Comp. Graph. Stat.* **14**, 610–625.

Kunihama, T., Herring, A. H., Halpern, C. T. and Dunson, D. B. (2016). Nonparametric Bayes modeling with sample survey weights. *Stat. Probabil. Lett.* **113**, 41–48.

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.

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

Murray, J. S. and Reiter, J. P. (2016). Multiple imputation of missing categorical and continuous values via bayesian mixture models with local dependence. *J. Am. Statist. Ass.* **111**, 1466–1476.

National Research Council (2010). *The prevention and treatment of missing data in clinical trials.* National Academic Press, Washington, DC.

Okello, A., Koivunen, J., Edison, P., Archer, H. A., Turkheimer, F. E., Nagren, K. U., Bullock, R., Walker, Z., Kennedy, A., Fox, N. C., Rossor, M. N., Rinne, J. O. and Brooks, D. J. (2009). Conversion of amyloid positive and negative mci to ad over 3 years an 11c-pib pet study. *Neurology* **73**, 754–760.

Pati, D., Dunson, D. B. and Tokdar, S. T. (2013). Posterior consistency in conditional distribution estimation. *J. Multivariate Anal.* **116**, 456–472.

Paton, N. I., Kityo, C., Hoppe, A., Reid, A., Kambugu, A. and Lugemwa, A. (2014). Assessment of second-line antiretroviral regimens for hiv therapy in africa. *New Engl. J. Med.* **371**, 234–247.

Rodriuez, A., Dunson, D. B. and Gelfand, A. E. (2009). Bayesian nonparametric functional

data analysis through density estimation. *Biometrika* **96**, 149–162.

Sethuraman, J. (1994). A constructive definition of dirichlet priors. *Stat. Sinica* **4**, 639–650.

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

Sinha, D., Ibrahim, J. G. and Chen, M. H. (2003). A Bayesian justification of cox's partial likelihood. *Biometrika* **90**, 629–641.

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. *Stat. Methods. Med. Res.* **16**, 219–242.

van Buuren, S. (2012). *Flexible imputation of missing data.* Chapman and Hall/CRC, Boca Raton, Florida.

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.

Walker, S. G., Damien, P., Laud, P. W. and Smith, A. F. (1999). Bayesian nonparametric inference for random distributions and related functions. *J. Roy. Stat. Soc. B Met.* **61**, 485–527.

White, I. R., Royston, P. and Wood, A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Stat. Med.* **30**, 377–399.

Appendix A

Detailed simulation design and results

In this appendix, we describe the simulation study mentioned in Section 4 in detail. We illustrate the performance of the proposed method in cases where MICE-FCS cannot draw from a Bayesian joint model. The situations considered are (i) linear regression with a quadratic term, (ii) linear regression with an interaction term, (iii) the Cox proportional hazards models, and (iv) logistic regression with a binary covariate. Throughout the simulation studies, we consider the situation outcomes of the substantive model $\boldsymbol{y}$ to be univariate. For simplicity, we also assume $\boldsymbol{y}$ contains no missing components; however, imputing from the predictive distribution (as in Step 8 in the MCMC algorithm in Section 3) allows us to address the case $\boldsymbol{y}$ contains missing components in a straightforward manner.

We consider the following throughout the simulation study: $N = 400$, the number of completely observed covariates $\boldsymbol{v}$ $(q = 1)$, and the number of incompletely observed covariates $\boldsymbol{w}$ $(p = 2)$. In order to compare the performance with MICE-FCS and SMC-FCS even when the normality assumptions are violated, we consider three cases, where the covariates $\boldsymbol{x}_i = (w_{i,1}, w_{i,2}, v_{i,1})'$ have a different data generating process from (a) a multivariate normal distribution, (b) a multivariate log-normal distribution, or (c) a multivariate normal mixture distribution. For (a), namely, the multivariate normal distribution, $\boldsymbol{x}_i$ is generated as $MVN(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$ for $i = 1, ..., N$, where $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ denotes the covariance structure with diagonal elements set to 1, and the pairwise off diagonal elements $corr(x_{ir}, x_{ir'})$ $(r = 1, 2, 3)$ are set to $\rho_{\boldsymbol{x}}^{|r-r'|}$. $\rho_{\boldsymbol{x}}$ is simulated from the uniform distribution over the interval $[0.2, 0.8]$. For (b), namely, the multivariate log-normal distribution, the missing variable $\boldsymbol{w}_i$ is simulated by exponentiating a draw from case (a). For (c), namely, the multivariate normal mixture distribution, $\boldsymbol{x}_i$ is generated from $MVN(\mathbf{0.5}, \boldsymbol{\Sigma}_{\boldsymbol{x}}^1)$ with probability 0.5 and from $MVN(\mathbf{-0.5}, \boldsymbol{\Sigma}_{\boldsymbol{x}}^2)$ with probability 0.5, where $\boldsymbol{\Sigma}_{\boldsymbol{x}}^1$ and $\boldsymbol{\Sigma}_{\boldsymbol{x}}^2$ are independently drawn in the same way as $\boldsymbol{\Sigma}_{\boldsymbol{x}}$.

We also assume the MAR missing mechanism. The elements of $\boldsymbol{y}$ are transformed to

$\lambda_{i,j} = U_{i,j} - \text{logit}^{-1}(y_i)$ $(j = 1, 2)$, where $U_{i,j}$ are *i.i.d.* uniform distributions over the interval $[0, 1]$, and the corresponding case of $w_{i,j}$ to the highest $(\chi/2)\%$ of each $\lambda_{i,1}$ and $\lambda_{i,2}$ are converted to be missing. We consistently set $\chi = 30$. For each simulation study setting, the missing components of $w_{i,j}$ are first imputed by $\boldsymbol{w}_i \sim N(\hat{\boldsymbol{\Gamma}}_l^{MLE} \boldsymbol{v}_i, \hat{\boldsymbol{\Phi}}_l^{MLE})$, where $\hat{\boldsymbol{\Gamma}}_l^{MLE}$ and $\hat{\boldsymbol{\Phi}}_l^{MLE}$ represents the maximum likelihood estimators of the complete case analysis.

We adopt the same default choices for hyperparameters as in Chung and Dunson (2009), namely, $L = 20$, $\mu_{\alpha_0} = 0$, $\sigma^2_{\alpha_0} = 1$, $\mu_{\psi_k} = 0$, $\sigma^2_{\psi_k} = 100$, and Ω^*_{km} are 50 equally spaced grid points in $(-3.5, 3.5)$, except for exponentiated $w_{i,1}$ in each simulation Scenario (b), where Ω^* are 50 equally spaced grid points in $(0, 10)$.

In order to confirm the performance of the proposed SB-MI under different model setups, we compare it with the MICE-FCS, SMC-FCS, and missForest algorithms. For FCS imputation approaches, we simulate 100 imputed datasets, and the estimates are computed using Rubin's rules. We employ, as FCS, a linear regression covariate model for continuous incomplete covariates and a logistic regression for binary ones.

A1. Linear regression with quadratic term

A1.1 *Simulation setup*

First, we simulate the case where the substantive model is a linear regression with normally distributed error terms, in which the covariates include a quadratic effect term. In this setting, the standard covariate model specification of MICE-FCS is incompatible. We specify the substantive model as follows:

$$y = \Gamma_0 + \Gamma_1 w_1 + \Gamma_2 w_2 + \Gamma_3 w_2^2 + \Gamma_4 v$$

with $\Gamma_0 = 1$, $\Gamma_1 = 1$, $\Gamma_2 = -1$, $\Gamma_3 = 1$, $\Gamma_4 = 1$, and $\epsilon \sim iid\ N(0, 0.5)$. These true coefficients are chosen as we consider a U-shaped association between the outcome y and missing variable w through the quadratic covariate effects.

In this simulation setup, the acceptance probability of the missing components discussed in Equation (2) can be written as

$$\min\left(\frac{\exp\left\{-\frac{1}{2\sigma_\epsilon^2}(y_i-\boldsymbol{\Gamma}^T\boldsymbol{x}_i^c)^2\right\}}{\exp\left\{-\frac{1}{2\sigma_\epsilon^2}(y_i-\boldsymbol{\Gamma}^T\boldsymbol{x}_i)^2\right\}},1\right)$$

where $\boldsymbol{\Gamma}=(\Gamma_0,\Gamma_1,\Gamma_2,\Gamma_3,\Gamma_4)^T$, $\boldsymbol{x}_i=(1,w_{1i},w_{2i},w_{2i}^2,v_{1i})$, and $\boldsymbol{x}_i^c$ is the vector $\boldsymbol{x}_i$ whose missing components are replaced by the candidate value. The missing values of w_2^2 are passively imputed as the squared of the imputed values of w_2.

A1.2 *Results*

Table A1 describes the results of the simulation, namely, the empirical mean, standard deviation, the coverage of nominal 95% CIs of the estimate, and the MSE from the true value of $\boldsymbol{\Gamma}$. The last row in each scenario shows each sum of the MSE ratio for MICE-FCS. With Scenario (a), namely, normally distributed covariates, SMC-FCS, missForest, and SB-MI give smaller MSEs for all Γs. Also, the CI coverage of SMC-FCS and SB-MI is very close to 0.95, although missForest results in poor CIs. However, since the imputaion model is not compatible with the substantive model, as expected, MICE-FCS results in biased estimates, not only for the quadratic term coefficient Γ_3 but also for the other coefficients. CI coverages are also slightly poor for all Γs. With Scenario (b), namely, log-normally distributed missing covariates, MICE-FCS, once again, results in severely biased estimates and poor empirical CI coverages. SMC-FCS gives comparatively correct estimates and CI coverage. missForest and SB-MI provide better estimates from the viewpoint of error from the true value, but the CI coverages of missForest are far from 0.95. Estimates of Γ_2 by SMC-FCS are more variable than those of SB-MI. With Scenario (c), namely, mixture of normally distributed covariates, MICE-FCS results in very biased estimates, and the CIs shows coverage of only $0.17-0.82$. SMC-FCS also provides slightly biased results for some Γs since the specified distributions

for the covariates are incorrect. CI coverages are also smaller than 0.95 for several Γs. SB-MI gives the smallest MSE and CI coverages are very close to 0.95 for all Γs. This indicates that the SB-MI is more robust to the complicated missing mechanism than SMC-FCS as well as MICE-FCS.

Table A1. Simulation results of case (i) - Linear regression with quadratic term

	Complete Case				MICE-FCS				SMC-FCS				missForest				SB-MI			
TRUE	Mean	(s.d.)	Cov	MSE	Mean	(s.d.)	Cov	MSE	Mean	(s.d.)	Cov	MSE	Mean	(s.d.)	Cov	MSE	Mean	(s.d.)	Cov	MSE
(a) Normal																				
$\Gamma_0 = 1$	1.465	(0.086)	0.136	0.346	1.075	(0.106)	0.862	0.023	1.000	(0.084)	0.948	0.010	0.973	(0.079)	0.858	0.010	1.029	(0.099)	0.940	0.011
$\Gamma_1 = 1$	0.848	(0.083)	0.583	0.040	0.914	(0.116)	0.899	0.028	1.001	(0.089)	0.961	0.010	1.045	(0.085)	0.801	0.016	0.987	(0.098)	0.955	0.010
$\Gamma_2 = -1$	-0.838	(0.091)	0.618	0.050	-0.812	(0.142)	0.733	0.061	-0.999	(0.099)	0.959	0.013	-1.017	(0.096)	0.904	0.017	-0.990	(0.106)	0.963	0.014
$\Gamma_3 = 1$	0.908	(0.044)	0.557	0.015	0.866	(0.072)	0.555	0.032	1.000	(0.049)	0.952	0.005	1.023	(0.047)	0.919	0.004	0.989	(0.077)	0.962	0.006
$\Gamma_4 = 1$	0.845	(0.083)	0.599	0.042	0.930	(0.107)	0.913	0.021	0.999	(0.083)	0.953	0.010	0.996	(0.080)	0.886	0.011	0.992	(0.098)	0.948	0.010
MSE ratio			2.970				1.000				0.290				0.352				0.308	
(b) Log-normal																				
$\Gamma_0 = 1$	1.515	(0.123)	0.235	0.455	1.707	(0.515)	0.781	1.092	1.013	(0.158)	0.916	0.194	1.003	(0.126)	0.941	0.020	1.002	(0.195)	0.963	0.037
$\Gamma_1 = 1$	0.966	(0.035)	0.841	0.004	1.362	(0.242)	0.766	0.305	0.999	(0.045)	0.919	0.046	1.000	(0.039)	0.936	0.002	0.997	(0.096)	0.950	0.034
$\Gamma_2 = -1$	-1.111	(0.077)	0.721	0.030	-2.042	(0.480)	0.658	2.090	-0.970	(0.110)	0.926	0.027	-1.008	(0.086)	0.938	0.012	-0.993	(0.177)	0.963	0.021
$\Gamma_3 = 1$	1.007	(0.006)	0.791	0.000	1.032	(0.034)	0.764	0.003	0.998	(0.009)	0.938	0.000	1.001	(0.007)	0.937	0.000	1.000	(0.050)	0.935	0.001
$\Gamma_4 = 1$	0.867	(0.082)	0.688	0.037	1.733	(0.438)	0.822	1.129	1.007	(0.101)	0.933	0.157	1.010	(0.080)	0.904	0.012	0.998	(0.127)	0.950	0.010
MSE ratio			0.114				1.000				0.092				0.010				0.022	
(c) Mixture of normals																				
$\Gamma_0 = 1$	1.319	(0.107)	0.437	0.213	2.730	(0.412)	0.352	4.476	0.881	(0.158)	0.867	0.038	0.801	(0.109)	0.452	0.062	1.001	(0.107)	0.941	0.026
$\Gamma_1 = 1$	0.902	(0.077)	0.806	0.025	0.470	(0.381)	0.791	0.492	0.986	(0.117)	0.938	0.019	1.106	(0.084)	0.603	0.028	0.995	(0.108)	0.949	0.018
$\Gamma_2 = -1$	-0.850	(0.093)	0.711	0.051	0.060	(0.354)	0.500	1.646	-1.108	(0.130)	0.852	0.030	-1.154	(0.096)	0.526	0.041	-0.994	(0.117)	0.958	0.021
$\Gamma_3 = 1$	0.983	(0.013)	0.801	0.001	0.585	(0.070)	0.186	0.230	1.001	(0.028)	0.955	0.001	1.010	(0.015)	0.812	0.000	0.999	(0.082)	0.962	0.001
$\Gamma_4 = 1$	0.906	(0.077)	0.800	0.023	0.690	(0.304)	0.818	0.211	1.164	(0.106)	0.697	0.042	1.093	(0.078)	0.613	0.023	0.999	(0.102)	0.951	0.015
MSE ratio			0.044				1.000				0.018				0.022				0.012	

Empirical mean, standard deviation, coverage of nominal 95% CIs, and mean squared error of the estimates, andsum of MSE Ratio of MICE-FCS basis from 1000 simulations are described. CC, complete case analysis; MICE-FCS, multiple imputation by chained equation - fully conditionalspecication; SMC-FCS, substantive model compatible - fully conditional specication; SB-MI, semiparametric Bayesian multiple imputation (Proposed)

A2. Linear regression with an interaction term

A2.1 *Simulation setup*

Next, we simulate the case where the substantive model is a linear regression with normally distributed error terms and the covariates include the cross-term effect. In this setting, the standard covariate model specification of MICE-FCS is incompatible. We consider the case where one of the incompletely observed covariates w_1 is binary, where $w_{i,1} = 1$ if the latent variable (which is simulated in each three case) $w^*_{i,1} > 0$ and $w_{i,1} = 0$ if the latent variable $w^*_{i,1} \leq 0$. Since w_1 is a binary, we do not exponentiate $w^*_{i,1}$ in the study for Scenario (b). We specify the substantive model as follows.

$$y = \Gamma_0 + \Gamma_1 w_1 + \Gamma_2 w_2 + \Gamma_3 w_1 w_2 + \Gamma_4 v_1 + \epsilon$$

with $\Gamma_0 = 1$, $\Gamma_1 = 1$, $\Gamma_2 = 1$, $\Gamma_3 = 1$, $\Gamma_4 = 1$, and $\epsilon \sim iid\ N(0, 0.5)$.

In this simulation setup, the acceptance probability of the missing components discussed in Equation (2) can be written as

$$\min\left(\frac{\exp\left\{-\frac{1}{2\sigma_\epsilon^2}(y_i - \boldsymbol{\Gamma}^T \boldsymbol{x}_i^c)^2\right\}}{\exp\left\{-\frac{1}{2\sigma_\epsilon^2}(y_i - \boldsymbol{\Gamma}^T \boldsymbol{x}_i)^2\right\}}, 1\right)$$

where $\boldsymbol{\Gamma} = (\Gamma_0, \Gamma_1, \Gamma_2, \Gamma_3, \Gamma_4)^T$, $\boldsymbol{x}_i = (1, w_{1i}, w_{2i}, w_{1i}w_{2i}, v_{1i})$, and $\boldsymbol{x}_i^c$ is the vector $\boldsymbol{x}_i$ whose missing components are replaced by the candidate value.

A2.2 *Results*

Table A2 describes the results of the simulation, including the empirical mean, standard deviation, the coverage of nominal 95% CIs of the estimate, and the MSE from the true value of $\boldsymbol{\Gamma}$. The last row of each scenario describes each sum of the MSE ratio for MICE-FCS. With Scenario (a), namely, the normally distributed covariates, SMC-FCS and SB-MI give the correct estimates. Both show empirical CI coverages of approximately 0.95. The MSEs are also similar to each other. Since the imputation model is incompatible with the substantive model, as expected, MICE-FCS, once again, results in biased estimates, and CI coverages are also considerably poor for all Γs. With Scenario (b), namely, the log-normally distributed missing covariates, MICE-FCS continues to be biased with incorrect empirical CI coverage. SMC-FCS gives biased estimates, with CI coverage of only $0.46 - 0.89$. However, SB-MI provides the estimate closest to the "true" value, the CI coverage being very close to 0.95. SB-MI seems to be more robust in situations where the distribution is skewed. With Scenario (c), namely, the mixture of normally distributed covariates, MICE-FCS again continues to give biased estimates ending in poor empirical CI coverages because of model incompatibility. SMC-FCS also provide slightly biased results for some Γs since the specified distributions for the covariates are incorrect. CI coverages are also slightly smaller than 0.95 for several Γs. On

the other hand, SB-MI gives estimates closely match the "true" value, and the CI coverages are also very close to 0.95 for all Γs. The estimates from missForest show somewhat larger MSEs. The results of the simulation for case (ii) indicates that the proposed SB-MI method is more robust than SMC-FCS and MICE-FCS in situations where the normality assumption is violated.

Table A2. Simulation results of case (ii) - Linear regression with interaction term

	Complete Case				MICE-FCS				SMC-FCS				missForest				SB-MI			
TRUE	Mean	(s.d.)	Cov	MSE	Mean	(s.d.)	Cov	MSE	Mean	(s.d.)	Cov	MSE	Mean	(s.d.)	Cov	MSE	Mean	(s.d.)	Cov	MSE
(a) Normal																				
$\Gamma_0 = 1$	1.283	(0.085)	0.379	0.162	1.389	(0.109)	0.277	0.268	0.999	(0.087)	0.938	0.011	1.012	(0.078)	0.881	0.014	1.000	(0.074)	0.938	0.006
$\Gamma_1 = 1$	0.900	(0.120)	0.901	0.033	0.872	(0.153)	0.886	0.053	1.002	(0.130)	0.950	0.025	1.007	(0.113)	0.849	0.038	1.014	(0.110)	0.943	0.011
$\Gamma_2 = 1$	0.864	(0.105)	0.781	0.046	0.822	(0.133)	0.736	0.068	1.001	(0.101)	0.951	0.015	1.122	(0.091)	0.589	0.035	1.005	(0.087)	0.954	0.010
$\Gamma_3 = 1$	0.917	(0.067)	0.809	0.018	1.056	(0.167)	0.920	0.046	1.000	(0.066)	0.956	0.007	0.971	(0.061)	0.842	0.008	0.986	(0.059)	0.932	0.015
$\Gamma_4 = 1$	1.036	(0.132)	0.932	0.033	0.892	(0.085)	0.755	0.026	0.997	(0.132)	0.949	0.023	0.883	(0.118)	0.709	0.054	0.996	(0.113)	0.935	0.010
MSE ratio			0.633				1.000				0.178				0.322				0.112	
(b) Log-normal																				
$\Gamma_0 = 1$	1.445	(0.126)	0.357	0.398	1.601	(0.152)	0.271	0.636	1.282	(0.116)	0.456	0.127	0.976	(0.110)	0.913	0.028	1.022	(0.066)	0.916	0.012
$\Gamma_1 = 1$	0.673	(0.159)	0.601	0.226	0.559	(0.192)	0.513	0.353	0.756	(0.163)	0.710	0.114	1.039	(0.147)	0.869	0.071	0.989	(0.091)	0.930	0.015
$\Gamma_2 = 1$	0.904	(0.080)	0.824	0.028	0.872	(0.096)	0.767	0.042	0.841	(0.081)	0.608	0.060	1.020	(0.079)	0.912	0.017	0.985	(0.043)	0.933	0.006
$\Gamma_3 = 1$	0.898	(0.065)	0.728	0.025	1.119	(0.101)	0.807	0.041	0.974	(0.069)	0.892	0.013	0.997	(0.059)	0.854	0.007	1.013	(0.047)	0.925	0.006
$\Gamma_4 = 1$	1.090	(0.084)	0.837	0.027	0.871	(0.078)	0.684	0.036	1.156	(0.086)	0.650	0.058	0.979	(0.083)	0.889	0.018	0.995	(0.036)	0.945	0.004
MSE ratio			0.635				1.000				0.336				0.127				0.038	
(c) Mixture of normals																				
$\Gamma_0 = 1$	1.201	(0.167)	0.851	0.122	1.310	(0.146)	0.561	0.188	0.929	(0.176)	0.932	0.056	1.113	(0.149)	0.884	0.057	0.999	(0.076)	0.955	0.006
$\Gamma_1 = 1$	0.835	(0.220)	0.910	0.119	0.859	(0.200)	0.900	0.081	1.046	(0.240)	0.948	0.093	0.912	(0.203)	0.904	0.087	1.019	(0.112)	0.935	0.013
$\Gamma_2 = 1$	1.002	(0.086)	0.959	0.010	0.880	(0.125)	0.849	0.041	0.986	(0.081)	0.948	0.010	1.105	(0.069)	0.553	0.023	1.003	(0.081)	0.955	0.007
$\Gamma_3 = 1$	0.996	(0.063)	0.963	0.006	1.095	(0.165)	0.903	0.048	0.970	(0.062)	0.919	0.007	0.966	(0.052)	0.757	0.006	0.986	(0.082)	0.928	0.011
$\Gamma_4 = 1$	0.989	(0.101)	0.953	0.014	0.923	(0.092)	0.883	0.019	1.052	(0.100)	0.940	0.027	0.920	(0.087)	0.812	0.021	0.993	(0.055)	0.929	0.004
MSE ratio			0.720				1.000				0.509				0.517				0.111	

Empirical mean, standard deviation, coverage of nominal 95% CIs, and mean squared error of the estimates, andsum of MSE Ratio of MICE-FCS basis from 1000 simulations are described. CC, complete case analysis; MICE-FCS, multiple imputation by chained equation - fully conditionalspecication; SMC-FCS, substantive model compatible - fully conditional specication; SB-MI, semiparametric Bayesian multiple imputation (proposed)

A3. Proportional hazards models

A3.1 *Simulation setup*

Next, we simulate a case where the substantive model is the proportional hazards models. In this setting, the standard covariate model specification of MICE-FCS is incompatible. We consider the case where one of the incompletely observed covariates w_1 is binary, where $w_{i,1} = 1$ if the latent variable (which is simulated in each three case) $w^*_{i,1} > 0$ and $w_{i,1} = 0$ if the latent variable $w^*_{i,1} \leq 0$. Since w_1 is binary, we do not exponentiate $w^*_{i,1}$ in the study for Scenario (b). We specify the substantive model of the hazard function as follows:

$$h(t|\boldsymbol{w}, v) = 0.002 \exp(\Gamma_1 w_1 + \Gamma_2 w_2 + \Gamma_3 v_1)$$

with $\Gamma_1 = 1$, $\Gamma_2 = 2$, and $\Gamma_3 = 4$. We generate censoring times from an exponential distribution with hazard $\lambda = 0.002$.

In this simulation setup, we assume a Weibull distribution for the hazard function, and the acceptance probability of the missing components discussed in Equation (2) can be written as

$$\min\left(\frac{\exp\left\{d\left[\boldsymbol{\Gamma}^T \boldsymbol{x}_i^c + \log(\lambda\alpha) + (\alpha-1)\log(\lambda t)\right] - \exp\left(\boldsymbol{\Gamma}^T \boldsymbol{x}_i^c\right)(\lambda t)^\alpha\right\}}{\exp\left\{d\left[\boldsymbol{\Gamma}^T \boldsymbol{x}_i + \log(\lambda\alpha) + (\alpha-1)\log(\lambda t)\right] - \exp\left(\boldsymbol{\Gamma}^T \boldsymbol{x}_i\right)(\lambda t)^\alpha\right\}}, 1\right)$$

where $\boldsymbol{\Gamma} = (\Gamma_0, \Gamma_1, \Gamma_2, \Gamma_3)^T$, $\boldsymbol{x}_i = (w_{1i}, w_{2i}, v_{1i})$, and $\boldsymbol{x}_i^c$ is the vector $\boldsymbol{x}_i$ whose missing components are replaced by the candidate value.

A3.2 *Results*

Table A3 shows the simulation results, including the empirical mean, standard deviation, coverage of nominal 95% CIs of the estimate, and MSE from the true value of $\boldsymbol{\Gamma}$ are described. The last row of each scenario in Table A3 provides each sum of the MSE ratio on MICE-FCS. With Scenario (a), namely, normally distributed covariates, SMC-FCS and SB-MI result in estimates with the smallest MSE for all Γs. Also, the CI coverage in both case is very close to 0.95. The MSE of SB-MI is somewhat smaller than that of SMC-FCS. On the other hand, MICE-FCS results in biased estimates for all the coefficients owing to the violation of model compatibility. Hence, CI coverages are also poor for all Γs. missForest gives relatively smaller MSEs, but the empirical coverages are very poor. With Scenario (b), namely, log-normally distributed missing covariates, MICE-FCS again shows severely biased estimates and poor empirical coverages. SMC-FCS gives biased estimates, and the CI coverages of Γs are much lower than 0.95. These biased results arise from the model incompatibility on FCS. missForest, once again, shows poor empirical coverages. Of all these results, SB-MI gives the most valid estimates, with the CI coverage being closest to 0.95. The MSEs from SB-MI are

much smaller than those from SMC-FCS. With Scenario (c), namely, a mixture of normally distributed covariates, where FCS does not satisfy the model compatibility assumption, the results are similar to Scenario (b).

Table A3. Simulation results of case (iii) - Proportional hazards models

	Complete Case				MICE-FCS				SMC-FCS				missForest				SB-MI			
TRUE	Mean	(s.d.)	Cov	MSE	Mean	(s.d.)	Cov	MSE	Mean	(s.d.)	Cov	MSE	Mean	(s.d.)	Cov	MSE	Mean	(s.d.)	Cov	MSE
(a) Normal																				
$\Gamma_0 = 1$	1.045	(0.201)	0.942	0.043	0.772	(0.205)	0.818	0.085	1.022	(0.188)	0.948	0.036	0.904	(0.169)	0.856	0.051	1.026	(0.157)	0.948	0.023
$\Gamma_1 = 2$	2.195	(0.182)	0.929	0.243	1.298	(0.215)	0.166	0.512	2.035	(0.162)	0.949	0.029	2.024	(0.144)	0.887	0.032	1.976	(0.132)	0.949	0.024
$\Gamma_2 = 4$	4.188	(0.295)	0.920	0.123	3.176	(0.257)	0.185	0.752	4.070	(0.254)	0.936	0.079	3.739	(0.207)	0.672	0.150	3.966	(0.137)	0.936	0.021
MSE ratio			0.304				1.000				0.107				0.173				0.051	
(b) Log-normal																				
$\Gamma_0 = 1$	1.050	(0.171)	0.939	0.034	0.730	(0.176)	0.653	0.107	0.872	(0.163)	0.933	0.189	0.815	(0.145)	0.690	0.069	1.032	(0.172)	0.932	0.016
$\Gamma_1 = 2$	2.203	(0.148)	0.918	0.632	0.700	(0.142)	0.050	1.718	1.744	(0.134)	0.542	0.201	1.815	(0.102)	0.576	0.111	1.980	(0.049)	0.941	0.013
$\Gamma_2 = 4$	4.292	(0.259)	0.910	0.152	2.372	(0.182)	0.100	2.725	3.674	(0.227)	0.634	0.199	3.247	(0.167)	0.152	0.736	4.022	(0.153)	0.940	0.047
MSE ratio			0.180				1.000				0.129				0.201				0.017	
(c) Mixture of normals																				
$\Gamma_0 = 1$	1.029	(0.517)	0.953	0.277	1.122	(0.292)	0.977	0.075	1.062	(0.481)	0.944	0.259	0.835	(0.468)	0.918	0.260	1.038	(0.162)	0.940	0.031
$\Gamma_1 = 2$	2.105	(0.263)	0.940	0.083	1.015	(0.205)	0.010	0.985	1.879	(0.214)	0.901	0.161	2.132	(0.182)	0.896	0.058	1.944	(0.126)	0.933	0.023
$\Gamma_2 = 4$	4.213	(0.426)	0.941	0.189	3.434	(0.248)	0.414	0.396	4.210	(0.319)	0.929	0.207	3.547	(0.241)	0.512	0.183	3.945	(0.130)	0.930	0.015
MSE ratio			0.377				1.000				0.430				0.344				0.048	

Empirical mean, standard deviation, coverage of nominal 95% CIs, and mean squared error of the estimates, andsum of MSE Ratio of MICE-FCS basis from 1000 simulations are described. CC, complete case analysis; MICE-FCS, multiple imputation by chained equation - fully conditionalspecication; SMC-FCS, substantive model compatible - fully conditional specication; SB-MI, semiparametric Bayesian multiple imputation (Proposed)

A4. Logistic regression with a binary covariate

A4.1 *Simulation setup*

Last, we simulate the case where the substantive model is a logistic regression with a binary outcome, in which the incomplete covariates include a binary variable. In this setting, the standard covariate model specification of MICE-FCS is incompatible and incapable of drawing from a Bayesian joint model. We consider the case where one of the incompletely observed covariates w_1 is binary, where $w_{i,1} = 1$ if the latent variable (which is simulated in each three case) $w^*_{i,1} > 0$ and $w_{i,1} = 0$ if the latent variable $w^*_{i,1} \leq 0$. Since w_1 is binary, we do not exponentiate $w^*_{i,1}$ in the study for Scenario (b). In addition, the outcome y is also binary; hence, the substantive model is a logistic regression. We specify the substantive model as follows:

$$\text{logit}(y = 1) = \Gamma_0 + \Gamma_1 w_1 + \Gamma_2 w_2 + \Gamma_3 v_1 + \epsilon$$

with $\Gamma_0 = 1$, $\Gamma_1 = 2$, $\Gamma_2 = -2$, and $\Gamma_3 = 3$.

In this simulation setup, the acceptance probability of the missing components discussed in Equation (2) can be written as

$$\min\left(\frac{\left\{\exp(\boldsymbol{\Gamma}^T\boldsymbol{x}_i^c)/\left[1+\exp(\boldsymbol{\Gamma}^T\boldsymbol{x}_i^c)\right]\right\}^{y_i}\left\{1/\left[1+\exp(\boldsymbol{\Gamma}^T\boldsymbol{x}_i^c)\right]\right\}^{1-y_i}}{\left\{\exp(\boldsymbol{\Gamma}^T\boldsymbol{x}_i)/\left[1+\exp(\boldsymbol{\Gamma}^T\boldsymbol{x}_i)\right]\right\}^{y_i}\left\{1/\left[1+\exp(\boldsymbol{\Gamma}^T\boldsymbol{x}_i)\right]\right\}^{1-y_i}},1\right)$$

where $\boldsymbol{\Gamma} = (\Gamma_0, \Gamma_1, \Gamma_2, \Gamma_3)^T$, $\boldsymbol{x}_i = (1, w_{1i}, w_{2i}, v_{1i})$, and $\boldsymbol{x}_i^c$ is the vector $\boldsymbol{x}_i$ whose missing components are replaced by the candidate value.

A4.2 *Results*

The results and discussions are described in the main paper and Table 1. In this supplementary material, we additionally provide Figure A1. Figure A1 presents the boxplot of biases in the estimates of coefficient Γ_0 compared with those of the "true" value $\Gamma_0 = 1$. Figure 3 indicates that the simulation using the proposed SB-MI method gives estimates most similar to the complete data for all scenarios.

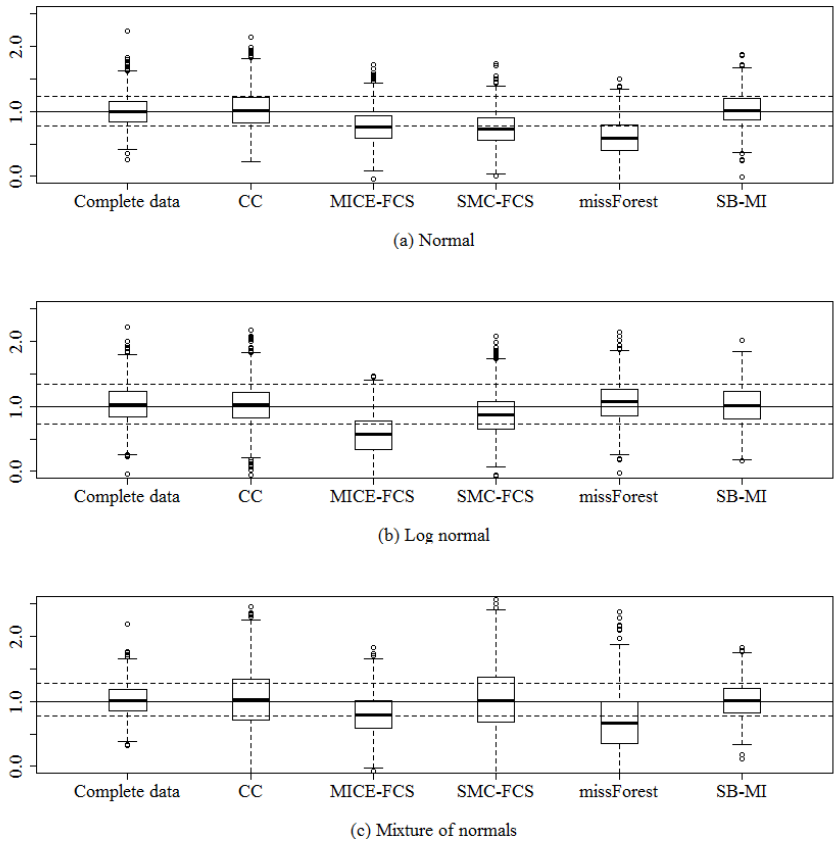

Fig.A1. The solid horizontal line is the "true" coefficient value $\Gamma_0 = 1$, and the dashed horizontal lines show empirical standard error ± 1. The boxes span the range from the 25th to the 75th percentiles, and the whiskers extend to an area no more than 1.5 times the range from the 25th to the 75th percentiles from the box. The circles above and below the whiskers represent outliers. CC: complete case analysis, MICE-FCS: multiple imputation by chained equation-fully conditional specification, SMC-FCS: substantive model compatible-fully conditional specification, SB-MI: semiparametric Bayes-multiple imputation (proposed method).