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

Ryo Kato · Takahiro Hoshino

1 Appendix1: 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 $\mathbf{y}$ to be univariate. For simplicity, we also assume $\mathbf{y}$ 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 $\mathbf{y}$ contains missing components in a straightforward manner.

We consider the following throughout the simulation study: $N = 400$, the number of completely observed covariates $\mathbf{v}$ ($q = 1$), and the number of incompletely observed covariates $\mathbf{w}$ ($p = 2$). 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 $\mathbf{x}_i = (w_{i,1}, w_{i,2}, v_{i,1})'$ 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, $\mathbf{x}_i$ is generated as $MVN(\mathbf{0}, \mathbf{\Sigma_x})$ for $i = 1, ..., N$, where $\mathbf{\Sigma_x}$ denotes the covariance structure with diagonal elements set to 1, and the pairwise off di-

R. Kato
Research Institute for Economics and Business Administration, Kobe University, 2-1 Rokkodai-cho, Nada-ku, Kobe, Japan
E-mail: kato.ryo@keio.jp

T. Hoshino (corresponding author)
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
E-mail: bayesian@jasmine.ocn.ne.jp

agonal elements $corr(x_{ir}, x_{ir'})$ $(r = 1, 2, 3)$ are set to $\rho_{\mathbf{x}}^{|r-r'|}$. $\rho_{\mathbf{x}}$ is simulated from the uniform distribution over the interval $[0.2, 0.8]$. For (b), namely, the multivariate log-normal distribution, the missing variable $\mathbf{w}_i$ is simulated by exponentiating a draw from case (a). For (c), namely, the multivariate normal mixture distribution, $\mathbf{x}_i$ is generated from $MVN(\mathbf{0.5}, \mathbf{\Sigma}_{\mathbf{x}}^1)$ with probability 0.5 and from $MVN(\mathbf{-0.5}, \mathbf{\Sigma}_{\mathbf{x}}^2)$ with probability 0.5, where $\mathbf{\Sigma}_{\mathbf{x}}^1$ and $\mathbf{\Sigma}_{\mathbf{x}}^2$ are independently drawn in the same way as $\mathbf{\Sigma}_{\mathbf{x}}$.

We also assume the MAR missing mechanism. The elements of $\mathbf{y}$ 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 $\mathbf{w}_i \sim N(\hat{\mathbf{\Gamma}}_l^{MLE}\mathbf{v}_i, \hat{\mathbf{\Phi}}_l^{MLE})$, where $\hat{\mathbf{\Gamma}}_l^{MLE}$ and $\hat{\mathbf{\Phi}}_l^{MLE}$ 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, $L = 20$, $\mu_{\alpha_0} = 0$, $\sigma_{\alpha_0}^2 = 1$, $\mu_{\psi_k} = 0$, $\sigma_{\psi_k}^2 = 100$, 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 this simulation study, the MCMC algorithm was run for 8,000 iterations, with the first 4,000 iterations excluded as a burn-in period. We confirmed the convergence using a diagnostic proposed by Geweke (1992).

In order to confirm the performance of the proposed SB-MI under different model setups, we compare it with the MICE-FCS, SMC-FCS, NP-MI, and missForest algorithms (Since NP-MI is developed to deal with categorical variable imputation and set up of Simulation-(i) do not include categorical variable, only simulation-(i) does not include NP-MI). 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. For NP-MI, MCMC algorithm was run for 8,000 iterations, with the first 4,000 iterations discarded, and we simulated 100 imputed datasets. The estimates from 100 datasets are integrated by Rubin's rules.

1.1 Linear regression with quadratic term

1.1.1 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:

$$y = \mathit{\Gamma}_0 + \mathit{\Gamma}_1 w_1 + \mathit{\Gamma}_2 w_2 + \mathit{\Gamma}_3 w_2^2 + \mathit{\Gamma}_4 v$$

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

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

where $\mathbf{\Gamma} = (\Gamma_0, \Gamma_1, \Gamma_2, \Gamma_3, \Gamma_4)^T$, $\mathbf{x}_i = (1, w_{1i}, w_{2i}, w_{2i}^2, v_{1i})$, and $\mathbf{x}_i^c$ is the vector $\mathbf{x}_i$ 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.

1.1.2 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 $\mathbf{\Gamma}$. 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.

1.2 Linear regression with an interaction term

1.2.1 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

$$\min\left(\frac{\exp\left\{-\frac{1}{2\sigma^2_\epsilon}(y_i - \mathbf{\Gamma}^T \mathbf{x}^c_i)^2\right\}}{\exp\left\{-\frac{1}{2\sigma^2_\epsilon}(y_i - \mathbf{\Gamma}^T \mathbf{x}_i)^2\right\}}, 1\right)$$

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

1.2.2 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 $\mathbf{\Gamma}$. 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. NP-MI also gives biased results with poor CIs. 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 $0.46 - 0.89$. NP-MI provides 18 times large MSE than the proposed. 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 and NP-MI again continues to give biased estimates ending in poor empirical CI coverages. 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.

1.3 Proportional hazards models

1.3.1 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:

$$h(t|\mathbf{w}, v) = 0.002 \exp(\mathit{\Gamma}_1 w_1 + \mathit{\Gamma}_2 w_2 + \mathit{\Gamma}_3 v_1)$$

with $\mathit{\Gamma}_1 = 1$, $\mathit{\Gamma}_2 = 2$, and $\mathit{\Gamma}_3 = 4$. 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

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

where $\mathbf{\Gamma} = (\mathit{\Gamma}_0, \mathit{\Gamma}_1, \mathit{\Gamma}_2, \mathit{\Gamma}_3)^T$, $\mathbf{x}_i = (w_{1i}, w_{2i}, v_{1i})$, and $\mathbf{x}_i^c$ is the vector $\mathbf{x}_i$ whose missing components are replaced by the candidate value.

1.3.2 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 $\mathbf{\Gamma}$ 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 $\mathit{\Gamma}$s. 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 $\mathit{\Gamma}$s. 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 $\mathit{\Gamma}s$ 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). Throughout this simulation study, NP-IV estimates are severely biased and MSE is at most 60 times larger than the proposed because the analysis model (proportional hazards model) is thought to be uncongenial to nonparametric imputation model.

We also provide Figure A1, which presents the boxplot of biases in the estimates of coefficient Γ_2 compared with those of the "true" value $\Gamma_2 = 2$. Figure A1 indicates that the simulation using the proposed SB-MI method gives estimates most similar to the complete data for all scenarios.

1.4 Logistic regression with a binary covariate

1.4.1 Simulation setup

Lastly, 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:

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

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

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

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

1.4.2 Results

The results and discussions are described in the main paper and Table 1. In this supplementary material, we additionally provide Figure A2. Figure A2 presents the boxplot of biases in the estimates of coefficient Γ_0 compared with those of the "true" value $\Gamma_0 = 1$.

As can be seen from the figure, complete case analysis in the logistic regression model does not have any bias. The robustness of logistic regression model to the complete case is proven by Vach and Blettner (1991). However,

its efficiency is inferior to the proposed imputation method due to the decrease in sample size. Therefore, the MSE of the proposed method is smaller than the complete case analysis (see also Table 1 in the main manuscript).

Figure A2 also indicates that the simulation using the proposed SB-MI method gives estimates most similar to the complete data for all scenarios. The sum of MSEs of the proposed method are 1,24, 1,41, 1.35 times as large as the complate data (full sample) estimates for Scenario (a), Scenario (b), and Scenario (c), respectively.

2 Appendix2: Acceptance rate of missing values

In our MCMC algorithm in Step 8, we use Metropolis-Hastings algorithm. Table A4 describes the average acceptance rate of the missing values. Since we employ $p(\mathbf{w}|\mathbf{v}, \boldsymbol{\vartheta}_m)$ as a proposal density and this seems to close to the desired distribution, the algorithm attains comparatively efficient sampling.

3 Appendix3: Sensitivity analysis on hyperparameters

We adopted the default choices for the hyperparameters according to the recommendation of Chung and Dunson (2009), which considered a variable selection problem. We partially introduced structure from Chung and Dunson (2009) to consider the dependence of stick-breaking weight on the covariate. Therefore, our model is not relevant to the hyperparameters on the variable selection. However, it is important to confirm the sensitivity to the hyperparameters, and we conducted sensitivity analysis.

Our predetermined hyperparameters are $\mu_{\psi_k} = 0$, $\sigma^2_{\psi_k} = 100$, $\mu_{\alpha_0} = 0$, and $\sigma^2_{\alpha_0} = 1$. Since $\mu_{\psi_k} = 0$, $\sigma^2_{\psi_k} = 100$ are set to be an uninformative prior on ψ_{lk}, they are reasonable. $\mu_{\alpha_0} = 0$ is also set to be uninformative, but $\sigma^2_{\alpha_0} = 1$ seems to be arbitrary. Then, we conducted sensitivity analysis for various values of $\sigma^2_{\alpha_0}$. Under simulation (ii)-(c) setting, we confirmed 6 types of hyperparameter settings, $\sigma^2_{\alpha_0}$=0.5, 1(default), 2, 5, 10, and 100 from 500 datasets with 8,000 iterations, with the first 4,000 iterations discarded as a burn-in. The results are described in Table A5. It tells us that the estimates are very similar each other even if we vary the value of $\sigma^2_{\alpha_0}$, indicating our model is robust to the changes in hyperparameter $\sigma^2_{\alpha_0}$.

References

Chung, Y., and Dunson, D. B. (2009). Nonparametric Bayes conditional distribution modeling with variable selection. *Journal of the American Statistical Association*, 104, 1646-1660.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.) *Bayesian Statistics, Vol4.*, Oxford University Press, New York.

Vach, W., and Blettner, M. (1991). Biased estimation of the odds ratio in case-control studies due to the use of ad hoc methods of correcting for missing values for confounding variables. *American journal of epidemiology*, 134, 895-907.

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; NP-MI, nonparametric multiple imputation; missForest, random forest approach for missing value prediction; SMC-FCS, substantive model compatible - fully conditional specication; SB-MI, semiparametric Bayesian multiple imputation (Proposed)

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

	Complete Case				MICE-FCS				SMC-FCS				NP-MI				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	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	1.066	(0.087)	0.837	0.016	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	0.923	(0.131)	0.903	0.031	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	0.815	(0.114)	0.600	0.064	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	1.202	(0.146)	0.679	0.102	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	0.985	(0.067)	0.951	0.007	0.883	(0.118)	0.709	0.054	0.996	(0.113)	0.935	0.010
MSE ratio				0.633				1.000				0.178				0.477				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	1.435	(0.127)	0.246	0.336	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	0.620	(0.162)	0.469	0.278	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	0.835	(0.076)	0.510	0.067	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	1.162	(0.081)	0.579	0.071	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	0.935	(0.065)	0.870	0.013	0.979	(0.083)	0.889	0.018	0.995	(0.036)	0.945	0.004
MSE ratio				0.635				1.000				0.336				0.691				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	1.007	(0.115)	0.913	0.021	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	1.000	(0.165)	0.936	0.041	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	0.854	(0.103)	0.641	0.041	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	1.190	(0.140)	0.652	0.080	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	0.931	(0.070)	0.831	0.015	0.920	(0.087)	0.812	0.021	0.993	(0.055)	0.929	0.004
MSE ratio				0.720				1.000				0.509				0.525				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; NP-MI, nonparametric multiple imputation; missForest, random forest approach for missing value prediction; SMC-FCS, substantive model compatible - fully conditional specication; SB-MI, semiparametric Bayesian multiple imputation (Proposed)

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

	Complete Case				MICE FCS				SMC FCS				NP MI				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	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	0.778	(0.204)	0.841	0.081	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	1.295	(0.216)	0.168	0.517	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	3.165	(0.257)	0.197	0.773	3.739	(0.207)	0.672	0.150	3.966	(0.137)	0.936	0.021
MSE ratio				0.304				1.000				0.107				1.016				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	0.706	(0.175)	0.615	0.116	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	0.695	(0.140)	0.000	1.739	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	2.369	(0.182)	0.000	2.731	3.247	(0.167)	0.152	0.736	4.022	(0.153)	0.940	0.047
MSE ratio				0.180				1.000				0.129				1.008				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	1.274	(0.287)	0.886	0.148	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	1.008	(0.207)	0.006	1.002	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	3.470	(0.250)	0.457	0.361	3.547	(0.241)	0.512	0.183	3.945	(0.130)	0.930	0.015
MSE ratio				0.377				1.000				0.430				1.038				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; NP-MI, nonparametric multiple imputation; missForest, random forest approach for missing value prediction; SMC-FCS, substantive model compatible - fully conditional specication; SB-MI, semiparametric Bayesian multiple imputation (Proposed)

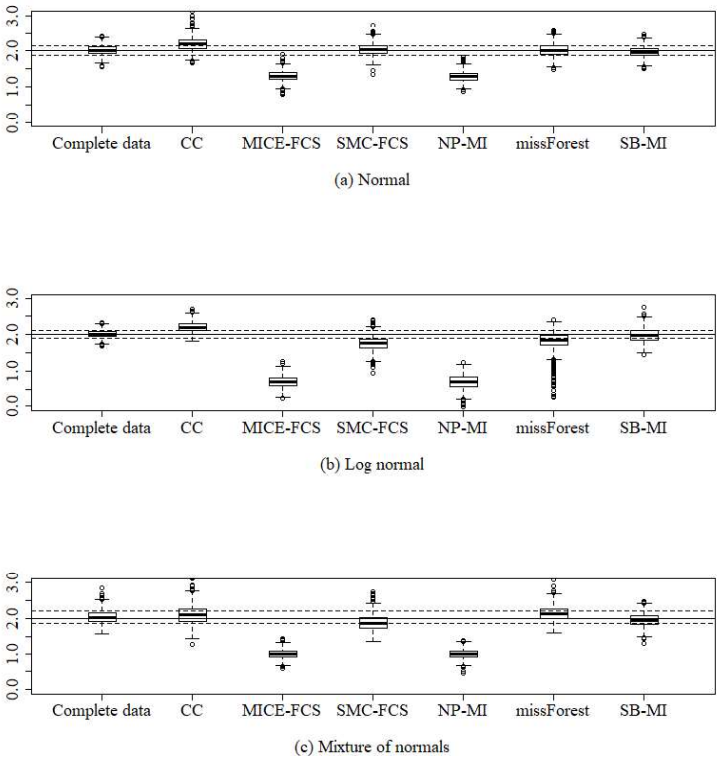

Fig.A1. The solid horizontal line is the "true" coefficient value $\Gamma_2 = 2$, and the dashed horizontal lines show empirical standard error ± 1. The boxes span the range from the 25th to the 75th percentiles, and the whiskers extend to an area no more than 1.5 times the range from the 25th to the 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, NP-MI: nonparametric multiple imputation, missForest: random forest approach for missing value prediction, SB-MI: semiparametric Bayes-multiple imputation (proposed method).

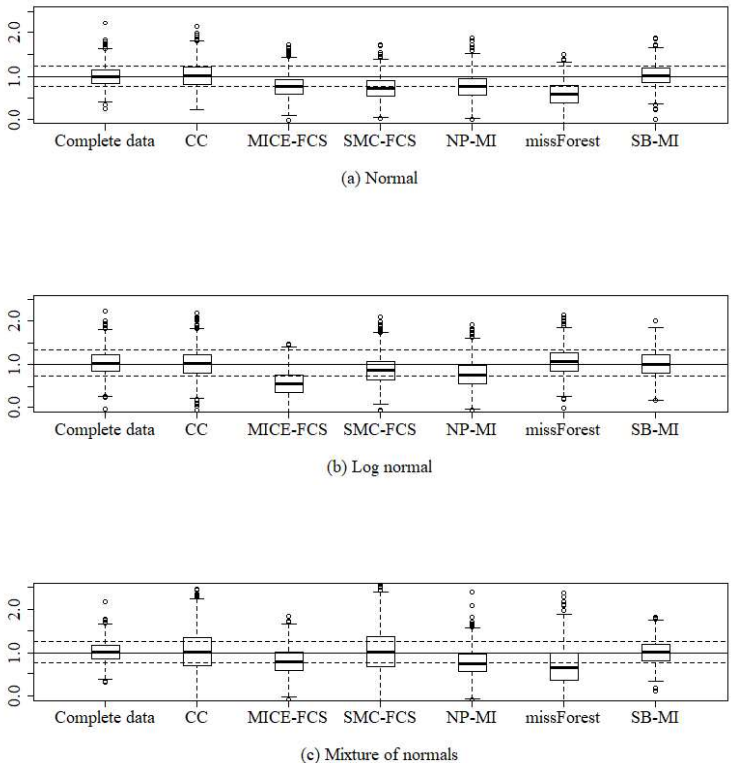

Fig.A2. The solid horizontal line is the "true" coefficient value $\Gamma_0 = 1$, and the dashed horizontal lines show empirical standard error ± 1. The boxes span the range from the 25th to the 75th percentiles, and the whiskers extend to an area no more than 1.5 times the range from the 25th to the 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, NP-MI: nonparametric multiple imputation, missForest: random forest approach for missing value prediction, SB-MI: semiparametric Bayes-multiple imputation (proposed method).

Table A4. Average acceptance rates of missing values in M-H step

	Simulation (i)	Simulation (ii)	Simulation (iii)	Simulation (iv)
	acceptance rate	acceptance rate	acceptance rate	acceptance rate
(a) Normal	66.4%	62.9%	50.0%	56.7%
(b) Log-normal	62.2%	60.6%	44.3%	48.2%
(c) Mixture of normals	65.9%	65.1%	51.2%	53.2%

The percentage shows the acceptance rate of M-H algorithm in Step 8 of the manuscript for each simulation settings. The rates are the average of the M-H acceptance for all the missing values in each simulation.

Table A5. Resutls of sensitivity analysis

	$\sigma^2_{\alpha_0} = 0.5$		$\sigma^2_{\alpha_0} = 1$		$\sigma^2_{\alpha_0} = 2$		$\sigma^2_{\alpha_0} = 5$		$\sigma^2_{\alpha_0} = 10$		$\sigma^2_{\alpha_0} = 100$	
	Mean	(s.d.)	Mean	(s.d.)	Mean	(s.d.)	Mean	(s.d.)	Mean	(s.d.)	Mean	(s.d.)
$\Gamma_0 = 1$	0.991	(0.075)	0.999	(0.076)	0.986	(0.075)	0.983	(0.075)	0.992	(0.075)	0.991	(0.075)
$\Gamma_1 = 1$	0.994	(0.109)	1.019	(0.112)	1.006	(0.108)	1.000	(0.110)	0.994	(0.108)	0.995	(0.108)
$\Gamma_2 = 1$	1.000	(0.082)	1.003	(0.081)	0.998	(0.080)	0.985	(0.085)	0.995	(0.088)	0.992	(0.088)
$\Gamma_3 = 1$	1.016	(0.082)	0.986	(0.082)	0.999	(0.085)	1.009	(0.084)	1.006	(0.081)	1.004	(0.087)
$\Gamma_4 = 1$	0.999	(0.051)	0.993	(0.055)	1.001	(0.059)	1.002	(0.059)	0.998	(0.057)	1.003	(0.056)

Empirical mean and standard deviation from 500 simulations are described.