Journal of Business & Economic Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ubes20

Imputations for High Missing Rate Data in Covariates via Semi-supervised Learning Approach

Wei Lan, Xuerong Chen, Tao Zou & Chih-Ling Tsai

To cite this article: Wei Lan, Xuerong Chen, Tao Zou & Chih-Ling Tsai (2021): Imputations for High Missing Rate Data in Covariates via Semi-supervised Learning Approach, Journal of Business & Economic Statistics, DOI: 10.1080/07350015.2021.1922120

To link to this article: https://doi.org/10.1080/07350015.2021.1922120

Imputations for High Missing Rate Data in Covariates via Semi-supervised Learning Approach

Wei Lan, Xuerong Chen, Tao Zou and Chih-Ling Tsai

Southwestern University of Finance and Economics, The Australian National University and University of California, Davis

Corresponding author Wei Lan lanwei@swufe.edu.cn

Abstract

Advancements in data collection techniques and the heterogeneity of data resources can yield high percentages of missing observations on variables, such as block-wise missing data. Under missing-data scenarios, traditional methods such as the simple average, *k*-nearest neighbor, multiple, and regression imputations may lead to results that are unstable or unable be computed. Motivated by the concept of semi-supervised learning (see, e.g., Zhu and Goldberg, 2009 and Chapelle et al., 2010), we propose a novel approach with which to fill in missing values in covariates that have high missing rates. Specifically, we consider the missing and non-missing subjects in any covariate as the unlabelled and labelled target outputs, respectively, and treat their corresponding responses as the unlabelled and labelled inputs. This innovative setting allows us to impute a large number of missing data without imposing any model assumptions. In addition, the resulting imputation has a closed form for continuous covariates, and it can be calculated efficiently. An analogous procedure is applicable for discrete covariates. We further employ the nonparametric techniques to show the theoretical properties of imputed covariates. Simulation studies and an online consumer finance example are presented to illustrate the usefulness of the proposed method.

KEY WORDS: Block-wise Missing; Cross-validation; High Missing Rate Data; Interchangeable Imputation; Semi-supervised Imputation

1 INTRODUCTION

Missing data is a common challenge encountered by researchers and practitioners. This is because ignoring missing values often results in loss of information and yields biased parameter estimates. To account for possible bias and retain the representativeness of the data, several traditional techniques have been proposed such as *k*-nearest neighbor (KNN, Troyanskaya et al., 2001), regression imputation (RI, Little and Rubin, 2002; Shao and Wang, 2002), multiple imputation (MI, Little and Rubin, 2002), random forest (RF, Stekhoven and Bühlmann, 2012), complete-case analysis (CC, see Little, 1992), full-information maximum likelihood estimation (FIME, Collins, et al., 2001; Myrtveit et al., 2001; Allison, 2012), weighted estimation equation (WEE, Robins et al., 1994; Lipsitz et al., 1999) and multiple robust estimation (MRE, Han and Wang, 2013; Chen and Haziza, 2017).

Among the aforementioned methods, FIME, WEE, and MRE mainly aim to alleviate bias and improve efficiency in parameter estimations. Hence, they often require model structure and additional assumptions. On the other hand, KNN, RF, MI and RI generally focus on imputation of the missing observations. These methods usually do not need to impose any model structure nor other assumptions. Although the popular CC method does not necessarily require imposing assumptions, it is not used for imputation. None of these methods are designed for addressing high missing rate data in covariates. Accordingly, their imputations can be unstable or even uncomputable.

Due to the rapid development of advanced technology and the heterogeneity of data resources, it is expected that data collection with high missing rates will become more prevalent. One example is data collected from multiple sources in which only a small portion of observations contain the complete information across all sources. In this case, the data has block-wise missing entries, which can result in high missing rates in covariates (see, e.g., Xiang et al., 2014 and Fang et al., 2019). Another example is data collected from online survey in which the respondents are not required to answer all questions. As a result, some

questions, such as income and age, can have low response rates due to privacy, and thus yields high missing rates (see, e.g., Nulty, 2008; Bollinger and Hirsch, 2013). This paper considers a real application for risk management in online consumer finance in which we assess the repayment ability of each loan applicant in order to determine their credit quota. Table 1 presents the data collected from five sources, yielding six distinct patterns of high missing rates; the total missing rate is 49.4%, and only 4.7% applicants have complete covariate information. A detailed description of this example can be found in Section 3.

Inspired by practical need, there are several methods proposed in the literature that manage high missing rate data such as block-wise missing data (see, e.g., Zhou et al., 2010, Yuan et al. 2012 and Xiang et al., 2014). These methods basically require some model assumptions in order to achieve their specific goals. To complement those approaches, we propose a new procedure motivated by semi-supervised learning (SSL; see, e.g., Zhu and Goldberg, 2009 and Chapelle et al., 2010). Specifically, we consider the missing and non-missing subjects in any covariate as the unlabelled and labelled target outputs, respectively. Then we treat their corresponding responses as the unlabelled and labelled inputs. Under this innovative setting, we are able to develop an imputation method that is applicable for high missing rate data without imposing any model structure assumptions. Since our proposed method derives from semi-supervised learning, we simply name it semi-supervised imputation (SSI).

The SSI method utilizes information from both observed and unobserved subjects across responses and covariates. Suppose that the i-th subject is missing at the j-th covariate (i.e., X_{ij}). We then formulate it as a function of the weighted average of the rest of the subjects in $\mathbb{X}_j$. The weight between X_{ij} and any other subject X_{lj} ($l \neq i$) is determined by their similarity (or distance) measure, which is constructed by using the X_{ij}'s and X_{lj}'s corresponding responses and their commonly observed covariates. We also allow the weight to

depend on scale-parameters for controlling the magnitude of similarity. As a result, we are able to obtain a closed-form imputation of X_{ij}.

Formulating the imputation task through the concept of semi-supervised learning has three important merits. First, SSL has the capability of handling large amounts of unlabeled data with promising results (e.g., see Zhu et al., 2003). Accordingly, SSI enables us to manage high missing rate data with a model-free process regardless of missing mechanisms such as missing completely at random, missing at random, and missing not at random. Second, SSI has the closed form for continuous covariates, which can be calculated efficiently even for a large number of missing subjects. In addition, this closed form allows us to develop theoretical properties of SSI. An analogous approach is applicable to discrete covariates. Third, SSI not only uses the information from the observed responses and covariates, but also utilizes the information from those covariates with missing subjects. Thus, SSI can yield more reliable results than commonly used methods such as KNN, RF, MI, and RI.

The rest of this article is organized as follows. Section 2 introduces SSI and demonstrates the theoretical properties of imputed values with continuous covariates and discrete covariates, respectively. In addition, the interchangeable imputation method for finding unknown scale-parameters is proposed. Based on both observed data and imputed values, we further introduce a linear regression model. This allows us not only to obtain regression parameter estimators and predictions, but also to propose a cross-validation approach for estimating unknown scale-parameters. Monte Carlo studies and an online consumer finance example are presented in Section 3, which indicate that SSI performs well. The article concludes with some discussion in Section 4, while all technical details are relegated to the supplementary material.

2 SEMI-SUPERVISED IMPUTATION FOR COVARIATES

This section contains three subsections, namely SSI for continuous covariates, SSI for discrete covariates, and scale-parameters selection.

2.1 SSI for continuous covariates

Consider a data set $\{(Y_i, X_{ij}), i = 1, \cdots, n, j = 1, \cdots, p\}$ containing n subjects and p covariates. Let $\mathbb{Y} = (Y_1, \cdots, Y_n)^\top \in \mathbb{R}^n$ be the response vector. We assume it to be continuous and fully observed throughout the entire article, while the discussions for partially observed $\mathbb{Y}$ are given in Section 4. Denote the i-th subject of the p covariates $X_i = (X_{i1}, \cdots, X_{ip})^\top \in \mathbb{R}^p$ and the j-th covariate for the n subjects $\mathbb{X}_j = (X_{1j}, \cdots, X_{nj})^\top \in \mathbb{R}^n$ as defined in Section 1. We also assume that all X_is are continuous and some of the X_{ij}s are missing. Moreover, define D_{ij} = 0 if covariate X_{ij} is missing, and D_{ij} = 1 otherwise.

We next adapt semi-supervised learning to fill in a larger number of missing values. Specifically, for every covariate $\mathbb{X}_j$ ($j = 1, \cdots, p$), we treat its missing and observed subjects as unlabeled and labeled target outputs, respectively, while we regard the corresponding responses in $\mathbb{Y}$ as unlabeled and labeled inputs. We then model the j-th covariate of the i-th subject, X_{ij}, as a weighted average of the remainder of the subjects in $\mathbb{X}_j$, where the weights are determined by similarity measures (or distances) between the i-th and any other subjects (see, e.g., Zhu et al., 2003; Belkin et al., 2006).

In general, if the two subjects' corresponding responses and covariates are similar, their distance should be small. Hence, for any two different subjects i_1 and i_2, we construct the weight matrix $A = (a_{i_1 i_2}) \in \mathbb{R}^{n \times n}$ given below (see, e.g., Zou et al., 2017).

$$a_{i_1 i_2} = K_{h_1}(Y_{i_1} - Y_{i_2}) \prod_{k \in D_{i_1} \cap D_{i_2}} K_{h_2}(X_{i_1 k} - X_{i_2 k}), \qquad (2.1)$$

where $K_{h_l}(\cdot) = K(\cdot / h_l)$, $K(\cdot)$ is a kernel function, h_l for l = 1 and 2 are bandwidth parameters, and $D_i = \{j : D_{ij} = 1\}$ contains the observed covariates at subject i.

When $\mathbb{X}_k$ is discrete, we define $K_{h_2}(X_{i_1k} - X_{i_2k}) = K_{h_2}(0)$ if X_{i_1k} and X_{i_2k} belong to the same category and $K_{h_2}(X_{i_1k} - X_{i_2k}) = K_{h_2}(1)$ otherwise (see, e.g., Rojagopalan and Lall, 1995). In this paper, we consider the popular kernel function used in machine learning, the radial basis function kernel (i.e., Gaussian kernel). Accordingly, the resulting weight is

$$a_{i_1i_2} = \exp\Big\{-\lambda_1(Y_{i_1} - Y_{i_2})^2 - \lambda_2 \sum_{k \in D_{i_1} \cap D_{i_2}} (X_{i_1k} - X_{i_2k})^2\Big\},$$

where $\lambda = (\lambda_1, \lambda_2)^\top$, $\lambda_j = h_j^{-1}$ (j = 1, 2) are unknown scale parameters that need to be determined from the data.

Based on the weight matrix A, we then propose a weighted average approach in order to impute missing observations in covariates. For any $j = 1, \cdots, p$, define $\mathcal{S}_{0j} = \{i : D_{ij} = 0\}$ and $\mathcal{S}_{1j} = \{i : D_{ij} = 1\}$, thus they represent the missing and non-missing subjects in covariate j, respectively. Accordingly, $X_{\mathcal{S}_{0j}j} = (X_{ij} : i \in \mathcal{S}_{0j})$ and $X_{\mathcal{S}_{1j}j} = (X_{ij} : i \in \mathcal{S}_{1j})$ are the missing and non-missing subsets of $\mathbb{X}_j$. Adopting the concept of semi-supervised learning in Zhu et al. (2003) and Wasserman (2007), we consider that the missing observation X_{ij} for $i \in \mathcal{S}_{0j}$ is closer to those observations that are nearer to subject i as induced by their associated weights in A. This motivates us to impute missing observations $X_{\mathcal{S}_{0j}j}$ by minimizing the following quadratic loss function with respect to X_{ij}^* under the constraint $X_{ij}^* = X_{ij}$ for the observed subjects (see, e.g., Zhu et al., 2003; Zhang and Li, 2006; Belkin et al., 2006; Du and Zhao, 2017):

$$\min_{X_{ij}^*, i=1,\cdots,n} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} w_{i_1i_2}(X_{i_1j}^* - X_{i_2j}^*)^2, \text{ subject to } \mathrm{X_{ij}^*}=\mathrm{X_{ij}} \text{ for any i} \in \mathcal{S}_{1j}, \qquad (2.2)$$

where $w_{i_1i_2} = a_{i_1i_2} / \sum_{i_2} a_{i_1i_2}$ are the row-normalized versions of $a_{i_1i_2}$ and they constitute the weighted matrix $W = (w_{i_1i_2})$.

To solve (2.2), we set the first derivative of the objective function (2.2) with respect to $X^*_{i_1 j}$, for any $i_1 \in \mathcal{S}_{0j}$, to be 0. As a result, $\sum_{i_2=1}^{n} w_{i_1 i_2}(X^*_{i_1 j} - X^*_{i_2 j}) = 0$ for any $i_1 \in \mathcal{S}_{0j}$. Note that W is a row-normalized matrix and $\sum_{i_2=1}^{n} w_{i_1 i_2} = 1$. Accordingly, we obtain the following imputation algorithm for every subject $i_1 \in \mathcal{S}_{0j}$,

$$X^*_{i_1 j} = \sum_{i_2 \in \mathcal{S}_{0j}} w_{i_1 i_2} X^*_{i_2 j} + \sum_{i_2 \in \mathcal{S}_{1j}} w_{i_1 i_2} X_{i_2 j}. \qquad (2.3)$$

Define $C_i = \sum_{i^* \in \mathcal{S}_{1j}} w_{ii^*} X_{i^* j}$, $C_{\mathcal{S}_{0j}} = (C_i : i \in \mathcal{S}_{0j})$, and $W_{\mathcal{S}_{0j}} = (w_{i_1 i_2} : i_1 \in \mathcal{S}_{0j}, i_2 \in \mathcal{S}_{0j})$. Then Equation (2.3) can be expressed as

$$X^*_{\mathcal{S}_{0j} j} = W_{\mathcal{S}_{0j}} X^*_{\mathcal{S}_{0j} j} + C_{\mathcal{S}_{0j}}, \qquad (2.4)$$

which has the closed form solution $X^*_{\mathcal{S}_{0j} j} = (I - W_{\mathcal{S}_{0j}})^{-1} C_{\mathcal{S}_{0j}}$, and $I - W_{\mathcal{S}_{0j}}$ is invertible by Condition (C5) (see, Du and Zhao, 2017). Hence, for $j = 1, \cdots, p$, one can impute the missing subjects in $X_{\mathcal{S}_{0j} j}$ without imposing any model assumptions. Furthermore, define $W_{\mathcal{S}_{0j}\mathcal{S}_{1j}} = (w_{i_1 i_2}, i_1 \in \mathcal{S}_{0j}, i_2 \in \mathcal{S}_{1j})$. As a result, $C_{\mathcal{S}_{0j}} = W_{\mathcal{S}_{0j}\mathcal{S}_{1j}} X_{\mathcal{S}_{1j} j}$, and according to (2.4), the missing subjects of $X_{\mathcal{S}_{0j} j}$ can be imputed by

$$\hat{X}_{\mathcal{S}_{0j} j} = (I - W_{\mathcal{S}_{0j}})^{-1} W_{\mathcal{S}_{0j}\mathcal{S}_{1j}} X_{\mathcal{S}_{1j} j}, \qquad (2.5)$$

which is a weighted average of the observed subjects in $X_{\mathcal{S}_{1j} j}$. Since the above imputation procedure is motivated from semi-supervised learning, we denote our proposed approach SSI as mentioned in Section 1. It is worth noting that SSI can yield reliable results in the imputation of a large number of missing subjects since it takes into account all information from both responses and covariates that are related to missing subjects. Simulation results in Section 3 support this finding.

For continuous covariates, we investigate the average performance of $\hat{X}_{\mathcal{S}_{0j} j}$ given below.

Theorem 1. Assume that Conditions (C1)-(C6) hold in Appendix A of the supplementary material. Then, for any given $\delta > 0$, there exist finite positive constants C_1 and C_2 such that, for each $j = 1, \cdots, p$,

$$P\Big(\Big|\frac{1}{|\mathcal{S}_{0j}|}\sum_{i\in\mathcal{S}_{0j}}(\hat{X}_{ij} - X_{ij})\Big| > \frac{\delta}{n}\Big) \le 6\exp\Big(-\frac{1}{2}\frac{\delta^2}{C_1 n + C_2\delta}\Big).$$

The above theorem implies that $|\mathcal{S}_{0j}|^{-1}\sum_{i\in\mathcal{S}_{0j}}(\hat{X}_{ij} - X_{ij}) \to_p 0$ as $n \to \infty$ under Conditions (C1)-(C6). Thus, the sample average of the imputed values via SSI consistently estimates the true average of the missing values.

Remark 1: To construct the weight matrix, we employ the widely used Gaussian kernel to measure the distance between any two different subjects. However, this does not exclude other possible kernel functions such as polynomial kernels. In comparing with that kernel, there are two major reasons for using the Gaussian kernel. First, the Gaussian kernel is equivalent to mapping the data into an infinitely differentiable function space. As a result, it can partially capture the properties of polynomial kernels with high polynomial degrees. Second, the Gaussian kernel has fewer tuning parameters to be calculated, so it can be computationally efficient in large data sets. Our simulation studies and real data analyses indicate that the Gaussian kernel performs satisfactorily.

In contrast to continuous covariates, we next study discrete covariates with missing subjects.

2.2 SSI for discrete covariates

Suppose that $\mathbb{X}_j$ is discrete and has C class labels $\{1, \cdots, C\}$. We then define label matrix $\overline{\mathbb{X}}_j = (\bar{X}_{ic,j}) \in \mathbb{R}^{n\times C}$, where $\bar{X}_{ic,j}$ denotes the probability of the j-th covariate and subject i in class c. Note that if $X_{\mathcal{S}_{1j}j}$ is observed, then, for any $i \in \mathcal{S}_{1j}$, we have $\bar{X}_{ic,j} = 1$ if $X_{ij} = c$, and $\bar{X}_{ic,j} = 0$ otherwise. In addition, define $\bar{X}_{\mathcal{S}_{0j}j} = (\bar{X}_{ic,j}, i \in \mathcal{S}_{0j}, 1 \le c \le C)$ and $\bar{X}_{\mathcal{S}_{1j}j} = (\bar{X}_{ic,j}, i \in \mathcal{S}_{1j}, 1 \le c \le C)$. We then adopt

Equation (2.4) and propose the following algorithm, which enables us to impute missing subjects efficiently.

(1.) Set an initial value of $\bar{X}_{\mathcal{S}_{0j}j}$, and name it as $\bar{X}^{(0)}_{\mathcal{S}_{0j}j}$. Then update $\bar{X}_{\mathcal{S}_{0j}j}$ by $\bar{X}^{(1)}_{\mathcal{S}_{0j}j} = W_{\mathcal{S}_{0j}} \bar{X}^{(0)}_{\mathcal{S}_{0j}j} + \bar{C}_{\mathcal{S}_{0j}}$ with $\bar{C}_{\mathcal{S}_{0j}} = W_{\mathcal{S}_{0j}\mathcal{S}_{1j}} \bar{X}_{\mathcal{S}_{1j}j}$. Subsequently normalize the rows of $\bar{X}^{(1)}_{\mathcal{S}_{0j}j}$.

(2.) In the *k*-th step, for given $\bar{X}^{(k)}_{\mathcal{S}_{0j}j}$, update $\bar{X}_{\mathcal{S}_{0j}j}$ by $\bar{X}^{(k+1)}_{\mathcal{S}_{0j}j} = W_{\mathcal{S}_{0j}} \bar{X}^{(k)}_{\mathcal{S}_{0j}j} + \bar{C}_{\mathcal{S}_{0j}}$, and normalize the rows of $\bar{X}^{(k+1)}_{\mathcal{S}_{0j}j}$.

(3.) Stop at the *K*-th step if $\| \bar{X}^{(K)}_{\mathcal{S}_{0j}j} - \bar{X}^{(K-1)}_{\mathcal{S}_{0j}j} \| < \epsilon$ for some pre-specified value ϵ, where $\|\cdot\|$ denotes the Euclidean norm of any arbitrary vector. Afterwards, normalize the rows of $\bar{X}^{(K)}_{\mathcal{S}_{0j}j}$. Finally, for any $i \in \mathcal{S}_{0j}$, obtain $\hat{X}_{ij} = \text{argmax}_{1 \le c \le C} \bar{X}^{(K)}_{ic,j}$.

The above algorithm generates imputations for discrete covariates. Note that, for each iteration, the algorithm has the closed form. Thus, it is convergent regardless of the initial value under Condition (C5) (see the discussion of Zhu et al., 2003). We next investigate the theoretical property of this algorithm.

Theorem 2. Assume that Conditions (C1)-(C6) hold in Appendix A of the supplementary material. Then, for any given $\bar{\delta} > 0$, there exist finite positive constants $\bar{C}_1$ and $\bar{C}_2$ such that, for each $j = 1, \cdots, p$,

$$P\Big(\Big| \frac{1}{|\mathcal{S}_{0j}|} \sum_{i \in \mathcal{S}_{0j}} (\bar{X}^{(K)}_{ic,j} - \bar{X}_{ic,j}) \Big| > \frac{\bar{\delta}}{n} \Big) \le 6 \exp\Big(-\frac{1}{2} \frac{\bar{\delta}^2}{\bar{C}_1 n + \bar{C}_2 \bar{\delta}} \Big),$$

where *K* is defined in the step (3) of the above algorithm.

The proof of Theorem 2 is quite similar to that of Theorem 1; we thus omit it to save space. This theorem shows that, for each of the discrete covariate, the sample average of the probability of imputed values belonging to class *c* via SSI consistently estimates the true average probability of missing values.

Remark 2: In the missing data with mixing continuous and discrete covariates, one can construct the weight matrix by integrating the weights from both continuous covariates and discrete covariates via (2.1) and the formula under (2.1), respectively. Afterwards, one can employ (2.5) and the algorithm proposed in Section 2.2 to impute missing values for continuous and discrete covariates, separately.

2.3 Scale-parameters Selection, Estimation and Prediction

In semi-supervised learning, one can assume that the scale parameter in the weight matrix is known (see, e.g., Zhu and Goldberg, 2009). Based on prior knowledge and previous experience, practitioners can take the same approach to implement SSI by using the two known scale parameters. However, this approach is subjective, which motivates us to propose a data-driven and model-free procedure. According to Condition (C2), we recommend using $\lambda = O(n^{1/(2d_0+1)})$, where $d_0 = 1 + \max_{1\le i\le n-1} |D_i \cap D_{i+1}|$ defined in Appendix A of the supplementary material. Let $\tau = \lambda n^{-1/(2d_0+1)}$, and then employ the interchangeable imputation method given below to select the optimal τ.

For the j-th covariate with given τ, we apply SSI to impute $X_{S_{0j}j}$ and obtain $\hat{X}^{\tau}_{S_{0j}j}$. Based on the imputed values of $\hat{X}^{\tau}_{S_{0j}j}$, we interchangeably treat $X_{S_{1j}j}$ as missing and employ SSI again to yield $\hat{X}^{\tau}_{S_{1j}j}$. This procedure allows us to compute the imputation error of $\hat{X}^{\tau}_{S_{1j}j}$, which is $\| \hat{X}^{\tau}_{S_{1j}j} - X_{S_{1j}j} \|^2$. Repeat the same procedure for $j = 1, \cdots, p$, and obtain the overall imputation error, $\mathcal{Q}(\tau) = \sum_{j=1}^{p} \| \hat{X}^{\tau}_{S_{1j}j} - X_{S_{1j}j} \|^2$. Finally, we select τ as $\hat{\tau} = \operatorname{argmin}_{\tau} \mathcal{Q}(\tau)$ by minimizing the overall imputation error with a grid-search over the pre-specified bounded region. As a result, the selected $\hat{\lambda} = \hat{\tau} n^{1/(2d_0+1)}$ should be of order $O(n^{1/(2d_0+1)})$.

After selecting scale parameters, one can not only impute missing values, but also conduct subsequent data analysis such as estimation, prediction, and classification. Since the focus of our paper is prediction, we first assume that

there exists a relationship between the responses and the imputed-observed covariates via the linear regression model, $Y_i = \hat{X}_i^\top(\hat{\tau})\beta + \epsilon_i$, where $\hat{X}_i(\hat{\tau}) = (\hat{X}_{i1}(\hat{\tau}), \cdots, \hat{X}_{ip}(\hat{\tau}))^\top$, $\hat{X}_{ij}(\hat{\tau}) = X_{ij}$ if the i-th subject in j-th covariate is observed for $i = 1, \cdots, n$ and $j = 1, \cdots, p$, and $\beta = (\beta_1, \cdots, \beta_p)^\top$ is the $p \times 1$ vector of unknown regression coefficients. In addition, ϵs are random errors with mean 0 and variance σ^2. Then, the resulting ordinary least squares estimator of β is $\hat{\beta}(\hat{\tau}) = (\mathbb{X}^\top(\hat{\tau})\mathbb{X}(\hat{\tau}))^{-1}\mathbb{X}^\top(\hat{\tau})\mathbb{Y}$, where $\mathbb{X}(\hat{\tau}) = (\hat{X}_1(\hat{\tau}), \cdots, \hat{X}_n(\hat{\tau}))^\top$. Then, for any given $\hat{X}_0(\hat{\tau})$, we obtain the prediction $\hat{Y}_0 = \hat{X}_0(\hat{\tau})^\top\hat{\beta}(\hat{\tau})$. It is worth noting that the scale-parameters selection process via the interchangeable imputation method does not require the model assumptions between Y_i and X_i. Suppose the model is set up for achieving the specific purpose mentioned above. Then it is of interest to incorporate the model structure in the scale-parameters selection process. Accordingly, we propose employing a cross-validation approach to select the scale parameters. To this end, calculate $\hat{\beta}_{(i)}(\tau)$ after removing the i-th subject from the data and compute $\hat{Y}_i(\tau) = \hat{X}_i^\top(\tau)\hat{\beta}_{(i)}(\tau)$. Then, select $\tilde{\tau}$ by minimizing the squared errors of predictions across all subjects $\sum_{i=1}^{n}(\hat{Y}_i(\tau) - Y_i)^2$ with a grid-search over the pre-specified bounded region.

3 NUMERICAL STUDIES

3.1 Simulation Studies

To assess the finite sample performance of the proposed imputation method, we consider the following simulation studies. The data are generated from a linear regression model, $Y_i = X_i^\top\beta + \epsilon_i = X_{ia}^\top\beta_a + X_{ib}^\top\beta_b + \epsilon_i$, for $i = 1, \cdots, n$, where $\beta = (\beta_a^\top, \beta_b^\top)$. Specifically, the discrete covariates, $X_{ia} \in \mathbb{R}^2$, are iid simulated from a binomial distribution with probability 0.5, while the continuous covariate, $X_{ib} \in \mathbb{R}^{p-2}$ with p = 10, are iid generated from a multivariate normal distribution with mean 0 and covariance matrix $\Sigma_X = (\sigma_{x,j_1j_2})$, where $\sigma_{x,j_1j_2} = 1$ for j_1 = j_2, and $\sigma_{x,j_1j_2} = \rho$ otherwise, and $0 < \rho < 1$. In addition, $\beta_a \in \mathbb{R}^2$, $\beta_b \in \mathbb{R}^{p-2}$, and $\beta_j = 1$ for $1 \le j \le p$. The random errors ϵ_i are iid simulated from a normal distribution with

mean 0 and variance σ^2, where σ^2 is chosen to control the coefficient of determination $R^2 = \mathrm{var}(X_i^\top \beta) / \mathrm{var}(Y_i)$. We consider data that are block-wise missing. This is a modification of the data type in Lin et al. (2020) with finite missing patterns, which reflects our real data structure. As mentioned in Section 2.1, we employ the Gaussian kernel of $K(\cdot)$ to construct the weight matrix defined in (2.1) throughout this entire section.

In this simulation setting, we consider 7 missing patterns: $\mathcal{D}_1 = \{1,2,3,10\}$, $\mathcal{D}_2 = \{4,5,6,10\}$, $\mathcal{D}_3 = \{7,8,9,10\}$, $\mathcal{D}_4 = \{1,2,3,4,5,6,10\}$, $\mathcal{D}_5 = \{1,2,3,7,8,9,10\}$, $\mathcal{D}_6 = \{4,5,6,7,8,9,10\}$, $\mathcal{D}_7 = \{1,2,3,4,5,6,7,8,9,10\}$, where the notation of missing pattern $\mathcal{D}_k$ is defined in Appendix A for $1 \le k \le 7$. Accordingly, if the i-th sample has missing observations, its missing pattern should be one of the aforementioned seven patterns. For example, if the i-th sample is missing with pattern $\mathcal{D}_1$, then X_{ij} with $3 < j < 10$ are all missing in this sample. Furthermore, X_{ip} is observed so that it can be used to generate the "missing at random" data. Such missing patterns are similar to that of our real example. Then we consider the three commonly used missing mechanisms: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Under MCAR, the missing observations in each sample are randomly generated with equal probability. Under MAR, however, the probability of the missing observation in the i-th sample is related to the observed covariate X_{ip} and it is $\{1+\exp(-0.5X_{ip} - \alpha_1)\}^{-1}$, where α_1 is chosen to control the missing rate being approximately 50%. On the other hand, under MNAR, the probability of the missing observation given Y_i in the i-th sample is related to the random error ϵ_i, which is $\{1+\exp(-\epsilon_i - \alpha_2)\}^{-1}$ and α_2 is selected to control the missing rate at approximately 50%. Lastly, the missing patterns $\mathcal{D}_j$ ($j = 1, \cdots, 7$) are randomly assigned to those missing samples. To assess the prediction accuracy, we further randomly split the entire sample into two parts, i.e., training sample (70%) and testing sample (30%).

To assess the performance of SSI, we conduct 1,000 realizations. For each realization $k = 1, \cdots, 1,000$, let $\mathbb{X}^{(k)}$, $\hat{\beta}^{(k)}$, and $\hat{Y}_{\text{test}}^{(k)}$ be its corresponding imputed

covariates, estimate of the parameter vector, and predictions on testing samples, respectively. In addition, let $\mathbb{X}$, β, and Y_{test} be the true covariates, true parameter vector, and true responses from testing samples, respectively. We next employ $\text{IA} = 1000^{-1}\sum_{k}||\mathbb{X}^{(k)} - \mathbb{X}||$, $\text{EA} = 1000^{-1}\sum_{k}||\hat{\beta}^{(k)} - \beta||$, and $\text{PA} = 1000^{-1}\sum_{k}||\hat{Y}_{\text{test}}^{(k)} - Y_{\text{test}}||$ to correspondingly assess the accuracy of imputations, regression estimates, and predictions. For the sake of comparison, we also include the commonly used methods, KNN, RF, MI, RI and FIME in our study. Here, KNN is implemented with the R package "caret", RF is implemented with the R package "missForest", RI and MI are implemented with the R package "mice", and FIME is implemented with the R package "lavvan". It is worth noting that KNN is not computable in some realizations due to high missing rate data. Hence, we only report computable results for KNN. In addition, "mice" is iterated for 50 times by burning-in the first 49 iterations to assure its stability. The rest of other methods are all implemented by using their corresponding default settings in R packages. Moreover, we use the interchangeable method and the cross-validation procedure mentioned in Section 2.3 to calculate the scale parameters in SSI by using a simple grid searching method with r ranging from 0 to 2, and we denote them SSI_1 and SSI_2, respectively. It is worth noting that SSI is a weighted average of the observed values and their associated weights are evaluated by using both observed and unobserved information. Hence, SSI is more informative than that of the mean imputation method, which is not included in simulation studies.

In the above setting, we use five components, R^2, ρ, n, missing mechanism, and accuracy measure, to study the performance of six imputation methods. As a result, there are various combinations via these five components. To save space, we only present two tables and show the averaged performance. Table 3 presents the results of IA, EA and PA under the three coefficients of determination ($R^2 = 0.3$, 0.6 and 0.9) and three missing mechanisms (MCAR, MAR, and MNAR), and they are obtained by averaging over the three

correlations ($\rho = 0.25$, 0.5 and 0.75) and three sample sizes (n = 500, 1,000 and 2,000). Note that FIME is only applicable for parameter estimation; hence, we display its IA and PA as "-". In imputation, SSI_1 performs the best, and SSI_1 is generally superior to SSI_2 in estimation when $R_2 = 0.3$ and $R_2 = 0.6$. In addition, both of them outperform the other four methods in both imputation and estimation. As for prediction, SSI_2 performs the best, and it is superior to SSI_1. This finding is not surprising since SSI_1 is mainly designed for imputation, while SSI_2 focuses on prediction. In sum, SSI_1 performs the best in imputation and well in estimation, while SSI_2 outperforms other methods in prediction. Finally, both SSI_1 and SSI_2 improve as R^2 increases. This finding is sensible since a stronger signal can lead to better imputation, estimation and prediction.

Next, Table 4 reports the results of three evaluation measures (IA, EA and PA) under the three correlation structures ($\rho = 0.25$, 0.5 and 0.75) and three missing mechanisms (MCAR, MAR and MNAR), and they are obtained by averaging over the three coefficients of determination ($R^2 = 0.3$, 0.6 and 0.9) and three sample sizes (n = 500, 1,000 and 2,000). Analogous to the results in Table 3, both SSI_1 and SSI_2 outperform the other methods in imputation and estimation. Although SSI_1 is slightly better than SSI_2 in imputation, SSI_2 performs the best in prediction. Note that both SSI_1 and SSI_2 improve in imputation and prediction as ρ becomes larger, but their performance gets worse in estimation. This finding indicates that multicollinearity can affect estimation rather than imputation and prediction, which is expected. Based on Tables 3 and 4, we conclude that SSI_1 and SSI_2 are generally superior to the other four methods in imputation, estimation and prediction. It is also worth noting that SSI_1 and SSI_2 have low variability in the imputation bias. This may be due to SSI using the information from observed responses and covariates as well as the information from covariates with missing values. In sum, SSI not only reduces the bias of imputation, estimation and prediction, but also yields lower variability in the imputation bias.

Finally, Figure 1 depicts the results of IA and PA for the three sample sizes (n = 500, 1,000 and 2,000) in Panels A and B, respectively, under MNAR. The IA and PA are obtained from SSI_1 and SSI_2 by averaging over the three correlations ($\rho = 0.25$, 0.5 and 0.75) and three coefficients of determination ($R^2 = 0.3$, 0.6 and 0.9). The results show that SSI_1 performs slightly better (weaker) than SSI_2 in imputation (prediction) as demonstrated in Tables 3 and 4. Additionally, they are comparable in estimation, but the EA plot is not presented here. These results also indicate that the performance of SSI_1 and SSI_2 improves as the sample size increases, which supports our theoretical findings. However, this improvement is limited. Our results show that SSI is steady even with n = 500. Under MCAR and MAR, we obtain similar results, which are omitted.

Upon the anonymous reviewers' suggestions, we assess the robustness of the proposed method against the non-normality of covariates, the selection of tuning parameters, and different missing mechanisms. To this end, we consider four additional simulation settings: (I) the covariate Xs being generated from a multivariate exponential distribution; (II) the selection of the tuning parameter τ; (III) the "non-ignorable missing data" setting modified from Kim and Yu (2011); (IV) the missing data with no pattern. The detailed simulation settings and results are given in Appendix C of the supplementary material. The results show similar patterns to those in Tables 3 and 4, which demonstrate the robustness of our method.

3.2 Real Data Analysis

To illustrate the usefulness of SSI, we consider a real example for risk management in online consumer finance, which was mentioned in the Introduction. The online consumer finance industry in China is growing rapidly in recent years and its market size is larger than 10 trillion. Unlike the traditional personal loan application process at commercial banks, verification of repayment ability is not feasible for online lenders since it would time consuming for applicants to prepare and submit documents, and also technically complicated

for the credit officers to validate them in a short time. Hence, how to measure customers' repayment ability via their personal attributes becomes critical.

Table 1 reports the data set collected from the five different resources with six missing patterns as well as the sample size and missing rate of each individual missing pattern. There is a total of $n = 2{,}390$ applicants and thirteen covariates, while the response variable is the log-income. There are eleven continuous covariates and two discrete covariates. One of the discrete covariates has four class labels and the other has three class labels. Detailed descriptions of the thirteen covariates are presented in Table 2. Accordingly, this is a typical block-wise missing data with high missing rates; the total missing rate is 49.4% and only 4.7% of applicants have complete information. To assess the accuracy of predictions, we randomly split the entire data set into a training sample (70%) and testing samples (30%). This procedure is repeated 100 times. In the m-th splitting, we calculate the mean of the prediction error in the testing sample based on the model estimated in the training sample data, and we denote it $\mu^{(m)}$. Then, we compute the mean and standard error of $\mu^{(m)}$ as $\mathrm{MEAN} = 100^{-1}\sum_{m=1}^{100}\mu^{(m)}$ and $\mathrm{SD} = \left\{100^{-1}\sum_{m=1}^{100}(\mu^{(m)} - \mathrm{MEAN})^2\right\}^{1/2}$. The scale-parameter vector λ of SSI is estimated via the interchangeable algorithm and the cross-validation methods, respectively, from the training sample. As in simulation studies, we denote them SSI_1 and SSI_2.

In addition to SSI_1 and SSI_2, we also consider the other four methods, KNN, RF, MI and RI. Table 5 reports the MEAN and SD of all five methods. It shows that SSI_2 has the smallest MEAN and SD as found in the simulation studies, while it is only slightly better than SSI_1. For the sake of comparison, we use SSI_2 as a baseline to calculate the percentage of reduction (PR) in MEAN and SD by comparing the other four methods to it. Table 5 demonstrates that SSI_2 yields more than 8% reductions in MEAN and SD, respectively, over each of other methods except for SSI_1. Based on the above findings, we conclude that SSI

produces reliable predictions. Accordingly, financial institutions can multiply the predicted income by a pre-specified constant to determine the optimal loan credit of each applicant. This empirical example suggests that SSI could play an important role in the online consumer finance industry.

4 CONCLUDING REMARKS

This paper proposes a semi-supervised imputation (SSI) method for data that have fully observed responses and covariates with high missing rates. The SSI is a model free method, which does not require any model assumptions. In addition, SSI is a computationally efficient method because it has a closed form when covariates are continuous and it can be calculated iteratively for discrete covariates. Moreover, we show the missing-averaged consistency of SSI, and we introduce an interchange algorithm to estimate the scale parameters of SSI. After filling in missing values, we establish the relationship between the responses and covariates via a linear regression model and introduce a cross-validation method to estimate scale parameters of SSI. Both simulation studies and a real application in online consumer finance demonstrate that SSI performs well.

It is worth noting that SSI is applicable when the response $\mathbb{Y}$ is partially observed. In this case, the weight matrix $\bar{A} = (\bar{a}_{i_1 i_2}) \in \mathbb{R}^{n \times n}$ can be defined as $\bar{a}_{i_1 i_2} = K_{h_1}(Y_{i_1} - Y_{i_2}) \prod_{k \in D_{i_1} \cap D_{i_2}} K_{h_2}(X_{i_1 k} - X_{i_2 k})$ if Y_{i_1} and Y_{i_2} are both observed, otherwise $\bar{a}_{i_1 i_2} = \prod_{k \in D_{i_1} \cap D_{i_2}} K_{h_2}(X_{i_1 k} - X_{i_2 k})$ if either Y_{i_1} or Y_{i_2} is missing. To impute missing observations in any given covariate, SSI does not utilize other covariates' information. Hence, we can follow an anonymous reviewer's suggestion to adapt the MI approach and perform the SSI imputation sequentially. Specifically, we first impute $\mathbb{X}_1$, and then impute $\mathbb{X}_2$ by taking into account the imputed $\mathbb{X}_1$. Repeat this procedure until $\mathbb{X}_p$ has been imputed. Subsequently, we restart a new iterative procedure, and begin to impute $\mathbb{X}_1$ by taking into account the imputed $\mathbb{X}_2, \cdots, \mathbb{X}_p$ obtained from the previous iterative process. Then impute $\mathbb{X}_2$

by taking into account the imputed $\mathbb{X}_1, \mathbb{X}_3, \cdots, \mathbb{X}_p$. Repeat this new iterative procedure *m* times, where *m* is a pre-specified number of iterations. We name this procedure Sequentially Semi-Supervised Imputation (SSSI), and our simulation results indicate that SSSI is slightly better than SSI; see Section D in the supplementary material.

To conclude this article, we identify five possible research avenues for future study. First, establish a theoretical framework to investigate the properties of SSSI. Second, extend SSI to discrete responses so that it can be used for classification. Third, study the efficiency and theoretical properties of imputation, estimation and prediction. Fourth, select the optimal kernel function and its parameters automatically for the weight matrix. Lastly, incorporate the relationship between the covariates (including the response variable) into the kernel function to construct weight matrix and improve efficiency. We believe these efforts can broaden the usefulness of our proposed SSI method.

REFERENCES

Allison, P. D. (2012). "Handling missing data by maximum likelihood," *SAS Global Forum Proceedings*, 1–21.

Belkin, M., Niyogi, P. and Sindhwani, V. (2006). "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *Journal of Machine Learning Research*, 7, 2399–2434.

Bollinger, C. and Hirsch, B. (2013). "Is earnings nonresponse ignorable," *The Review of Economics and Statistics*, 95, 407–416.

Chen, S. and Haziza, D. (2017). "Multiply robust imputation procedures for the treatment of item nonresponse in surveys," *Biometrika*, 104, 439–453.

Chen, X., Wan, A. and Zhou, Y. (2015). "Efficient quantile regression analysis with missing observations," *Journal of the American Statistical Association,*, 110, 723–741.

Chapelle, O., Scholkopf, B. and Zien, A. (2010). *Semi-Supervised Learning,* Cambridge: The MIT Press.

Collins, L. M., Schafer, J. L. and Kam, C. M. (2001). “A comparison of inclusive and restrictive strategies in modern missing data procedures,” *Psychological Methods,* 23, 330–351.

Du, C. and Zhao, Y. (2017). “On consistency of graph-based semi-supervised learning,” *arXiv: Machine Learning*.

Fan, J., Gijbels, I. and King, M. (1997). “Local likelihood and local partial likelihood in hazard regression,” *Annals of Statistics*, 25, 1661–1690.

Fang, F., Lan, W., Tong, J. and Shao, J. (2019). “Model averaging for prediction with fragmentary data,” *Journal of Business & Economic Statistics*, 37, 517–527.

Han, P. and Wang, L. (2013). “Estimation with missing data: beyond double robustness,” *Biometrika*, 100, 417–430.

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

Lafferty, J. and Wasserman, L. (2007). “Statistical analysis of semi-supervised regression,” *Advances in Neural Information Processing Systems*, 801–808.

Little, R. J. A. (1992). “Regression with missing *X*: A review,” *Journal of the American Statistical Association*, 87, 1227–1237.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical analysis with missing data (Second Edition),* New York: Wiley.

Lin, H., Liu, W. and Lan, W. (2020). “Regression analysis with individual-specific patterns of missing covariates,” *Journal of Business & Economic Statistics*, In Press.

Lipsitz, S. R., Ibrahim, J. G. and Zhao, L. P. (1999). “A weighted estimating equation for missing covariate data with properties similar to maximum likelihood,” *Journal of the American Statistical Association*, 94, 1147–1160.

Myrtveit, I., Stensrud, E. and Olsson, U. H. (2001). “Analyzing data sets with missing data: an empirical evaluation of imputation methods and likelihood-based methods,” *IEEE Transactions on Software Engineering*, 27, 999–1013.

Nulty, D. D. (2008). “The adequacy of response rates to online and paper surveys: what can be done?” *Assessment & Evaluation in Higher Education*, 33, 301–314.

Rajagopalan, B. and Lall, U. (1995). “A kernel estimator for discrete distributions,” *Journal of Nonparametric Statistics*, 4, 409–426.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). “Estimation of regression coefficients when some regressors are not always observed,” *Journal of the American Statistical Association*, 89, 846–866.

Shao, J. and Wang, H. (2002). “Sample correlation coefficients based on survey data under regression imputation,” *Journal of American Statistical Association*, 97, 544-552.

Stekhoven, D. and Bühlmann, P. (2012). “MissForest non-parametric missing value imputation for mixed-type data,” *Bioinformatics*, 28, 112–118.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, R. (2001). “Missing value estimation methods for DNA microarrays,” *Bioinformatics*, 17, 520–525.

Xiang, S., Yuan, L., Fan, W., Wang, Y., Thompson, P. M. and Ye, J. (2014). “Bi-level multi- source learning for heterogeneous block-wise missing data,” *Neuroimage*, 102, 192–206.

Yuan, L., Wang, Y., Thompson, P. M., Narayan, V. A. and Ye, J. (2012). “ Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data,” *Neuroimage*, 61, 622–632.

Zhang, X. and Li, W. S. (2006). “Hyperparameter learning for graph based semi-supervised learning algorithms,” *Advances in Neural Information Processing Systems*, 19, 1595–1592.

Zhang, Z. (2016). “Missing data imputation: focusing on single imputation,” *Annals of Translational Medicine*, 4, 1–8.

Zhou, Y., Litter, R. J. A. and Kalbeisch, J. D. (2010). “Block-conditional missing at random models for missing data,” *Statistical Science*, 25, 517–532.

Zhu, X., Ghahramani, Z. and Lafferty, J. (2003). “Semi-supervised learning using Gaussian fields and Harmonic functions,” *International Conference on Machine Learning*, 118, 912–919.

Zhu, X. and Goldberg, A. B. (2009). *Introduction to Semi-Supervised Learning,* San Rafael: Morgan & Claypool.

Zou, T., Lan, W., Wang, H. and Tsai., C. L. (2017). “Covariance regression analysis,” *Journal of the American Statistical Association*, 112, 266–281.

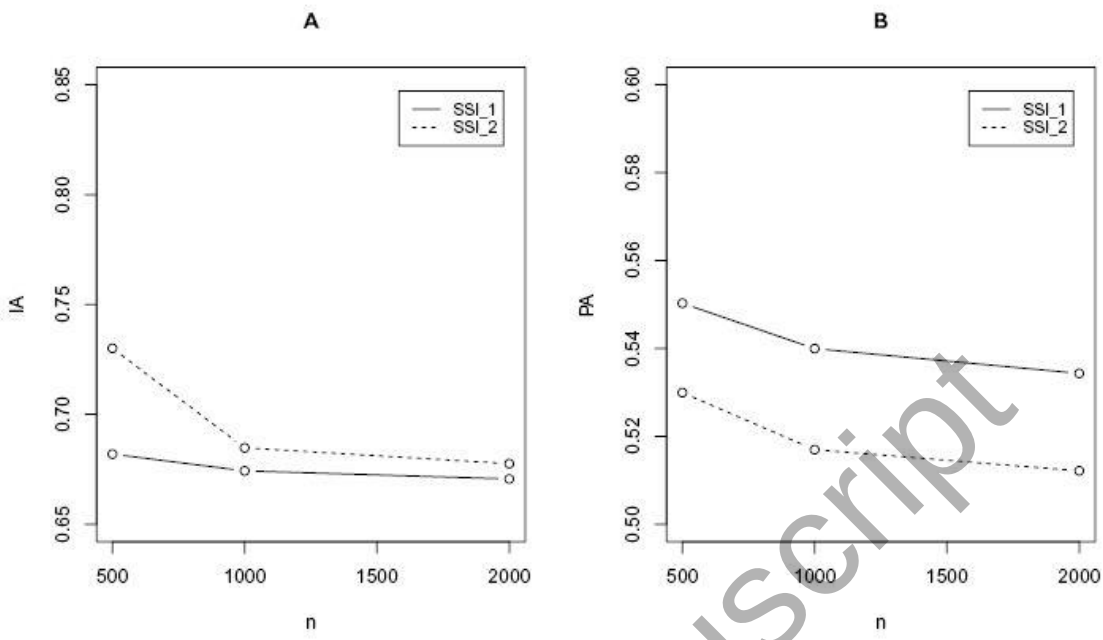

Fig. 1 The plots of IA (panel A) and PA (panel B) versus the sample sizes n = 500, 1,000, and 2,000 under MNAR. The IA and PA are obtained from SSI_1 and SSI_2 by averaging over three correlations ($\rho = 0.25$, 0.5 and 0.75) and three coefficients of determination ($R^2 = 0.3$, 0.6 and 0.9).

Table 1 Sample sizes and missing rates of the six missing patterns for the empirical data, where “1” represents the data being available in the corresponding source, and “0” otherwise.

Pattern	Credit card	e-shopping record in Taobao	Cell phone usage record	Credit bureau information	Fraud record	Sample size	Missing rate (%)
1	1	1	1	1	1	113	0.0
2	1	1	1	1	0	29	15.4
3	1	1	0	1	1	231	15.4
4	1	1	1	0	0	220	30.8
5	1	1	0	0	0	1,161	46.2
6	0	1	0	0	0	636	84.6
					Total	2,390	49.4

Table 2 Detailed descriptions of the 13 covariates including data type, data source and variable definition. Note that the variables “credit-use-number” and “ credit-use-amount” are discrete, and they have four and three class labels, respectively.

Covariate	Data Type	Data Source	Definition
credit-min	Continuous	Credit card	the minimum credit limit among all credit cards
credit-max	Continuous	Credit card	the maximum credit limit among all credit cards
credit-union	Continuous	Credit card	the total amount of credit limit of credit cards
credit-use-number	Discrete (four class labels)	Credit card	average transaction number of credit cards
credit-use-amount	Discrete (three class labels)	Credit card	average transaction amount of credit cards
shopping-record-amount	Continuous	E-shopping record	average transaction amount of e-shopping
shopping-record-number	Continuous	E-shopping record	average transaction number of e-shopping
phone-use-plansum	Continuous	Cell-phone usage	total mobile package fee over the last 6 months
phone-use-totalsum	Continuous	Cell-phone usage	total mobile cost over the last 6 months
payment-amount	Continuous	Credit bureau	average payment for commercial banks
card-number	Continuous	Credit bureau	number of active bank cards

Covariate	Data Type	Data Source	Definition
approval-number	Continuous	Fraud record	approval number for non-commercial banks
fraud-score	Continuous	Fraud record	fraud score evaluated by credit company

Table 3 The simulation results of IA, EA and PA under the three coefficients of determination (R^2=0.3, 0.6 and 0.9) and three missing mechanisms (MCAR, MAR, and MNAR), and they are obtained by averaging over three correlations ($\rho = 0.25$, 0.5 and 0.75) and three sample sizes (n = 500, 1,000 and 2,000). The value in parentheses is the averaged standard error. Note that FIME is only applicable for parameter estimation, hence, we display its IA and PA as "-".

		MCAR			MAR			MNAR		
R^2	Methods	IA	EA	PA	IA	EA	PA	IA	EA	PA
0.3	KNN	0.79(0.197)	0.76(0.061)	0.83(0.098)	1.02(0.216)	0.74(0.062)	0.97(0.124)	0.79(0.219)	0.76(0.064)	0.97(0.094)
	RF	0.74(0.228)	0.78(0.027)	0.78(0.081)	0.75(0.234)	0.78(0.027)	0.79(0.085)	0.74(0.233)	0.78(0.027)	0.78(0.083)
	MI	1.15(0.347)	0.79(0.024)	0.80(0.085)	1.15(0.347)	0.79(0.025)	0.80(0.088)	1.15(0.351)	0.79(0.023)	0.79(0.088)
	RI	0.86(0.262)	0.81(0.141)	0.97(0.399)	0.86(0.237)	1.28(0.959)	2.12(1.962)	0.82(0.266)	0.94(0.498)	1.39(1.347)
	FIME	-	0.76(0.088)	-	-	0.76(0.048)	-	-	0.76(0.042)	-
	SSI_1	0.68(0.157)	0.70(0.069)	0.83(0.083)	0.68(0.158)	0.70(0.075)	0.83(0.087)	0.68(0.157)	0.70(0.065)	0.81(0.087)
	SSI_2	0.70(0.149)	0.72(0.046)	0.77(0.079)	0.70(0.149)	0.72(0.051)	0.78(0.083)	0.73(0.148)	0.72(0.054)	0.77(0.081)
0.6	KNN	0.79(0.192)	0.70(0.077)	0.73(0.096)	0.86(0.197)	0.70(0.069)	0.73(0.094)	0.79(0.213)	0.71(0.068)	0.80(0.091)
	RF	0.74(0.229)	0.73(0.027)	0.54(0.088)	0.75(0.234)	0.73(0.028)	0.54(0.090)	0.74(0.235)	0.73(0.028)	0.54(0.093)
	MI	1.15(0.351)	0.74(0.022)	0.57(0.094)	1.15(0.348)	0.74(0.022)	0.57(0.092)	1.15(0.349)	0.74(0.021)	0.57(0.096)
	RI	0.86(0.260)	0.73(0.093)	0.66(0.252)	0.88(0.255)	0.81(0.341)	0.89(0.894)	0.80(0.234)	1.13(0.723)	1.79(2.039)

		MCAR			MAR			MNAR		
	FIME	-	0.69(0.045)	-	-	0.69(0.048)	-	-	0.69(0.047)	-
	SSI_1	0.68(0.158)	0.65(0.053)	0.54(0.072)	0.68(0.158)	0.65(0.055)	0.54(0.072)	0.68(0.159)	0.65(0.055)	0.54(0.074)
	SSI_2	0.68(0.157)	0.66(0.053)	0.52(0.075)	0.68(0.157)	0.66(0.051)	0.52(0.072)	0.69(0.161)	0.65(0.058)	0.52(0.076)
0.9	KNN	0.80(0.192)	0.66(0.062)	0.48(0.125)	0.81(0.197)	0.76(0.072)	0.48(0.119)	-	-	-
	RF	0.74(0.229)	0.69(0.028)	0.29(0.107)	0.75(0.234)	0.69(0.027)	0.31(0.111)	0.75(0.234)	0.69(0.027)	0.31(0.121)
	MI	1.15(0.349)	0.70(0.021)	0.34(0.121)	1.15(0.352)	0.70(0.021)	0.34(0.119)	1.15(0.345)	0.71(0.061)	0.35(0.122)
	RI	0.86(0.261)	0.67(0.049)	0.36(0.168)	0.85(0.264)	0.66(0.045)	0.34(0.142)	0.75(0.222)	1.92(1.709)	2.63(2.486)
	FIME	-	0.64(0.044)	-	-	0.64(0.043)	-	-	0.64(0.045)	-
	SSI_1	0.67(0.162)	0.63(0.042)	0.27(0.083)	0.67(0.162)	0.63(0.041)	0.27(0.083)	0.67(0.161)	0.63(0.042)	0.28(0.087)
	SSI_2	0.68(0.155)	0.62(0.053)	0.26(0.079)	0.68(0.161)	0.62(0.053)	0.27(0.083)	0.68(0.155)	0.61(0.057)	0.27(0.083)

Table 4 The simulation results of IA, EA and PA under the three correlations ($\rho = 0.25$, 0.5 and 0.75) and three missing mechanisms (MCAR, MAR, and MNAR), and they are obtained by averaging over the three coefficients of determination (R^2=0.3, 0.6 and 0.9) and three sample sizes (n = 500, 1,000 and 2,000). The value in parentheses is the averaged standard error. Note that FIME is only applicable for parameter estimation; hence, we display its IA and PA as "-".

		MCAR			MAR			MNAR		
ρ	Methods	IA	EA	PA	IA	EA	PA	IA	EA	PA
0.25	KNN	1.00(0.066)	0.67(0.075)	0.79(0.223)	1.03(0.089)	0.84(0.072)	0.74(0.239)	1.04(0.110)	0.80(0.074)	0.85(0.208)
	RF	1.01(0.056)	0.71(0.048)	0.62(0.178)	1.02(0.060)	0.71(0.051)	0.64(0.178)	1.02(0.059)	0.71(0.048)	0.63(0.175)
	MI	1.54(0.095)	0.73(0.042)	0.67(0.163)	1.54(0.094)	0.73(0.045)	0.67(0.165)	1.54(0.094)	0.73(0.043)	0.67(0.161)
	RI	1.15(0.101)	0.70(0.104)	0.75(0.372)	1.14(0.103)	0.77(0.318)	1.06(1.265)	1.13(0.098)	1.28(0.829)	2.73(2.219)
	FIME	-	0.65(0.078)	-	-	0.65(0.061)	-	-	0.65(0.067)	-
	SSI_1	0.85(0.038)	0.60(0.038)	0.61(0.208)	0.85(0.037)	0.60(0.036)	0.61(0.211)	0.85(0.038)	0.60(0.039)	0.61(0.203)
	SSI_2	0.85(0.038)	0.61(0.058)	0.58(0.192)	0.86(0.040)	0.61(0.056)	0.59(0.194)	0.86(0.039)	0.60(0.061)	0.59(0.192)
0.5	KNN	0.81(0.072)	0.76(0.065)	0.72(0.411)	0.88(0.142)	0.89(0.054)	0.75(0.285)	0.86(0.111)	0.73(0.061)	0.86(0.298)
	RF	0.75(0.046)	0.73(0.043)	0.54(0.211)	0.76(0.046)	0.74(0.043)	0.53(0.211)	0.76(0.044)	0.74(0.041)	0.54(0.208)
	MI	1.19(0.082)	0.75(0.037)	0.56(0.199)	1.19(0.089)	0.75(0.039)	0.57(0.201)	1.19(0.087)	0.75(0.036)	0.57(0.195)
	RI	0.88(0.095)	0.74(0.099)	0.65(0.369)	0.89(0.098)	0.86(0.486)	1.02(1.448)	0.88(0.085)	1.01(0.587)	1.53(1.661)

		MCAR			MAR			MNAR		
	FIME	-	0.70(0.053)	-	-	0.70(0.057)	-	-	0.71(0.055)	-
	SSI_1	0.71(0.035)	0.66(0.039)	0.54(0.234)	0.71(0.035)	0.66(0.038)	0.54(0.238)	0.71(0.035)	0.66(0.039)	0.54(0.227)
	SSI_2	0.72(0.037)	0.67(0.047)	0.51(0.216)	0.71(0.035)	0.67(0.050)	0.51(0.217)	0.72(0.042)	0.67(0.05)	0.51(0.215)
0.75	KNN	0.57(0.076)	0.84(0.059)	0.66(0.430)	0.68(0.217)	0.92(0.069)	0.59(0.301)	0.58(0.123)	0.86(0.062)	0.68(0.285)
	RF	0.49(0.028)	0.75(0.042)	0.46(0.235)	0.49(0.029)	0.75(0.039)	0.48(0.235)	0.50(0.027)	0.75(0.039)	0.49(0.233)
	MI	0.71(0.061)	0.76(0.036)	0.48(0.227)	0.72(0.063)	0.76(0.038)	0.48(0.226)	0.72(0.061)	0.76(0.035)	0.48(0.219)
	RI	0.55(0.076)	0.78(0.135)	0.58(0.387)	0.57(0.087)	0.98(0.808)	0.91(1.27)	0.55(0.083)	1.08(0.948)	1.16(1.353)
	FIME	-	0.74(0.079)	-	-	0.74(0.051)	-	-	0.74(0.048)	-
	SSI_1	0.47(0.024)	0.71(0.040)	0.49(0.260)	0.47(0.027)	0.72(0.051)	0.48(0.259)	0.47(0.025)	0.71(0.044)	0.48(0.245)
	SSI_2	0.49(0.036)	0.72(0.041)	0.46(0.235)	0.49(0.038)	0.72(0.042)	0.46(0.237)	0.51(0.101)	0.72(0.047)	0.46(0.234)

Table 5 The mean (MEAN) and standard error (SD) of the prediction errors obtained from KNN, MI, RI, MissForest, SSI_1 and SSI_2, and the percentage reduction (PR) in Mean and SD that is achieved by SSI_2 in comparison to the given method.

	KNN	MI	RI	RF	SSI_1	SSI_2
Mean	0.311	0.474	0.360	0.306	0.282	0.280
PR (%)	10.21	40.93	22.22	8.02	0.72	0
SD	0.053	0.077	0.070	0.052	0.049	0.048
PR (%)	9.43	37.66	31.42	8.42	2.08	0