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

Wei Lan¹, Xuerong Chenª, Tao Zoub, and Chih-Ling Tsaić

¹School of Statistics and Center of Statistical Research, Southwest University of Finance and Economics, Chengdu, China; ²Research School of Finance, Actuarial Studies & Statistics, The Australian National University, Canberra, Australia; ³Graduate School of Management, University of California, Davis, CA

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 to be computed. Motivated by the concept of semi-supervised learning, 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 nonmissing subjects in any covariate as the unlabeled and labeled target outputs, respectively, and treat their corresponding responses as the unlabeled and labeled 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.

1. Introduction

Missing data are 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 Litter 1992), full-information maximum likelihood estimation (FIME; Collins, Schafer, and Kam 2001; Myrervist, Stensrud, and Olsson 2001; Allison 2012), weighted estimation equation (WEE; Robins, Rotnitzky, A. and Zhao 1994; Lipsitz, Ibrahim, and Zhao 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; 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 article 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% applicant 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, Litter, and Kalbeisch 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, Scholkopf, and Zien 2010). Specifically, we consider the missing and non-missing subjects in any covariate as the unlabeled and labeled target outputs, respectively. Then we treat their corresponding responses as the unlabeled and labeled 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 SSL, we simply name it semi-supervised imputation (SSI).

The SSI method uses information from both observed and unobserved subjects across responses and covariates. Suppose that the $i$th subject is missing at the $j$th covariate $X_j = (X_{ij1}, \ldots, X_{ijn})^\top \in \mathbb{R}^n$ (i.e., $X_i$). We then formulate it as a function of the weighted average of the rest of the subjects in $X_j$. The weight between $X_j$ and any other subject $X_{ij}$ ($l \neq i$) is determined by their similarity (or distance) measure, which is constructed by using the $X_{ij}$’s and $X_j$’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 $X_j$.

Formulating the imputation task through the concept of SSL has three important merits. First, SSL has the capability of handling large amounts of unlabeled data with promising results (see, e.g., Zhu, Ghahramani, and Lafferty 2003). Accordingly, SSI enables us to manage high missing rate data without imposing any model structure assumptions. Since our proposed method performs well. The article concludes with some discussion in Section 4, while all technical details are relegated to the supplementary material.

2. SSI 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 dataset $\{(Y_i, X_{ij}), i = 1, \ldots, n, j = 1, \ldots, p\}$ containing $n$ subjects and $p$ covariates. Let $\mathbb{Y} = (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^n$ 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 $X_i = (X_{i1}, \ldots, X_{ip})^\top \in \mathbb{R}^p$ and the $j$th covariate for the $n$ subjects $X_j = (X_{ij1}, \ldots, X_{ijn})^\top \in \mathbb{R}^n$ as defined in Section 1. We also assume that all $X_i$s 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 SSL to fill in a larger number of missing values. Specifically, for every covariate $X_j$ ($j = 1, \ldots, p$), 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 $X_j$, where the weights are determined by similarity measures (or distances) between the $i$th and any other subjects (see, e.g., Zhu, Ghahramani, and Lafferty 2003; Belkin, Niyogi, and Sindhwani 2006).

In general, if the two subjects’ corresponding responses and covariates are similar, then their distance should be small. Hence, for any two different subjects $i_1$ and $i_2$, we construct the weight matrix $A = (a_{i_1i_2}) \in \mathbb{R}^{n \times n}$ given below (see, e.g., Zou et al. 2017).

$$a_{i_1i_2} = K_{h_1}(Y_{i_1} - Y_{i_2}) \prod_{k \in D_{i_1} \cap D_{i_2}} K_{h_2}(X_{i_1k} - X_{i_2k}),$$  \hspace{1cm} (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 $X_j$ is discrete,

| 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                           | 0                       | 0                         | 2            | 220         | 30.8            |
| 5       | 1           | 1                           | 0                       | 0                         | 0            | 1161        | 46.2            |
| 6       | 0           | 1                           | 0                       | 0                         | 0            | 636         | 84.6            |
|         |             |                             |                         |                           |              | Total       | 2390            | 49.4            |
we define $K_{h_2}(X_{i,k} - X_{i,k}) = K_{h_2}(0)$ if $X_{i,k}$ and $X_{i,k}$ belong to the same category and $K_{h_2}(X_{i,k} - X_{i,k}) = K_{h_2}(1)$ otherwise (see, e.g., Rajagopalan and Lall 1995). In this article, we consider the popular kernel function used in machine learning, the radial basis function kernel (i.e., Gaussian kernel). Accordingly, the resulting weight is

$$a_{i_1,i_2} = \exp \left\{ -\lambda_1(Y_{i_1} - Y_{i_2})^2 - \lambda_2 \sum_{k \in D_{i_1} \cap D_{i_2}} (X_{i,k} - X_{i,k})^2 \right\},$$

where $\lambda = (\lambda_1, \lambda_2)^T$, $\lambda_j = h_j^{-1} (j = 1, 2)$ 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 $j = 1, \ldots, p$, define $S_{ij} = \{i : D_{ij} = 0\}$ and $S_{ij} = \{i : D_{ij} = 1\}$, thus they represent the missing and non-missing subjects in covariate $j$, respectively. Accordingly, $X_{S_{ij}} = (X_{ij} : i \in S_{ij})$ and $X_{S_{ij}} = (X_{ij} : i \in S_{ij})$ are the missing and non-missing subsets of $X_j$. Adopting the concept of SSL in Zhu, Ghahramani, and Lafferty (2003) and Lafferty and Wasserman (2007), we consider that the missing observation $X_j$ for $i \in S_{ij}$ 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 $X_{S_{ij}}$ by minimizing the following quadratic loss function with respect to $X^*_j$ under the constraint $X^*_j = X_j$ for any $i \in S_{ij}$,

$$\min_{X^*_j} \sum_{i_1 = 1}^n \sum_{i_2 = 1}^n w_{i_1,i_2} (X^*_{ij} - X_{ij})^2, \quad \text{subject to} \quad X^*_j = X_j \text{ for any } i \in S_{ij}, \quad (2.2)$$

where $w_{i_1,i_2} = a_{i_1,i_2} / \sum_{i_2} a_{i_1,i_2}$ are the row-normalized versions of $a_{i_1,i_2}$ and they constitute the weighted matrix $W = (w_{i_1,i_2})$.

To solve (2.2), we set the first derivative of the objective function (2.2) with respect to $X^*_{ij}$ for any $i_1 \in S_{ij}$, to be 0. As a result, $\sum_{i_2 = 1}^n w_{i_1,i_2} (X^*_{ij} - X_j) = 0$ for any $i_1 \in S_{ij}$. 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 $i_1 \in S_{ij}$,

$$X^*_{ij} = \sum_{i_2 \in S_{ij}} w_{i_1,i_2} X^*_{ij} + \sum_{i_2 \in S_{ij}} w_{i_1,i_2} X_{ij}. \quad (2.3)$$

Define $C_i = \sum_{i_2 \in S_{ij}} w_{i_1,i_2} X^*_{ij}$, $C_{S_{ij}} = (C_i : i \in S_{ij})$, and $W_{S_{ij}} = (w_{i_1,i_2} : i_1 \in S_{ij}, i_2 \in S_{ij})$. Then Equation (2.3) can be expressed as

$$X^*_{S_{ij}} = W_{S_{ij}} X_{S_{ij}} + C_{S_{ij}}, \quad (2.4)$$

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

$$\hat{X}_{S_{ij}} = (I - W_{S_{ij}})^{-1} W_{S_{ij}} S_{ij} X_{S_{ij}}, \quad (2.5)$$

which is a weighted average of the observed subjects in $X_{S_{ij}}$. Since the above imputation procedure is motivated from SSL, we denote our proposed approach SSI as mentioned in Section 1. It is worth noting that SSL 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 $\hat{X}_{S_{ij}}$ 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 $j = 1, \ldots, p$,

$$P\left( \left| \frac{1}{|S_{ij}|} \sum_{i \in S_{ij}} (\hat{X}_j - X_j) \right| > \frac{\delta}{n} \right) \leq 6 \exp \left( -\frac{1}{2} \left( \frac{\delta^2}{C_1 n + C_2 n^2} \right) \right).$$

The above theorem implies that $|S_{ij}|^{-1} \sum_{i \in S_{ij}} (\hat{X}_j - X_j) \to_p 0$ 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 datasets. 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 $X_j$ is discrete and has $C$ class labels $\{1, \ldots, C\}$. We then define label matrix $\hat{X}_j = (\hat{X}_{i,j}) \in \mathbb{R}^{n \times C}$, where $\hat{X}_{i,j}$ denotes the probability of the $j$th covariate and subject $i$ in class $c$. Note that if $X_{S_{ij}}$ is observed, then, for any $i \in S_{ij}$, we have $\hat{X}_{i,j} = 1$ if $X_{ij} = c$, and $\hat{X}_{i,j} = 0$ otherwise. In addition, define $\bar{X}_{S_{ij}} = (\bar{X}_{i,j} : i \in S_{ij}, 1 \leq c \leq C)$ and $\hat{X}_{S_{ij}} = (\hat{X}_{i,j} : i \in S_{ij}, 1 \leq c \leq C)$. 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 $\hat{X}_{S_{ij}}$, and name it as $\hat{X}_{S_{ij}}^{(0)}$. Then update $\hat{X}_{S_{ij}}$ by $\hat{X}_{S_{ij}}^{(1)} = W_{S_{ij}} \hat{X}_{S_{ij}}^{(0)} + \bar{C}_{S_{ij}}$, with $\bar{C}_{S_{ij}} = W_{S_{ij}} S_{ij} \hat{X}_{S_{ij}}$. Subsequently normalize the rows of $\hat{X}_{S_{ij}}^{(1)}$. ...
2. In the kth step, for given \( \hat{X}_{i,j}^{(k)} \), update \( \hat{X}_{S_{i,j}} \) by \( \hat{X}_{S_{i,j}}^{(k+1)} = W_{S_{i,j}} \hat{X}_{i,j}^{(k)} + \hat{c}_{S_{i,j}} \), and normalize the rows of \( \hat{X}_{S_{i,j}}^{(k+1)} \).

3. Stop at the kth step if \( \| \hat{X}_{S_{i,j}}^{(k)} - \hat{X}_{S_{i,j}}^{(k-1)} \| < \epsilon \) for some prespecified value \( \epsilon \), where \( \| \| \) denotes the Euclidean norm of any arbitrary vector. Afterwards, normalize the rows of \( \hat{X}_{S_{i,j}}^{(k)} \). Finally, for any \( i \in S_{i,j} \), obtain \( \hat{X}_i = \text{argmax}_{1 \leq c \leq \hat{c}_{i,j}^{(k)}} \hat{X}_{i,j}^{(k)} \).

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, Ghahramani, and Lafferty 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 \( \delta > 0 \), there exist finite positive constants \( C_1 \) and \( C_2 \) such that, for each \( j = 1, \ldots, p \),

\[
P \left( \frac{1}{|S_{i,j}|} \sum_{i \in S_{i,j}} (\hat{X}_{i,j}^{(K)} - \hat{X}_{i,j}) > \frac{\delta}{n} \right) \leq 6 \exp \left( -\frac{1}{2} \frac{\delta^2}{C_1 \epsilon n + C_2} \right),
\]

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 SSI, 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 \( d_0 = 1 + \max_{1 \leq i \leq n} |D_i \cap D_{i+1}| \) 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 \( \tau \).

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

After selecting scale parameters, one cannot only impute missing values, but also conduct subsequent data analysis such as estimation, prediction, and classification. Since the focus of our article is prediction, we first assume that there exists a relationship between the responses and the imputed-observed covariates via the linear regression model, \( Y_i = \hat{X}_{i,0}^{(c)}(\tau) \beta + \epsilon_i \), where \( \hat{X}_{i,0}^{(c)}(\tau) = (\hat{X}_{i,1}(\tau), \ldots, \hat{X}_{i,p}(\tau))^\top \).

### 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, \( Y_i = X_i^{\top} \beta + \epsilon_i \), where \( \beta = (\beta_1, \ldots, \beta_p)^\top \) is the \( p \times 1 \) vector of unknown regression coefficients. In addition, \( \epsilon_i \)s are random errors with mean 0 and variance \( \sigma^2 \). Then, the resulting ordinary least-square estimator of \( \beta \) is \( \hat{\beta}(\tau) = (\hat{X}_{i,0}^{(c)}(\tau) \hat{X}_{i,0}^{(c)}(\tau))^\top \hat{Y}_{i,0}^{(c)}(\tau) \). Then, for any given \( \hat{X}_{i,0}^{(c)}(\tau) \), we obtain the prediction \( \hat{Y}_{i,0}^{(c)}(\tau) = \hat{X}_{i,0}^{(c)}(\tau) \hat{\beta}(\tau) \). 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 \( \hat{\beta}(\tau) \) after removing the jth subject from the data and compute \( \hat{Y}_{i,0}^{(c)}(\tau) = \hat{X}_{i,0}^{(c)}(\tau) \hat{\beta}(\tau) \). Then, select \( \tau \) by minimizing the squared errors of predictions across all subjects \( \sum_{i=1}^{n} (\hat{Y}_{i,0}^{(c)}(\tau) - Y_i)^2 / n \) with a grid-search over the prespecified bounded region.

In this simulation setting, we consider 7 missing patterns:

\( \mathcal{D}_1 = \{1, 2, 3, 10\} \), \( \mathcal{D}_2 = \{4, 5, 6, 10\} \), \( \mathcal{D}_3 = \{7, 8, 9, 10\} \), \( \mathcal{D}_4 = \{1, 2, 3, 4, 5, 6, 10\} \), \( \mathcal{D}_5 = \{1, 2, 3, 7, 8, 9, 10\} \), \( \mathcal{D}_6 = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\} \), \( \mathcal{D}_7 = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\} \).
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 amount 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 |
| Approval-number                | Continuous     | Fraud record     | Approval number for non-commercial banks |
| Fraud-score                    | Continuous     | Fraud score      | 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, 1000, and 2000).

| $R^2$ | Methods | IA | EA | PA |
|-------|---------|----|----|----|
| 0.3   | KNN     | 0.79(0.197) | 0.76(0.061) | 0.83(0.098) |
|       | RF      | 0.74(0.228) | 0.78(0.027) | 0.78(0.081) |
|       | MI      | 1.15(0.347) | 0.79(0.024) | 0.80(0.085) |
|       | RI      | 0.86(0.262) | 0.81(0.141) | 0.97(0.399) |
|       | FIME    | 0.76(0.088) | 0.76(0.088) | - |
|       | SS1     | 0.68(0.157) | 0.70(0.069) | 0.83(0.083) |
|       | SS2     | 0.70(0.149) | 0.72(0.046) | 0.77(0.079) |
| 0.6   | KNN     | 0.79(0.192) | 0.70(0.077) | 0.73(0.096) |
|       | RF      | 0.74(0.229) | 0.73(0.027) | 0.54(0.088) |
|       | MI      | 1.15(0.351) | 0.74(0.022) | 0.57(0.094) |
|       | RI      | 0.86(0.260) | 0.73(0.093) | 0.66(0.252) |
|       | FIME    | 0.69(0.045) | 0.69(0.045) | - |
|       | SS1     | 0.68(0.158) | 0.65(0.053) | 0.54(0.072) |
|       | SS2     | 0.68(0.157) | 0.66(0.053) | 0.52(0.075) |
| 0.9   | KNN     | 0.80(0.192) | 0.68(0.062) | 0.48(0.125) |
|       | RF      | 0.74(0.229) | 0.69(0.028) | 0.29(0.107) |
|       | MI      | 1.15(0.349) | 0.70(0.021) | 0.34(0.121) |
|       | RI      | 0.86(0.261) | 0.67(0.049) | 0.36(0.168) |
|       | FIME    | 0.64(0.044) | 0.64(0.044) | - |
|       | SS1     | 0.67(0.162) | 0.63(0.042) | 0.27(0.083) |
|       | SS2     | 0.68(0.155) | 0.62(0.053) | 0.26(0.079) |

NOTE: 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 ‘-‘. 
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 $\tau$ ranging from 0 to 2, and we denote them SSI1 and SSI2, respectively. It is worth noting that SSI is a weighted average of the observed values and their associated weights are evaluated by 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, $R^2$, $\rho$, and $n$, 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 ($n = 500, 1000$, and $2000$).

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 ($n = 500, 1000$, and $2000$). Analogous to the results in Table 3, both SSI1 and SSI2 outperform the other methods in imputation and estimation. Although SSI1 is slightly better than SSI2 in imputation, SSI2 performs the best in prediction. Note that both SSI1 and SSI2 improve in imputation and prediction as $\rho$ 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 SSI1 and SSI2 are generally superior to the other four methods in imputation, estimation and prediction. It is also worth noting that SSI1 and SSI2 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 ($n = 500, 1000$, and $2000$) in Panels A and B, respectively, under MNAR. The IA and PA are obtained from SSI1 and SSI2 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 SSI1 performs slightly better (weaker) than SSI2 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 SSI1 and SSI2 improves as the sample size increases, which supports our theoretical findings.

| $\rho$ | Methods | IA | EA | PA |
|-------|---------|----|----|----|
| 0.25  | KNN     | 1.00(0.066) | 0.670(0.075) | 0.790(0.223) |
|       | RF      | 1.01(0.056) | 0.710(0.048) | 0.620(0.178) |
|       | MI      | 1.540(0.095) | 0.730(0.042) | 0.670(0.163) |
|       | RI      | 1.15(0.101) | 0.700(0.104) | 0.750(0.372) |
|       | FIME    | -       | 0.650(0.078) | -       |
|       | SSI1    | 0.85(0.038) | 0.600(0.038) | 0.610(0.208) |
|       | SSI2    | 0.85(0.038) | 0.610(0.058) | 0.580(0.192) |
|       | MI      | 1.190(0.082) | 0.750(0.037) | 0.550(0.199) |
|       | RI      | 0.88(0.095) | 0.740(0.099) | 0.650(0.369) |
|       | FIME    | -       | 0.700(0.053) | -       |
|       | SSI1    | 0.71(0.035) | 0.660(0.039) | 0.540(0.234) |
|       | SSI2    | 0.720(0.037) | 0.670(0.047) | 0.510(0.216) |
|       | MI      | 0.570(0.076) | 0.840(0.059) | 0.660(0.450) |
|       | RI      | 0.490(0.028) | 0.750(0.042) | 0.460(0.235) |
|       | FIME    | -       | 0.740(0.079) | -       |
|       | SSI1    | 0.470(0.024) | 0.710(0.040) | 0.490(0.260) |
|       | SSI2    | 0.490(0.036) | 0.720(0.041) | 0.460(0.235) |

Table 4. The simulation results of IA, EA, and PA under the three correlations ($\rho = 0.25, 0.5, 0.75$) and three missing mechanisms (MCAR, MAR, and MNAR), and they are obtained by averaging over the three coefficients of determination ($R^2=0.3, 0.6, 0.9$) and three sample sizes ($n = 500, 1000$, and $2000$).

| $\rho$ | Methods | IA | EA | PA |
|-------|---------|----|----|----|
| 0.25  | KNN     | 1.00(0.066) | 0.670(0.075) | 0.790(0.223) |
|       | RF      | 1.01(0.056) | 0.710(0.048) | 0.620(0.178) |
|       | MI      | 1.540(0.095) | 0.730(0.042) | 0.670(0.163) |
|       | RI      | 1.15(0.101) | 0.700(0.104) | 0.750(0.372) |
|       | FIME    | -       | 0.650(0.078) | -       |
|       | SSI1    | 0.85(0.038) | 0.600(0.038) | 0.610(0.208) |
|       | SSI2    | 0.85(0.038) | 0.610(0.058) | 0.580(0.192) |
|       | MI      | 1.190(0.082) | 0.750(0.037) | 0.550(0.199) |
|       | RI      | 0.88(0.095) | 0.740(0.099) | 0.650(0.369) |
|       | FIME    | -       | 0.700(0.053) | -       |
|       | SSI1    | 0.71(0.035) | 0.660(0.039) | 0.540(0.234) |
|       | SSI2    | 0.720(0.037) | 0.670(0.047) | 0.510(0.216) |
|       | MI      | 0.570(0.076) | 0.840(0.059) | 0.660(0.450) |
|       | RI      | 0.490(0.028) | 0.750(0.042) | 0.460(0.235) |
|       | FIME    | -       | 0.740(0.079) | -       |
|       | SSI1    | 0.470(0.024) | 0.710(0.040) | 0.490(0.260) |
|       | SSI2    | 0.490(0.036) | 0.720(0.041) | 0.460(0.235) |

NOTE: 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 “-“.

This finding is sensible since a stronger signal can lead to better imputation, estimation and prediction.
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 \( X_i \)’s being generated from a multivariate exponential distribution; (ii) the selection of the tuning parameter \( \tau \); (iii) the “non-ignorable missing data” setting modified from Kim and Yu (2011); and (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 dataset 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 \( n = 2390 \) applicants and 13 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 13 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 dataset 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 

\[
\text{MEAN} = \frac{1}{100} \sum_{m=1}^{100} \mu^{(m)} \\
\text{SD} = \left\{ \frac{1}{100} \sum_{m=1}^{100} (\mu^{(m)} - \text{MEAN})^2 \right\}^{1/2}
\]

The scale-parameter vector \( \lambda \) 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\(_1\) as a baseline to calculate the percentage reduction (PR) in MEAN and SD by comparing the other four methods to it.
Table 5 demonstrates that SSI2 yields more than 8% reductions in MEAN and SD, respectively, over each of other methods except for SSI1. Based on the above findings, we conclude that SSI produces reliable predictions. Accordingly, financial institutions can multiply the predicted income by a prespecified 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 article proposes a 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 $Y$ is partially observed. In this case, the weight matrix $ar{A}$ can be defined as $\bar{a}_{i12} = K_h(Y_i - Y_{i2}) \prod_{k \in D_1 \cap D_2} K_h(X_{i1k} - X_{i2k})$ if $Y_i$ and $Y_{i2}$ are both observed, otherwise $\bar{a}_{i12} = \prod_{k \in D_1 \cap D_2} K_h(X_{i1k} - X_{i2k})$ if either $Y_i$ or $Y_{i2}$ is missing. To impute missing observations in any given covariate, SSI does not use 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 $X_1$, and then impute $X_2$ by taking into account the imputed $X_1$. Repeat this procedure until $X_p$ has been imputed. Subsequently, we restart a new iterative process, and begin to impute $X_1$ by taking into account the imputed $X_2, \ldots, X_p$ obtained from the previous iterative process. Then impute $X_2$ by taking into account the imputed $X_1, X_3, \ldots, X_p$. Repeat this new iterative procedure $m$ times, where $m$ is a prespecified 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.

**Supplementary Material**

The online supplementary material includes four components. Appendix A presents six technical conditions, Appendix B provides the proof of Theorem 1, Appendix C provides additional simulations to assess the robustness of SSI against data non-normality, tuning parameter selection and different missing mechanisms, and Appendix D presents simulation results for sequentially semi-supervised imputation, SSSI. Note that the conditions are used only for the theoretical proofs, and not for the practical imputations of SSI.

**Acknowledgments**

The authors are grateful to the editor, associate editor, and anonymous referees for their insightful comments and constructive suggestions.

**Funding**

Wei Lan’s research was supported by the National Natural Science Foundation of China (NSFC,71991472, 12171395, 11931014, 71532001), the Joint Lab of Data Science and Business Intelligence at Southwestern University of Finance and Economics, and the Fundamental Research Funds for the Central Universities (JBK1806002). Xuerong Chen’s research was supported by the National Natural Science Foundation of China (NSFC,11871402,11931014) and the Fundamental Research Funds for the Central Universities (JBK180602). Tao Zou’s research was supported by ANU College of Business and Economics Early Career Researcher Grant, the RSFAS Cross Disciplinary Grant.

**References**

Allison, P. D. (2012), “Handling Missing Data by Maximum Likelihood,” SAS Global Forum Proceedings, pp. 1–21. [1282]

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. [1283,1284]

Bollinger, C., and Hirsch, B. (2013), “Is Earnings Nonresponse Ignorable,” The Review of Economics and Statistics, 95, 407–416. [1282]

Chen, S., and Haziza, D. (2017), “Multiply Robust Imputation Procedures for the Treatment of Item Nonresponse in Surveys,” Biometrika, 104, 439–453. [1282]

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. [1283]

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. [1282]

Du, C., and Zhao, Y. (2017), “On Consistency of Graph-based Semi-supervised Learning,” Machine Learning, arXiv:1703.06177. [1284]

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. [1282]

Han, P., and Wang, L. (2013), “Estimation With Missing Data: Beyond Double Robustness,” Biometrika, 100, 417–430. [1282]

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. [1288]

Lafferty, J., and Wasserman, L. (2007), “Statistical Analysis of Semi-Supervised Regression,” in Advances in Neural Information Processing Systems, Proceedings of the 20th International Conference on Neural Information Processing Systems, 801–808. [1284]
Little, R. J. A. (1992), “Regression With Missing X’: A Review,” *Journal of the American Statistical Association*, 87, 1227–1237. [1282]

Little, R. J. A., and Rubin, D. B. (2002), *Statistical Analysis With Missing Data* (2nd ed.), New York: Wiley. [1282]

Lin, H., Liu, W., and Lan, W. (2020), “Regression Analysis With Individual-Specific Patterns of Missing Covariates,” *Journal of Business & Economic Statistics*, 39, 179–188. [1285]

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. [1282]

Myrthveit, 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. [1282]

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. [1282]

Rajagopalan, B., and Lall, U. (1995), “A Kernel Estimator for Discrete Distributions,” *Journal of Nonparametric Statistics*, 4, 409–426. [1284]

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. [1282]

Shao, J., and Wang, H. (2002), “Sample Correlation Coefficients Based on Survey Data Under Regression Imputation,” *Journal of American Statistical Association*, 97, 544–552. [1282]

Stekhoven, D., and Bühlmann, P. (2012), “MissForest Non-parametric Missing Value Imputation for Mixed-type Data,” *Bioinformatics*, 28, 112–118. [1282]

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Alm, R. (2001), “Missing Value Estimation Methods for DNA Microarrays,” *Bioinformatics*, 17, 520–525. [1282]

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. [1282]

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. [1282]

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. [1284]

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. [1282]

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. [1283,1284,1285]

Zhu, X., and Goldberg, A. B. (2009), *Introduction to Semi-Supervised Learning*, San Rafael: Morgan & Claypool. [1283,1285]

Zou, T., Lan, W., Wang, H., and Tsai, C.-L. (2017), “Covariance Regression Analysis,” *Journal of the American Statistical Association*, 112, 266–281. [1283]