Statistica Sinica **30** (2020), 631-651
doi:https://doi.org/10.5705/ss.202018.0008

IMPUTED FACTOR REGRESSION FOR HIGH-DIMENSIONAL BLOCK-WISE MISSING DATA

Yanqing Zhang[1], Niansheng Tang[1] and Annie Qu[2]

[1] *Yunnan University and* [2] *University of Illinois at Urbana-Champaign*

Abstract: Block-wise missing data are becoming increasingly common in high-dimensional biomedical, social, psychological, and environmental studies. As a result, we need efficient dimension-reduction methods for extracting important information for predictions under such data. Existing dimension-reduction methods and feature combinations are ineffective for handling block-wise missing data. We propose a factor-model imputation approach that targets block-wise missing data, and use an imputed factor regression for the dimension reduction and prediction. Specifically, we first perform screening to identify the important features. Then, we impute these features based on the factor model, and build a factor regression model to predict the response variable based on the imputed features. The proposed method utilizes the essential information from all observed data as a result of the factor structure of the model. Furthermore, the method remains efficient even when the proportion of block-wise missing is high. We show that the imputed factor regression model and its predictions are consistent under regularity conditions. We compare the proposed method with existing approaches using simulation studies, after which we apply it to data from the Alzheimer's Disease Neuroimaging Initiative. Our numerical results confirm that the proposed method outperforms existing competitive approaches.

Key words and phrases: Alzheimer's disease, Alzheimer's Disease Neuroimaging Initiative, block-wise missing data, data imputation, dimension reduction, factor model, principal component.

1. Introduction

Factor models play an important role in simultaneously extracting predictive information and modeling the commonality and dependence of observed data. The essential idea is to combine all predictors to construct low-dimensional latent factor variable, without loss of information. Moreover, a factor model can be utilized in a factor regression model, which associates the response variables and the covariate information through latent factors. This is especially effective when the data are high-dimensional. For this reason, the factor regression model has been widely employed in many fields, including economics (Stock and Watson

(2002)) and biomedical studies (West (2003); Rai and Daume (2008); Carvalho et al. (2008)).

The existing literature on factor regression models includes the works of Stock and Watson (2002), Artis, Banerjee and Marcellino (2005), Forni et al. (2005), Bair et al. (2006), Anderson and Vahid (2007), Bai and Ng (2008), Giannone, Reichlin and Small (2008), Kneip and Sarda (2011), Pan et al. (2015), Guo, Ahn and Zhu (2015), Zhu et al. (2017), Fan, Lian and Wang (2016), and Fan, Xue and Yao (2017). Specifically, Stock and Watson (2002) apply the principal components (PC) from a large number of predictors to estimate the factors when forecasting a single time series. Bair et al. (2006) develop a joint model based on factors estimated by the supervised PC. Kneip and Sarda (2011) apply the factor model to decompose the predictors into two parts: uncorrelated random components, reflecting common factors, and specific variabilities of the explanatory variables. Pan et al. (2015) propose an additive hazards model with latent variables from a factor model. Zhu et al. (2017) propose a multiscale weighted PC regression for performing matrix decompositions for both dimension reduction and feature extraction. Fan, Xue and Yao (2017) analyze the projected PC for a semiparametric factor model and develop sufficient forecasting using a sliced inverse regression.

However, most existing factor regression models are built on fully observed data. The statistical properties of factor regression models are not well studied for block-wise missingness. In the event of block-wise missing predictors, a large portion of predictors is missing for one or more blocks of the sources data, as indicated in Figure 1. This can be caused by high measurement costs, poor data quality, or the noncompliance of participants. Block-wise missing data are prevalent in multisource high-dimensional data, especially in the biomedical, social, psychological and environmental science fields. Therefore, it is important to develop dimension-reduction methods and to achieve accurate prediction power for block-wise missing data.

Recent works on solving block-wise missing data problems include those of Zhou, Little and Kalbfleisch (2010), Yuan et al. (2012), Xiang et al. (2014), Thung et al. (2014), Li et al. (2014), and Liu et al. (2017). For example, Liu et al. (2017) develop a hypergraph classification model based on the availability of different modalities from incomplete multi-modality data. Yuan et al. (2012) propose an incomplete multi-source feature learning method (iMSF) that performs feature learning for each disjoint group independently, and then combines these results. However, the iMSF does not provide a consistent prediction model

for a unified data source across different groups. This makes it difficult to predict when the testing data involve a different data source combination to that of the training data. Xiang et al. (2014) propose an incomplete source-feature selection method (iSFS) to address the above problems. They first partition subjects into several views, according to the availability of data modalities, and then build a bi-level (i.e., both feature-level and source-level) feature learning model to learn the optimal weights for the features and views. Although the iMSF and iSFS both avoid imputing missing data, they do not fully utilize the observed data from other groups when they build the models in each group, potentially leading to inefficient estimations. Moreover, the accelerated gradient method in the iSFS incurs a high computational cost for high-dimensional parameters and multi-view feature learning models.

We propose a novel imputation method, referred to as factor-model imputation, for carrying out block-wise missing value imputations and performing dimension reduction using a factor regression. First, we first apply the sure screening method to obtain those features related to the response variables, based on block-wise missing data. Then, we use the dependence of the selected covariates to impute the missing values. Consequently, we build an imputed factor regression model for predicting a response variable. From a theoretical viewpoint, we provide the convergence rates of the imputation and estimators of the factors and factor loadings. In addition, we achieve consistency of the factor regression coefficient estimators and the predictions. Simulation studies and a real example using data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) show that the proposed method has good prediction accuracy for finite samples.

The main merits of the proposed method are summarized as follows. It fully utilizes information on correlated predictors to impute missing data through factor modeling. In contrast to traditional imputation methods, such as the expectation-maximization method and inverse probability weighting method (Robins, Rotnitzky and Zhao (1994)), the proposed method does not rely on the missing mechanism and the probability of missingness. Therefore, it is robust against the misspecified probability of missingness. Moreover, the proposed method adopts the PC method, which offers several computational advantages. The PC method is asymptotically equivalent to the maximum likelihood method under normal random errors (Chamberlain and Rothschild (1983); Bai (2003)). However, the maximum likelihood method is not feasible for large-dimensional factor models, owing to the large number of parameters involved. In addition, the proposed method compares favorably with those of Yuan et al. (2012) and

Xiang et al. (2014), because it fully extracts the information of all observed data and has better predictive accuracy, even when the proportion of missing data is quite high. Most importantly, the proposed factor regression models for prediction apply the observed part of the testing data only and do not need to impute the missing part of the testing data. This is quite different from the standard imputation methods, such as the zero imputation, k-nearest neighbors imputation, and inverse probability weighting method.

The proposed method targeting block-wise missing data is motivated by the ADNI data (Jack et al. (2008)) published in 2003, followed by the ADNI-1, ADNI-GO, and ADNI-2 groups. The primary goal is to test whether serial biological clinical markers and neuropsychological assessments can be combined to measure the progression of mild-cognitive impairment (MCI) and early Alzheimer's disease (AD). To facilitate AD research using multi-source data, the ADNI study has been collecting data from various sources, including cerebrospinal fluid (CSF) biomarkers, positron emission tomography (PET), magnetic resonance imaging (MRI), and microarray gene expression profile data (GENE). Unfortunately, not all samples in the ADNI study are fully collected, because they are from different sources. Thus, the existence of block-wise missing data is a major challenge. Figure 2 shows the missing structure of baseline ADNI-2 data, with block-wise missing features. The data consist of high-dimensional variables from MRI or GENE data, which often contain irrelevant variables corresponding to the response variable. Moreover, these variables are correlated with each other, because they are from various sources. For more detail on the ADNI data, see `www.adni.loni.ucla.edu`.

The remainder of the paper is organized as follows. Section 2 provides the background for block-wise missing data and the factor regression model. Section 3 presents the proposed method and the theoretical results. Section 4 compares the proposed approach with existing works using simulation studies. Section 5 illustrates an application of the proposed method to ADNI data. Concluding remarks and a discussion are provided in Section 6. All technical derivations are provided in the Supplementary Material.

2. Background

2.1. Data structure for block-wise missing data

In this subsection, we introduce some notation and the structure of the block-wise missing data. Let $\{\boldsymbol{x}_i, y_i\}$ $(i = 1, \ldots, n)$ be a set of independent and

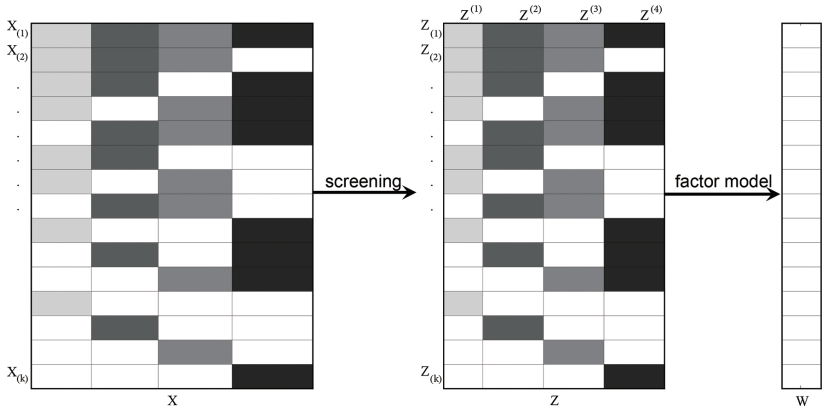

Figure 1. An illustration of the proposed method, with multi-source block-wise missing data with four sources. The blank region indicates missing data from the corresponding source.

identically distributed (i.i.d.) observations of random variables $\{\boldsymbol{x}, y\}$, where y is the response variable without missing, and $\boldsymbol{x}$ denotes the p-dimensional predictor variables from K different data sources that could be block-wise missing; that is, a certain block of predictors could be missing for some subjects (see Figure 1). We assume that each participant has at least one observed data source; that is, there are $2^K - 1$ possible missing patterns: the number of all possible combinations of K data sources, except for the one with all data sources missing.

Denote $\mathbf{Y} = (y_1, \ldots, y_n)^\top$ and $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top = (\mathbf{X}_1, \ldots, \mathbf{X}_p)$ as an $n \times p$ design matrix. Let X_{ij} be the jth variable measured for the ith subject. Thus, $\boldsymbol{x}_i = (X_{i1}, \ldots, X_{ip})^\top$. The predictors $\mathbf{X}$ can be divided into $k (k \leq 2^K - 1)$ groups according to different missing patterns, denoted as $\mathbf{X}_{(j)} = \{\mathbf{X}_{o(j)}, \mathbf{X}_{m(j)}\}$, for $j = 1, \ldots, k$, where $\mathbf{X}_{o(j)} \in \mathbb{R}^{n_j \times p_j}$ is a matrix of observed covariates in the jth group, and $\mathbf{X}_{m(j)} \in \mathbb{R}^{n_j \times (p - p_j)}$ is a matrix of missing covariates in the jth group. Figure 1 illustrates the case when the number of data sources $K = 4$ and $k = 2^K - 1$ groups. For example, the first three sources are observed for $\mathbf{X}_{(2)}$, but the fourth source is missing. In contrast, only the fourth source is observed in group $\mathbf{X}_{(k)}$. We consider the case of $p \gg n$, which is very common in imaging and genetics studies with a high dimension of predictors that exceeds the sample size.

2.2. Factor regression model

In this subsection, we provide the background of the factor regression model.

See also Stock and Watson (2002), West (2003), Rai and Daume (2008), and Carvalho et al. (2008).

The factor regression model consists of two models: one for extracting information from high-dimensional predictors, and one for predictions using latent factors; that is,

$$\begin{cases} \boldsymbol{x}_i = \boldsymbol{\Lambda}\mathbf{w}_i + \boldsymbol{\nu}_i, \\ y_i = \boldsymbol{\alpha}^\top \mathbf{w}_i + \varepsilon_i, \end{cases} \tag{2.1}$$

where $\boldsymbol{\alpha}$ is an $r \times 1$ coefficient vector corresponding to an $r \times 1$ latent factor vector $\mathbf{w}_i$, $\boldsymbol{\Lambda} = (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_p)^\top$ is a $p \times r$ matrix of factor loadings corresponding to the number of factors, ε_i is a random error with a zero mean and a finite variance, and $\boldsymbol{\nu}_i$ is a $p \times 1$ random error vector with a zero mean and a covariance matrix with bounded eigenvalues. The first part of model (2.1) is a latent factor model that attributes a common structure in $\boldsymbol{x}_i$ to underlying factors, and isolates any variation that is purely idiosyncratic in the error $\boldsymbol{\nu}_i$. In the second part of model (2.1), the response variable y_i is conditionally independent of the variables $\boldsymbol{x}_i$, given the latent factor variables $\mathbf{w}_i$. In addition, the original design variables $\boldsymbol{x}_i$ provide information on the latent variables through the factor model, but do not form part of the regression (West (2003)). In matrix form, the factor regression models can be rewritten as

$$\begin{cases} \mathbf{X} = \mathbf{W}\boldsymbol{\Lambda}^\top + \boldsymbol{\nu}, \\ \mathbf{Y} = \mathbf{W}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}, \end{cases} \tag{2.2}$$

where $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_n)^\top$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$, and $\boldsymbol{\nu} = (\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_n)^\top$. The purpose of the factor model is to construct low-dimensional latent factor variables $\mathbf{W}$ by extracting the underlying structure from the predictors.

Two major approaches are commonly used to fit factor regression model. The first is the Bayesian method, given certain assumptions on the prior distribution (West (2003); Rai and Daume (2008); Carvalho et al. (2008)). However, this type of method is usually sensitive to the distribution assumptions. An alternative is the least squares method, with constraints. Specifically, we estimate $\mathbf{w}_i$ and $\boldsymbol{\Lambda}$ by minimizing the least squares objective function (Stock and Watson (2002)):

$$\begin{aligned} Q(\mathbf{W}, \boldsymbol{\Lambda}) &= (np)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{p} (X_{ij} - \boldsymbol{\lambda}_j^\top \mathbf{w}_i)^2 \\ &= (np)^{-1} \operatorname{tr}\left\{ (\mathbf{X} - \mathbf{W}\boldsymbol{\Lambda}^\top)^\top (\mathbf{X} - \mathbf{W}\boldsymbol{\Lambda}^\top) \right\}, \end{aligned} \tag{2.3}$$

subject to $\mathbf{W}^\top\mathbf{W}/n = \mathbf{I}_r$ and $\boldsymbol{\Lambda}^\top\boldsymbol{\Lambda}$ being diagonal, where $\operatorname{tr}(\cdot)$ denotes the matrix

trace and $\mathbf{I}_r$ denotes an $r \times r$ identity matrix.

By replacing $\mathbf{\Lambda}$ in (2.3) with $\mathbf{X}^\top\mathbf{W}(\mathbf{W}^\top\mathbf{W})^{-1}$ derived from (2.2), minimizing (2.3) is equivalent to maximizing $\operatorname{tr}\{\mathbf{W}^\top\mathbf{X}\mathbf{X}^\top\mathbf{W}\}$, subject to $\mathbf{W}^\top\mathbf{W}/n = \mathbf{I}_r$. Then, the problem becomes one of finding the principal components of $\mathbf{X}\mathbf{X}^\top$. Thus, the estimated factor matrix $\widehat{\mathbf{W}}$ is $\sqrt{n}$ times the eigenvectors corresponding to the first r largest eigenvalues of the matrix $\mathbf{X}\mathbf{X}^\top/(np)$, in decreasing order, and the corresponding estimator of $\mathbf{\Lambda}$ is $\widehat{\mathbf{\Lambda}} = \mathbf{X}^\top\widehat{\mathbf{W}}/n$. The PC method is easy to compute and only involves the eigenvalue calculation of an $n \times n$ matrix, where n is smaller than the dimension p. Moreover, this estimator is asymptotically equivalent to the maximum likelihood estimator under the normality assumption (Chamberlain and Rothschild (1983); Bai (2003)). Therefore, we adopt this method to fit the factor regression models, because the computational cost is relatively low.

The normalization restriction of $\mathbf{W}^\top\mathbf{W}/n = \mathbf{I}_r$ and the diagonal constraint of $\mathbf{\Lambda}^\top\mathbf{\Lambda}$ ensure the identifiability of the spaces spanned by the columns of $\mathbf{W}$ and by the columns of $\mathbf{\Lambda}$. Furthermore, they guarantee that $\mathbf{W}$ and $\mathbf{\Lambda}$ are estimable up to an invertible matrix transformation. Here, $\widehat{\mathbf{W}}^\top\widehat{\mathbf{W}}/n = \mathbf{I}_r$ holds, by construction, and $\widehat{\mathbf{\Lambda}}^\top\widehat{\mathbf{\Lambda}} = \widehat{\mathbf{V}}$, where $\widehat{\mathbf{V}}$ is an $r \times r$ diagonal matrix of the first r largest eigenvalues of $\mathbf{X}\mathbf{X}^\top/(np)$, in decreasing order. Thus, the estimators $\widehat{\mathbf{W}}$ and $\widehat{\mathbf{\Lambda}}$ satisfy the restrictions. Consequently, the factors can be estimated asymptotically with rotation. Bai and Ng (2013) provide several conditions under which the true factors and true matrix of factor loadings can be estimated asymptotically without rotation. These conditions include $\mathbf{W}^\top\mathbf{W}/n = \mathbf{I}_r$, and that $\mathbf{\Lambda}^\top\mathbf{\Lambda}$ is a diagonal matrix with distinct entries. In this study, however, we only consider the estimators with rotation, because this will not affect the model interpretation. Once we obtain the estimator of $\mathbf{W}$, the estimator $\widehat{\boldsymbol{\alpha}} = (\widehat{\mathbf{W}}^\top\widehat{\mathbf{W}})^{-1}\widehat{\mathbf{W}}^\top\mathbf{Y}$. Thus, we can predict y_i by model (2.1), given predictor $\boldsymbol{x}_i$ and estimators $\{\widehat{\mathbf{\Lambda}}, \widehat{\boldsymbol{\alpha}}\}$.

3. Proposed Method

3.1. Factor-model imputation

In this section, we present the proposed factor-model imputation approach that targets high-dimensional block-wise missing data. High-dimensional data often contain variables with redundant information, where most of the features might not be relevant to the response variable. This could lead to selecting an improper factor model and producing a biased imputation. Thus, we first select those features that are related to the response variable, before imputing

the missing data. We apply correlation screening to select features based on the observed data. Specifically, we standardize each predictor using the observed values in the corresponding predictor. Define

$$\omega_j = n_{oj}^{-1}\mathbf{X}_{oj}^{\top}\mathbf{Y}_{oj} \text{ for } j = 1, \ldots, p,$$

where n_{oj} is the number of observed values in the jth variable, $\mathbf{X}_{oj}$ is an $n_{oj} \times 1$ vector containing the observed values of the jth variable, and $\mathbf{Y}_{oj}$ is the corresponding response variable vector. We obtain a p-dimensional vector $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_p)^{\top}$, which separately measures the association between a predictor and $\mathbf{Y}$, and is equivalent to the marginal correlation of the predictors, with the response variable rescaled by the standard deviation of the response. Therefore, $\boldsymbol{\omega}$ can be applied to select relevant features. We define a submodel

$$\mathcal{M}_q = \{1 \le i \le p : |\omega_i| \text{ is among the } q \text{ largest of } \boldsymbol{\omega}\},$$

where q can be larger or smaller than the sample size n; for instance, we may choose a conservative q of $n/\log(n)$ or $n-1$, depending on the order of the sample size n.

Screening enables us to choose predictors that are more relevant to $\mathbf{Y}$, thus ensuring effective factor-model imputation and prediction. Moreover, we reduce the dimensionality from a huge p to a relatively large q. Consequently, we can obtain additional relevant features $\mathbf{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)^{\top} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_q)$, a subset columns of $\mathbf{X}$ corresponding to $\mathcal{M}_q$. Similarly to Fan and Lv (2008), we can show that the screening method based on the observed data satisfies the sure-screening property (see the Supplementary Material for technical details).

After sure screening, $\mathbf{Z}$ still contains block-wise missing variables, and the number of data sources may decrease. For simplicity, we assume that $\mathbf{Z}$ still consists of K data sources in total and k data groups, denoted as $\mathbf{Z}_{(j)} = \{\mathbf{Z}_{o(j)}, \mathbf{Z}_{m(j)}\}$, for $j = 1, \ldots, k$. Here, $\mathbf{Z}_{o(j)} \in \mathbb{R}^{n_j \times q_j}$ is a matrix with observed data in the jth group, and $\mathbf{Z}_{m(j)} \in \mathbb{R}^{n_j \times (q-q_j)}$ is an unavailable matrix with missing values in the jth group; see Figure 1 for an example. We assume there exists a group of subjects $\mathbf{Z}_{(1)}$, where all predictors are observed; that is, $q_1 = q$, and the proportion of missing data is $(n - n_1)/n$.

We propose a new imputation approach for missing data that integrates multiple incomplete-block data. One challenge here is that the screening features $\mathbf{Z}$ are from multiple sources, and could be correlated with each other. Therefore, it is important to incorporate the dependence between features when estimating estimate missing data.

We define a factor model among the screening features as

$$\mathbf{Z} = \mathbf{W}\mathbf{\Lambda}^\top + \boldsymbol{e}, \tag{3.1}$$

where $\mathbf{W}$ is an $n \times r$ factor variable matrix, $\mathbf{\Lambda}$ is a $q \times r$ matrix of factor loadings, r is the number of factors, $\boldsymbol{e} = (\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n)^\top$ is an $n \times q$ matrix of random errors, and $\boldsymbol{e}_i$ has a zero mean and a covariance matrix with bounded eigenvalues. The factor models corresponding to the block-wise observed and missing data are

$$\mathbf{Z}_{o(j)} = \mathbf{W}_{(j)}\mathbf{\Lambda}_{o(j)}^\top + \boldsymbol{e}_{o(j)}, \tag{3.2}$$

$$\mathbf{Z}_{m(j)} = \mathbf{W}_{(j)}\mathbf{\Lambda}_{m(j)}^\top + \boldsymbol{e}_{m(j)}, \tag{3.3}$$

respectively, where $\mathbf{W}_{(j)} \in \mathbb{R}^{n_j \times r}$ is the jth sub-matrix of $\mathbf{W}$ according to missing patterns, and $\mathbf{\Lambda}_{o(j)} \in \mathbb{R}^{q_j \times r}$ and $\mathbf{\Lambda}_{m(j)} \in \mathbb{R}^{(q-q_j)\times r}$ are the coefficient matrices corresponding to the block-wise observed data $\mathbf{Z}_{o(j)}$ and missing data $\mathbf{Z}_{m(j)}$, respectively.

In model (3.2), the observed covariate $\mathbf{Z}_{o(j)}$ is used to estimate $\mathbf{W}_{(j)}$. With the estimator of $\mathbf{W}_{(j)}$, the missing values $\mathbf{Z}_{m(j)}$ in model (3.3) can be imputed if $\mathbf{\Lambda}$ is known. Thus, it is critical to obtain an estimator of $\mathbf{\Lambda}$. Based on the complete data $\mathbf{Z}_{(1)}$, we consider the following least squares objective function:

$$Q(\mathbf{W}_{(1)}, \mathbf{\Lambda}) = (n_1 q)^{-1} \operatorname{tr}\{(\mathbf{Z}_{(1)} - \mathbf{W}_{(1)}\mathbf{\Lambda}^\top)^\top(\mathbf{Z}_{(1)} - \mathbf{W}_{(1)}\mathbf{\Lambda}^\top)\}, \tag{3.4}$$

subject to $\mathbf{W}_{(1)}^\top\mathbf{W}_{(1)}/n_1 = \mathbf{I}_r$ and $\mathbf{\Lambda}^\top\mathbf{\Lambda}$ being diagonal. See Section 2.2 for the calculation of the estimated factor matrix $\widetilde{\mathbf{W}}_{(1)}$ and the loading matrix $\widetilde{\mathbf{\Lambda}}$, where the matrix $\widetilde{\mathbf{W}}_{(1)}$ consists of $\sqrt{n_1}$ times the eigenvectors corresponding to the first r largest eigenvalues of the matrix $\mathbf{Z}_{(1)}\mathbf{Z}_{(1)}^\top/(n_1 q)$, in decreasing order, and $\widetilde{\mathbf{\Lambda}} = \mathbf{Z}_{(1)}^\top\widetilde{\mathbf{W}}_{(1)}/n_1$.

Using the estimator $\widetilde{\mathbf{\Lambda}}$, we obtain the estimators of submatrices $\mathbf{\Lambda}_{o(j)}$ and $\mathbf{\Lambda}_{m(j)}$, denoted as $\widetilde{\mathbf{\Lambda}}_{o(j)}$ and $\widetilde{\mathbf{\Lambda}}_{m(j)}$, respectively. By model (3.2), we obtain $\widetilde{\mathbf{W}}_{(j)} = \mathbf{Z}_{o(j)}\widetilde{\mathbf{\Lambda}}_{o(j)}(\widetilde{\mathbf{\Lambda}}_{o(j)}^\top\widetilde{\mathbf{\Lambda}}_{o(j)})^{-1}$. Then, the missing values in (3.3) are imputed as $\widetilde{\mathbf{Z}}_{m(j)} = \widetilde{\mathbf{W}}_{(j)}\widetilde{\mathbf{\Lambda}}_{m(j)}^\top$, for $j = 2, \ldots, k$. The observed and imputed data for the jth data group are $\mathbf{Z}_{o(j)}$ and $\widetilde{\mathbf{Z}}_{m(j)}$, respectively, and

$$\widehat{\mathbf{Z}} = \{\{\mathbf{Z}_{o(j)}, \widetilde{\mathbf{Z}}_{m(j)}\} : j = 1, \ldots, k\} \tag{3.5}$$

is referred to as the factor-model imputation. The imputing process only involves the eigenvalue calculation of an $n_1 \times n_1$ matrix and the inversion of $r \times r$ matrices $(\widetilde{\mathbf{\Lambda}}_{o(j)}^\top\widetilde{\mathbf{\Lambda}}_{o(j)})^{-1}$. Thus the computational cost is relatively low, because n_1 and r are smaller than the original dimension q. Moreover, the imputed features retain the correlation information, and hence preserve the full information of the

observed data for predictions. The consistency of the factor-model imputation is described in Section 3.3.

The proposed imputation method applies the factor structure of multi-source predictors to estimate missing covariates using the PC method. The proposed method is attractive because it does not rely on the missing mechanism or the specified missing probability. Essentially, our method only requires a group of subjects with screening predictors that are fully observed in order to provide factor structure information for all covariates. The factor-model imputation can also be extended to include the case without fully observed subject information. See Section 3.4 for further details.

Note that the number of factors could be unknown, and thus must be estimated. Prior studies on determining the number of factors include the works of Bai and Ng (2002), Alessi, Barigozzi and Capasso (2010), and Ahn and Horenstein (2013). In our numerical studies, we follow the information criterion proposed by Bai and Ng (2002),

$$IC(t) = \ln\{\widehat{Q}(\mathbf{Z}_{(1)})_t\} + tg(n_1, q), \tag{3.6}$$

where $\widehat{Q}(\mathbf{Z}_{(1)})_t$ is the minimum value of function (3.4) with t factors and $g(n, q)$ is a penalty function, for example, $g(n, q) = ((n + q)/(nq)) \ln(nq/(n + q))$. We can obtain the number of factors r by minimizing (3.6).

3.2. Prediction and implementation

In this subsection, we predict responses following the imputed factor regression. Similarly to Section 2.2, we estimate $\mathbf{\Lambda}$ and $\boldsymbol{\alpha}$ in the factor regression models (2.2) using the imputed covariables $\widehat{\mathbf{Z}}$; that is, $\widehat{\mathbf{\Lambda}} = \widehat{\mathbf{Z}}^\top\widehat{\mathbf{W}}/n$ and $\widehat{\boldsymbol{\alpha}} = (\widehat{\mathbf{W}}^\top\widehat{\mathbf{W}})^{-1}\widehat{\mathbf{W}}^\top\mathbf{Y}$, where the matrix $\widehat{\mathbf{W}}$ is $\sqrt{n}$ times the eigenvectors corresponding to the first r largest eigenvalues of the matrix $\widehat{\mathbf{Z}}\widehat{\mathbf{Z}}^\top/(nq)$, in decreasing order. The number of factors r is defined as the minimizer of the criterion in (3.6), based on the imputed variables. Consequently, we can predict y_i using $\widehat{y}_i = \widehat{\boldsymbol{\alpha}}^\top\widehat{\mathbf{w}}_i$.

The following algorithm provides the specific implementation. Given training data $\{\mathbf{X}, \mathbf{Y}\}$ and testing data $\mathbf{x}$, we make a prediction as follows.

Implementation: Prediction based on imputed factor regression models

1. Standardize the columnwise components of $\mathbf{X}$ to be mean zero and standard deviation one on the observed training data: $\mathbf{X}_j \leftarrow (\mathbf{X}_j - \overline{\mathbf{X}}_j)/\mathrm{std}(\mathbf{X}_j)$; and center response $\mathbf{Y}$: $\mathbf{Y} \leftarrow \mathbf{Y} - \overline{\mathbf{Y}}$, where $\overline{\mathbf{X}}_j$ and $\overline{\mathbf{Y}}$ are the means of the columnwise observed components $\mathbf{X}_j$ and $\mathbf{Y}$, respectively.

2. Perform sure screening to obtain the significant features $\mathbf{Z}$ and construct the factor-model imputed predictors $\widehat{\mathbf{Z}}$ using (3.5).

3. Fit the factor regression model based on $\widehat{\mathbf{Z}}$ and obtain the estimators $\widehat{\mathbf{\Lambda}} = \widehat{\mathbf{Z}}^\top\widehat{\mathbf{W}}/n$ and $\widehat{\boldsymbol{\alpha}} = (\widehat{\mathbf{W}}^\top\widehat{\mathbf{W}})^{-1}\widehat{\mathbf{W}}^\top\mathbf{Y}$.

4. Standardize each component of $\mathbf{x}$: $\mathbf{x}_j \leftarrow (\mathbf{x}_j - \overline{\mathbf{X}}_j)/\mathrm{std}(\mathbf{X}_j)$, and obtain $\mathbf{z}$ according to the screening feature in step 2.

5. Based on $\widehat{\mathbf{\Lambda}}$ in step 3, obtain the factor values $\widehat{\mathbf{w}} = \mathbf{z}_o\widehat{\mathbf{\Lambda}}_o(\widehat{\mathbf{\Lambda}}_o^\top\widehat{\mathbf{\Lambda}}_o)^{-1}$, where $\mathbf{z}_o$ consists of observed values and $\widehat{\mathbf{\Lambda}}_o$ is the sub-matrix of $\widehat{\mathbf{\Lambda}}$ corresponding to $\mathbf{z}_o$. Then, predict the response value $\widehat{\mathbf{y}} = \overline{\mathbf{Y}} + \widehat{\boldsymbol{\alpha}}^\top\widehat{\mathbf{w}}$.

One advantage of the proposed method is that the prediction process does not require the imputation of missing data for the testing data, but does for the observed part $\mathbf{z}_o$, owing to the nature of the factor model.

3.3. Theoretical properties

In this subsection, we establish the theoretical properties of the proposed method. We assume that both the sample size n and the dimensionality p go to infinity. Therefore, the number of samples n_1 with all complete observations also goes to infinity. Note that once the number of factors in the factor model is estimated consistently, the following asymptotic results hold by a conditioning argument. Throughout this paper, the number of factors r is assumed to be known and fixed as n and p increase.

We define T_j as the collection of row indices corresponding to subjects in the jth data group $\mathbf{Z}_{(j)}$, and M_{mj} as the collection of column indices corresponding to variables with missingness in the jth data group $\mathbf{Z}_{(j)}$. The details of the assumptions and the proof for the established theorems are provided in the Supplementary Material. The following theorem gives the convergence rate of the factor-model imputation.

Theorem 1. *Under conditions* (C1)–(C5)*, for* $t \in T_j$ *and* $i \in M_{mj}$ *(*$j = 2, \ldots, k$*), we have*

$$\widetilde{\boldsymbol{\lambda}}_i^\top\widetilde{\mathbf{w}}_t - \boldsymbol{\lambda}_i^\top\mathbf{w}_t = O_p\left\{\sqrt{q_j}(\min(\sqrt{n_1 q_j}, q))^{-1}\right\}.$$

Theorem 1 implies that the convergence rate of the estimator of $\boldsymbol{\lambda}_i^\top\mathbf{w}_t$ as an imputation is $\min\{\sqrt{n_1}, q/\sqrt{q_j}\}$. When the factor loadings $\boldsymbol{\lambda}_i$ are all known, $\boldsymbol{\lambda}_i^\top\mathbf{w}_t$ can be estimated as the rate of convergence $\sqrt{n_1}$. The rate of convergence $\min\{\sqrt{n_1}, q/\sqrt{q_j}\}$ implies that the factor loadings are estimated. Based on

the imputed values $\widehat{\mathbf{Z}}$, the convergence rates of the factor estimators and factor loading matrix estimators are established in the following theorem.

Theorem 2. *Under conditions* (C1)–(C5), *we have*

(i) $\widehat{\mathbf{w}}_t - \mathbf{H}^\top \mathbf{w}_t = O_p\left\{(\min(\sqrt{q}, \sqrt{n_1}))^{-1}\right\}$, *for* $t = 1, \ldots, n$;

(ii) $\widehat{\boldsymbol{\lambda}}_i - \mathbf{H}^{-1}\boldsymbol{\lambda}_i = O_p\left\{(\min(\sqrt{q}, \sqrt{n_1}))^{-1}\right\}$, *for* $i = 1, \ldots, q$,

where $\mathbf{H} = (\boldsymbol{\Lambda}^\top\boldsymbol{\Lambda}/q)(\mathbf{W}^\top\widehat{\mathbf{W}}/n)\widehat{\mathbf{V}}^{-1}$, *and* $\widehat{\mathbf{V}}$ *is an* $r \times r$ *diagonal matrix of the first* r *largest eigenvalues of* $\widehat{\mathbf{Z}}\widehat{\mathbf{Z}}^\top/(nq)$, *in decreasing order.*

Theorem 2 indicates that there exists a rotation matrix $\mathbf{H}$, such that $\widehat{\mathbf{w}}_t$ is an estimator of $\mathbf{H}^\top\mathbf{w}_t$ and $\widehat{\boldsymbol{\lambda}}_i$ is an estimator of $\mathbf{H}^{-1}\boldsymbol{\lambda}_i$. Moreover, $\widehat{\mathbf{W}}\widehat{\boldsymbol{\Lambda}}^\top$ is an estimator of $\mathbf{W}\boldsymbol{\Lambda}^\top$. In a regression analysis, using $\mathbf{W}$ as the regressor gives the same predicted value as using $\mathbf{WH}$ as the regressor, because $\mathbf{W}$ and $\mathbf{WH}$ span the same space. The following theorem indicates the consistency of the regression coefficient estimators and the prediction.

Theorem 3. *Under conditions* (C1)–(C6), *as* $p, n \to \infty$, *we have*

(i) $\widehat{\boldsymbol{\alpha}} - \mathbf{H}^{-1}\boldsymbol{\alpha} \xrightarrow{\mathrm{p}} \mathbf{0}$;

(ii) $\widehat{\boldsymbol{\alpha}}^\top\widehat{\mathbf{w}}_t - \boldsymbol{\alpha}^\top\mathbf{w}_t \xrightarrow{\mathrm{p}} 0$ *for* $t = 1, \ldots, n$.

3.4. Factor-model imputation without fully observed subject information

In this subsection, we develop the factor-model imputation method that targets the screening features $\mathbf{Z}$ that do not have fully observed subject information. Without a group of completely observed subjects, we utilize the correlation information from the block-wise observed data to obtain an efficient imputation using an iterative procedure.

Specifically, the factor model (3.1) can be expressed according to the observed and missing data blocks from different sources, as follows:

$$\mathbf{Z}_o^{(i)} = \mathbf{W}_o^{(i)}\boldsymbol{\Lambda}^{(i)\top} + \boldsymbol{e}_o^{(i)}, \tag{3.7}$$

$$\mathbf{Z}_m^{(i)} = \mathbf{W}_m^{(i)}\boldsymbol{\Lambda}^{(i)\top} + \boldsymbol{e}_m^{(i)}, \tag{3.8}$$

for $i = 1, \ldots, K$, where $\mathbf{Z}_o^{(i)} \in \mathbb{R}^{n_o^{(i)} \times q^{(i)}}$ and $\mathbf{Z}_m^{(i)} \in \mathbb{R}^{n_m^{(i)} \times q^{(i)}}$ consist of the observed and missing parts of $\mathbf{Z}$, respectively, corresponding to the ith data source, $\mathbf{Z}^{(i)} = \{\mathbf{Z}_o^{(i)}, \mathbf{Z}_m^{(i)}\} \in \mathbb{R}^{n \times q^{(i)}}$ (see Figure 1), $\mathbf{W}_o^{(i)} \in \mathbb{R}^{n_o^{(i)} \times r}$ and $\mathbf{W}_m^{(i)} \in \mathbb{R}^{n_m^{(i)} \times r}$ are sub-matrices of $\mathbf{W}$ corresponding to $\mathbf{Z}_o^{(i)}$ and $\mathbf{Z}_m^{(i)}$, respectively, $\boldsymbol{\Lambda}^{(i)} \in$

$\mathbb{R}^{q^{(i)}\times r}$ is the sub-matrix coefficient corresponding to the ith data source, and $\mathbf{\Lambda} = (\mathbf{\Lambda}^{(1)\top}, \ldots, \mathbf{\Lambda}^{(K)\top})^\top$.

Based on model (3.7) and the PC method, we obtain the sub-matrix estimator $\widetilde{\mathbf{\Lambda}}^{(i)}$, consisting of $\sqrt{n_o^{(i)}}$ times the eigenvectors corresponding to the first r largest eigenvalues of the matrix $\mathbf{Z}_o^{(i)\top}\mathbf{Z}_o^{(i)}/(n_o^{(i)}q^{(i)})$, in decreasing order (Stock and Watson (2002)). Thus, $\widetilde{\mathbf{\Lambda}} = (\widetilde{\mathbf{\Lambda}}^{(1)\top}, \ldots, \widetilde{\mathbf{\Lambda}}^{(K)\top})^\top$. By model (3.2), we obtain the estimated factor $\widetilde{\mathbf{W}}_{(j)}$ via $\sqrt{n_j}$ times the eigenvectors corresponding to the first r largest eigenvalues of $\mathbf{Z}_{o(j)}\mathbf{Z}_{o(j)}^\top/(n_j q_j)$, in decreasing order. Thus, $\widetilde{\mathbf{W}} = (\widetilde{\mathbf{W}}_{(1)}^\top, \ldots, \widetilde{\mathbf{W}}_{(k)}^\top)^\top$. In addition, we can use model (3.3) to estimate the missing values $\widetilde{\mathbf{Z}}_{m(j)} = \widetilde{\mathbf{W}}_{(j)}\widetilde{\mathbf{\Lambda}}_{m(j)}^\top$, for $j = 1, \ldots, k$.

Because $\widetilde{\mathbf{W}}$ and $\widetilde{\mathbf{\Lambda}}$ are block-wise obtained, based on the different observed data blocks, we can improve the estimation by using an iterative method to fully extract the integral correlation information of all predictors. Specifically, let $\widehat{\mathbf{Z}}^0 = \{\{\mathbf{Z}_{o(j)}, \widetilde{\mathbf{Z}}_{m(j)}\} : j = 1, \ldots, k\}$ as the initial value. At the tth iteration, the estimator $\widehat{\mathbf{W}}^t$ consists of $\sqrt{n}$ times the eigenvectors corresponding to the first r largest eigenvalues of $\widehat{\mathbf{Z}}^{t-1}\widehat{\mathbf{Z}}^{t-1\top}/(nq)$, in decreasing order, where $\widehat{\mathbf{Z}}^{t-1}$ represents the $(t-1)$th iteration value. Then, $\widehat{\mathbf{\Lambda}}^t = \widehat{\mathbf{Z}}^{t-1\top}\widehat{\mathbf{W}}^t/n$. Based on (3.3), we obtain $\widehat{\mathbf{Z}}_{m(j)}^t = \widehat{\mathbf{W}}_{(j)}^t\widehat{\mathbf{\Lambda}}_{m(j)}^{t\top}$, for $j = 1, \ldots, k$. Thus, the tth iterative imputation is $\widehat{\mathbf{Z}}^t = \{\{\mathbf{Z}_{o(j)}, \widehat{\mathbf{Z}}_{m(j)}^t\} : j = 1, \ldots, k\}$. We set $\|\widehat{\mathbf{Z}}^t - \widehat{\mathbf{Z}}^{t-1}\| < c$ as the convergence condition. The iterative procedure converges quickly and achieves effective imputation power, as showed in Section 4.2. Incorporating the imputed data, we follow Section 3.2 to predict the response values.

4. Simulation Study

In this section, we perform simulations to compare the proposed method (FR-FI) with four competing methods: the iMSF method (Yuan et al. (2012)), iSFS method (Xiang et al. (2014)), factor regression model with zero imputation (FR-ZERO), and k-nearest neighbor imputation (FR-KNN) (Hastie et al. (1999)). For the FR-ZERO method, we fill in the missing entries using zero, and then analyze the imputed data using a factor regression and the sure-screening method. The FR-KNN approach applies the k-nearest neighbor imputation (Hastie et al. (1999)) to missing data and builds the factor regression utilizing the sure-screening method.

4.1. Study I

The first simulation study is designed to evaluate the finite-sample performance of the proposed method. We assume a p-dimensional predictor $\boldsymbol{x}$ from $K = 4$ different data sources, where $\boldsymbol{x} = (\boldsymbol{x}^{(1)\top}, \boldsymbol{x}^{(2)\top}, \boldsymbol{x}^{(3)\top}, \boldsymbol{x}^{(4)\top})^\top$, and the corresponding dimension of $\boldsymbol{x}^{(j)}$ is s_j, with $p = \sum_{j=1}^4 s_j$. However, only part of $\boldsymbol{x}^{(j)}$ is related to the response variable y, denoted as $\boldsymbol{x}^{*(j)}$. Let $\boldsymbol{x}^* = (\boldsymbol{x}^{*(1)\top}, \boldsymbol{x}^{*(2)\top}, \boldsymbol{x}^{*(3)\top}, \boldsymbol{x}^{*(4)\top})^\top$ be s-dimensional relevant predictors associated with y_i, and $\boldsymbol{x}^{*c}$ be $(p-s)$-dimensional variables unrelated to y.

We generate the data set $\{y_i, \boldsymbol{x}_i\}$ based on the factor structure $y_i = \boldsymbol{\alpha}^\top \mathbf{w}_i + \varepsilon_i$, $\boldsymbol{x}_i^* = \boldsymbol{\Lambda}\mathbf{w}_i + \boldsymbol{\nu}_i$, and $\boldsymbol{x}_i^{*c} \sim N(\mathbf{0}, \mathbf{I}_{p-s})$, for $i = 1, \ldots, n$, where $\mathbf{w}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, $\varepsilon_i \sim N(0,1)$, and $\boldsymbol{\nu}_i \sim N(\mathbf{0}, \mathbf{I}_s)$. We set the number of factors $r = 5$, parameters $\boldsymbol{\alpha} = (0.8, 0.5, 0.3, 0.1, 0.1)^\top$, and the (l, j) component of $\boldsymbol{\Sigma}$ to $0.7^{|l-j|}$. We generate an $(s \times r)$-dimensional matrix $\boldsymbol{\Lambda}$ of factor loadings, with each component following a standard normal distribution.

We consider two settings, with $p = 1{,}000$ or $4{,}000$. When $p = 1{,}000$, we let $s_1 = s_2 = s_3 = s_4 = 250$, and when $p = 4{,}000$, we let $s_1 = s_2 = s_3 = s_4 = 1{,}000$. In both settings, the first 25 variables of $\boldsymbol{x}^{(j)}$ in each data source are relevant to the response y, and are denoted as $\boldsymbol{x}^{*(j)}$. That is, $s = 100$. We consider the sample size $n = 200$ and different missing mechanisms.

In the following, we first construct the block-wise missing data with the missing completely at random (MCAR) mechanism, such that 80% or 40% of the entire samples is completely observed, and the remainder of the sample is split into $2^4 - 2$ missing patterns, with an equal probability. Moreover, we consider the missing not at random (MNAR) mechanism to construct the missing data, using a variable δ_i to indicate the missing pattern of the ith sample. The variable δ_i is generated from a multinomial distribution with $\mathbf{P}(\boldsymbol{x}_i^*) = (P_1(\boldsymbol{x}_i^*), \ldots, P_{15}(\boldsymbol{x}_i^*))$, where $P_j(\boldsymbol{x}_i^*) = g_j(\boldsymbol{x}_i^*) / \sum_{l=1}^{15} g_l(\boldsymbol{x}_i^*)$ and $g_j(\boldsymbol{x}_i^*) = \boldsymbol{\gamma}_j \boldsymbol{x}_i^*$ for $j = 1, \ldots, 15$. For simplicity, we consider two settings. One has $\boldsymbol{\gamma}_j = (0.1, \ldots, 0.1)$, for $j = 1, \ldots, 14$, and $\boldsymbol{\gamma}_{15} = (1, \ldots, 1)$, with an average missing rate of 58%. The other has $\boldsymbol{\gamma}_j = (0.1, \ldots, 0.1)$, for $j = 1, \ldots, 15$, with an average missing rate of 90%. For each simulation, we employ 80% of the sample as training data, and the remaining 20% as testing data. We perform 100 simulation runs for each method.

Table 1 summarizes the performance of each method based on the mean squared error (MSE) of the predicted values, where the MSE is defined as $n^{-1} \sum_i (y_i - \widehat{y}_i)^2$. Note that the FR-FI is more robust than the other methods under various missing rates and missing mechanisms. Specifically, the FR-FI produces

Table 1. Mean and standard deviation (SD) of mean squared error (MSE) for prediction in simulation study I.

	MCAR				MNAR			
Missing rate	20%		60%		58%		90%	
Method	Mean	SD	Mean	SD	Mean	SD	Mean	SD
$(n, p, s) = (200, 1{,}000, 100)$								
FR-FI	**1.054**	0.242	**1.073**	0.228	**1.055**	0.234	**1.019**	0.452
FR-ZERO	1.238	0.273	1.390	0.318	1.394	0.311	1.464	0.302
FR-KNN	1.384	0.067	1.414	0.063	1.290	0.026	1.255	0.019
iSFS	1.357	0.356	1.462	0.388	1.475	0.360	1.490	0.410
iMSF	1.660	0.343	1.847	0.482	1.724	0.366	1.842	0.424
$(n, p, s) = (200, 4{,}000, 100)$								
FR-FI	**1.043**	0.225	**1.085**	0.232	**1.050**	0.225	**1.163**	0.254
FR-ZERO	1.181	0.291	1.396	0.323	1.386	0.316	1.448	0.368
FR-KNN	1.115	0.007	1.200	0.031	1.295	0.059	1.194	0.045
iSFS	1.310	0.305	1.444	0.341	1.485	0.361	1.543	0.406
iMSF	1.715	0.410	1.874	0.437	1.748	0.364	1.905	0.433

Table 2. Mean and standard deviation (SD) of mean squared error (MSE) for prediction in simulation study II.

	MCAR		MNAR	
Method	Mean	SD	Mean	SD
FR-FI	**1.066**	0.229	**1.097**	0.262
FR-ZERO	1.408	0.340	1.313	0.314
FR-KNN	1.421	0.022	1.383	0.026
iSFS	1.517	0.363	1.456	0.393
iMSF	1.668	0.380	1.674	0.409

a smaller MSE than those of the other methods in all settings, especially when the missing rate is as high as 60% (MCAR) and 90% (MNAR), and the dimension $p = 4{,}000$ is also high. In these scenarios, the FR-FI is 33% better than the iSFS and 73% better than the iMSF in terms of the MSEs. When the missing rate increases, the MSEs of the iMSF and iSFS methods also increase. In contrast, the MSEs of the FR-FI method increase only slightly when the missing rate increases. In summary, this simulation study indicates that the FR-FI is able to achieve better prediction power for multi-source block-wise missing data.

4.2. Study II

The second simulation study is designed to evaluate the finite-sample performance of the proposed method when each subject has missing data. We still

denote this as FR-FI. Similarly to Study I, we generate data $\{y_i, \boldsymbol{x}_i\}$ based on the setting $(n, p, s) = (200, 1{,}000, 100)$, and construct block-wise missing data with $2^4 - 2$ different missing patterns; that is, the number of all possible combinations of four data sources, except for the two combinations of the completely missing and completely observed sources. We generate the MCAR case, with equal probability for 14 missing patterns and the MNAR case. For the MNAR, we consider the parameters $\boldsymbol{\gamma}_{jt}$ of the indicator variable δ_i as 0.1 times the number of the observed data sources in the jth missing pattern, for $j = 1, \ldots, 14$ and $t = 1, \ldots, s$. We conduct 100 replications.

Table 2 summarizes the performance of each method under the MCAR and MNAR cases. It is clear that the FR-FI exhibits the best overall performance in terms of the MSE. Specifically, the MSE of the FR-FI is smaller than that of any of the other methods under both missing mechanisms. The MSEs of the other methods are all above 1.3. In contrast, the proposed method is able to reduce the MSE by more than 20%. In particular, it is 56% (MCAR) and 53% (MNAR) better than the iMSF, indicating that the FR-FI is able to improve the prediction accuracy by incorporating important information from correlated predictors.

5. Real-Data Application

In this section, we apply the proposed method to the baseline ADNI-2 data set (Jack et al. (2008)). The goal of this study is to predict the mini-mental state examination (MMSE) score, a significant criterion used to categorize different Alzheimer's disease (AD) stages. Using this prediction procedure, we are able to evaluate a patient's disease progression based on the predicted MMSE.

AD is the most common form of dementia and results in the loss of memory and impaired of cognitive and language skills. AD is the sixth-leading cause of death in the United States. On the other hand, there is no effective prevention, treatment, or way to slow the progression of the disease. The number of AD patients has increased exponentially as a result of the aging population, causing a socioeconomic burden to families and society (Brookmeyer et al. (2007)). The initial ADNI study was launched in 2003 for AD, and was later followed by ADNI-1, ADNI-GO, and ADNI-2.

In the ADNI data, high-dimensional variables are collected from multiple sources to aid researchers and clinicians to develop new treatments for AD and to monitor patients' disease progression, in addition to decreasing the time and cost of clinical trials. However, the ADNI data contain block-wise missing data, for

Table 3. Statistics of the baseline ADNI-2 data set and the data sources used in our evaluations: "o" denotes the observed data, "–" denotes the missing data, and the number in parentheses is the number of features in each source.

Missing pattern	CSF (3)	PET (243)	MRI (317)	GENE (49386)	# of samples
I	–	o	–	o	21
II	o	o	–	o	13
III	o	o	o	o	467
IV	–	o	o	o	118
V	–	–	o	o	39
VI	–	–	o	–	13
VII	–	o	o	–	47
VIII	o	o	o	–	356
# of samples	836	1,022	1,040	658	1,074

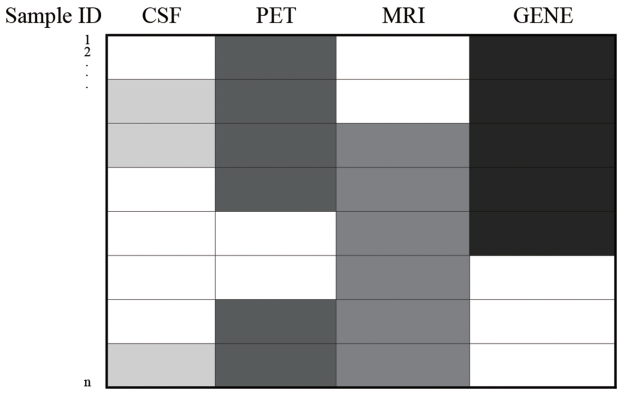

Figure 2. An illustration of multi-source block-wise missing data in the baseline ADNI-2 data, where there are 1,074 participants in total, and four sources (CSF, PET, MRI, and GENE). The blank region indicates missing data from corresponding sources.

several reasons: low-quality data sources of certain samples might be discarded, some data collection (e.g., PET scans) may be too costly for every participant, and participants may not be willing to provide certain measurements (e.g., owing to a lack of consent, participant attrition, or noncompliance with a long scan) (Yuan et al. (2012)). The missing data often emerge in a block-wise fashion.

The proposed method is motivated by the multi-source block-wise missing data. We use the baseline ADNI-2 data set to build factor regression models, with the MMSE score as the response and features from four data sources: three CSF features, 243 PET features, 317 MRI features, and 49,386 GENE features; that is, $p = 49{,}949$. The data contain 1,074 subjects with eight different missing patterns; see Figure 2 and Table 3 for an illustration of the multi-source block-

Table 4. Mean and standard deviation (SD) of mean square error (MSE) and the relative improvement (RIMSE) of the proposed method over existing methods in terms of the mean MSE for ADNI data.

	80% training rate			90% training rate		
Method	Mean	SD	RIMSE	Mean	SD	RIMSE
FR-FI	**5.5765**	1.0511	–	**5.4931**	1.6009	–
FR-ZERO	6.1632	0.9657	10.5%	6.1858	1.6344	12.6%
FR-KNN	6.2669	1.0965	12.4%	6.0526	1.4362	10.2%
iSFS	5.7129	0.9499	2.5%	5.9301	1.7763	8.0%
iMSF	5.8137	1.0118	4.3%	5.7453	1.4704	4.6%

wise missing data and the sample size information of the ADNI-2 data. The sample group with all complete observed features includes 467 subjects, and the total missing rate is about 56.5%. The data are randomly split into an 80% or 90% training set, with the remaining data as testing data. The random split is replicated 30 times. The average numbers of factors selected from the 80% training data and 90% training data are about 15.6 and 17.5, respectively. We compare the proposed method with existing methods in Section 4.

Table 4 indicates that the proposed FR-FI method has the best performance. Specifically, the FR-FI method produces the smallest MSEs among all competing methods. The FR-FI method improves on the MSEs of the FR-ZERO and FR-KNN methods by more than 10%, thus also demonstrating better imputation power. In addition, the FR-FI method improves on the MSEs of the iSFS and iMSF methods by more than 2%, illustrating that the FR-FI method has better prediction accuracy as a result of the joint modeling and incorporating the correlation information of predictors.

6. Discussion

We have proposed a new factor-model imputation method for block-wise missing data that builds an imputed factor regression model. A unique contribution of our method is that we utilize the correlation information between predictors to impute the missing data using a factor structure model. When we have a group of completely observed subjects, the proposed method extracts the correlation information from these subjects in order to estimate the missing values. If we do not have fully observed subjects, we estimate the missing values through iterative factor-model imputing. The advantages of the proposed method include that it does not rely on the missing mechanism or on the missing

probability, and that it has a relatively low computational cost. Moreover, the proposed method extracts information from all available data sources to build imputed factor regression models efficiently. In the prediction process, we only apply the observed part of testing data, which does not require imputing the missing part of the testing data, owing to the nature of factor regression models.

We also show the theoretical properties of the proposed method, along with its numerical performance. We demonstrate the theoretical convergence rate and consistency of the estimators. Our simulation studies show that the proposed method is robust under different missing rates and missing mechanisms, compared with existing approaches. The proposed method demonstrates excellent performance, even when the missing rate is high. In addition, the proposed method outperforms competing methods when applied to the ADNI-2 data.

In the proposed method, we only consider a linear model for the dimension reduction. The proposed factor-model imputation can be extended to more complex predicting models, such as nonlinear regressions and nonparametric regressions. In addition, it would be worth investigating the properties of the iterative factor-model imputation without completely observed subjects.

Supplementary Material

The online Supplementary Material contains all technical conditions and proofs.

Acknowledgments

The authors thank the Associate Editor and two anonymous reviewers for their suggestions and helpful feedback which improved the paper significantly. This research was supported by National Science Foundation Grants (DMS14 15308 and DMS1613190), China Postdoctoral Science Foundation Grants (2017M 623077), and National Natural Science Foundation of P.R. China (11601471, 11731011, 11871420).

References

Ahn, S. C. and Horenstein, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica* **81**, 1203–1227.

Alessi, L., Barigozzi, M. and Capasso, M. (2010). Improved penalization for determining the number of factors in approximate factor models. *Statistics and Probability Letters* **80**, 1806–1813.

Anderson, H. and Vahid, F. (2007). Forecasting the volatility of Australian stock returns: Do

common factors help. *Journal of Business & Economic Statistics* **25**, 76–90.

Artis, M., Banerjee, A. and Marcellino, M. (2005). Factor forecasts for the UK. *Journal of Forecasting* **24**, 279–298.

Bai, J. and Ng, S. (2002). Determing the number of factors in approximate factor models. *Econometrica* **70**, 191–221.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica* **71**, 135–171.

Bai, J. and Ng, S. (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics* **146**, 304–317.

Bai, J. and Ng, S. (2013). Principal components estimation and identification of static factors. *Journal of Econometrics* **176**, 18–29.

Bair, E., Hastie, T., Paul, D. and Tibshirani, R. (2006). Prediction by supervised principal components. *Journal of the American Statistical Association* **101**, 119–137.

Brookmeyer, R., Johnson, E., Ziegler-Graham, K. and Arrighi, H. M. (2007). Forecasting the global burden of Alzheimer's disease. *Alzheimers Dement* **3**, 186–191.

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q. and West, M. (2008). High-dimensional sparse factor modeling: Applications in gene expression genomics. *Journal of the American Statistical Association* **103**, 1438–1456.

Chamberlain, G. and Rothschild, M. (1983). Arbitrage, factor structure and mean-variance analysis in large asset markets. *Econometrica* **51**, 1305–1324.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh-dimensional feature space. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)* **70**, 849–911.

Fan, J., Liao, Y. and Wang, W. (2016). Projected principal component analysis in factor models. *The Annals of Statistics* **44**, 219–254.

Fan, J., Xue, L. and Yao, J. (2017). Sufficient forecasting using factor models. *Journal of Econometrics* **201**, 292–306.

Forni, M., Hallin, M., Lippi, M. and Reichlin, L. (2005). The generalized dynamic factor model: One-sided estimation and forecasting. *Journal of the American Statistical Association* **100**, 830–840.

Giannone, D., Reichlin, L. and Small, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economic* **55**, 665–676.

Guo, R., Ahn, M. and Zhu, H. (2015). Spatially weighted principal component analysis for imaging classification. *Journal of Computational and Graphical Statistics* **24**, 274–296.

Hastie, T., Tibshirani, R., Sherlock, G., Eisen, M., Brown, P. and Botstein, D. (1999). Imputing missing data for gene expression arrays. Technical Report, Division of Biostatistics, Stanford University.

Jack, C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P. J., Whitwell, J. L. and Ward, C. (2008). The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging* **27**, 685–691.

Kneip, A. and Sarda, P. (2011). Factor models and variable selection in high-dimensional regression analysis. *The Annals of Statistics* **39**, 2410–2447.

Li, Z., Li, Q., Han, C. and Li, B. (2014). A hybrid approach for regression analysis with block missing data. *Computational Statistics and Data Analysis* **75**, 239–247.

Liu, M., Zhang, J., Yap, P. T. and Shen, D. (2017). View-aligned hypergraph learning for Alzheimer's disease diagnosis with incomplete multi-modality data. *Medical Image Analysis* **36**, 123–134.

Pan, D., He, H., Song, X. and Sun, L. (2015). Regression analysis of additive hazards model with latent variables. *Journal of the American Statistical Association* **110**, 1148–1159.

Rai, P. and Daume, H. (2008). The infinite hierarchical factor regression model. In: *Advances in Neural Information Processing Systems*, 1321–1328.

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.

Stock, J. H. and Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association* **97**, 1167–1179.

Thung, K. H., Wee, C., Yap, P. and Shen, D. (2014). Neurodegenerative disease diagnosis using incomplete multi-modality data via matrix shrinkage and completion. *NeuroImage* **91**, 386–400.

West, M. (2003). Bayesian regression analysis in the "large p, small n" paradigm. *Bayesian Statistics* **7**, 723–732.

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.

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

Zhu, H., Shen, D., Peng, X. and Liu, L. Y. (2017). MWPCR: Multiscale weighted principal component regression for high-dimensional prediction. *Journal of the American Statistical Association* **112**, 1009–1021.

Department of Statisitcs, Yunnan University, Kunming 650091, China.

E-mail: zyqznl2010@126.com

Department of Statistics, Yunnan University, Kunming 650091, China.

E-mail: nstang@ynu.edu.cn

Department of Statistics University of Illinois at Urbana-Champaign, 725 S. Wright Street Champaign, IL 61820, USA.

E-mail: anniequ@illinois.edu

(Received January 2018; accepted May 2018)