Annals of the Institute of Statistical Mathematics (2023) 75:463–492
https://doi.org/10.1007/s10463-022-00851-5

Matrix completion under complex survey sampling

Xiaojun Mao[1] · Zhonglei Wang[2] · Shu Yang[3]

Received: 16 November 2021 / Revised: 12 August 2022 / Accepted: 17 August 2022 /
Published online: 19 September 2022

Abstract

Multivariate nonresponse is often encountered in complex survey sampling, and simply ignoring it leads to erroneous inference. In this paper, we propose a new matrix completion method for complex survey sampling. Different from existing works either conducting row-wise or column-wise imputation, the data matrix is treated as a whole which allows for exploiting both row and column patterns simultaneously. A column-space-decomposition model is adopted incorporating a low-rank structured matrix for the finite population with easy-to-obtain demographic information as covariates. Besides, we propose a computationally efficient projection strategy to identify the model parameters under complex survey sampling. Then, an augmented inverse probability weighting estimator is used to estimate the parameter of interest, and the corresponding asymptotic upper bound of the estimation error is derived. Simulation studies show that the proposed estimator has a smaller mean squared error than other competitors, and the corresponding variance estimator performs well. The proposed method is applied to assess the health status of the U.S. population.

Keywords Asymptotic upper bound · Augmented inverse probability weighting estimator · Low-rank structure · Missingness at random

✉ Zhonglei Wang
wangzl@xmu.edu.cn

Xiaojun Mao
maoxj@sjtu.edu.cn

Shu Yang
syang24@ncsu.edu

[1] School of Mathematical Sciences, Ministry of Education Key Laboratory of Scientific and Engineering Computing, Shanghai Jiao Tong University, Shanghai 200240, People's Republic of China

[2] Wang Yanan Institute for Studies in Economics and School of Economics, Xiamen University, Xiamen 361005, Fujian, People's Republic of China

[3] Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA

1 Introduction

Survey sampling serves as a golden standard for estimating population parameters. However, survey data analysis becomes increasingly challenging due to the inevitable multivariate nonresponse (Keiding and Louis, 2016), leading to complex "Swiss cheese" patterns in the sample. This occurs due to item nonresponse, when individuals provide answers to partial but not all questions. Moreover, response rates may vary across questions, and some of them are low. This phenomenon is not an exception but is universally encountered in practice (Elliott and Valliant, 2017). Inference ignoring the nonresponse is questionable (Rubin, 1976).

1.1 Existing works

Imputation is widely used to handle item nonresponse, and existing methods for multivariate missingness can be categorized into two types: row-wise imputation and column-wise imputation. Multiple imputation (Clogg et al., 1991; Fay, 1992; Kim et al., 2006; Meng, 1994; Nielsen, 2003; Rubin, 1976; Wang and Robins, 1998; Yang and Kim, 2016) can be viewed as a row-wise imputation method, and it models the joint distribution of all variables and generates imputed values by a posterior predictive distribution. However, multiple imputation suffers from model misspecification and is computationally intensive, especially for large surveys. Hot deck imputation (Andridge and Little, 2010; Chen and Shao, 2000; Fuller and Kim, 2005; Kim and Fuller, 2004), on the other hand, is a column-wise imputation method. For unit i with nonresponse y_{ij} of the jth question, hot deck imputation searches among the units responding the jth question (referred to as donors for the jth question), and imputes y_{ij} by the responses from its neighbors based on a certain distance metric. Although hot deck imputation is easy to implement, it is hard to find a good distance metric, and it is also computationally inefficient for large surveys since it is conducted to impute the nonresponses for each question.

Compared with existing methods using either parametric models or a prespecified distance, we treat the sample data matrix as a whole and apply matrix completion (Cai and Zhou, 2016; Candès and Recht, 2009; Koltchinskii et al., 2011; Mazumder et al., 2010; Negahban and Wainwright, 2012; Robin et al., 2020) for imputation, which exploits both row and column patterns of the sample data matrix simultaneously. Imputation by matrix completion has gained growing attention in survey sampling due to its flexibility. Davenport et al. (2014) proposed 1-bit matrix completion for sample data matrices with binary responses. Zhang et al. (2020) adopted a classical probabilistic matrix factorization for imputation. Sengupta et al. (2021) explored machine learning algorithms for social surveys. However, neither the sampling mechanism nor the demographic covariates were incorporated in the existing works.

1.2 Our contribution

In this paper, we adopt a column-space-decomposition model (Mao et al., 2019) for the sample data matrix incorporating easy-to-obtained demographic data as covariates. Besides, sampling weights are involved in the proposed risk function to adjust for the selection bias due to complex survey sampling. Most works in the matrix completion literature assume uniform missingness (or generally missingness completely at random), which, however, is unlikely to hold for survey data. Instead, the response probabilities are estimated for the proposed method. After imputing missing values in the sample data matrix, we adopt an augmented inverse probability weighting (AIPW) estimator (Qin et al., 2017; Robins et al., 1995) to estimate the population parameters, and the asymptotic error bounds of the AIPW estimator are also established.

The proposed method differs from existing works in the following aspects. First, instead of proposing an identification condition on the sample data matrix as Mao et al. (2019), we propose it for the finite population model, which is more appropriate for survey sampling; see (2) and its discussion for details. Since only a sample is available, the proposed identification condition cannot be used to guarantee the uniqueness of the estimated model parameters. To circumvent this identification issue, we have *innovatively* proposed a projection technique, and the theoretical properties of the proposed method have been investigated as well; see Sects. 2.2–2.3 and Theorem 1 for details. Second, in the literature of matrix completion, existing works mainly focused on imputing missing values for the target matrix, and researchers are often interested in the average squared loss of a “completed matrix”. However, under survey sampling, we are more interested in estimating the population parameters, such as the population mean of the responses to a specific question. Thus, we propose to use matrix completion as an intermediate tool to impute missing values in the sample data matrix. To alleviate the bias of the estimated model parameters, an AIPW estimator is used to estimate the parameters of interest. Please notice that in order to guarantee a design-unbiased estimator for the population parameters, we need to incorporate the uncertainty due to the sampling design, which is also new in the literature of matrix completion. Besides, we have also considered a plug-in variance estimator for the AIPW estimator, but up to our knowledge, there is no variance estimator in the literature of matrix completion. Third, using matrix completion for imputing missing values is new under survey sampling. Different from existing works either conducting row-wise or column-wise imputation when dealing with missing data problems, the data matrix is treated as a whole which allows for exploiting both row and column patterns simultaneously. Besides, we have compared the matrix completion techniques with existing works in Sect. 2.4. Numerical results also demonstrate that the proposed estimator outperforms its competitors in terms of estimation efficiency, and the relative bias of the plug-in variance estimator is also small.

The proposed method achieves the following advantages:

- First, it is computationally efficient compared with the multiple imputation and hot deck imputation, especially for large surveys. Based on the column-space-decomposition model, we have modified the objective function so that a closed-form solution is available to recover the sample data matrix, and only one singular value decomposition (SVD) of an $n \times L$ matrix is required, where n is the sample size, and L is the number of questions in the sample.

- Second, it is a multi-purpose imputation method. We impute all nonresponse simultaneously, and it is particularly attractive for a comprehensive analysis by the imputed sample data matrix.
- Third, compared with parametric methods, we only assume a low-rank structure to achieve robustness; see Remark 1 in Sect. 2 for details.
- Last but not least, we do not assume any restricted nonresponse missing patterns (Fletcher Mercaldo and Blume, 2018; Molenberghs et al., 1998) for the survey data, so the proposed method can be widely applied to handle multivariate nonresponse under complex survey sampling.

The rest of the article is structured as follows. Section 2 provides the basic setup and estimation procedure of the proposed method. Section 3 establishes the theoretical properties of the proposed method. Simulation studies are conducted in Sect. 4 to illustrate the advantage of the proposed method compared with other competitors, and the performance of the variance estimator is tested as well. Section 5 presents the analysis of the National Health and Nutrition Examination Survey (NHANES) Questionnaire Data. Some concluding remarks are given in Sect. 6. Additional materials are relegated to the appendices.

2 Basic setup

2.1 Model

Consider a finite population $U_N = \{(\boldsymbol{x}_i, \boldsymbol{y}_i) : i = 1, \dots, N\}$ of size N, where $\boldsymbol{x}_i^{\mathrm{T}} = (x_{i1}, \dots, x_{id}) \in \mathbb{R}^d$ is a covariate vector of length d associated with the ith unit, and $\boldsymbol{y}_i^{\mathrm{T}} = (y_{i1}, \dots, y_{iL}) \in \mathbb{R}^L$ is the response of interest for L questions. The goal is to estimate $\theta_j = N^{-1} \sum_{i=1}^N y_{ij}$ for $j = 1, \dots, L$.

Assume that the finite population is a realization of the following super-population model,

$$\boldsymbol{Y}_N = \boldsymbol{A}_N + \boldsymbol{\epsilon}_N, \tag{1}$$

where $\boldsymbol{Y}_N = (y_{ij}) \in \mathbb{R}^{N \times L}$, $\boldsymbol{A}_N \in \mathbb{R}^{N \times L}$ represents the structural component, $\boldsymbol{\epsilon}_N = (\epsilon_{ij}) \in \mathbb{R}^{N \times L}$ is a matrix of independent errors with $E(\epsilon_{ij}) = 0$ and $E(\epsilon_{ij}^2) < \sigma_0^2$ for $i = 1, \dots, N$ and $j = 1, \dots, L$, and σ_0^2 is a constant with respect to N and L. Following Candès and Recht (2009), we assume that $\boldsymbol{A}_N$ has a low-rank structure due to underlying clusters of individuals and sections of questions; also see Davenport and Romberg (2016), Robin et al. (2020) and van der Linden and Hambleton (2013) for details. To further incorporate the covariates $\boldsymbol{X}_N = (x_{ij}) \in \mathbb{R}^{N \times d}$, we consider the following column-space-decomposition model (Mao et al., 2019),

$$\boldsymbol{A}_N = \boldsymbol{X}_N \boldsymbol{\beta}^* + \boldsymbol{B}_N^*, \tag{2}$$

where $\boldsymbol{\beta}^* = (\beta_{ij})$ is a $d \times L$ coefficient matrix, and $\boldsymbol{B}_N^* = (b_{ij})$ is an $N \times L$ low-rank matrix, inherited from the low-rank structure of $\boldsymbol{A}_N$. To avoid identification issues for $(\boldsymbol{\beta}^*, \boldsymbol{B}_N^*)$, we assume that $\boldsymbol{X}_N^{\mathrm{T}} \boldsymbol{B}_N^* = \mathbf{0}$; see Sect. 2.2 for details. For simplicity of notation, we omit the dependence of the elements on N in the matrices.

Remark 1 Under survey sampling, the finite population is often assumed to be a random sample from a super-population model; refer to Section 2.2 of Chen et al. (2020), Section 1.3.1 of Fuller (2009), Section 1 of Tan (2013) and Section 2 of Wu (2003) for details. Different from existing works, we consider an additional fixed effect $\boldsymbol{B}_N^*$ in the super-population model (2). For example, in the NHANES Questionnaire Data (https://www.cdc.gov/nchs/nhanes), there exist different sections of questions, such as health and nutrition status and education, and it is reasonable to assume a fixed effect for the responses in the same section. Besides, there may also exist common effects for respondents sharing certain hidden common characters. Instead of specifying $\boldsymbol{B}_N^*$ parametrically, we only assume a low-rank structure for it to achieve robustness. There have been several works taking specific distributions of responses into account (Alaya and Klopp, 2019; Fan et al., 2019; Robin et al., 2020) under general setups, but it is beyond our scope to investigate those methods under complex survey sampling.

In practice, it is both time-consuming and expensive to conduct a census, and survey sampling serves as a golden standard to estimate population parameters. Assume that a sample of size n is generated by a probability sampling design (Fuller, 2009, Chapter 1). For $i = 1, \dots, N$, let $\{I_i : i = 1, \dots, N\}$ be the sampling indicators, and $\pi_i = E(I_i \mid U_N)$ be the corresponding inclusion probability, where $I_i = 1$ if the ith unit is sampled and 0 otherwise, and the expectation is taken with respect to the sampling design conditional on the finite population. Without loss of generality, assume that the first n subjects of the finite population are sampled, and denote $\boldsymbol{M}_n$ and $\boldsymbol{M}_N$ to be generic sample and population data matrices, respectively. If the sample data were fully observed, we could use the following Horvitz–Thompson estimator (Horvitz and Thompson, 1952) to estimate θ_j:

$$\widehat{\theta}_j = \frac{1}{N} \sum_{i=1}^{N} \frac{I_i}{\pi_i} y_{ij}. \tag{3}$$

It follows that $\widehat{\theta}_j$ is a design-unbiased estimator of θ_j given the finite population, that is, $E(\widehat{\theta}_j \mid U_N) = \theta_j$.

However, nonresponse is common in survey sampling (Keiding and Louis, 2016), and ignoring the nonresponse leads to erroneous inference. In the presence of nonresponse in the sample data matrix $\boldsymbol{Y}_n$, we propose to impute the nonresponse simultaneously by the following risk function:

$$R^*(\boldsymbol{\beta}, \boldsymbol{B}_n) = \frac{1}{NL}\sum_{i=1}^{N}\frac{I_i}{\pi_i}\sum_{j=1}^{L}\left\{\frac{r_{ij}}{p_{ij}}y_{ij} - (\boldsymbol{X}_N\boldsymbol{\beta})_{ij} - b_{ij}\right\}^2 \tag{4}$$

$$= \frac{1}{NL}\left\|\boldsymbol{D}_n^{-1/2}\left(\boldsymbol{R}_n \circ \boldsymbol{P}_n^{\dagger} \circ \boldsymbol{Y}_n - \boldsymbol{X}_n\boldsymbol{\beta} - \boldsymbol{B}_n\right)\right\|_F^2, \tag{5}$$

where r_{ij} is the response indicator with response probability p_{ij} associated with y_{ij}, $(\boldsymbol{M})_{ij} = m_{ij}$ for a generic matrix $\boldsymbol{M} = (m_{ij})$, "$\circ$" is the Hadamard product with $\boldsymbol{A} \circ \boldsymbol{B} = (a_{ij}b_{ij})$ for two matrices $\boldsymbol{A} = (a_{ij})$ and $\boldsymbol{B} = (b_{ij})$ of the same dimension, $\boldsymbol{R}_n = (r_{ij}) \in \{0, 1\}^{n \times L}$ is the response indicator matrix with $r_{ij} = 1$ if y_{ij} is observed and $r_{ij} = 0$ otherwise, $\boldsymbol{P}_n^{\dagger} = (p_{ij}^{-1}) \in \mathbb{R}^{n \times L}$, $\|\boldsymbol{M}\|_F = (\sum_{i=1}^{n}\sum_{j=1}^{L} m_{ij}^2)^{1/2}$ is the Frobenius norm of an $n \times L$ matrix $\boldsymbol{M} = (m_{ij})$, and $\boldsymbol{D}_n = \mathrm{diag}(\pi_1, \ldots, \pi_n)$ is a diagonal matrix with π_i being the (i, i)th entry. Ideally, the risk function is associated with estimation of $(\boldsymbol{\beta}^*, \boldsymbol{B}_n^*)$, where $\boldsymbol{B}_n^*$ is the sample counterpart of $\boldsymbol{B}_N^*$. If the sampling mechanism is non-informative, there is no need to adjust sampling weights for estimating $(\boldsymbol{\beta}^*, \boldsymbol{B}_n^*)$ in (4). Adjusting for sampling weights, however, achieves two goals. First, the expectation of (4) is the population risk function

$$R(\boldsymbol{\beta}, \boldsymbol{B}) = \frac{1}{NL}E\left\|\boldsymbol{R}_N \circ \boldsymbol{P}_N^{\dagger} \circ \boldsymbol{Y}_N - \boldsymbol{X}_N\boldsymbol{\beta} - \boldsymbol{B}\right\|_F^2, \tag{6}$$

so we target for estimating parameters for the population data matrix not for the sample data matrix, where $\boldsymbol{R}_N$ and $\boldsymbol{P}_N^{\dagger}$ are the corresponding population versions of $\boldsymbol{R}_n$ and $\boldsymbol{P}_n^{\dagger}$, respectively. Second, it allows for informative sampling, under which the empirical risk function without sampling weights is biased of the population risk function; refer to Pfeffermann (1993) for details.

Remark 2 Another commonly used population risk function which leads to unbiased minimizer $(\boldsymbol{\beta}^*, \boldsymbol{B}_N^*)$ would be $E\|\boldsymbol{R}_N \circ \boldsymbol{P}_N^* \circ (\boldsymbol{Y}_N - \boldsymbol{X}_N\boldsymbol{\beta} - \boldsymbol{B})\|_F^2$ where $\boldsymbol{P}_N^* = (p_{ij}^{-1/2}) \in \mathbb{R}^{N \times L}$. In this paper, we adopt the former one so that we can completely separate the loss into two parts by utilizing the orthogonality between $\boldsymbol{\beta}^*$ and $\boldsymbol{B}_N^*$. Due to the existence of $\boldsymbol{R}_N \circ \boldsymbol{P}_N^*$, we cannot eliminate the inner product term for the later risk function, when separating the $\boldsymbol{X}_N\boldsymbol{\beta}$ and $\boldsymbol{B}$ parts.

2.2 Non-identifiability of $(\boldsymbol{\beta}^*, \boldsymbol{B}_n^*)$

In the population risk function (6), $\boldsymbol{X}_N^{\mathrm{T}}\boldsymbol{B}_N^* = \boldsymbol{0}$ guarantees that $(\boldsymbol{\beta}^*, \boldsymbol{B}_N^*)$ is identifiable and the unique minimizer of (6); see Proposition 1 of Mao et al. (2019) for details. Moreover, the decomposition of $\boldsymbol{A}_N$ into $\boldsymbol{X}_N\boldsymbol{\beta}^* \in \mathcal{C}(\boldsymbol{X}_N)$ and $\boldsymbol{B}_N^* \in \mathcal{N}(\boldsymbol{X}_N)$ gives benefits for showing theoretical properties of the estimators and encourages an efficient algorithm allowing for a closed-form solution, where $\mathcal{C}(\boldsymbol{X}_N)$ is the linear space spanned by the columns of $\boldsymbol{X}_N$, and $\mathcal{N}(\boldsymbol{X}_N)$ is the orthogonal complement of $\mathcal{C}(\boldsymbol{X}_N)$. However, the same decomposition technique may fail to guarantee identification of the model parameters in the sample risk function $R^*(\boldsymbol{\beta}, \boldsymbol{B}_n)$ in (5) since $(\boldsymbol{D}_n^{-1/2}\boldsymbol{X}_n)^{\mathrm{T}}(\boldsymbol{D}_n^{-1/2}\boldsymbol{B}_n) = \boldsymbol{X}_n^{\mathrm{T}}\boldsymbol{D}_n^{-1}\boldsymbol{B}_n$ may not be a zero matrix. Even for simple random sampling with $\pi_i = n/N$ for $i = 1, \ldots, N$, we cannot ensure $\boldsymbol{X}_n^{\mathrm{T}}\boldsymbol{D}_n^{-1}\boldsymbol{B}_n = Nn^{-1}\boldsymbol{X}_n^{\mathrm{T}}\boldsymbol{B}_n = \boldsymbol{0}$. Thus, there is no space restriction for both $\boldsymbol{\beta}$ and $\boldsymbol{B}_n$ in $R^*(\boldsymbol{\beta}, \boldsymbol{B}_n)$. Specifically, for any $(\boldsymbol{\beta}, \boldsymbol{B}_n)$ and nonzero $\boldsymbol{\beta}_1$, we always have $R^*(\boldsymbol{\beta}, \boldsymbol{B}_n) = R^*(\boldsymbol{\beta} + \boldsymbol{\beta}_1, \boldsymbol{B}_n - \boldsymbol{X}_n\boldsymbol{\beta}_1)$.

To deal with the lack of identifiability, we decompose

$$
\begin{aligned}
&\boldsymbol{D}_n^{-1/2}(\boldsymbol{R}_n \circ \boldsymbol{P}_n^{\dagger} \circ \boldsymbol{Y}_n - \boldsymbol{X}_n\boldsymbol{\beta} - \boldsymbol{B}_n) \\
&\quad = \boldsymbol{D}_n^{-1/2}\left(\boldsymbol{R}_n \circ \boldsymbol{P}_n^{\dagger} \circ \boldsymbol{Y}_n\right) - \boldsymbol{D}_n^{-1/2}\boldsymbol{X}_n\boldsymbol{\beta} - \mathcal{P}_{\boldsymbol{D}_n^{-1/2}\boldsymbol{X}_n}(\boldsymbol{D}_n^{-1/2}\boldsymbol{B}_n) \\
&\qquad - \mathcal{P}^{\perp}_{\boldsymbol{D}_n^{-1/2}\boldsymbol{X}_n}(\boldsymbol{D}_n^{-1/2}\boldsymbol{B}_n),
\end{aligned}
$$

where $\mathcal{P}_{\boldsymbol{D}_n^{-1/2}\boldsymbol{X}_n} = \boldsymbol{D}_n^{-1/2}\boldsymbol{X}_n(\boldsymbol{X}_n^{\mathrm{T}}\boldsymbol{D}_n^{-1}\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^{\mathrm{T}}\boldsymbol{D}_n^{-1/2}$, $\mathcal{P}^{\perp}_{\boldsymbol{D}_n^{-1/2}\boldsymbol{X}_n} = \boldsymbol{I} - \mathcal{P}_{\boldsymbol{D}_n^{-1/2}\boldsymbol{X}_n}$ and $\boldsymbol{I}$ is the $n \times n$ identity matrix. Denote

$$
\boldsymbol{\beta}^{*\prime} = \boldsymbol{\beta}^* + (\boldsymbol{X}_n^{\mathrm{T}}\boldsymbol{D}_n^{-1}\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n^{\mathrm{T}}\boldsymbol{D}_n^{-1}\boldsymbol{B}_n^*, \quad \boldsymbol{B}_n^{*\prime} = \mathcal{P}^{\perp}_{\boldsymbol{D}_n^{-1/2}\boldsymbol{X}_n}(\boldsymbol{D}_n^{-1/2}\boldsymbol{B}_n^*).
$$

Then, we have $\boldsymbol{B}_n^{*\prime} \in \mathcal{N}(\boldsymbol{D}_n^{-1/2}\boldsymbol{X}_n)$, so we can decompose the objective function $R^*(\boldsymbol{\beta}, \boldsymbol{B}_n)$ as

$$
\begin{aligned}
R^*(\boldsymbol{\beta}, \boldsymbol{B}_n) &= R^*(\boldsymbol{\beta}', \boldsymbol{B}_n') \\
&= \frac{1}{NL}\left[\left\| \mathcal{P}_{\boldsymbol{D}_n^{-1/2}\boldsymbol{X}_n}\left\{\boldsymbol{D}_n^{-1/2}\left(\boldsymbol{R}_n \circ \boldsymbol{P}_n^{\dagger} \circ \boldsymbol{Y}_n\right)\right\} - \boldsymbol{D}_n^{-1/2}\boldsymbol{X}_n\boldsymbol{\beta}'\right\|_F^2\right. \\
&\qquad \left. + \left\| \mathcal{P}^{\perp}_{\boldsymbol{D}_n^{-1/2}\boldsymbol{X}_n}\left\{\boldsymbol{D}_n^{-1/2}\left(\boldsymbol{R}_n \circ \boldsymbol{P}_n^{\dagger} \circ \boldsymbol{Y}_n\right)\right\} - \boldsymbol{B}_n'\right\|_F^2\right].
\end{aligned}
$$

It can be seen that $\boldsymbol{\beta}^{*\prime}$ and $\boldsymbol{B}_n^{*\prime}$ are the unique minimizers of $R^*(\boldsymbol{\beta}', \boldsymbol{B}_n')$. Although $\boldsymbol{\beta}^*$ and $\boldsymbol{B}_n^*$ cannot be uniquely determined, we ensure that $\boldsymbol{X}_n\boldsymbol{\beta}^{*\prime} + \boldsymbol{D}_n^{1/2}\boldsymbol{B}_n^{*\prime} = \boldsymbol{X}_n\boldsymbol{\beta}^* + \boldsymbol{B}_n^*$, which is sufficient to estimate the parameters of interest θ_j for $j = 1, \ldots, L$. Therefore, in what follows, we focus on estimating $\boldsymbol{\beta}^{*\prime}$ and $\boldsymbol{B}_n^{*\prime}$.

2.3 Estimation of $\boldsymbol{\beta}^{*\prime}$ and $\boldsymbol{B}_n^{*\prime}$

Since $\boldsymbol{P}_n = (p_{ij}) \in \mathbb{R}^{n \times L}$ is unknown, we adopt the assumption of missingness at random (Rubin, 1976) and consider a maximum likelihood estimator $\widehat{\boldsymbol{P}}_n$ of $\boldsymbol{P}_n$ based on the following logistic regression model:

$$
p_{ij} = p_{ij}(\boldsymbol{x}_i) = \frac{\exp\left\{\left(1, \boldsymbol{x}_i^{\mathrm{T}}\right)\boldsymbol{\gamma}_{\cdot j}\right\}}{1 + \exp\left\{\left(1, \boldsymbol{x}_i^{\mathrm{T}}\right)\boldsymbol{\gamma}_{\cdot j}\right\}}, \tag{7}
$$

where $\boldsymbol{\gamma}_{\cdot j} \in \mathbb{R}^{d+1}$ is the parameter vector specific for the jth column of $\boldsymbol{Y}_n$. Once $\widehat{\boldsymbol{P}}_n$ is estimated, we consider

$$\widehat{R}^*(\boldsymbol{\beta}', \boldsymbol{B}_n') = \frac{1}{NL}\Bigg[\left\|\mathcal{P}_{\boldsymbol{D}_n^{-1/2}\boldsymbol{X}_n}\left\{\boldsymbol{D}_n^{-1/2}\left(\boldsymbol{R}_n \circ \widehat{\boldsymbol{P}}_n^{\dagger} \circ \boldsymbol{Y}_n\right)\right\} - \boldsymbol{D}_n^{-1/2}\boldsymbol{X}_n\boldsymbol{\beta}'\right\|_F^2 \\ + \left\|\mathcal{P}^{\perp}_{\boldsymbol{D}_n^{-1/2}\boldsymbol{X}_n}\left\{\boldsymbol{D}_n^{-1/2}\left(\boldsymbol{R}_n \circ \widehat{\boldsymbol{P}}_n^{\dagger} \circ \boldsymbol{Y}_n\right)\right\} - \boldsymbol{B}_n'\right\|_F^2\Bigg],$$

where $\widehat{\boldsymbol{P}}_n^{\dagger}$ is the matrix containing the inverse of the estimated response probabilities. Since $\boldsymbol{\beta}'$ and $\boldsymbol{B}_n'$ are high-dimensional, directly minimizing $\widehat{R}^*(\boldsymbol{\beta}', \boldsymbol{B}_n')$ would often result in over-fitting. To avoid such an issue, we incorporate penalty terms for those two parameters and consider

$$(\widehat{\boldsymbol{\beta}}', \widehat{\boldsymbol{B}}_n') = \underset{\substack{\boldsymbol{\beta}' \in \mathbb{R}^{d\times L} \\ \boldsymbol{B}_n' \in \mathcal{N}(\boldsymbol{D}_n^{-1/2}\boldsymbol{X}_n)}}{\arg\min} \quad \widehat{R}^*(\boldsymbol{\beta}', \boldsymbol{B}_n') \\ + \tau_1\|\boldsymbol{\beta}'\|_F^2 + \tau_2\left\{\alpha\|\boldsymbol{B}_n'\|_* + (1-\alpha)\|\boldsymbol{B}_n'\|_F^2\right\}, \tag{8}$$

where $\|\boldsymbol{M}\|_* = \mathrm{trace}(\sqrt{\boldsymbol{M}^{\mathrm{T}}\boldsymbol{M}})$ is the nuclear norm of a generic matrix $\boldsymbol{M}$, and τ_1, $\tau_2 > 0$ along with $0 \le \alpha \le 1$ are regularization parameters. Since $\boldsymbol{B}_N^*$ is assumed to be low-rank, $\boldsymbol{B}_n^*$ is also low-rank and $\mathrm{rank}(\boldsymbol{B}_n^{*\prime}) = \mathrm{rank}(\boldsymbol{B}_n^*)$. Similar to the rank sum norm, the nuclear norm also encourages a low-rank solution; see Candès and Recht (2009) for details. The two additional Frobenius norm penalties for $\boldsymbol{\beta}'$ and $\boldsymbol{B}_n'$ are applied to improve finite sample performance (Zou and Hastie, 2005; Harchaoui et al., 2012; Li et al., 2012; Sun and Zhang, 2012; Kim et al., 2015; Mao et al., 2019).

It is essentially a ridge regression problem to estimate $\boldsymbol{\beta}'$, and we have

$$\widehat{\boldsymbol{\beta}}' = \left(\boldsymbol{X}_n^{\mathrm{T}}\boldsymbol{D}_n^{-1}\boldsymbol{X}_n + NL\tau_1\boldsymbol{I}\right)^{-1}\boldsymbol{X}_n^{\mathrm{T}}\boldsymbol{D}_n^{-1}\left(\boldsymbol{R}_n \circ \widehat{\boldsymbol{P}}_n^{\dagger} \circ \boldsymbol{Y}_n\right).$$

To obtain $\widehat{\boldsymbol{B}}_n'$, following the same argument in Proposition 2 of Mao et al. (2019), we can extend the searching domain for $\boldsymbol{B}_n' \in \mathcal{N}(\boldsymbol{D}_n^{-1/2}\boldsymbol{X}_n)$ in the minimization problem (8) to be $\boldsymbol{B}_n' \in \mathbb{R}^{n\times L}$. This allows us to express the solution $\widehat{\boldsymbol{B}}_n'$ in a closed form. Let $\boldsymbol{U\Sigma V}^{\mathrm{T}}$ be the SVD of a generic matrix $\boldsymbol{M}$, where $\boldsymbol{\Sigma} = \mathrm{diag}(\{\sigma_i\})$. For $c > 0$, define a singular value soft-thresholding operator $\mathcal{T}_c$ by $\mathcal{T}_c(\boldsymbol{M}) = \boldsymbol{U}\mathrm{diag}(\{(\sigma_i - c)_+\})\boldsymbol{V}^{\mathrm{T}}$, where $x_+ = \max(x, 0)$. It can be shown that the solution $\widehat{\boldsymbol{B}}_n'$ in (8) is

$$\widehat{\boldsymbol{B}}_n' = \frac{1}{1 + (1-\alpha)NL\tau_2}\mathcal{T}_{\alpha NL\tau_2/2}\left[\mathcal{P}^{\perp}_{\boldsymbol{D}_n^{-1/2}\boldsymbol{X}_n}\left\{\boldsymbol{D}_n^{-1/2}\left(\boldsymbol{R}_n \circ \widehat{\boldsymbol{P}}_n^{\dagger} \circ \boldsymbol{Y}_n\right)\right\}\right].$$

Following the common practice in matrix completion (Mazumder et al., 2010), we obtain tuning parameters τ_1, τ_2 and α by a 5-fold cross validation procedure. Cross validation is widely adopted to choose the tuning parameters; see a voluminous literature such as Bi et al. (2017), Liu et al. (2020), Mao et al. (2019, 2021) and Mazumder et al. (2010). After obtaining $(\widehat{\boldsymbol{\beta}}', \widehat{\boldsymbol{B}}'_n)$, an estimator of $\boldsymbol{A}_n$ is

$$\widehat{\boldsymbol{A}}_n = \boldsymbol{X}_n\widehat{\boldsymbol{\beta}}' + \boldsymbol{D}_n^{1/2}\widehat{\boldsymbol{B}}'_n. \tag{9}$$

2.4 Comparison with some existing approaches

It is worth comparing the proposed matrix completion method with existing approaches for imputation under survey sampling. The multiple imputation (Rubin, 1978) assumes a joint model of $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ and uses all available variables for imputation. However, fully parametric modeling is sensitive to model misspecification, and the computation may be cumbersome when the dimension of $\boldsymbol{y}_i$ is large. To impute each missing item, hot deck imputation chooses an observed datum as a “donor” based on a specific distance using some fully observed auxiliary information. For hot deck imputation, an underlying regression model, $f_j(\boldsymbol{x}_i)$, is (implicitly) assumed for the item y_{ij}. Therefore, only the fully observed auxiliary information $\boldsymbol{x}_i$ is used for imputing y_{ij} but not y_{ik} with $k \neq j$.

For the proposed approach, we do not make restrictive parametric model assumptions. Instead, the low-rank structure of $\boldsymbol{A}_N$ suggests a general decomposition of $\boldsymbol{A}_N$ to be $\boldsymbol{A}_N = \boldsymbol{U}_N\boldsymbol{V}_N^{\mathrm{T}}$, where $\boldsymbol{U}_N \in \mathbb{R}^{N\times r_{A_N}}$ and $\boldsymbol{V}_N \in \mathbb{R}^{L\times r_{A_N}}$ are two hidden matrices. Due to the low-rank assumption, we have $r_{\boldsymbol{A}_N} \ll N$ and $r_{\boldsymbol{A}_N} \ll L$. In our column-space-decomposition model, we enforce part of the hidden matrix $\boldsymbol{U}_N$ to be a fully observed matrix $\boldsymbol{X}_N \in \mathbb{R}^{N\times d}$ and denote the corresponding part in $\boldsymbol{V}_N$ to be $\boldsymbol{\beta}^*$, where $\boldsymbol{\beta}^*$ is just a different notation and still totally unspecified. Thus, the decomposition could be written as $\boldsymbol{A}_N = (\boldsymbol{X}_N, \boldsymbol{U}_N^*)(\boldsymbol{\beta}^*, \boldsymbol{V}_N^*)^{\mathrm{T}}$ with $\boldsymbol{B}_N^* = \boldsymbol{U}_N^*\boldsymbol{V}_N^{*\,\mathrm{T}}$. In a general setting, the only restriction for $\boldsymbol{U}_N^*$ is $\mathrm{rank}(\boldsymbol{X}_N, \boldsymbol{U}_N^*) = r_{\boldsymbol{A}_N}$, which means that each column of $\boldsymbol{U}_N^*$ cannot be fully expressed by the columns in $\boldsymbol{X}_N$. However, it allows for $\mathrm{cor}(\boldsymbol{X}_N, \boldsymbol{U}_N^*) \neq \boldsymbol{0}$. Then, it is difficulty to identify the hidden matrix $\boldsymbol{U}_N^*$ under the general setting. Thus, we restrict the column space of $\boldsymbol{U}_N^*$ to be orthogonal to the column space of $\boldsymbol{X}_N$. Fortunately, the number of covariates d is usually fixed and $d \ll r_{\boldsymbol{A}_N}$, so we would not lose too much freedom for $\boldsymbol{U}_N^*$.

2.5 Estimation of $\boldsymbol{\theta}_j$

After imputation, we can estimate θ_j by the Horvitz–Thompson estimator (3) applied to the imputed dataset. However, it is well known that the estimated low-rank matrix $\widehat{\boldsymbol{B}}'_n$ is biased when n is finite due to regularization (Carpentier and Kim, 2018; Chen et al., 2019; Foucart et al., 2017; Mazumder et al., 2010). Therefore, the resulting imputation estimator is biased. Researchers have proposed different procedures to alleviate or eliminate the bias. Mazumder et al. (2010) suggested a post-processing step by re-estimating the estimated singular values without any theoretical guarantee. Foucart et al. (2017) proposed an algorithm based on projection onto the max-norm ball to de-bias the estimator under non-uniform and deterministic sampling patterns. Carpentier and Kim (2018) considered an estimator using an iterative hard thresholding method and showed that the entry-wise bias is small when the sampling design is Gaussian. More recently, Chen et al. (2019) developed a de-biasing procedure using a similar idea to de-biasing LASSO estimators and showed nearly optimal properties for the resulting estimator. Despite these advances in literature, the scenarios considered by the existing works are restricted to deterministic sampling, Gaussian sampling or missing completely at random, which are not applicable under survey sampling.

We use a simple strategy borrowing the idea from the AIPW literature (Qin et al., 2017; Robins et al., 1994) and consider an AIPW estimator of θ_j as

$$\widehat{\theta}_{j,AIPW} = \frac{1}{\widehat{N}} \sum_{i=1}^{N} \frac{I_i}{\pi_i} \left\{ \frac{r_{ij}(y_{ij} - \widehat{a}_{ij})}{\widehat{p}_{ij}} + \widehat{a}_{ij} \right\}, \tag{10}$$

and it solves $\sum_{i=1}^{N} I_i \pi_i^{-1} \{ r_{ij} \widehat{p}_{ij}^{-1} (y_{ij} - \theta) - \widehat{p}_{ij}^{-1} (r_{ij} - \widehat{p}_{ij})(\widehat{a}_{ij} - \theta) \} = 0$, where where $\widehat{N} = \sum_{i=1}^{N} I_i \pi_i^{-1}$ is the estimated population size, and $\widehat{p}_{ij}$ and $\widehat{a}_{ij}$ are the (i, j)th element of $\widehat{\boldsymbol{P}}_n$ and $\widehat{\boldsymbol{A}}_n$, respectively. Under regularity conditions, it can be shown that

$$\widehat{\theta}_{j,AIPW} = \frac{1}{N} \sum_{i=1}^{N} \frac{I_i}{\pi_i} \left\{ \frac{r_{ij}(y_{ij} - \widehat{a}_{ij})}{p_{ij}} + \widehat{a}_{ij} \right\} + o_p(1), \tag{11}$$

when the response model (7) is correctly specified. Since the leading term of (11) is unbiased for θ_j conditional on the estimated $\{\widehat{a}_{ij} : i = 1, \dots, n; j = 1, \dots, L\}$ if the response model is correctly specified, $\widehat{\theta}_{j,AIPW}$ is asymptotically unbiased. In the simulation study, we have compared the proposed estimator (10) with other estimators, and numerical results show that the proposed estimator is more efficient.

3 Asymptotic properties

In this section, we first study the asymptotic properties of the estimator $\widehat{\boldsymbol{A}}_n$ in (9) under the logistic regression model (7). Then, we establish the average convergence rate of $\widehat{\theta}_{j,AIPW} - \theta_j$ for $j = 1, \dots, L$.

For asymptotic inference, we follow the framework of Isaki and Fuller (1982) and assume that both the population size N and the sample size n diverge to infinity. Let $\|\boldsymbol{M}\| = \sigma_{\max}(\boldsymbol{M})$ and $\|\boldsymbol{M}\|_\infty = \max_{i,j} |m_{ij}|$ be the spectral and the maximum norms of a generic matrix $\boldsymbol{M} = (m_{ij})$, respectively. We use "$\asymp$" to represent the asymptotic equivalence in order, that is, $a_n \asymp b_n$ is equivalent to $a_n = O(b_n)$ and $b_n = O(a_n)$. The technical conditions are delegated to Appendix A.

For any $\delta_\sigma > 0$, positive constants C_d, C_g, C and $t \in (d + 3, +\infty)$, define

$$\Delta\left(\delta_\sigma, t\right) = \max\left\{N^{1/2}n^{-1}L^{-1}\log^{1/2}(n)p_{\min}^{-1/2},\right.$$
$$\left.N^{1/2}n^{-5/4}L^{-1/4}\log^{1/2}(L)\log^{\delta_\sigma/4}(n)t^{1/2}p_{\min}^{-3/2}\right\}, \tag{12}$$

and $\eta_{n,L}(\delta_\sigma, t) = 4/(n+L) + 4C_d t\exp\{-t/2\} + 4/L + C\log^{-\delta_\sigma}(n)$, where $p_{\min}$ is positive constant satisfying $p_{\min} \le \min\{p_{i,j}\}$; see Condition C6 in Appendix A for details. We can verify that $\lim_{t\to\infty}\{\lim_{n,L\to\infty}\eta_{n,L}(\delta_\sigma, t)\} = 0$, and we implicitly assume that $L \to \infty$ in the following analysis. If $n^{1/2}L^{-3/2}\log(n)p_{\min}^2 \ge (d+3)$, by choosing t such that

$$d + 3 < t < n^{1/2}L^{-3/2}\log(n)p_{\min}^2, \tag{13}$$

we can show $\sup_t \Delta(\delta_\sigma, t) \asymp N^{1/2}n^{-1}L^{-1}\log^{1/2}(n)p_{\min}^{-1/2}$, which is denoted by $\Delta(\delta_\sigma)$. The requirement $n^{1/2}L^{-3/2}\log(n)p_{\min}^2 \ge (d+3)$ is easy to fulfill as long as n is large enough. A similar condition is also considered by Robin et al. (2020) and Zhang et al. (2020).

Theorem 1 *Assume Conditions C1–C6 given Appendix A, $p_{\min}^{-1} = O(L\log^{-1}(n+L))$ and the logistic model (7) hold. Choose t as (13), $\tau_1 \asymp N^{-1}nL^{-1}\log^{-1/2}(n)\Delta(\delta_\sigma)$, $1-\alpha \asymp (nL)^{-1}$, $\tau_2 \asymp p_{\min}^{-3/2}N^{-1}n^{1/4}L^{-1/4}\log^{1/2}(L)\log^{\delta_\sigma/3}(n)$ in (8) for any $\delta_\sigma > 0$. Then, for some positive constant C_1 and C_2, with probability at least $1-\eta_{n,L}(\delta_\sigma, t)$, we have*

$$(mL)^{-1}\left\|\widehat{\boldsymbol{\beta}}' - \boldsymbol{\beta}^{*\prime}\right\|_F^2 \le C_1 r_{\boldsymbol{B}_N}L^{-1}\log(n)p_{\min}^{-1},$$
$$(nL)^{-1}\left\|\widehat{\boldsymbol{B}}_n' - \boldsymbol{B}_n^{*\prime}\right\|_F^2 \le C_2 r_{\boldsymbol{B}_N}Nn^{-1}L^{-1}\log(n)p_{\min}^{-1}.$$

A proof of Theorem 1 is given in Appendix C.1. Theorem 1 implies that as $\lim_{t\to\infty}\{\lim_{n,L\to\infty}\eta_{n,L}(\delta_\sigma, t)\} = 0$,

$$\|\widehat{\boldsymbol{B}}_n' - \boldsymbol{B}_n^{*\prime}\|_F^2 = O_p\left\{r_{\boldsymbol{B}_N}Nn^{-1}L^{-1}\log(n)p_{\min}^{-1}\right\},$$
$$(mL)^{-1}\|\widehat{\boldsymbol{\beta}}' - \boldsymbol{\beta}^{*\prime}\|_F^2 = O_p\left\{r_{\boldsymbol{B}_N}L^{-1}\log(n)p_{\min}^{-1}\right\}(nL)^{-1}.$$

As we pointed out in Sect. 2.2, even with the knowledge of $(\boldsymbol{\beta}^{*\prime}, \boldsymbol{B}_n^{*\prime})$, we cannot recover $(\boldsymbol{\beta}^*, \boldsymbol{B}_n^*)$ exactly. Fortunately, we have $\widehat{\boldsymbol{A}}_n = \boldsymbol{X}_n\widehat{\boldsymbol{\beta}}' + \boldsymbol{D}_n^{1/2}\widehat{\boldsymbol{B}}_n'$, which enables us to derive the asymptotic bound for $(nL)^{-1}\|\widehat{\boldsymbol{A}}_n - \boldsymbol{A}_n\|_F^2$ given in the following theorem.

Theorem 2 *Assume that the conditions in Theorem 1 hold. For a positive constant C_3, with probability at least $1-\eta_{n,L}(\delta_\sigma, t)$, we have*

$$(nL)^{-1}\left\|\widehat{\boldsymbol{A}}_n-\boldsymbol{A}_n\right\|_F^2 \le C_3 r_{\boldsymbol{B}_N} L^{-1}\log(n)p_{\min}^{-1}.$$

A brief proof of Theorem 2 can be found in Appendix C.2. The term $(nL)^{-1}\|\widehat{\boldsymbol{A}}_n-\boldsymbol{A}_n\|_F^2$ has the same order with upper bound of $(mL)^{-1}\|\widehat{\boldsymbol{\beta}}'-\boldsymbol{\beta}^{*\prime}\|_F^2$. To ensure the convergence of $(nL)^{-1}\|\widehat{\boldsymbol{A}}_n-\boldsymbol{A}_n\|_F^2$, we only require that $n=O\{\exp(r_{\boldsymbol{B}_N}^{-1}Lp_{\min})\}$ which is quite mild. Under survey sampling, it is reasonable to assume that $p_{\min}\asymp 1$, especially when the participants are awarded. Thus, the assumption that $p_{\min}^{-1}=O(L\log^{-1}(n+L))$ is easy to fulfill as long as L is large enough. Besides, the convergence rate for $(nL)^{-1}\|\widehat{\boldsymbol{A}}_n-\boldsymbol{A}_n\|_F^2$ can be simplified as $r_{\boldsymbol{B}_N}L^{-1}\log(n)$ if $p_{\min}\asymp 1$.

Theorem 3 *Assume that the conditions in Theorem 1 and Condition C7 in Appendix A hold and* $p_{\min}\asymp 1$. *Then, we have*

$$L^{-1}\sum_{j=1}^{L}(\widehat{\theta}_{j,AIPW}-\theta_j)^2=O_p\{r_{\boldsymbol{B}_N}L^{-1}\log(n)\}.$$

A proof for Theorem 3 is given in Appendix C.3. By Theorem 3, the mean squared difference between $\widehat{\theta}_{j,AIPW}$ and θ_j among the L questions is bounded by $O_p\{r_{\boldsymbol{B}_N}L^{-1}\log(n)\}$. To ensure the convergence of $L^{-1}\sum_{j=1}^{L}(\widehat{\theta}_{j,AIPW}-\theta_j)^2$, similarly as before, we only require that $n=o\{\exp(r_{\boldsymbol{B}_N}^{-1}L)\}$ which is quite mild. Although we have discussed the average convergence rate for the AIPW estimator $\widehat{\theta}_{j,AIPW}$ for $j=1,\dots,L$ in Theorem 3, it is not easy to improve the convergence rate for each $\widehat{\theta}_{j,AIPW}$. Up to our knowledge, there do not exist column-wise or element-wise convergence results in the literature of matrix completion. Thus, it is hard to establish a limiting distribution for each $\widehat{\theta}_{j,AIPW}$, and this topic will be pursued in the future. Besides, an unbiased variance estimator of $\widehat{\theta}_{j,AIPW}$ is also intractable, and we propose to use a plug-in variance estimator instead. For example, under Poisson sampling, a plug-in variance estimator for the AIPW estimator is

$$\begin{aligned}\widehat{V}_{j,poi}=N^{-2}\sum_{i=1}^{N}\frac{I_i r_{ij}(1-p_{ij})}{\pi_i^2 p_{ij}^2}(y_{ij}-\widehat{a}_{ij})^2\\ +N^{-2}\sum_{i=1}^{N}\frac{I_i r_{ij}(1-\pi_i)}{p_{ij}\pi_i^2}(y_{ij}-\widehat{\theta}_{j,AIPW})^2.\end{aligned}$$

We also discuss the plug-in variance estimators for other commonly used sampling designs in Appendix D.

For the NHANES and other national surveys, a stratified multi-stage sampling design is used, and only the final weight is available. Then, it is impossible to derive the above plug-in variance estimator. For such a design, we consider the modified balanced repeated replication (Rao and Shao, 1999) for the variance estimation; a brief discussion about this method is provided in Appendix E.

4 Simulation

In this section, we compare the proposed estimator with its competitors under stratified two-stage cluster sampling. The corresponding R code can be found in https://github.com/mxjki/Matrix_Completion_For_Complex_Survey_with_Multivariate_Missingness. See Section S1 of the Supplementary Material for simulation under single-stage sampling.

We use (1) and (2) to generate a finite population U_N, consisting of $H = 15$ strata. For $h = 1, \dots, H$, generate $N_h \sim \mathrm{Poi}(10) + 20$ and $N_{hi} \sim \mathrm{Poi}\{10(\zeta_h + \xi_{hi})\} + 40$, where N_h is number of clusters of the hth stratum, N_{hi} is the size of the ith cluster for $i = 1, \dots, N_h$, $\zeta_h \sim \mathrm{Exp}(3)$ is the stratum effect, $\xi_{hi} \sim \mathrm{Exp}(3)$ is the cluster effect independent with the stratum effect, $\mathrm{Poi}(\lambda)$ is a Poisson distribution with parameter λ, and $\mathrm{Exp}(\lambda)$ is the exponential distribution with rate parameter λ. The population size is $N = \sum_{h=1}^{H} \sum_{i=1}^{N_h} N_{hi} = 19{,}658$ in the simulation study.

We set the dimension of covariate to be $d = 20$, number of questions to be $L = 500$, and the rank of $\boldsymbol{B}_N^*$ to be $m = 10$. The elements of $\boldsymbol{x}_{hij}$ are independently generated by $\mathcal{N}(0.5, 1^2) + (\zeta_h + \xi_{hi})/2$ for $j = 1, \dots, N_{hi}$, and the elements of $\boldsymbol{\beta}^*$ are independently generated by $\mathcal{N}(0.5, 1^2)$. To generate $\boldsymbol{B}_N^*$, we first generate an $N \times m$ matrix $\boldsymbol{B}_L$ and an $m \times L$ matrix $\boldsymbol{B}_R$, where the elements of $\boldsymbol{B}_L$ is generated by $\mathcal{N}(1, 3^2) + (\zeta_h + \xi_{hi})/3$ for the row vector corresponds to $\boldsymbol{x}_{hij}$, and the elements of $\boldsymbol{B}_R$ are independently generated by $\mathcal{N}(1, 3^2)$. Then, $\boldsymbol{B}_N^* = \mathcal{P}_{\boldsymbol{X}_N}^{\perp} \boldsymbol{B}_L \boldsymbol{B}_R$. For the random errors in (1), we consider the following two scenarios. For Scenario I, we generate $\epsilon_{ij} \sim N(0, 12^2)$ for $i = 1, \dots, N$ and $j = 1, \dots, L$. Scenario I corresponds to the case where the stratum and cluster effects of the study variable are explained by the covariates the random errors are randomly generated, and it corresponds to the model we have assumed in (1). For Scenario II, $\epsilon_{kj} = 10\zeta'_{h(k)} + 9\xi_{hi(k)} + \eta_{kj}$ for $k = 1, \dots, N$ and $j = 1, \dots, L$, where $h(k)$ and $hi(k)$ are the stratum and cluster indexes for the kth element, respectively. Under the second scenario, however, random errors involve additional stratum and cluster effects, so they are not independent. We consider the Scenario II to test the robustness of the proposed method.

Within each stratum, we consider a two-stage sampling design. In the first stage, two clusters are selected using probability-proportional-to-size sampling with selection probability proportional to the cluster size. Within each selected cluster, simple random sampling is conducted to draw a sample of size n_c. We consider two different sample sizes $n_c = 10$ and $n_c = 20$ for the second stage sampling. The following estimation methods are compared:

I Hot deck imputation (HDI). For each item with $r_{ij} = 0$, we use y_{kj} as the imputed value, where $\boldsymbol{x}_k$ is nearest to $\boldsymbol{x}_j$ among $\{\boldsymbol{x}_l : r_{lj} = 1\}$ in terms of the Euclidean norm. Treating the imputed values as observed ones, we estimate θ_j by (3).

Table 1 Summary of MSE for different estimation methods under stratified two-stage sampling and different scenarios

		Stat	HDI	MI	IPW	AIPWLR	AIPWMC	Full
I	$n_c = 10$	Mean	24.95	6.54	7.14	7.53	4.59	3.49
		SE	10.84	2.00	2.69	2.91	1.58	1.24
	$n_c = 20$	Mean	23.91	3.06	3.49	3.55	2.25	1.72
		SE	10.81	0.86	1.30	1.37	0.78	0.64
II	$n_c = 10$	Mean	25.14	6.28	7.26	7.65	5.55	3.71
		SE	10.63	1.88	2.65	2.90	1.91	1.28
	$n_c = 20$	Mean	24.42	3.08	3.75	3.81	2.75	2.01
		SE	10.47	0.80	1.36	1.43	0.89	0.66

Table 2 Mean and the standard error of the relative bias for the 500 variance estimators using the modified balanced repeated replication method

	Scenario I		Scenario II	
	Mean	SE	Mean	SE
$n_c = 10$	0.01	0.12	0.09	0.16
$n_c = 20$	0.04	0.09	0.07	0.13

II Multiple imputation (MI). We adopt the multivariate imputation by chained equations (MICE) by van Buuren and Groothuis-Oudshoorn (2011). MICE fully specifies the conditional distribution for the missing data and uses a posterior predictive distribution to generate imputed values for the nonresponse items. However, it is impossible for MICE to impute all missing responses in $\boldsymbol{Y}_n$ at the same time due to the computational issues. For comparison, we only use the first 20 items of $\boldsymbol{Y}_n$ to specify the conditional distribution for MICE and generate imputed values for the corresponding nonresponse. Then, we can use (3) to estimate θ_j.

III Inverse probability weighting method (IPW). For $j = 1, \dots, L$, a logistic regression model (7) is fitted. Then, θ_j is estimated by $\widehat{\theta}_{j,IPW} = N^{-1} \sum_{i=1}^{n} r_{ij}\widehat{p}_{ij}^{-1} y_{ij}$.

IV AIPW estimator using a linear regression model (AIPWLR). For $j = 1, \dots, L$, consider the following linear regression model:

$$y_{ij} = \phi_{0j} + \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\phi}_{1j}, \tag{14}$$

and the parameters in (14) are estimated by

$$(\widehat{\phi}_{0j}, \widehat{\boldsymbol{\phi}}_{1j}) = \underset{(\phi_{0j}, \boldsymbol{\phi}_{1j})}{\arg\min} \sum_{i=1}^{n} \frac{r_{ij}}{\pi_i \widehat{p}_{ij}} (y_{ij} - \phi_{0j} - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\phi}_{1j})^2.$$

Then, we can use the AIPW estimator based on the linear model (14) to estimate θ_j.

V AIPW estimator using the proposed method (IPWMC).

For the response model (7), we use $\mathrm{logit}(p_{ij}) = \gamma_0 + \sum_{j=1}^{3} \gamma_j x_{ij}$, where $\mathrm{logit}(p) = \log(p) - \log(1-p)$, $\gamma_0 \sim \mathcal{N}(0.3, 0.1^2)$ and $\gamma_j \sim \mathcal{N}(-0.1, 0.1^2)$ for $j = 1, \ldots, 3$. The response rate is about 0.53. We also consider the Horvitz–Thompson estimator (Full) in (3) using the fully observed data for comparison.

We conduct 1000 Monte Carlo simulations. Table 1 summarizes the mean and standard error of MSEs of different questions for different methods. Under the stratified two-stage sampling, the AIPW estimator using the proposed method performs best among the estimation methods since it has smallest mean and standard error of MSEs. Although the Horvitz–Thompson estimator is better than the AIPW estimator using the proposed method, it cannot be used in practice in that the fully observed sample is not available. Besides, compared with the case under Scenario II, the AIPW estimator using the proposed method performs better under Scenario I.

We also test the modified balanced repeated replication method (Rao and Shao, 1999) for variance estimation of the AIPW estimator using the proposed method, and the Monte Carlo mean and standard error of the relative bias for the 500 variance estimators is shown in Table 2. The mean of the relative bias of the variance estimators are reasonably small, and the standard error of the relative bias decreases as the sample size increases under both scenarios. Compared with Scenario II, the variance estimator has a smaller relative bias under Scenario I. In addition, the variance estimator is conservative regardless of the sample size and scenarios. Thus, the performance of the modified balanced repeated replication method is satisfactory for estimating the variance of the AIPW estimator using the proposed method.

Remark 3 The simulation shows that the proposed AIPWMC estimator outperforms the IPW counterpart numerically in terms of the mean squared error. However, theoretically, it is hard to compare the asymptotic efficiency of the two estimators in our context. First, even for a parametric model, Qin et al. (2017) pointed out in their Theorem 1 that there are no general results for comparing the asymptotic variance between the IPW and AIPW estimators. Besides, the AIPWMC estimator relies on semiparametric matrix completion, and no theoretical results exist for column-wise or element-wise analysis in the literature of matrix completion, as we have mentioned in the preceding section. Therefore, we will leave the research question as a future research topic.

5 Application

NHANES is a well-structured program to assess the health and nutrition status of children and adults in the United States. The survey combines physical examinations and questionnaires and, therefore, can provide a thorough and detailed health status assessment. Moreover, analyzing annual data provides the trend of health status of the entire population over time, so it is important for policy makers.

A stratified multi-stage sampling has been conducted to obtain the NHANES samples. There are about 15 strata, formed by state-level health-related variables such as death rate and infant mortality rate. The primary sampling unit consists of counties and is selected by probability-proportional-to-size sampling, and some primary sampling units with large measure of size are selected with certainty. The second stage is conducted by selecting area segments, consisting of census blocks based on the 2000 census data. The third stage selects dwelling units, and the fourth one is a selection of eligible members. See Chen et al. (2020) for details about the sampling design. Thus, participants are nationally representative. Data are released in a two-year cycle to guarantee statistically stable estimates, and there are two PSUs selected within each stratum in every two-year sample. The available *Questionnaire Data* dates back to 1999, and the most updated one is the 2015–2016 cycle. This data contains family-level information including food security status as well as individual level information including dietary behavior and alcohol use. The questionnaire has evolved since 1999, and new questions have been added subsequently.

Unfortunately, the analysis of the NHANES is challenging due to the complex study designs and multivariate missingness, and almost all of the health-related questions suffer from missingness. In the 2015–2016 Questionnaire Data, for example, there are 245 questions, which can be answered by all participants. There are 47 demographic questions, among which 21 fully answered, including age, gender, race-ethnicity and education. There are 198 health-related questions, and 136 questions have nonresponse. Moreover, the response rates of 53 questions are less than 0.95.

Remark 4 In the NHANES Questionnaire Data, some questions are skipped. For example, if the response of “Had at least 12 alcohol drinks in any one year?” is “Yes”, the participant may skip the question “Had at least 12 alcohol drinks in the entire lifetime?”. If a question is skipped, it is not regarded as an item non-response. However, it is not the main topic of our study to investigate missingness of “skipped questions”, so we omit them in our analysis. We have carefully reviewed all questions for the NHANES 2015–2016 Questionnaire Data and picked 198 health-related questions, which should not be skipped. We focus ourselves on those questions in this paper, and it would be an interesting topic to study the missingness incorporating the “skipped questions” in the future.

Based on the 198 selected questions from the 2015–2016 cycle, Table 3 shows the number of questions among them and the percentage of questions with response rates less than 0.95 since 1999. Although the number of “non-skipped” questions has increased since 1999, there are more of them with low response rates. Thus, ignoring the missingness may result in more questionable inference.

Table 3 Number of questions (No. qn) among the 198 selected ones from the 2015–2016 cycle and the percent (Pct) of questions with response rates less than 0.95 since 1999

	99–00	01–02	03–04	05–06	07–08	09–10	11–12	13–14	15–16
No. qn	81	108	101	153	167	147	149	175	198
Pct	0.09	0.09	0.12	0.20	0.18	0.23	0.20	0.29	0.27

“99–00” represents the 1999–2000 cycle

Table 4 Estimation results for three questions

Items	Res	Stat	Estimation methods					
			MI	HDI	IPW	AIPWLR	AIPWNI	AIPWMC
I	0.99	Mean	198.79	198.86	197.38	198.79	198	198.60
		SE	–	–	4.10	4.10	–	3.79
II	0.94	Mean	207.79	200.13	198.40	207.83	207.83	201.50
		SE	–	–	28.63	28.68	–	21.94
III	0.86	Mean	7.86	7.16	7.77	7.86	0.40	7.79
		SE	–	–	1.63	1.63	–	1.48

"I" is about the self-reported greatest weights in pounds. "II" is about the money spent on eating out. "III" is about the monthly family income

We apply the proposed method to analyze the longitudinal NHANES Questionnaire Data. The goal is two-fold. First, we are interested in estimating the population mean of health-related questions, such as immunization and diabetes, based on the most updated NHANES 2015–2016 Questionnaire Data. Second, we analyze the longitudinal trend for some selected questions, and the results are shown in Section S2 of the Supplementary Material. We focus on the age group from 20 to 59 since the corresponding participants should answer all of the 198 selected questions. Covariates, including age, gender, marital status, ethnic group and education, are used in the analysis.

For estimating the population mean of each question, we consider estimation methods in Sect. 4, and the covariates are standardized. Since the population size is unavailable, we use $\widehat{N} = \sum_{i=1}^{n} w_i$ instead, where w_i is the sampling weight of the ith subject incorporating the sampling design as well as calibration (Fuller, 2009).

Table 4 shows estimation results for three representative items with response rates 0.99, 0.94 and 0.86, respectively so the selected questions are representative in terms of response rates. We apply the modified balanced repeated replication method to estimate the variance of the corresponding estimators except for the multiple imputation and hot deck imputation. The variance of the multiple imputation is not estimated due to computational complexity. The variance estimation method is not applicable for the hot deck imputation since the imputed values are not affected by the repetition procedure.

When the response rate is high, different estimators are similar. As the response rate decreases, multiple imputation performs similarly as the AIPW estimator using linear regression, and they are different from the remaining three. Compared with others, the estimator by the hot deck imputation is different from the other four, especially when the response rate is below 90%. Besides, the variance of the inverse probability weighting method is similar with the AIPW estimator using linear regression, and the AIPW estimator using the proposed method is more efficient than these two estimators regardless of the response rate.

We analyze the result in Table 4 based on the estimators by the AIPW estimator using the proposed method, and we would analyze the trend with respect to these three questions in the next section. Question I, “Up to the present time, what is the most {you have/SP has} ever weighed?”, is about the self-reported greatest weight in pounds, and the response ranges from 75 to 559. The estimated self-reported greatest weight is about 198.6 pounds for the people in the age group from 20 to 59, and the corresponding standard error is 3.79 pounds. Question II, “During the past 30 days, how much money {did your family/did you} spend on eating out? Please include money spent in cafeterias at work or at school or on vending machines, for all family members. (You can tell me per week or per month.)”, is about the money spent on eating out, and the response value ranges from \$0 to \$3000. The estimated average is \$201.5 with standard error \$21.94. It indicates that people spent about \$201.5 on eating out. The third question, “Monthly family income (reported as a range value in dollars)”, is about the monthly family income, and the response rate of this question is quite low. Instead of reporting the actual income, there are 12 levels for the response: level 1 corresponds to \$0–\$399, and level 12 to more than \$8400. The estimated average is about level 8. Thus, the monthly family income is about \$3750–\$4599.

6 Concluding remarks

We have proposed a new imputation method for survey sampling by assuming a low-rank structure for the super-population model and incorporating fully observed auxiliary information. Asymptotic properties of the proposed method are investigated. One of the major advantages of the proposed method is that we can impute all nonresponse simultaneously for the whole sample data matrix consisting of complex missingness patterns. Two different variance estimators are suggested. Simulation studies demonstrate that the proposed method is more accurate than some commonly used alternatives, including inverse probability weighting method and multiple imputation, for estimating all items, and the variance estimator is satisfactory when the sample size is large.

Our framework can also be extended in the following directions. First, we have considered missingness at random; however, in some situations, the missingness of y_{ij} may depend on its own value, leading to missingness not at random (Rubin, 1976); that is, y_{ij} is also involved in the response probability (7). In this case, we will consider the instrumental variable approach (Wang et al., 2014; Yang et al., 2019) or stringent parametric model assumptions (Chang and Kott, 2008; Kim and Yu, 2011; Tang et al., 2003) for identification and estimation. Second, even though we have proposed an efficient estimator using matrix completion and derived the asymptotic bounds, its asymptotic distribution is not completely developed, which will be our future work. Third, because causal inference of treatment effects can be viewed as a missing data problem, it is intriguing to develop matrix completion to deal with a partially observed confounder matrix, which is ubiquitous in practice but has received little attention in the literature (Yang et al., 2019).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10463-022-00851-5.

Acknowledgements We are grateful to two referees and the Associate Editor for their constructive comments which have greatly improved the paper. Mao is partially supported by NSFC (No.: 12001109, 92046021) and the Science and Technology Commission of Shanghai Municipality grant 20dz1200600. Wang is partially supported by NSFC (No.: 11901487, 72033002) and the Fundamental Scientific Center of National Natural Science Foundation of China Grant No. 71988101. Yang is partly supported by the NSF DMS 1811245, NIH 1R01AG066883 and 1R01ES031651.

Appendix

A Technical conditions

The technical conditions needed for our analysis are given as follows.

C1 (a) The random errors $\{\epsilon_{ij} : i = 1, \ldots, N; j = 1, \ldots, L\}$ in (2) are independently distributed random variables such that $E(\epsilon_{ij}) = 0$ and $E(\epsilon_{ij}^2) = \sigma_{ij}^2 < \infty$ for all i, j. (b) For some finite positive constants c_σ and η, $\max_{i,j} E|\epsilon_{ij}|^l \le \frac{1}{2} l! c_\sigma^2 \eta^{l-2}$ for any positive integer $l \ge 2$.

C2 The inclusion probability satisfies $\pi_i \asymp nN^{-1}$ for $i = 1, \ldots, N$.

C3 The population design matrix $\boldsymbol{X}_N$ is of size $N \times d$ such that $N > d$. Moreover, there exists a positive constant a_x such that $\|\boldsymbol{X}_N\|_\infty \le a_x$ and $\boldsymbol{X}_N^{\mathrm{T}} \boldsymbol{D}_N \boldsymbol{X}_N$ is invertible, where $\boldsymbol{D}_N$ is a diagonal matrix with π_i as its (i, i)th entry. Furthermore, there exists a symmetric matrix $\boldsymbol{S}_X$ with $\sigma_{\min}(\boldsymbol{S}_X) \asymp 1 \asymp \|\boldsymbol{S}_X\|$ such that $n_0^{-1} \boldsymbol{X}_N^{\mathrm{T}} \boldsymbol{D}_N \boldsymbol{X}_N \to \boldsymbol{S}_X$ as $N \to \infty$, where $n_0 = \sum_{i=1}^N \pi_i$ is the expected sample.

C4 There exists a positive constant a such that $\max\{\|\boldsymbol{X}_N \boldsymbol{\beta}^*\|_\infty, \|\boldsymbol{A}_N\|_\infty\} \le a$.

C5 The indicators of observed entries $\{r_{ij} : i = 1, \ldots, N; j = 1, \ldots, L\}$ are mutually independent, $r_{ij} \sim \mathrm{Bern}(p_{ij})$ for $p_{ij} \in (0, 1)$ and are independent of $\{\epsilon_{ij}\}_{i,j=1}^{N,L}$ given $\boldsymbol{X}_N$. Furthermore, for $i = 1, \ldots, N$ and $j = 1, \ldots, L$, $\Pr(r_{ij} = 1 | \boldsymbol{x}_i, y_{ij}) = \Pr(r_{ij} = 1 | \boldsymbol{x}_i)$ follows the logistic regression model (7).

C6 There exists a lower bound $p_{\min} \in (0, 1)$ such that $\min_{i,j}\{p_{ij}\} \ge p_{\min} > 0$, where $p_{\min}$ is allowed to depend on n and L. The number of questions $L \le n$.

C7 The sampling design satisfies that $N^{-1} \sum_{i=1}^N y_i \pi_i^{-1} = O_p(n^{-1/2})$ if $N^{-1} \sum_{i=1}^N y_i^2$ is asymptotically bounded.

Condition C1(a) is a common regularity condition for the measurement errors in $\boldsymbol{\epsilon}_N$, and C1(b) is the Bernstein condition (Koltchinskii et al., 2011). Condition C2 is widely used in survey sampling and regulates the inclusion probabilities of a sampling design (Fuller, 2009). In Condition C3, the requirement $N > d$ is easily met as the number of questions in a survey is usually fixed, and the population size is often larger than the number of questions. As the dimension of $n_0^{-1} \boldsymbol{X}_N^{\mathrm{T}} \boldsymbol{D}_N \boldsymbol{X}_N$ is fixed at $d \times d$, it is mild to assume $\boldsymbol{X}_N^{\mathrm{T}} \boldsymbol{D}_N \boldsymbol{X}_N$ to be invertible, and there exists a symmetric matrix $\boldsymbol{S}_X$ as the limit of $n_0^{-1} \boldsymbol{X}_N^{\mathrm{T}} \boldsymbol{D}_N \boldsymbol{X}_N$. Please notice that we do not assume randomness for generating $\boldsymbol{X}_N$, and it is a common assumption for design-based framework. Furthermore, the sample size is often larger than the number of questions, that is, $n > d$, and it is not hard to show that together with Condition C2, the probability limit of $n^{-1}\boldsymbol{X}_n^{\mathrm{T}}\boldsymbol{X}_n$ is also $\boldsymbol{S}_X$ under regularity conditions. The order of $\sigma_{\min}(\boldsymbol{S}_X)$ and $\|\boldsymbol{S}_X\|$ equals to 1 is due to $\|\boldsymbol{X}_N\|_\infty < \infty$. Condition C4 is also standard in the matrix completion literature (Cai and Zhou, 2016; Koltchinskii et al., 2011; Negahban and Wainwright, 2012). Especially, it is reasonable to assume all the responses are bounded in survey sampling. Condition C5 describes the independent Bernoulli model for the response indicator of observing y_{ij}, where the probability of observation p_{ij} follows the logistic model (7). In Condition C6, the lower bound $p_{\min}$ is allowed to go to 0 with n and L growing. This condition is more general than we need for a typical survey, and $p_{\min} \asymp 1$ suffices. Typically, the number of questions L grows slower than the number of participants n in survey sampling. Thus, the assumption that $L \leq n$ is quite mild. Condition C7 is a mild restriction on the estimator for the population mean, and it can be satisfied under general sampling designs. To get general results, we do not make any assumptions for the asymptotic relationship between the population size N and the sample size n; see Theorem 1 for details. We can make further assumptions for the sample sizes to guarantee certain convergence properties; see the discussion of Theorem 3.

B Lemmas

Under the logistic model (7), together with the results in Mao et al. (2019) and Sweeting (1980), it can been shown that for all $t > d + 3$, there exist some positive constants C_g, C and C_d such that $\Pr\{\sum_{ij}(1/\widehat{p}_{ij} - 1/p_{ij})^2 \geq C_g p_{\min}^{-3} t\} \leq C_d t\exp\{-t/2\}$ $\exp\{-t/2\} + L\max_j \sup_t |\Pr\{\sum_i(1/\widehat{p}_{ij} - 1/p_{ij})^2 \geq t\} - \Pr(\chi^2_{d+1} \geq C_g p_{\min}^{-3} t)|$. Then, $\max_j \sup_t |\Pr\{\sum_i(1/\widehat{p}_{ij} - 1/p_{ij})^2 \geq t\} - \Pr(\chi^2_{d+1} \geq C_g p_{\min}^{-3} t)| \leq L^{-2}$ and $C_d t\exp\{-t/2\}$ is a function independent of n and L such that $\lim_{t\to\infty} t\exp\{-t/2\} = 0$.

Write $\boldsymbol{J}_{ij} = \boldsymbol{e}_i(n_1)\boldsymbol{e}_j^{\mathrm{T}}(n_2)$, where $\boldsymbol{e}_i(n) \in \mathbb{R}^n$ is the standard basis vector with the i-th element being 1 and the rest being 0. Now we present several lemmas.

Lemma 1 *Under Conditions C2 and C3 and Poisson sampling, we have*

$$n^{-1}\boldsymbol{X}_n^{\mathrm{T}}\boldsymbol{X}_n = \boldsymbol{S}_X + o_p(1). \tag{15}$$

Proof of Lemma 1 Denote $\mathbf{e}_i$ to be a column vector of length d with jth element being 1 and others being 0. Recall that $n_0 = \sum_{k=1}^N \pi_k$ is the expected sample size. For $i = 1, \ldots, d$ and $j = 1, \ldots, d$, consider

$$\begin{aligned} E(n_0^{-1}\boldsymbol{e}_i^{\mathrm{T}}\boldsymbol{X}_n^{\mathrm{T}}\boldsymbol{X}_n\boldsymbol{e}_j) &= n_0^{-1}E\left(\sum_{k=1}^N I_k x_{ki}x_{kj}\right) = n_0^{-1}\sum_{i=1}^N x_{ki}\pi_k x_{kj} \\ &= n_0^{-1}\boldsymbol{e}_i^{\mathrm{T}}\boldsymbol{X}_N^{\mathrm{T}}\boldsymbol{D}_N\boldsymbol{X}_N\boldsymbol{e}_j, \end{aligned} \tag{16}$$

where the expectation is take with respect to the sampling design, and $\boldsymbol{x}_i$ is the ith row of $\boldsymbol{X}_N$.

Under Poisson sampling, we have

$$\begin{aligned} \operatorname{var}(n_0^{-1}\boldsymbol{e}_i^{\mathrm{T}}\boldsymbol{X}_n^{\mathrm{T}}\boldsymbol{X}_n\boldsymbol{e}_j) &= n_0^{-2}\sum_{k=1}^{N}\pi_k(1-\pi_k)x_{k,i}x_{k,j} \\ &< n_0^{-2}\sum_{k=1}^{N}\pi_k x_{k,i}x_{k,j} \\ &= n_0^{-1}\left(n_0^{-1}\boldsymbol{e}_i^{\mathrm{T}}\boldsymbol{X}_N^{\mathrm{T}}\boldsymbol{D}_N\boldsymbol{X}_N\boldsymbol{e}_j\right). \end{aligned} \tag{17}$$

By a similar argument in Wang et al. (2019), we can show that $n/n_0 \to 1$ in probability as $N \to \infty$ under Condition C2. By Condition C3, (16) and (17), we have proved Lemma 1. □

Lemma 2 *Let* $\Psi^{(1)} = \sum_{ij} r_{ij}\epsilon_{ij}\boldsymbol{J}_{ij}/(nL\widehat{p}_{ij}\pi_{ij}^{1/2})$. *Under Conditions C1, C5 and C6 and Poisson sampling, for some positive constants* C_1, c_σ, η, δ_σ *and all* $t > d+3$, *we have*

$$\begin{aligned} &\left\|\Psi^{(1)}\right\| \\ &\quad\le C_1 \max\left\{N^{1/2}n^{-1}L^{-1}\log^{1/2}(n)p_{\min}^{-1/2}, N^{1/2}n^{-5/4}L^{-1/4}\log^{1/2}(L)\log^{\delta_\sigma/4}(n)t^{1/2}p_{\min}^{-3/2}\right\} \end{aligned}$$

holds with probability at least $1 - 1/(n+L) - C_d t\exp\{-t/2\} - 1/L - 12c_\sigma^2\eta^2\log^{-\delta_\sigma}(n)$.

Lemma 3 *Let* $\Psi^{(2)} = \sum_{ij} a_{ij}(r_{ij}/p_{ij} - 1)\boldsymbol{J}_{ij}/(nL\pi_{ij}^{1/2})$. *Under Conditions C4–C6 and Poisson sampling, for some positive constants* C_2, *we have*

$$\left\|\Psi^{(2)}\right\| \le C_2 N^{1/2}n^{-1}L^{-1}\log^{1/2}(n)p_{\min}^{-1/2}$$

holds with probability at least $1 - 1/(n+L)$.

Lemma 4 *Let* $\Psi^{(3)} = \sum_{ij} a_{ij}(r_{ij}/\widehat{p}_{ij} - r_{ij}/p_{ij})\boldsymbol{J}_{ij}/(nL\pi_{ij}^{1/2})$. *Under Conditions C4 and C6 and Poisson sampling, for some positive constants* C_3, δ_σ *and all* $t > d+3$, *we have*

$$\left\|\Psi^{(3)}\right\| \le C_3 N^{1/2}n^{-5/4}L^{-1/4}\log^{1/2}(L)\log^{\delta_\sigma/4}(n)t^{1/2}p_{\min}^{-3/2}$$

holds with probability at least $1 - C_d t\exp\{-t/2\} - 1/L$.

It is easy to show Lemma 2–4 by the proof of Lemma S4.1–S4.3 in the supplementary material of Mao et al. (2019).

C Proofs

C.1 Proof of Theorem 1

With the definition of $\Delta(\delta_\sigma, t)$ in (12), under Conditions C1–C6 and Poisson sampling, together with Lemmas 2–4, We have for a positive constant C_0,

$$\left\|\Psi^{(1)}\right\| + \|\Psi^{(2)}\| + \left\|\Psi^{(3)}\right\| \leq C_0\Delta(\delta_\sigma, t),$$

with probability at least $1 - 2/n - 2C_d t\exp\{-t/2\} - 2/L - 12c_\sigma^2\eta^2\log^{-\delta_\sigma}(n)$.

Let $\boldsymbol{X}_n' = \boldsymbol{D}_n^{-1/2}\boldsymbol{X}_n$ and $\eta_{n,L}(\delta_\sigma, t) = 4/(n+L) + 4C_d t\exp\{-t/2\} + 4/L + C\log^{-\delta_\sigma}(n)$ for a positive constant C. By choosing t as (13), $\tau_1 \asymp N^{-1}nL^{-1}\log^{-1/2}(n)\Delta(\delta_\sigma)$ and $\tau_2 \asymp \eta_g^{-1/2}N^{-1}n^{1/4}L^{-1/4}\log^{1/2}(L)\log^{\delta_\sigma/3}(n)$, where $\Delta(\delta_\sigma) = N^{1/2}n^{-1}L^{-1}\log^{1/2}(n)p_{\min}^{-1/2}$ and $1 - \alpha \asymp (nL)^{-1}$ in (8) for any $\delta_\sigma > 0$, together with Condition C2 and Poisson sampling, it follows the same proof with the proof of Corollary 1 in Mao et al. (2019) that, for some constants C_1 and C_2, with probability at least $1 - \eta_{n,L}(\delta_\sigma, t)$,

$$\text{both} \quad \frac{1}{nL}\left\|\boldsymbol{X}_n\widehat{\boldsymbol{\beta}}' - \boldsymbol{X}_n\boldsymbol{\beta}^{*\prime}\right\|_F^2 \quad \text{and} \quad \frac{1}{nL}\left\|\widehat{\boldsymbol{B}}_n' - \boldsymbol{B}_n^{*\prime}\right\|_F^2 \leq C_1 r_{\boldsymbol{B}_N} N n^{-1}L^{-1}\log(n)p_{\min}^{-1}.$$

Thus it is easy to obtain that

$$\frac{1}{mL}\left\|\widehat{\boldsymbol{\beta}}' - \boldsymbol{\beta}^{*\prime}\right\|_F^2 \leq C_2 r_{\boldsymbol{B}_N} L^{-1}\log(n)p_{\min}^{-1},$$

under Condition C3. □

C.2 Proof of Theorem 2

Due to the observations that

$$\left\|\widehat{\boldsymbol{A}}_n - \boldsymbol{A}_n\right\|_F^2 \leq \left\|\boldsymbol{X}_n\widehat{\boldsymbol{\beta}}' - \boldsymbol{X}_n\boldsymbol{\beta}^{*\prime}\right\|_F^2 + \left\|\boldsymbol{D}_n^{-1/2}\widehat{\boldsymbol{B}}_n' - \boldsymbol{D}_n^{-1/2}\boldsymbol{B}_n^{*\prime}\right\|_F^2,$$

together with Theorem 1, it is easy to obtain the result under Condition C2 and Poisson sampling. □

C.3 Proof of Theorem 3

Denote

$$\tilde{\theta}_{j,AIPW} = N^{-1}\sum_{i=1}^{N}\frac{I_i}{\pi_i}\left\{\frac{r_{ij}(y_{ij} - a_{ij})}{p_{ij}} + a_{ij}\right\}, \tag{18}$$

$$\theta_{j,AIPW}^{\dagger} = N^{-1}\sum_{i=1}^{N}\frac{I_i}{\pi_i}\left\{\frac{r_{ij}(y_{ij} - \widehat{a}_{ij})}{p_{ij}} + \widehat{a}_{ij}\right\} \tag{19}$$

for $j = 1, \ldots, L$. The difference between $\widehat{\theta}_{j,AIPW}$ in (10) and $\tilde{\theta}_{j,AIPW}$ in (18) is that we use estimators $\widehat{N}$, $\widehat{p}_{ij}$ and $\widehat{a}_{ij}$ for $\widehat{\theta}_{j,AIPW}$ but use true values N, p_{ij} and a_{ij} for $\tilde{\theta}_{j,AIPW}$.

The difference between $\widehat{\theta}_{j,AIPW}$ and $\theta^{\dagger}_{j,AIPW}$ in (19) is that we use $\widehat{N}$ for $\widehat{\theta}_{j,AIPW}$, but use N for $\theta^{\dagger}_{j,AIPW}$.

First, we prove

$$\tilde{\theta}_{j,AIPW} - \theta_j = O_p(n^{-1/2}). \tag{20}$$

Consider

$$E(\tilde{\theta}_{j,AIPW}) = E\{E(\tilde{\theta}_{j,AIPW} \mid \{I_i\})\} = N^{-1}\sum_{i=1}^{N}\frac{E(I_i)}{\pi_i}(y_{ij}) = \theta_j, \tag{21}$$

where the first equality holds due to $E(r_{ij}) = p_{ij}$. Next, we derive the variance of $\tilde{\theta}_{j,AIPW}$. Specifically, we have

$$\begin{aligned} \operatorname{var}(\tilde{\theta}_{j,AIPW}) &= \frac{1}{N^2}E\left(\operatorname{var}\left[\sum_{i=1}^{N}\frac{I_i}{\pi_i}\left\{\frac{r_{ij}(y_{ij}-a_{ij})}{p_{ij}} + a_{ij}\right\} \mid S\right]\right) \\ &\quad + \frac{1}{N^2}\operatorname{var}\left(E\left[\sum_{i=1}^{N}\frac{I_i}{\pi_i}\left\{\frac{r_{ij}(y_{ij}-a_{ij})}{p_{ij}} + a_{ij}\right\} \mid S\right]\right) = V_{1,j} + V_{2,j}, \end{aligned} \tag{22}$$

where $S = \{I_i \,:\, i = 1, \ldots, N\}$.

Because $E(r_{ij}) = p_{ij}$, we have

$$\operatorname{var}\left[\sum_{i=1}^{N}\frac{I_i}{\pi_i}\left\{\frac{r_{ij}(y_{ij}-a_{ij})}{p_{ij}} + a_{ij}\right\} \mid S\right] = \sum_{i=1}^{N}\frac{I_i(1-p_{ij})}{\pi_i^2 p_{ij}}(y_{ij}-a_{ij})^2.$$

Thus, we have

$$V_{1,j} = \frac{1}{N^2}\sum_{i=1}^{N}\frac{1-p_{ij}}{\pi_i}(y_{ij}-a_{ij})^2 = O_p(n^{-1}), \tag{23}$$

where the last equality holds by Conditions C1, C2 and the strong law of large numbers (Athreya and Lahiri, 2006). Notice that

$$E\left[\sum_{i=1}^{N}\frac{I_i}{\pi_i}\left\{\frac{r_{ij}(y_{ij}-a_{ij})}{p_{ij}} + a_{ij}\right\} \mid S\right] = \sum_{i=1}^{N}\frac{I_i y_{ij}}{\pi_i}.$$

By the models (1)–(2) and Condition C4, we can show that $N^{-1}\sum_{i=1}^{N} y_i^2$ is asymptotically bounded. Thus, by Condition C7, we have

$$V_{2,j} = O_p(n^{-1}) \tag{24}$$

By (21)–(24), we have shown (20).

Next, we show that

$$\frac{1}{L}\sum_{j=1}^{L}(\theta_{j,AIPW}^{\dagger}-\tilde{\theta}_{j,AIPW})^2=O_p\{r_{\boldsymbol{B}_N}L^{-1}\log(n)\}. \tag{25}$$

Consider

$$\theta_{j,AIPW}^{\dagger}-\tilde{\theta}_{j,AIPW}=\frac{1}{N}\sum_{i=1}^{N}\frac{I_i}{\pi_i}\left\{\frac{r_{ij}(y_{ij}-a_{ij})(p_{ij}-\widehat{p}_{ij})}{p_{ij}\widehat{p}_{ij}}+\frac{(r_{ij}-\widehat{p}_{ij})(a_{ij}-\widehat{a}_{ij})}{\widehat{p}_{ij}}\right\}. \tag{26}$$

Consider

$$E\left\{\frac{1}{N}\sum_{i\in S}\pi_i^{-1}\frac{r_{ij}(y_{ij}-a_{ij})}{p_{ij}}\right\}=\frac{1}{N}\sum_{i=1}^{N}(y_{ij}-a_{ij})=O_p(N^{-1/2}), \tag{27}$$

$$\mathrm{var}\left\{\frac{1}{N}\sum_{i\in S}\pi_i^{-1}\frac{r_{ij}(y_{ij}-a_{ij})}{p_{ij}}\right\}=O_p(n^{-1}), \tag{28}$$

where the asymptotic order in (28) holds due to Condition C7 and $N^{-1}\sum_{i=1}^{N}(y_{ij}-a_{ij})^2$ is asymptotically bounded in probability since $\{\epsilon_{ij}:j=1,\dots,N\}$ are independent and their variances are uniformly bounded. By (27) and (28), we have

$$\frac{1}{N}\sum_{i=1}^{N}\frac{I_i}{\pi_i}\frac{r_{ij}(y_{ij}-a_{ij})}{p_{ij}}=O_p(n^{-1/2}). \tag{29}$$

Because the response model (7) for p_{ij} is assumed to be correctly specified, and p_{ij} is bounded away from 0 by Condition C6, we have $\widehat{p}_{ij}^{-1}(p_{ij}-\widehat{p}_{ij})=O_p(1)$ uniformly for $i=1,\dots,N$. Thus, by (29), we have

$$N^{-1}\sum_{i=1}^{N}\frac{I_i}{\pi_i}\frac{r_{ij}(y_{ij}-a_{ij})(p_{ij}-\widehat{p}_{ij})}{p_{ij}\widehat{p}_{ij}}=O_p(n^{-1/2}). \tag{30}$$

By (26) and (30), we have

$$\theta_{j,AIPW}^{\dagger}-\tilde{\theta}_{j,AIPW}=O_p(n^{-1/2})+\frac{1}{N}\sum_{i=1}^{N}\frac{I_i}{\pi_i}\frac{(r_{ij}-\widehat{p}_{ij})(a_{ij}-\widehat{a}_{ij})}{\widehat{p}_{ij}}. \tag{31}$$

Since $\widehat{p}_{ij}=p_{ij}+o_p(1)$, we have

$$\frac{r_{ij}-\widehat{p}_{ij}}{\widehat{p}_{ij}}=\{1+o_p(1)\}\frac{r_{ij}-p_{ij}}{p_{ij}}. \tag{32}$$

Since $p_{ij}\geq p_{\min}>0$ by Condition C6, we have

$$\frac{r_{ij} - \widehat{p}_{ij}}{\widehat{p}_{ij}} = O_p(1) \tag{33}$$

uniformly for $i = 1, \dots, N$.

Thus, by Condition C2, (31) and (33), we have

$$\begin{aligned} \frac{1}{L}\sum_{j=1}^{L}(\theta^{\dagger}_{j,AIPW} - \tilde{\theta}_{j,AIPW})^2 &\leq O_p(n^{-1}) + \frac{C_6 O_p(1)}{Ln^2}\sum_{j=1}^{L}\left\{\sum_{i=1}^{n}(a_{ij} - \widehat{a}_{ij})\right\}^2, \\ &\leq O_p(n^{-1}) + \frac{O_p(1)}{Ln}\sum_{j=1}^{L}\sum_{i=1}^{n}(a_{ij} - \widehat{a}_{ij})^2 \\ &= O_p(n^{-1}) + \frac{O_p(1)}{nL}\left\|\widehat{A}_n - A_n\right\|_F^2, \end{aligned} \tag{34}$$

where $\pi_i \geq C_6^{-1}nN^{-1}$ for $C_6 > 0$ by Condition C2, and we have assumed that the first n subjects are sampled. By (20), (34) and Theorem 2 and the fact that $L \leq n$, we have proved (25).

By Condition C4 and the fact that $E(\epsilon_{ij}^2) < \sigma_0^2$ uniformly, θ_j is uniformly bounded for $j = 1, \dots, L$ in probability. Thus, by (20) and (25), we conclude that

$$\frac{1}{L}\sum_{j=1}^{L}(\theta^{\dagger}_{j,AIPW})^2 = O_p\{r_{\boldsymbol{B}_N}L^{-1}\log(n)\} \tag{35}$$

By Condition C7, we conclude that $\widehat{N}N^{-1} = 1 + O_p(n^{-1/2})$. Consider

$$\frac{1}{L}\sum_{j=1}^{L}(\widehat{\theta}_{j,AIPW} - \theta^{\dagger}_{j,AIPW})^2 = \frac{O_p(n^{-1})}{L}\sum_{j=1}^{L}(\theta^{\dagger}_{j,AIPW})^2 = o_p\{r_{\boldsymbol{B}_N}L^{-1}\log(n)\}, \tag{36}$$

where the first equality holds since $\widehat{N}N^{-1} = 1 + O_p(n^{-1/2})$ uniformly for $j = 1, \dots, L$, and the second equality holds by (35). Thus, by (34) and (36), we have proved Theorem 3. □

D Plug-in variance estimators

When deriving the plug-in variance estimator, we ignore the variability for estimating $\widehat{a}_{ij}$. First, we consider the plug-in variance estimator under Poisson sampling. For $j = 1, \dots, L$, let

$$g_j(\theta) = \frac{1}{N}\sum_{i=1}^{N}\frac{I_i}{\pi_i}\left\{\frac{r_{ij}(y_{ij} - a_{ij})}{p_{ij}} + a_{ij} - \theta\right\}$$

be the estimating function for the AIPW estimator with $\widehat{a}_{ij}$ replaced by a_{ij}. Let $\tilde{\theta}_{j,AIPW}$ solves $g_j(\theta) = 0$, and we use a variance estimator of $\tilde{\theta}_{j,AIPW}$ to approximate that of $\widehat{\theta}_{j,AIPW}$.

It can be shown that $\tilde{\theta}_{j,AIPW} - \theta_j = O_p(n^{-1/2})$, so we have

$$0 = g_j(\tilde{\theta}_{j,AIPW}) = g_j(\theta_j) + g_j'(\theta_j)(\tilde{\theta}_{j,AIPW} - \theta_j) + o_p(n^{-1/2}), \tag{37}$$

where $g_j'(\theta) = -N^{-1} \sum_{i=1}^{N} I_i \pi_i^{-1}$ is the derivative of $g_j(\theta)$. By a similar argument for (27)–(28), we can show that $g_j'(\theta_j) \to -1$ in probability. Besides, by (37), we have

$$(\tilde{\theta}_{j,AIPW} - \theta_j) = -\{g_j'(\theta_j)\}^{-1} g_j(\theta_j) + o_p(n^{-1/2}) = g_j(\theta_j) + o_p(n^{-1/2}).$$

Thus, the variance of $\tilde{\theta}_{j,AIPW}$ can be estimated by the one of $g_j(\theta_j)$.

Consider

$$\begin{aligned} \mathrm{var}\{g_j(\theta_j)\} &= N^{-2} E\left(\mathrm{var}\left[\sum_{i=1}^{N} \frac{I_i}{\pi_i} \left\{ \frac{r_{ij}(y_{ij} - a_{ij})}{p_{ij}} + a_{ij} - \theta_j \right\} \mid S \right] \right) \\ &\quad + N^{-2} \mathrm{var}\left(E\left[\sum_{i=1}^{N} \frac{I_i}{\pi_i} \left\{ \frac{r_{ij}(y_{ij} - a_{ij})}{p_{ij}} + a_{ij} - \theta_j \right\} \mid S \right] \right) \\ &= V_{1,j} + V_{2,j}. \end{aligned} \tag{38}$$

Since $E(r_{ij}) = p_{ij}$, we have

$$\mathrm{var}\left[\sum_{i=1}^{N} \frac{I_i}{\pi_i} \left\{ \frac{m_{ij}(y_{ij} - a_{ij})}{p_{ij}} + a_{ij} - \theta_j \right\} \mid S \right] = \sum_{i=1}^{N} \frac{I_i(1 - p_{ij})}{\pi_i^2 p_{ij}} (y_{ij} - a_{ij})^2.$$

Thus, we have

$$V_{1,j} = N^{-2} \sum_{i=1}^{N} \frac{1 - p_{ij}}{\pi_i p_{ij}} (y_{ij} - a_{ij})^2, \tag{39}$$

and it can be estimated by

$$\widehat{V}_{1,j} = N^{-2} \sum_{i=1}^{N} \frac{r_{ij}(1 - p_{ij})}{\pi_i^2 p_{ij}^2} (y_{ij} - a_{ij})^2. \tag{40}$$

Notice that

$$E\left[\sum_{i=1}^{N} \frac{I_i}{\pi_i} \left\{ \frac{r_{ij}(y_{ij} - a_{ij})}{p_{ij}} + a_{ij} - \theta_j \right\} \mid S \right] = \sum_{i=1}^{N} \frac{I_i}{\pi_i} (y_{ij} - \theta_j).$$

Under Poisson sampling,

$$\mathrm{var}\left(E\left[\sum_{i=1}^{N} \frac{I_i}{\pi_i} \left\{ \frac{r_{ij}(y_{ij} - a_{ij})}{p_{ij}} + a_{ij} - \theta_j \right\} \mid S \right] \right) = \sum_{i=1}^{N} \frac{1 - \pi_i}{\pi_i} (y_{ij} - \theta_j)^2.$$

Thus,

$$V_{2,j} = N^{-2} \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i}(y_{ij} - \theta_j)^2, \tag{41}$$

and it can be estimated by

$$\widehat{V}_{2,j} = N^{-2} \sum_{i=1}^{N} \frac{I_i r_{ij}(1-\pi_i)}{p_{ij}\pi_i^2} y_{ij}^2. \tag{42}$$

By (38)–(42) and plugging in $\widehat{a}_{ij}$ and $\widehat{\theta}_{j,AIPW}$ for a_{ij} and θ_j, respectively, the plug-in variance estimator for $\widehat{\theta}_{j,AIPW}$ is

$$\begin{aligned}\widehat{V}_{j,poi} = {} & N^{-2} \sum_{i=1}^{N} \frac{I_i r_{ij}(1-p_{ij})}{\pi_i^2 p_{ij}^2}(y_{ij} - \widehat{a}_{ij})^2 \\ & + N^{-2} \sum_{i=1}^{N} \frac{I_i r_{ij}(1-\pi_i)}{p_{ij}\pi_i^2}(y_{ij} - \widehat{\theta}_{j,AIPW})^2\end{aligned}$$

under Poisson sampling.

Use a similar argument, we can show that the plug-in variance estimator is

$$\begin{aligned}\hat{V}_{j,srs} = {} & n^{-2} \sum_{i=1}^{n} \frac{I_i r_{ij}(1-\hat{p}_{ij})}{\hat{p}_{ij}^2}(y_{ij} - \widehat{a}_{ij})^2 + n^{-1}(1 - nN^{-1}) \\ & \left[n^{-1} \sum_{i=1}^{N} \frac{I_i r_{ij}(y_{ij} - \widehat{\theta}_{j,AIPW})^2}{\hat{p}_{ij}} - \left(n^{-1} \sum_{i=1}^{N} \frac{I_i r_{ij}(y_{ij} - \widehat{\theta}_{j,AIPW})}{\hat{p}_{ij}} \right)^2 \right]\end{aligned}$$

under simple random sampling, and the one is

$$\begin{aligned}\hat{V}_{j,pps} = {} & N^{-2} \sum_{i \in S} \frac{r_{ij}(1-\hat{p}_{ij})}{(nq_i)^2 \hat{p}_{ij}^2}(y_{ij} - \widehat{a}_{ij})^2 + \{n(n-1)\}^{-1} \\ & \left\{ \sum_{i \in S} \frac{r_{ij}(y_{ij} - \widehat{\theta}_{j,AIPW})^2}{\hat{p}_{ij} q_i^2} - n^{-1} \left(\sum_{i \in S} \frac{r_{ij}(y_{ij} - \widehat{\theta}_{j,AIPW})}{\hat{p}_{ij} q_i} \right)^2 \right\}\end{aligned}$$

under probability-proportional-to-size sampling, where S is the index set of the sample, and q_i is the selection probability of the ith element.

E Balanced repeated replication method

Consider a stratified multi-stage sampling design with two clusters selected per stratum for the first stage. Denote w_{hik} to be the survey weight associated with y_{hik}, the kth sample element in the ith cluster of the hth stratum. The basic idea of the modified balanced repeated replication is to use the same estimation method based on the "reconstruct" the survey weight, that is,

$$w_{hik}^{(r)} = w_{hik}(1 + \epsilon\delta_{rh}),$$

where $\epsilon \in (0, 1)$ is a predefined constant, and $\delta_{rh} = 1$ or $\delta_{rh} = -1$ for the rth repetition. A set of R repetitions is said to be balanced if $\sum_{r=1}^{R} \delta_{rh}\delta_{rh'} = 0$ for $h \neq h'$. The $R \times H$ matrix $(\delta_{rh})_{R\times H}$ can be obtained from a Hadamard matrix. Please check Rao and Shao (1999) for details about the modified balanced repeated replication method. In the simulation study and real data application, we choose $\epsilon = 1/2$ and $R = H$.

References

Alaya, M. Z., Klopp, O. (2019). Collective matrix completion. *Journal of Machine Learning Research, 20*(148), 1–43.

Andridge, R. R., Little, R. J. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review, 78*(1), 40–64.

Athreya, K. B., Lahiri, S. N. (2006). *Measure theory and probability theory*. New York: Springer.

Bi, X., Qu, A., Wang, J., Shen, X. (2017). A group-specific recommender system. *Journal of the American Statistical Association, 112*(519), 1344–1353.

Cai, T. T., Zhou, W.-X. (2016). Matrix completion via max-norm constrained optimization. *Electronic Journal of Statistics, 10*(1), 1493–1525.

Candès, E. J., Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational Mathematics, 9*(6), 717–772.

Carpentier, A., Kim, A. K. (2018). An iterative hard thresholding estimator for low rank matrix recovery with explicit limiting distribution. *Statistica Sinica, 28*, 1371–1393.

Chang, T., Kott, P. S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika, 95*, 555–571.

Chen, J., Shao, J. (2000). Nearest neighbor imputation for survey data. *Journal of Official Statistics, 16*, 113–131.

Chen, T. C., Clark, J., Riddles, M. K., Mohadjer, L. K., Fakhouri, T. H. I. (2020). National health and nutrition examination survey, 2015–2018: Sample design and estimation procedures. *National Center for Health Statistics. Vital Health Stat, 2*(184), 1–26.

Chen, Y., Fan, J., Ma, C., Yan, Y. (2019). Inference and uncertainty quantification for noisy matrix completion. *Proceedings of the National Academy of Sciences, 116*(46), 22931–22937.

Chen, Y., Li, P., Wu, C. (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association, 115*(532), 2011–2021.

Clogg, C. C., Rubin, D. B., Schenker, N., Schultz, B., Weidman, L. (1991). Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression. *Journal of the American Statistical Association, 86*, 68–78.

Davenport, M. A., Romberg, J. (2016). An overview of low-rank matrix recovery from incomplete observations. *IEEE Journal of Selected Topics in Signal Processing, 10*(4), 608–622.

Davenport, M. A., Plan, Y., van den Berg, E., Wootters, M. (2014). 1-bit matrix completion. *Information and Inference, 3*(3), 189–223.

Elliott, M. R., Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science, 32*(2), 249–264.

Fan, J., Gong, W., Zhu, Z. (2019). Generalized high-dimensional trace regression via nuclear norm regularization. *Journal of Econometrics, 212*(1), 177–202.

Fay, R. E. (1992). When are inferences from multiple imputation valid? *Proceedings of the survey research methods section of the American Statistical Association*, 227–232. American Statistical Association.

Fletcher Mercaldo, S., Blume, J. D. (2018). Missing data and prediction: The pattern submodel. *Biostatistics, 21*(2), 236–252.

Foucart, S., Needell, D., Plan, Y., Wootters, M. (2017). De-biasing low-rank projection for matrix completion. *Wavelets and sparsity XVII*, Vol. 10394, p. 1039417. International Society for Optics and Photonics.

Fuller, W. A. (2009). *Sampling statistics*. Hoboken, NJ: Wiley.

Fuller, W. A., Kim, J. K. (2005). Hot deck imputation for the response model. *Survey Methodology, 31*, 139.

Harchaoui, Z., Douze, M., Paulin, M., Dudik, M., Malick, J. (2012). Large-scale image classification with trace-norm regularization. *2012 IEEE conference on computer vision and pattern recognition*, 3386–3393. IEEE.

Horvitz, D. G., Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association, 47*(260), 663–685.

Isaki, C. T., Fuller, W. A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association, 77*, 89–96.

Keiding, N., Louis, T. A. (2016). Perils and potentials of self-selected entry to epidemiological studies and surveys. *Journal of the Royal Statistical Society*: *Series A (Statistics in Society)*, *179*, 319–376.

Kim, E., Lee, M., Oh, S. (2015). Elastic-net regularization of singular values for robust subspace learning. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 915–923.

Kim, J. K., Fuller, W. (2004). Fractional hot deck imputation. *Biometrika, 91*, 559–578.

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

Kim, J. K., Brick, J., Fuller, W. A., Kalton, G. (2006). On the bias of the multiple-imputation variance estimator in survey sampling. *Journal of the Royal Statistical Society*: *Series B (Statistical Methodology)*, *68*, 509–521.

Koltchinskii, V., Lounici, K., Tsybakov, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Annals of Statistics, 39*(5), 2302–2329.

Li, H., Chen, N., Li, L. (2012). Error analysis for matrix elastic-net regularization algorithms. *IEEE Transactions on Neural Networks and Learning Systems, 23*(5), 737–748.

Liu, W., Mao, X., Wong, R. K. W. (2020). Median matrix completion: From embarrassment to optimality. Proceedings of the 37th International Conference on Machine Learning, Vol. 119, 294–6304.

Mao, X., Chen, S. X., Wong, R. K. (2019). Matrix completion with covariate information. *Journal of the American Statistical Association, 114*(525), 198–210.

Mao, X., Wong, R. K., Chen, S. X. (2021). Matrix completion under low-rank missing mechanism. *Statistica Sinica, 31*(4), 2005–2030.

Mazumder, R., Hastie, T., Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research, 11*, 2287–2322.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science, 9*, 538–558.

Molenberghs, G., Michiels, B., Kenward, M. G., Diggle, P. J. (1998). Monotone missing data and pattern-mixture models. *Statistica Neerlandica, 52*(2), 153–161.

Negahban, S., Wainwright, M. J. (2012). Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research, 13*(1), 1665–1697.

Nielsen, S. F. (2003). Proper and improper multiple imputation. *International Statistical Review, 71*, 593–607.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review, 61*, 317–337.

Qin, J., Zhang, B., Leung, D. H. (2017). Efficient augmented inverse probability weighted estimation in missing data problems. *Journal of Business & Economic Statistics, 35*(1), 86–97.

Rao, J. N. K., Shao, J. (1999). Modified balanced repeated replication for complex survey data. *Biometrika, 86*(2), 403–415.

Robin, G., Klopp, O., Josse, J., Moulines, É., Tibshirani, R. (2020). Main effects and interactions in mixed and incomplete data frames. *Journal of the American Statistical Association, 115*(531), 1292–1303.

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

Robins, J. M., Rotnitzky, A., Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association, 90*, 106–121.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*, 581–592.

Rubin, D. B. (1978). Multiple imputations in sample surveys—a phenomenological Bayesian approach to nonresponse. *Proceedings of the survey research methods section of the American Statistical Association*, Vol. 1, 20–34. American Statistical Association.

Sengupta, N., Srebro, N., Evans, J. (2021). Simple surveys: Response retrieval inspired by recommendation systems. *Social Science Computer Review, 39*(1), 105–129.

Sun, T., Zhang, C.-H. (2012). Calibrated elastic regularization in matrix completion. *Advances in Neural Information Processing Systems, 25*, 863–871.

Sweeting, T. (1980). Uniform asymptotic normality of the maximum likelihood estimator. *Annals of Statistics,* 1375–1381.

Tan, Z. (2013). Simple design-efficient calibration estimators for rejective and high-entropy sampling. *Biometrika, 100*(2), 399–415.

Tang, G., Little, R. J., Raghunathan, T. E. (2003). Analysis of multivariate missing data with nonignorable nonresponse. *Biometrika, 90*, 747–764.

van Buuren, S., Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(3), 1–67.

van der Linden, W. J., Hambleton, R. K. (2013). *Handbook of modern item response theory*. New York, NY: Springer.

Wang, N., Robins, J. M. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika, 85*, 935–948.

Wang, S., Shao, J., Kim, J. K. (2014). An instrument variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica, 24*(3), 1097–1116.

Wang, Z., Peng, L., Kim, J. K. (2022). Bootstrap inference for the finite population mean under complex sampling designs. *Journal of the Royal Statistical Society*: *Series B (Statistical Methodology)*, Accepted.

Wu, C. (2003). Optimal calibration estimators in survey sampling. *Biometrika, 90*(4), 937–951.

Yang, S., Kim, J. K. (2016). A note on multiple imputation for method of moments estimation. *Biometrika, 103*(1), 244–251.

Yang, S., Wang, L., Ding, P. (2019). Causal inference with confounders missing not at random. *Biometrika, 106*(4), 875–888.

Zhang, C., Taylor, S. J., Cobb, C., Sekhon, J. (2020). Active matrix factorization for surveys. *Annals of Applied Statistics, 14*(3), 1182–1206.

Zou, H., Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*: *Series B* (*Statistical Methodology*), *67*(2), 301–320.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.