Causal Inference and Matrix Completion with Correlated Incomplete Data

by

Zhaohan Sun

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2022

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner: **Mireille Schnitzer**
Associate Professor, Department of Mathematics and Statistics,
University of Montreal

Supervisor(s): **Yeying Zhu**
Associate Professor, Department of Statistics and Actuarial Science,
University of Waterloo

Joel A. Dubin
Professor, Department of Statistics and Actuarial Science,
University of Waterloo

Internal Member: **Changbao Wu**
Professor, Department of Statistics and Actuarial Science,
University of Waterloo

Mu Zhu
Professor, Department of Statistics and Actuarial Science,
University of Waterloo

Internal-External Member: **Yaoliang Yu**
Associate Professor, Cheriton School of Computer Science,
University of Waterloo

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Missing data problems are frequently encountered in biomedical research, social sciences, and environmental studies. When data are missing completely at random, a *complete-case analysis* may be the easiest approach. However, when data are *missing not completely at random*, ignoring the missing values will result in biased estimators. There has been a lot of work in handling missing data in the last two decades, such as likelihood based methods, imputation methods, and bayesian approaches. The so-called *matrix completion algorithm* is one of the imputation approaches that has been widely discussed in the missing data literature. However, in a longitudinal setting, limited efforts have been devoted to using covariate information to recover the outcome matrix via matrix completion, when the response is subject to missingness.

In Chapter 1, the basic definition and concepts of different types of correlated data are introduced, and matrix completion algorithms as well as the semiparametric approaches are also introduced for handling missingness in the literature of correlated data analysis. The definition of robust estimation and interference in causal inference are also presented in this chapter.

In Chapter 2, we consider the prediction of missing responses in a longitudinal dataset via matrix completion. We propose a fixed effects longitudinal low-rank model which incorporates both subject-specific and time-specific covariates. The missingness mechanism is allowed to be missing at random, and the inverse probability weighting approach is utilized to debias the traditional quadratic loss in the matrix completion literature. To solve the optimization problem, a two-step optimization algorithm is proposed which provides good statistical properties for the estimation of the fixed effects and the low-rank term. In the theoretical investigation, the non-asymptotic error bounds on the fixed effects and the low-rank term are presented. We illustrate the finite sample performance of the proposed algorithm via simulation studies and apply our method to both a Covid-19 and PM2.5 emissions dataset.

In Chapter 3, we consider the partial interference setting, that is, the whole population can be partitioned into clusters where the outcome of each unit depends on the intervention on other units within the same cluster, but not on the units in different clusters. We also assume that the confounders are subject to nonignorable missingness. We propose three distinct consistent estimators for the direct, indirect, total, and overall effect of the intervention on the outcome, and derive the asymptotic results accordingly. A comprehensive simulation study is carried out as well to investigate the finite sample properties of the proposed estimators. We illustrate the proposed methods by analyzing the data

collected from an Acid Rain Program, which was launched to reduce air pollution in the USA by encouraging the scrubber's installation on power plants, where the records of some operating characteristics of the power generating facilities are subject to missingness.

In Chapter 4, we focus on the estimation of network causal effects. Under the setting of nonignorable missing confounders, we develop a multiply robust estimation procedure that gains extra protection against model misspecification. Compared with doubly robust estimators proposed in Chapter 3, the proposed multiply robust estimators are consistent if either one pair of the propensity score of treatment and missingness mechanism, or the joint model of confounders and the outcome, is correctly specified. The finite performance of the proposed methods under different missingness rates and cluster sizes is investigated, and we further illustrate the proposed methods with the same real data used in Chapter 3.

We conclude this thesis and discuss the future work in Chapter 5. Specifically, in Section 5.1, we summarize the contributions of the chapters in this thesis. In Section 5.2, we discuss the extension of Chapter 2, where the construction of confidence intervals for the low-rank term and the estimated fixed effects are investigated. Finally, in Section 5.3, we briefly discuss the potential extensions of Chapters 3 and 4 to a more general setting.

Acknowledgements

First, I am extremely grateful to my supervisors Dr. Yeying Zhu and Dr. Joel A. Dubin. I would not have been able to complete this thesis without their inspiring advice and guidance. Their invaluable suggestions and relentless support carried me through all stages of finishing writing this thesis. In the past four years, they are great supervisors and friends that have always been here to help me in different aspects of my life and research. I deeply appreciate all the efforts that they have done to help me overcome the encountered obstacles.

I want to express my gratitude to my committee members, Dr. Mireille Schnitzer, Dr. Changbao Wu, Dr. Mu Zhu, and Dr. Yaoliang Yu for their valuable suggestions and comments on this thesis.

I want to thank Dr. Pengfei Li, Dr. Cecilia Cotton, and Dr. Shai Ben-David for the valuable knowledge and experience that they shared during their wonderful lectures. My sincere thanks to Dr. Peijun Sang, who gave me a lot of invaluable suggestions and insights during my Ph.D. study. I also want to thank Dr. Lan Liu for introducing me to the area of causal inference, helping me strengthen the foundations, and guiding me to explore further as a researcher.

I would also like to give special thanks to my dear friends, Fangya Mao, Xiyue Han, Chi-Kuang Yeh, Yechao Meng, Cong Jiang, Chris Salahub, Sheng Wang, Ce Yang, Wenyuan Li, Hongda Hu, Yuyu Chen, Qiuqi Wang, Mingren Yin, Kecheng Li, Minghui Gao, Wenling Zhang, Trang Bui, Yuying Xie, Jialin Li, Gui Li, Jinzhou Li, Muzhu Hong, Weijie Tang, Wei Feng, Wenzhong Han, Yan Li. Thank you for all your help, encouragement, and support along this journey.

Last but not least, I would like to express my deepest gratitude to my parents Suqin Zheng and Falong Sun, who have always been the backbone of my life and given me everything to follow my dreams. I owe them everything and can never thank them enough.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Correlated Data

Generally, correlated data can be classified into different types according to the corresponding types of correlation. In Sections 1.1.1 and 1.1.2, we introduce the basic definitions and steps of statistical analysis of two types of correlated data: longitudinal data and network data. In Section 1.2, we introduce the three most widely-adopted missing data patterns and possible reasons that lead to missingness. In Section 1.3, matrix completion algorithms are introduced for handling large incomplete data matrices. In Section 1.4, we provide an introduction to the robust estimation, and we introduce network data in causal inference literature in Section 1.5.

1.1.1 Longitudinal Data

A longitudinal study refers to a research design that involves repeated observations of the same variables over short or long periods of time across a sample of units. Following Diggle et al. [2002], let Y_{ij} and x_{ij} denote a response variable and a length p vector of explanatory variables for a unit j at time t, respectively, where $i = 1, 2, \cdots m$ and $j = 1, 2, \cdots, n_i$. Let $Y_i = (Y_{i1}, Y_{i2}, \cdots Y_{in_i})^T$ be the vector of repeated outcomes for subject i. Assume $E(Y_i) = \mu_i$, $Var(Y_i) = V_i$ be the mean and covariance matrix for the outcome of subject i, respectively, where $Cov(Y_{ij}, Y_{ik}) = v_{ijk}$. The total number of responses of all units across all time points is denoted by $N = \sum_{i=1}^{m} n_i$.

A simple but widely utilized model for conducting longitudinal data analysis is a multiple linear regression model that has the following expression:

$$Y_{ij} = x_{ij}^T \beta + \epsilon_{ij},$$

where $\beta = (\beta_1, \beta_2, \cdots \beta_p)$ is a length p vector of unknown regression coefficients, and ϵ_{ij} are random variables representing model error terms, which account for measurement error and other sources of random variation, including the within-unit variation of the responses in this fixed effects model specification. The regression model has the following matrix form:

$$Y_i = X_i^T \beta + \epsilon_i,$$

where X_i is a $n_i \times p$ matrix and $\epsilon_i = (\epsilon_{i1}, \epsilon_{i2}, \cdots \epsilon_{in_i})^T$.

The primary interest of longitudinal studies lies in the investigation of the effects of change over time, and there are many merits of conducting such studies. First, they are more powerful than a *cross-sectional study* when the interest is to explore how responses vary over time with covariates. For example, a longitudinal study can help reduce the burden of collecting a sizable number of subjects required for cross-sectional studies. Besides, the variability caused by unmeasured characteristics such as environmental exposures can be controlled in the estimation from a longitudinal study. However, such effects of environmental exposures can obscure the estimation in cross-sectional studies. Second, a longitudinal study can separate the cohort and time effects because it can focus on each unit's response trajectories, but the cross-sectional study may not be suitable in the case when the variation among people is large.

With repeated measurements, numerous methods have been proposed to facilitate longitudinal data analysis. A simple, but limited, strategy is to reduce the repeated values into one or two summary statistics, then analyze each summary variable as a function of covariates. There are also different approaches to model the individual responses Y_{ij} with covariates x_{ij} such as modeling the mean of the responses, conditional expectation of Y_{ij} given the subject-specific covariates, or utilizing the *transition models* when responses are binary variables such as the following logistic regression model (Albert [2000]):

$$\log\left\{\frac{\Pr(Y_{ij}|Y_{i,j-1}, \cdots Y_{i1}, x_{ij})}{1 - \Pr(Y_{ij}|Y_{i,j-1}, \cdots Y_{i1}, x_{ij})}\right\} = x_{ij}^T \beta + \alpha Y_{i,j-1}.$$

Another widely used model is the linear mixed effects model (Laird and Ware [1982]), where both fixed and random effects are included in the model. For each individual i, the model has the following expression:

$$Y_i = X_i\alpha + Z_i b_i + e_i, \quad i = 1, 2, \cdots m$$

where α denotes a $p \times 1$ vector of fixed effects, b_i is a $k \times 1$ vector of random effects that follows multivariate normal distribution $\mathcal{N}(\mathbf{0}, D)$, D is a $k \times k$ positive-definite covariance matrix, Z_i is a known $n_i \times k$ design matrix of random effects, e_i is assumed to follow multivariate normal distribution $\mathcal{N}(0, R_i)$, and R_i is a $n_i \times n_i$ positive-definite covariance matrix. Then it can be shown that the marginal distribution of Y_i follows $Y_i \sim \mathcal{N}(X_i\alpha, Z_i D Z_i^T + R_i)$. In practice, the model can be fitted with various packages including *lme4* package in R (R Core Team [2019]). More examples and detailed explanations of our setting for longitudinal studies are presented in Chapter 2.

1.1.2 Network Data

A network (or graph) $G = (V, E)$ is a mathematical structure consisting of a set V nodes and a set E of edges, where elements of E are unordered pairs $\{u, v\}$ of distinct vertices $u, v \in V$. The number of vertices $\mathcal{N}_v = |V|$ and the number of edges $\mathcal{N}_e = |E|$ are called the order and size of the graph G, respectively. The network density is defined by the ratio of the number of edges $|E|$ and the potential number of edges, that is, $2|E|/(n(n-1))$. The following adjacency matrix A can be used to represent the structure of the network.

$$A_{ij} = \begin{cases} 1 & \text{edges exist from node } v_i \text{ to node } v_j \\ 0 & \text{otherwise.} \end{cases}$$

Note that the network defined above is named a so-called *binary network*, and it does not work for *weighted network* where the entries of A may take any positive values.

Conventional regression analysis is not suitable for network data without any justification. To see this, assume a network can be disconnected into different sub-networks or groups such as hospitals in a local district where each hospital serves as its own sub-network. Let Y_{ij} be the j^{th} observation in the i^{th} group, $j = 1, 2, \cdots n_i$, and $i = 1, 2, \cdots K$, that is, there are K disjoint groups with n_i units in each group. Suppose

$$Y_{ij} = x_{ij}^T\beta + \epsilon_{ij},$$

and

$$Cov(Y_{ij}, Y_{ik}) = \sigma^2\rho,$$

where x_{ij} is length p explanatory variable of unit j in the group i, $\beta = (\beta_1, \beta_2, \cdots \beta_p)$ is a vector of unknown regression coefficients of length p, ϵ_{ij} is the random error with zero mean and variance σ^2, and the covariance between any pair of the responses in the same group equals $\sigma^2\rho$, $0 \leq \rho \leq 1$. Let $\hat{\beta}$ be the ordinary least square(OLS) estimator of β. The OLS estimator is unbiased. However, ignoring the dependency structure within the groups will result in several problems. On one hand, the variance estimator of $\hat{\beta}$ is incorrect. On the other hand, using $\hat{\beta}$ as the estimator will result in a loss of efficiency, that is, the variance of $\hat{\beta}$ is greater than that of the best-unbiased estimator (Ntani et al. [2021]). Further discussion of this group network setting will be left to Section 1.5 and Chapter 3.

Many models and methods have been proposed to handle network data. The classical models usually assume a likelihood function for the network data with some underlying parameters. The most fundamental probabilistic model is Erdős-Rényi model, which was first proposed by Erdős and Rényi [1960], where the presence of the edges between all nodes are assumed to be *i.i.d.* Bernoulli random variables, and the model of the probability density of the network has the following expression:

$$\Pr(A|\theta) = \prod_{i,j} \theta^{A_{ij}}(1-\theta)^{1-A_{ij}},$$

where θ is the unknown parameter in the Bernoulli distribution. Several extended models based on Erdős-Rényi model have also been proposed such as p_1, p_2 model, and exponential (family) random graph models (ERGMs)(see Kolaczyk and Csárdi [2014] for more details).

Latent variable models aim to explain the underlying structure of the network through some additional modeling. The stochastic block model introduced by Holland et al. [1983] is one of the most widely used latent variable models that can detect network structures. Assume that the nodes in a network can be partitioned into K clusters $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \cdots \mathcal{C}_K\}$. Let block density matrix $B^{K\times K}$ be a probability matrix with each of its entries $B(i,j)$ equal to the probability of any nodes $u \in \mathcal{C}_i$ and $v \in \mathcal{C}_j$ being connected by an edge. Then, the model can be expressed as

$$g(\mu) = \sum_i \sum_j C_{ij} B_{ij} B_{ji},$$

where C_{ij} represents the cluster assignment matrix with $C_{ij} = 1$ if node i belongs to cluster $\mathcal{C}_j$ and 0 otherwise, $\mu = E(A)$, and $g(\cdot)$ is a link function. The probabilistic graphical models such as the latent space model introduced by Hoff et al. [2002] and the latent factor model by Hoff [2009] have also been proposed to help explain the latent structure of the network.

1.2 Missing Data Patterns

According to Rubin [1976], and still emphasized presently, there are three types of missingness: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). The missingness mechanism is MCAR if the missingness does not depend on either observed or missing values, the missingness mechanism is MAR if the missingness does not depend on the missing values, and MNAR refers to a missing data process that can depend on both observed and missing values.

There are different reasons that may lead to missingness in correlated data. One reason is non-response. For example, in the literature on survey sampling, the participant may feel uncomfortable or highly sensitive when answering some of the questions in a questionnaire. To be more specific, there are two types of non-response in the context of a network: unit non-response and item non-response, where the unit non-response refers to the case when both the outcome and out-going edges are completely missing for a unit, and the item non-response refers to that either the outcome or certain out-going edges are missing. In the case of longitudinal data, the non-response can be further distinguished by including partial non-response, which is characterized by time dependency, and means that for some units only at certain time points of the intended data collection are available. This is often due to panel mortality or attrition, which results in completely missing cases after a certain time point(Huisman and Steglich [2008]). In a longitudinal study, one omnipresent reason is the existence of dropouts, where dropouts may occur when some participants experience adverse treatment effects, or some participants change their living location and can no longer participate in a given study. It is also common that intermittent missingness happens, for example, in a blood test experiment, a participant may skip a few tests occasionally during the study period. A further complication is when a person is observed and either the response or at least one predictor is not available, for example, one or two predictors are not available due to a lab error.

Examining the patterns of missingness can be performed in different ways. Plotting trajectories of the response variable or the proportion of missingness across time is a simple approach. It is also appealing to perform a formal statistical test for the types of missingness on a given dataset. Little [1988] proposed a global test statistic for MCAR where the asymptotic null distribution and small sample distribution are given when data are subject to monotone missingness. Qu and Song [2002] proposed a testing procedure for MCAR missingness mechanism with quadratic inference functions. In practice, it may be difficult to distinguish between MAR and MNAR, one possible way is to conduct sensitivity analysis, which is used for examining and quantifying the effects of departures of different assumptions. There are various ways to conduct sensitivity analysis such as the *pattern* *mixture model* approach and *selection model* approach (Molenberghs and Verbeke [2000]). Ma et al. [2005] extended the idea of *index of sensitivity to nonignorability* (ISNI) to longitudinal data with nonignorable drop-outs. Since the literature on ISNI methods is still evolving, there is no standard rule to perform the analysis.

One naive method to handle missingness is to delete all the units that contain missing values and use only those units with complete observed data to conduct statistical analysis, which is known as *complete case analysis*. In practice, the complete case analysis may be acceptable if those with any incomplete data comprise less than 5% of the original sample size. In a longitudinal study, instead of ignoring all of those subjects' visits, one may use *available case analysis* by just deleting the visits at which the missingness happens. Both cases work well if the proportion of missingness is small. However, both complete case analysis and available case analysis are no longer suitable if the proportion of missingness is large or the missingness mechanism is not MCAR, as such deletion can result in severely biased estimators of mean and regression coefficients. When data are not missing completely at random, various methods have been proposed to handle this issue such as inverse probability weighting, EM algorithm, imputation, and maximum likelihood-based method; see Tang and Ju [2018] for a review of more state-of-the-art approaches.

1.3 Matrix Completion

Matrix completion problems are frequently encountered in recommendation systems, computer vision, and system identification studies. The goal of matrix completion is to estimate unobserved elements in a matrix through observed elements. Without any restriction, this problem is NP-hard. However, a data matrix usually has some special properties, and these properties make it possible to perform matrix completion. For example, low rank is one of these properties. If a low-rank assumption is made, matrix completion can be characterized as the following optimization problem:

$$\begin{aligned} &minimize && rank(X) \\ &subject\, to && X_{ij} = M_{ij},\ for\ \ (i,j) \in \Omega, \end{aligned}$$

where X is the unknown matrix, M is the true matrix, and Ω is the set of locations at which entries are observed. Let $\mathcal{P}_\Omega : \mathcal{R}^{n\times n} \rightarrow \mathcal{R}^{n\times n}$ be the orthogonal projection onto the subspace of matrices which vanishes outside Ω (i.e., $\mathcal{P}_\Omega = X\ if\ \ (i,j) \in \Omega$, and $\mathcal{P}_\Omega = 0$ *otherwise*), so that the information about M can be given by $\mathcal{P}_\Omega(X)$.

Since the optimization problem above cannot be solved in practice, Candès and Recht

[2009] proposed recovering the unknown matrix by minimizing the following nuclear norm of X:

$$\begin{array}{ll} minimize & \|X\|_\star \\ subject\,to & \mathcal{P}_\Omega(X) = \mathcal{P}_\Omega(M), \end{array}$$

where $\|X\|_\star$ is the nuclear norm which is defined as the summation of singular values of X. Then, after convex relaxation, the optimization problem can be solved by semi-definite programming. Candès and Recht [2009] proved that under an incoherence assumption and uniformly at random sampling scheme, with large probability, the solution to the optimization program is unique and equal to the true matrix, provided that the number of observed entries obeys $m \geqslant \mathcal{O}(n^{5/4}log(n))$, where m is the number of observed entries, and n is the maximum of the number of rows and the number of columns. This is the first theoretical result that shows that the lower bound of sampling complexity can be utilized to evaluate whether a matrix can be exactly recovered, which opens the possibility of exact matrix recovery. Candès and Tao [2010] improved their results by making a strong incoherence assumption on matrices. With this assumption, the lower bound of sampling complexity was further improved to $\mathcal{O}(nlog^2(n))$; this improvement is significant because the strong incoherence assumption is satisfied by most matrices. More importantly, $n^{5/4}$ is optimized to n, which can greatly reduce the sampling complexity for high-dimensional matrices.

Although the feasibility of the exact matrix recovery has been proven, in practice, it is more important to find an effective algorithm for solving the problem. Since minimization of the nuclear norm problem can be transferred into a semi-definite programming problem, it is natural to use a semi-definite toolkit package (e.g. SDPT3, SeDuMi) to get a solution, but there are some issues when applying these tools. On one hand, most of these methods use the interior point algorithm to solve convexity optimization problems. When the dimension of the matrix is high, computing the Newton direction is time-consuming, while another problem lies in the fact that the condition number can be really large when using the conjugate gradient method to calculate the direction of Newton's step, which results in unstable numerical results. On the other hand, these general optimization methods are inefficient because they do not use the low-rank property of the true matrix.

The Singular Value Thresholding (SVT) algorithm proposed by Candès and Tao [2010] is the first spectral decomposition algorithm. The original idea is straightforward, as they used matrix factorization to get a deeper understanding of the internal matrix structure and generation process so that they can recover the true matrix. The optimization problem with convexity relaxation has the following expression:

$$\begin{aligned} &minimize \quad \tau\|X\|_\star + \frac{1}{2}\|X\|_F^2 \\ &subject\, to \quad \mathcal{P}_\Omega(X) = \mathcal{P}_\Omega(M), \end{aligned}$$

where $\|.\|_F$ is the Frobenius norm, which is defined as the square root of the sum of squares of the elements of the inside matrix. By Lagrange dual method,

$$\begin{aligned} &X^k = D_\tau(Y^{k-1}) \\ &Y^k = Y^{k-1} + \delta_k \mathcal{P}_\Omega(M - X^k), \end{aligned}$$

where D_τ is an operator for each $\tau > 0$. Assume $X = U\Sigma V^T$, and the singular values σ_i are positive. Then we have

$$\begin{aligned} &D_\tau(X) = UD_\tau(\Sigma)V^T \\ &D_\tau(\Sigma) = diag(max\{\sigma_i - \tau, 0\}). \end{aligned}$$

It has been proved when the learning rate δ_k is smaller than a Lipschitz constant, the sequence $\{X^k\}$ will converge to the true matrix M. In each iteration, both $\{X^k\}$ and $\{Y^k\}$ are estimators of the true matrix, since the estimation of $\{Y^k\}$ can be affected by small singular values. Singular value thresholding was utilized to remove those small singular values, with only key components being kept in each step.

However, one of the constraints in previous results is that all observed elements in the data matrix should be accurate, which is too idealistic in practice. Candes and Plan [2010] utilized the following additive noise model:

$$\mathcal{P}_\Omega(Y) = \mathcal{P}_\Omega(M) + \mathcal{P}_\Omega(Z),$$

where $Y \in R^{n_1 \times n_2}$ is the observed data matrix, M is the true matrix, and Z is a noise term that may be stochastic or deterministic. They showed that if there exists a constant δ such that $||\mathcal{P}_\Omega(Z)||_F \leq \delta$, then the estimator $\hat{M}$ obeys

$$\|M - \hat{M}\|_F \leq 4\sqrt{\frac{C_p \min(n_1, n_2)}{p}}\delta + 2\delta,$$

i.e., the error is proportional to the magnitude of the noise, such that the error is small when the noise level is small.

In recent years, driven by the aforementioned fundamental theoretical studies, various estimation and identification methods have been proposed when the objective matrix is subject to missingness in matrix completion literature. Mazumder et al. [2010] proposed an iterative algorithm *Soft-Impute* for computing low-rank approximations for incomplete matrices, with convex relaxations of rank penalty term, on a grid of tuning parameter values. The algorithm alternates between imputing missing values from a current singular value decomposition, and updating the singular value decomposition using the current imputed matrix. In the noiseless setting, Keshavan et al. [2010] proposed a three-step algorithm called *Optspace*, which avoids the time-consuming SVD decomposition in every iteration of the algorithm; they also showed that the unknown matrix can be exactly recovered from $\mathcal{O}(n\log(n))$ entries, given that the true matrix is sufficiently unstructured and the rank is $\mathcal{O}(1)$.

1.4 Robust Estimation

In causal inference literature, to reduce the bias of the average treatment effect estimators, various methods have been proposed to adjust for confounders. One of the most popular classes of methods is the *inverse probability weighting* (Rosenbaum and Rubin [1983]) approach. The idea is to create a pseudo-population by weighting each subject by the inverse of the probability of receiving the treatment conditional on the confounders, such that the association between the treatment and confounders can be removed. In an observational study, the propensity score model is generally unknown and needs to be estimated. However, in practice, the relationship between the covariates and the treatment can be complicated and may be difficult to correctly specify the propensity score model. In fact, the IPW estimators are highly sensitive to the specification and the estimation of the propensity score models. To avoid the risk of biased estimators due to the incorrect specification of propensity score models, robust estimation has gained more and more popularity in causal inference and missing data problems. Robust estimation aims at adding protection against model misspecification by allowing the specification of multiple candidate working models.

In Chapter 3, we propose a set of doubly robust estimators for four types of network causal effects (the definitions of the network causal effects are given in 3). The doubly robust estimator was first proposed by Robins et al. [1994] and Rotnitzky et al. [1998] in the form of an augmented IPW estimator in missing data models. The methodology was further discussed in Robins et al. [2000], Lunceford and Davidian [2004], Bang and Robins [2005], and Kang and Schafer [2007]. Typically, a doubly robust (DR) estimator requires the estimation of two nuisance functionals: the propensity score and an outcome regression. In missing data problems, an estimator is DR if it is consistent when either the missingness mechanism or the conditional distribution of the outcome data is correctly specified. In causal inference problems, an estimator is DR if it is consistent when either the conditional probability of receiving the treatment or the conditional distribution of the outcome given treatment and confounders is consistently estimated. When data are independent, it is also well-known that the estimator achieves the semiparametric efficiency bound if both models are correctly specified.

In Chapter 4, we propose a set of multiply robust estimators. The multiply robust estimators have been studied in Han and Wang [2013], Han [2014a,b] and Han [2018] based on the empirical likelihood approach. Compared with doubly robust estimators, the multiply robust (MR) estimators add more protection against model misspecification by allowing the postulation of a set of candidate parametric working models. An estimator is MR if it remains consistent if any one of the candidate models, either for the propensity score model or for the regression model, is correctly specified. When data are missing not at random, Li et al. [2020a] constructed the multiply robust estimators by proposing the calibration constraints directly on the score equations for the parameter of interest under multiple working models.

1.5 Causal Inference with Interference

Over the past decade, the problem of inference in the treatment effect when data are subject to missingness has drawn a great amount of attention. According to Rubin [1976], there are two types of missingness: ignorable and nonignorable missingness. Ignorable missingness refers to missingness that is independent of the missing values, and nonignorable missingness refers to missingness that is dependent on the missing values. The inference for nonignorable missingness is more challenging than ignorable missingness because the full data distribution is not fully identifiable without any assumptions, sometimes very restrictive. Following Yang et al. [2019], we consider the *group-level outcome-independent missingness assumption*, where the missingness is independent of the outcome conditional on confounders and treatment, which is plausible when the covariates are collected at the beginning of the study, and the outcome is collected long after the covariates are measured.

In most of the aforementioned work in handling missing data in the literature, methods rely on the Stable Unit Treatment Value Assumption (SUTVA) Rubin [1980]. SUTVA states that (i) the potential outcome of each unit is unaffected by the treatment assignment of any other unit, and (ii) there are no different versions of each treatment level. However, the first assumption, which is known as the **no interference assumption** Cox [1958], can be violated in some scenarios. For example, in 1990, the Acid Rain Program was launched to reduce ambient PM2.5 (atmospheric particulate matter (PM) that has a diameter of fewer than 2.5 micrometers) by assigning power plants to install scrubber facilities Zigler et al. [2016]. The monitored reduction of SO_2 emission data at the location of one power plant not only depends on its own scrubbers' installation but may also be affected by the intervention of power plants upwind. Another example comes from the nationally representative US Population Assessment of Tobacco and Health (PATH) Study Hyland et al. [2017], where researchers were interested in evaluating the influence of Electronic Nicotine Delivery Systems (ENDS) and pharmaceutical cessation aids on persistent abstinence from cigarette smoking and reduced cigarette consumption. In this study, it has been revealed that one individual's marital satisfaction and family members' smoking status can affect this individual's smoking cessation, that is, the smoking cessation of one individual may be affected by the intervention of other family members. More examples can also be found in biomedical research, public health sciences, and social networking studies.

Various identification and estimation methods have been proposed in the scenario when interference exists but the confounders and outcome are fully observed. Generally, there are two types of interference: *full interference* and *partial interference*. Full interference happens when the potential outcome of a unit is affected by the intervention of any other unit that interferes with this unit. The network interference structure can be represented by an adjacency matrix: the entries of which take the value on $\{0, 1\}$ (e.g., if the unit i is affected by the intervention on the individual j, then the entry of the matrix in i^{th} row and j^{th} column equals one; otherwise, the value of the entry equals zero). Partial interference is a special case of full interference where the adjacency matrix follows block diagonal structure, the entry of the matrix is equal to zero if the unit of the corresponding row and the unit of the corresponding column are not in the same block, that is, the interference may happen between units in the same block but not between units in different blocks. In this chapter, we consider the latter type of interference and focus on the semiparametric estimation of four network treatment effects: the direct effect, indirect effect, total effect, and overall effect (Tchetgen Tchetgen and VanderWeele [2012], Hudgens and Halloran [2008], Liu et al. [2019], Papadogeorgou et al. [2019]). The definitions and illustrations of these treatment effects are presented in Section 3.2. Bhattacharya et al. [2020] proposed a general method for estimating causal effects under data dependence when the structure of this dependence is not known a priori. Imai et al. [2021] proposed consistent estimators for direct and spillover effects under the stratified interference assumption. Giffin et al. [2020] proposed a generalized propensity score and a computational algorithm for estimating the spillover effects.

Chapter 2

Noisy Matrix Completion for Longitudinal Data

2.1 Introduction

In modeling longitudinal data, there is an increasing interest in estimating the unknown parameters in the models, when the responses and/or the covariates are subject to missingness. Fitzmaurice et al. [2012] discussed different types of missingness mechanisms and proposed several methods for dealing with the missingness accordingly. Generally, different methods are reliable under different missingness patterns. When data are subject to ignorable missingness, various approaches have been proposed to handle the missingness. For example, multiple imputation (MI) replaces the missing values with plausible values multiple times. MI includes parametric approaches and non-parametric approaches. Multivariate imputation by chained equations (MICE)(Buuren and Groothuis-Oudshoorn [2010]) is one of the parametric methods which, with the conditional distribution, regresses each variable based on other variables during the imputation procedures. Maximum likelihood approachDempster et al. [1977], Ibrahim [1990], Eekhout et al. [2015], fully Bayesian inference such as Gibbs sampling Rubin [1976], and semiparametric methods Zhao et al. [1996], Robins et al. [1994] have also been proposed for estimating the parameters of interest.

In this chapter, we focus on imputing the missing responses with longitudinal data and deriving the upper bounds for the estimation error based on matrix completion theories. To analyze the longitudinal data, we utilize a longitudinal low-rank model based on the linear fixed effects model. The unit- and time-specific covariates are included in the proposed model to improve the imputation accuracy. The linear fixed effects model with the unit- and time-specific covariates has been widely utilized for modeling longitudinal data in environmental studies, health research, and econometric problems. The following are some examples that use the model or variants of it.

Example 1: Conroy et al. [2002] developed multiple models for the analysis of recapture data for 2678 serins ringed in north-eastern Spain since 1985. The objective of the research was to explore the predictive relationship between the survival of serins and the unit-specific and time-specific covariates. Time-specific covariates included different types of weather conditions. Individual covariates included body mass, wing length, interactions between body mass and environmental factors, etc. A number of plausible models with different combinations of the unit- and time-specific covariates were formed, and the Akaike Information Criterion(AIC) was used to rank the fitted models.

Example 2: In econometric literature, there is some interest in investigating the effects of the macroeconomic and bank-specific covariates on the non-performing loans of a bank. For example, Mehmood et al. [2013] used a fixed effects model to model the effects of macroeconomic factors(e.g., interest rate, and GDP) and bank-specific covariates such as market share of the bank in the banking market, return on assets of the bank, return on equity and statuary liquidity requirements on non-performing loans in Pakistan from 2003 to 2012.

Example 3: Berry et al. [2004] utilized a modified empirical differentiated products demand model on second-choice automotive purchases data. The data contains both product-specific covariates and consumers' characteristics, and the interaction terms between consumer tastes and product characteristics were included in the model to determine substitution patterns. The authors also proposed moments estimators for the unknown parameters in the linear fixed effects model and showed that the limiting distributions of the proposed estimators follow normal distributions.

In this study, we develop a two-step estimation procedure to estimate the fixed effects and the low-rank term in the proposed model. We further show the performance of the proposed methods in terms of the estimation error via both the theoretical analysis and the simulation studies.

The rest of the chapter is organized as follows. In Section 2.2, we introduce the notation and assumptions. In Section 2.3, we propose a longitudinal low-rank model and a two-step estimation algorithm for solving the optimization problem. In Section 2.4, we present the non-asymptotic error bounds for the estimated outcome matrix, the first-order fixed effects, and the low-rank term. The finite sample performance of the proposed algorithm is illustrated through simulation studies in Section 2.5, and the proposed algorithm is applied to both the Covid-19 and the SO_2 emissions dataset in Section 2.6.

2.2 Notations and Setup

Let $Y \in \mathbb{R}^{N\times T}$ denote the outcome matrix, the values of which may be subject to missingness, N represents the number of units, and T represents the number of time points. Let $X \in \mathbb{R}^{N\times P}$ denote the unit-specific covariate matrix, and $Z \in \mathbb{R}^{T\times Q}$ denote the time-specific covariate matrix. Let $R \in \mathbb{R}^{N\times T}$ denote the missingness indicator, more specifically, $R_{ij} = 1$ if Y_{ij} is observed, and $R_{ij} = 0$ if Y_{ij} is missing. Let Y_i^O and Y_i^M denote the vector of observed and missing responses on the i^{th} subject, respectively. Here, we only consider that the outcome is subject to missingness and the covariate matrices X and Z are assumed deterministic and complete. Let $L \in \mathbb{R}^{N\times T}$ denote a low-rank matrix, and r_L denote the rank of L, where we assume $r_L << min(N, T)$. Let $e_1, e_2 \cdots e_N$ denote the standard basis of $\mathbb{R}^N$. Let $\mathcal{E}_{ij}$ be a matrix with all entries equal to zero except that its $(i, j)^{\text{th}}$ entry is equal to 1, where $1 \le i \le N$ and $1 \le j \le T$. We consider the following decomposition of the target outcome matrix:

$$Y_{N\times T} = X_{N\times P} H_{P\times Q} Z'_{T\times Q} + L_{N\times T} + \epsilon_{N\times T}, \tag{2.1}$$

where $H_{P\times Q}$ is a fixed effects coefficient matrix, and $\epsilon_{N\times T}$ is a random error term. Let $\epsilon_{ij} = \epsilon_{ij}^{(1)} + \epsilon_{ij}^{(2)} \quad \forall 1 \le i \le N, 1 \le j \le T$, where $\epsilon_{ij}^{(1)}$ represents presumed serial correlation and $\epsilon_{ij}^{(2)}$ represents the random noise term.

Our model shares some similarities with the model proposed by Athey et al. [2018], in which the time-varying covariates are also included. Robin et al. [2020] also proposed a similar decomposition, where the matrix of interest is decomposed as the summation of the main effects and the low-rank term, and a negative quasi-likelihood function is utilized as the loss function. Our model differs from the previous work in two main aspects. First, we assume that the low-rank term includes higher-order main and interaction terms, and we let the random error term be a combination of serial correlation and random noise instead of the random noise only. Under this assumption, we focus on estimating the main and first-order interaction terms. Second, we consider the missingness mechanism as MAR (Little and Rubin [2019]), i.e., $\Pr(R_{ij} = 1|Y, X, Z) = \Pr(R_{ij} = 1|Y^O, X, Z)$. In this chapter, not only are the statistical guarantees for the estimator of L shown, but the statistical properties for the estimator of the fixed effects coefficient matrix are also provided. The proposed method is also similar to recovering the principal components of the data matrix (Candès et al. [2011]), but our focus here is to impute the missing values.

Let $A, B \in \mathbb{R}^{N\times T}$. The Kronecker product is denoted by $A \otimes B$, and the trace inner product is defined as

$$\langle \mathbf{A}, \mathbf{B}\rangle := \operatorname{tr}\left(\mathbf{A}\mathbf{B}^{\top}\right).$$

The following matrix norms are used for the remainder of this paper.

1. **The Schatten p-norms:** The singular values of the matrix A are denoted by σ_i, $1 \leq i \leq \min(N, T)$, and $\|A\|_p = \left(\sum_{i=1}^{\min(N,T)} \sigma_i^p(A)\right)^{\frac{1}{p}}$.

2. **Nuclear norm:** $\|A\|_* = \operatorname{trace}\left(\sqrt{A^T A}\right) = \sum_{i=1}^{\min\{N,T\}} \sigma_i(A)$.

3. **Operator norm:**

$$\|A\|_{op} = \sup\left\{\frac{\|Av\|}{\|v\|} : v \in V \text{ with } v \neq 0\right\}.$$

   If we specifically choose the Euclidean norm on both R^N and R^T, then the matrix norm given to a matrix A is the square root of the largest eigenvalue of the matrix $A^T A$. This is equivalent to the largest singular value of A.

4. **Frobenius norm:** The Frobenius norm is a special case of $\mathbf{L_{p,q}}$ norm when $p = q = 2$. This norm can be defined in various ways:

$$\|A\|_{\mathrm{F}} = \sqrt{\sum_{i=1}^{N}\sum_{j=1}^{T} |a_{ij}|^2} = \sqrt{\operatorname{trace}\left(A^T A\right)} = \sqrt{\sum_{i=1}^{\min\{N,T\}} \sigma_i^2(A)}.$$

5. $\mathbf{L_\infty}$ **norm:** $\|A\|_\infty = \max_{1\leq i\leq N, 1\leq j\leq T} \sum_{i=1}^{N}\sum_{i=1}^{T} \mid a_{ij} \mid$.

6. $\mathbf{L_1}$ **norm:** $\|A\|_1 = \max_{1\leq i\leq N, 1\leq j\leq T} \sum_{i=1}^{N}\sum_{i=1}^{T} |a_{ij}|$.

7. $\mathbf{L_2(\Pi)}$ **norm:** $\|A\|_{L_2(\Pi)} = \sqrt{E\langle A, X\rangle^2}$, where X is sampled from a probability measure on $\mathbb{R}^{N\times T}$.

Here, we also make the following assumptions:

Assumption 1 (covariate terms constraints)**.** *There exists constants C_x, C_z, and C_{xz} such that $\|X\|_\infty < C_x$, $\|Z\|_\infty < C_z$, $\|Z \otimes X\|_\infty < C_{xz}$, and $\|L\|_\infty < C_L$.*

Assumption 1 is an extension of the conditions that $\|L\|_\infty < \infty$, $\|X\|_\infty < C_x$, and $\|Z\|_\infty < C_z$, which guarantees that the fixed effects term is finite.

Assumption 2 (propensity score constraints)**.** *There exist positive constants p_1 and p_2 such that $0 < p_1 \leq \mathcal{P}_{ij} = \Pr(R_{ij} = 1|Y^O_{i.}, X_{i.}, Z_{j.}) \leq p_2 < 1$ a.s., where $\Pr(R_{ij} = 1|Y^O, X, Z)$, or $\mathcal{P}$ for short, is the propensity score model (or matrix) for missingness.*

Assumption 2 states that every entry in the matrix has a positive probability to be either observed or missing, which rules out the case when some individuals at some time points can never be observed or always be observed.

Assumption 3 (random error term constraints)**.** *(a) $E(\epsilon) = \mathbf{0}, \|\mathbf{E}(\epsilon^{\mathbf{2}})\|_\infty < \infty$. (b) $\{\epsilon^{(1)}_{ij}\}_{1\leq i\leq N, 1\leq j\leq T}$ are σ sub-Gaussian random variables, and $\{\epsilon^{(1)}_{i.}\}_{1\leq i\leq N}$ are independent of each other. (c) $\{\epsilon^{(2)}_{ij}\}_{1\leq i\leq N, 1\leq j\leq T}$ are i.i.d. τ sub-Gaussian random variables.*

Assumption 3 (a) states that the random error is centered and has a finite variance. In (b), both $\epsilon^{(1)}$ and $\epsilon^{(2)}$ follow sub-Gaussian distributions.

2.3 A Longitudinal Low-rank Model and a Two-step Estimation Algorithm

The classic inverse probability weighting (IPW) estimator was first proposed by Horvitz and Thompson [1952] in the survey sampling literature. The idea is to create a pseudo-population by weighting each subject by the inverse of the conditional probability of receiving the treatment. The classic IPW can also be extended to the setting when the outcome is subject to missingness. The widely used way to correct the bias is to weight each subject by the inverse of the propensity score for missingness, i.e., $1/\Pr(R = 1|Y^O, X, Z)$. The weighted objective function has the following representation:

$$\arg\min_{H,L} E\left\{\frac{R}{\Pr(R = 1|Y^O, X, Z)}\|Y - (XHZ' + X\alpha\mathbf{1_T}' + \mathbf{1_N}\beta' Z' + L)\|_F^2\right\} \\ +\lambda_H\|H\|_1 + \lambda_L\|L\|_\star, \tag{2.2}$$

where $\alpha_{P\times 1}$ denotes a vector of unknown parameters for fixed unit effects, $\beta_{1\times Q}$ denotes fixed time effects, H is the coefficient matrix for the first-order interaction terms among time-specific and unit-specific covariates, L is a low-rank matrix representing the higher-order main effects and interaction effects, and $\mathbf{1_N}$ and $\mathbf{1_T}$ are column vectors with all entries equal to one, the dimensions of which are N and T, respectively. The penalty terms $\lambda_H\|H\|_1$ and $\lambda_L\|L\|_\star$ are incorporated to avoid overfitting, and $\lambda_H, \lambda_L > 0$ are the respective regularization parameters.

To reduce the computational burden, notice that the previous optimization problem can be expressed in a more compact way: the terms $X\alpha\mathbf{1_T}^T$ and $\mathbf{1_N}\beta^T Z^T$ can be incorporated into the term XHZ^T, and the optimization problem is equivalent to the following representation:

$$\arg\min_{H,L} E\left\{\frac{R}{\Pr(R=1|Y^O,X,Z)}\|(\tilde{X}\tilde{H}\tilde{Z}'+L)-Y\|_F^2\right\}+\lambda_H\|\tilde{H}\|_1+\lambda_L\|L\|_\star,$$

where

$$\tilde{X}=\left(X_{N\times P} \vdots I_{N\times N}\right)\begin{pmatrix} I_{P\times P} & \mathbf{0_{P\times 1}} \\ \mathbf{0_{N\times P}} & \mathbf{1_N}\end{pmatrix}, \tag{2.3}$$

$$\tilde{Z}=\left(Z_{T\times Q} \vdots I_{T\times T}\right)\begin{pmatrix} I_{Q\times Q} & \mathbf{0_{Q\times 1}} \\ \mathbf{0_{T\times Q}} & \mathbf{1_T}\end{pmatrix}, \tag{2.4}$$

$$\tilde{H}=\begin{pmatrix} H & \alpha \\ \beta^T & 0\end{pmatrix}, \tag{2.5}$$

$I_{N\times N}$ and $I_{T\times T}$ are identity matrices, and $\mathbf{0_{P\times 1}}$, $\mathbf{0_{N\times P}}$, $\mathbf{0_{Q\times T}}$, and $\mathbf{0_{1\times Q}}$ are matrices with all entries equal to zero. For the rest of this paper, we use X, H and Z instead of $\tilde{X}$,$\tilde{H}$ and $\tilde{Z}$, respectively, to avoid excess notation. Let

$$F(H,L)=\frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{T}\frac{R_{ij}}{\hat{\Pr}(R_{ij}=1|Y_i^O,X_{i.},Z_{j.})}\{(XHZ^T)_{ij}+L_{ij}-Y_{ij}\}^2,$$

and

$$F_{\bar{\lambda}}(H,L)=F(H,L)+\lambda_H\|H\|_1+\lambda_L\|L\|_\star,$$

where $\bar{\lambda}=(\lambda_H,\lambda_L)$, and $\hat{\Pr}(R_{ij}=1|Y_i^O,X_{i.},Z_{j.})$ is a consistent estimator of $\Pr(R_{ij}=1|$ $Y_i^O,X_{i.},Z_{j.})$, which can be obtained via maximum likelihood. We then aim to minimize $F_{\bar{\lambda}}(H,L)$ under the constraints that $\|H\|_\infty<\infty$ and $\|L\|_\infty<\infty$, i,e., $(\hat{H},\hat{L})\in\arg\min_{H,L}$

$F_{\bar{\lambda}}(H, L)$. We will further explain the constraints on the parameter space in Section 2.4.

Let $M = f_{(x,z)}(H) + L = XHZ^T + L$, where $f_{(x,z)}(H) = XHZ^T$. Let $M^* = XH^*Z^T + L^*$ denote the true value of M. Let $\hat{H}$, $\hat{L}$ and $\hat{M}$ be the estimators of H, L, and M, respectively. The minimization of $F(H, L)$ can be solved by different algorithms. For example, stochastic gradient descent, proximal gradient descent, and the Adagrad algorithm (see Bottou et al. [2018] for a review of more advanced optimization algorithms) can all be utilized to solve this optimization problem. Coordinate gradient descent (CGD) proposed by Yun and Toh [2011] is one of those algorithms with Q-linear convergence rate, which can also be extended to the setting when there are multiple variables that need to be updated. The minimization of the above convex optimization problem can be solved by the adaptive CGD algorithm in an iterative way. Note that Athey et al. [2018] also mentioned that iterative coordinate descent can be used to solve such a convex function. Here, we provide the details of the estimation procedure, and focus on showing how the unit-specific and time-specific covariates can benefit the imputation accuracy in a longitudinal setting. In each iteration step (n), we update H and L by the following rule: first, the search direction for H is

$$\begin{aligned} d_H^{(n)} &\in \underset{d \in R^{P\times Q}}{\arg\min} \left\{ F(\hat{H}^{(n)} + d, \hat{L}^{(n)}) + \lambda_H ||\hat{H}^{(n)} + d||_1 \right\} \\ &\in \underset{d \in R^{P\times Q}}{\arg\min} \frac{1}{NT} \sum_{i=1}^{N} \sum_{j=1}^{T} \frac{R_{ij}}{\hat{\Pr}(R_{ij} = 1 | X_{i.}, Z_{j.})} \{ f_{(x,z)}(d)_{ij} - (Y_{ij} - \hat{M}_{ij}^{(n)}) \}^2 + \lambda_H \|\hat{H}^{(n)} + d\|_1. \end{aligned} \tag{2.6}$$

Notice that the l_1 penalty term in Equation (2.6) includes the summation of the current update $\hat{H}^{(n)}$ and the search direction for $\hat{H}^{(n)}$. To simplify the problem, we rewrite the above optimization problem in the following way such that it can be directly solved with well-developed packages in R, such as the *glmnet* package in R Hastie and Qian [2014]:

$$\begin{aligned} H_{temp}^{(n+1)} &\in \underset{H \in R^{P\times Q}}{\arg\min} \frac{1}{NT} \sum_{i=1}^{N} \sum_{j=1}^{T} \frac{R_{ij}}{\hat{\Pr}(R_{ij} = 1 | Y_i^O, X_{i.}, Z_{j.})} \left[f_{(x,z)}(H)_{ij} - \{ (Y_{ij} - M_{ij}^{(n)}) + f_{(x,z)}(H^{(n)})_{ij} \} \right]^2 \\ &\quad + \lambda_H \|H\|_1 \\ &\in \underset{H \in R^{P\times Q}}{\arg\min} \frac{1}{NT} \sum_{i=1}^{N} \sum_{j=1}^{T} \frac{R_{ij}}{\hat{\Pr}(R_{ij} = 1 | Y_i^O, X_{i.}, Z_{j.})} \left\{ f_{(x,z)}(H)_{ij} - (Y_{ij} - L_{ij}^{(n)}) \right\}^2 + \lambda_H \|H\|_1. \end{aligned} \tag{2.7}$$

Specifically, the above minimization problem is equivalent to a weighted lasso problem with l_1 penalty term, which can be solved numerically with *glmnet*. Thus, the estimated search direction is calculated by $\hat{d}_H^{(n)} = \hat{H}_{temp}^{(n+1)} - \hat{H}^{(n)}$, and the intermediate estimator $\hat{M}^{(n),int}$ of M is equal to

$$\hat{M}^{(n),int} = f_{x,z}(\hat{H}^{(n)} + \tau_H^{(n)} \hat{d}_H^{(n)}) + L^{(n)},$$

where $\tau_H^{(n)}$ is the step size that is selected by the following steps of the adapted *Armijo rule* Bertsekas [1997]. First, choose $\tau_H^0 > 0$ as an initial step size, and let $\tau_H^{(n)}$ be the largest element of $\{\tau_H^0 \beta^j\}_{j=0,1\ldots}$ satisfying

$$F(\hat{H}^{(n)} + \tau_H^{(n)} \hat{d}_H^{(n)}, \hat{L}^{(n)}) + \lambda_H \|\hat{H}^{(n)} + \tau_H^{(n)} \hat{d}_H^{(n)}\|_1 \leq F(\hat{H}^{(n)}, \hat{L}^{(n)}) + \lambda_H \|\hat{H}^{(n)}\|_1 + \tau_H^{(n)} \sigma \Delta_H^{(n)}, \tag{2.8}$$

where $0 < \beta < 1$, $0 < \sigma < 1$, $0 < \gamma < 1$,

$$\begin{aligned} \Delta_H^{(n)} = & -\frac{2}{NT} \sum_{i=1}^N \sum_{j=1}^T \frac{R_{ij}}{\hat{\Pr}(R_{ij} = 1 | Y_i^O, X_{i.}, Z_{j.})} (Y_{ij} - \hat{M}_{ij}^{(n)})(f_{(x,z)}(\hat{d}_H^{(n)})_{ij}) \\ & + \gamma \frac{1}{NT} \sum_{i=1}^N \sum_{j=1}^T \frac{R_{ij}}{\hat{\Pr}(R_{ij} = 1 | Y_i^O, X_{i.}, Z_{j.})} (f_{(x,z)}(\hat{d}_H^{(n)})_{ij})^2 \\ & + \lambda_H (\|\hat{H}^{(n)} + \hat{d}_H^{(n)}\|_1 - \|\hat{H}^{(n)}\|_1). \end{aligned} \tag{2.9}$$

The first and second terms in $\Delta_H^{(n)}$ are the first and second order derivatives of the objective function, respectively. The above Armijo rule is a popular inexact line search condition, where larger step sizes are accepted if we choose larger γ and smaller σ. The basic intuition of the Armijo rule is to find a small step size such that the objective function has a sufficient decrease in each iteration. Let $d_H = 0$, and we have the following representation for the search direction of the low-rank matrix L:

$$\begin{aligned} d_L^{(n)} \in \underset{d \in R^{N \times T}}{\arg\min} \frac{1}{NT} \sum_{i=1}^N \sum_{j=1}^T \frac{R_{ij}}{\hat{\Pr}(R_{ij} = 1 | Y_i^O, X_{i.}, Z_{j.})} \left\{ d_{ij} - \left(Y_{ij} - \hat{M}_{ij}^{(n),int} \right) \right\}^2 \\ + \lambda_L \|\hat{L}^{(n)} + d\|_\star. \end{aligned} \tag{2.10}$$

Then,

$$\begin{aligned} L_{temp}^{(n+1)} \in \underset{L \in R^{N \times T}}{\arg\min} \frac{1}{NT} \sum_{i=1}^N \sum_{j=1}^T \frac{R_{ij}}{\hat{\Pr}(R_{ij} = 1 | Y_i^O, X_{i.}, Z_{j.})} \left[L_{ij} - \left\{ Y_{ij} - f_{(x,z)}(\hat{H}^{(n+1)}) \right\} \right]^2 \\ + \lambda_L \|L\|_\star. \end{aligned} \tag{2.11}$$

The above minimization problem is equivalent to a weighted matrix completion problem, where each individual is weighted by the inverse of estimated propensity score for missingness, i.e., $1/\hat{\Pr}(R_{ij} = 1|Y_i^O, X_{i.}, Z_{j.})$. Notice that $L^{(n)}$ is estimable as we restrict the summation of the square loss only over those observed entries. Then, the estimated search direction for L is $\hat{d}_L^{(n)} = \hat{L}_{temp}^{(n+1)} - \hat{L}^{(n)}$, and we have the following expression of the estimator for M in the next iteration:

$$\hat{M}^{(n+1)} = f_{x,z}(\hat{H}^{(n)} + \tau_H^{(n)} \hat{d}_H^{(n)}) + (\hat{L}^{(n)} + \tau_L^{(n)} \hat{d}_L^{(n)}),$$

where $\tau_L^{(n)}$ is the step size selected by the following adapted Armijo rule. Similarly as before, we first choose $\tau_L^0 > 0$ as an initial step size, and let $\tau_L^{(n)}$ be the largest element of $\{\tau_L^0 \beta^j\}_{j=0,1\dots}$ satisfying

$$\begin{aligned} &F(\hat{H}^{(n)} + \tau_H^{(n)} \hat{d}_H^{(n)}, \hat{L}^{(n)} + \tau_L^{(n)} \hat{d}_L^{(n)}) + \lambda_L \|\hat{L}^{(n)} + \tau_L^{(n)} \hat{d}_L^{(n)}\|_\star \\ \leq &F(\hat{H}^{(n)} + \tau_H^{(n)} \hat{d}_H^{(n)}, \hat{L}^{(n)}) + \lambda_L \|\hat{L}^{(n)}\|_\star + \tau_L^{(n)} \sigma \Delta_L^{(n)}, \end{aligned} \tag{2.12}$$

where $0 < \beta < 1$, $0 < \sigma < 1$, $0 < \gamma < 1$, and

$$\begin{aligned} \Delta_L^{(n)} = &-\frac{2}{NT} \sum_{i=1}^{N} \sum_{j=1}^{T} \frac{R_{ij}}{\hat{\Pr}(R_{ij} = 1|Y_i^O, X_{i.}, Z_{j.})} (Y_{ij} - \hat{M}_{ij}^{(n),int})(\hat{d}_{L,ij}^{(n)}) \\ &+ \gamma \frac{1}{NT} \sum_{i=1}^{N} \sum_{j=1}^{T} \frac{R_{ij}}{\hat{\Pr}(R_{ij} = 1|Y_i^O, X_{i.}, Z_{j.})} (\hat{d}_{L,ij}^{(n)})^2 \\ &+ \lambda_L (\|\hat{L}^{(n)} + \hat{d}_L^{(n)}\|_\star - \|\hat{L}^{(n)}\|_\star). \end{aligned} \tag{2.13}$$

For the sake of clarity, the steps of the proposed algorithm are summarized as below.

2.4 Theoretical Results

In this section, we aim to derive the l_2 estimation error of the estimated parameters when serial correlation exists. Notice that related results have also been discussed in some previous work, e.g., Klopp et al. [2014], Klopp et al. [2017], Athey et al. [2018], Robin et al. [2019] and Hamidi and Bayati [2019]. The main difference lies in that we need to account for the stochastic error term of the serial correlation in our setting, as well as provide the non-asymptotic error bounds for the estimated fixed effects. For the sake of

Algorithm 1 Two-step Matrix Completion Algorithm

```
Input: Unit-specific covariates $X$, time-specific covariates $Z$, and observed outcome matrix $Y$.
Initialize: $\hat{H}^{(0)}$ and $\hat{L}^{(0)}$ are generated via ordinary least squares estimation.
for $\frac{\|\hat{H}^{(n+1)}-\hat{H}^{(n)}\|_F}{\|\hat{H}^{(n)}\|_F} > \epsilon$ or $\frac{\|\hat{L}^{(n+1)}-\hat{L}^{(n)}\|_F}{\|\hat{L}^{(n)}\|_F} > \epsilon$ do
    Step 1: Compute $\hat{H}^{(n+1)}$ by Equation (2.7) with glmnet package in R;
    Step 1.5 : Compute intermediate estimator $\hat{M}^{(n),int}$;
    Step 2: Compute $\hat{L}^{(n+1)}$ by Equation (2.11) with softImpute algorithm.
end for
Output: Compute $\hat{Y} = X\hat{H}Z' + \hat{L}$.
```

completeness, we first prove the global convergence of our algorithm, which is summarized in Theorem 2.4. Since the updates of $\hat{H}$ and $\hat{L}$ in each iteration satisfy the line search condition, the objective function is always non-increasing after the updates of each of the parameters. The proof of Lemma 2.1 is an adaptive version of Proposition 3.1 in Tseng and Yun [2009].

Lemma 2.1. *The $F_{\bar{\lambda}}(H^n, L^n)$ is monotonically non-increasing.*

The proof of Lemma 2.1 is given in Section 2.8. It follows that the updates of estimated unknown parameters H and L belong to the level set defined as

$$lev(F_{\bar{\lambda}}) = \{(H^{(n)}, L^{(n)}) | F_{\bar{\lambda}}(H^{(n)}, L^{(n)}) \leq F_{\bar{\lambda}}(H^0, L^0)\},$$

where (H^0, L^0) is the starting point of the proposed algorithm. Since the objective function is non-increasing according to Lemma 2.1, we proceed by showing the following Lemma, which is the condition for proving the existence of minimizer stated in Lemma 2.3.

Lemma 2.2. *The level sets $lev(F_{\bar{\lambda}})$ are compact. The proof is given is Section 2.8.*

Lemma 2.3. *There exists at least one minimizer $(H^\star, L^\star)$ for the objective function $F_{\bar{\lambda}}(H, L)$, i.e., $\forall H \in R^{P\times Q}$, and $L \in R^{N\times T}$, $F_{\bar{\lambda}}(H, L) \geq F_{\bar{\lambda}}(H^\star, L^\star)$.*

The main tool for proving Lemma 2.3 is Weierstrass's Theorem (Apostol [1974]), which states that every continuous function in a compact set attains its minimum, the formal proof for the lemma is shown in Section 2.8.

Theorem 2.4. Under Assumptions 1 and 2, assume $\{H^{(n)}, L^{(n)}\}$ are generated by the proposed algorithm, then, every cluster point of $\{H^{(n)}, L^{(n)}\}$ is a stationary point of $F_{\bar{\lambda}}(H, L)$.

Next, to derive the estimation error bounds for $\hat{Y}$, $\hat{H}$, and $\hat{L}$, we proceed by the following steps. First, we start by showing that the error bounds for $\hat{Y}$ depend on the summation of two terms: the first term is proportional to $\|\Delta H\|_1$, and the second term is proportional to $\|\Delta L\|_F$, given that the regularization parameters satisfy $\lambda_H \geq \| \sum_{i=1}^N \sum_{j=1}^T \epsilon_{ij} R_{ij} \mathcal{E}_{ij} \|_\infty$, and $\lambda_L \geq \| \sum_{i=1}^N \sum_{j=1}^T \epsilon_{ij} R_{ij} \mathcal{E}_{ij} / \hat{\Pr}(R_{ij} = 1 | Y_i^O, X_{i.}, Z_{j.}) \|_{op}$. Second, we show the probabilistic upper bounds for the stochastic errors $\| \sum_{i=1}^N \sum_{j=1}^T \epsilon_{ij} R_{ij} \mathcal{E}_{ij} \|_\infty$ and $\| \sum_{i=1}^N \sum_{j=1}^T \epsilon_{ij} R_{ij} \mathcal{E}_{ij} / \hat{\Pr}(R_{ij} = 1 | Y_i^O, X_{i.}, Z_{j.}) \|_{op}$. In the proposed algorithm, we require the regularization parameters λ_H and λ_L to be greater than these two upper bounds, respectively. The expression of the non-asymptotic error bounds is presented in Lemma 4 and Lemma 5, respectively. Thirdly, we define the following two constrained sets with respect to H and L:

$$\mathcal{C}_H(\theta_H) := \left\{ H \in R^{P \times Q} \middle| \|H\|_1 \leq 1, \|H\|_{L_2(\Pi)}^2 \geq \theta_H \right\},$$

and

$$\mathcal{C}_L(r, \theta_L) := \left\{ L \in R^{N \times T} \middle| \|L\|_\infty \leq 1, \|L\|_{L_2(\Pi)}^2 \geq \theta_L, \|L\|_\star \leq \sqrt{r} \|L\|_F \right\}.$$

We will present that the restricted strong convexity property (RSC) holds on these two constrained sets. Such property was first proven to hold in matrix completion problems by Negahban and Wainwright [2012], and similar work can also be found in Klopp et al. [2014] and Athey et al. [2018]. Roughly speaking, we will show if $\Delta H \in \mathcal{C}_H(\theta_H)$ and $\Delta L \in \mathcal{C}_L(r, \theta_L)$, with high probability, $E(\|\Delta L\|_F^2)$ is smaller than $\|\Delta L\|_F^2$ with an additional term, and $E(\|\Delta X H Z^T\|_F^2)$ is smaller than $\|\Delta X H Z^T\|_F^2$ with an additional term.

Lemma 2.5. With probability $1 - \exp(-t)$, the stochastic error $\| \sum_{i=1}^N \sum_{j=1}^T \epsilon_{ij} R_{ij} \mathcal{E}_{ij} \|_\infty$ has the following probabilistic upper bound:

$$\| \sum_{i=1}^N \sum_{j=1}^T \epsilon_{ij} R_{ij} \mathcal{E}_{ij} \|_\infty \leq 2 \sqrt{\log(2NT) + t} \left(\sqrt{T} \sigma + \tau \right).$$

Lemma 2.5 shows the explicit expression of the probabilistic upper bounds for the stochastic error with respect to λ_H. Notice that the above expression contains the term $\sqrt{T}\sigma + \tau$, which indicates that the infinity norm of this stochastic error term is proportional to the magnitude of the variance of the random error. Thus, as we will show in the following theorems, the additional term $\sqrt{T}\sigma + \tau$ deteriorates the estimation error of the estimated fixed effects, the low-rank term, and the imputed outcome matrix.

Lemma 2.6. With probability $1 - \exp(-t)$, the probabilistic upper bound for the stochastic error $\| \sum_{i=1}^{N} \sum_{j=1}^{T} \epsilon_{ij} R_{ij} \mathcal{E}_{ij} / \hat{\Pr}(R_{ij} = 1 | Y_i^O, X_{i.}, Z_{j.}) \|_{op}$ has the following expression:

$$
\begin{aligned}
&\| \sum_{i=1}^{N} \sum_{j=1}^{T} \epsilon_{ij} \frac{R_{ij}}{\hat{\Pr}(R_{ij} = 1 | Y_i^O, X_{i.}, Z_{j.})} \mathcal{E}_{ij} \|_{op} \\
\leq & (c_1^\star + c_2^\star) \max \Bigg[\left\{ \frac{16 \max\{T \log(N), N \log(T)\} \sigma^2 + 2 \max(N, T) \tau^2}{p_1} \right\} \times \sqrt{t + log(N+T)}, \\
& \left\{ \frac{16 \sigma^2 e \max\{T \log(N), N \log(T)\}}{p_1} + \frac{2 \max(N, T) \sqrt{T} \tau}{p_1} \log\left(\frac{\sqrt{T} p_1}{\tau} \right) \right\} \times \left(t + log(N+T) \right) \Bigg].
\end{aligned}
$$

Lemma 2.6 shows the expression of the upper bound for the stochastic error with respect to λ_L. The main idea is to decompose the random error term into the serial correlation and the independent random noise components and apply Bernstein's inequality to both parts. The idea is similar to some of the previous work in the matrix completion literature, but differs in that such decomposition provides more information on the dependency of the upper bounds for $\hat{H}$ and $\hat{L}$ on the covariance structure of the serial correlation and the variance of random noise. To be more specific, the above representation of the stochastic error is the maximum of the two terms, where both terms consist of the variance of serial correlation and the random noise. Thus, it implies that the operator norm of this stochastic error is positively correlated with the variance of the random error. With such decomposition, similar results can also be achieved by extending the model to the setting when the random error also includes some subordinate random effects. Since this is beyond the scope of the paper, we will leave it as future work.

Theorem 2.7. Assume $\lambda_H \geq 2 \| \sum_{i=1}^{N} \sum_{j=1}^{T} \epsilon_{ij} R_{ij} \mathcal{E}_{ij} \|_\infty / NT$, and
$\lambda_L \geq 4 \| \sum_{i=1}^{N} \sum_{j=1}^{T} \epsilon_{ij} \frac{R_{ij}}{\Pr(R_{ij}=1|X_{i.}, Z_{j.})} \mathcal{E}_{ij} \|_{op} / NT$, we have the following representation of the upper bound for $\hat{\mathbf{M}}$.

$$
\frac{1}{NT} \left\| \frac{R \circ (\hat{\mathbf{M}} - \mathbf{M}^\star)}{\hat{\mathcal{P}}} \right\|_F^2 \leq 6 \lambda_L \sqrt{2 r_L} \|\Delta L\|_F + 2 \lambda_H \|\Delta H\|_1.
$$

Interpretation of Theorem 2.7: The above error bound, which quantifies the l_2 estimation error of the imputed matrix on the observed data, can be decomposed as the summation of the error bound for the estimated fixed effects matrix and the error bound for the estimated low-rank matrix. However, the estimated error for the whole outcome matrix (or the root mean square error) cannot be directly obtained through the above inequality. To make progress, in Theorem 2.9, we will show that the restricted strong convexity (RSC) holds on the constraint set defined for $\hat{H}$, with the tuning parameter λ_H satisfying the condition in Lemma 2.8. Similarly, in Theorem 2.11, we will show that the RSC holds on the subspace of the constraint set of $\hat{L}$, where the definition of the constraint set for $\hat{H}$ and $\hat{L}$ will be given in the proof of Lemma 2.8 and Theorem 2.11, respectively. Then, the error bound for the estimated outcome matrix can easily be obtained via the summation of the error bound for the estimated fixed effects and the low-rank matrix.

Lemma 2.8. Assume $\lambda_H \geq 6C_xC_z\Big(2C_L/p_1 + \|\sum_{i=1}^N\sum_{j=1}^T R_{ij}\epsilon_{ij}\mathcal{E}_{ij}/\mathcal{P}_{ij}\|_\infty\Big)/(NT)$, we have the following inequality:

$$\frac{1}{2}\sum_{i=1}^N\sum_{j=1}^T\frac{R_{ij}}{\mathcal{P}_{ij}}\langle X\Delta HZ^T,\mathcal{E}_{ij}\rangle^2 \leq 4\lambda_H\|H^\star\|_1.$$

Theorem 2.9. Suppose $\lambda_H \geq 6C_xC_z\Big(2C_L/p_1 + \|\sum_{i=1}^N\sum_{j=1}^T R_{ij}\epsilon_{ij}\mathcal{E}_{ij}/\mathcal{P}_{ij}\|_\infty\Big)/(NT)$, the following probabilistic upper bound holds.

$$\Pr\Big\{\frac{p_1}{2}\|X\Delta HZ^T\|_F^2 > \sum_{i=1}^N\sum_{j=1}^T R_{ij}\langle X\Delta HZ^T,\mathcal{E}_{ij}\rangle^2 + \frac{c_H^\star}{p_1}(E\|\sum_{i=1}^N\sum_{j=1}^T\zeta_{ij}R_{ij}\mathcal{E}_{ij}\|_\infty)^2\Big\}$$
$$\leq \quad \frac{1}{(N+T)^2},$$

where $c_H^\star$ is a large enough constant, the value $c_H^\star$ is shown in Section 2.8.

Assume $c_H^\star$ is defined the same as that in Theorem 2.9. Combining the results of Lemma 2.8 and Theorem 2.9, with probability $1-1/(N+T)^2$, we have the following upper bound for the square of the root mean square error (RMSE) of $X\Delta H Z^T$,

$$\begin{aligned}
\frac{1}{NT}\|X\Delta H Z^T\|_F^2 &\le \frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{T} R_{ij}\langle X\Delta H Z^T, \mathcal{E}_{ij}\rangle^2 + \frac{c_H^\star}{NTp_1}(E||\sum_{i=1}^{N}\sum_{j=1}^{T}\zeta_{ij}R_{ij}\mathcal{E}_{ij}||_\infty)^2 \\
&\le \frac{1}{NT}4\lambda_H||H^\star||_1 + \frac{c_H^\star}{NTp_1}(E||\sum_{i=1}^{N}\sum_{j=1}^{T}\zeta_{ij}R_{ij}\mathcal{E}_{ij}||_\infty)^2 \\
&\le \frac{24C_HC_xC_z}{NT}\bigg(\frac{2C_L}{p_1} + ||\frac{\sum_{i=1}^{N}\sum_{j=1}^{T}R_{ij}\epsilon_{ij}\mathcal{E}_{ij}}{\mathcal{P}_{ij}}||_\infty\bigg) + \frac{c_H^\star}{NTp_1} \\
&\le \frac{24C_HC_xC_z}{NT}\bigg\{\frac{2C_L}{p_1} + \frac{2\sqrt{\log(2NT)+t}(\sqrt{T}\sigma+\tau)}{p_1}\bigg\} + \frac{c_H^\star}{NTp_1} \quad (2.14)
\end{aligned}$$

where ζ_{ij} are i.i.d. Rademacher random variables. As shown above, the square of RMSE of the estimated fixed effects term will converge to zero as $\max\{N, T\}$ goes to infinity.

Lemma 2.10. Suppose $\lambda_L \ge 6\bigg(2C_L\sqrt{\max(N,T)}/p_1 + \|\sum_{i=1}^{N}\sum_{j=1}^{T} R_{ij}\epsilon_{ij}\mathcal{E}_{ij}/\hat{\mathcal{P}}_{ij}\|_{op}\bigg)/(NT)$, then the following inequality holds:

$$\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{T}\frac{R_{ij}}{\hat{\mathcal{P}}_{ij}}\langle \Delta L, \mathcal{E}_{ij}\rangle^2 \le 2\lambda_L\|\Delta L\|_\star.$$

Theorem 2.11. *Assume* $\lambda_L \ge 6\bigg(2C_L\sqrt{\max(N,T)}/p_1 + \|\sum_{i=1}^{N}\sum_{j=1}^{T} R_{ij}\epsilon_{ij}\mathcal{E}_{ij} /\hat{\mathcal{P}}_{ij}\|_{op}\bigg)/(NT)$, *then we have the following representation of probabilistic upper bound:*

$$\begin{aligned}
&\Pr\bigg\{\frac{p_1}{2}\|\Delta L\|_F^2 > \sum_{i=1}^{N}\sum_{j=1}^{T}R_{ij}\langle \Delta L, \mathcal{E}_{ij}\rangle^2 + \frac{c_L^\star r_{L^\star}}{p_1}(E\|\sum_{i=1}^{N}\sum_{j=1}^{T}\zeta_{ij}R_{ij}\mathcal{E}_{ij}\|_{op})^2\bigg\} \\
\le\ & \frac{1}{(N+T)^2},
\end{aligned}$$

where c_L^* is some numerical constants. Let the value of λ_L equal to $6\bigg(2C_L\sqrt{\max(N,T)}/p_1 + || \sum_{i=1}^{N}\sum_{j=1}^{T} R_{ij}\epsilon_{ij}\mathcal{E}_{ij}/\mathcal{P}_{ij}||_{op}\bigg)/(NT)$. Then, with probability $1-1/(N+T)^2$, we have the following upper bound for the square of the RMSE of the estimator $\hat{L}$ by Lemma 2.10:

$$\begin{aligned}
\frac{1}{NT}\|\Delta L\|_F^2 &\leq \frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{T} R_{ij}\langle \Delta L, \mathcal{E}_{ij}\rangle^2 + \frac{c_L^\star r_{L^\star}}{NTp_1}(E\|\sum_{i=1}^{N}\sum_{j=1}^{T}\zeta_{ij}R_{ij}\mathcal{E}_{ij}\|_{op})^2 \\
&\leq \frac{4\lambda_L p_2}{NT}\|\Delta L\|_\star + \frac{c_L^\star r_{L^\star}}{NTp_1}(E\|\sum_{i=1}^{N}\sum_{j=1}^{T}\zeta_{ij}R_{ij}\mathcal{E}_{ij}\|_{op})^2 \\
&\leq \frac{16\sqrt{2r_{L^\star}}\lambda_L p_2}{NT}\|\Delta L\|_F + \frac{c_L^\star r_{L^\star}}{NTp_1}(E\|\sum_{i=1}^{N}\sum_{j=1}^{T}\zeta_{ij}R_{ij}\mathcal{E}_{ij}\|_{op})^2 \\
&\leq \frac{16\sqrt{2r_{L^\star}}p_2}{NT}\left(\frac{\sqrt{\max(N,T)}C_L}{p_1} + \|\frac{\sum_{i=1}^{N}\sum_{j=1}^{T} R_{ij}\epsilon_{ij}\mathcal{E}_{ij}}{\hat{\mathcal{P}}_{ij}}\|_{op}\right) + \frac{c_L^\star min(N,T) r_{L^\star}}{max(N,T)p_1}
\end{aligned} \tag{2.15}$$

Remark: Here, we assume that $\min\{N,T\}$ is equal to $(\max\{N,T\})^\delta$, where $\delta \in (0,1)$ is a constant. For example, in the scenario that the number of units increases faster than the number of time periods, we let both T and N increase, but T increases only at a certain rate of N.

Theorem 2.9 establishes the upper bound for the root mean square error of the main and the first-order interaction terms in the proposed model, i.e., $\|X\Delta HZ^T\|_F^2/NT$. Note that on the right-hand side of the inequality 2.14, the upper bound contains two terms. In the first term, the numerator has a logarithmic factor $2\sqrt{\log(2NT)+t}$ multiplied by the magnitude of the random error $\sqrt{T}\sigma + \tau$. In both the first and the second terms, the denominator is a polynomial function of NT. Therefore, the square of the RMSE will converge to zero as the dimension of the matrix becomes sufficiently large.

Theorem 2.11 establishes the upper bound for the root mean square error of the low-rank term, i.e., $\|\Delta L\|_F^2/NT$. The derived upper bound in inequality 2.15 has two terms, where the numerator in the first term includes $\sqrt{\max(N,T)}$ and an operator norm, which is bounded above by the polynomial function NT. Besides, the second term converges to zero as the $\max(N,T)$ becomes sufficiently large. Therefore, the RMSE will converge to zero as the dimension of the outcome matrix grows.

Notice that the quantification of the statistical variation of the estimators is beyond the scope of this chapter, we will leave it as a future research topic and briefly discuss potential extensions in Section 2.7 and Chapter 5.

2.5 Simulation Studies

In this section, we study the finite sample performance of the proposed two-step algorithm via a comprehensive simulation study. As shown in the theoretical analysis, the performance of the proposed algorithm is dependent on the dimension of the outcome matrix and the probability of missingness. To investigate the ability of the proposed algorithm to handle the missingness, we consider two different missing data patterns, MCAR and MAR, under three different missing rates: 30%, 50%, and 70%. We conduct the simulation study in the following steps.

First, we start by considering the scenario when the outcome of 100 individuals has evaluated over 30 time points under 30% missing rate. We generate a unit-specific covariate matrix $X_{100\times5}$ and a time-specific covariate matrix $Z_{30\times5}$. The unit-specific covariate matrix is generated from $\mathcal{N}(0.5, 0.1)$, and the time-specific covariate matrix is generated from $\mathcal{N}(1.5, 0.1)$, where $\mathcal{N}(\mu, \sigma^2)$ denotes normal distribution with mean μ and variance σ^2. We generate a low-rank matrix $L_{N\times T}$ with rank 2. We generate the random noise term by $\mathcal{N}(0, \tau^2)$, where τ^2 is the variance of noise term that is calculated by setting the signal-to-noise ratio (SNR) to be 1, i.e., $SNR = E(signal^2)/\tau^2 = 1$ where $E(signal^2) = \sum_{i=1}^{N}\sum_{j=1}^{T}((XHZ^T + L)_{ij} - \mu_{sig})^2/(NT-1)$, and $\mu_{sig} = \sum_{i=1}^{N}\sum_{j=1}^{T}(XHZ^T + L)_{ij}/(NT)$. The serial correlation term is generated by $\mathcal{N}(\mu, \Sigma)$, where $\Sigma_{ij} = \sigma^2\rho^{|i-j|}$, $\sigma^2 = \tau^2$ and $\rho = 0.5$. To obtain the performance of the proposed method under different levels of correlation, we also show the results of $\rho = 0.2$ and $\rho = 0.8$. To investigate the effect of including the serial correlation in the proposed longitudinal low-rank model, we generate the complete outcome matrix Y from $\mathcal{N}(XHZ^T + L, \mathcal{E})$ in two scenarios. In the first scenario, $\mathcal{E}_{100\times30}$ is the variance of random noise itself. In the second scenario, $\mathcal{E}_{100\times30}$ is the summation of the covariance of the serial correlation and the variance of random noise.

Second, under MCAR, the missingness indicator matrix $R_{N\times T}$ is generated for three different settings from the Bernoulli distribution with probability 0.3, 0.5 and 0.7, respectively. In the scenario when the data are assumed MAR, for simplicity, we assume the propensity score of the missingness as the logistic model: $\text{logit}\{\Pr(R_{ij} = 1|Y_i^O, X_{i.}, Z_{j.})\} = \beta_0 + \beta_1 X_{i.} + \beta_2 Z_{j.}$. In order to compare the results with those in the scenario when data are MCAR, let $\beta_0 = -0.7$, $\beta_1 = (-0.6, -0.5, -0.4, -0.3, -0.2)$ and $\beta_2 = (-0.1, 0, 0.1, 0.2, 0.3)$, such that the total missing rate is 70%. Similarly, when missing rate is 50%, let $\beta_0 = -0.61$, $\beta_1 = (-0.51, -0.41, -0.31, -0.21, -0.11)$ and $\beta_2 = (-0.01, 0.09, 0.19, 0.29, 0.39)$. When missing rate is 30%, let $\beta_0 = -0.54$, $\beta_1 = (-0.44, -0.34, -0.24, -0.14, -0.04)$ and $\beta_2 = (0.06, 0.16, 0.26, 0.36, 0.46)$. Based on the missingness probability matrix R, the training set indices are generated, and the set of the remaining entries of the outcome matrix, which contains all the true values of the outcome matrix, is used for testing.

Third, the estimators $\hat{\beta}$ and $\hat{L}$ are calculated via the proposed two-step algorithm. Note that although the proposed algorithm provides restrictions on the choice of tuning parameters, in practice, we find the optimal values of tuning parameters via K-fold cross-validation. More specifically, we choose the tuning parameters λ_H and λ_L by 10-fold cross-validation over the two-dimensional grid of 10 different values of λ_L and 100 different values of λ_H that are evenly distributed on the log scale. The maximum value of the sequence of tuning parameters λ_L is computed via the singular value of the original matrix with missing values replaced by zero. We also consider the dimension of the outcome matrix to be 1000 × 30, 1000 × 100, and repeat the above procedures 200 times accordingly.

To provide more insight into how the unit-specific and the time-specific covariates information can benefit the accuracy of imputation, we compare the proposed approach with two existing imputation methods: a) a penalized estimator for contaminated incomplete outcome matrix without utilizing the covariate information, proposed by Koltchinskii et al. [2011]) multivariate imputation by chained equations (MICE) by Buuren and Groothuis-Oudshoorn [2010] with the unit- and time- specific covariates, which has also been widely used in drawing imputations from different types of datasets, including longitudinal data. In MICE, the number of draws is set to 5 in our setting. Across the whole simulation, we use root mean square error (RMSE) (or test error) and mean square error (MSE) to evaluate the performance of the algorithms, the definitions of which are presented below:

$$MSE = \frac{\|R \circ (Y - \hat{Y})\|_F^2}{\|R \circ Y\|_F^2}, \quad RMSE = \frac{\|(Y - \hat{Y})\|_F}{\sqrt{NT}}. \tag{2.16}$$

where $\circ$ denotes Hadamard product (also known as element-wise product). The MSE measures the bias of the estimated observed values, and the RMSE measures the performance of the algorithm on the estimation of the whole matrix.

For the scenarios when the error term is only the random noise, and the dimension of the outcome matrix is 100 × 30, 1000 × 30 and 1000 × 100, the MSE, RMSE, and empirical standard errors are presented in Tables 2.1, 2.2 and 2.3, respectively. The simulation results, where the serial correlation was included in the model, are presented in Table 2.4. First, for the proposed algorithm and the other two methods, note that both the MSE and RMSE decrease as the values of N and T increase, that is, the proposed algorithm as well as the two competing methods perform better as the dimension of the outcome matrix increases, which is consistent with our theoretical results (e.g., under 30 % missing rate and the MCAR missingness mechanism, the MSE of the proposed algorithm are 0.013 and 0.009, and the RMSE are 0.439 and 0.433, when the dimension of the outcome matrix are 100 × 30 and 1000 × 30, respectively). Second, the estimators of the proposed algorithm perform well compared with the other two methods in terms of having smaller MSE and RMSE. Third, it is worth noting that the multiple imputation approach has inferior performance in terms of having a large MSE compared with the other two methods in all the scenarios. We speculate one of the reasons is that we utilize the general fixed effects model during the MI procedure instead of a low-rank model, which treats the low-rank term as the intercept term when fitting the regression models. Besides, we did not perform variable selection here, which can lead to biased estimators for both coefficient matrices H and L. Again, we aim to show that the proposed two-step estimation procedure that utilized both the unit-specific and the time-specific covariates can produce estimators with smaller MSE and RMSE compared with (i) other existing matrix completion algorithms that do not use such information, or (ii) other imputation algorithms in the longitudinal setting that do not consider the low-rank property of the original matrix.

Table 2.1: MSE and RMSE (empirical standard errors are in brackets) of the proposed two-step matrix completion algorithm (TSMC), multiple imputation (MI), and the traditional matrix completion algorithm (TMC) when the dimension of response matrix is 100 × 30.

			MCAR			MAR	
		70%	50%	30%	70%	50%	30%
TSMC	MSE	0.019[0.005]	0.012[0.002]	0.013[0.012]	0.035[0.022]	0.0126[<0.001]	0.015[0.004]
	RMSE	0.780[0.102]	0.502[0.044]	0.439[0.122]	0.947[0.212]	0.684[0.023]	0.665[0.097]
MI	MSE	0.502[0.053]	0.488[0.051]	0.482[0.052]	0.475[0.064]	0.485[0.059]	0.477[0.057]
	RMSE	1.337[0.131]	1.137[0.110]	1.137[0.110]	1.335[0.143]	1.124[0.110]	1.124[0.111]
TMC	MSE	0.078[0.041]	0.037[0.080]	0.051[0.058]	0.070[0.081]	0.032[0.040]	0.038[0.071]
	RMSE	1.507[0.366]	0.988[0.100]	0.994[0.145]	1.237[0.251]	0.032[0.040]	1.114[0.112]

2.6 Application

2.6.1 Covid-19 data

In this section, we apply the proposed methods on a Covid-19 dataset to estimate the potential contaminated data. We first describe the Covid-19 dataset and discuss the reasons for missingness or contamination; we also introduce the chosen unit-specific and time-specific covariates, and their potential influence on the outcome variable. Second, we apply the proposed low-rank model and the algorithm to the data. The data are collected from 304 main cities in China from January 19, 2020, to February 29, 2020. We consider

Table 2.2: MSE and RMSE (empirical standard errors are in brackets) of the proposed two-step matrix completion algorithm (TSMC), the multiple imputation (MI), and the traditional matrix completion algorithm (TMC) when the dimension of response matrix is 1000×30.

		MCAR			MAR		
		70%	50%	30%	70%	50%	30%
TSMC	MSE	0.022[0.013]	0.010[<0.001]	0.009[0.007]	0.024[0.004]	0.015[0.001]	0.014[0.001]
	RMSE	0.751[0.155]	0.473[0.022]	0.433[0.129]	0.780[0.057]	0.681[0.031]	0.659[0.032]
MI	MSE	0.464[0.034]	0.470[0.050]	0.469[0.049]	0.483[0.045]	0.471[0.048]	0.473[0.050]
	RMSE	1.431[0.094]	1.426[0.130]	1.426[0.140]	1.466[0.049]	1.437[0.138]	1.438[0.139]
TMC	MSE	0.059[0.062]	0.037[0.060]	0.044[0.045]	0.035[0.010]	0.037[0.060]	0.035[0.050]
	RMSE	1.158[0.187]	1.02[0.090]	0.989[0.131]	0.938[0.116]	1.121[0.092]	1.099[0.085]

Table 2.3: MSE and RMSE (empirical standard errors are in brackets) of the proposed two-step matrix completion algorithm (TSMC), the multiple imputation (MI), and the traditional matrix completion algorithm (TMC) and the test error when the dimension of response matrix is 1000×100.

		MCAR			MAR		
		70%	50%	30%	70%	50%	30%
TSMC	MSE	0.019[0.002]	0.013[0.001]	0.008[0.006]	0.024[0.004]	0.015[0.001]	0.014[0.001]
	RMSE	0.797[0.060]	0.638[0.026]	0.433[0.144]	0.800[0.049]	0.684[0.023]	0.673[0.028]
MI	MSE	0.443[0.043]	0.488[0.058]	0.422[0.037]	0.478[0.061]	0.437[0.077]	0.470[0.039]
	RMSE	1.367[0.115]	1.437[0.152]	1.436[0.132]	1.479[0.131]	1.350[0.140]	1.351[0.138]
TMC	MSE	0.041[0.040]	0.035[0.040]	0.046[0.050]	0.041[0.014]	0.032[0.041]	0.035[0.030]
	RMSE	1.133[0.060]	1.049[0.060]	1.020[0.120]	1.133[0.157]	1.275[0.160]	1.157[0.049]

the recorded number of cases per day as the outcome variable. We exclude the responses from the city of Wuhan as part of the dependent variable in the modeling because of the unique epidemic pattern of the virus in the city, though other information from Wuhan may play a role in the analysis, as can be seen below.

Regarding the explanatory variables, in the dataset, four important unit-specific covariates are included: the population density, the number of doctors, the gross domestic product(GDP), and the distance between cities and Wuhan. These factors may affect the change in the number of confirmed cases. For example, if the population density is high, then the virus can spread among individuals with higher speed and probability; the number of doctors affects the speed of admission and the treatment of patients; GDP is a measure of economic performance, which affects both the level of government assistance and the

Table 2.4: MSE and RMSE (empirical standard errors are in brackets) of the proposed two-step matrix completion algorithm (TSMC), the multiple imputation (MI) and the traditional matrix completion algorithm (TMC) when the random error term contains serial correlation, and (a) the dimension of response matrix is 1000 × 100, the missing rate is 30%, and the missingness mechanism is MCAR; (b) the dimension of response matrix is 1000 × 30, the missing rate is 50%, and the missingness mechanism is MAR; (c) the dimension of response matrix is 100 × 30, the missing rate is 30%, and the missingness mechanism is MCAR.

		(a)	(b)	(c)
TSMC	MSE	0.007[0.002]	0.015[0.001]	0.015[0.008]
	RMSE	0.455[0.114]	0.615[0.034]	0.440[0.150]
MI	MSE	0.464[0.034]	0.470[0.050]	0.469[0.049]
	RMSE	1.431[0.094]	1.426[0.130]	1.426[0.140]
TMC	MSE	0.041[0.038]	0.035[0.057]	0.052[0.080]
	RMSE	1.035[0.080]	1.115[0.095]	1.090[0.166]

ability to deal with an emergency, and the distance from Wuhan influence the transmission of the virus among cities.

We also include five important time-specific covariates in the data: the weather condition, the daily confirmed cases in Wuhan, the day indicating the time that the pandemic lied in the study period, and the day effect, where the day effect is defined as a categorical variable indicating the time period the cities' lockdown; the weather condition includes wind speed and temperature in Wuhan, which affects the social activities and the population flow, and the number of daily confirmed cases in Wuhan is included as the early cases in the other cities could be traced back to the patients out of Wuhan in the early stage.

For the Covid-19 dataset, it has been recognized that with high probability, the records of infections may be subject to missingness or contamination. For example, Hao et al. [2020] stated that the estimated ascertained rate of the confirmed cases is 87% in Wuhan. On one hand, it is inevitable that the individuals with asymptotic symptoms or in the incubation period are unlikely to be recorded. On the other hand, the missingness may occur due to concern for self-isolated infected patients providing the data during the panic time at the beginning of the pandemic. Besides, the improper collection of data, as well as the limited capacity of testing can also result in potential missingness. In the application, we consider the daily confirmed cases that are equal to zero as missing values. It is plausible because we observe that in most of the cities, the number of recorded cases starts from zero and then goes up to higher values. At the beginning of the pandemic, it is counter-intuitive that the number of cases suddenly drops down to zero. Therefore, it is possible that the zero values after non-zero values are incorrectly recorded. Moreover, the imputed values that are close to zero will be approximated as zero, because the number of cases is an integer. If the neighbor values are zero, then the estimated value would be close to zero and be approximated as zero. To model the missingness mechanism, in Figure 2.1, we present the proportion of daily non-zero confirmed cases in different tiers of the cities divided according to the GDP across the time periods that are partitioned by the time-specific indicator variable. Notice that Tier-1 cities have the highest GDP, and there are four of them in China; Tier-1.5 cities are the next highest in GDP, and there are 15 of them; there are 60 cities in Tier-2, and the rest of the cities are classified as lower tier cities. As shown in Figure 2.1, Tier-1.5 cities have the largest proportion of observations, and lower-tier cities have a relatively smaller proportion of observations. Besides, all the cities achieved their peak between the lockdown date of the Hubei province and mid-February. For cities in all tiers except Tier-1, the rate of observations first increases until mid-February, and then decreases afterward, indicating the effects of city lockdown as well as the other time-specific covariates on the spread of the virus. Notice that Tier-1.5 cities have a relatively larger rate of missingness because smaller cities have fewer health resources and lagging information. To reduce the computational burden, we include the indicator variable for tiers and the indicator variable for the time period as the explanatory variables in the logistic model for the missingness mechanism. The parameters and the matrix of the probability of the observations are estimated through the maximum likelihood approach, accordingly. It is also worth noting that the outcome data are count data, that is, all entries in the outcome matrix are positive integer numbers, and such data may need proper modeling. For example, Cao and Xie [2015] proposed two sets of efficient algorithms for recovering incomplete data sets under the Poisson measurements assumption. In the real data application, to avoid complex modeling and assumptions for the outcome data, together with that the fact that the proportion of the imputed entries that are negative is small ($< 1\%$) compared with the size of the whole dataset, we simply put the negative values to be equal to zero.

Figure 2.2 displays four cities' trajectories of daily confirmed cases. For Tier-1 cities such as Beijing and Shanghai, two curves have similar patterns. During the time period from Feb 2 to Feb 29, the number of confirmed cases increased before mid-February, reached a peak at around Feb 16, and then decreased afterward, which indicates the potential positive effects of the policies proposed by the Chinese government such as the cities' lockdown, the public health intervention, and better health resource allocation.

For estimation, we use 10-fold cross-validation to train the model and tune the hyperparameters. Since there is 31% missingness in the dataset, at each time, 30% of observations are chosen as the validation set, and the remaining observations are used for training. The unknown parameters in the missingness mechanism model are estimated from the training set via the maximum likelihood approach. The estimated main and interaction effects are summarized in Table 2.6, where the last row of the table displays the estimated main effects of unit-specific covariates, and the last column of the table presents the estimated main effects of time-specific covariates. Notice that the estimated effects of two indicator variables for city and province lockdown are negative, which indicates that the intervention of the lockdown is effective in preventing the transmission of people among the cities as well as the spread of the virus. Besides, the interaction effects between Wuhan's lockdown and the density are negative, but the interaction between Hubei's lockdown and the density is positive, indicating that Wuhan's lockdown may be more effective in reducing the number of cases infected. It is also worth noting that the effects of the distance from Wuhan are large compared with other effects, which implies that the cities far from Wuhan are less affected by the outbreak of the virus in Wuhan.

Figure 2.1: Plot of proportion of observed data for Tier-1,Tier-1.5, Tier-2, and lower-tier cities across four different pandemic time periods.

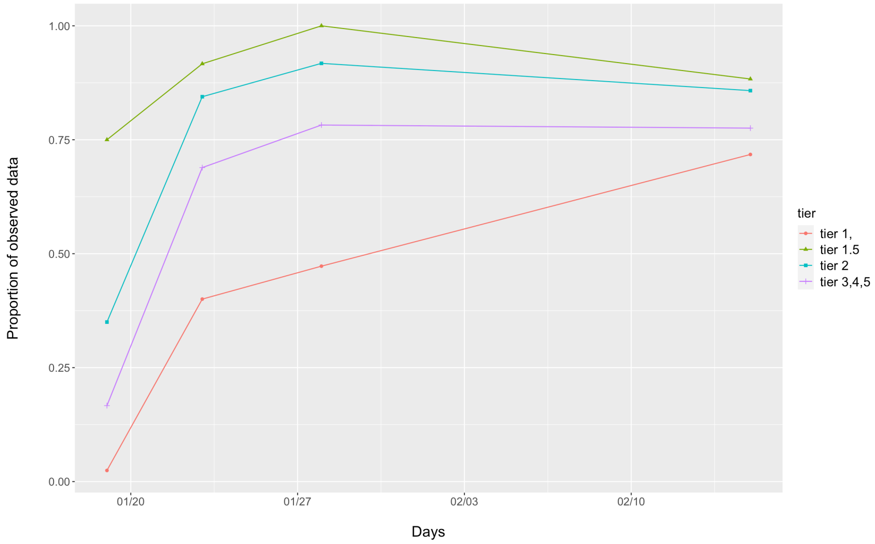


Figure 2.2: Curves of confirmed cases for the four example cities: Beijing, Shanghai, Qianjiang and Anqing, where the solid red line, green line, blue line, and purple line represent the city Anqing, Beijing, Qianjiang, and Shanghai, respectively.

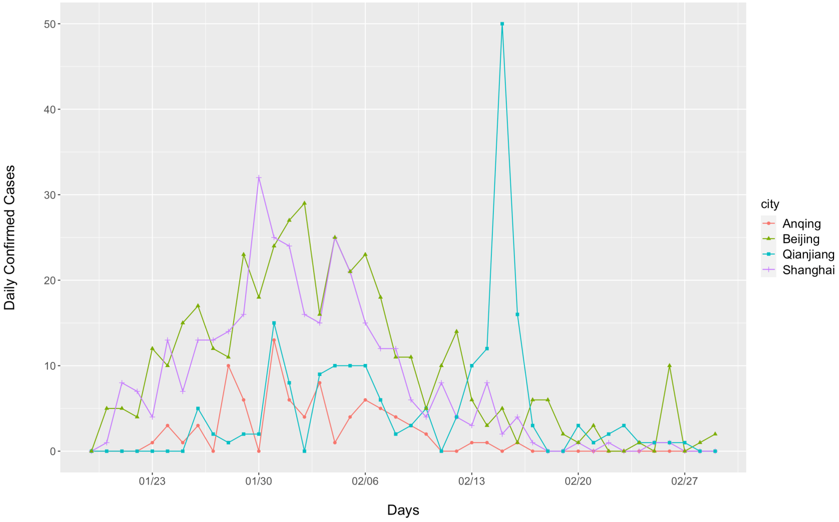


Figure 2.3: Curves of confirmed cases for the four example cities: Beijing, Shanghai, Qianjiang and Anqing, where the solid blue line represents the original daily confirmed cases over 42 days, the red solid line represents the 7-day moving average, and the green dashed line represents the predictions on each day using the proposed algorithm.

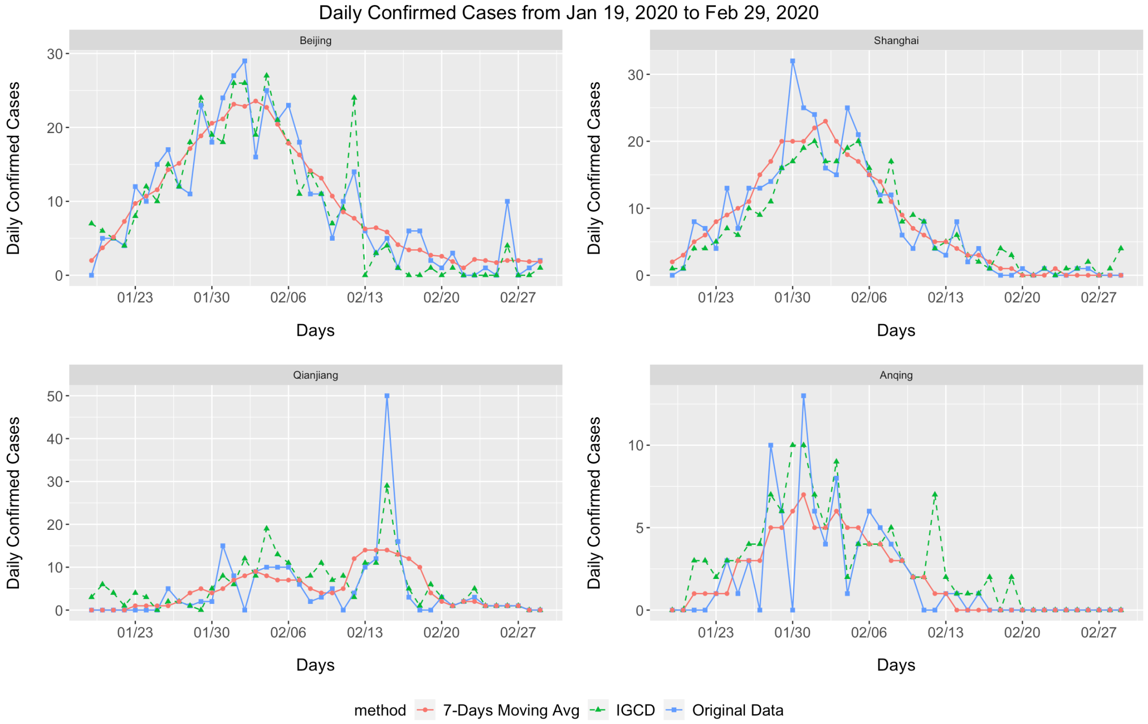


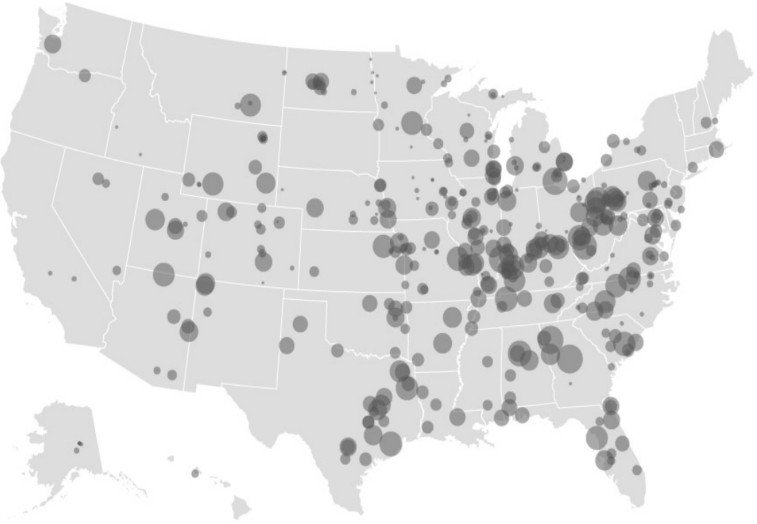
Table 2.5: Estimation of main and interaction effects of unit-specific covariates and time-specific covariates in Covid-19 data.

	Density	Hospital	GDP	Distance from Wuhan	**Main Effects**
Day	0.26	0	0.02	-0.13	1.21
Wuhan lockdown	-0.31	0	-0.12	-1.38	-1.77
Hubei lockdown	0.4	0	0.07	1.76	-1.96
Daily cases in Wuhan	0	-0.18	0.11	-0.45	0.49
Temperature	-0.19	0.2	0	-0.78	-0.64
Wind speed	0.8	-0.08	0.4	2.3	-2.48
Main Effects	0	-0.74	0	-3.27	NA

2.6.2 SO_2 emissions data

In this section, we applied the proposed methods to power plants' SO_2 emissions dataset. The monthly emissions data are collected from 1256 coal-fired electricity-generating units (EGUs) across the U.S.A (Zigler et al. [2016]), where there is 16.26% missingness in the outcomes. Coal-fired power plants are one of the primary sources of electrical generation in 19 states in the U.S.A. Figure 2.4 shows the distribution of coal-fired power plants and the corresponding linked monitors. We include five unit-specific covariates: the heat input rate, the capacity of the EGU, the number of scrubbers (flue-gas desulfurization equipment or controls) installed, the sulfur content, and the operation time, as these characteristics may affect the change of the amount of emissions of SO_2. For example, the sulfur content refers to the amount of sulfur per ton of coals. The installation of scrubbers is a primary strategy for power plants to reduce the emissions of SO_2. The heat input rate and the capacity of the EGU are the measures of the power of the EGUs, which further affects the amount of pollution they can generate per unit of time. For the operation time, the more time the EGUs are operated, the more pollutants they may generate.

Figure 2.4: Map of the distribution of the power plants in the U.S.A. Each gray circle corresponds to the area covered by a monitor around a power plant.

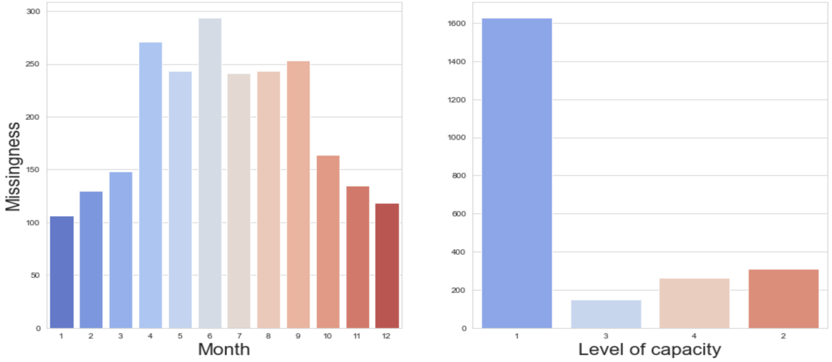


Source by: https://www.washingtonpost.com/graphics/national/power-plants.

We include three time-specific covariates from the dataset: the average temperature by month, the average precipitation by month, and the quarter indicator. The average temperature and precipitation may affect the spread of SO_2. The quarter indicator, which is defined as the categorical variable indicating the quarter that the month belongs to, is included because it may affect the investment of the power plants on the maintenance and update of the generators, which further affects the generation capacity of the power plants Henry and Pratson [2019].

For the SO_2 emissions data, the monthly records of the amount of emissions may be subject to missingness due to different reasons. For example, the missingness may be caused by monitor failures and errors, power outages, system crashes, undetectable pollutant levels, and filter changes (Imtiaz and Shah [2008]). To model the missingness mechanism, we include both the unit-specific and time-specific covariates as explanatory variables in the logistic model. The parameters and the matrix of the probability of the observations are estimated through the regularized maximum likelihood approach described in Section 2.3. The left subplot in Figure 2.5 shows the number of missing values across different months, where the second and the third quarter contain relatively more missing values, and the right subplot shows the number of unobserved values within different levels of capacity.

Figure 2.5: Plot of the number of unobserved data across 12 months, and for different levels of capacity.



The estimated main and interaction effects are summarized in Figure 2.6. Notice that the estimated main effect of the number of scrubbers is negative, which indicates that installing scrubbers has a positive effect on reducing the ambient SO_2 emissions. The estimated main effect of operation time is zero, which implies that it may have less effect on SO_2 emissions compared with the other unit-specific covariates. For time-specific covari-

Table 2.6: Estimation of main and interaction effects of unit-specific covariates and time-specific covariates in SO_2 emissions data.

	Heat input rate	Capacity	Scrubbers	Sulfer content	Operation time	**Main Effects**
Quarter 1	24.71	0	-7.83	24.37	10.44	-8.85
Quarter 2	11.53	4.64	12.40	34.82	-26.26	0
Quarter 3	0	13.32	-10.59	15.21	0	0
Quarter 4	-6.84	0	-54.36	9.51	8.54	4.89
Temperature	0	0	0	1.8	0	0
Precipitation	4.06	0	0	0	0	-8.35
Main Effects	44.96	149.50	-131.79	205.95	0	NA

ates, the estimated main effects of temperature and precipitation, and the interaction effect of the temperature and the precipitation with unit-specific covariates are relatively low, indicating that the two time-specific covariates may have relatively low effect on the ambient SO_2 emissions. Besides, the interaction effects between the number of scrubbers and the second quarter are positive, but the interaction effects between the number of scrubbers and the third quarter are negative, which indicates that the scrubbers' installation may have a greater effect on reducing the ambient SO_2 in the third quarter of the year than in the second quarter. It is also worth noting that the estimated main and interaction effects of weather conditions are smaller compared with the other covariates, which implies that recorded SO_2 emissions are less affected by the weather conditions. The cross-validated errors of the proposed algorithm, MI, and TMC are 0.75, 1.38, and 1.69, respectively.

2.7 Discussion

In this Chapter, we proposed a fixed effect low-rank model for longitudinal data and a corresponding iterative algorithm for imputing the missing outcomes. In Section 2.4, we showed the non-asymptotic error bounds for the estimated main effects, interaction effects, the low-rank term, and the imputed outcome matrix.

The novel feature of this paper lies in two main aspects. First, we consider the random error term as the summation of the serial correlation and the random noise and focus on estimating the fixed effects and the low-rank term. In Section 2.4, we showed that the non-asymptotic error bounds for the low-rank term as well as the imputed matrix are dependent on the magnitude of the variance of the random term. Notice that we used multivariate normal distribution as an example to illustrate our methods in Section 2.4 to derive the closed form of the upper bound, but it can be extended to other settings with various distributional assumptions. Second, we consider the MAR missingness mechanism and utilized the inverse probability weighting approach to reduce the bias.

One limitation of the proposed algorithm is that our method did not estimate the covariance matrix of the serial correlation. Incorporating the spatial correlation into consideration may improve the performance of our algorithm. Also, throughout the paper, we did not consider the inference problem, although the variation of the low-rank term and the fixed effect term can be measured using a re-sampling technique such as the bootstrap. However, the confidence interval may be too wide, and thus, be sub-optimal. When there is no covariate information and the data are missing completely at random, Chen et al. [2019] proposed a debiased estimator for the low-rank term. How to construct the confidence intervals for both the fixed effects and the low-rank term with missing not at random correlated data is left for future investigation. In data application, we treat the zero values as missing values, but the records with higher values could also be incorrect. Therefore, it is interesting to consider the measurement error instead of missing values in future research.

Incorporating the time-specific and unit-specific covariates is an important avenue for optimizing the conventional matrix completion algorithm when applying it to a longitudinal dataset. Beyond the unit-specific and time-specific covariates, it is also a promising direction to extend the proposed methods to a longitudinal dataset with unit-level time-varying covariates to recover the missing outcomes.

2.8 Proof of Theorems

This section contains all the proofs of Theorems and Lemmas.

Lemma 2.1, 2.2, and 2.3 are necessary conditions for Theorem 1. Since the updates of $\hat{H}$ and $\hat{L}$ in each iteration satisfy the line search condition, the objective function is always non-increasing after the updates of each of the parameters. The proof of Lemma 2.1 is adapted from Proposition 3.1 in Tseng and Yun [2009].

2.8.1 Proof of Lemma 2.1

Proof. Assume

$$d_H^{(n)} \in \underset{d \in R^{P \times Q}}{\arg\min} \left\{ F(H^{(n)} + d, L^{(n)}) + \lambda_H \|H + d\|_1 \right\},$$

and

$$d_L^{(n)} \in \arg\min_{d \in R^{N\times T}} \Big\{ F(H^{(n+1)}, L^{(n)} + d) + \lambda_L \|L + d\|_\star \Big\}.$$

Let α_1, $\alpha_2 \in (0,1)$ denote step sizes for H and L respectively, and $H^{(n+1)} = H^{(n)} + \alpha_1 d_H^{(n)}$. Then we have

$$\begin{aligned}
& F_{\bar{\lambda}}(H^{(n)} + \alpha_1 d_H^{(n)}, L^{(n)} + \alpha_2 d_L^{(n)}) - F_{\bar{\lambda}}(H^{(n)}, L^{(n)}) \\
= & \Big\{ F_{\bar{\lambda}}(H^{(n)} + \alpha_1 d_H^{(n)}, L^{(n)} + \alpha_2 d_L^{n}) - F_{\bar{\lambda}}(H^{(n)} + \alpha_1 d_H^{(n)}, L^{(n)}) \Big\} \\
& + \Big\{ F_{\bar{\lambda}}(H^{(n)} + \alpha_1 d_H^{(n)}, L^{(n)}) - F_{\bar{\lambda}}(H^{(n)}, L^{(n)}) \Big\}.
\end{aligned}$$

The remaining proof follows from the proof of Lemma 1 in Tseng and Yun [2009]. For the first term in the above equation, we have

$$\begin{aligned}
& F_{\bar{\lambda}}(H^{(n)} + \alpha_1 d_H^{(n)}, L^{(n)} + \alpha_2 d_L^{(n)}) - F_{\bar{\lambda}}(H^{(n)} + \alpha_1 d_H^{(n)}, L^{(n)}) \\
= & F(H^{(n+1)}, L^{(n)} + \alpha_L^{(n)} d_L^{(n)}) - F(H^{(n+1)}, L^{(n)}) + \lambda_L \Big\{ \|\alpha_2(L^{(n)} + d_L^{(n)}) + (1-\alpha_2) d_L^{(n)}\|_\star \Big\} \\
& -\lambda_L \|L^{(n)}\|_\star \\
\leq & -\frac{2}{NT} \sum_{i=1}^{N} \sum_{j=1}^{T} \frac{R_{ij}}{\Pr(R_{ij} = 1 | Y_i^O, X_{i.}, Z_{j.})} \Big\{ Y_{ij} - (XH^{(n+1)}Z^T + L^{(n)}) \Big\} \Big(\alpha_2 d_L^{(n)} \Big) \\
& +\lambda_L \Big\{ \alpha_2 \|L^{(n)} + d_L^{(n)}\|_\star + (1-\alpha_2)\|L^{(n)}\|_\star \Big\} - \lambda_L \|L^{(n)}\|_\star \\
= & \alpha_2 \Big[-\frac{2}{NT} \sum_{i=1}^{N} \sum_{j=1}^{T} \frac{R_{ij}}{\Pr(R_{ij} = 1 | Y_i^O, X_{i.}, Z_{j.})} \Big\{ Y_{ij} - (XH^{(n+1)}Z^T + L^{(n)}) \Big\} \Big(d_L^{(n)} \Big) \\
& +\lambda_L \Big\{ \|L^{(n)} + d_L^{(n)}\|_\star - \|L^{(n)}\|_\star \Big\} \Big].
\end{aligned}$$

By convexity of the nuclear norm and l_1 norm, we have

$$\begin{aligned}
& -\frac{2}{NT}\sum_{i=1}^{N}\sum_{j=1}^{T}\frac{R_{ij}}{\Pr(R_{ij}=1|Y_i^O, X_{i.}, Z_{j.})}\Big\{Y_{ij}-(XH^{(n+1)}Z^T+L^{(n)})\Big\}(d_L^n) \\
& +\lambda_L\Big\{\|L^{(n)}+d_L^{(n)}\|_\star-\|L^{(n)}\|_\star\Big\} \\
\leq \quad & -\frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{T}\frac{R_{ij}}{\Pr(R_{ij}=1|Y_i^O, X_{i.}, Z_{j.})}\Big(d_L^{(n)}\Big)^2 < 0.
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
& F_{\bar\lambda}(H^n+\alpha_1 d_H^n, L^n)-F_{\bar\lambda}(H^n, L^n) \\
\leq \quad & \alpha_1\Big[-\frac{2}{NT}\sum_{i=1}^{N}\sum_{j=1}^{T}\frac{R_{ij}}{\Pr(R_{ij}=1|Y_i^O, X_{i.}, Z_{j.})}\Big\{Y_{ij}-(XH^{(n)}Z^T+L^{(n)})\Big\}\Big(Xd_H^{(n)}Z^T\Big) \\
& +\lambda_H\Big\{\|H^{(n)}+d_H^{(n)}\|_1-\|H^{(n)}\|_1\Big\}\Big],
\end{aligned}$$

and

$$\begin{aligned}
& -\frac{2}{NT}\sum_{i=1}^{N}\sum_{j=1}^{T}\frac{R_{ij}}{\Pr(R_{ij}=1|Y_i^O, X_{i.}, Z_{j.})}\Big\{Y_{ij}-(XH^{(n)}Z^T+L^{(n)})\Big\}\Big(Xd_H^{(n)}Z^T\Big) \\
& +\lambda_H\Big\{\|H^{(n)}+d_H^{(n)}\|_1-\|H^{(n)}\|_1\Big\} \\
\leq \quad & -\frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{T}\frac{R_{ij}}{\Pr(R_{ij}=1|Y_i^O, X_{i.}, Z_{j.})}\Big(d_H^{(n)}\Big)^2 < 0.
\end{aligned}$$

□

2.8.2 Proof of Lemma 2.2

Proof. Since the objective function $F_{\bar\lambda}(H, L)$ is lower bounded by constant 0, together with the fact that $F_{\bar\lambda}(H, L)$ is continuous on $R^{P\times Q}\times R^{N\times T}$, then the level sets $lev(F_{\bar\lambda})$ are closed.

We now prove the boundedness of $lev(F_{\bar{\lambda}})$ by contradiction. Assume that $\forall\ r_0 > 0$, $\exists H^\dagger, L^\dagger$ s.t. $\|H^\dagger - H^0\|^2 + ||L^\dagger - L^0||^2 > r_0^2$. WLOG, assume $||H^\dagger - H^0||^2 > r_0^2/2$, then we have

$$\begin{aligned} F_{\bar{\lambda}}(H^\dagger, L^\dagger) &= \frac{1}{NT}\|\frac{1}{\Pr(R=1|Y^O, X, Z)}(XH^\dagger Z^T + L^\dagger - Y)\|_F^2 + \lambda_H\|H^\dagger\|_1 + \lambda_L\|L^\dagger\|_\star \\ &= \frac{1}{NT}\|\frac{1}{\Pr(R=1|Y^O, X, Z)}\{X(H^\dagger - H^0)Z^T + (L^\dagger - L^0) + (XH^0Z^T + L^0 - Y)\}\|_F^2 \\ &\quad +\lambda_H\|H^\dagger - H^0 + H^0\|_1 + \lambda_L\|L^\dagger - L^0 + L^0\|_\star \\ &> F_{\bar{\lambda}}(H^0, L^0), \end{aligned}$$

which contradicts with the definition of level sets $lev(F_{\bar{\lambda}})$. Thus, the level sets $lev(F_{\bar{\lambda}})$ are bounded.

□

2.8.3 Proof of Lemma 2.3

Proof. By Weierstrass's Theorem and Lemma 2.2, since the level sets $\{(H^{(n)}, L^{(n)}) \mid F_{\bar{\lambda}}(H^{(n)}, L^{(n)}) \leq F_{\bar{\lambda}}(H^0, L^0)\}$ are compact sets, and $F_{\bar{\lambda}}(H, L)$ is a continuous function, then $F_{\bar{\lambda}}(H, L)$ attains its minimum on this level set, that is, there exists $(H^\star, L^\star)$ such that $F_{\bar{\lambda}}(H^{(n)}, L^{(n)}) \geq F_{\bar{\lambda}}(H^\star, L^\star) \quad \forall \quad n \in \mathcal{Z}^+$. □

2.8.4 Proof of Theorem 2.4

Proof. Suppose $(H^\dagger, L^\dagger)$ is an accumulation point of $(H^{(n)}, L^{(n)})$, let $(H^{(n_k)}, L^{(n_k)})$ be a subsequence of $(H^{(n)}, L^{(n)})$ that converges to $(H^\dagger, L^\dagger)$. Thus, the subsequence of search direction $(d_H^{(n_k)}, d_L^{(n_k)}) \to 0$, and we have $(d_H^\dagger, d_L^\dagger) = 0$, by Lemma 2 of Tseng and Yun [2009], $(H^\dagger, L^\dagger)$ is a stationary point of $F_{\bar{\lambda}}(H, L)$. Since $F_{\bar{\lambda}}(H^{(n)}, L^{(n)})$ is monotonically non-increasing and bounded from below, then by Lemma 2.3, $\lim_{n\to\infty} F_{\bar{\lambda}}(H^{(n_k)}, L^{(n_k)}) = F_{\bar{\lambda}}(H^\dagger, L^\dagger)$, i.e., the limit of objective function exists and must converge to the global optimum.

□

2.8.5 Proof of Lemma 2.5

Proof. Recall that $\epsilon_{ij} = \epsilon_{ij}^{(1)} + \epsilon_{ij}^{(2)} \quad \forall 1 \leq i \leq N, 1 \leq j \leq T$, where $\epsilon_{ij}^{(1)}$ stands for the serial correlation and $\epsilon_{ij}^{(2)}$ represents the random noise term.

$$
\begin{aligned}
& \Pr\Big\{ \| \sum_{i=1}^{N} \sum_{j=1}^{T} \epsilon_{ij} R_{ij} \mathcal{E}_{ij} \|_\infty > 2t \Big\} \\
= \ & \Pr\Big\{ \| \sum_{i=1}^{N} \sum_{j=1}^{T} (\epsilon_{ij}^{(1)} + \epsilon_{ij}^{(2)}) R_{ij} \mathcal{E}_{ij} \|_\infty > 2t \Big\} \\
\leq \ & \Pr\Big\{ \| \sum_{i=1}^{N} \sum_{j=1}^{T} \epsilon_{ij}^{(1)} R_{ij} \mathcal{E}_{ij} ||_\infty + \| \sum_{i=1}^{N} \sum_{j=1}^{T} \epsilon_{ij}^{(2)} R_{ij} \mathcal{E}_{ij} \|_\infty > 2t \Big\} \\
\leq \ & \Pr\Big\{ \| \sum_{i=1}^{N} \sum_{j=1}^{T} \epsilon_{ij}^{(1)} R_{ij} \mathcal{E}_{ij} \|_\infty > t \Big\} + \Pr\Big\{ \| \sum_{i=1}^{N} \sum_{j=1}^{T} \epsilon_{ij}^{(2)} R_{ij} \mathcal{E}_{ij} \|_\infty > t \Big\} \\
\leq \ & \Pr\Big\{ \max_{1 \leq i \leq N} | \sum_{j=1}^{T} \epsilon_{ij} | > t \Big\} + \Pr\Big\{ \max_{1 \leq i,j \leq N \vee T} |\epsilon_{ij}^{(2)}| > t \Big\} \\
\leq \ & 2N \exp\{-\frac{t^2}{2T\sigma^2}\} + 2NT \exp(-\frac{t^2}{2\tau^2}).
\end{aligned}
$$

Thus, with probability $1 - 2\exp(-t)$, we have

$$
\| \sum_{i=1}^{N} \sum_{j=1}^{T} \epsilon_{ij} R_{ij} \mathcal{E}_{ij} \|_\infty \leq 2\sqrt{\log(2NT) + t} \Big(\sqrt{T}\sigma + \tau \Big).
$$

□

2.8.6 Proof of Lemma 2.6

Proof. In order to prove Lemma 2.6, we need to utilize the following Bernstein inequality. Recall that ψ_1 norm is defined as follows,

$$U_i = inf\left\{K > 0 : Eexp(\frac{\|Z\|}{K}) \leq e\right\}.$$

Let $Z_1, Z_2 \cdots Z_n$ be i.i.d random matrices with dimension $m_1 \times m_2$ that satisfies $E(Z) = 0$. Suppose that $U_i < U$ for some constant U, then there exists a constant $c^\star > 0$ such that for all $t > 0$, with probability $1 - e^{-t}$,

$$\|\sum_{i=1}^{n} Z_i\|_{op} \leq c^\star \max\left\{\sigma_z\sqrt{t + log(m_1 + m_2)}, U\log(\frac{U}{\sigma_z})(t + \log(m_1 + m_2))\right\},$$

where

$$\sigma_z = \max\left\{\|\sum_{k=1}^{n} E(Z_k Z_k^T)\|_{op}, \|\sum_{k=1}^{n} E(Z_k^T Z_k)\|_{op}\right\},$$

and

$$\|Z\|_{op} = \lambda_{max}(Z).$$

Then, we have the following representation for the stochastic error:

$$\begin{aligned}
Z_i &= \sum_{j=1}^{T} \epsilon_{ij} \frac{R_{ij}}{\Pr(R_{ij} = 1|Y_i^O, X_{i.}, Z_{j.})}\mathcal{E}_{ij} \\
&= \sum_{j=1}^{T} (\epsilon_{ij}^{(1)} + \epsilon_{ij}^{(1)}) \frac{R_{ij}}{\Pr(R_{ij} = 1|Y_i^O, X_{i.}, Z_{j.})}\mathcal{E}_{ij} \\
&= \sum_{j=1}^{T} \epsilon_{ij}^{(1)} \frac{R_{ij}}{\Pr(R_{ij} = 1|Y_i^O, X_{i.}, Z_{j.})}\mathcal{E}_{ij} + \sum_{j=1}^{T} \epsilon_{ij}^{(2)} \frac{R_{ij}}{\Pr(R_{ij} = 1|Y_i^O, X_{i.}, Z_{j.})}\mathcal{E}_{ij} \\
&= Z_i^{(1)} + Z_i^{(2)},
\end{aligned}$$

where $Z_i^{(1)} = \sum_{j=1}^T \epsilon_{ij}^{(1)} \frac{R_{ij}}{\Pr(R_{ij}=1|Y_i^O, X_{i.}, Z_{j.})} \mathcal{E}_{ij}$ and $Z_i^{(2)} = \sum_{j=1}^T \epsilon_{ij}^{(2)} \frac{R_{ij}}{\Pr(R_{ij}=1|Y_i^O, X_{i.}, Z_{j.})} \mathcal{E}_{ij}$. Notice that

$$\begin{aligned} \| \sum_{i=1}^N \sum_{j=1}^T \epsilon_{ij} \frac{R_{ij}}{\Pr(R_{ij}=1|Y_i^O, X_{i.}, Z_{j.})} \mathcal{E}_{ij} \|_{op} &= \| \sum_{i=1}^N Z_i \|_{op} \\ &\leq \| \sum_{i=1}^N Z_i^{(1)} \|_{op} + \| \sum_{i=1}^N Z_i^{(2)} \|_{op} \end{aligned}$$

In the remaining proof of this Lemma, we aim to provide upper bounds for both $\| \sum_{i=1}^N Z_i^{(1)} \|_{op}$ and $\| \sum_{i=1}^N Z_i^{(2)} \|_{op}$. Let $K_1 = 16\sigma^2 e \max\{T\log(N), N\log(T)\}/p_1$. Then, we have

$$\begin{aligned} & E\exp\left\{ \frac{\| \sum_{j=1}^T \epsilon_{ij}^{(1)} \frac{R_{ij}}{\Pr(R_{ij}=1|Y_i^O, X_{i.}, Z_{j.})} \mathcal{E}_{ij} \|_{op}}{K_1} \right\} \\ \leq\ & E\exp(\frac{\| \sum_{j=1}^T \epsilon_{ij}^{(1)} R_{ij} \mathcal{E}_{ij} \|_{op}}{K_1 p_1}) \\ \leq\ & E\exp(\frac{\sum_{j=1}^T \| \epsilon_{ij}^{(1)} R_{ij} \mathcal{E}_{ij} \|_{op}}{K_1 p_1}) \\ \leq\ & E\exp(\frac{\mathbf{1}^T \cdot \epsilon_{i.}^{(1)}}{K_1 p_1}) \\ =\ & \exp\left\{ \frac{\mathbf{1}^T \Sigma \mathbf{1}}{2(K_1 p_1)^2}) \right\} \\ \leq\ & \exp\left\{ \frac{T^2\sigma^2}{2K_1^2 p_1^2} \right\} \leq e, \end{aligned}$$

and

$$
\begin{aligned}
\| \sum_{i=1}^{N} E\{Z_i^{(1)}(Z_i^{(1)})^T\}\|_{op} &= \| \sum_{i=1}^{N} E(\sum_{j=1}^{T} \frac{R_{ij}}{\mathcal{P}_{ij}} \epsilon_{ij}^2 e_i(N) e_i(N)^T)\|_{op} \\
&\leq E(\max_{1\leq j\leq T} \frac{\epsilon_{ij}^2}{4\sigma^2}) \frac{4\sigma^2}{p_1} \| \sum_{i=1}^{N} E\{\sum_{j=1}^{T} R_{ij} e_i(N) e_i^T(N)\}\|_{op} \\
&\leq \frac{8N}{p_1} \log E\left[\exp\left\{\frac{1}{4} \max_{1\leq j\leq T} (\frac{\epsilon_{ij}^2}{\sigma^2})\right\}\right] \\
&\leq \frac{8N}{p_1} \log E\left[\max_{1\leq j\leq T} \exp\left\{\frac{1}{4}(\frac{\epsilon_{ij}^2}{\sigma^2})\right\}\right] \\
&\leq \frac{8N}{p_1} \log \sum_{1\leq j\leq T} E\left[\exp\left\{\frac{1}{4}(\frac{\epsilon_{ij}^2}{\sigma^2})\right\}\right] \\
&= \frac{8N}{p_1} \log \sum_{1\leq j\leq T} \frac{1}{\sqrt{1 - 4\cdot\frac{1}{2}}} \\
&\leq \frac{16N\log(T)\sigma^2}{p_1}.
\end{aligned}
$$

Similarly, we can get the following upper bound for $\| \sum_{i=1}^{N} E\{(Z_i^{(1)})^T(Z_i^{(1)})\}\|_{op}$:

$$
\| \sum_{i=1}^{N} E\{(Z_i^{(1)})^T(Z_i^{(1)})\}\|_{op} \leq \frac{16T\log(N)\sigma^2}{p_1}.
$$

Thus, we have

$$
\sigma_{Z^{(1)}} \leq \frac{16\max\{T\log(N), N\log(T)\}\sigma^2}{p_1^2}.
$$

By applying the matrix version of Bernstein inequality, there exists a constant $c_1^\star$, such that with probability $1-e^{-t}$,

$$
\begin{aligned}
\| \sum_{i=1}^{N} Z_i^{(1)}\|_{op} \leq\ & c_1^\star \max\left\{ \frac{16\max\{T\log(N), N\log(T)\}\sigma^2}{p_1^2} \sqrt{t + log(N+T)}, \right. \\
& \left. \frac{16\sigma^2 e \max\{T\log(N), N\log(T)\}}{p_1} \Big(t + log(N+T) \Big) \right\}.
\end{aligned}
$$

Let $K_2 = 2\max(N,T)\sqrt{T}\tau/p_1$, since $\{\epsilon_{ij}^{(2)}\}_{1\leq i\leq N, 1\leq j\leq T}$ are i.i.d. $\tau-$Sub Gaussian random variables, we have

$$\begin{aligned}
&E\exp\bigg\{\frac{\|\sum_{j=1}^{T}\epsilon_{ij}^{(2)}\frac{R_{ij}}{\Pr(R_{ij}=1|Y_i^O,X_{i.},Z_{j.})}\mathcal{E}_{ij}\|_{op}}{K_2}\bigg\} \\
\leq\ & \prod_{i=1}^{N}\bigg\{E\exp\bigg(\frac{\|\epsilon_{ij}^{(2)}R_{ij}\mathcal{E}_{ij}\|_{op}}{K_2 p_1}\bigg)\bigg\} \\
\leq\ & \exp\bigg(\frac{T\tau^2}{2K_2p_1}\bigg)\leq e,
\end{aligned}$$

and

$$\begin{aligned}
\|\sum_{i=1}^{N}E\{Z_i^{(2)}(Z_i^{(1)})^T\}\|_{op} &= \|\sum_{i=1}^{N}E(\sum_{j=1}^{T}\frac{R_{ij}}{\mathcal{P}_{ij}}\epsilon_{ij}^2 e_i(N)e_i(N)^T)\|_{op} \\
&\leq \bigg\|\frac{1}{p_1^2}\sum_{i=1}^{N}\sum_{j=1}^{T}E\bigg\{R_{ij}e_i(N)e_i^T(N)\bigg\}E\bigg(\epsilon_{ij}^2\bigg)\bigg\|_{op} \\
&\leq \bigg\|\frac{\sigma^2}{p_1}\sum_{i=1}^{N}\sum_{j=1}^{T}E\bigg\{R_{ij}e_i(N)e_i^T(N)\bigg\}\bigg\|_{op} \\
&\leq \frac{2N\tau^2}{p_1}.
\end{aligned}$$

Similarly, we have

$$\|\sum_{i=1}^{N}E\{Z_i^{(2)}(Z_i^{(1)})^T\}\|_{op}\leq\frac{2T\tau^2}{p_1}.$$

Thus,

$$\sigma_{Z^{(2)}}\leq 2\tau^2\max(N,T)/p_1.$$

Applying the matrix Bernstein Inequality, there exists a constant $c_2^\star$ such that with probability $1-e^{-t}$,

$$\begin{aligned}\|\sum_{i=1}^{N} Z_i^{(2)}\|_{op} &\leq c_2^\star \max\Big\{\frac{2\max(N,T)\tau^2}{p_1}\sqrt{t+log(N+T)}, \\ &\qquad \frac{2\max(N,T)\sqrt{T}\tau}{p_1}\log\Big(\frac{\sqrt{T}p_1}{\tau}\Big)\Big(t+log(N+T)\Big)\Big\}.\end{aligned}$$

Combining the above inequalities, with probability $1-e^{-t}$,

$$\begin{aligned}&\|\sum_{i=1}^{N} Z_i\|_{op} \leq (c_1^\star + c_2^\star)\max\Big[\Big\{\frac{16\max\{T\log(N), N\log(T)\}\sigma^2 + 2\max(N,T)\tau^2}{p_1}\Big\}\cdot \\ &\sqrt{t+log(N+T)}, \\ &\Big\{\frac{16\sigma^2 e\max\{T\log(N), N\log(T)\}}{p_1} + \frac{2\max(N,T)\sqrt{T}\tau}{p_1}\log\Big(\frac{\sqrt{T}p_1}{\tau}\Big)\Big\}\cdot \\ &\Big(t+log(N+T)\Big)\Big].\end{aligned}$$

□

2.8.7 Proof of Theorem 2.7

Proof. Notice that

$$\begin{aligned}F_{\bar\lambda}(H,L) &= F(H,L) + \lambda_H\|H\|_1 + \lambda_L\|L\|_\star \\ &= \frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{T}\frac{R_{ij}}{\hat{\Pr}(R_{ij}=1|Y_i^O, X_{i.}, Z_{j.})}\{(XHZ^T)_{ij} + L_{ij} - Y_{ij}\}^2 + \lambda_H\|H\|_{1,e} + \lambda_L\|L\|_\star.\end{aligned}$$

Inspired by Theorem 1-4 in Koltchinskii et al. [2011] and main results in Klopp et al. [2014], this proof contains two main sections. In the first section, we show that $\|\hat{H}-H^\star\|_F^2$, $\|\hat{L}-L^\star\|_F^2$ and $\|\hat{M}-M^\star\|_F^2$ are upper bounded, where $\hat{H}$ and $\hat{L}$ are estimated parameters in the proposed algorithm, $H^\star$ and $L^\star$ are true parameters, and $\hat{M} = X\hat{H}Z^T + \hat{L}$, $M^\star = XH^\star Z^T + L^\star$. In the second part, we present that under some mild conditions, $E[\|\hat{H} - H^\star\|_F^2]$ and $E[\|\hat{L}-L^\star\|_F^2]$ are upper bounded by some constants with high probability.

By definition of $\hat{H}$ and $\hat{L}$, we have

$$F(\hat{H},\hat{L}) + \lambda_H \|\hat{H}\|_1 + \lambda_L \|\hat{L}\|_\star \leq F(H^\star, L^\star) + \lambda_H \|H^\star\|_1 + \lambda_L \|L^\star\|_\star.$$

Rearranging terms in the above inequality, we have

$$\begin{aligned} & \frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{T}\frac{R_{ij}}{\hat{\Pr}(R_{ij}=1|Y_i^O, X_{i.}, Z_{j.})}\{Y_{ij}-\hat{M}_{ij}\}^2 \\ - & \frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{T}\frac{R_{ij}}{\hat{\Pr}(R_{ij}=1|Y_i^O, X_{i.}, Z_{j.})}\{Y_{ij}-M^\star_{ij}\}^2 \\ \leq & \lambda_H(\|H^\star\|_1 - \|\hat{H}\|_1) + \lambda_L(\|L^\star\|_\star - \|\hat{L}\|_\star), \end{aligned}$$

and

$$\begin{aligned} & \frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{T}\frac{R_{ij}}{\hat{\Pr}(R_{ij}=1|Y_i^O, X_{i.}, Z_{j.})}\langle M^\star - \hat{M}, \mathcal{E}_{ij}\rangle^2 \\ \leq & -\frac{2}{NT}\sum_{i=1}^{N}\sum_{j=1}^{T}\frac{R_{ij}}{\hat{\Pr}(R_{ij}=1|Y_i^O, X_{i.}, Z_{j.})}\epsilon_{ij}\langle M^\star - \hat{M}, \mathcal{E}_{ij}\rangle \\ & +\lambda_H(\|H^\star\|_1 - \|\hat{H}\|_1) + \lambda_L(\|L^\star\|_\star - \|\hat{L}\|_\star) \\ = & -\frac{2}{NT}\sum_{i=1}^{N}\sum_{j=1}^{T}\frac{R_{ij}}{\hat{\Pr}(R_{ij}=1|Y_i^O, X_{i.}, Z_{j.})}\epsilon_{ij}\langle (XH^\star Z^T) - (X\hat{H}Z^T), \mathcal{E}_{ij}\rangle \\ & -\frac{2}{NT}\sum_{i=1}^{N}\sum_{j=1}^{T}\frac{R_{ij}}{\hat{\Pr}(R_{ij}=1|Y_i^O, X_{i.}, Z_{j.})}\epsilon_{ij}\langle L^\star - \hat{L}, \mathcal{E}_{ij}\rangle \\ & +\lambda_H(\|H^\star\|_1 - \|\hat{H}\|_1) + \lambda_L(\|L^\star\|_\star - \|\hat{L}\|_\star), \end{aligned}$$

where ϵ is the random error matrix. Let

$$\begin{aligned} U_1 = & -\frac{2}{NT}\sum_{i=1}^{N}\sum_{j=1}^{T}\frac{R_{ij}}{\hat{\Pr}(R_{ij}=1|Y_i^O, X_{i.}, Z_{j.})}\epsilon_{ij}\langle (XH^\star Z^T) - (X\hat{H}Z^T), \mathcal{E}_{ij}\rangle \\ & +\lambda_H(\|H^\star\|_1 - \|\hat{H}\|_1), \end{aligned}$$

and

$$U_2 \;\; = \;\; -\frac{2}{NT}\sum_{i=1}^{N}\sum_{j=1}^{T}\frac{R_{ij}}{\hat{\Pr}(R_{ij}=1|Y_i^O, X_{i.}, Z_{j.})}\epsilon_{ij}\langle L^\star - \hat{L}, \mathcal{E}_{ij}\rangle + \lambda_L(\|L^\star\|_\star - \|\hat{L}\|_\star).$$

By duality of trace norm, we have

$$\begin{aligned} U_2 \;\; &= \;\; -\frac{2}{NT}\sum_{i=1}^{N}\sum_{j=1}^{T}\frac{R_{ij}}{\Pr(R_{ij}=1|Y_i^O, X_{i.}, Z_{j.})}\epsilon_{ij}\langle L^\star - \hat{L}, \mathcal{E}_{ij}\rangle \\ &\quad +\lambda_L(\|L^\star\|_\star - \|\hat{L}\|_\star) \\ &\leq \;\; \frac{2}{NT}\|\Delta L\|_\star\|\frac{\sum_{i=1}^{N}\sum_{j=1}^{T}\epsilon_{ij}R_{ij}\mathcal{E}_{ij}}{\hat{\mathcal{P}}}\|_{op} + \lambda_L(\|L^\star\|_\star - \|\hat{L}\|_\star). \end{aligned}$$

Assume

$$\lambda_L \geq 4\|\frac{\sum_{i=1}^{N}\sum_{j=1}^{T}\epsilon_{ij}R_{ij}\mathcal{E}_{ij}}{\hat{\mathcal{P}}}\|_{op}/(NT).$$

Then,

$$U_2 \leq \frac{3}{2}\lambda_L\|\Delta L\|_\star \leq 6\lambda_L\sqrt{2r_L}\|\Delta L\|_F.$$

Also, by duality between l_1 norm and l_∞ norm, we have the following upper bound for U_1:

$$\begin{aligned} U_1 \;\; &= \;\; -\frac{2}{NT}\sum_{i=1}^{N}\sum_{j=1}^{T}\frac{R_{ij}}{\hat{\Pr}(R_{ij}=1|Y_i^O, X_{i.}, Z_{j.})}\epsilon_{ij}\langle(XH^\star Z^T) - (X\hat{H}Z^T), \mathcal{E}_{ij}\rangle \\ &\quad +\lambda_H(\|H^\star\|_1 - \|\hat{H}\|_1) \\ &\leq \;\; \frac{2}{NT}\|\frac{\sum_{i=1}^{N}\sum_{j=1}^{T}\epsilon_{ij}R_{ij}\mathcal{E}_{ij}}{\hat{\mathcal{P}}}\|_\infty\|X\Delta HZ^T\|_1 + \lambda_H(\|H^\star\|_1 - \|\hat{H}\|_1) \\ &\leq \;\; \frac{2}{NT}C_xC_z\|\frac{\sum_{i=1}^{N}\sum_{j=1}^{T}\epsilon_{ij}R_{ij}\mathcal{E}_{ij}}{\hat{\mathcal{P}}}\|_\infty\|\Delta H\|_1 + \lambda_H(\|H^\star\|_1 - \|\hat{H}\|_1). \end{aligned}$$

Assume

$$\lambda_H \geq \frac{2}{NT}C_xC_z\|\frac{\sum_{i=1}^{N}\sum_{j=1}^{T}\epsilon_{ij}R_{ij}\mathcal{E}_{ij}}{\hat{\mathcal{P}}}\|_\infty.$$

Then,

$$\begin{aligned} U_1 &\leq \frac{2}{NT} C_x C_z \| \frac{\sum_{i=1}^N \sum_{j=1}^T \epsilon_{ij} R_{ij} \mathcal{E}_{ij}}{\hat{\mathcal{P}}} \|_\infty \|\Delta_H\|_1 + \lambda_H \|\Delta_H\|_1 \\ &\leq 2\lambda_H \|\Delta H\|_1. \end{aligned}$$

Combining inequalities for U_1 and U_2, we have

$$\begin{aligned} & \frac{1}{NT} \left\| \frac{R(\hat{M} - M^\star)}{\hat{\mathcal{P}}} \right\|_F^2 \\ = & \frac{1}{NT} \sum_{i=1}^N \sum_{j=1}^T \frac{R_{ij}}{\hat{\Pr}(R_{ij} = 1 \mid X_{i.}, Z_{j.})} \langle M^\star - \hat{M}, \mathcal{E}_{ij} \rangle^2 \\ \leq & 6\lambda_L \sqrt{2r_L} \|\Delta L\|_F + 2\lambda_H \|\Delta H\|_1. \end{aligned}$$

□

2.8.8 Proof of Lemma 2.8

Proof. By optimality condition in Theorem 3.1.24 in Nesterov [2018], $\exists g_1^\star \in \partial \|H^\star\|_1$, $g_2^\star \in \partial \|L^\star\|_\star$ such that

$$\langle \frac{2}{NT} \sum_{i=1}^N \sum_{j=1}^T \frac{R_{ij}}{\mathcal{P}_{ij}} (X\hat{H}Z^T + \hat{L} - Y), X(H^\star - \hat{H})Z^T \rangle + \lambda_H \langle g_1^\star, H^\star - \hat{H} \rangle \geq 0.$$

Since $\langle g_1^\star, \hat{H} - H^\star \rangle \geq \|\hat{H}\|_1 - \|H^\star\|_1$, we have

$$
\begin{aligned}
\lambda(\|\hat{H}\|_1 - \|H^\star\|_1) &\leq \langle \frac{2}{NT} \sum_{i=1}^{N} \sum_{j=1}^{T} \frac{R_{ij}}{\mathcal{P}_{ij}} (X\hat{H}Z^T + \hat{L} - Y), X\Delta H Z^T \rangle \\
&= \langle \frac{2}{NT} \sum_{i=1}^{N} \sum_{j=1}^{T} \frac{R_{ij}}{\mathcal{P}_{ij}} \{(X\hat{H}Z^T + \hat{L}) - (XH^\star Z^T + \hat{L})\}, X\Delta H Z^T \rangle \\
&\quad + \langle \frac{2}{NT} \sum_{i=1}^{N} \sum_{j=1}^{T} \frac{R_{ij}}{\mathcal{P}_{ij}} \{(XH^\star Z^T + \hat{L}) - (XH^\star Z^T + L^\star)\}, X\Delta H Z^T \rangle \\
&\quad + \langle \frac{2}{NT} \sum_{i=1}^{N} \sum_{j=1}^{T} \frac{R_{ij}\epsilon_{ij}\mathcal{E}_{ij}}{\mathcal{P}_{ij}}, X\Delta H Z^T \rangle \\
&\leq \frac{2}{NT} \sum_{i=1}^{N} \sum_{j=1}^{T} \frac{-R_{ij}}{\mathcal{P}_{ij}} \|\langle X\Delta H Z^T, \mathcal{E}_{ij} \rangle\|^2 \\
&\quad + \frac{2}{NT} \sum_{i=1}^{N} \sum_{j=1}^{T} \|X\Delta H Z^T\|_1 \frac{2C_L}{p_1} \\
&\quad + \frac{2}{NT} \sum_{i=1}^{N} \sum_{j=1}^{T} \|X\Delta H Z^T\|_1 \| \frac{\sum_{i=1}^{N} \sum_{j=1}^{T} R_{ij}\epsilon_{ij}\mathcal{E}_{ij}}{\mathcal{P}_{ij}} \|_\infty \\
&\leq \frac{2}{NT} \sum_{i=1}^{N} \sum_{j=1}^{T} C_x C_z \|\Delta H\|_1 \frac{2C_L}{p_1} \\
&\quad + \frac{2}{NT} \sum_{i=1}^{N} \sum_{j=1}^{T} C_x C_z \|\Delta H\|_1 \| \frac{\sum_{i=1}^{N} \sum_{j=1}^{T} R_{ij}\epsilon_{ij}\mathcal{E}_{ij}}{\mathcal{P}_{ij}} \|_\infty.
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
&\lambda(\|\hat{H}\|_1 - \|H^\star\|_1) \\
\leq\ & \frac{2}{NT} C_x C_z (\|\hat{H}\|_1 + \|H^\star\|_1) \bigg(\frac{2C_L}{p_1} + \| \frac{\sum_{i=1}^{N} \sum_{j=1}^{T} R_{ij}\epsilon_{ij}\mathcal{E}_{ij}}{\mathcal{P}_{ij}} \|_\infty \bigg).
\end{aligned}
$$

Let $\lambda_H \geq 3 \frac{2}{NT} C_x C_z \bigg(\frac{2C_L}{p_1} + \| \frac{\sum_{i=1}^{N} \sum_{j=1}^{T} R_{ij}\epsilon_{ij}\mathcal{E}_{ij}}{\mathcal{P}_{ij}} \|_\infty \bigg)$. Then, we have $\|\hat{H}\|_1 \leq 2\|H^\star\|_1$.

Let $\tilde{M}^{(1)} = XH^\star Z^T + \hat{L}$, and $\tilde{M}^{(2)} = X\hat{H}Z^T + L^\star$, we have

$$\begin{aligned}
\frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{T}\frac{R_{ij}}{\mathcal{P}_{ij}}\Big(\langle \tilde{M}^{(1)} - \hat{M}, \mathcal{E}_{ij})\rangle\Big) &\leq -\frac{2}{NT}\sum_{i=1}^{N}\sum_{j=1}^{T}\frac{R_{ij}}{\mathcal{P}_{ij}}\langle X(H^\star - \hat{H})Z^T, \epsilon_{ij}\mathcal{E}_{ij}\rangle \\
&\quad +\lambda(\|H^\star\|_1 - \|\hat{H}\|_1) \\
&\leq \lambda(\|H^\star\|_1 + \|\hat{H}\|_1) + \lambda_H\|H^\star\|_1 \\
&= 4\lambda_H.
\end{aligned}$$

Notice that

$$\begin{aligned}
LHS &\geq \frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{T}\frac{R_{ij}}{\mathcal{P}_{ij}}\Big(\langle \tilde{M}^{(1)} - \hat{M}, \mathcal{E}_{ij}\rangle\Big) \\
&= \frac{1}{2}\frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{T}\frac{R_{ij}}{\mathcal{P}_{ij}}\Big(\langle \tilde{M}^{(1)} - M^\star + M^\star - \hat{M}, \mathcal{E}_{ij})\rangle\Big) \\
&\geq \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{T}\frac{R_{ij}}{\mathcal{P}_{ij}}\Big(\langle \tilde{M}^{(1)} - M^\star, \mathcal{E}_{ij}\rangle^2 + \langle M^\star - \hat{M}, \mathcal{E}_{ij}\rangle^2\Big) \\
&\geq \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{T}\frac{R_{ij}}{\mathcal{P}_{ij}}\langle X\Delta HZ^T, \mathcal{E}_{ij}\rangle^2.
\end{aligned}$$

Combining the above inequalities, we have

$$\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{T}\frac{R_{ij}}{\mathcal{P}_{ij}}\langle X\Delta HZ^T, \mathcal{E}_{ij}\rangle^2 \leq 4\lambda_H\|H^\star\|_1$$

□

2.8.9 Proof of Theorem 2.9

Proof. Define $\|B\|_{L_2(\Pi)}$ as

$$\|B\|^2_{L_2(\Pi)} = E(\sum_{i=1}^{N}\sum_{j=1}^{T} R_{ij}\langle B, \mathcal{E}_{ij}\rangle^2).$$

Recall that the constraint set $\mathcal{C}_H(\theta_H)$ is defined as follows,

$$\mathcal{C}_H(\theta_H) := \left\{ H \in R^{P\times Q} \middle| \|H\|_1 \leq 1, \|H\|^2_{L_2(\Pi)} \geq \theta_H \right\}.$$

Let $\theta_H = 8c_H^2 \log(N{+}T)/\log(6/5)$. If $\|\Delta H\|^2_{L_2(\Pi)} > \theta_H$, then $\Delta H/(3a_H \vee c_H) = \Delta H/(3a_H) \in \mathcal{C}_H(\theta_H)$, and $\|\Delta H/(3a_H)\|_1 \leq 1$. Denote ζ_{ij} be i.i.d. Rademacher random variables. In the remaining of this proof, we show that with high probability, the following probabilistic upper bounds holds,

$$\Pr\left\{ \frac{p_1}{2}\|X\Delta HZ^T\|_F^2 > \sum_{i=1}^{N}\sum_{j=1}^{T} R_{ij}\langle X\Delta HZ^T, \mathcal{E}_{ij}\rangle^2 + \frac{c_H^\star}{p_1}E(\|\sum_{i=1}^{N}\sum_{j=1}^{T}\zeta_{ij}R_{ij}\mathcal{E}_{ij}\|_\infty)^2 \right\}$$
$$\leq \quad 2\exp(-c_H^\star log(\frac{6}{5})\theta_H),$$

where $c_L^\star$ is a constant. Define the bad event as

$$\mathcal{B} := \left\{ \exists A \in \mathcal{C}_H(\theta_H) \ s.t. \ \|XAZ^T\|^2_{L_2(\Pi)} - \sum_{i=1}^{N}\sum_{j=1}^{T} R_{ij}\langle XAZ^T, \mathcal{E}_{ij}\rangle^2 > \frac{1}{2}\|XAZ^T\|^2_{L_2(\Pi)} + \theta_H \right\}.$$

To make progress, we need to bound the probability of this bad event. Let $\xi = \frac{6}{5}$, we first define the subset of constraint set as

$$\mathcal{C}'(\theta_H, K) := \left\{ A \in \mathcal{C}_H(\theta_H) \mid K \leq \|XAZ^T\|^2_{L_2(\Pi)} \leq \xi K \right\}.$$

Next, define the subset of bad event $\mathcal{B}$ as

$$\mathcal{B}_l := \left\{ \exists A \in \mathcal{C}(\theta_H, K) \ s.t. \ \|XAZ^T\|^2_{L_2(\Pi)} - \sum_{i=1}^{N}\sum_{j=1}^{T} R_{ij}\langle XAZ^T, \mathcal{E}_{ij}\rangle^2 > \frac{1}{2}\|XAZ^T\|^2_{L_2(\Pi)} + \theta_H \right\}.$$

Then, $\mathcal{C}_H(\theta_H) = \bigcup_{l=1}^{\infty} \mathcal{C}'(\theta_H, \xi^{l-1}\theta_H)$, if $\exists A \in \mathcal{C}_H(\theta_H)$, then $\exists l$ s.t. $A \in \mathcal{C}'(\theta_H, \xi^{l-1}\theta_H)$. Define

$$Z_K := \sup_{A \in \mathcal{C}'(r,\theta_L,K)} \left\{ \|XAZ^T\|_{L_2(\Pi)}^2 - \sum_{i=1}^{N}\sum_{j=1}^{T} R_{ij}\langle XAZ^T, \mathcal{E}_{ij}\rangle^2 \right\},$$

and

$$\tilde{Z}_K := \sup_{A \in \mathcal{C}'(r,\theta_L,K)} \left\{ \left| \|XAZ^T\|_{L_2(\Pi)}^2 - \sum_{i=1}^{N}\sum_{j=1}^{T} R_{ij}\langle XAZ^T, \mathcal{E}_{ij}\rangle^2 \right| \right\}.$$

We aim to prove the following inequality via Massart's Inequality:

$$\Pr\left\{ Z_K \geq \frac{5\xi K}{24} + 8C_xC_zE\left(\left\| \sum_{i=1}^{N}\sum_{j=1}^{T} \zeta_{ij}R_{ij}\mathcal{E}_{ij} \right\| \right) \right\} \leq exp\{-c * log(\xi)\theta_L\}.$$

Notice that $\Pr(Z_K > t) \leq \Pr(\tilde{Z}_K > t)$. Thus, if $\tilde{Z}_K$ holds for the above inequality, Z_K also satisfies it. To utilize Massart's Inequality, we need to bound the $E(\tilde{Z}_K)$ as well as the variance term $Var(\tilde{Z}_K)$.

By symmetrization argument and Talagrand's contraction inequality, we have

$$\begin{aligned}
E(\tilde{Z}_K) &\leq 2E\left\{ \sup_{A \in \mathcal{C}'(\theta_H,K)} \left| \sum_{i=1}^{N}\sum_{j=1}^{T} \zeta_{ij}R_{ij}\langle XAZ^T, \mathcal{E}_{ij}\rangle^2 \right| \right\} \\
&\leq 8E\left\{ \sup_{A \in \mathcal{C}'(\theta_H,K)} \left| \sum_{i=1}^{N}\sum_{j=1}^{T} \zeta_{ij}R_{ij}\langle XAZ^T, \mathcal{E}_{ij}\rangle \right| \right\} \\
&\leq 8E\left\{ \sup_{A \in \mathcal{C}'(\theta_H,K)} C_xC_z\|A\|_1 \left\| \sum_{i=1}^{N}\sum_{j=1}^{T} \zeta_{ij}R_{ij}\mathcal{E}_{ij} \right\|_\infty \right\} \\
&\leq 8E\left\{ \sup_{A \in \mathcal{C}'(\theta_H,K)} C_xC_z \left\| \sum_{i=1}^{N}\sum_{j=1}^{T} \zeta_{ij}R_{ij}\mathcal{E}_{ij} \right\|_\infty \right\} \\
&\leq 8C_xC_zE\left\{ \left\| \sum_{i=1}^{N}\sum_{j=1}^{T} \zeta_{ij}R_{ij}\mathcal{E}_{ij} \right\|_\infty \right\}.
\end{aligned}$$

For the variance term,

$$
\begin{aligned}
\sup_{A\in\mathcal{C}'(r,\theta_L,K)} Var(\tilde{Z}_K) &= \sup_{A\in\mathcal{C}'(\theta_H,K)} Var\Big\{\|XAZ^T\|^2_{L_2(\Pi)} - \sum_{i=1}^{N}\sum_{j=1}^{T} R_{ij}\langle XAZ^T,\mathcal{E}_{ij}\rangle^2\Big\} \\
&\le \sup_{A\in\mathcal{C}'(\theta_H,K)} E\Big\{\Big\|\sum_{i=1}^{N}\sum_{j=1}^{T} R_{ij}\langle XAZ^T,\mathcal{E}_{ij}\rangle\Big\|^4\Big\} \\
&\le NT \sup_{A\in\mathcal{C}'(r,\theta_L,K)} E\Big\{\Big\|\sum_{i=1}^{N}\sum_{j=1}^{T} R_{ij}\langle XAZ^T,\mathcal{E}_{ij}\rangle\Big\|^2\Big\} \\
&\le NT(\xi^l\theta_L).
\end{aligned}
$$

Thus, by Massart's theorem, we have

$$
\begin{aligned}
&\Pr\Big\{\tilde{Z^K} > \frac{5}{24}\xi^l\theta_L + 8C_xC_zE\Big(\Big\|\sum_{i=1}^{N}\sum_{j=1}^{T}\zeta_{ij}R_{ij}\mathcal{E}_{ij}\Big\|_\infty\Big)^2\Big\} \\
\le\ & \exp(-c^\star\theta_L\xi^l).
\end{aligned}
$$

The union bound implies

$$
\begin{aligned}
\Pr\Big(\mathcal{B}\Big) &\le \sum_{l=1}^{\infty}\Big\{\Pr(\mathcal{B}_l)\Big\} \\
&\le \frac{\exp\{-c^\star\log(\xi)\theta_L\}}{1-\exp\{-c^\star\log(\xi)\theta_L\}} \\
&\le 2\exp\{-c^\star\log(\xi)\theta_L\} \\
&\le \frac{1}{(N+T)^2}.
\end{aligned}
$$

□

2.8.10 Proof of Lemma 2.10

Proof. By optimality condition, we have

$$\langle \frac{2}{NT}\sum_{i=1}^{N}\sum_{j=1}^{T}\frac{R_{ij}}{\mathcal{P}_{ij}}(X\hat{H}Z^T+\hat{L}-Y), L^\star-\hat{L}\rangle+\lambda_H\langle g_2^\star, L^\star-\hat{L}\rangle \geq 0.$$

Since $\langle g_2^\star, \hat{L}-L^\star\rangle \geq \|\hat{L}\|_\star-\|L^\star\|_\star$, we have

$$\begin{aligned}
\lambda_L(\|\hat{L}\|_\star-\|L^\star\|_\star) &\leq \langle \frac{2}{NT}\sum_{i=1}^{N}\sum_{j=1}^{T}\frac{R_{ij}}{\mathcal{P}_{ij}}(X\hat{H}Z^T+\hat{L}-Y), \Delta L\rangle \\
&= \langle \frac{2}{NT}\sum_{i=1}^{N}\sum_{j=1}^{T}\frac{R_{ij}}{\mathcal{P}_{ij}}\{(X\hat{H}Z^T+\hat{L})-(X\hat{H}Z^T+L^\star)\}, \Delta L\rangle \\
&\quad +\langle \frac{2}{NT}\sum_{i=1}^{N}\sum_{j=1}^{T}\frac{R_{ij}}{\mathcal{P}_{ij}}\{(X\hat{H}Z^T+L^\star)-(XH^\star Z^T+L^\star)\}, \Delta L\rangle \\
&\quad +\langle \frac{2}{NT}\sum_{i=1}^{N}\sum_{j=1}^{T}\frac{R_{ij}\epsilon_{ij}\mathcal{E}_{ij}}{\mathcal{P}_{ij}}, \Delta L\rangle \\
&\leq \frac{2}{NT}\sum_{i=1}^{N}\sum_{j=1}^{T}\frac{-R_{ij}}{\mathcal{P}_{ij}}\|\langle \Delta L, \mathcal{E}_{ij}\rangle\|^2 \\
&\quad +\frac{2}{NT}\|\Delta L\|_\star\frac{2C_L\max(N,T)}{p_1} \\
&\quad +\frac{2}{NT}\|\Delta L\|_\star\|\frac{\sum_{i=1}^{N}\sum_{j=1}^{T}R_{ij}\epsilon_{ij}\mathcal{E}_{ij}}{\mathcal{P}_{ij}}\|_{op}.
\end{aligned}$$

Let $\lambda_L \geq 3\frac{1}{NT}\left(\frac{2C_L\max(N,T)}{p_1} + \|\frac{\sum_{i=1}^N\sum_{j=1}^T R_{ij}\epsilon_{ij}\mathcal{E}_{ij}}{\mathcal{P}_{ij}}\|_{op}\right)$. Then, we have

$$\begin{aligned}
\frac{2}{3}\lambda_L\|\Delta L\|_\star &\geq \lambda_L(\|\hat{L}\|_\star - \|L^\star\|_\star)\\
&= \lambda_L(\|L^\star + \mathcal{P}_{L^{\star\perp}}(\Delta L) + \mathcal{P}_{L^\star}(\Delta L)\|_\star - \|L^\star\|_\star)\\
&\geq \lambda_L(\|L^\star + \mathcal{P}_{L^{\star\perp}}(\Delta L)\|_\star - \|\mathcal{P}_{L^\star}(\Delta L)\|_\star - \|L^\star\|_\star)\\
&= \lambda_L(\|L^\star\|_\star + \|\mathcal{P}_{L^{\star\perp}}(\Delta L)\|_\star - \|\mathcal{P}_{L^\star}(\Delta L)\|_\star - \|L^\star\|_\star)\\
&= \lambda_L(\|\mathcal{P}_{L^{\star\perp}}(\Delta L)\|_\star - \|\mathcal{P}_{L^\star}(\Delta L)\|_\star),
\end{aligned}$$

and

$$3(\|\mathcal{P}_{L^{\star\perp}}(\Delta L)\|_\star - \|\mathcal{P}_{L^\star}(\Delta L)\|_\star) \leq 2(\|\mathcal{P}_{L^{\star\perp}}(\Delta L)\|_\star + \|\mathcal{P}_{L^\star}(\Delta L)\|_\star).$$

Together with the fact that $\|\mathcal{P}_{L^\star}(\Delta L)\|_F \leq \|\Delta L\|_F$ and $rank(\mathcal{P}_{L^\star}(\Delta L)) \leq 2r_{L^\star}$, we have

$$\|\Delta L\|_\star \leq \sqrt{32 r_{L^\star}}\|\Delta L\|_F.$$

Similarly, we can have the following upper bounds for $\frac{1}{NT}\frac{R_{ij}}{\mathcal{P}_{ij}}\langle\Delta L, \mathcal{E}_{ij}\rangle$:

$$\begin{aligned}
\frac{1}{NT}\sum_{i=1}^N\sum_{j=1}^T\frac{R_{ij}}{\mathcal{P}_{ij}}\left(\langle\tilde{M}^{(2)} - \hat{M}, \mathcal{E}_{ij})\rangle\right) &\leq -\frac{2}{NT}\sum_{i=1}^N\sum_{j=1}^T\frac{R_{ij}}{\mathcal{P}_{ij}}\langle L^\star - \hat{L}, \epsilon_{ij}\mathcal{E}_{ij}\rangle\\
&\quad +\lambda(\|L^\star\|_\star - \|\hat{L}\|_\star)\\
&\leq \|\Delta L\|_\star\frac{2}{NT}\|\frac{\sum_{i=1}^N\sum_{j=1}^T R_{ij}\epsilon_{ij}\mathcal{E}_{ij}}{\mathcal{P}_{ij}}\|_{op} + \lambda_L\|L^\star\|_\star\\
&= \|\Delta L\|_\star\left(\frac{2}{NT}\|\frac{\sum_{i=1}^N\sum_{j=1}^T R_{ij}\epsilon_{ij}\mathcal{E}_{ij}}{\mathcal{P}_{ij}}\|_{op} + \lambda_L\right)\\
&\leq 2\lambda_L\|\Delta L\|_\star.
\end{aligned}$$

Then,

$$\begin{aligned}
LHS &\geq \frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{T}\frac{R_{ij}}{\mathcal{P}_{ij}}\Big(\langle \tilde{M}^{(2)} - \hat{M}, \mathcal{E}_{ij}\rangle\Big) \\
&= \frac{1}{2}\frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{T}\frac{R_{ij}}{\mathcal{P}_{ij}}\Big(\langle \tilde{M}^{(2)} - M^{\star} + M^{\star} - \hat{M}, \mathcal{E}_{ij})\rangle\Big) \\
&\geq \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{T}\frac{R_{ij}}{\mathcal{P}_{ij}}\Big(\langle \tilde{M}^{(2)} - M^{\star}, \mathcal{E}_{ij}\rangle^2 + \langle M^{\star} - \hat{M}, \mathcal{E}_{ij}\rangle^2\Big) \\
&\geq \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{T}\frac{R_{ij}}{\mathcal{P}_{ij}}\langle \Delta L, \mathcal{E}_{ij}\rangle^2.
\end{aligned}$$

□

2.8.11 Proof of Theorem 2.11

Proof. Recall the definition of $\|B\|_{L_2(\Pi)}$ is

$$\|B\|^2_{L_2(\Pi)} = E(\sum_{i=1}^{N}\sum_{j=1}^{T} R_{ij}\langle B, \mathcal{E}_{ij}\rangle^2),$$

and the constraint set $\mathcal{C}_L(r, \theta_L)$ is defined as follows:

$$\mathcal{C}_L(r, \theta_L) := \Big\{ L \in R^{N\times T} \|L\|_\infty \leq 1, \|L\|^2_{L_2(\Pi)} \geq \theta_L, \|L\|_\star \leq \sqrt{r}\|L\|_F \Big\}.$$

Let $\eta = 32 r_{L^\star}$, $\theta_L = 8c_L^2 \log(N+T)/\log(6/5)$, and ζ_{ij} be i.i.d. Rademacher random variables. For the rest of the proof, we show that with high probability, the following probabilistic upper bounds holds,

$$\begin{aligned}
&\Pr\Big\{\frac{p_1}{2}\|\Delta L\|_F^2 > \sum_{i=1}^{N}\sum_{j=1}^{T} R_{ij}\langle \Delta L, \mathcal{E}_{ij}\rangle^2 + \frac{c_L^\star r_{L^\star}}{p_1}E(\|\sum_{i=1}^{N}\sum_{j=1}^{T}\zeta_{ij}R_{ij}\mathcal{E}_{ij}\|_{op})^2\Big\} \\
\leq\ & 2\exp(-c_L^\star log(\frac{6}{5})\theta_L),
\end{aligned}$$

where $c_L^\star$ is a constant. Define the bad event as

$$\mathcal{B} := \left\{\exists A \in \mathcal{C}(r, \theta_L) \ s.t. \ \|A\|_{L_2(\Pi)}^2 - \sum_{i=1}^{N}\sum_{j=1}^{T} R_{ij}\langle A, \mathcal{E}_{ij}\rangle^2 > \frac{1}{2}\|A\|_{L_2(\Pi)}^2 + \theta_L\right\}.$$

Similarly, as in the proof of Theorem 3, to proceed, we need to bound the probability of this bad event. Let $\xi = \frac{6}{5}$, we first define the subset of constraint set as

$$\mathcal{C}'(r, \theta_L, K) := \left\{A \in \mathcal{C}_L(r, \theta_L) | K \leq \|A\|_{L_2(\Pi)}^2 \leq \xi K\right\}.$$

Next, we define the subset of bad event $\mathcal{B}$ as

$$\mathcal{B}_l := \left\{\exists A \in \mathcal{C}(r, \theta_L, K) \ s.t. \ \|A\|_{L_2(\Pi)}^2 - \sum_{i=1}^{N}\sum_{j=1}^{T} R_{ij}\langle A, \mathcal{E}_{ij}\rangle^2 > \frac{1}{2}\|A\|_{L_2(\Pi)}^2 + \theta_L\right\}.$$

Then, $\mathcal{C}(r, \theta_L) = \bigcup_{l=1}^{\infty} \mathcal{C}'(r, \theta_L, \xi^{l-1}\theta_L)$, if $\exists A \in \mathcal{C}(r, \theta_L)$, then $\exists l$ s.t. $A \in \mathcal{C}'(r, \theta_L, \xi^{l-1}\theta_L)$.

Define

$$Z_K := \sup_{A \in \mathcal{C}'(r, \theta_L, K)} \left\{\|A\|_{L_2(\Pi)}^2 - \sum_{i=1}^{N}\sum_{j=1}^{T} R_{ij}\langle A, \mathcal{E}_{ij}\rangle^2\right\},$$

and

$$\tilde{Z}_K := \sup_{A \in \mathcal{C}'(r, \theta_L, K)} \left\{\left|\|A\|_{L_2(\Pi)}^2 - \sum_{i=1}^{N}\sum_{j=1}^{T} R_{ij}\langle A, \mathcal{E}_{ij}\rangle^2\right|\right\}$$

We aim to prove the following inequality via Massart's Inequality:

$$\Pr\left(Z_K \geq \frac{5\xi K}{24} + \frac{c_L^\star r_{L^\star}}{p_1} E(\|\sum_{i=1}^{N}\sum_{j=1}^{T} \zeta_{ij} R_{ij}\mathcal{E}_{ij}\|_{op})^2\right) \leq exp\{-c * log(\xi)\theta_L\}.$$

Notice that $\Pr(Z_K > t) \leq \Pr(\tilde{Z}_K > t)$, thus if $\tilde{Z}_K$ holds for the above inequality, Z_K also satisfies it. To utilize Massart's Inequality, we need to bound the $E(\tilde{Z}_K)$ as well as the variance term $Var(\tilde{Z}_K)$.

By symmetrization argument and Talagrand's contraction inequality, we have

$$
\begin{aligned}
E(\tilde{Z}_K) &\leq 2E\Big\{\sup_{A\in\mathcal{C}'(r,\theta_L,K)}\Big|\sum_{i=1}^{N}\sum_{j=1}^{T}\zeta_{ij}R_{ij}\langle A,\mathcal{E}_{ij}\rangle^2\Big|\Big\} \\
&\leq 8E\Big\{\sup_{A\in\mathcal{C}'(r,\theta_L,K)}\Big|\sum_{i=1}^{N}\sum_{j=1}^{T}\zeta_{ij}R_{ij}\langle A,\mathcal{E}_{ij}\rangle\Big|\Big\} \\
&\leq 8E\Big\{\sup_{A\in\mathcal{C}'(r,\theta_L,K)}\|A\|_\star\Big\|\sum_{i=1}^{N}\sum_{j=1}^{T}\zeta_{ij}R_{ij}\mathcal{E}_{ij}\Big\|_{op}\Big\} \\
&\leq 8E\Big\{\sup_{A\in\mathcal{C}'(r,\theta_L,K)}\sqrt{\eta}||A||_F\Big\|\sum_{i=1}^{N}\sum_{j=1}^{T}\zeta_{ij}R_{ij}\mathcal{E}_{ij}\Big\|_{op}\Big\} \\
&\leq 8E\Big\{\sup_{A\in\mathcal{C}'(r,\theta_L,K)}\sqrt{\frac{\eta}{p_1}}||A||_F\Big\|\sum_{i=1}^{N}\sum_{j=1}^{T}\zeta_{ij}R_{ij}\mathcal{E}_{ij}\Big\|_{op}\Big\} \\
&= 8\sqrt{\frac{\eta}{p_1}}\sqrt{\xi K}E\Big\{\Big\|\sum_{i=1}^{N}\sum_{j=1}^{T}\zeta_{ij}R_{ij}\mathcal{E}_{ij}\Big\|_{op}\Big\} \\
&\leq \frac{1}{2}\Big\{\frac{5}{12}(\xi^l\theta_L)+\frac{64\eta}{p_1}\Big\|\sum_{i=1}^{N}\sum_{j=1}^{T}\zeta_{ij}R_{ij}\mathcal{E}_{ij}\Big\|_{op}\Big\}.
\end{aligned}
$$

For the variance term,

$$
\begin{aligned}
\sup_{A\in\mathcal{C}'(r,\theta_L,K)}Var(\tilde{Z}_K) &= \sup_{A\in\mathcal{C}'(r,\theta_L,K)}Var\Big\{\|A\|_{L_2(\Pi)}^2-\sum_{i=1}^{N}\sum_{j=1}^{T}R_{ij}\langle A,\mathcal{E}_{ij}\rangle^2\Big\} \\
&\leq \sup_{A\in\mathcal{C}'(r,\theta_L,K)}E\Big\{\Big\|\sum_{i=1}^{N}\sum_{j=1}^{T}R_{ij}\langle A,\mathcal{E}_{ij}\rangle\Big\|^4\Big\} \\
&\leq NT\sup_{A\in\mathcal{C}'(r,\theta_L,K)}E\Big\{\Big\|\sum_{i=1}^{N}\sum_{j=1}^{T}R_{ij}\langle A,\mathcal{E}_{ij}\rangle\Big\|^2\Big\} \\
&\leq NT(\xi^l\theta_L).
\end{aligned}
$$

Thus, by Massart's theorem, we have

$$\begin{aligned}
&\Pr\bigg\{\tilde{Z^K} > \frac{5}{24}\xi^l\theta_L + \frac{c_L^\star r_{L^\star}}{p_1}E(\|\sum_{i=1}^N\sum_{j=1}^T \zeta_{ij}R_{ij}\mathcal{E}_{ij}\|_{op})^2\bigg\} \\
\leq\ & \exp(-c^\star\theta_L\xi^l).
\end{aligned}$$

The union bound implies

$$\begin{aligned}
\Pr\bigg(\mathcal{B}\bigg) &\leq \sum_{l=1}^{\infty}\bigg\{\Pr(\mathcal{B}_l)\bigg\} \\
&\leq \frac{\exp\{-c^\star\log(\xi)\theta_L\}}{1-\exp\{-c^\star\log(\xi)\theta_L\}} \\
&\leq 2\exp\{-c^\star\log(\xi)\theta_L\} \\
&\leq \frac{1}{(N+T)^2}.
\end{aligned}$$

□

Chapter 3

Estimation of Network Causal Effects with Confounders Missing Not at Random

3.1 Introduction

There is limited work discussing the estimation of network treatment effect when confounders are subject to nonignorable missingness. Sun and Liu [2021] proposed a doubly robust estimator when data were subject to nonignorable missingness, where the data are assumed to be independent and the interference was ignored in the data application. Unlike Sun and Liu [2021], we study the partial interference setting and propose three pairs of semiparametric estimators: inverse probability weighting (IPW), regression, and doubly robust (DR) estimators for the four types of network treatment effects. Here, the regression estimator is an estimator for the average potential outcome in the causal inference framework and is different from that in survey sampling literature. Compared to the non-interference setting, there are several challenges in this new setting. For example, the IPW and DR estimators require the joint modeling of unit-level propensity scores, where the existence of extreme probabilities may lead to high-variance estimators. To circumvent the problem of varying cluster sizes, we propose self-normalized IPW and DR estimators, which can be viewed as stabilized versions of and have smaller variances than their respective conventional estimators. In this paper, we consider the direct interference pattern, where the outcome of one unit is affected by other units only through their treatment assignment, and the missingness of one unit is not affected by other units. To model the joint propensity score of missingness and treatment, we provide a concrete example in Section 3.3.1 to better illustrate the procedure of the model specification and the parameter estimation.

The rest of the chapter is organized as follows: we start by introducing the notation and assumptions in Section 3.2. The construction of the IPW estimator is presented in Section 3.3. The performance of the proposed methods is further illustrated via simulation studies in Section 3.4. Since this is ongoing work, additional simulation studies and analysis of real data are under construction. We leave the discussion of the construction of the regression estimator and the doubly robust estimator in Section 3.6.

3.2 Notation and Assumptions

Following Perez-Heydrich et al. [2014] and Tchetgen Tchetgen and VanderWeele [2012], we consider a finite population of size N, which can be partitioned into K mutually exclusive groups, and each group i has N_i units, where $1 \leq i \leq K$ and $1 \leq N_i \leq N - K + 1$. Let $X_{ij} = (X_{1ij}, X_{2ij}, \cdots X_{pij})$ denote p-dimensional confounders of unit j in the group i, the values of which may be subject to missingness, and let $X_i = (X_{i1}, X_{i2}, \cdots X_{iN_i})$ be the confounders of all units in the group i. Let $R_{ij} = 1$ if X_{ij} is complete and $R_{ij} = 0$ if X_{ij} is missing. In this paper, we consider the single missingness pattern, that is, $R_{ij} = 0$ if any components of the confounders of unit j in the group i are missing. We leave the extension to multiple missingness patterns as a future research direction. Let Y_{ij} and A_{ij} denote the observed outcome and treatment status for unit j in the group i, respectively, and $Y_i = (Y_{i1}, Y_{i2}, \cdots, Y_{iN_i})$ and $A_i = (A_{i1}, A_{i2}, \cdots, A_{iN_i})$ denote the vectors of observed outcome and treatment indicators, respectively, for all units in the group i. Assume $A_{i(-j)} = (A_{i1}, \cdots, A_{ij-1}, A_{ij+1}, \cdots A_{iN_i})$ is the vector of treatment status for all units in the group i except for individual j. Let a_{ij}, $a_{i(-j)}$ and a_i denote possible values of A_{ij} , $A_{i(-j)}$ and A_i, respectively. Suppose $\mathcal{A}(n)$ is the set of vectors of all possible treatment assignments of length n. Then, there are 2^{N_i} possible treatment assignments in group i, and $a_i \in \mathcal{A}(N_i)$.

We consider the scenario when only X is missing while the other variables are fully observed. Suppose the treatment is assigned with an α-strategy where every unit in a group is assigned to the treatment with average treatment allocation probability α. In a randomized trial, the probability of being treated is fully determined by the treatment allocation probability. However, in observational studies, whether or not an individual receives the treatment assignment is not only determined by the treatment allocation strategy but dependent on his/her choice of participation in the study. To avoid the additional modeling of participation (probability of the participation status given confounders), we will model the probability of the treatment indicator conditional on confounders directly. Let $Y_{ij}(a_i)$ and $Y_{ij}(a_{ij}, a_{i-j})$ denote the potential outcome for unit j in the group i under treatment allocation a_i, where $a_i = (a_{i1}, \ldots a_{i,j-1}, a_{ij}, a_{i,j+1}, \ldots a_{i,N_i})$, and denote $Y_i(a_i)$ as the vector of potential outcomes for all units in the group i under strategy α. Let $\bar{Y}_i(a, \alpha) = N_i^{-1} \sum_{j=1}^{N_i} \sum_{a_{i(-j)} \in \mathcal{A}(N_i-1)} Y_{ij}(a, a_{i(-j)}) P_{\alpha,x}(a_{i(-j)})$ denote the average potential outcome for group i, where $P_{\alpha,x}(a_{i(-j)}) = \Pr(A_{i(-j)} = a_{i(-j)} | A_{ij} = a_{ij}, X_i)$ is the probability of $A_{i(-j)} = a_{i(-j)}$ in group i. Let $\bar{Y}_i(\alpha) = N_i^{-1} \sum_{j=1}^{N_i} \sum_{a_i \in \mathcal{A}(N_i)} Y_{ij}(a_i) P_{\alpha,x}(a_{i(-j)})$ denote the marginal average potential outcome for group i.

There are different treatment allocation strategies that can be deployed. For example, in the cholera vaccine study Hudgens and Halloran [2008], $P_{\alpha,x}(a_{i(-j)}) = \prod_{j' \neq j} \alpha^{a_{ij'}} (1 - \alpha)^{1-a_{ij'}}$, that is, the treatment allocation strategy does not depend on individuals' characteristics. Once the individuals choose to participate in the trial, they are classified into different groups. For each group, the individuals are assigned the vaccine randomly with probability α. On the other hand, in the Acid Rain Program study, the intervention of scrubber installation is encouraged by federal regulations. However, the assignment of the intervention is also affected by the characteristics of power plants such as size and heat input. To account for the influence of confounders on the treatment allocation probability, Papadogeorgou et al. [2019] models $P_{\alpha,x}(a_{i(-j)})$ as $\text{logit}\{P_{\alpha,x}(a_{i(-j)})\} = \xi_i^\alpha + \delta X_i$, where $\xi_i^{(\alpha)}$ satisfying $(N_i^{-1}) \sum_{i=1}^{N_i} \text{expit}(\xi_i^{(\alpha)} + \delta X_i) = \alpha$ is a parameter to be estimated, and δ is some pre-fixed value, more details can be found in Section 3.5. We adopt the same modeling strategy in this paper.

The average potential outcome and the marginal average potential outcome are defined as $\mu_{a\alpha} = E(\bar{Y}_i(a, \alpha))$ and $\mu_\alpha = E(\bar{Y}_i(\alpha))$, respectively. Then, the direct effect (or the population average treatment effect) is defined as

$$\overline{\text{DE}}(\alpha) = E(\bar{Y}_i(1, \alpha) - \bar{Y}_i(0, \alpha)) = \mu_{1\alpha} - \mu_{0\alpha}. \tag{3.1}$$

For the acid rain program study, the direct effect represents the difference in the amount of PM2.5 emissions when a power generating facility is equipped with scrubbers compared to when scrubbers are not installed. For policies α_0 and α_1, the indirect effect is defined as

$$\overline{\text{IE}}(\alpha_0, \alpha_1) = E(\bar{Y}_i(0, \alpha_1) - \bar{Y}_i(0, \alpha_0)) = \mu_{0\alpha_1} - \mu_{0\alpha_0}, \tag{3.2}$$

which represents the difference in the amount of PM2.5 emissions when a power generating facility is not equipped under a different treatment allocation policy. The total effect is denoted by

$$\overline{\mathrm{TE}}(\alpha_1, \alpha_1) = E(\bar{Y}_i(1, \alpha_1) - \bar{Y}_i(0, \alpha_0)) = \mu_{1\alpha_1} - \mu_{0\alpha_0}, \tag{3.3}$$

which corresponds to the combination of both the direct effect and the indirect effect. The overall effect is defined as

$$\overline{\mathrm{OE}}(\alpha_0, \alpha_1) = E(\bar{Y}_i(\alpha_1) - \bar{Y}_i(\alpha_0)) = \mu_{\alpha_1} - \mu_{\alpha_0}, \tag{3.4}$$

which represents the difference in the amount of PM2.5 emissions for units under one coverage probability of scrubbers' installation compared to units with another level of coverage probability. In Section 3.3, we focus on estimating the direct effect. The estimation of the other network causal effects can be approached in a similar way, and we present the results of four types of network causal effects in the data application. For the purpose of estimation with incomplete confounders, we assume the type of interference to be direct interference (see more details in Ogburn et al. [2020]), where the interference happens only through the effect of the treatment assignment on other units in the same clusters. We leave more sophisticated interference types such as contagion interference and allocation interference as future research directions. Throughout this paper, we also make the following assumptions:

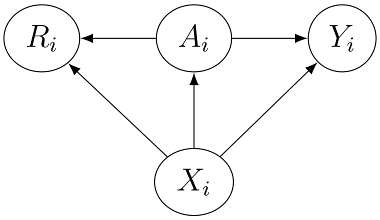
1. **Causal Consistency Assumption:** a subject's potential outcome under their observed treatment assignment is equal to the outcome that will actually be observed, that is, $Y_{ij} = \sum_{a_i \in \mathcal{A}(N_i)} 1(A_i = a_i) Y_{ij}(a_i)$.

2. **Exchangeability Assumption:** for each group, the treatment vector A_i is assumed to be conditionally independent with potential outcomes given confounders X_i, that is, $A_i \perp\!\!\!\perp Y_i(a_i) | X_i$.

3. **Positivity Assumption:** $\Pr(A_{ij} = a_{ij} | X_i) > 0$ and $\Pr(R_{ij} = 1 | A_i, X_i) > 0$ for all A_i and X_i.

It has been shown in Ding and Geng [2014] that without any assumptions, the joint distribution of (A_i, Y_i, X_i) is not fully identifiable. For identification purposes, we assume that the missingness mechanism for confounders X_i satisfies the group-level outcome-independent missingness assumption, which is modified from the outcome-dependent missingness assumption in Yang et al. [2019], that is, $R_i \perp\!\!\!\perp Y_i | A_i, X_i$. The associated causal diagram is shown in Figure 3.1. The assumption is plausible if the confounders are measured long before the outcome data are collected. For example, as mentioned in Yang et al. [2019], the potentially exposed children and their neighborhoods were more carefully measured than those that were not at risk of exposure in the water crisis study in Flint, Michigan U.S., which implies that the missingness may depend on both the measured confounders and the exposure status. In addition, the health status of the children was tested long after the confounders (e.g., age) were collected. Thus, the missingness of confounders is independent of the outcome conditional on all the other relevant information including observed confounders and exposure status.

Figure 3.1: The causal diagram for the group-level outcome independent missingness assumption, where the dashed line represents the conditional independence between Y_i and R_i, $1 \leq i \leq N_i$.



3.3 Estimation

3.3.1 Inverse Probability Weighting

In this section, we propose an IPW estimator for network treatment effects when confounders are subject to a nonignorable missingness mechanism. The idea of constructing the IPW estimator is to weigh each individual with the inverse of the probability of receiving the treatment, such that the association between confounders and treatment assignment can be removed. However, since X is not fully observed, $\Pr(A|X)$ is not estimable without further adjustment. Besides, even if the data are fully observed, we cannot directly apply the IPW as the data are not independent within the same group. To address this, we utilize the inverse of the group-level joint propensity score of treatment assignment and missingness mechanism as the weight for each subject, that is, $1/\Pr(A_i, R_{ij}|X_i), i = 1, 2, ...K, j = 1, 2, ...N_i$, and replace the number of individuals in each group by the sum of inverse joint propensity scores within the group.

To obtain the group-joint propensity score modeling for the treatment and missingness, we assume $P(R_{ij} = 1|A_{ij}, X_{ij})$ is correctly specified as $P(R_{ij} = 1|A_{ij}, X_{ij}; \gamma)$, and $P(A_{ij} = 1|R_{ij} = 1, X_{ij})$ is correctly specified as $P(A_{ij} = 1|R_{ij} = 1, X_{ij}; \delta)$. The unknown parameter γ can be estimated using generalized methods of moments, and δ can be estimated via the maximum likelihood approach. After obtaining the estimates of $P(R_{ij} = 1|A_{ij}, X_{ij})$ and $P(A_{ij} = 1|R_{ij} = 1, X_{ij})$, the joint density of treatment A and missingness R conditional on confounders X can be parameterized as follows Chen [2007]:

$$\Pr(a, r|x) = \frac{\psi(a, a_0, r, r_0|x) \Pr(r|a_0, x) \Pr(a|r_0, x)}{\sum_{r=0}^{1} \sum_{a=0}^{1} \psi(a, a_0, r, r_0|x) \Pr(r|a_0, x) \Pr(a|r_0, x)}, \tag{3.5}$$

where $r_0 = 1$, $a_0 = 1$, and

$$\psi(a, a_0, r, r_0|x) = \frac{\Pr(r|a, x) \Pr(r_0|a_0, x)}{\Pr(r|a_0, x) \Pr(r_0|a, x)}.$$

We then construct the IPW estimator for the population average potential outcome with estimated parameters as

$$\hat{\mu}_{a\alpha}^{ipw} = \frac{1}{K} \sum_{i=1}^{K} \hat{Y}_i^{IPW}(a, \alpha) = \frac{1}{K} \sum_{i=1}^{K} \frac{\sum_{j=1}^{N_i} \hat{w}_{ij}^{a\alpha} Y_{ij}}{\sum_{j=1}^{N_i} \hat{w}_{ij}^{a\alpha}}; \tag{3.6}$$

and estimator of the marginal average potential outcome. is defined as

$$\hat{\mu}_{\alpha}^{ipw} = \frac{1}{K} \sum_{i=1}^{K} \hat{Y}_i^{IPW}(\alpha) = \frac{1}{K} \sum_{i=1}^{K} \frac{\sum_{j=1}^{N_i} \hat{w}_{ij}^{\alpha} Y_{ij}}{\sum_{j=1}^{N_i} \hat{w}_{ij}^{\alpha}}, \tag{3.7}$$

where

$$\hat{w}_{ij}^{a\alpha} = \sum_{j=1}^{N_i} \frac{1(A_{ij} = a)1(R_{ij} = 1)P_{\alpha,x}(A_{i(-j)})}{\hat{\Pr}(A_i, R_{ij}|X_i; \hat{\delta}, \hat{\gamma})}, \tag{3.8}$$

$$\hat{w}_{ij}^{\alpha} = \sum_{j=1}^{N_i} \frac{1(A_{ij} = a)1(R_{ij} = 1)P_{\alpha,x}(A_i)}{\hat{\Pr}(A_i, R_{ij}|X_i; \hat{\delta}, \hat{\gamma})}, \tag{3.9}$$

where $\hat{Y}_i^{IPW}(a, \alpha) = \sum_{j=1}^{N_i} \hat{w}_{ij}^{\alpha} Y_{ij} / \sum_{j=1}^{N_i} \hat{w}_{ij}^{\alpha}$, and $\sum_{j=1}^{N_i} \hat{w}_{ij}^{\alpha} Y_{ij} / \sum_{j=1}^{N_i} \hat{w}_{ij}^{\alpha}$. Assume the estimation equations for δ and γ are $\sum_{i=1}^{K} \psi_\delta(O_i; \delta) = 0$ and $\sum_{i=1}^{K} \psi_\gamma(O_i; \gamma) = 0$, respectively, where $O_i = (X_i, Y_i, A_i, R_i)$. Let $\psi_a^{IPW}(O_i; \mu_{1\alpha}, \gamma, \delta) = \hat{Y}_i^{IPW}(1, \alpha) - \mu_{1\alpha}$, and $\psi_a^{IPW}(O_i; \mu_{0\alpha}, \gamma, \delta) = \hat{Y}_i^{IPW}(0, \alpha) - \mu_{0\alpha}$. Then the estimated parameter $\hat{\boldsymbol{\theta}}^{\boldsymbol{IPW}} = (\hat{\delta}, \hat{\gamma}, \hat{\mu}_{1\alpha}^{IPW}, \hat{\mu}_{0\alpha}^{IPW})$ is a solution of the following estimation equations:

$$\sum_{i=1}^{K} \boldsymbol{\psi}^{IPW}(O_i; \boldsymbol{\theta}) = \mathbf{0}, \tag{3.10}$$

where $\boldsymbol{\psi}^{IPW}(O_i;\boldsymbol{\theta}) = \{\psi_\delta(O_i;\delta), \psi_\delta(O_i;\gamma), \psi_a^{IPW}(O_i;\mu_{1\alpha},\gamma,\delta), \psi_a^{IPW}(O_i;\mu_{0\alpha},\gamma,\delta)\}^T$. The true values of unknown parameter $\boldsymbol{\theta} = (\delta,\gamma,\mu_{1\alpha},\mu_{0\alpha})$ is the solution to $\int \boldsymbol{\psi}^{IPW}(o;\theta)dF(o;\theta) = 0$, where F denotes the cumulative function of O_i. We show an example below of using the equations (3.6) and (3.7) to obtain the IPW estimators. For example, we may assume:

- $\text{logit}\{\Pr(A_{ij} = 1|X_{ij} = x_{ij}, R_{ij} = 1, b_i; \delta)\} = \delta_0 + \boldsymbol{\delta_1 x_{ij}} + b_i$,
- $\text{logit}\{\Pr(R_{ij} = 1|A_{ij} = a_{ij}, X_{ij} = x_{ij}; \gamma)\} = \gamma_0 + \boldsymbol{\gamma_1} a_{ij} + \boldsymbol{\gamma_2 x_{ij}}$,
- $Y_{ij} = \beta_0 + \beta_1 + \boldsymbol{\beta_2} \cdot (a_{ij}, f_a(a_{i(-j)}))^T + \beta_3 \boldsymbol{x_{ij}} + \zeta_i$,
- $\text{logit}P_{\alpha,x}(a_{i(-j)}) = \xi + \boldsymbol{\delta x_{ij}}$,

where $f_a(\cdot)$ is a summary function that represents the other units' effect on the same cluster, $\zeta_i \sim \mathcal{N}(0,\sigma_\zeta^2)$ and $b_i \sim \mathcal{N}(0,\sigma_b^2)$ represent the random effects that introduce the dependency between units in the same cluster. Assume $\hat{p}_{ar}(X_i) = \Pr(A_{ij'}, R_{ij}|X_{ij}, b_i, \hat{\delta}, \hat{\gamma})$. The estimate of the joint probability of $\Pr(A_i, R_{ij}|X_i)$ can be obtained by

$$\hat{\Pr}(A_i, R_{ij}|X_i, \hat{\delta}, \hat{\gamma}) = \int_{-\infty}^{\infty} \prod_{j'=1}^{N_i} \hat{p}_{ar}(X_i)^{A_{ij}} (1 - \hat{p}_{ar}(X_i))^{1-A_{ij}} f(b_i) db_i, \tag{3.11}$$

and $\hat{p}_{ar}(X_i)$ is obtained by equation (3.5). The consistency and asymptotic normality of the IPW estimator are presented in Theorem 3.1 below.

Theorem 3.1. *Under assumptions 1-3, if* $P(A_{ij} = 1|R_{ij} = 1, X_{ij}; \delta)$ *and* $P(R_{ij} = 1|A_{ij} = 1, X_{ij}; \gamma)$ *are correctly specified, then IPW estimator* $\hat{\mu}_{a\alpha}^{IPW}$ *is consistent for* $\mu_{a\alpha}$*, and*

$$\sqrt{K}(\hat{\boldsymbol{\theta}}^{\boldsymbol{IPW}} - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma^{IPW}),$$

as K *goes to infinity, where* $\Sigma^{IPW} = U(\boldsymbol{\theta})^{-1}V(\boldsymbol{\theta})\{U(\boldsymbol{\theta})^{-1}\}^T$, $U(\boldsymbol{\theta}) = E\{-\partial\boldsymbol{\psi}^{IPW}(O_i;\boldsymbol{\theta})/\partial\boldsymbol{\theta}^T\}$, *and* $V(\boldsymbol{\theta}) = E\{\boldsymbol{\psi}^{IPW}(O_i;\boldsymbol{\theta})\boldsymbol{\psi}^{IPW}(O_i;\boldsymbol{\theta})^T\}$.

The proof of Theorem 3.1 is given in Section 3.7.

3.3.2 Regression

In this section, we first introduce the idea of constructing the estimation equation for the regression estimator. Notice that by exchangeability assumption, we can estimate the causal effect by regressing Y on A and X, and then marginalize over X if confounders are fully observed. However, if confounders are subject to nonignorable missingness, we cannot directly estimate the causal effect because we do not know the distribution of the confounders in the whole population. To recover the full data distribution, we need to obtain the joint probability density of X and Y conditional on A and $R = 0$ for each group, which requires estimation of the odds ratio model and the group level joint probability density of X and Y conditional on A and $R = 1$. First, we model $f(Y_{ij}|A_i, X_{ij}, R_{ij} = 1)$ as $f(Y_{ij}|A_i, X_{ij}, R_{ij} = 1; \beta)$, and model $f(X_{ij}|A_i, R_{ij} = 1)$ as $f(X_{ij}|A_i, R_{ij} = 1; \beta)$. We then define the odds ratio function $OR(X, Y|A)$ as

$$\begin{aligned}
OR(X_{ij}, Y_{ij}|A_i) &= \log \frac{f(X_{ij}, Y_{ij}|A_i, R_{ij} = 0) f(X_{ij} = x_0, Y_{ij}|A_i, R_{ij} = 1)}{f(X_{ij}, Y_{ij}|A_i, R_{ij} = 1) f(X_{ij} = x_0, Y_{ij}|A_i, R_{ij} = 0)} \\
&= \log \frac{P(R_{ij} = 0|A_i, X_{ij}, Y_{ij}) P(R_{ij} = 1|A, X_{ij} = x_0, Y_{ij} = 0)}{P(R_{ij} = 1|A, X_{ij}, Y_{ij}) P(R_{ij} = 0|A_i, X_{ij} = x_0, Y_{ij} = 0)} \\
&= \log \frac{P(R_{ij} = 0|A_i, X_{ij}) P(R_{ij} = 1|A_i, X_{ij} = x_0)}{P(R_{ij} = 1|A_i, X_{ij}) P(R_{ij} = 0|A_i, X_{ij} = x_0)} \\
&= \log \frac{f(X_{ij}|A_i, R_{ij} = 0) f(X_{ij} = x_0|A_i, R_{ij} = 1)}{f(X_{ij}|A_i, R_{ij} = 1) f(X_{ij} = x_0|A_i, R_{ij} = 0)}, \qquad (3.12)
\end{aligned}$$

where $1 \leq i \leq K$, $1 \leq j \leq N_i$, and x_0 is an arbitrary fixed constant. For simplicity, we let $x_0 = 0$ in the subsequent sections. Since the last equation does not depend on Y, we can simplify the notation $OR(X_{ij}, Y_{ij}|A_i)$ as $OR(X_{ij}|A_i)$, which we model as $OR(X_{ij}|A_i; \zeta)$. Thus, we can parameterize the joint probability density of X and Y conditional on A and $R = 0$ as

$$f(X_{ij}, Y_{ij}|A_i, R_{ij} = 0; \beta, \zeta) = \frac{\exp\{OR(X_{ij}|A_i; \zeta)\} f(X_{ij}, Y_{ij}|A_i, R_{ij} = 1; \beta)}{E[\exp\{OR(X_{ij}|A; \zeta)\}|A_i, R_{ij} = 1; \beta]}. \qquad (3.13)$$

The parameter ζ can be estimated from the equation $\sum_{i=1}^{K} \psi_\zeta(O_i; \zeta) = 0$, where $\psi_\zeta(O_i; \zeta)$ has the following expression:

$$(1 - R_i)^T \left\{ l(A_i, Y_i) - E\{l(A_i, Y_i)|A_i, R_i = 0; \hat{\beta}, \zeta\} \right\}, \qquad (3.14)$$

where $l(A,Y)$ are pre-defined vectorized differentiable functions, and

$$E\{l(A_i,Y_{ij})|A_i,R_{ij}=0;\hat{\beta},\zeta\} = \frac{\int \exp\{OR(X_{ij}|A_i;\zeta)\} f(X_{ij},Y_{ij}|A_i,R_{ij}=1;\beta) l(A_i,Y_{ij}) dy}{E[\exp\{OR(X_{ij}|A_i;\zeta)\}|A_i,R_{ij}=1;\beta]}. \tag{3.15}$$

Then, $f(X_{ij},Y_{ij}|A_i;\beta,\zeta)$ can be obtained by $f(X_{ij},Y_{ij}|A_i,R_{ij}=0;\beta,\zeta)$ and $f(X_{ij},Y_{ij}|A_i,R_{ij}=1;\beta,\zeta)$ because R and A are fully observed. Note that $f(Y_{ij}|A_i,X_{ij};\beta,\zeta) \propto f(X_{ij},Y_{ij}|A_i;\beta,\zeta)$, thus we can obtain the regression estimators. More specifically, let $g_{ij}(a_i,x_i) = E(Y_{ij}|A_i = a_i, X_i = x_i, R_{ij}=1)$, then the regression estimators for the average potential outcome and the marginal average potential outcome have the following expressions:

$$\begin{aligned}
\hat{\mu}^{reg}_{a\alpha} &= \frac{1}{K}\sum_{i=1}^{K} \hat{Y}^{reg}_i(a,\alpha) \\
&= \frac{1}{K}\sum_{i=1}^{K}\Big[\frac{1}{N_i}\sum_{j=1}^{N_i}\sum_{a_{i(-j)}} 1(R_{ij}=1)\hat{g}_{ij}(a_i,X_i;\hat{\beta})P_{\alpha,x}(a_{i(-j)}) \\
&\quad + \frac{1}{N_i}\sum_{j=1}^{N_i}\sum_{a_{i(-j)}} 1(R_{ij}=0)P_{\alpha,x}(a_{i(-j)})E\{\hat{g}_{ij}(a_i,X_i;\hat{\beta})|A_i=a_i,R_{ij}=0;\hat{\beta},\hat{\zeta}\}\Big], \qquad (3.16)
\end{aligned}$$

and

$$\begin{aligned}
\hat{\mu}^{reg}_{\alpha} &= \frac{1}{K}\sum_{i=1}^{K} \hat{Y}^{reg}_i(\alpha) \\
&= \frac{1}{K}\sum_{i=1}^{K}\Big[\frac{1}{N_i}\sum_{j=1}^{N_i}\sum_{a_i} 1(R_{ij}=1)\hat{g}_{ij}(a_i,X_i;\hat{\beta})P_{\alpha,x}(a_i) \\
&\quad + \frac{1}{N_i}\sum_{j=1}^{N_i}\sum_{a_i} 1(R_{ij}=0)P_{\alpha,x}(a_i)E\{\hat{g}_{ij}(a_i,X_i;\hat{\beta})|A_i=a_i,R_{ij}=0;\hat{\beta},\hat{\zeta}\}\Big], \qquad (3.17)
\end{aligned}$$

where $\hat{\beta}$ is the MLE of β, and $\hat{\zeta}$ is an estimator of ζ by equation (3.14). Assume the estimation equation for β is $\sum_{i=1}^{K}\psi_{\beta}(O_i;\beta)=0$. Let $\psi^{reg}_a(O_i;\mu_{1\alpha},\zeta,\beta) = \hat{Y}^{reg}_i(1,\alpha)-\mu_{1\alpha}$, and $\psi^{reg}_a(O_i;\mu_{0\alpha},\zeta,\beta) = \hat{Y}^{reg}_i(0,\alpha)-\mu_{0\alpha}$ denote the estimation equations for $\mu_{1\alpha}$ and $\mu_{0\alpha}$, respectively. Then the estimated parameter $\hat{\boldsymbol{\theta}}^{\boldsymbol{reg}} = (\hat{\beta},\hat{\zeta},\hat{\mu}^{reg}_{1\alpha},\hat{\mu}^{reg}_{0\alpha})$ is a solution of the following estimation equations:

$$\sum_{i=1}^{K} \boldsymbol{\psi}^{reg}(O_i; \boldsymbol{\theta}) = \mathbf{0}. \tag{3.18}$$

The consistency of the regression estimators is shown in Theorem 3.2 below.

Theorem 3.2. *If the baseline outcome regression model $f(Y_{ij}|A_i, X_{ij}R_{ij} = 1; \beta)$ and the conditional distribution of observed confounders $f(X_{ij}|A_i, R_{ij} = 1)$ are correctly specified, then the regression estimator $\hat{\mu}^{reg}_{a\alpha}$ is consistent for $\mu_{a\alpha}$, and*

$$\sqrt{K}(\hat{\boldsymbol{\theta}}^{\boldsymbol{reg}} - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma^{reg}),$$

as K goes to infinity, where $\Sigma^{reg} = U(\boldsymbol{\theta})^{-1}V(\boldsymbol{\theta})\{U(\boldsymbol{\theta})^{-1}\}^T$, $U(\boldsymbol{\theta}) = E\{-\partial\boldsymbol{\psi}^{reg}(O_i; \boldsymbol{\theta})/\partial\boldsymbol{\theta}^T\}$, and $V(\boldsymbol{\theta}) = E\{\boldsymbol{\psi}^{reg}(O_i; \boldsymbol{\theta})\boldsymbol{\psi}^{reg}(O_i; \boldsymbol{\theta})^T\}$.

The proof of Theorem 3.2 is given in Section 3.7.

3.3.3 Doubly Robust Estimation

In this section, we aim to propose a DR estimator $\hat{\mu}_{a\alpha}$ in the sense that it is consistent if either the propensity score models or the baseline regression models, but not necessarily both, are correctly specified. The typical DR estimator involves two parts, where the first part is the regression term, and the second part is the inverse probability weighted residuals of the regression estimator. In our case, when confounders are missing not at random, both the propensity score of missingness and the outcome regression model contain the odds ratio model. Therefore, the specification of propensity score models and the regression models are not independent. Hence, the specification of the propensity score models and the regression model cannot be fully separated. More specifically, in our setting, the specification of both the propensity score of missingness and the modeling of probability $f(x_{ij}, y_{ij}|a_i, r_{ij} = 0)$ requires the modeling of the odds ratio $OR(x_{ij}|a_i)$, that is, $OR(x_{ij}|a_i)$ lies in the intersection of the IPW estimator and the regression estimator. Thus, to construct the DR estimator, we assume $OR(x|a; \zeta)$ is always correctly specified, and the proposed DR estimator is consistent if either the IPW models, $\Pr(a_{ij} = 1|r_{ij} = 1, x_{ij}; \delta)$ and $\Pr(r_{ij} = 1|a_{ij} = 1, x_{ij}; \gamma)$, are correctly specified or the outcome regression model conditional on the observed values, $f(y_{ij}|a_i, x_{ij}, r_{ij} = 1, \beta)$ and $f(x_{ij}|a_i, r_{ij} = 1, \beta)$, are correctly specified, where $\gamma = (\gamma^{'}, \xi)$. ξ is obtained by solving the equation $\sum_{i=1}^{K} \psi_\zeta(O_i; \xi) = 0$, where $\psi_\xi(O_i;\xi) = \mathbf{v}^T(A_i, R_i, X_i; \hat{\gamma}', \xi)\phi(A_i, Y_i; \hat{\beta}, \xi)$,

$$\mathbf{v}(A_i, R_i, X_i; \hat{\gamma}', \hat{\xi}) = \begin{bmatrix} \frac{R_{i1}}{\Pr(R_{i1}=1|A_{i1}, X_{i1}; \hat{\gamma}', \hat{\xi})} - 1 \\ \frac{R_{i2}}{\Pr(R_{i2}=1|A_{i2}, X_{i2}; \hat{\gamma}', \hat{\xi})} - 1 \\ \cdots \\ \frac{R_{iN_i}}{\Pr(R_{iN_i}=1|A_{iN_i}, X_{iN_i}; \hat{\gamma}', \hat{\xi})} - 1 \end{bmatrix}, and \tag{3.19}$$

$$\phi(A_i, Y_i; \hat{\beta}, \hat{\xi}) = \begin{bmatrix} l(A_{i1}, Y_{i1}) - \mathbf{E}\{l(A_{i1}, Y_{i1}) \mid A_i, R_{i1} = 0; \hat{\beta}, \hat{\xi}\} \\ l(A_{i2}, Y_{i2}) - \mathbf{E}\{l(A_{i2}, Y_{i2}) \mid A_i, R_{i2} = 0; \hat{\beta}, \hat{\xi}\} \\ \cdots \\ l(A_{iN_i}, Y_{iN_i}) - \mathbf{E}\{l(A_{iN_i}, Y_{iN_i}) \mid A_i, R_{iN_i} = 0; \hat{\beta}, \hat{\xi}\} \end{bmatrix} \tag{3.20}$$

Since the confounders are not fully observed, we replace the weights in the residuals of the regression estimator in the traditional DR estimator by the joint probability of A and R conditional on X, and the construction of the regression estimator term is similar as that in Section 3.3.2. Hence, the DR estimator has the following representation:

$$\begin{aligned} \hat{\mu}^{dr}_{a\alpha} =& \frac{1}{K}\sum_{i=1}^{K}\Bigg[\sum_{j=1}^{N_i}\Bigg[E\{\hat{h}^a_{ij}(A_i, X_i, Y_i)|A_i = a_i, R_{ij} = 0; \hat{\delta}, \hat{\gamma}, \hat{\beta}\} \\ &+ \frac{1(R_{ij} = 1)}{\Pr(R_{ij}|A_i = a_i, X_i; \gamma)}\Bigg\{\hat{h}^a_{ij}(A_i, X_i, Y_i) - E\{\hat{h}^a_{ij}(A_i, X_i, Y_i)|A_i = a_i, R_{ij} = 0; \hat{\delta}, \hat{\gamma}, \hat{\beta}\}\Bigg\}\Bigg]\Bigg], \end{aligned} \tag{3.21}$$

and

$$\begin{aligned} \hat{\mu}^{dr}_{\alpha} =& \frac{1}{K}\sum_{i=1}^{K}\Bigg[\sum_{j=1}^{N_i}\Bigg[E\{\hat{h}_{ij}(A_i, X_i, Y_i)|A_i = a_i, R_{ij} = 0; \hat{\delta}, \hat{\gamma}, \hat{\beta}\} \\ &+ \frac{1(R_{ij} = 1)}{\Pr(R_{ij}|A_i, X_i; \gamma)}\Bigg\{\hat{h}_{ij}(A_i, X_i, Y_i) - E\{\hat{h}_{ij}(A_i, X_i, Y_i)|A_i, R_{ij} = 0; \hat{\delta}, \hat{\gamma}, \hat{\beta}\}\Bigg\}\Bigg]\Bigg], \end{aligned} \tag{3.22}$$

where $h^a_{ij}(A_i, X_i, Y_i; \delta, \gamma, \beta)$, $h_{ij}(A_i, X_i, Y_i; \delta, \gamma, \beta)$ have the following expressions:

$$h^a_{ij}(A_i, X_i, Y_i; \delta, \gamma, \beta) = w^{a\alpha}_{ij}(Y_{ij} - g_{ij}(A_i, X_i; \beta)) + \frac{1}{N_i}\sum_{a_{i(-j)}} P_{\alpha,x}(a_{i(-j)})g_{ij}(a, X_i, \beta), \tag{3.23}$$

$$h_{ij}(A_i, X_i, Y_i; \delta, \gamma, \beta) = w_{ij}^{\alpha}(Y_{ij} - g_{ij}(A_i, X_i; \beta)) + \frac{1}{N_i} \sum_{a_i} P_{\alpha,x}(a_i) g_{ij}(a, X_i, \beta), \tag{3.24}$$

$$w_{ij}^{a\alpha} = \frac{1(A_{ij} = a) P_{\alpha,x}(A_i)}{\Pr(A_i | X_i; \delta, \gamma)} \Big/ \sum_{j=1}^{N_i} \frac{1(A_{ij} = a) P_{\alpha,x}(A_i)}{\Pr(A_i | X_i; \delta, \gamma)}, \tag{3.25}$$

and

$$w_{ij}^{\alpha} = \frac{P_{\alpha,x}(A_i)}{\Pr(A_i | X_i; \delta, \gamma)} \Big/ \sum_{j=1}^{N_i} \frac{P_{\alpha,x}(A_i)}{\Pr(A_i | X_i; \delta, \gamma)}. \tag{3.26}$$

The basic idea of the construction of $h_{ij}^{a}(A_i, X_i, Y_i; \delta, \gamma, \beta)$ lies in that equation (3.23) is a doubly robust estimator of $\mu_{a\alpha}$, in the sense that the expectation of $h^a(A, X, Y)$ converges to $\mu_{a\alpha}$ when data are fully observed, that is, $E\{\sum_{j=1}^{N_i} h_{ij}^{a}(A_i, X_i, Y_i; \delta, \gamma, \beta)\} = \mu_{a\alpha}$, if either the IPW models or regression models are correctly specified. When the propensity score models are correctly specified,

$$\begin{aligned} & E\{\sum_{j=1}^{N_i} h_{ij}^{a}(A_i, X_i, Y_i; \delta, \gamma, \beta)\} \\ = \; & E\Big[\sum_{j=1}^{N_i} \Big\{ w_{ij}^{a\alpha} Y_{ij} + \sum_{a_{i(-j)}} P_{\alpha,x}(A_{i(-j)}) g_{ij}(a, X_i, \beta) - w_{ij}^{a\alpha} g_{ij}(A_i, X_i; \beta) \Big\}\Big] \\ = \; & \mu_{a\alpha}^{ipw}. \end{aligned}$$

Similarly, when the baseline regression models have been correctly specified, it is obvious to see that the expectation of the first part of the equation (3.23) is equal to zero, and the expectation of the second part is the regression estimator, that is, $E\{\sum_{j=1}^{N_i} h_{ij}^{a}(A_i, X_i, Y_i; \delta, \gamma, \beta)\} = \mu_{a\alpha}^{reg}$. Therefore, to obtain the doubly robust estimator when confounders are subject to non-ignorable missingness, we can replace the observed outcomes and regression estimator in the conventional residuals weighted DR estimator by $h_{ij}^{a}(A_i, X_i, Y_i; \delta, \gamma, \beta)$ and $E\{h_{ij}(X_i, Y_i; \delta, \gamma, \beta) | A_i, R_{ij} = 0\}$, respectively. Hence, the constructed estimator can be viewed as a new residuals weighted DR estimator, where the weights are the inverse of the estimated propensity score of missingness, and the residuals come from the constructed $h_{ij}^{a}(A_i, X_i, Y_i; \delta, \gamma, \beta)$ instead of the regression estimator. By allowing either one of the sets of models to be correctly specified, the doubly robustness property can also be achieved accordingly. For estimating α, δ, and β, the estimators can be obtained by the maximum likelihood approach using observed data when $R = 1$. The estimation equation for α, δ, β can be written as $\sum_{i=1}^{K} \psi_{\alpha}(O_i; \alpha) = 0$, $\sum_{i=1}^{K} \psi_{\delta}(O_i; \delta) = 0$, and $\sum_{i=1}^{K} \psi_{\beta}(O_i; \beta) = 0$, respectively. Let $\psi_a^{dr}(O_i;\mu_{1\alpha},\xi,\beta) = \hat{\mu}^{dr}(1,\alpha) - \mu_{1\alpha}$, and $\psi_a^{dr}(O_i;\mu_{0\alpha},\xi,\beta) = \hat{\mu}^{dr}(0,\alpha) - \mu_{0\alpha}$ denote the estimation equations for $\mu_{1\alpha}$ and $\mu_{0\alpha}$, respectively. Then the estimated parameter $\hat{\boldsymbol{\theta}}^{\boldsymbol{dr}} = (\hat{\alpha},\hat{\delta},\hat{\beta},\hat{\xi},\hat{\mu}_{1\alpha}^{dr},\hat{\mu}_{0\alpha}^{dr})$ is a solution of the following estimation equations:

$$\sum_{i=1}^{K} \boldsymbol{\psi}^{dr}(O_i;\boldsymbol{\theta}) = \mathbf{0}. \tag{3.27}$$

The consistency and asymptotic normality of the proposed DR estimator are summarized in the following theorem.

Theorem 3.3. *Assume $OR(X_{ij}|A_i;\xi)$ is correctly specified, if either (a) $\Pr(R_{ij}=1|A_{ij}=1,X_i;\gamma')$ and $\Pr(A_{ij}=1|R_{ij}=1,X_{ij};\delta)$ or (ii) $f(Y_{ij}|A_i,X_{ij},R_{ij}=1;\beta)$ and $f(X_{ij}|A_i,R_{ij}=1;\beta)$ are correctly specified; then the DR estimator $\hat{\mu}_{a\alpha}^{dr}$ is consistent for μ_a, where $\hat{\mu}_{a\alpha}^{dr}$ is obtained from equation (3.21), and*

$$\sqrt{K}(\hat{\boldsymbol{\theta}}^{\boldsymbol{dr}} - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}(\mathbf{0},\Sigma^{dr}),$$

as K goes to infinity, where $\Sigma^{dr} = U(\boldsymbol{\theta})^{-1}V(\boldsymbol{\theta})\{U(\boldsymbol{\theta})^{-1}\}^T$, $U(\boldsymbol{\theta}) = E\{-\partial\boldsymbol{\psi}^{dr}(O_i;\boldsymbol{\theta})/\partial\boldsymbol{\theta}^T\}$, and $V(\boldsymbol{\theta}) = E\{\boldsymbol{\psi}^{dr}(O_i;\boldsymbol{\theta})\boldsymbol{\psi}^{dr}(O_i;\boldsymbol{\theta})^T\}$.

The proof of Theorem 3.3 is given in Section 3.7.

3.4 Simulation

In this section, we conduct a simulation study to illustrate the proposed IPW, regression, and DR estimator. First, under setting 1 (S1), we generate a population of size $N = 10000$, and randomly classify the whole population into $K = 100$ mutually exclusive groups with $N_i = 100$ individuals in each group. For each individual, the covariate X_1 is generated from the Bernoulli distribution with probability 0.5, and the covariate X_2 is generated from the standard normal distribution. To generate treatment and missing indicators, we assume two logistic models. We assume a mixed effects logistic model $\text{logit}\{\Pr(A_{ij}=1 \mid R_{ij}=1, X_{ij}=x_{ij}, b_i)\} = 0.1 + 0.2x_{1ij} - 0.1x_{2ij} + b_i$ for the propensity score of treatment, where b_i is the group-level random effect term generated from $\mathcal{N}(0,\sigma^2)$, i.e., the normal distribution with mean 0 and variance σ^2. We assume a logistic model $\text{logit}\{\Pr(R_{ij}=1 \mid A_{ij}=a_{ij}, X_{ij}=x_{ij})\} = 1 + a_{ij} + 1.2x_{1ij} - 0.5x_{2ij}$ for the propensity score model of missingness. Then the joint propensity score for treatment and missingness with random effects $\Pr(a_{ij}, r_{ij}=1|x_{ij},b_i)$ is calculated according to equation (3) in Chen [2007], and the missingness indicator R_{ij} and the treatment indicator A_{ij} are generated from the joint propensity score, which is given by the following equation:

$$\Pr\left(A_i, R_{ij}=1 \mid X_i\right)=\int \prod_{j'=1}^{n_i}\left\{h_{ij'}^{11}(b_i)\right\}^{A_{ij}}\left\{h_{ij'}^{01}\left(b_i\right)\right\}^{(1-A_{ij})} f_b\left(b_i;\sigma^2\right) db_i,$$

where $h_{ij}^{ar}(b_i) = \Pr(a_{ij} = a, r_{ij} = r|x_{ij}, b_i)$. Finally, we generate the outcome Y_{ij} from $Y_{ij} = 1 + 2x_{1ij} - 3x_{2ij} + 0.5x_{1ij}x_{2ij} + 2a_{ij} + 3p_i(a_i) + \epsilon_{ij}$, where $p_i(a_i)$ is the proportion of units in group i that receive the treatment, and $\{\epsilon_{ij}\}_{1\leq i\leq K,1\leq j\leq N_i}$ are i.i.d. random noise terms generated by $\mathcal{N}(0, \sigma^2)$. We consider four settings for the number of observations and the variance of the group effect:

(S1) $N = 10000$, $K = 50$, and $\sigma^2 = 0.25$;

(S2) $N = 10000$, $K = 50$, and $\sigma^2 = 0.16$;

(S3) $N = 12000$, $K = 200$, and $\sigma^2 = 0.25$;

(S4) $N = 12000$, $K = 200$, and $\sigma^2 = 0.16$.

Second, letting $\alpha = 0.5$, the propensity score models are correctly specified as $\Pr(a_{ij} = 1|r_{ij} = 1, x_{ij}; \delta) = \text{logit}^{-1}(\delta_1 + \delta_2 x_{1ij} + \delta_3 x_{2ij} + b_i)$ and $\Pr(r = 1|a, x; \gamma) = \text{logit}^{-1}(\gamma_1 + \gamma_2 a + \gamma_1 x_{1ij} + \gamma_2 x_{2ij})$. The parameter $\delta = (\delta_1, \delta_2, \delta_3)$ and $\gamma = (\gamma_1, \gamma_2, \gamma_3, \gamma_4)$ are estimated with *lme4* Bates et al. [2015] package in *R* R Core Team [2019]. The regression model is correctly specified as $E(y_{ij}|a_i, x_i, r_{ij} = 1) = \beta_0+\beta_1 a_{ij}+\beta_2 p_i(a_i)+\beta_3 x_{1ij}+\beta_4 x_{2ij}+\beta_5 x_{1ij}x_{2ij}$. The odds ratio function is correctly specified as $\eta(x_{ij}|a_{ij}; \zeta) = exp(\zeta_1 x_1 + \zeta_2 x_2)$. The parameters β are estimated by MLE. Then, the IPW estimator, regression estimator, and doubly robust estimators (DR-TT: when all models are correctly specified), i.e., $\hat{\mu}^{ipw}_{a\alpha}$, $\hat{\mu}^{reg}_{a\alpha}$, $\hat{\mu}^{dr}_{a\alpha}$, are calculated according to equations (3.6), (3.16), and (3.21), respectively.

Third, in the scenario when there exists model misspecification, DR estimator DR-TF is calculated when the outcome model is misspecified as $E(y_{ij}|a_i, x_i, r_{ij} = 1) = \beta_0 + \beta_1 a_{ij} + \beta_3 x_{1ij} + \beta_4 x_{2ij} + \beta_5 x_{1ij}x_{2ij}$, and the propensity score models are correctly specified.

The DR estimator DR-FT is calculated when the propensity score of treatment is incorrectly specified as $\Pr(a_{ij} = 1|r_{ij} = 1, x_{ij}; \delta) = \text{logit}^{-1}(\delta_1 + \delta_2 x_{1ij} + b_i^{(1)})$, and the outcome model is correctly specified.

Finally, the DR estimator DR-FF is calculated when the propensity score for treatment is incorrectly specified as $\Pr(a_{ij} = 1|r_{ij} = 1, x_{ij}; \delta) = \text{logit}^{-1}(\delta_1 + \delta_2 x_{1ij} + b_i^{(1)})$, and the outcome model is misspecified as $E(y_{ij}|a_i, x_i, r_{ij} = 1) = \beta_0 + \beta_1 a_{ij} + \beta_3 x_{1ij} + \beta_4 x_{2ij} + \beta_5 x_{1ij} x_{2ij}$.

The simulation study is repeated 1000 times, and the estimators of $\mu_{1\alpha}$, $\mu_{0\alpha}$, and $\bar{D}E(\alpha)$ are summarized in the following tables.

Table 3.1: Bias, empirical standard error (se), and standard deviation [in brackets] for IPW estimators, regression estimators, and doubly robust estimators under S1, S2, S3, and S4.

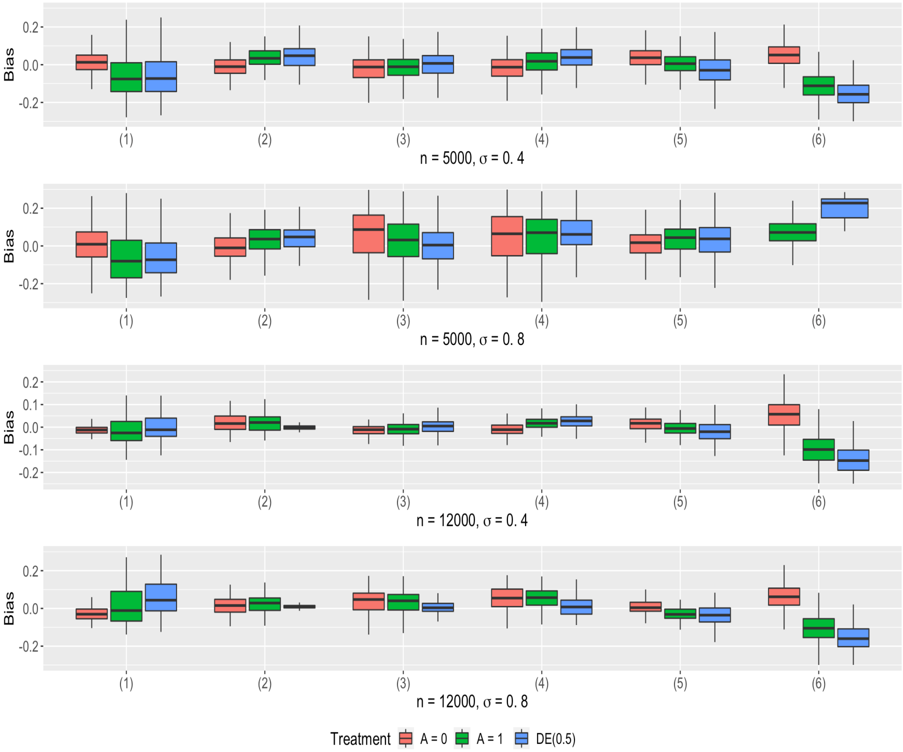
		IPW	REG	DR-TT	DR-TF	DR-FT	DR-FF
S1	$\mu_{1,0.5}$	0.005[0.051]	-0.037[0.013]	-0.062[0.073]	0.043[0.072]	0.081[0.046]	-0.102[0.191]
	$\mu_{0,0.5}$	0.003[0.059]	-0.003[0.047]	-0.023[0.082]	0.023[0.080]	0.030[0.067]	0.054[0.072]
	$\bar{\mathrm{D}}\mathrm{E}(0.5)$	0.002[0.082]	-0.033[0.130]	-0.040[0.072]	0.020 [0.068]	0.050[0.091]	-0.156[0.104]
	$se(\mu_{1,0.5})$	0.152	0.030	0.078	0.103	0.102	0.181
	$se(\mu_{0,0.5})$	0.110	0.032	0.102	0.110	0.085	0.138
	$se(\bar{\mathrm{D}}\mathrm{E}(0.5))$	0.114	0.069	0.080	0.075	0.113	0.166
S2	$\mu_{1,0.5}$	0.035[0.109]	0.034[0.059]	-0.018[0.042]	0.017[0.046]	-0.023[0.047]	0.060[0.121]
	$\mu_{0,0.5}$	0.018[0.122]	-0.008[0.057]	-0.005[0.042]	-0.007[0.042]	0.041[0.061]	-0.796[0.092]
	$\bar{\mathrm{D}}\mathrm{E}(0.5)$	0.016[0.118]	0.042[0.077]	0.020[0.052]	0.019 [0.058]	-0.019[0.080]	0.860[0.142]
	$se(\mu_{1,0.5})$	0.302	0.079	0.054	0.052	0.041	0.084
	$se(\mu_{0,0.5})$	0.111	0.073	0.057	0.057	0.069	0.094
	$se(\bar{\mathrm{D}}\mathrm{E}(0.5))$	0.234	0.069	0.067	0.064	0.088	0.171
S3	$\mu_{1,0.5}$	-0.005[0.180]	0.019[0.046]	-0.009[0.041]	0.037[0.027]	-0.004[0.031]	-0.105[0.081]
	$\mu_{0,0.5}$	-0.012[0.041]	0.021[0.042]	-0.014[0.029]	-0.020[0.028]	0.028[0.032]	0.049[0.081]
	$\bar{\mathrm{D}}\mathrm{E}(0.5)$	0.007[0.099]	-0.002[0.010]	0.006[0.035]	0.057 [0.032]	-0.03[0.041]	-0.155[0.072]
	$se(\mu_{1,0.5})$	0.154	0.036	0.041	0.037	0.029	0.090
	$se(\mu_{0,0.5})$	0.045	0.038	0.042	0.037	0.018	0.110
	$se(\bar{\mathrm{D}}\mathrm{E}(0.5))$	0.137	0.089	0.041	0.051	0.033	0.080
S4	$\mu_{1,0.5}$	0.110[0.162]	0.022[0.046]	-0.030[0.018]	0.010[0.022]	-0.011[0.032]	0.053[0.085]
	$\mu_{0,0.5}$	-0.018[0.035]	0.013[0.038]	-0.002[0.017]	-0.001[0.017]	0.028[0.034]	0.053[0.085]
	$\bar{\mathrm{D}}\mathrm{E}(0.5)$	0.128[0.185]	0.008[0.019]	-0.028[0.024]	0.011 [0.033]	-0.040[0.042]	0.160[0.074]
	$se(\mu_{1,0.5})$	0.225	0.048	0.037	0.027	0.037	0.103
	$se(\mu_{0,0.5})$	0.149	0.048	0.037	0.027	0.028	0.120
	$se(\bar{\mathrm{D}}\mathrm{E}(0.5))$	0.249	0.015	0.041	0.040	0.038	0.090

The bias and empirical coverages of the proposed estimators are shown in Table 3.1, Table 3.2, and Figure 3.2, where the 95% Wald-type confidence intervals are constructed

Table 3.2: Coverage Probability (%) of IPW estimators, regression estimators, and doubly robust estimators under S1, S2, S3, and S4.

		IPW	REG	DR-TT	DR-TF	DR-FT	DR-FF
S1	$\mu_{1,0.5}$	97.0	91.0	94.7	92.8	98.5	0
	$\mu_{0,0.5}$	96.0	91.0	91.2	94.0	95.5	0
	$\bar{\mathrm{DE}}(0.5)$	95.5	89.5	90.5	92.5	93.0	0
S2	$\mu_{1,0.5}$	98.5	89.5	98.5	93.5	90.5	0
	$\mu_{0,0.5}$	95.5	92.5	91.5	95.5	90.5	0
	$\bar{\mathrm{DE}}(0.5)$	98.5	92.0	93.0	92.0	88.5	0
S3	$\mu_{1,0.5}$	95.5	93.0	96.2	94.8	96.8	59.8
	$\mu_{0,0.5}$	96.0	93.0	95.0	93.6	94.8	69.8
	$\bar{\mathrm{DE}}(0.5)$	97.0	95.5	97.2	96.0	89.2	56.6
S4	$\mu_{1,0.5}$	93.0	90.0	97.0	92.0	92.5	50.6
	$\mu_{0,0.5}$	97.0	96.0	98.5	96.5	91.5	57.4
	$\bar{\mathrm{DE}}(0.5)$	92.0	92.5	96.5	92.5	85.5	56.0

Figure 3.2: Bias of the (1) IPW, (2) Regression, and DR estimators for $\mu_{1,0.5}$ and $\mu_{0,0.5}$ under four scenarios: (3) both propensity and outcome regression models are correctly specified; (4) propensity score models are correctly specified; (5) when only outcome model is correctly specified; (6) when neither the outcome regression model nor the IPW models are correctly specified (The boxplot of $\hat{\mu}^{dr}_{0,0.5}$ in the second scenario is dropped because the absolute value of the bias is greater than 0.2).



according to the asymptotic distribution of the proposed estimators in Theorems 3.1,3.2, and 3.3. When both the IPW models and the regression model are correctly specified, the IPW, regression, and DR estimators all perform well in terms of having small bias and variances, and the DR robust estimator DR-TT has the smallest variance among all the estimators.When the outcome regression model is correctly specified, the regression estimator has the smallest variance among all the proposed estimators. When both the model of the propensity score for treatment and the regression model are misspecified, the bias of $\hat{\mu}^{dr}_{0\alpha}$ has the same magnitude with the other estimators when either the IPW or the regression model, but not both, is correctly specified. As shown in Table 3.2, all three estimators achieve nominal levels when corresponding models are correctly specified, which indicates the standard error formulas proposed in Theorem 3.1, 3.2, and 3.3 are valid; when both propensity score and regression models are misspecified, the coverage probability of the DR estimator decreases to zero.

3.5 Application

Particulate matter 2.5 (PM2.5) refers to tiny particles or droplets in the air that can affect people's short-term or even long-term health conditions such as respiratory issues, increased mortality from lung cancer, and heart disease. A primary strategy to achieve the reduction of ambient PM2.5 is the installation of flue-gas desulfurization equipment or controls ("scrubbers") to reduce the sulfur dioxide (SO_2), nitrogen dioxide (NO_x), and carbon dioxide (CO_2) emissions, which are three main air pollutants that mediate the changes of PM2.5. In 1990, the U.S. Clean Air Act (CAA) amendments launched the Acid Rain Program (ARP) to reduce the emissions of the ambient SO_2, NO_x, and CO_2 by regulating those power plants to install scrubbers on coal-fired electricity-generating units (EGUs).

In this section, we apply the proposed estimators on a dataset from the U.S. EPA's Air Quality System (AQS) to estimate the causal effect of installing the scrubbers on the reduction of ambient NO_x emissions. The dataset contains monthly emissions data from 1218 EGUs in the U.S. in 2004. The 1218 EGUs are classified into 40 clusters through the linkage algorithm by Zigler et al. [2016]. Among 1218 EGUs, 913 EGUs installed the scrubbers and 205 EGUs did not install the scrubbers. Five important characteristics of the EGUs and the atmosphere are included in the data analysis: the heat input rate, the capacity of the EGU, the amount of coal, the operation time, and the average temperature in the previous year. As these characteristics affect both the emissions of NO_x and the installation of scrubbers, they are treated as confounders of the causal relationship between the outcome and the treatment. For the treatment (scrubbers' installation) coverage probability, the average proportion of treated units among all the groups is 66.79%. For the missingness in confounders, there is 15.76% missingness in the amount of coal, 2.25% missingness in operation time, 21.69% missingness in heat input rate, 0.99% in capacity, and 24.90% missingness in total. The missingness can be caused by multiple reasons such as monitor errors (e.g., the operation time exceeds a certain time), the failure of recording, and the changes in the filter. Since the units with fewer NO_x emissions are less likely to disclose the baseline characteristics, and records of the baseline characteristics were measured long before the emissions of NO_x took place, it is plausible that the outcome-independence assumption holds.

We assume a linear fixed effect regression model for the outcome and assume different mixed effects, and logistic regression models, for the propensity score of treatment and missingness, respectively. To account for the dependency of the treatment allocation strategy on the baseline characteristics, following Papadogeorgou et al. [2019], we assume $P_{\alpha,x}(a_i)$ as the logistic regression model: $\text{logit}\{P_{\alpha,x}(a_i)\} = \xi_i^\alpha + \sum_{j=1}^6 \beta_i X_{ij}$, where ξ is estimated by solving the following equation,

$$\frac{1}{N_i}\sum_{j=1}^{N_i} \text{expit}\left(\xi_i^\alpha + \beta_j X_{ij}\right) = \alpha.$$

Since in the dataset, there are over 80% of units lying in the clusters with the average proportion of units with scrubbers installed ranging from 0.3 to 0.8, we consider values of α varying from 0.3 to 0.8.

Figure 3.3 presents the results of IPW, regression, and DR estimators for the direct effect DE(α) across different values of α. When $\alpha = 0.3$, $\bar{DE}(\alpha)$ of the IPW, regression and DR estimators are -29.45, -49.25, and -34.72, respectively. When $\alpha = 0.8$, DE(α) of the IPW, regression and DR estimators are -19.55, -21.99, and -17.44, respectively. The estimated $\bar{DE}(\alpha)$ are negative for all three estimators across different α values, indicating that the intervention of installing scrubbers has a positive effect on reducing the emissions of the tons of NO$_x$, and there would be approximately 30 tons of NO$_x$ emissions fewer per unit among the units with scrubbers installed compared to the units without scrubbers installed. As α increases, $\bar{DE}(\alpha)$ has an increasing trend, which implies that the intervention at one EGU is beneficial for the reduction of the emissions of NO$_x$, but the effect is smaller when the proportion of the units within the same cluster that have scrubbers installed becomes larger.

Let α be the average treatment coverage probability among 40 clusters, i.e., $\alpha_0 = 0.67$; in this setting, IPW, regression, and DR estimates for the indirect ($\bar{IE}(0.6, \alpha_1)$), total effect ($\bar{TE}(0.6, \alpha_1)$), and overall effect ($\bar{OE}(\alpha_1)$) are given in Figure 3.4. The indirect effect has a decreasing trend when $\alpha_1 - \alpha_0$ increases, and it is positive when $\alpha_1 < 0.67$ and negative when $\alpha_1 > 0.67$. For example, when $\alpha_1 = 0.46$, the DR estimate is 15.14, and 95 % CI is (11.19, 19.76), which suggests that there would be 15.14 more tons of NO$_x$ emissions if scrubbers had not been installed for units within groups with 46% coverage compared to that with groups with 67% coverage of scrubber installation; when $\alpha_1 = 0.75$, the estimate is -6.84 for DR estimator, and the corresponding 95 % CI is (-8.82,-5.40), which implies that we would expect 3.90 tons of NO$_x$ emissions fewer if scrubbers had not been installed for units within groups with 75% coverage compared to groups with average coverage probability. It is also worth noting that the confidence intervals would decrease as the difference between α_0 and α_1 decreases, which indicates that the decrease in the variance of the estimator of $\mu_{0\alpha}$ and $\mu_{1\alpha}$ cannot offset the increase in the correlation between the two estimators, because both the estimators are dependent on the treatment allocation function $P_{\alpha,x}(a_{(i(-j)})$. The total effect of the proposed estimators, which combine both the direct and the indirect effects, have a slightly decreasing trend when α increases. For example, when $\alpha = 0.5$, the IPW, regression, and DR estimates are -18.15, -25.70, and -16.91, and the corresponding 95% CIs are (-21.38, -14.43), (-30.27, -21.64), and (-19.78, -14.05), respectively; when $\alpha = 0.75$, the IPW, regression, and DR estimates are -25.56, -28.35, and -22.29, and the corresponding 95% CIs are (-29.44,-21.71), (-31.56, -24.85), and (-24.75,-19.83), respectively. Therefore, all three estimators indicate that there would be fewer NO$_x$ emissions if the scrubber had been installed in a unit within groups with higher coverage probability compared to the unit without scrubbers installed in groups with lower coverage probability. The estimates of the overall effect of scrubber installation, which quantify the difference in the tons of NO$_x$ emissions under two treatment allocation strategies, suggest that there would be fewer NO$_x$ emissions within groups with a higher average percentage of the coverage of scrubber installation.

Figure 3.3: Estimates and 95% Wald-type confidence intervals of $\bar{DE}(\alpha)$ of scrubber installation for (1) IPW, (2) Regression, and (3) DR estimators with $\alpha \in (0.3, 0.8)$, where the shadow area represents the pointwise confidence intervals.

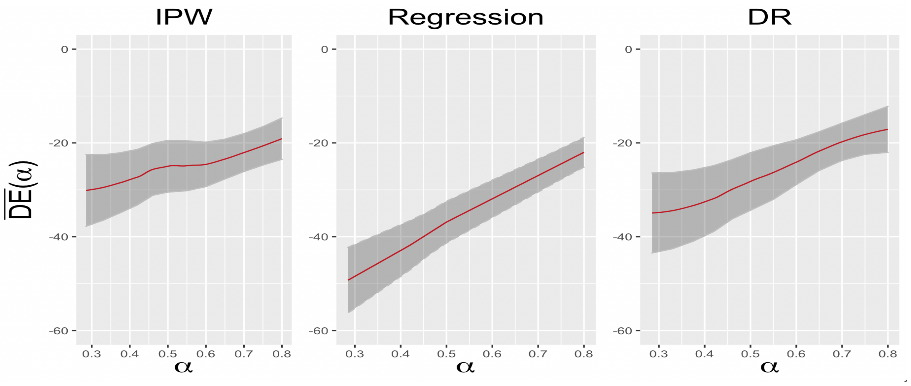
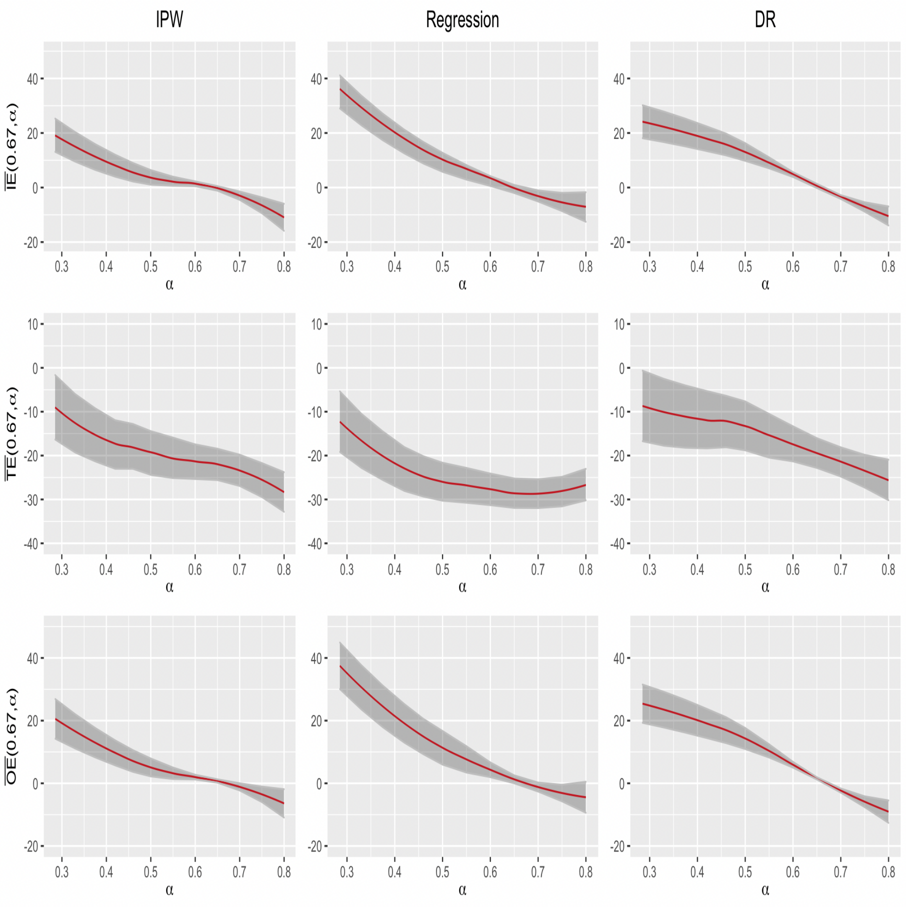


Figure 3.4: Estimates and 95% Wald-type confidence intervals of $I\bar{E}(0.67, \alpha)$, $T\bar{E}(0.67, \alpha)$, and $O\bar{E}(0.67, \alpha)$ of scrubbers' installation for (1) IPW, (2) Regression, and (3) DR estimators with $\alpha \in (0.3, 0.8)$, where the shadow area represents the pointwise confidence intervals.



3.6 Discussion

In this paper, we constructed three consistent estimators: IPW, regression, and DR estimators for four types of network causal effects: the direct, indirect, total and overall effects when the confounders are missing not at random. Under the group-level outcome-independent missingness assumption, the IPW and regression estimators are consistent and asymptotically normal if the corresponding models are correctly specified, and the consistency of the DR estimator requires that either the joint modeling of the propensity score of treatment and missingness or the outcome regression model, but not necessarily both, are correctly specified.

The proposed doubly robust estimator is based on the conventional DR estimators in Kang and Schafer [2007] without interference. In the setting where interference exists, the methodology avoids the assumption of SUTVA, and can recover the causal effect in the whole population. To solve the problem of the extreme group-level joint propensity scores, we propose self-normalized estimators to reduce the variance. In the real application, we classify the power-generating facilities into different groups based on their geographical locations and apply three proposed estimators on the incomplete clustered data. We further show that the intervention of installing scrubbers has a positive effect on reducing the emissions of NO_x which, in turn, may potentially reduce the ambient PM2.5 and the concentrations of ozone. The effect of scrubbers' installation on one power-generating facility decreases as the number of treated facilities increases within the same group.

One limitation of the proposed methods is that they can cause increased computational burden when the number of units in each group increases, and it may not be suitable to implement the estimators when the confounders are high-dimensional. Adapting the proposed methods to handle high-dimensional data is an interesting direction for future research. Another limitation lies in that the estimators do not work well for the cases when the proportion of missingness is large. In this study, we only consider the setting under partial interference only; how to draw causal effects in the network of general interference remains uncovered, and we leave it as another future research. Besides, it is interesting to consider multiple missingness patterns so that the information of confounders can be fully utilized.

3.7 Proof of Theorems

3.7.1 Proof of Theorem 3.1

Proof: In this section, we show the consistency of $\hat{\mu}_{a\alpha}^{ipw}$. First, we notice that:

$$
\begin{aligned}
& \frac{1}{K}\sum_{i=1}^{K}\frac{1}{N_i}\sum_{j=1}^{N_i}\frac{1(A_{ij}=a)1(R_{ij}=1)Y_{ij}P_{\alpha,x}(A_{i(-j)})}{\hat{\Pr}(A_i,R_{ij}|X_i)} \\
\xrightarrow{p} \quad & E\Bigg[\frac{1}{N_i}\sum_{j=1}^{N_i}\frac{1(A_{ij}=a)1(R_{ij}=1)Y_{ij}P_{\alpha,x}(A_{i(-j)})}{\Pr(A_i,R_{ij}|X_i)}\Bigg] \\
= \quad & E\Bigg[\frac{1}{N_i}\sum_{j=1}^{N_i}E\Bigg[E\Bigg\{\frac{1(R_{ij}=1)}{\Pr(R_{ij}|A_i,X_i,Y_{ij})}|A,X,Y\Bigg\}\frac{1(A_{ij}=a)Y_{ij}P_{\alpha,x}(A_{i(-j)})}{\Pr(A_i|X_i)}\Bigg|X\Bigg]\Bigg] \\
= \quad & E\Bigg[\frac{1}{N_i}\sum_{j=1}^{N_i}E\Bigg\{\frac{1(A_{ij}=a)Y_{ij}P_{\alpha,x}(A_{i(-j)})}{\Pr(A_i|X_i)}\Bigg|X\Bigg\}\Bigg] \\
= \quad & E\Bigg[\frac{1}{N_i}\sum_{j=1}^{N_i}\sum_{a_i}\frac{1(a_{ij}=a)Y_{ij}(a_i)\pi(a_{i(-j)};\alpha)}{\Pr(A_i=a_i|X_i)}\Pr(A_i=a_i|X_i)\Bigg] \\
= \quad & E\Bigg[\frac{1}{N_i}\sum_{j=1}^{N_i}\sum_{a_{i(-j)}}Y_{ij}(a_i)\pi(a_{i(-j)};\alpha)\Bigg] \\
= \quad & E\Big[\bar{Y}_i(a,\alpha)\Big]=\mu_{a\alpha},
\end{aligned}
$$

and

$$
E\sum_{j=1}^{N_i}\frac{1(A_{ij}=a)P_{\alpha,x}(A_{i(-j)})}{\hat{\Pr}(A_i,R_{ij}|X_i)}=N_i.
$$

Therefore, we have

$$
E\Bigg[\sum_{j=1}^{N_i}\Bigg\{\frac{1(A_{ij}=a)1(R_{ij}=1)Y_{ij}P_{\alpha,x}(A_{i(-j)})}{\hat{\Pr}(A_i,R_{ij}|X_i)}-\frac{1(A_{ij}=a)P_{\alpha,x}(A_{i(-j)})}{\hat{\Pr}(A_i,R_{ij}|X_i)}\mu_{a\alpha}\Bigg\}\Bigg]=0.
$$

3.7.2 Proof of Theorem 3.2

Proof: The consistency of $\hat{\mu}^{reg}_{a\alpha}$ follows below.

$$
\begin{aligned}
\hat{\mu}^{reg}_{a\alpha} &= \frac{1}{K}\sum_{i=1}^{K} \hat{Y}_i^{reg}(a,\alpha) \\
&= \frac{1}{K}\sum_{i=1}^{K}\Big[\frac{1}{N_i}\sum_{j=1}^{N_i}\sum_{a_{i(-j)}} 1(R_{ij}=1)\hat{g}(a_i,X_i;\hat{\beta})P_{\alpha,x}(a_{i(-j)}) \\
&+ \frac{1}{N_i}\sum_{j=1}^{N_i}\sum_{a_{i(-j)}} 1(R_{ij}=0)\pi(a_{i(-j)};\alpha)E\{\hat{g}(a_i,X_i;\hat{\beta})|A_i=a_i,R_{ij}=0;\hat{\beta},\hat{\zeta}\}\Big] \\
&\xrightarrow{p} E\Big[\frac{1}{N_i}\sum_{j=1}^{N_i}\sum_{a_{i(-j)}} 1(R_{ij}=1)\hat{g}(a_i,X_i;\hat{\beta})P_{\alpha,x}(a_{i(-j)}) \\
&+ \frac{1}{N_i}\sum_{j=1}^{N_i}\sum_{a_{i(-j)}} 1(R_{ij}=0)P_{\alpha,x}(A_{i(-j)})E\{\hat{g}(a_i,X_i;\hat{\beta})|A_i=a_i,R_{ij}=0;\hat{\beta},\hat{\zeta}\}\Big] \\
&= E\Big[E\Big[\frac{1}{N_i}\sum_{j=1}^{N_i}\sum_{a_{i(-j)}} 1(R_{ij}=1)E\{\hat{g}(a_i,X_i;\hat{\beta})|A_i,R_{ij}=1;\hat{\beta},\hat{\zeta}\}P_{\alpha,x}(a_{i(-j)}) \\
&+ \frac{1}{N_i}\sum_{j=1}^{N_i}\sum_{a_{i(-j)}} 1(R_{ij}=0)\pi(a_{i(-j)};\alpha)E\{\hat{g}(a_i,X_i;\hat{\beta})|A_i,R_{ij}=0;\hat{\beta},\hat{\zeta}\}\Big|A_i\Big]\Big] \\
&= E\Big[E\Big[\frac{1}{N_i}\sum_{j=1}^{N_i}\sum_{a_{i(-j)}} R_{ij}E\{\hat{g}(a_i,X_i;\hat{\beta})|A_i,R_{ij};\hat{\beta},\hat{\zeta}\}P_{\alpha,x}(a_{i(-j)}) \\
&+ \frac{1}{N_i}\sum_{j=1}^{N_i}\sum_{a_{i(-j)}} (1-R_{ij})P_{\alpha,x}(a_{i(-j)})E\{\hat{g}(a_i,X_i;\hat{\beta})|A_i,R_{ij};\hat{\beta},\hat{\zeta}\}\Big|A_i\Big]\Big] \\
&= E\Big[E\Big[\frac{1}{N_i}\sum_{j=1}^{N_i}\sum_{a_{i(-j)}} E\{\hat{g}(a_i,X_i;\hat{\beta})|A_i,R_{ij};\hat{\beta},\hat{\zeta}\}P_{\alpha,x}(A_{i(-j)})\Big|A_i\Big]\Big] \\
&= E\Big[\frac{1}{N_i}\sum_{j=1}^{N_i}\sum_{a_{i(-j)}} E\{Y_{ij}(a_{ij},a_{i(-j)})|X_i;\hat{\beta},\hat{\zeta}\}P_{\alpha,x}(a_{i(-j)})\Big] = \mu_{a\alpha}.
\end{aligned}
$$

3.7.3 Proof of Theorem 3.3

Proof: If (a) $\Pr(R_{ij}=1|A_i=1,X_i;\delta)$ and $\Pr(A_i|R_i=1,X_i;\gamma)$ are correctly specified, the consistency of $\hat{\mu}^{dr}_{a\alpha}$ follows by the law of large numbers and the outcome independent missingness assumption,

$$
\begin{aligned}
&\hat{\mu}^{dr}_{a\alpha}\\
&=\frac{1}{K}\sum_{i=1}^{K}\frac{1}{N_i}\Bigg[\sum_{j=1}^{N_i}\Bigg[E\{\hat{h}^a_{ij}(A_i,X_i,Y_i)|A_i=a_i,R_{ij}=0;\hat{\delta},\hat{\gamma},\hat{\beta}\}\\
&+\frac{1(R_{ij}=1)}{\hat{\Pr}(R_{ij}|A_i=a_i,X_i;\hat{\gamma})}\Bigg\{\hat{h}^a_{ij}(A_i,X_i,Y_i)-E\{\hat{h}^a_{ij}(A_i,X_i,Y_i)|A_i=a_i,R_{ij}=0;\hat{\delta},\hat{\gamma},\hat{\beta}\}\Bigg\}\Bigg]\Bigg]\\
&\xrightarrow{p} E\Bigg[\frac{1}{N_i}\sum_{j=1}^{N_i}\Bigg[E\{\hat{h}^a_{ij}(A_i,X_i,Y_i)|A_i=a_i,R_{ij}=0;\hat{\delta},\hat{\gamma},\hat{\beta}\}\\
&+\frac{1(R_{ij}=1)}{\Pr(R_{ij}|A_i=a_i,X_i;\gamma)}\Bigg\{\hat{h}^a_{ij}(A_i,X_i,Y_i)-E\{\hat{h}^a_{ij}(A_i,X_i,Y_i)|A_i=a_i,R_{ij}=0;\hat{\delta},\hat{\gamma},\hat{\beta}\}\Bigg\}\Bigg]\Bigg]\\
&=E\Bigg[\frac{1}{N_i}\sum_{j=1}^{N_i}\frac{1(A_{ij}=a)R_{ij}Y_{ij}P_{\alpha,x}(A_{i(-j)})}{\Pr(R_{ij}|A_i,X_i;\hat{\gamma})\Pr(A_i|X_i;\hat{\delta},\hat{\gamma})}\\
&+\sum_{j=1}^{N_i}\frac{1(R_{ij}=1)}{\Pr(R_{ij}|A_i=a_i,X_i;\hat{\gamma})}\Bigg\{1-\frac{1(A_{ij}=a)P_{\alpha,x}(A_{i(-j)})}{\Pr(A_i|X_i;\hat{\gamma},\hat{\delta})}\Bigg\}g_{ij}(a,X_i;\hat{\beta})\\
&+\sum_{j=1}^{N_i}\Bigg\{1-\frac{1(R_{ij}=1)}{\Pr(R_{ij}|A_i=a_i,X_i;\hat{\gamma})}\Bigg\}E\{\hat{h}^a_{ij}(A_i,X_i,Y_i)|A_i=a_i,R_{ij}=0;\hat{\delta},\hat{\gamma},\hat{\beta}\}\Bigg]\\
&=E\Bigg[\frac{1}{N_i}\sum_{j=1}^{N_i}\sum_{a_i}\frac{1(a_{ij}=a)Y_{ij}(a_i)P_{\alpha,x}(A_{i(-j)})}{\Pr(A_i=a_i|X_i)}\Bigg]=\mu_{a\alpha}.
\end{aligned}
$$

If (b) the regression model $f(x, y|a, r = 1)$ is correctly specified, for any user-defined differentiable function $h(A, X, Y)$, we have

$$\begin{aligned}
E\{h(a_i, X_i, Y_i) \mid a_i, r_i = 0\} =& \frac{\iint \eta\left(r_i = 0, r_{i0}, x_i, x_{i0} \mid a_i\right) f(x_i, y_i \mid a_i, r_i = 1) h(a_i, x_i, y_i) dxdy}{E\left\{\eta\left(r_i = 0, r_{i0}, x_i, x_{i0} \mid a_i\right) \mid a_i, r_i = 1\right\}} \\
=& \frac{E\left\{\eta\left(r_i = 0, r_{i0}, X_i, x_0 \mid a\right) h(a_i, X_i, Y_i) \mid a_i, r_i = 1\right\}}{E\left\{\eta\left(r_i = 0, r_{i0}, X_i, x_{i0} \mid a_i\right) \mid a_i, r_i = 1\right\}} \\
=& \frac{E\left\{R_i \eta\left(r_i = 0, r_0, X_i, x_0 \mid a_i\right) h(a_i, X_i, Y_i) \mid a_i\right\}}{E\left\{R_i \eta\left(r_i = 0, r_{i0}, X_i, x_{i0} \mid a_i\right) \mid a_i\right\}}.
\end{aligned}$$

Therefore, it is straightforward to see

$$\begin{aligned}
& E\left\{R_i \eta\left(r_i = 0, r_{i0}, X_i, x_{i0} \mid A_i\right) h(A_i, X_i, Y_i) \mid A_i\right\} \\
=& E\left\{R_i \eta\left(r_i = 0, r_{i0}, X_i, x_{i0} \mid A_i\right) \mid A_i\right\} E\{h(A_i, X_i, Y_i) \mid A_i, R_i = 0\},
\end{aligned}$$

and

$$E\{R_i \eta\left(r_i = 0, r_{i0}, X_i, x_{i0} \mid A_i\right) \{h(A_i, X_i, Y_i) - E\{l(A_i, X_i, Y_i) \mid A_i, R_i = 0\}\} = 0.$$

Hence, we can replace $h(A, X, Y)$ by $\hat{h}(A, X, Y)$ in the above equation, and the consistency of $\hat{\mu}_{a\alpha}$ follows below.

$$
\begin{aligned}
&\hat{\mu}_{a\alpha}^{dr} \\
&= \frac{1}{K}\sum_{i=1}^{K}\frac{1}{N_i}\Bigg[\sum_{j=1}^{N_i}\Bigg[E\{\hat{h}_{ij}^{a}(A_i, X_i, Y_i)|A_i = a_i, R_{ij} = 0; \hat{\delta}, \hat{\gamma}, \hat{\beta}\} \\
&+ \frac{1(R_{ij} = 1)}{\Pr(R_{ij}|A_i = a_i, X_i; \hat{\gamma})}\Bigg\{\hat{h}_{ij}^{a}(A_i, X_i, Y_i) - E\{\hat{h}_{ij}^{a}(A_i, X_i, Y_i)|A_i = a_i, R_{ij} = 0; \hat{\delta}, \hat{\gamma}, \hat{\beta}\}\Bigg\}\Bigg]\Bigg] \\
&\xrightarrow{p} E\Bigg[\frac{1}{N_i}\sum_{j=1}^{N_i}\Bigg[E\{\hat{h}_{ij}(a, X_{ij}, Y_i)|A_i = a_i, R_{ij} = 0; \hat{\delta}, \hat{\gamma}, \hat{\beta}\} \\
&+ \frac{1(R_{ij} = 1)}{\Pr(R_{ij}|A_i = a_i, X_i; \hat{\gamma})}\Bigg\{\hat{h}_{ij}^{a}(A_i, X_i, Y_i) - E\{\hat{h}_{ij}^{a}(A_i, X_{ij}, Y_{ij})|A_i = a_i, R_{ij} = 0; \hat{\delta}, \hat{\gamma}, \hat{\beta}\}\Bigg\}\Bigg]\Bigg] \\
&= E\Bigg[\frac{1}{N_i}\sum_{j=1}^{N_i}\Bigg[1(R_{ij} = 1)\frac{\Pr(R_{ij} = 0|A_i = a_i, X_i; \hat{\gamma})}{\Pr(R_{ij} = 1|A_i = a_i, X_i; \hat{\gamma})} \\
&\quad \Bigg\{\hat{h}_{ij}^{a}(A_i, X_i, Y_i) - E\{\hat{h}_{ij}^{a}(A_i, X_i, Y_i)|A_i = a_i, R_{ij} = 0; \hat{\delta}, \hat{\gamma}, \hat{\beta}\}\Bigg\} \\
&+ 1(R_{ij} = 1)\Bigg\{\hat{h}_{ij}^{a}(A_i, X_i, Y_i) - E\{\hat{h}_{ij}^{a}(A_i, X_i, Y_i)|A_i = a_i, R_{ij} = 0; \hat{\delta}, \hat{\gamma}, \hat{\beta}\}\Bigg\} \\
&+ E\{\hat{h}_{ij}^{a}(A_i, X_i, Y_i)|A_i = a_i, R_{ij} = 0; \hat{\delta}, \hat{\gamma}, \hat{\beta}\}\Bigg]\Bigg] \\
&\hspace{20em} \textit{(by outcome independent missingness assumption)} \\
&= E\Bigg[\frac{1}{N_i}\sum_{j=1}^{N_i}\sum_{a_{i(-j)}}\Bigg[R_{ij}\hat{h}_{ij}^{a}(A_i, X_i, Y_i) + (1 - R_{ij})E\{\hat{h}_{ij}^{a}(A_i, X_i, Y_i)|A_i, R_{ij} = 0; \hat{\beta}, \hat{\xi}\}\Bigg]P_{\alpha,x}(a_{i(-j)})\Bigg] \\
&= E\Bigg[\frac{1}{N_i}\sum_{j=1}^{N_i}\sum_{a_{i(-j)}} E\{\hat{g}_{ij}(a_{i(-j)}, a_{ij}, X_i; \hat{\beta})|A_i, R_{ij}\}P_{\alpha,x}(a_{i(-j)})\Bigg] \\
&= \mu_{a\alpha}.
\end{aligned}
$$

Chapter 4

Mutiply Robust Estimation of Network Causal Effects

4.1 Introduction

Robust estimation in causal inference has drawn a great amount of interest in the past ten years. Many doubly robust estimators have been proposed under different settings. An estimator is doubly robust if it is consistent when either a propensity score model, or an outcome regression model, but not necessarily both, are correctly specified. Bang and Robins [2005] proposed doubly robust estimators when the outcome is possibly missing at random. Miao and Tchetgen Tchetgen [2016] utilized a shadow variable to construct the doubly robust estimators when the outcome is subject to nonignorable missingness. Shardell et al. [2015] proposed a doubly robust augmented IPW (AIPW) estimator for the effect of a time-varying exposure on the outcomes in a longitudinal study with dropout and truncation by death. Tan [2020], Ning et al. [2020], and Tang et al. [2022] proposed a doubly robust estimation procedure for drawing causal effects in the high-dimensional setting.

There has been an increasing interest in the field of multiply robust estimators in recent years, which are extensions of doubly robust estimators. Multiply robust estimators are constructed to provide more protection against model misspecification. An estimator is multiple robust if it is consistent when any one of the candidate models, either a propensity score model or an outcome regression model, is correctly specified. Han and Wang [2013] proposed a multiply robust estimator for the population mean of the response, given that the outcome is missing at random. The estimator is consistent if at least one of the candidate models, either for a propensity score or outcome regression, is correctly specified. Han [2014a,b] further improved their multiply robust estimators by proposing a new computational method, which overcomes the problems of multiple roots and non-convergence. Li et al. [2020a] showed that the multiply robust estimators are special cases of doubly robust estimators, and proposed a model mixing procedure for combining multiple candidate models. Zhang et al. [2019] proposed an empirical likelihood based approach to both testing and estimation of the treatment effect in non-randomized pretest-posttest studies. Following Han and Wang [2013], we consider the estimation of network treatment effects when confounders are missing not at random.

In this chapter, we extend the multiply robust estimation with incomplete data to the partial interference setting, where data can be grouped into disjoint clusters and observations within the same cluster are correlated. We consider both the scenario when data are fully observed and the scenario when confounders are subject to nonignorable missingness. The proposed methods are based on the empirical likelihood approach. In the first scenario when data are complete, the conventional multiply robust estimation procedure is adjusted by replacing the unit-level propensity score and outcome regression models with the group-level propensity score models and outcome regression models. In the second scenario when confounders are subject to nonignorable missingness, we developed a novel estimation procedure to overcome the difficulty of estimating the expectation of both outcome regression models and the propensity scores.

The chapter is organized as follows. In Section 4.2, we introduce the notations and assumptions. In Section 4.3, we introduce the developed multiply robust estimators based on clustered data with and without missingness. The theoretical properties of the proposed estimators are established, and the cluster-based bootstrapping method is utilized to estimate the variance of the proposed estimators. In Section 4.4, we conduct a series of simulation studies. In Section 4.5, we further illustrate the proposed estimators with a real dataset on a network emissions application, and the summary of the paper is given in Section 4.6.

4.2 Notation and Assumptions

Assume there are K disjoint groups of units. For $i = 1, 2, ...K$, let n_i denote the number of units in group i. Let $Y_i = (Y_{i1}, Y_{i2}, ..., Y_{in_i})$ denote the observed outcome in group i, $A_i = (A_{i1}, A_{i2}, ..., A_{in_i})$ the binary treatment indicator for group i, $\boldsymbol{X_{ij}} = (X_{ij,1}, X_{ij,2}, ..., X_{ij,p})^T$ denote the vector of confounders of subject j in group i that may be subject to missingness, and R_{ij} the indicator of observing $\boldsymbol{X_{ij}}$, where $R_{ij} = 1$ if any component of X_{ij} is missingness and $R_{ij} = 0$ if otherwise. Let $Y_{ij}(a_i)$ and $Y_{ij}(a_{ij}, a_{i(-j)})$ denote the potential outcome for unit j in group i under treatment allocation a_i. The average potential outcome is defined as $\mu_{a\alpha} = E(\bar{Y}_i(a, \alpha))$, where $\bar{Y}_i(a, \alpha) = n_i^{-1} \sum_{j=1}^{n_i} \sum_{a_{i(-j)} \in \mathcal{A}(N_i - 1)} Y_{ij}(a, a_{i(-j)}) P_{\alpha,x}(a_{i(-j)})$, and $P_{\alpha,x}(a_{i(-j)}) = \Pr(A_{i(-j)} = a_{i(-j)} | A_{ij} = a_{ij}, X_i)$ is the treatment allocation probability, which may depend on the confounders in an observational study. The direct effect (or the population average treatment effect) is defined as $\overline{\mathrm{DE}}(\alpha) = E(\bar{Y}_i(1, \alpha) - \bar{Y}_i(0, \alpha)) = \mu_{1\alpha} - \mu_{0\alpha}$, the indirect effect is defined as $\overline{\mathrm{IE}}(\alpha_0, \alpha_1) = E(\bar{Y}_i(0, \alpha_1) - \bar{Y}_i(0, \alpha_0)) = \mu_{0\alpha_1} - \mu_{0\alpha_0}$, the total effect is defined as $\overline{\mathrm{TE}}(\alpha_1, \alpha_1) = E(\bar{Y}_i(1, \alpha_1) - \bar{Y}_i(0, \alpha_0)) = \mu_{1\alpha_1} - \mu_{0\alpha_0}$, and the overall effect is defined as $\overline{\mathrm{OE}}(\alpha_0, \alpha_1) = E(\bar{Y}_i(\alpha_1) - \bar{Y}_i(\alpha_0)) = \mu_{\alpha_1} - \mu_{\alpha_0}$. More discussion on these causal effect estimands can be found in Section 3.2.

Assume the interference type is partial and direct interference, that is, there is no interference between units in different groups, and the outcome of one unit is affected only through the treatment assignment of another unit. This assumption is plausible if the groups are geographically separated, and the treatment assignment of one unit is dependent only on its own characteristics. We also make the causal consistency assumption, positivity assumption, and the group-level outcome-independent missingness assumption throughout the chapter. Formal definitions of these assumptions are given in Section 3.2.

4.3 Estimation

4.3.1 Multiply Robust with Interference

In this section, we assume that the interference exists and the data are fully observed. To construct the multiply robust estimator, we postulate multiple candidate propensity score models $\mathcal{P}^a = \{\pi_A^j(\boldsymbol{x_i}; \alpha^j),\ j = 1, 2, ...J\}$ for $\Pr(A_{ij} = a \mid \boldsymbol{X_i} = \boldsymbol{x_i})$, and multiple candidate models $\mathcal{A} = \{f_{ij}^l(y_{ij}|a_i, \boldsymbol{x_i}; \beta^l),\ l = 1, 2, ...L\}$ for the conditional distribution of the outcome. Let $\{m_{ij}^l(a_i, \boldsymbol{x_i}; \beta^l),\ l = 1, 2, ...L\}$ denote the set of corresponding outcome regression models for the conditional expectation of the outcome on the treatment and confounders, i.e., $E(Y_{ij}|A_i = a_i, \boldsymbol{X_i} = \boldsymbol{x_i})$. Let b_i and ξ_i denote the normally and independently distributed random effect terms in the propensity score models and outcome models respectively, that induce the correlation among units within the same cluster. Usually, each α^j is estimated by maximizing the following binomial log-likelihood,

$$\sum_{i=1}^{K} \log\{f^j(A_i|\boldsymbol{X_i}; \alpha, \psi_b)\},$$

where

$$f^j(A_i|\boldsymbol{X_i};\alpha,\psi_b) = \int \prod_{j=1}^{n_i} (\pi_A^j(\boldsymbol{X_i};\alpha^j))^{A_{ij}}(1-\pi_A^j(\boldsymbol{X_i};\alpha^j))^{1-A_{ij}} f_b(b_i;\psi_b)db_i, \tag{4.1}$$

and $f_b(b_i;\psi_b)$ is the normal distribution with mean zero and variance ψ_b. Similarly, we utilize the maximum likelihood approach to obtain the estimates of β^l in each of the candidate outcome regression models. In the literature, the most popular class of outcome models for correlated data is the linear mixed effect model. Thus, we assume a linear mixed effects model for the outcome model which has the following representation:

$$Y_{ij} = \beta_0^l + \beta_1^l A_{ij} + \beta_2^l f(A_{i(-j)}) + \beta_3^l \boldsymbol{X_{ij}} + \beta_4^l \boldsymbol{X_{i(-j)}} + \xi_i + \epsilon_{ij}, \tag{4.2}$$

where ϵ_{ij} are random error terms, ξ_i is the random effect term that induces the dependency among units in the same cluster, and $f(\cdot)$ is a summary statistics function of the treatment vector that introduces the interference among units. Without loss of generality, we assume that the true propensity score model is $\pi_A^1(x_i;\alpha^1)$, and the true outcome regression model is $m^1(a_i,x_i;\beta^1)$. Let $w_{ij}^a(\boldsymbol{X_i}) = P_{\alpha,x}(A_{i(-j)})/f(A_{ij}=a, A_{i(-j)}|\boldsymbol{X_i})$, where $a \in \{0,1\}$. It is straightforward to verify that

$$E\Bigg[w_{ij}^a(X_i)\Bigg\{\pi_A^j(X_i;\hat{\alpha}^j) - E\Bigg(\pi_A^j(\boldsymbol{X_i};\hat{\alpha}^j)\Bigg)\Bigg\}|A_{ij}=a\Bigg] = 0 \tag{4.3}$$

$$E\Bigg[w_{ij}^a(\boldsymbol{X_i})\Bigg\{m_{ij}^l(A_i,\boldsymbol{X_i};\hat{\beta}^l) - E\Bigg(m_{ij}^l(A_i,\boldsymbol{X_i};\hat{\beta}^l)\Bigg)\Bigg\}|A_{ij}=a\Bigg] = 0 \tag{4.4}$$

Let $\hat{\theta}^j(\hat{\alpha}^j) = 1/K\sum_{i=1}^K 1/n_i \sum_{j=1}^{n_i} \pi_A^j(X_i;\hat{\alpha}^j)$ and $\hat{\eta}^l(\hat{\beta}^l) = 1/K\sum_{i=1}^K 1/(n_i)\sum_j \sum_{a_{i(-j)}} m_{ij}^l(A_{ij}=a, A_{i(-j)}=a_{i(-j)};\hat{\beta}^l)P_{\alpha,x}(a_{i(-j)})$ be the empirical means for expectations $E\Bigg(\pi_A^j(X_i;\hat{\alpha}^j)\Bigg)$ and $E\Bigg(m_{ij}^l(A_{ij},A_{i(-j)},X_i;\hat{\beta}^l)|A_{ij}=a\Bigg)$, respectively. We replace the inner expectation in (4.3) and (4.4) by the empirical estimates $\hat{\theta}^j(\hat{\boldsymbol{\alpha}}^j)$ and $\hat{\eta}^l(\hat{\boldsymbol{\beta}}^l)$, and utilize the empirical likelihood approach to estimate the weights by maximizing $\prod_{i,j} w_{ij}^a$ with the following constraints:

$$\begin{aligned} &\sum\sum_{i,j:A_{ij}=a} w_{ij}^{a} = 1, \quad \sum\sum_{i,j:A_{ij}=a} w_{ij}^{a} \pi^{j}\left(\boldsymbol{X_i}; \hat{\boldsymbol{\alpha}}^{j}\right) = \hat{\theta}^{j}(\hat{\boldsymbol{\alpha}}^{j}) \quad (j = 1, \ldots, J), \\ &\sum\sum_{i,j:A_{ij}=a} w_{ij}^{a} m_{ij}^{l}\left(A_i, \boldsymbol{X_i}; \hat{\boldsymbol{\beta}}^{l}\right) = \hat{\eta}^{l}(\hat{\boldsymbol{\beta}}^{l}) \quad (l = 1, \ldots, L), \end{aligned} \tag{4.5}$$

for $a \in \{0, 1\}$. Then, the empirical likelihood weights are obtained through the Lagrange multiplier's method, which has the following representation:

$$\widehat{w}_{ij}^{a} = \frac{1}{\sum_i n_{i,a}} \frac{1}{1 + \widehat{\boldsymbol{\lambda}}^{T} \boldsymbol{g}_{ij}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}})}, \tag{4.6}$$

where $n_{i,a}$ is the number of observations in the group i such that $A_{ij} = a$, $\widehat{\boldsymbol{\alpha}} = \left\{(\widehat{\boldsymbol{\alpha}}^1)^\top, (\widehat{\boldsymbol{\alpha}}^2)^\top, \ldots, (\widehat{\boldsymbol{\alpha}}^J)^\top\right\}^\top$, $\widehat{\boldsymbol{\beta}} = \left\{\left(\widehat{\boldsymbol{\beta}}^1\right)^\top, \left(\widehat{\boldsymbol{\beta}}^2\right)^\top, \ldots, \left(\widehat{\boldsymbol{\beta}}^L\right)^\top\right\}^\top$,

$$\boldsymbol{g}_{ij}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}) = \left\{ \begin{array}{c} \pi_A^1\left(\mathbf{x}_i; \widehat{\boldsymbol{\alpha}}^1\right) - \widehat{\theta}^1 \\ \vdots \\ \pi_A^J\left(\mathbf{x}_i; \widehat{\boldsymbol{\alpha}}^J\right) - \widehat{\theta}^J \\ m_{ij}^1\left(\mathbf{x}_i; \widehat{\boldsymbol{\beta}}^1\right) - \widehat{\eta}^1 \\ \vdots \\ m_{ij}^J\left(\mathbf{x}_i; \widehat{\boldsymbol{\beta}}^L\right) - \widehat{\eta}^L \end{array} \right\},$$

and $\hat{\lambda}$ is the solution of

$$U(\widehat{\boldsymbol{\lambda}}) = \sum\sum_{i,j:A_{ij}=a} \frac{\boldsymbol{g}_{ij}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}})}{1 + \widehat{\boldsymbol{\lambda}} \boldsymbol{g}_{ij}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}})} = \mathbf{0}. \tag{4.7}$$

The proposed multiply robust estimator $\hat{\mu}_{a\alpha}^{\mathrm{mr}}$ for $\mu_{a\alpha}^{0}$ is given by solving the following estimation equation:

$$\sum_{i=1}^{K} W_i(\boldsymbol{Y}_i - \mu_{i,a\alpha}) = \mathbf{0}, \tag{4.8}$$

where $W_i = \mathrm{diag}\{\hat{w}_{i1}^{a} 1(A_{i1} = a), \hat{w}_{i2} 1(A_{i2} = a), \cdots, \hat{w}_{in_i}^{a} 1(A_{ij} = a)\}$, and $\boldsymbol{\mu}_{a\alpha} = (\mu_{1,a\alpha}, \mu_{2,a\alpha}, \cdots, \mu_{K,a\alpha})^T$. Then, under the regularity conditions similar to Theorem 2 in Han and Wang

[2013], we have the following theorem,

Theorem 4.1. *(Multiple robustness) Under suitable regularity conditions, if either (1) any one candidate model of the propensity score of treatment is correctly specified, or (2) any one of the candidates' outcome models $f(y_{ij}|\boldsymbol{a}_i, \boldsymbol{x}_i)$ are correctly specified, then $\hat{\boldsymbol{\mu}}^{\mathbf{mr}}_{\mathbf{a}\alpha} \to \boldsymbol{\mu}^{\mathbf{0}}_{\mathbf{a}\alpha}$ in probability as $K \to \infty$.*

4.3.2 Multiply Robust Estimator with Confounders Missing Not at Random

The methodology presented above is to estimate the network causal effects with complete data. In this section, we extend the idea to the setting when confounders are subject to nonignorable missingness. In equations (4.3) and (4.4), if we replace the group-level propensity score of treatment with the group-level joint propensity score of treatment and missingness, the two equations will still hold. Therefore, to obtain the multiply robust estimators, the specification of a set of paired propensity score models of treatment given observed data and the propensity score models of missingness are required for estimation. First, we postulate multiple models $\mathcal{P}^a = \{\pi^j_A(x; \alpha^j),\ \ j = 1, 2, ...J\}$ for propensity score of treatment $\Pr(A_{ij} = 1 \mid R_{ij} = 1, \boldsymbol{X_i} = \boldsymbol{x_i}, b_i)$, $\mathcal{P}^r = \{\pi^j_R(x; \gamma^j),\ \ j = 1, 2, ...J\}$ for $\Pr(R_{ij} = 1 \mid A_i = a_i, \boldsymbol{X_{ij}} = \boldsymbol{x_{ij}})$. We postulate a set of candidate outcome models $\mathcal{A}^y = \{f^l(y|a, x, r = 1; \beta^{ly}),\ \ l = 1, 2, ...L\}$ for the outcome regression model $f(Y|A, X, R = 1)$, and a set of models $\mathcal{A}^x = \{f^{(l)}(a; \beta^{lx}),\ \ l = 1, 2, ...L\}$ for the distribution of confounders $f(X|A, R = 1)$, where b_i are group-level random effect terms. Let $\boldsymbol{\beta^l} = (\beta^{lx}, \beta^{ly})^T$. Each pair of α^j, γ^j are estimated through maximizing the following binomial likelihood,

$$\prod_{i=1}^{K} \int \prod_{j=1}^{n_i} \left\{h^{11}_{ij}(b_i)\right\}^{A_{ij}R_{ij}} \left\{h^{01}_{ij}\left(b_i\right)\right\}^{(1-A_{ij})R_{ij}} f_b\left(b_i; \sigma^2\right) db_i,$$

where $h^{ar}_{ij}(b_i) = \Pr(a_{ij} = a, r_{ij} = r|\boldsymbol{x_{ij}}, b_i)$. Similarly, under the group-level outcome-independent missingness assumption, each β^l is estimated through the maximum likelihood approach based on observed data. In order to recover the conditional expectation of the outcome given the treatment and confounders from the observed samples

$(i=1,2,\cdots K,\ \ j=1,2,\cdots m_i)$, we impose the following constraints on the weights w_{ij}:

$$\begin{aligned}
&\sum_{i=1}^{K}\sum_{j=1}^{m_i} w_{ij}^{a}=1, \quad w_{ij}^{a}>0,\\
&\sum_{i=1}^{K}\sum_{j=1}^{m_i} w_{ij}^{a}\pi_{AR}^{j}\left(\hat{\boldsymbol{\alpha}}^{\boldsymbol{j}},\hat{\boldsymbol{\gamma}}^{\boldsymbol{j}};\boldsymbol{X_i}\right)=\hat{\theta}_{AR}^{j}(\hat{\boldsymbol{\alpha}}^{j},\hat{\boldsymbol{\gamma}}^{j}) \quad (j=1,\ldots,J)\\
&\sum_{i=1}^{K}\sum_{j=1}^{m_i} w_{ij}^{a}m_{ij}^{l}\left(\hat{\boldsymbol{\beta}}^{\boldsymbol{l}};A_i,\boldsymbol{X_i}\right)=\hat{\eta}^{l}(\hat{\boldsymbol{\beta}}^{l}) \quad (l=1,\ldots,L),
\end{aligned}$$

where $\hat{\theta}_{AR}^{j} = \sum_{i=1}^{K}\sum_{j=1}^{n_i}\pi_{AR}^{j}\left(\boldsymbol{X_i};\hat{\boldsymbol{\alpha}}^{j},\hat{\boldsymbol{\gamma}}^{j}\right)$, and $\hat{\eta}^{l} = 1/K\sum_{i=1}^{K}1/n_i\sum_{j=1}^{n_i}\sum_{a_{i(-j)}} m_{ij}^{l} \left(A_{ij}=a, A_{i(-j)}=a_{i(-j)},\boldsymbol{X_i};\hat{\boldsymbol{\beta}}^{l}\right) P_{\alpha,x}(a_{i(-j)})$, and $\pi_{AR}^{j}\left(\hat{\boldsymbol{\alpha}}^{\boldsymbol{j}},\hat{\boldsymbol{\gamma}}^{\boldsymbol{j}};\boldsymbol{X_i}\right)$ is obtained by equation (3.11). Here, the first constraint is imposed for regularization, and the second to fourth constraints equate the weighted average of each parametric function evaluated at the observed samples to the corresponding unweighted sample mean. From the above constraints, we can see $\hat{\theta}_{R}^{j}(\hat{\boldsymbol{\gamma}}^{j})$, $\hat{\theta}_{A}^{j}(\hat{\boldsymbol{\alpha}}^{j})$ and $\hat{\eta}^{l}(\hat{\boldsymbol{\beta}}^{l})$ are not estimable through sample average, because X_i are subject to missingness. To make progress, we propose a novel procedure here to estimate the expectation of model averages in this section. It is worth noting that Li et al. [2020b] proposed a similar procedure to recover the population mean when the outcome is subject to missingness. The main difference lies in two main aspects. First, the outcome is missing not at random, but the confounders are assumed to be fully observed in their setting. Thus, the expectation of their outcome models can be directly estimated through the sample average. Second, they did not consider the cluster structure within the dataset. Moreover, in order to guarantee consistency, one outcome model and one propensity score model have to be correctly specified simultaneously in Li et al. [2020b], whereas in our setting, the consistency of the estimator follows if one set of working models, either the set of the propensity score models, or the set of the joint models of outcome and confounders, but not necessarily the both, is correctly specified.

To proceed, we first notice that, for any vectorized differentiable function $l(a,x,y)$, we have

$$f(x_{ij},y_{ij}\mid a_i,r_{ij}=0)=\frac{\eta\left(r_{ij}=0,r_0,x_{ij},x_0=0\mid a_i\right)f(x_{ij},y_{ij}\mid a_i,r_{ij}=1)}{E\left\{\eta\left(r_{ij}=0,r_0,X_{ij},x_0\mid a_i\right)\mid a_i,r_{ij}=1\right\}}, \tag{4.9}$$

and

$$\begin{aligned}E\{l(a_{ij}, X_{ij}, Y_{ij}) \mid a_i, r_{ij}=0\} &= \frac{\iint \eta\left(r_{ij}=0, r_0, x_{ij}, x_0 \mid a_i\right) f(x_{ij}, y_{ij} \mid a_i, r_{ij}=1) l(a_{ij}, x_{ij}, y_{ij}) dxdy}{E\left\{\eta\left(r_{ij}=0, r_0, x_{ij}, x_0 \mid a_i\right) \mid a_i, r_{ij}=1\right\}} \\ &= \frac{E\left\{\eta\left(r_{ij}=0, r_0, X_{ij}, x_0 \mid a_i\right) l(a_{ij}, X_{ij}, Y_{ij}) \mid a_i, r_{ij}=1\right\}}{E\left\{\eta\left(r_{ij}=0, r_0, X_{ij}, x_0 \mid a_i\right) \mid a_i, r_{ij}=1\right\}} \\ &= \frac{E\left\{R_{ij}\eta\left(r_{ij}=0, r_0, X_{ij}, x_0 \mid a_i\right) l(a_{ij}, X_{ij}, Y_{ij}) \mid a_i\right\}}{E\left\{R_{ij}\eta\left(r_{ij}=0, r_0, X_{ij}, x_0 \mid a_i\right) \mid a_i\right\}},\end{aligned} \tag{4.10}$$

where

$$\begin{aligned}\eta\left(r_{ij}, r_0, x_{ij}, x_0 \mid a_i\right) &= \frac{\Pr(r_{ij} \mid a_i, x_{ij}) \Pr\left(r_0 \mid a_i, x_0\right)}{\Pr\left(r_0 \mid a_i, x_{ij}\right) \Pr\left(r_{ij} \mid a_i, x_0\right)} & (4.11) \\ &= \frac{\Pr(r_{ij} \mid a_i, x_{ij}, y_{ij}) \Pr\left(r_0 \mid a_i, x_0, y_{ij}\right)}{\Pr\left(r_0 \mid a_i, x_{ij}, y_{ij}\right) \Pr\left(r_{ij} \mid a_i, x_0, y_{ij}\right)} & (4.12) \\ &= \frac{f(x_{ij}, y_{ij} \mid a_i, r_0) f\left(x_0, y_0 \mid a_i, r_{ij}\right)}{f\left(x_{ij}, y_{ij} \mid a_i, r_{ij}\right) f\left(x_0, y_0 \mid a_i, r_0\right)}, & (4.13)\end{aligned}$$

$x_0 = y_0 = r_0 = 0$. Let $l_{ij}(\boldsymbol{X_i}) = \Pr(A_i, R_{ij} \mid \boldsymbol{X_i})$, then $\hat{\theta}_{AR},\ \ j \in \{1, 2, ..J\},\ \ l \in \{1, 2, ...L\}$ are obtained by

$$\frac{1}{K}\sum_{i=1}^{K}\frac{1}{n_i}\sum_{j=1}^{n_i}\left[(1-R_{ij})E\left\{\hat{\pi}^j_{AR}(\boldsymbol{X_i}; \boldsymbol{\hat{\alpha}^l}, \boldsymbol{\hat{\gamma}^l}) \mid a_i, r_{ij}=0; \hat{\beta}^l, \hat{\alpha}^j\right\} + R_{ij}\hat{\pi}^j_{AR}(A_i, \boldsymbol{X_i}; \boldsymbol{\hat{\gamma}^l})\right], \tag{4.14}$$

and $\hat{\eta}^l_{AR},\ \ l \in \{1, 2, ...L\}$ are obtained by

$$\frac{1}{K}\sum_{i=1}^{K}\frac{1}{n_i}\sum_{j=1}^{n_i}\left[(1-R_{ij})E\left\{\hat{m}^l_{ij}(A_i, \boldsymbol{X_i}; \boldsymbol{\hat{\beta}^l}) \mid a_i, r_{ij}=0; \boldsymbol{\hat{\beta}^l}, \boldsymbol{\hat{\beta}'^l}\right\} + R_{ij}\hat{m}^l_{ij}(A_i, \boldsymbol{X_i}; \boldsymbol{\hat{\beta}^l})\right], \tag{4.15}$$

where each $\hat{\beta'}^l$ is estimated by solving

$$\sum_{i=1}^{K}\frac{1}{n_i}\sum_{j=1}^{n_i}[(1-R_{ij})\{l(A_i, Y_i) - E\{l(A_{ij}, Y_{ij}) \mid A_i, R_{ij}=0; \hat{\beta}^l, \hat{\beta'}^l\}\}] = 0. \tag{4.16}$$

To avoid excess notations, we use β^l instead of (β^l, β'^l) for the rest of the chapter. Let

$$\hat{\boldsymbol{\alpha}}^{\mathrm{T}} = \left\{ \left(\hat{\alpha}^1\right)^{\mathrm{T}}, \ldots, \left(\hat{\alpha}^J\right)^{\mathrm{T}} \right\}, \quad \hat{\boldsymbol{\gamma}}^{\mathrm{T}} = \left\{ \left(\hat{\gamma}^1\right)^{\mathrm{T}}, \ldots, \left(\hat{\gamma}^J\right)^{\mathrm{T}} \right\}$$

$$\hat{\boldsymbol{\beta}}^{\mathrm{T}} = \left\{ \left(\hat{\beta}^1\right)^{\mathrm{T}}, \ldots, \left(\hat{\beta}^L\right)^{\mathrm{T}} \right\}, \text{ and}$$

$$\boldsymbol{g}_{ij}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\beta}}) = \left\{ \begin{array}{c} \pi_A^1\left(\mathbf{x}_i; \widehat{\boldsymbol{\alpha}}^1, \widehat{\boldsymbol{\gamma}}^1\right) - \widehat{\theta}^1_{AR} \\ \vdots \\ \pi_R^J\left(\mathbf{x}_i; \widehat{\boldsymbol{\alpha}}^J, \widehat{\boldsymbol{\gamma}}^J\right) - \widehat{\theta}^J_{AR} \\ m_{ij}^1\left(a_i, \mathbf{x}_i; \widehat{\boldsymbol{\beta}}^1\right) - \widehat{\eta}^1 \\ \vdots \\ m_{ij}^J\left(a_i, \mathbf{x}_i; \widehat{\boldsymbol{\beta}}^L\right) - \widehat{\eta}^L \end{array} \right\}. \tag{4.17}$$

If $\rho^T = (\rho^1, ..., \rho^{J+L})$ is a (J + L)-dimensional vector satisfying the equation

$$\sum_{i=1}^{K} \sum_{j=1}^{m_i} \frac{\hat{g}_{ij}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}})}{1 + \rho^{\mathrm{T}} \hat{g}_{ij}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}})} = 0, \tag{4.18}$$

then, by empirical likelihood theory, the solution to equation (4.17) is given by

$$\hat{w}_{ij}^a = \frac{1}{\sum_i n_{i,ar}} \frac{1}{1 + \hat{\boldsymbol{\rho}}^{\mathrm{T}} \hat{\boldsymbol{g}}_{ij}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}})}$$

with

$$1 + \hat{\boldsymbol{\rho}}^{\mathrm{T}} \hat{\boldsymbol{g}}_{ij}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}}) > 0, \quad j = 1, \ldots, n_i,$$

where $n_{i,ar}$ is the number of observations in the group i satisfying $A_{ij} = a$ and $R_{ij} = 1$. In practice, directly solving equation (4.18) may lead to multiple roots. Therefore, in order to guarantee uniqueness of the $\boldsymbol{\rho}$, we utilize the methods proposed by Han [2014a] to estimate $\boldsymbol{\rho}$ by minimizing the convex function $F(\rho) = -\sum_{i,j} [\log\{1 + \boldsymbol{\rho}^{\mathrm{T}} \hat{\boldsymbol{g}}_{ij}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}})\}]$. Then, the proposed multiply robust estimator $\hat{\boldsymbol{\mu}}_{a\alpha}^{\mathrm{mr}}$ for $\boldsymbol{\mu}_{a\alpha}^0$ is obtained by solving the following estimating equation:

$$\sum_{i=1}^{K} W_i(\boldsymbol{Y}_i - \mu_{i,a\alpha}) = \mathbf{0}, \tag{4.19}$$

where $W_i = \text{diag}\{\hat{w}_{i1}^a 1(A_{i1} = a)1(R_{i1} = 1), \hat{w}_{i2}^a 1(A_{i2} = a)1(R_{i2} = 1), \cdots, \hat{w}_{in_i}^a 1(A_{ij} = a)1(R_{in_i} = 1)\}$, and $\boldsymbol{\mu}_{a\alpha} = (\mu_{1,a\alpha}, \mu_{2,a\alpha}, \cdots, \mu_{K,a\alpha})^T$.

Theorem 4.2. *(Multiple robustness) Under suitable regularity conditions, if either (1) any one pair of candidate propensity score models, i.e., propensity score of treatment and propensity score of missingness, are correctly specified, or (2) any one pair of the candidate odds ratio model $\eta(r, r_0, x, x_0)$ and the candidate joint density model $f(x, y|\boldsymbol{a}, r = 1)$ are correctly specified, then $\hat{\boldsymbol{\mu}}_{a\alpha}^{mr} \to \boldsymbol{\mu}_{a\alpha}^0$ in probability as $K \to \infty$.*

The above multiple robustness theorem is different from that when data are fully observed. Notice that the consistency of the proposed estimator only requires the correct specification of either the propensity score model of treatment or the joint density model of confounders and the outcome. When confounders are missing not at random, the joint propensity score of missingness and the treatment requires the specification of the propensity score model of treatment and the propensity score model of missingness. The joint model of the confounders and the outcome conditional on missing data requires both the specification of the odds ratio model and the specification of the joint model of the outcome and the confounders given observed data.

Under partial interference, we assumed a linear and a logistic mixed effects model for the outcome regression model and the propensity score of treatment. Although the cluster structure does not explicitly affect the derivation of the proposed multiply robust estimator after integration, as only marginal distribution models are utilized in the estimation procedure, the limiting distribution of the proposed estimators is dependent on the covariance structure. Since the derivation of the asymptotic variance is challenging and is beyond the scope of this paper, we leave it as a future research topic. To circumvent the difficulty of variance estimation, we estimate the standard errors of the estimators by a cluster-based bootstrapping procedure. More specifically, we take a simple random sample with the replacement of K clusters from the original K clusters in the study population to form a bootstrap sample $\{O_i = (Y_{ij}, X_{ij}, A_{ij}, R_{ij}), \quad i = 1, 2, \cdots K, \quad j = 1, 2 \cdots n_i\}$. Calculate $\hat{\mu}_n^{\text{mr}}$ based on the B (e.g., $B = 9999$) bootstrap samples. Then, under some regularity conditions, we have $var(\hat{\mu}_n^{\text{mr}}) \to se^2(\hat{\mu}^{\text{mr}})$ in probability as $n \to \infty$ (see more details in Chen et al. [2021]).

4.4 Simulation

In this section, we study the finite performance of the proposed estimators. We first generate a population with $K = 200$ groups, and $n_i = 40$ units in each of the groups. For each unit, the covariate X_1 is generated from the Bernoulli distribution with expectation 0.5, and X_2 is generated by the standard normal distribution.

In Scenario 1, we consider the case when data are fully observed. Let $\alpha = 0.5$. The treatment indicators are generated from a mixed effects logistic model $\text{logit}\{\Pr(A_{ij} = 1 \mid X_{ij} = x_{ij}, b_i)\} = -0.1 + 0.1x_{1ij} - 0.2x_{2ij} - 0.15x_{1ij}x_{2ij} + b_i$, where b_i is group-level random effect terms generated from $\mathcal{N}(0, 0.5)$. In Scenario 2, we assume the propensity score of treatment as a mixed effects logistic model $\text{logit}\{\Pr(A_{ij} = 1 \mid X_{ij} = x_{ij}, R_{ij} = 1, b_i)\} = 0.2 + 0.1x_{1ij} - 0.1x_{2ij} - 0.1x_{1ij}x_{2ij} + b_i$. We let the model of propensity score of missingness as $\text{logit}\{\Pr(R_{ij} = 1 \mid A_{ij} = a_{ij}, X_{ij} = x_{ij}\} = 1 + 1.2a_{ij} - 0.5x_{1ij} + 1x_{2ij}$. The missingness rate is 15% under such a setting.

In both scenarios, the outcomes are generated from $y_{ij} = 1 + 2a_{ij} + 2p_i(a_i) - 3x_{1ij} + 0.5x_{2ij} + 2x_{1ij}x_{2ij} + \xi_i + \epsilon_{ij}$, where $\xi_i \sim \mathcal{N}(0, 0.25)$ is the group level random effect terms generated from $\mathcal{N}(0, 0.25)$, $p_i(a_i)$ is the proportion of units in group i that receive the treatment, and $\{\epsilon_{ij}\}_{1\le i\le K, 1\le j\le N_i}$ are i.i.d. random noise terms generated by $\mathcal{N}(0, 0.25)$.

Each candidate regression models and the propensity score models are fit, and the parameters α^j, β^l, and γ^j are estimated through the maximum likelihood approach. $\hat{m}^l_{ij}(A_i, \boldsymbol{X_i}; \hat{\beta})$, $\hat{\pi}^j_A(\boldsymbol{X_i}; \hat{\alpha})$ and $\hat{\pi}^j_{AR}(\boldsymbol{X_i}; \hat{\alpha}, \hat{\gamma})$ are calculated accordingly, where $\hat{m}^l_{ij}(A_i, \boldsymbol{X_i}; \hat{\beta})$ are calculated by $\hat{f}_{ij}(Y_i|A_i, \boldsymbol{X_i}; \hat{\beta}) = \int \hat{f}_{ij}(Y_i|A_i, \boldsymbol{X_i}; \hat{\beta}, \xi_i)dP(\xi_i)$

The MR estimator is calculated according to the equations shown in Section 3.1. For both Scenarios 1 and 2, three misspecified outcome models $E[Y_{ij} \mid A_i, X_i] = \beta_0 + \beta_1 A_{ij} + \beta_2 x_{1ij}$, $E[Y_{ij} \mid A_i, X_i] = \beta_0 + \beta_1 A_{ij} + \beta_2 x_{2ij}$, and $E[Y_{ij} \mid A_i, X_i] = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij}$ are fit, and the parameters are calculated via MLE.

In Scenario 1, three misspecified propensity score models for treatment $\text{logit}\{\Pr(A_{ij} = 1 \mid X_{ij} = x_{ij}, b_i^{(1)})\} = \alpha_0 + \alpha_1 x_{1ij} + b_i$, $\text{logit}\{\Pr(A_{ij} = 1 \mid X_{ij} = x_{ij})\} = \alpha_0 + \alpha_1 x_{2ij} + b_i$, $\text{logit}\{\Pr(A_{ij} = 1 \mid X_{ij} = x_{ij})\} = \alpha_0 + \alpha_1 |x_{1ij}| + b_i$ are fit and parameters are estimated through MLE. In Scenario 2, when confounders are subject to missingness, two misspecified propensity score models for missingness $\text{logit}\{\Pr(R_{ij} = 1 \mid A_{ij} = a_{ij}, X_{ij} = x_{ij})\} = \gamma_0 + \gamma_1 x_{1ij}$ and $\text{logit}\{\Pr(R_{ij} = 1 \mid A_{ij} = a_{ij}, X_{ij} = x_{ij})\} = \gamma_0 + \gamma_1 x_{2ij}$ are fit, and one propensity score model for treatment $\text{logit}\{\Pr(A_{ij} = 1 \mid X_{ij} = x_{ij}, r_{ij} = 1, b_i)\} = \alpha_0 + \alpha_1 x_{1ij} + b_i$ is fit. The odds ratio model is specified $\eta(r = 0, r_0, x, x_0; \beta^{'}) = \exp(\beta^{'}_1 x_1 + \beta^{'}_1 x_2)$, and the parameters in the model are estimated by equation 4.16.

To further illustrate the performance of the proposed estimators under different missingness rates, we consider Scenario 3, where the propensity score and the regression models are assumed to be the same as those in Scenario 2, but the true parameters in the propensity score of missingness are set to be $\gamma^0 = (0.35, 1, -0.5, 1)$. In scenario 3, the missingness rate is 25%.

The simulation study is repeated 200 times under Scenarios 1, 2, and 3. The bias and variance are summarized in Tables 4.1 and 4.2.

The results of the proposed multiply robust estimators for both Scenarios S1 and S2 with (1) $K = 150$, $n_i = 30$ and (2) $K = 250$, $n_i = 50$ are summarized in Table 4.1, and the results for scenario 3 are summarized in 4.2. Notice that the proposed estimators achieve smaller biases, if at least one set of models, either the set of the regression models (outcome regression models and the distribution models of observed confounders) or the set of propensity score models (propensity score of treatment and the missingness mechanism), is correctly specified, which is consistent with Theorems 4.1 and 4.2. The proposed estimators have relatively smaller biases for the scenario when both one set of correct propensity score models and one set of correct joint models of the outcome and the confounders are included in the candidate working models. When the missingness rate is 25%, the proposed estimators have a relatively larger bias than those under 15% missingness rate, which implies that the estimators perform betters with a smaller missingness rate. In Scenario 2-(2), the estimators have slightly smaller bias and variance compared to Scenario 2-(1) due to the larger number of observations in each cluster. The results show that the proposed estimators have improved the doubly robust estimators by providing extra robustness against model misspecification.

4.5 Application

In this section, we apply the proposed estimators on the NO_x emissions data to quantify the causal effects of scrubbers' installation on the reduction of the amount of NO_x. The motivation and introduction of the data can be found in Section 3.5. In this study, the outcome variable Y is the amount of NO_x emissions in each of 1218 coal-fired power generating units and is measured across 24 months from Jan 2004 to Dec 2005 in the U.S. The study population is split into disjoint clusters by their geographical locations and the linkage algorithm (see more details in Section 3.5). We let $\boldsymbol{X} = (X_1, X_2, X_3, X_4, X_5)^T$ be the confounders that are subject to missingness, where X_1 represents the heat input rate, X_2 denotes the capacity of the EGU, X_3 represents the operation time, X_4 denotes the amount of coal, and X_5 denotes the average temperature. For the missingness rate, there is 24.90% missingness in confounders in total. Specifically, there is 15.76% missingness in the amount of coal, 2.25% missingness in operation time, 21.69% missingness in heat input rate, 0.99% in capacity. The potential reasons for missingness can be found in Section 3.5.

Table 4.1: Bias, standard deviation (SD), and empirical standard error (SE) for the proposed multiply robust estimators under scenarios S1 and S2 with 15% missingness rate.

		MR-TFF-TFF			MR-TFF-FFF			MR-FFF-TFF			MR-FFF-FFF		
		Bias	SD	SE	Bias	SD	SE	Bias	SD	SE	Bias	SD	SE
S1-(1)	$\mu_{1,0.5}$	0.011	0.037	0.048	0.012	0.048	0.043	-0.034	0.037	0.044	-0.102	0.987	1.165
	$\mu_{0,0.5}$	-0.009	0.029	0.033	-0.034	0.042	0.051	0.057	0.032	0.035	0.054	1.781	1.527
	$\bar{\text{D}}$E(0.5)	0.021	0.054	0.069	0.057	0.065	0.082	-0.092	0.061	0.082	-0.156	1.769	1.975
S2-(1)	$\mu_{1,0.5}$	0.016	0.048	0.054	0.017	0.092	0.095	0.007	0.084	0.078	-0.907	2.968	3.268
	$\mu_{0,0.5}$	-0.006	0.033	0.045	0.092	0.060	0.073	0.015	0.072	0.078	-3.787	4.825	4.359
	$\bar{\text{D}}$E(0.5)	0.023	0.064	0.067	-0.075	0.098	0.114	0.022	0.092	0.102	2.880	5.728	5.418
S1-(2)	$\mu_{1,0.5}$	0.014	0.045	0.064	0.019	0.010	0.088	0.041	0.055	0.051	0.934	3.982	3.768
	$\mu_{0,0.5}$	-0.005	0.029	0.025	0.084	0.071	0.083	0.023	0.068	0.058	2.384	4.788	5.265
	$\bar{\text{D}}$E(0.5)	0.019	0.056	0.065	-0.065	0.092	0.090	0.018	0.072	0.065	-1.451	5.752	6.418
S2-(2)	$\mu_{1,0.5}$	0.010	0.043	0.054	0.014	0.062	0.078	0.038	0.076	0.085	1.892	3.452	3.916
	$\mu_{0,0.5}$	-0.009	0.040	0.045	0.082	0.068	0.063	0.014	0.072	0.078	-2.182	4.129	4.378
	$\bar{\text{D}}$E(0.5)	0.020	0.056	0.060	-0.068	0.078	0.076	0.024	0.089	0.095	4.047	5.420	5.619

Table 4.2: Bias, standard deviation (SD), and empirical standard error (SE) for the proposed multiply robust estimators under scenarios S3 with 25% missingness rate.

		MR-TFF-TFF			MR-TFF-FFF			MR-FFF-TFF			MR-FFF-FFF		
		Bias	SD	SE	Bias	SD	SE	Bias	SD	SE	Bias	SD	SE
S3-(1)	$\mu_{1,0.5}$	0.026	0.045	0.052	0.045	0.072	0.069	0.036	0.061	0.068	2.208	3.368	3.646
	$\mu_{0,0.5}$	-0.017	0.038	0.044	-0.052	0.068	0.078	-0.065	0.055	0.067	-1.152	2.814	3.421
	$\bar{\mathrm{DE}}(0.5)$	0.043	0.062	0.073	0.097	0.084	0.092	0.101	0.068	0.074	3.360	3.826	4.189
S3-(2)	$\mu_{1,0.5}$	0.018	0.058	0.071	0.032	0.064	0.058	0.024	0.052	0.056	1.201	2.213	3.646
	$\mu_{0,0.5}$	-0.012	0.045	0.055	-0.042	0.076	0.072	-0.067	0.048	0.062	-1.041	2.825	2.457
	$\bar{\mathrm{DE}}(0.5)$	0.030	0.072	0.086	0.074	0.082	0.080	0.091	0.067	0.078	2.242	2.961	3.082

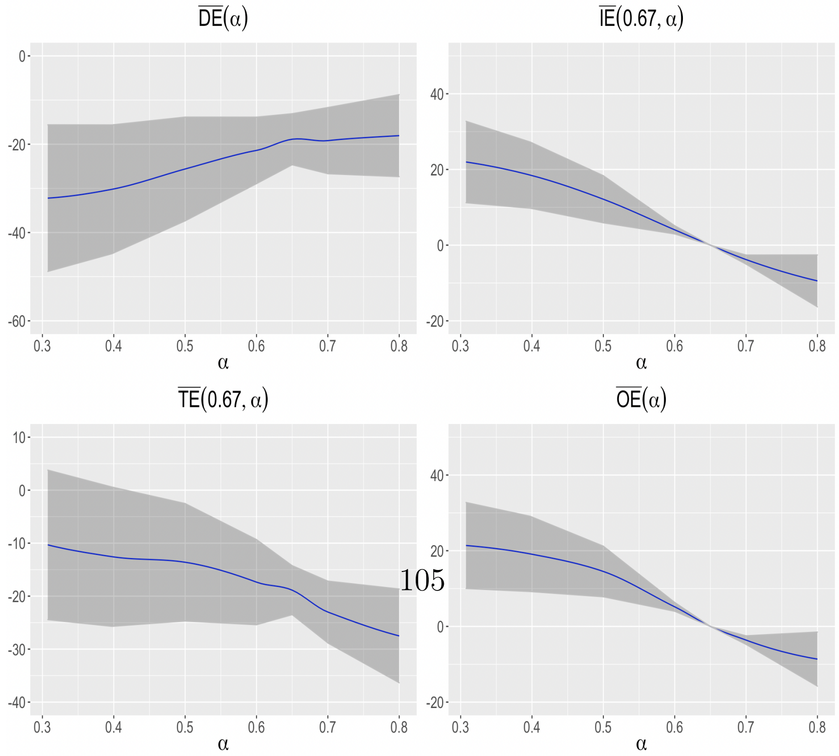
We consider two sets of propensity score models, wherein the first model of the propensity score of treatment includes all main effects. In the second model of the propensity score of treatment, we first include all the first-order interaction terms and utilize LASSO (Tibshirani [1996]) to reduce the magnitude of the coefficients, where the tuning parameter is chosen via cross-validation. Specifically, the first set contains one linear logistic model for the propensity score of missingness and one linear mixed effects logistic model for the propensity score of treatment, that is, 1) $\text{logit}\{\pi_R^{(1)}(\boldsymbol{X_{ij}}; \gamma^1)\} = \gamma_0^1 + \gamma_1^1 A_{ij} + \boldsymbol{\gamma_2^1 X_{ij}}$, and $\text{logit}\{\pi_A^{(1)}(\boldsymbol{X_{ij}}; \alpha^1)\} = \alpha_0^1 + \boldsymbol{\alpha}^1 \boldsymbol{X_{ij}} + b_i^{(1)}$. In the second set, the selected interaction terms are included in the model of the propensity score of treatment. Let I denote the set of the pair of indices of the selected first-order interaction terms using LASSO, then we have 2) $\text{logit}\{\pi_R^{(1)}(\boldsymbol{X_{ij}}; \gamma^1)\} = \gamma_0^1 + \gamma_1^1 A_{ij} + \boldsymbol{\gamma_2^1 X_{ij}}$, and $\text{logit}\{\pi_A^{(2)}(\boldsymbol{X_{ij}}; \alpha^{(2)})\} = \alpha_0^{(2)} + \boldsymbol{\alpha_1}^{(2)} \boldsymbol{X_{ij}} + \sum_{(k,v)\in I} \alpha_{(k,v)}^{(2)} X_{ijk} X_{ijv} + b_i^{(2)}$. For the joint model of confounders and the outcome, we first assume the density function of confounders $f(x_{ij}|a_i, r_{ij} = 1)$ as $\mathcal{N}(\mu_x, \sigma_x^2)$, where $\mathcal{N}(\mu_x, \sigma_x^2)$ is the normal density function with mean μ_x and variance σ_x^2. We consider the outcome regression model as $Y_{ij} = \beta_0 + \beta_2 A_{ij} + \beta_3 f_s(A_i) + \boldsymbol{\beta_4 X_{ij}} + \xi_i + \epsilon_{ij}$, where $f_x(A_i)$ is the summary function that represents the proportion of the treated units in the group i, ϵ_{ij} are i.i.d. normally distributed random errors with mean zero and variance σ_e^2, and the density of ξ_i is assumed to be $\mathcal{N}(0, \sigma_\xi^2)$ that introduce the dependency between the units in the same cluster.

The estimated direct effect, indirect effect, total effect, and overall effect are summarized in Table 4.3 and Figure 4.1. The standard errors as well as the 95% bootstrap confidence intervals are estimated based on the cluster-based bootstrapping method. As we can see from the above table, the estimated network causal effects show similar patterns with the results presented in Section 3.5 using the proposed doubly robust estimator. For example, for all α values, the estimated direct effects are smaller than zero, which is consistent with the results in Section 3.5 that the scrubbers' installation indeed has a positive effect on the reduction of the amount of NO_2. As α increases, the estimated direct effect increases, implying that the effect of the installation of scrubbers decreases as the number of treated units in the same group increases. Compared with the DR estimators, MR estimators can provide more flexibility in the model specifications, hence it can lead to more reliable estimates.

Table 4.3: Estimated $\bar{\mathrm{D}}\mathrm{E}(\alpha)$ (direct effect), $\bar{\mathrm{I}}\mathrm{E}(0.67, \alpha)$ (indirect effect), $\bar{\mathrm{T}}\mathrm{E}(0.67, \alpha)$ (total effect), $\bar{\mathrm{O}}\mathrm{E}(\alpha)$ (overall effect), and standard errors (SE) for the proposed multiply robust estimators.

	$\bar{\mathrm{D}}\mathrm{E}(\alpha)$		$\bar{\mathrm{I}}\mathrm{E}(0.67, \alpha)$		$\bar{\mathrm{T}}\mathrm{E}(0.67, \alpha)$		$\bar{\mathrm{O}}\mathrm{E}(\alpha)$	
α	EST	SE	EST	SE	EST	SE	EST	SE
0.30	-32.043	4.318	21.954	2.825	-10.089	3.642	21.476	2.966
0.40	-30.875	3.806	18.574	2.1647	-13.501	3.539	18.793	2.548
0.50	-25.365	3.035	12.190	1.669	-13.274	2.762	14.571	1.787
0.60	-22.315	2.208	4.073	0.369	-18.241	2.513	5.652	0.385
0.67	-18.871	1.503	0	0	-18.871	1.202	0	0
0.70	-20.148	2.234	-3.906	0.398	-24.055	1.867	-3.326	0.370
0.80	-17.943	2.382	-9.438	1.775	-27.381	2.260	-8.671	1.851

Figure 4.1: Estimates and 95% bootstrap CIs of $\bar{DE}(\alpha)$, $\bar{IE}(0.67, \alpha)$, $\bar{TE}(0.67, \alpha)$, and $\bar{OE}(\alpha)$ of scrubbers' installation for MR estimators with $\alpha \in (0.3, 0.8)$, where the shadow area represents the pointwise confidence intervals.



4.6 Conclusion

In this chapter, we developed the multiply robust estimators for estimating network causal effects under partial interference when confounders are missing not at random. The proposed estimators can provide extra protection against model misspecification. Compared with doubly robust estimators where the consistency is achieved if any one of the two sets of working models is correctly specified, the extra robustness in multiply robust estimators comes from the specification of multiple working models, and the consistency is achieved if any one set of the candidate models is correctly specified.

In the current literature on multiply robust estimation when data are subject to missingness, little effort has been devoted to the setting of interference. One of the main challenges is to derive the asymptotic variance. To overcome the difficulty, we used a cluster-based bootstrapping method to calculate the standard errors. In the data application, we provide more robust evidence that the installation of scrubbers' installation has a positive effect on reducing the ambient PM2.5. It is also worth noting that the proposed estimators can have a large bias if none of the working models is correctly specified. Therefore, the selection of candidate working models is also important in practice.

There are several possible future research directions. First, the multiply robustness of the proposed estimator with variable selection methods remains to be further investigated. Second, under the current setting when the clustered data are missing not at random, it is of interest to explore the derivation of closed-form asymptotic variance estimators, and the efficiency of the proposed estimators also remains to be investigated. This paper considers the missingness in confounders, where the unit is assumed to be missing if any one component of the confounders of that unit is missing, it is possible to consider the multiple missingness pattern so that the information of each unit can be utilized more efficiently. Moreover, the performance of the proposed estimators in a more generalized interference pattern such as general interference is also worth investigating, and we leave it as a future research topic.

4.7 Proof of Theorems

In this section, we present the proof of Theorems 4.1 and 4.2.

4.7.1 Proof of Theorem 4.1

Without loss of generality, assume that $\pi_A^{(1)}(x;\alpha)$ and $f^{(1)}(a,x;\beta)$ are correctly specified models for the propensity score of treatment and the outcome regression model, respectively. Let $\alpha_\star$ and $\beta_\star$ be the corresponding true parameters. let $\hat{\alpha}^j$ and $\hat{\beta}^l$ be the estimated parameters for $\pi_A^{(l)}(x;\alpha)$ and $f^{(l)}(a,x;\beta)$, respectively. Assume there exist $\alpha_\star^j$ and $\beta_\star^l$ such that $\hat{\alpha}^l \xrightarrow{p} \alpha_\star^l$ and $\hat{\alpha}^l \xrightarrow{p} \alpha_\star^l$ as $K \to \infty$. First, we show that the consistency of $\hat{\mu}_{a\alpha}^{\mathrm{mr}}$ when $\pi_A^{(1)}(x;\alpha)$ is correctly specified. Note that when $\pi_A^{(1)}(x;\alpha)$ is correctly specified, we have $\hat{\alpha}^1 \to \alpha_\star^1$ in probability, and $\hat{\alpha}_\star^1 = \alpha_\star$. Following the derivations similar to Han [2014a,b], we notice that

$$\begin{aligned}
& \frac{1}{\sum_i n_{i,a}} \sum_{i=1}^{K} \sum_{j=1}^{n_{i,a}} \frac{\widehat{\boldsymbol{g}_{ij}}(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}}) P_\alpha(A_i) / \pi_{i,j}^1\left(\widehat{\boldsymbol{\alpha}}^1\right)}{1+\boldsymbol{\lambda}^{\mathrm{T}} \widehat{\boldsymbol{g}}_{ij}(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}}) P_\alpha(A_i) / \pi_{i,j}^1\left(\widehat{\boldsymbol{\alpha}}^1\right)} \\
= \; & \frac{1}{\theta^1\left(\widehat{\boldsymbol{\alpha}}^1\right)} \frac{1}{\sum_i n_{i,a}} \sum_{i=1}^{K} \sum_{j=1}^{n_{i,a}} \frac{\widehat{\boldsymbol{g}}_{ij}(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}}) P_\alpha(A_i)}{1+\frac{\pi_{i,j}^1(\widehat{\boldsymbol{\alpha}}^1)-\theta^1(\widehat{\boldsymbol{\alpha}}^1)}{\theta^1(\widehat{\boldsymbol{\alpha}}^1)} P_\alpha(A_i) + \left\{\frac{\boldsymbol{\lambda}}{\theta^1(\widehat{\boldsymbol{\alpha}}^1)}\right\}^{\mathrm{T}} \widehat{\boldsymbol{g}}_{ij}(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}}) P_\alpha(A_i)} \\
= \; & \frac{1}{\theta^1\left(\widehat{\boldsymbol{\alpha}}^1\right)} \frac{1}{\sum_i n_{i,a}} \sum_{i=1}^{K} \sum_{j=1}^{n_{i,a}} \frac{\widehat{\boldsymbol{g}}(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}}) P_\alpha(A_i)}{1+\left\{\frac{\lambda_1+1}{\theta^1(\widehat{\boldsymbol{\alpha}}^1)}, \frac{\lambda_2}{\theta^1(\widehat{\boldsymbol{\alpha}}^1)}, \ldots, \frac{\lambda_{J+K}}{\theta^1(\widehat{\boldsymbol{\alpha}}^1)}\right\} \widehat{\boldsymbol{g}_{ij}}(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}}) P_\alpha(A_i)}.
\end{aligned}$$

Therefore, the estimated weights $\hat{w}_{ij}^a$ has the following representation,

$$\frac{1}{\sum_i n_{i,a}} \frac{\theta^1\left(\widehat{\boldsymbol{\alpha}}^1\right) / \pi_{i,j}^1\left(\widehat{\boldsymbol{\alpha}}^1\right)}{1+\widehat{\boldsymbol{\lambda}}^{\mathrm{T}} \widehat{\boldsymbol{g}}_{ij}(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\gamma}}) / \pi_{i,j}^1\left(\widehat{\boldsymbol{\alpha}}^1\right)}, \tag{4.20}$$

and the consistency of $\mu_{a\alpha}^{\mathrm{mr}}$ follows:

$$\begin{aligned}
& \sum_{i=1}^{K}\sum_{j=1}^{n_i} 1(A_{ij}=a)\hat{w}_{ij}^{a} Y_{ij}(A_i) \\
= \quad & \sum_{i=1}^{K} \frac{\theta^1\left(\widehat{\boldsymbol{\alpha}}^1\right)}{\sum_i n_{i,a}} \sum_{j=1}^{n_i} \frac{1(A_{ij}=a)P_\alpha(A_i)/\pi_{i,j}^1\left(\widehat{\boldsymbol{\alpha}}^1\right)}{1+\widehat{\lambda}^{\mathrm{T}}\widehat{\boldsymbol{g}}_i(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}})P_\alpha(A_i)/\pi_{i,j}^1\left(\widehat{\boldsymbol{\alpha}}^1\right)} Y_{ij}(A_i) \\
= \quad & \frac{1}{K}\sum_{i=1}^{K} \frac{\theta^1\left(\widehat{\boldsymbol{\alpha}}^1\right)}{n_{i,a}} \sum_{j=1}^{n_i} \frac{1(A_{ij}=a)P_\alpha(A_i)/\pi_{i,j}^1\left(\widehat{\boldsymbol{\alpha}}^1\right)}{1+\widehat{\lambda}^{\mathrm{T}}\widehat{\boldsymbol{g}}_i(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}})P_\alpha(A_i)/\pi_{i,j}^1\left(\widehat{\boldsymbol{\alpha}}^1\right)} Y_{ij}(A_i) + o_p(1) \\
\xrightarrow{p} \quad & E\left[\frac{1}{n_i}\sum_{j=1}^{n_i} \frac{1(A_{ij}=a)Y_{ij}(A_i)P_\alpha(A_i)}{\pi_{i,j}^1(\boldsymbol{\alpha}_\star)}\right].
\end{aligned}$$

Second, if $f^{(1)}(a,x;\beta)$ is correctly specified, we have $\hat{\beta}^1 \to \beta^1_\star$ in probability, and $\beta^1_\star = \beta_\star$. The consistency of $\hat{\mu}^{\mathrm{mr}}_{a\alpha}$ follows:

$$
\begin{aligned}
& \sum_{i=1}^{K}\sum_{j=1}^{n_i} 1(A_{ij}=a)\hat{w}^a_{ij}Y_{ij}(A_i) \\
= \;& \sum_{i=1}^{K}\sum_{j=1}^{n_i} 1(A_{ij}=a)\hat{w}^a_{ij}\{Y_{ij}(A_i) - m^1_{ij}(A_i, X_{ij};\hat{\beta}^1)\} \\
& + \frac{1}{K}\sum_{i=1}^{K}\frac{1}{n_i}\sum_{j=1}^{n_i}\sum_{a_{i(-j)}} m^1_{ij}(a, a_{i(-j)}, X_{ij};\hat{\beta}^1)P_\alpha(a_{i(-j)}) \\
= \;& \sum_{i=1}^{K}\sum_{j=1}^{n_i} 1(A_{ij}=a)\hat{w}^a_{ij}\{Y_{ij}(A_i) - m^1_{ij}(A_i, X_{ij};\hat{\beta}^1)\} \\
& + E\left[\frac{1}{n_i}\sum_{j=1}^{n_i}\sum_{a_{i(-j)}} m^1_{ij}(a, a_{i(-j)}, X_{ij};\beta_\star)P_\alpha(a_{i(-j)})\right] + o_p(1) \\
= \;& \frac{1}{\sum_i n_{i,a}}\sum_{i=1}^{K}\sum_{j=1}^{n_i}\frac{1(A_{ij}=a)\{E^1(Y_{ij}(a_i)|A_i=a_i, X_{ij}) - m^1_{ij}(A_i, X_{ij};\beta_\star)\}}{1+\boldsymbol{\lambda}_\star^T\boldsymbol{g}_{ij}(\boldsymbol{\alpha}_\star, \boldsymbol{\beta}_\star)} \\
& + E\left[\frac{1}{n_i}\sum_{j=1}^{n_i}\sum_{a_{i(-j)}} m^1_{ij}(a, a_{i(-j)}, X_{ij};\beta_\star)P_\alpha(a_{i(-j)})\right] + o_p(1) \\
\xrightarrow{p} \;& \frac{1}{\Pr(A=a)}E\left[\frac{1(A_{ij}=a)\{E^1(Y_{ij}(a_i)|A_i=a_i, X_{ij}) - m^1_{ij}(A_i, X_{ij};\beta_\star)\}}{1+\boldsymbol{\lambda}_\star^T\boldsymbol{g}_{ij}(\boldsymbol{\alpha}_\star, \boldsymbol{\beta}_\star)}\right] \\
& + E\left[\frac{1}{n_i}\sum_{j=1}^{n_i}\sum_{a_{i(-j)}} m^1_{ij}(a, a_{i(-j)}, X_{ij};\beta_\star)P_\alpha(a_{i(-j)})\right] \\
= \;& \frac{1}{\Pr(A=a)}E\left[\frac{1}{n_{i,a}}\sum_{j=1}^{n_i}\frac{1(A_{ij}=a)\{E^1(Y_{ij}(a_i)|A_i=a_i, X_{ij}) - m^1_{ij}(A_i, X_{ij};\beta_\star)\}}{1+\boldsymbol{\lambda}_\star^T\boldsymbol{g}_{ij}(\boldsymbol{\alpha}_\star, \boldsymbol{\beta}_\star)}\right] + \mu^0_{a\alpha} \\
= \;& \mu^0_{a\alpha}.
\end{aligned}
$$

4.7.2 Proof of Theorem 4.2

Assume that $\pi^{(1)}_A(x;\alpha)$ and $\pi^{(1)}_R(a,x;\gamma)$ are correctly specified propensity score of treatment and propensity score of missingness, respectively. Let $\eta^{(1)}(r,x;\beta)$ and $f^{(1)}(a,x;\beta)$ be the correctly specified odds ratio model and joint model of confounders and the outcome. Let $\alpha_\star$ and $\beta_\star$ be the corresponding true parameters. let $\hat{\alpha}^1$, $\hat{\gamma}^1$, and $\hat{\beta}^1$ be the estimated parameters. First, we show that the consistency of $\hat{\mu}^{\mathrm{mr}}_{a\alpha}$ when $\pi_A^{(1)}(x;\alpha)$ and $\pi_R^{(1)}(a,x;\gamma)$ are correctly specified. Then, we have both $\hat{\alpha}^1 \to \alpha^1_\star$ and $\hat{\gamma}^1_\star \to \gamma^1_\star$ in probability as $K \to \infty$. It is straightforward to observe that $\alpha_\star = \alpha^1_\star$ and $\gamma_\star = \gamma^1_\star$. Similar to the derivations in Theorem 4.1, we have

$$
\begin{aligned}
& \frac{1}{\sum_i n_{i,ar}} \sum_{i=1}^{K} \sum_{j=1}^{n_{i,ar}} \frac{\widehat{\boldsymbol{g}_{ij}}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}) P_\alpha(A_i) / \pi^1_{i,j}\left(\widehat{\alpha}^1, \widehat{\gamma}^1\right)}{1 + \boldsymbol{\lambda}^{\mathrm{T}} \widehat{\boldsymbol{g}}_{ij}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}) P_\alpha(A_i) / \pi^1_{i,j}\left(\widehat{\alpha}^1, \widehat{\gamma}^1\right)} \\
= \; & \frac{1}{\theta^1\left(\widehat{\boldsymbol{\alpha}}^1, \widehat{\boldsymbol{\gamma}}^1\right)} \frac{1}{\sum_i n_{i,a}} \sum_{i=1}^{K} \sum_{j=1}^{n_{i,ar}} \frac{\widehat{\boldsymbol{g}}_{ij}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}) P_\alpha(A_i)}{1 + \frac{\pi^1_{i,j}(\widehat{\alpha}^1, \widehat{\gamma}^1) - \theta^1(\widehat{\alpha}^1, \widehat{\gamma}^1)}{\theta^1(\widehat{\alpha}^1, \widehat{\gamma}^1)} P_\alpha(A_i) + \left\{ \frac{\boldsymbol{\lambda}}{\theta^1(\widehat{\alpha}^1, \widehat{\gamma}^1)} \right\}^{\mathrm{T}} \widehat{\boldsymbol{g}}_{ij}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}) P_\alpha(A_i)} \\
= \; & \frac{1}{\theta^1\left(\widehat{\alpha}^1, \widehat{\gamma}^1\right)} \frac{1}{\sum_i n_{i,ar}} \sum_{i=1}^{K} \sum_{j=1}^{n_{i,ar}} \frac{\widehat{\boldsymbol{g}}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}) P_\alpha(A_i)}{1 + \left\{ \frac{\lambda_1 + 1}{\theta^1(\widehat{\alpha}^1, \widehat{\gamma}^1)}, \frac{\lambda_2}{\theta^1(\widehat{\alpha}^1, \widehat{\gamma}^1)}, \ldots, \frac{\lambda_{J+K}}{\theta^1(\widehat{\alpha}^1, \widehat{\gamma}^1)} \right\} \widehat{\boldsymbol{g}_{ij}}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}) P_\alpha(A_i)}.
\end{aligned}
$$

Therefore, we have the following representation of the estimated weights $\hat{w}^{ar}_{ij}$:

$$
\frac{1}{\sum_i n_{i,ar}} \frac{\theta^1\left(\widehat{\boldsymbol{\alpha}}^1, \widehat{\boldsymbol{\gamma}}^1\right) / \pi^1_{i,j}\left(\widehat{\alpha}^1, \widehat{\gamma}^1\right)}{1 + \widehat{\boldsymbol{\lambda}}^{\mathrm{T}} \widehat{\boldsymbol{g}}_{ij}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}) / \pi^1_{i,j}\left(\widehat{\alpha}^1, \widehat{\gamma}^1\right)}, \tag{4.21}
$$

and the consistency of $\mu^{\mathrm{mr}}_{a\alpha}$ follows:

$$
\begin{aligned}
& \sum_{i=1}^{K} \sum_{j=1}^{n_i} 1(A_{ij} = a) 1(R_{ij} = 1) \hat{w}^{ar}_{ij} Y_{ij}(A_i) \\
= \; & \sum_{i=1}^{K} \frac{\theta^1\left(\widehat{\boldsymbol{\alpha}}^1, \widehat{\boldsymbol{\gamma}}^1\right)}{\sum_i n_{i,ar}} \sum_{j=1}^{n_i} \frac{1(A_{ij} = a) 1(R_{ij} = 1) P_\alpha(A_i) / \pi^1_{i,j}\left(\widehat{\alpha}^1, \widehat{\gamma}^1\right)}{1 + \widehat{\lambda}^{\mathrm{T}} \widehat{\boldsymbol{g}}_{ij}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}) P_\alpha(A_i) / \pi^1_{i,j}\left(\widehat{\alpha}^1, \widehat{\gamma}^1\right)} Y_{ij}(A_i) \\
\xrightarrow{p} \; & E\left[\frac{1}{n_i} \sum_{j=1}^{n_i} \frac{1(A_{ij} = a) 1(R_{ij} = 1) Y_{ij}(A_i) P_\alpha(A_i)}{\pi^1_{i,j}(\widehat{\alpha}^1, \widehat{\gamma}^1)} \right] \\
= \; & \mu^0_{a\alpha}.
\end{aligned}
$$

Second, if $f^{(1)}(a,x;\beta)$,and $\eta^{(1)}(r,x;\beta')$ are correctly specified, we have $\hat{\beta}^1 \xrightarrow{p} \beta^1_\star$, $\widehat{\beta'}^1 \xrightarrow{p} {\beta'}^1_\star$ as $K \to \infty$, $\beta^1_\star = \beta_\star$ and ${\beta'}^1_\star = {\beta'}_\star$. The consistency of $\hat{\mu}^{\mathrm{mr}}_{a\alpha}$ follows:

$$
\begin{aligned}
& \sum_{i=1}^{K}\sum_{j=1}^{n_i} 1(A_{ij}=a)1(R_{ij}=1)\hat{w}^{ar}_{ij}Y_{ij}(A_i) \\
= \;& \sum_{i=1}^{K}\sum_{j=1}^{n_i} 1(A_{ij}=a)1(R_{ij}=1)\hat{w}^{ar}_{ij}\{Y_{ij}(A_i)-m^1_{ij}(A_i,X_{ij};\hat{\beta}^1)\} \\
& +\sum_{i=1}^{K}\sum_{j=1}^{n_i} 1(A_{ij}=a)1(R_{ij}=1)w^{ar}_{ij}m^1_{ij}(a,a_{i(-j)},X_{ij};\hat{\beta}^1) \\
= \;& \sum_{i=1}^{K}\sum_{j=1}^{n_i} 1(A_{ij}=a)1(R_{ij}=1)\hat{w}^{ar}_{ij}\{Y_{ij}(A_i)-m^1_{ij}(A_i,X_{ij};\hat{\beta}^1)\} \\
& +\frac{1}{K}\sum_{i=1}^{K}\frac{1}{n_i}\sum_{j=1}^{n_i}\sum_{a_{i(-j)}}\Bigg[1(R_{ij}=1)m^1_{ij}(a,a_{i(-j)},X_{ij};\hat{\beta}^1) \\
& +1(R_{ij}=0)E\Bigg\{m^1_{ij}(a,a_{i(-j)},X_{ij};\hat{\beta}^1)\Bigg|A_{i(-j)}=a_{i(-j)},X_{ij},R_{ij}=0;\hat{\beta}^{(1)},\widehat{\beta'}^1\Bigg\}\Bigg]P_\alpha(a_{i(-j)}) \\
= \;& \sum_{i=1}^{K}\sum_{j=1}^{n_i} 1(A_{ij}=a)1(R_{ij}=1)\hat{w}^{a}_{ij}\{Y_{ij}(A_i)-m^1_{ij}(A_i,X_{ij};\hat{\beta}^1)\} \\
& +E\Bigg[\frac{1}{n_i}\sum_{j=1}^{n_i}\sum_{a_{i(-j)}} m^1_{ij}(a,a_{i(-j)},X_{ij};\boldsymbol{\beta}_\star)P_\alpha(a_{i(-j)})\Bigg]+o_p(1) \\
= \;& \frac{1}{\sum_i n_{i,a}}\sum_{i=1}^{K}\sum_{j=1}^{n_i}\frac{1(A_{ij}=a)1(R_{ij}=1)\{E^1(Y_{ij}(a_i)|A_i=a_i,X_{ij})-m^1_{ij}(A_i,X_{ij};\boldsymbol{\beta}_\star **966)\}}{1+\boldsymbol{\lambda}^T_\star\boldsymbol{g}_{ij}(\boldsymbol{\alpha}_\star,\boldsymbol{\beta}_\star,\boldsymbol{\gamma}_\star)} \\
& +E\Bigg[\frac{1}{n_i}\sum_{j=1}^{n_i}\sum_{a_{i(-j)}} m^1_{ij}(a,a_{i(-j)},X_{ij};\boldsymbol{\beta}_\star)P_\alpha(a_{i(-j)})\Bigg]+o_p(1) \\
\xrightarrow{p} \;& \frac{1}{\Pr(A=a)}E\Bigg[\frac{1(A_{ij}=a)1(R_{ij}=1)\{E^1(Y_{ij}(a_i)|A_i=a_i,X_{ij})-m^1_{ij}(A_i,X_{ij};\boldsymbol{\beta}_\star)\}}{1+\boldsymbol{\lambda}^T_\star\boldsymbol{g}_{ij}(\boldsymbol{\alpha}_\star,\boldsymbol{\beta}_\star,\boldsymbol{\gamma}_\star)}\Bigg] \\
& +E\Bigg[\frac{1}{n_i}\sum_{j=1}^{n_i}\sum_{a_{i(-j)}} m^1_{ij}(a,a_{i(-j)},X_{ij};\boldsymbol{\beta}_\star)P_\alpha(a_{i(-j)})\Bigg] \\
= \;& \frac{1}{\Pr(A=a)}E\left[\frac{1}{n_i}\sum_{j=1}^{n_i}\frac{1(A_{ij}=a)1(R_{ij}=1)\{E^1(Y_{ij}(a_i)|A_i=a_i,X_{ij})-m^1_{ij}(A_i,X_{ij};\boldsymbol{\beta}_\star)\}}{1+\boldsymbol{\lambda}^T_\star\boldsymbol{g}_{ij}(\boldsymbol{\alpha}_\star,\boldsymbol{\beta}_\star,\boldsymbol{\gamma}_\star)}\right]+\mu^0_{a\alpha} \\
= \;& \mu^0_{a\alpha}.
\end{aligned}
$$

Chapter 5

Discussion and Future Work

5.1 Discussion

In this thesis, we proposed different approaches for handling missingness in correlated data. As mentioned in Chapter 1, conventional regression methods may not be suitable for correlated data, even if the data are complete. When data are subject to missingness, simply ignoring the missing values may result in biased estimators when the missingness mechanism is not MCAR, which is often the case in correlated data. To address these issues, in Chapter 2, we discussed how to handle missingness in longitudinal data, and in Chapters 3 and 4, we investigated how to deal with confounders missing not at random with the goal of estimating causal effects under partial interference.

In the first project, in Chapter 2, we aim to impute missing longitudinal outcomes. First, we proposed a longitudinal low-rank model that utilizes both unit-specific and time-specific covariates. Second, based on the conventional matrix completion algorithms and the LASSO algorithm, a two-step estimation procedure was proposed to solve the optimization problem. In theoretical analysis, it was shown that including the unit-specific and time-specific covariates is beneficial for improving the imputation accuracy. In simulation studies, it was further illustrated that the proposed low-rank longitudinal model is better than both the conventional matrix completion and multiple imputation method. Finally, we applied the proposed method on both Covid-19 and SO_2 emissions datasets.

In the second project, in Chapter 3, we focused on drawing causal effects from incomplete network data, where we consider that the confounders are subject to nonignorable missingness and the outcome is fully observed. We proposed three sets of estimators: IPW, regression, and doubly robust estimators. The IPW and regression estimators are consistent and asymptotically normal if the corresponding working model is correctly specified. The doubly robust estimator is consistent if either the propensity score models or the distribution of observed data is correctly specified. We applied the proposed estimators on the NO_x emissions dataset and presented the estimated causal effects based on the different rates of treatment coverage.

The third project, covering Chapter 4, is an extension of the second project, where we proposed multiply robust estimators by constructing the constraints based on a new estimation procedure. The performance of the proposed estimators is illustrated via simulation studies under different combinations of working models, cluster sizes, and the missingness rate. In addition, the method was applied to the NO_x emissions data analyzed in Chapter 3.

Section 5.2 below is an extension of Chapter 2, where we investigate the construction of confidence intervals for both the low-rank term and the main effects as well as the interaction effects among the covariates. In Section 5.3, we discuss a few potential extensions of Chapters 3 and 4 in the general interference setting.

5.2 Future Work on Matrix Completion

5.2.1 Introduction

In Chapter 2, we showed the non-asymptotic error bounds for the estimated main effects, interaction effects, the low-rank term, and the whole imputed outcome matrix. The predefined constants in the error bounds may result in a gap between the proposed theoretical guarantee and the optimal guarantee. Besides, it is also important to quantify the statistical variation of the estimated matrix to provide more confidence in the estimates.

In recent years, there is an increasing interest in deriving the confidence interval for the imputed low-rank matrix estimated by noisy matrix completion algorithms. Xia [2019] introduced a two-step procedure to construct confidence regions of the singular subspace via double-sample splitting. Chen et al. [2019] constructed de-biased estimators for the low-rank factors, under some mild conditions of sample size and noise level, they showed that the proposed de-shrunken estimators can be decomposed as the summation of a residual matrix and a matrix that follows Gaussian distribution. They further presented the distribution of the estimator for the original low-rank matrix by combining the estimators of these two low-rank factors. Xia and Yuan [2019] proposed a double-sample debiasing and spectral projection procedure to produce an unbiased and asymptotically normal estimator from the original estimator for the linear forms of large matrices. They also presented the confidence interval and hypothesis testing procedure for the linear forms. Cai et al. [2020] proposed an inferential procedure that can adapt automatically to unknown noise distributions.

In most of the aforementioned work of constructing the confidence intervals for the low-rank matrices, very limited efforts have been devoted to constructing the coefficient matrices where the covariate information is included in the low-rank models, let alone the cases where both the unit- and time-specific covariate are included in the model. One way to construct the confidence region is to decompose the low-rank term into two low-rank factors, derive the distribution of the estimators of the two low-rank factors, and then try to derive the distributions of the original low-rank term based on the distribution of these two terms. Finally, the confidence region for the main and interaction effects can be constructed in a similar way, where a de-biasing procedure should be applied. In Section 5.2.2, we will first briefly recall some notations. In Section 5.2.3, we will show the steps of the estimation algorithm as well as the final debiased estimators for the low-rank term and the fixed effects.

5.2.2 Notation

Following Chen et al. [2019], assume the true outcome matrix, the residual matrix, and the observed outcome matrix are denoted by $Y^\star$, E, and Y, respectively, where the entries of $Y^\star$ may be subject to missingness, and the residual matrix includes random noise and serial correlation. The rank of the true outcome matrix is assumed to be r, and the singular value decomposition of Y is given by $Y^\star = U^\star \Sigma^\star V^{\star T}$. Let $A^\star = U^\star \Sigma^{\star 1/2}$ and $B^\star = V^\star \Sigma^{\star 1/2}$. Let P be the probability matrix with each entry presenting the true probabilities of observations of the corresponding entry in the outcome matrix. The definitions of the matrix norms are the same as those introduced in Section 2.2.

5.2.3 Algorithm

Instead of using the convex relaxation of the low-rank penalty term in Equation 5.1, we decompose the low-rank term L into two low-rank factors A and B. Then, the objective function has the following representation:

$$\arg\min_{H,L} E\left\{\frac{R}{\Pr(R=1|X,Z)}||Y-(XHZ^T+AB^T)||_F^2\right\}+\lambda_H||H||_1+\frac{\lambda_L}{2}(||A||_F^2+||B||_F^2).$$

Assume $\hat{H}$, $\hat{A}$, and $\hat{B}$ are estimators for the true fixed effects matrix $H^\star$ and the true low rank term $A^\star$ and $B^\star$ estimated via any nonconvex algorithms (e.g., Zhao et al. [2015]), respectively. Then the debiased estimators H^d, A^d, B^d and L^d can be represented in the following way:

$$vec(H^d) = vec(\hat{H}) + \frac{1}{NT}Mvec(Z\otimes X)^T\{vec(Y) - vec(Z\otimes X)vec(\hat{H}) - vec(L)\}, \quad (5.1)$$

$$A^d = U(\Sigma + M\mathcal{I}_r)^{\frac{1}{2}}, \quad (5.2)$$

$$B^d = V(\Sigma + M\mathcal{I}_r)^{\frac{1}{2}}, \quad (5.3)$$

$$L^d = \hat{L} + \frac{1}{NT}M\{Y - XHZ^T - \hat{L}\} \quad or \quad L^d = A^d B^{d1/2}, \quad (5.4)$$

where M is some unknown matrix that needs proper construction. The basic intuition comes from the *debiased lasso*, where M is a good approximation of the precision matrix $E\{(Z\otimes X)(Z\otimes X)^T\}$ (e.g. see Javanmard and Montanari [2014] for more details). To derive the confidence intervals for the low-rank term L, we first have the following theorem to infer the distribution of the estimators of two low-rank factors A and B.

Theorem 5.1. *The errors of the debiased estimators A^d and B^d have the following decomposition:*

$$A^dH^d - A^\star = \mathcal{Z}_\mathcal{A} + \Psi_A,$$

and

$$B^dH^d - B^\star = \mathcal{Z}_\mathcal{B} + \Psi_B,$$

where rows of $\mathcal{Z}_\mathcal{A}$ and $\mathcal{Z}_\mathcal{B}$ are independent and follow a normal distribution with zero mean and finite variance, and H^d is the rotation matrix that satisfies the following condition:

$$H^d \in \arg\min_{H^d\in R^{r\times r}} \|A^dH^d - A^\star\|_F^2 + \|B^dH^d - B^\star\|_F^2.$$

The residual term Ψ_A and Ψ_B are $o(c^\star)$, where $c^\star$ is some constant relevant to the probability matrix and the correlation structure of residual term E.

With the above theorem, $\hat{L} - L^\star$ can be represented by Z_A, Z_B, $A^\star$ and $B^\star$, it can be shown $\hat{L} - L^\star$ also follows the normal distribution with zero mean and finite variance, and the variance of $\hat{L} - L^\star$ can be derived accordingly.

5.3 Future Work on Network Causal Effects

In Chapters 3 and 4, we discuss the robust estimation of network causal effects under the partial interference setting. However, for some real applications, it may be more reasonable to assume the general interference setting, where the interference can take place between any pair of units in the population. For example, for social network data, each person in the network can not only be affected by his/her own friends, but also by the friends of friends. In such a case, it may not be appropriate to assume that the population can be grouped into disjoint clusters, and the group-level propensity scores are not directly applicable to the estimation of causal effects without any adjustment.

Notice that in Chapters 3 and 4, the effect of the neighborhood treatment information is included in the outcome regression models through a function that maps the treatment vector to the proportion of treated units in the same cluster. The summary function is based on domain knowledge and can lead to biased estimators if it is incorrectly specified (Sävje [2021]). In the setting of general interference, the structure of the network can be complicated, and it can be difficult to correctly capture all the neighborhood information if using such a mapping function because the effect of each neighbor on the unit can be different. Thus, it is interesting to further explore a more robust way to include neighborhood information in the general network setting. For example, one may consider assigning a weight to each neighborhood of units, where the values of those weights are dependent on the social or geographical distances between the unit and each of its neighbors.

Moreover, even in the partial interference setting, the parametric modeling and assumption can be sophisticated as the number of units in each cluster becomes large. The covariates in the regression models and the propensity score models can be high-dimensional because the outcome and the treatment assignment of each unit may be dependent on its neighbors' covariates. Therefore, a potential research topic would be using machine learning or deep learning methods to assist in the estimation of the propensity score and the regression models. It is also of interest to further construct the doubly robust and multiple robust estimators based on the estimated models, and explore the semiparametric efficiency and the convergence rate of those estimators.

References

Paul S Albert. A transitional model for longitudinal binary data subject to nonignorable missing data. *Biometrics*, 56(2):602–608, 2000. 2

Tom M Apostol. Mathematical analysis. *Addison-Wesley, Boston*, 1974. 21

Susan Athey, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. Matrix completion methods for causal panel data models. Technical report, National Bureau of Economic Research, 2018. 14, 18, 20, 22

Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005. 9, 90

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. 76

Steven Berry, James Levinsohn, and Ariel Pakes. Differentiated products demand systems from a combination of micro and macro data: The new car market. *Journal of Political Economy*, 112(1):68–105, 2004. 13

Dimitri P Bertsekas. Nonlinear programming. *Journal of the Operational Research Society*, 48(3):334–334, 1997. 19

Rohit Bhattacharya, Daniel Malinsky, and Ilya Shpitser. Causal inference under interference and network uncertainty. In *Uncertainty in Artificial Intelligence*, pages 1028–1038. PMLR, 2020. 11

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018. 18

S van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical software*, 45(3):1–68, 2010. 12, 28

Changxiao Cai, H Vincent Poor, and Yuxin Chen. Uncertainty quantification for nonconvex tensor completion: Confidence intervals, heteroscedasticity and optimality. In *International Conference on Machine Learning*, pages 1271–1282. PMLR, 2020. 114

Emmanuel J Candes and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010. 8

Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009. 6, 7

Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010. 7

Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011. 15

Yang Cao and Yao Xie. Poisson matrix completion. *2015 IEEE International Symposium on Information Theory (ISIT)*, 64(6):1841–1845, 2015. 32

Jiahua Chen, Pengfei Li, Yukun Liu, and James V Zidek. Composite empirical likelihood for multisample clustered data. *Journal of Nonparametric Statistics*, 33(1):60–81, 2021. 99

Yun Hua Chen. A semiparametric odds ratio model for measuring association. *Biometrics*, 63(2):413–421, 2007. 68, 75

Yuxin Chen, Jianqing Fan, Cong Ma, and Yuling Yan. Inference and uncertainty quantification for noisy matrix completion. *Proceedings of the National Academy of Sciences*, 116(46):22931–22937, 2019. 39, 113, 114

Michael J Conroy, Juan Carlos Senar, and Jordi Domènech. Analysis of individual-and time-specific covariate effects on survival of serinus serinus in north-eastern spain. *Journal of applied Statistics*, 29(1-4):125–142, 2002. 13

D.R. Cox. *Planning of Experiments.* Wiley, 1958. 11

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977. 12

Peter Diggle, Peter J Diggle, Patrick Heagerty, Kung-Yee Liang, Patrick J Heagerty, Scott Zeger, et al. *Analysis of longitudinal data*. Oxford University Press, 2002. 1

Peng Ding and Zhi Geng. Identifiability of subgroup causal effects in randomized experiments with nonignorable missing covariates. *Statistics in Medicine*, 33(7):1121–1133, 2014. 66

Iris Eekhout, Craig K Enders, Jos WR Twisk, Michiel R de Boer, Henrica CW de Vet, and Martijn W Heymans. Analyzing incomplete item scores in longitudinal data by including item score information as auxiliary variables. *Structural Equation Modeling: A Multidisciplinary Journal*, 22(4):588–602, 2015. 12

Paul Erdős and Alfréd Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60, 1960. 4

Garrett M Fitzmaurice, Nan M Laird, and James H Ware. *Applied longitudinal analysis*, volume 998. John Wiley & Sons, 2012. 12

Andrew Giffin, Brian Reich, Shu Yang, and Ana Rappold. Generalized propensity score approach to causal inference with spatial interference. *arXiv preprint arXiv:2007.00106*, 2020. 11

Nima Hamidi and Mohsen Bayati. On low-rank trace regression under general sampling distribution. *arXiv preprint arXiv:1904.08576*, 2019. 20

Peisong Han. A further study of the multiply robust estimator in missing data analysis. *Journal of Statistical Planning and Inference*, 148:101–110, 2014a. 10, 91, 98, 107

Peisong Han. Multiply robust estimation in regression analysis with missing data. *Journal of the American Statistical Association*, 109(507):1159–1173, 2014b. 10, 91, 107

Peisong Han. Calibration and multiple robustness when data are missing not at random. *Statistica Sinica*, 28(4):1725–1740, 2018. 10

Peisong Han and Lu Wang. Estimation with missing data: beyond double robustness. *Biometrika*, 100(2):417–430, 2013. 10, 90, 91, 94

Xingjie Hao, Shanshan Cheng, Degang Wu, Tangchun Wu, Xihong Lin, and Chaolong Wang. Reconstruction of the full transmission dynamics of covid-19 in wuhan. *Nature*, 584(7821):420–424, 2020. 31

Trevor Hastie and Junyang Qian. Glmnet vignette. *Retrieved June*, 9(2016):1–30, 2014. 18

Candise L. Henry and Lincoln F. Pratson. Differentiating the effects of climate change-induced temperature and streamflow changes on the vulnerability of once-through thermoelectric power plants. *Environmental Science & Technology*, 53(7):3969–3976, 2019. 37

Peter D Hoff. Multiplicative latent factor models for description and prediction of social networks. *Computational and Mathematical Organization Theory*, 15(4):261, 2009. 4

Peter D Hoff, Adrian E Raftery, and Mark S Handcock. Latent space approaches to social network analysis. *Journal of the american Statistical association*, 97(460):1090–1098, 2002. 4

Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983. 4

Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47 (260):663–685, 1952. 16

M.G. Hudgens and M.E. Halloran. Toward causal inference with interference. *Journal of the American Statistical Association*, 103:832–842, 2008. 11, 65

Mark Huisman and Christian Steglich. Treatment of non-response in longitudinal network studies. *Social Networks*, 30(4):297–308, 2008. 5

Andrew Hyland, Bridget K Ambrose, Kevin P Conway, Nicolette Borek, Elizabeth Lambert, Charles Carusi, Kristie Taylor, Scott Crosse, Geoffrey T Fong, K Michael Cummings, et al. Design and methods of the population assessment of tobacco and health (path) study. *Tobacco Control*, 26(4):371–378, 2017. 11

Joseph G Ibrahim. Incomplete data in generalized linear models. *Journal of the American Statistical Association*, 85(411):765–769, 1990. 12

Kosuke Imai, Zhichao Jiang, and Anup Malani. Causal inference with interference and noncompliance in two-stage randomized experiments. *Journal of the American Statistical Association*, 116(534):632–644, 2021. 11

SA Imtiaz and SL Shah. Treatment of missing values in process data analysis. *The Canadian Journal of Chemical Engineering*, 86(5):838–858, 2008. 37

Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014. 115

Joseph DY Kang and Joseph L Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 22(4):523–539, 2007. 9, 84

Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE transactions on information theory*, 56(6):2980–2998, 2010. 9

Olga Klopp, Karim Lounici, and Alexandre B Tsybakov. Robust matrix completion. *Probability Theory and Related Fields*, 169(1-2):523–564, 2017. 20

Olga Klopp et al. Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, 20(1):282–303, 2014. 20, 22, 48

Eric D Kolaczyk and Gábor Csárdi. *Statistical analysis of network data with R*, volume 65. Springer, 2014. 4

Vladimir Koltchinskii, Karim Lounici, Alexandre B Tsybakov, et al. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011. 28, 48

Nan M Laird and James H Ware. Random-effects models for longitudinal data. *Biometrics*, pages 963–974, 1982. 2

Wei Li, Yuwen Gu, and Lan Liu. Demystifying a class of multiply robust estimators. *Biometrika*, 107(4):919–933, 2020a. 10, 91

Wei Li, Shu Yang, and Peisong Han. Robust estimation for moment condition models with data missing not at random. *Journal of Statistical Planning and Inference*, 207:246–254, 2020b. 96

Roderick JA Little. A test of missing completely at random for multivariate data with missing values. *Journal of the American statistical Association*, 83(404):1198–1202, 1988. 5

Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019. 15

Lan Liu, Michael G Hudgens, Bradley Saul, John D Clemens, Mohammad Ali, and Michael E Emch. Doubly robust estimation in observational studies with partial interference. *Stat*, 8(1):e214, 2019. 11

Jared K Lunceford and Marie Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine*, 23(19):2937–2960, 2004. 9

Guoguang Ma, Andrea B. Troxel, and Daniel F. Heitjan. *Statistics in Medicine*, 24:2129–50, 2005. 6

Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research*, 11: 2287–2322, 2010. 9

Bilal Mehmood, Zahid Irshad Younas, and Nisar Ahmed. Macroeconomic and bank specific covariates of non-performing loans (npls) in pakistani commercial banks: Panel data evidence. *Journal of Emerging Economies and Islamic Research*, 1(3):34–48, 2013. 13

Wang Miao and Eric J Tchetgen Tchetgen. On varieties of doubly robust estimators under missingness not at random with a shadow variable. *Biometrika*, 103(2):475–482, 2016. 90

Geert Molenberghs and Geert Verbeke. *Linear mixed models for longitudinal data*. Springer, 2000. 6

Sahand Negahban and Martin J Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research*, 13(1):1665–1697, 2012. 22

Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018. 51

Yang Ning, Peng Sida, and Kosuke Imai. Robust estimation of causal effects via a high-dimensional covariate balancing propensity score. *Biometrika*, 107(3):533–554, 2020. 90

Georgia Ntani, Hazel Inskip, Clive Osmond, and David Coggon. Consequences of ignoring clustering in linear regression. *BMC Medical Research Methodology*, 21(1):1–13, 2021. 4

Elizabeth L Ogburn, Ilya Shpitser, and Youjin Lee. Causal inference, social networks and chain graphs. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(4):1659–1676, 2020. 66

Georgia Papadogeorgou, Fabrizia Mealli, and Corwin M Zigler. Causal inference with interfering units for cluster and population level treatment allocation programs. *Biometrics*, 75(3):778–787, 2019. 11, 65, 81

Carolina Perez-Heydrich, Michael G Hudgens, M Elizabeth Halloran, John D Clemens, Mohammad Ali, and Michael E Emch. Assessing effects of cholera vaccination in the presence of interference. *Biometrics*, 70(3):731–741, 2014. 64

Annie Qu and Peter X-K Song. Testing ignorable missingness in estimating equation approaches for longitudinal data. *Biometrika*, 89(4):841–850, 2002. 5

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019. URL https://www.R-project.org/. 3, 76

Geneviève Robin, Julie Josse, Éric Moulines, and Sylvain Sardy. Low-rank model with covariates for count data with missing values. *Journal of Multivariate Analysis*, 173: 416–434, 2019. 20

Geneviève Robin, Olga Klopp, Julie Josse, Éric Moulines, and Robert Tibshirani. Main effects and interactions in mixed and incomplete data frames. *Journal of the American Statistical Association*, 115(531):1292–1303, 2020. 14

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994. 9, 12

James M Robins, Andrea Rotnitzky, and Mark van der Laan. On profile likelihood: comment. *Journal of the American Statistical Association*, 95(450):477–482, 2000. 9

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983. 9

Andrea Rotnitzky, James M Robins, and Daniel O Scharfstein. Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the american statistical association*, 93(444):1321–1339, 1998. 9

Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. 5, 10, 12

Donald B Rubin. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593, 1980. 10

Fredrik Sävje. Causal inference with misspecified exposure mappings. *arXiv preprint arXiv:2103.06471*, 2021. 116

Michelle Shardell, Gregory E Hicks, and Luigi Ferrucci. Doubly robust estimation and causal inference in longitudinal studies with dropout and truncation by death. *Biostatistics*, 16(1):155–168, 2015. 90

Zhaohan Sun and Lan Liu. Semiparametric inference of causal effect with nonignorable missing confounders. *Statistica Sinica*, 31:1–20, 2021. 63

Zhiqiang Tan. Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data. *The Annals of Statistics*, 48(2):811–837, 2020. 90

Dingke Tang, Dehan Kong, Wenliang Pan, and Linbo Wang. Ultra-high dimensional variable selection for doubly robust causal inference. *Biometrics*, 2022. 90

Niansheng Tang and Yuanyuan Ju. Statistical inference for nonignorable missing-data problems: a selective review. *Statistical Theory and Related Fields*, 2(2):105–133, 2018. 6

E.J. Tchetgen Tchetgen and T.J. VanderWeele. On causal inference in the presence of interference. *Statistical Methods in Medical Research*, 21:55–75, 2012. 11, 64

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996. 104

Paul Tseng and Sangwoon Yun. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, 117(1-2):387–423, 2009. 21, 39, 40, 42

Dong Xia. Confidence region of singular subspaces for low-rank matrix regression. *IEEE Transactions on Information Theory*, 65(11):7437–7459, 2019. 113

Dong Xia and Ming Yuan. Statistical inferences of linear forms for noisy matrix completion. *arXiv preprint arXiv:1909.00116*, 2019. 113

Shu Yang, Linbo Wang, and Peng Ding. Causal inference with confounders missing not at random. *Biometrika*, 106(4):875–888, 2019. 10, 66

Sangwoon Yun and Kim-Chuan Toh. A coordinate gradient descent method for l1-regularized convex minimization. *Computational Optimization and Applications*, 48(2): 273–307, 2011. 18

Shixiao Zhang, Peisong Han, and Changbao Wu. Empirical likelihood inference for non-randomized pretest-posttest studies with missing data. *Electronic Journal of Statistics*, 13(1):2012–2042, 2019. 91

L. Zhao, S. Lipsitz, and D. Lew. Regression analysis with missing covariate data using estimating equations. *Biometrics*, 52:1165–1182, 1996. 12

Tuo Zhao, Zhaoran Wang, and Han Liu. A nonconvex optimization framework for low rank matrix estimation. *Advances in Neural Information Processing Systems*, 28:559–567, 2015. 115

Corwin Matthew Zigler, Chanmin Kim, Christine Choirat, John Barrett Hansen, Yun Wang, Lauren Hund, Jonathan Samet, Gary King, and Francesca Dominici. Causal inference methods for estimating long-term health effects of air quality regulations. Technical Report 187, 2016. 11, 36, 80