Inference for Low-rank Completion without Sample Splitting with Application to Treatment Effect Estimation

Jungjun Choi
Department of Economics, Rutgers University
Hyukjun Kwon
Department of Economics, Rutgers University
Yuan Liao
Department of Economics, Rutgers University

Saturday 26th February, 2022

Abstract

This paper studies the inferential theory for estimating low-rank matrices. It also provides an inference method for the average treatment effect as an application. We show that the least square estimation of eigenvectors following the nuclear norm penalization attains the asymptotic normality. The key contribution of our method is that it does not require sample splitting. In addition, this paper allows dependent observation patterns and heterogeneous observation probabilities. Empirically, we apply the proposed procedure to estimating the impact of the presidential vote on allocating the U.S. federal budget to the states.

Keywords: Matrix completion; Nuclear norm penalization; Two-step least squares estimation; Leave-one-out method; Approximate factor model; Causal inference

1 Introduction

The task of imputing the missing entries of a partially observed matrix, often dubbed as *matrix completion*, is widely applicable in various areas. In addition to the well-known application to recommendation systems (e.g., the Netflix problem), this problem is applied in a diverse array of science and engineering such as collaborative filtering, system identification, social networks recovery and causal inference.

One of the common assumptions for identification is that the matrix is of (approximately) low-rank compared to its dimension. In this paper, we focus on the following approximate low-rank model with a strong factor structure:

$$Y = M + \mathcal{E} \approx \beta F' + \mathcal{E}, \tag{1.1}$$

where Y is an $N \times T$ data matrix which is subject to missing, M is a latent matrix of interest, and $\mathcal{E}$ represents a noise contamination. Importantly, M is assumed to be an approximate low-rank matrix having an approximate factor structure $M \approx \beta F'$, where β is factor loadings and F is latent factors. In addition, we allow some entries of Y to be unobserved by defining an indicator ω_{it}, which equals one if the (i,t) element of Y is observed, and zero otherwise. In this practical setting, we provide the inferential theory for each entry of M, regardless of whether its corresponding entry in Y is observed or not.

One of the widely used methods for the low-rank matrix completion is the nuclear norm penalization and it has been intensively studied in the last decade. Candès and Recht (2009), Candes and Plan (2010), Koltchinskii et al. (2011), Negahban and Wainwright (2012), and Chen et al. (2020b) provide statistical rates of convergence for the nuclear norm penalized estimator and a branch of studies including Beck and Teboulle (2009), Cai et al. (2010), Mazumder et al. (2010), Ma et al. (2011), and Parikh and Boyd (2014) provide algorithms to compute the nuclear norm penalized estimator. However, research

on inference is still limited. This is because the shrinkage bias caused by the penalization, as well as the lack of the closed-form expression of the estimator, hinders the distributional characterization of the estimator.

Very recently, some studies proposed the ways of achieving unbiased estimation for the inference of the nuclear norm penalized estimator.[1] Chen et al. (2019) and Xia and Yuan (2021) explicitly subtract the estimator of the bias from the initial estimator and exploit the projection method to control the rank of the debiased estimator for reducing the variance. Chernozhukov et al. (2019, 2021) propose a two-step least square procedure with sample splitting, which estimates the factors and loadings successively using the least square estimations. However, this method exploits sample splitting which has several costs. First, sample splitting restricts the shape of groups when we conduct inference about the group averages.[2] Second, it dramatically inflates computational costs in multiple tests. Third, sample splitting may degrade the estimation quality when the sample size is not large enough. In addition, we may have different estimates for the same target parameter depending on how the sample is split.

We contribute to the literature by providing an inferential theory of the low-rank estimation without sample splitting. Our estimation procedure consists of the following main steps:

1. Using the full sample of observed Y, compute the nuclear norm penalized estimator $\widetilde{M}$ and use the left singular vectors of $\widetilde{M}$ as the initial estimator for β.
2. To estimate F, regress the observed Y onto the initial estimator for β.
3. To re-estimate β, regress the observed Y on the estimator for F.
4. The product of the estimators in Steps 2 and 3 is the final estimator for M.

Since we skip sample splitting and simply use the full (observed) sample in every step of our procedure, we need an alternative approach to show the negligibility of the potential

[1] In this paper, the "unbiased estimation" means the estimation having only higher-order biases.

[2] For example, conducting inference about the time average of a certain i, or the cross-sectional average of several periods is not allowed in Chernozhukov et al. (2019, 2021).

bias terms (for which Chernozhukov et al. (2019, 2021) use the sample splitting). We make use of a hypothetical *leave-one-out* estimator. It is an auxiliary estimator, which is to be shown that it is i) asymptotically equivalent to the initial estimator for β in Step 1 and ii) independent of the sample used in the least squares estimation, namely, the sample in period t.[3] Using the leave-one-out estimator, we can separate out the part in the initial estimator for β, which is correlated with the sample in period t. Once we separate out the correlated part, we can enjoy a similar effect to the sample splitting. And we show the separated correlated part is sufficiently small. Importantly, the leave-one-out estimator only appears in the proof as an auxiliary point of the initial estimator for β, so we do not need to compute it in the estimation procedure, which allows us to remove the sample splitting step without implementing any additional steps. That is, only the two-step least squares step is enough to achieve the unbiased estimation for the inference.

The idea of the leave-one-out estimator has been employed in other recent works such as Ma et al. (2019); Chen et al. (2019, 2020a,b) as well. In particular, Chen et al. (2019) pioneered using this idea to convex relaxation of low-rank inference. We highlight that our leave-one-out "auxiliary estimator" is defined differently from theirs. In both Chen et al. (2019, 2020b) and this paper, the leave-one-out estimator is to be (hypothetically) calculated by using the gradient descent iteration from the leave-one-out problem, which rules out, for example, samples in period t. However, the crucial difference is, unlike Chen et al. (2019, 2020b) who derive the stopping point from the problem using the full sample, we define the stopping point from the leave-one-out problem. Note that the gradient descent iteration must stop where the gradient of the loss function becomes sufficiently "small". If this stopping point depends on the sample in period t, the leave-one-out estimator using this stopping point may not be truly independent of the sample in period t. This is crucial otherwise it does not assure the independence between the leave-one-out estimator and

[3] In Step 2, we run the least square regressions for each $t \leq T$. So, we define the hypothetical leave-one-out estimator for each $t \leq T$.

samples of period t. While this change causes some nontrivial difficulties in the proof, we successfully resolve the problems.

Another contribution of this paper is that our inference procedure allows more general data-observation patterns than the one commonly adopted in the matrix completion literature and exploits a weighting method in the objective function to incorporate the heterogeneous observation probability. Furthermore, we accommodate the correlated observation pattern by assuming the cluster structure and allowing dependences within a cluster. Moreover, we provide the inferential theory for the average treatment effect estimator as an application.

Empirically, we apply the proposed procedure to making inference for the impact of the presidential vote on allocating the U.S. federal budget to the states. We find the states that supported the incumbent president in past presidential elections tend to receive more federal funds and this tendency is stronger for the loyal states than the swing states. In addition, this tendency is stronger after the 1980s.

This paper is organized as follows. Section 2 provides the model and the estimation procedure as well as our strategy achieving the unbiased estimation. Section 3 gives the asymptotic results of our estimator. Section 4 provides the inferential theory for the average treatment effect estimator as an application. Section 5 presents an empirical study about the impact of the president on allocating the U.S. federal budget to the states to illustrate the use of our inferential theory. Section 6 includes the simulation studies. Section 7 concludes. All proofs are relegated to the Appendix in the supplement.

There are a few words on our notation. For any matrix A, we use $\|A\|_F$, $\|A\|$, and $\|A\|_*$ to denote the Frobenius norm, operator norm, and nuclear norm respectively. $\|A\|_{2,\infty}$ denotes the largest l_2 norm of all rows of a matrix A. $\text{vec}(A)$ is the vector constructed by stacking the columns of the matrix A in order. For any vector B, $\text{diag}(B)$ is the diagonal matrix whose diagonal entries are B. $a \asymp b$ means a/b and b/a are $O_P(1)$.

2 Model and Estimation

We consider the following nonparametric panel model subject to missing data problem:

$$y_{it} = h_t(\zeta_i) + \varepsilon_{it},$$

where y_{it} is the scalar outcome for a unit i in a period t, $h_t(\cdot)$ is a time-varying nonparametric function, ζ_i is a unit-specific latent state variable, ε_{it} is the noise, and $\omega_{it} = 1\{y_{it} \text{ is observed}\}$.[4] Here, $\{h_t(\cdot), \zeta_i, \varepsilon_{it}\}$ are unobservable. In the model, the (latent) unit states ζ_i have a time-varying effect on the outcome variable through $h_t(\cdot)$. This model can be written in (1.1) using the sieve representation. Suppose the function $h_t(\cdot)$ has the following sieve approximation:

$$h_t(\zeta_i) = \sum_{r=1}^{K} \kappa_{t,r}\phi_r(\zeta_i) + M_{it}^R = \beta_i' F_t + M_{it}^R = M_{it}^\star + M_{it}^R,$$

where $\beta_i = (\phi_1(\zeta_i), \ldots, \phi_K(\zeta_i))'$ and $F_t = (\kappa_{t,1}, \ldots, \kappa_{t,K})'$. Here, M_{it}^R is the sieve approximation error and, for all $1 \leq r \leq K$, $\phi_r(\zeta_i)$ is the sieve transformation of ζ_i using the basis function $\phi_r(\cdot)$ and $\kappa_{t,r}$ is the sieve coefficient. Then, $h_t(\zeta_i) = M_{it}$ can be successfully represented as the approximate factor structure.[5]

In matrix form, we can represent the model as

$$Y = M + \mathcal{E} = M^\star + M^R + \mathcal{E} = \beta F' + M^R + \mathcal{E}, \tag{2.1}$$

where we denote $Y = [y_{it}]_{N\times T}$, $M = [M_{it}]_{N\times T}$, $M^\star = [M_{it}^\star]_{N\times T}$, $M^R = [M_{it}^R]_{N\times T}$, $\beta = [\beta_1, \ldots, \beta_N]'$, $F = [F_1, \ldots, F_T]'$, and $\mathcal{E} = [\varepsilon_{it}]_{N\times T}$. Note that Y and $\mathcal{E}$ are incomplete matrices which have missing components while M is a complete matrix.

Let $\mathcal{M} \coloneqq \{\beta, F, M^R\}$ be the set of random matrices that compose M. In the paper, we allow the heterogeneous observation probability, i.e., $P(\omega_{it} = 1) = p_i$ and denote

[4] Trivially, our theory holds for the model of $y_{it} = h_i(\eta_t) + \varepsilon_{it}$ as well. We omit it for brevity.

[5] Although we consider the nonparametric panel model in the paper, our inferential theory covers other approximate factor models having the form (1.1) also. Please refer to Section A.3 of the supplement.

$\Pi = \text{diag}(p_1, \ldots, p_N)$. Here, we shall assume the sieve dimension K is pre-specified by researchers and propose some data-driven ways of choosing K in Section A.4 of the supplement.

2.1 Nuclear norm penalized estimation with inverse probability weighting

One of the commonly used methods for the low-rank matrix completion is the nuclear norm penalization. Most of the previous works about this method assume the homogeneous observation probability and consider the following convex program:

$$\underset{A \in \mathbb{R}^{N \times T}}{\arg\min} \frac{1}{2} \|\Omega \circ (A - Y)\|_F^2 + \lambda \|A\|_* \tag{2.2}$$

where $\Omega = [\omega_{it}]_{N \times T}$ and $\circ$ denotes the Hadamard product.

However, in the case of the heterogeneous observation probability, using the objective function (2.2) may not be an optimal choice. If we use the objective function (2.2), the estimation errors of individuals with low observation probability would be factored less into minimizing squared errors, and it may debase the estimation quality.

To avoid this problem, this paper utilizes the inverse probability weighting scheme, referred to as inverse propensity scoring (IPS) or inverse probability weighting in causal inference literature (e.g., Imbens and Rubin (2015), Little and Rubin (2019), Schnabel et al. (2016)), in the following way:

$$\widetilde{M} := \underset{A \in \mathbb{R}^{N \times T}}{\arg\min} \frac{1}{2} \|\widehat{\Pi}^{-\frac{1}{2}} \Omega \circ (A - Y)\|_F^2 + \lambda \|A\|_* \tag{2.3}$$

where $\widehat{\Pi} = \text{diag}(\widehat{p}_1, \ldots, \widehat{p}_N)$, and $\widehat{p}_i = \frac{1}{T} \sum_{t=1}^{T} \omega_{it}$ for each $i \leq N$. As noted in Ma and Chen (2019), this inverse probability weighting debiases the objective function itself. If there is heterogeneity in the observation probability, $\|\Pi^{-\frac{1}{2}} \Omega \circ (A - Y)\|_F^2$ is an unbiased

estimates of $\|A-Y\|_F^2$, which we would use if there is no missing entry, in the sense that $\mathbb{E}_\Omega[\|\Pi^{-\frac{1}{2}}\Omega\circ(A-Y)\|_F^2] = \|A-Y\|_F^2$, while $\|\Omega\circ(A-Y)\|_F^2$ is biased.[6] Figure A.1 in the supplement shows that using the inverse probability weighting improves the estimation quality of the nuclear norm penalized estimation when units may have different observation probabilities.

2.2 Estimation procedure

Although the inverse probability weighting enhances the estimation quality, the weighting alone cannot guarantee the asymptotic normality of the estimator because of the shrinkage bias. To achieve the unbiased estimation having the asymptotic normality, we run the two-step least squares procedure. As noted previously, our estimation does not have the sample splitting steps. Our estimation algorithm is as follows:

Algorithm 1 Constructing the estimator for M.

Step 1 Compute the initial estimator $\widetilde{M}$ using the nuclear norm penalization.

Step 2 Let $\widetilde{\beta}$ be $N\times K$ matrix whose columns are $\sqrt{N}$ times the top K left singular vectors of $\widetilde{M}$.

Step 3 For each $t\le T$, run OLS to get $\widehat{F}_t = \left(\sum_{j=1}^N \omega_{jt}\widetilde{\beta}_j\widetilde{\beta}_j'\right)^{-1}\sum_{j=1}^N \omega_{jt}\widetilde{\beta}_j y_{jt}$.

Step 4 For each $i\le N$, run OLS to get $\widehat{\beta}_i = \left(\sum_{s=1}^T \omega_{is}\widehat{F}_s\widehat{F}_s'\right)^{-1}\sum_{s=1}^T \omega_{is}\widehat{F}_s y_{is}$.

Step 5 The final estimator $\widehat{M}_{it}$ is $\widehat{\beta}_i'\widehat{F}_t$ for all (i,t).

The nuclear norm penalized estimator $\widetilde{M}$ can be estimated by using many existing algorithms for the nuclear norm penalization in the literature.[7] After deriving the initial estimator of loadings from the nuclear norm penalized estimator $\widetilde{M}$, we estimate latent factors and loadings using the two-step least squares procedure. The final estimator of M is then the product of the estimates for latent factors and loadings.[8]

[6] If there is no heterogeneity in the observation probability, we have $\mathbb{E}_\Omega\left[\|\Omega\circ(A-Y)\|_F^2\right] = p\|A-Y\|_F^2$ and so $p^{-1}\|\Omega\circ(A-Y)\|_F^2$ is an unbiased estimate. For details, please refer to Ma and Chen (2019).

[7] For instance, we use the proximal gradient method (Parikh and Boyd, 2014) in the simulation study.

[8] In fact, Algorithm 1 gives the estimator of $M^\star$. However, because M is well approximated by $M^\star$, the

2.3 A general discussion of the main idea

It is well-known that the nuclear-norm penalized estimator $\widetilde{M}$, like other penalized estimators, is subject to shrinkage bias which complicates statistical inference. To resolve this problem, we use the two-step least squares procedure, i.e., Steps 3 and 4 in Algorithm 1. In showing the asymptotic normality of the resulting estimator $\widehat{M}$, a key challenge is to show the following term is asymptotically negligible:

$$R_t = \frac{1}{\sqrt{N}} \sum_{j=1}^{N} \omega_{jt} \varepsilon_{jt} (\widetilde{\beta}_j - H_1' \beta_j)$$

where H_1 is some rotation matrix.[9] This term represents the effect of the bias of the nuclear-norm penalization since $\widetilde{\beta}_j$ is derived from the nuclear-norm penalized estimator. Chernozhukov et al. (2019, 2021) resort to sample splitting to show the asymptotic negligibility of R_t.

2.3.1 A new auxiliary leave-one-out method

Motivated by Chen et al. (2020b), we show the asymptotic negligibility of R_t without sample splitting by using two hypothetical estimators which are asymptotically equivalent to the nuclear norm penalized estimator $\widetilde{\beta}$. Namely, we consider a hypothetical non-convex iteration procedure for the low-rank regularization, where singular vectors are iteratively solved as the solution and show that this procedure can be formulated as the following two problems:

$$\begin{aligned} L^{full}(B, F) &= \frac{1}{2} \| \Pi^{-\frac{1}{2}} \Omega \circ (BF' - Y) \|_F^2 + \frac{\lambda}{2} \|B\|_F^2 + \frac{\lambda}{2} \|F\|_F^2 \\ &= \frac{1}{2} \| \Pi^{-\frac{1}{2}} \Omega \circ (BF' - Y) \|_{F,(-t)}^2 + \frac{1}{2} \| \Pi^{-\frac{1}{2}} \Omega \circ (BF' - Y) \|_{F,t}^2 + \frac{\lambda}{2} \|B\|_F^2 + \frac{\lambda}{2} \|F\|_F^2 \end{aligned} \tag{2.4}$$

estimator of $M^\star$ works as the estimator of M in the paper. Hence, we regard the estimator from Algorithm 1 as the estimator of M here.

[9] Another term $\frac{1}{\sqrt{N}} \sum_{j=1}^{N} (\omega_{jt} - p_j) \beta_j F_t' H_1'^{-1} (\widetilde{\beta}_j - H_1' \beta_j)$ is also to be shown negligible, but the argument is similar to that of R_t.

$$L^{(-t)}(B,F) = \frac{1}{2}\|\Pi^{-\frac{1}{2}}\Omega \circ (BF' - Y)\|^2_{F,(-t)} + \frac{1}{2}\|BF' - M^\star\|^2_{F,t} + \frac{\lambda}{2}\|B\|^2_F + \frac{\lambda}{2}\|F\|^2_F. \quad (2.5)$$

Here, $\|\cdot\|_{F,(-t)}$ denotes the Frobenius norm which is computed ignoring t-th column and $\|\cdot\|_{F,t}$ is the Frobenius norm of only t-th column. Note that the only difference between (2.4) and (2.5) is that the t-th column of the goodness of fit part in (2.4) is replaced by its conditional expectation in (2.5). So, $\{\omega_{jt}, \varepsilon_{jt}\}_{j\leq N}$ is excluded from the problem (2.5).[10]

Both hypothetical problems should be computed iteratively until the gradients of the non-convex loss functions become "sufficiently small." However, the gradients do not monotonically decrease as iteration proceeds since the problem is non-convex. So, one cannot let it iterate until convergence is reached, but has to stop at the point where the gradient is small enough. Fix t of interest and suppose we iterate both problems τ_t times, where τ_t depends on t. Denote the "solutions" of (2.4) and (2.5) as $\breve{\beta}^{full,t}$ and $\breve{\beta}^{(-t)}$ respectively, where they are the τ_t-th iterates. Hence, they share the same stopping point τ_t. Noticeably, although $\breve{\beta}^{full,t}$ is a solution for the full sample problem (2.4), it depends on t through τ_t.

It is crucial to derive the stopping point τ_t and there is an essential difference in the way of selecting the stopping point between Chen et al. (2019, 2020b) and this paper. Note that, even if one estimator is derived from the leave-one-out problem (2.5), it may not be independent of $\{\omega_{jt}, \varepsilon_{jt}\}_{j\leq N}$, if the stopping point τ_t is derived from the full sample problem (2.4). Therefore, while they derived the stopping point from the full sample problem (2.4), we derive the stopping point from the leave-one-out problem (2.5). Hence, it ensures that the estimator $\breve{\beta}^{(-t)}$ using this stopping point is independent of $\{\omega_{jt}, \varepsilon_{jt}\}_{j\leq N}$. This introduces nontrivial technical challenges. Namely, τ_t, being derived from the problem $L^{(-t)}(B,F)$, depends on t, so the "full-problem" solution $\breve{\beta}^{full,t}$ would therefore also depend on t. We derive the uniform convergence of both $\breve{\beta}^{full,t}$ and $\breve{\beta}^{(-t)}$ uniformly in $t = 1, ..., T$.

Being equipped with these two auxiliary non-convex estimators, we can bound R_t in

[10] Namely, $\{\omega_{jt}, \varepsilon_{jt}\}_{j\leq N}$ are replaced by $\{p_j, 0\}_{j\leq N}$.

the following scheme:

1. First, decompose R_t into two terms:

$$\begin{aligned} R_t &= \frac{1}{\sqrt{N}} \sum_{j=1}^{N} \omega_{jt} \varepsilon_{jt} (\widetilde{\beta}_j - H_1' \beta_j) \\ &= \underbrace{\frac{1}{\sqrt{N}} \sum_{j=1}^{N} \omega_{jt} \varepsilon_{jt} (\widetilde{\beta}_j - \breve{\beta}_j^{(-t)})}_{:=a} + \underbrace{\frac{1}{\sqrt{N}} \sum_{j=1}^{N} \omega_{jt} \varepsilon_{jt} (\breve{\beta}_j^{(-t)} - H_1' \beta_j)}_{:=b}. \end{aligned}$$

2. $\max_t \|b\| = o_P(1)$ can be shown using the independence between $\breve{\beta}^{(-t)}$ and $\{\omega_{jt}\varepsilon_{jt}\}_{j \le N}$, which is along the same line as sample splitting.
3. In addition, $\max_t \|a\| = o_P(1)$ comes from the following two rationales:

   (a) $\breve{\beta}^{full,t} \approx \breve{\beta}^{(-t)}$

   Their loss functions (2.4) and (2.5) are very similar and they share the same stopping point τ_t. Therefore, $\max_t \|\breve{\beta}^{full,t} - \breve{\beta}^{(-t)}\|$ is sufficiently small. Following the guidance of Chen et al. (2020b), we apply the mathematical induction.

   (b) $\widetilde{\beta} \approx \breve{\beta}^{full,t}$

   Note that $\breve{\beta}^{full,t}$ is derived from the non-convex problem (2.4) and $\widetilde{\beta}$ comes from the nuclear norm penalization (2.3). Although the loss functions (2.3) and (2.4) are seemingly distinct, their penalty terms are closely related in the sense that

$$\|A\|_* = \inf_{B \in \mathbb{R}^{N \times K}, F \in \mathbb{R}^{T \times K}: BF' = A} \left\{ \frac{1}{2} \|B\|_F^2 + \frac{1}{2} \|F\|_F^2 \right\}.$$

   Hence, $\max_t \|\widetilde{\beta} - \breve{\beta}^{full,t}\|$ is sufficiently small. A technical innovation is that $\breve{\beta}^{full,t}$ depends on t, so the uniformity is crucially relevant.

In this way, we can successfully show the negligibility of R_t uniformly in t without resorting to sample splitting.

2.3.2 Singular vector estimation is unbiased

From Algorithm 1, we see that there is no explicit debias step. In fact, in terms of estimating the singular vector space, the sigular vector estimator from the least square estimation following the nuclear norm penalization, $\widehat{F}_t$, is unbiased (up to a rotation).

To see this, note that the estimation of F_t has the following maximization problem:

$$\widehat{F}_t \coloneqq \arg\max_{f \in \mathbb{R}^K} Q_t(f, \widetilde{\beta})$$

where $Q_t(f, B) = -\frac{1}{N}\sum_{j=1}^{N} \omega_{jt}(y_{jt} - f'b_j)^2$, $B = (b_1, \ldots, b_N)'$ and b_j are K dimensional vectors. In this step, β is the nuisance parameter and F_t is the parameter of interest. By Taylor expansion, we have, for some invertible matrix A,

$$\begin{aligned}
&\sqrt{N}(\widehat{F}_t - H_1^{-1}F_t) \\
&\quad = -\sqrt{N}A^{-1}\frac{\partial Q_t(H_1^{-1}F_t, \beta H_1)}{\partial f} - \sqrt{N}A^{-1}\frac{\partial^2 Q_t(H_1^{-1}F_t, \beta H_1)}{\partial f \partial \mathrm{vec} b}\mathrm{vec}(\widetilde{\beta} - \beta H_1) + o_P(1).
\end{aligned} \tag{2.6}$$

The first term is the score which leads to the asymptotic normality and the second term represents the effect of the β estimation which is subject to the shrinkage bias. The second term can be decomposed into two parts: one is the terms like R_t which is negligible as we show in the previous section, and another is $\sqrt{N}\varphi H_1^{-1}F_t$ where $\varphi = -A^{-1}H_1'\frac{1}{N}\sum_{j=1}^{N} p_j\beta_j(\widetilde{\beta}_j - H_1'\beta_j)'$. Although the term $\sqrt{N}\varphi H_1^{-1}F_t$ is non-negligible, it has a useful feature, that is, it is on the space of $H_1^{-1}F_t$. Making use of this fact, (2.6) can be re-written as follows:

$$\sqrt{N}(\widehat{F}_t - H_2F_t) = \underbrace{-\sqrt{N}A^{-1}\frac{\partial Q_t(H_1^{-1}F_t, \beta H_1)}{\partial f}}_{\text{asymptotically normal}} + o_P(1)$$

by defining $H_2 \coloneqq (I_K + \varphi)H_1^{-1}$. Note that the non-negligible bias term is absorbed by the rotation matrix H_2, and thus $\widehat{F}_t$ can unbiasedly estimate F_t up to this new rotation. Then, in Step 4 of Algorithm 1, $\widehat{\beta}$, the least square estimator using $\widehat{F}$ as a regressor, can

unbiasedly estimate β_i up to the rotation since $\widehat{F}_t$ has only a higher order bias now. As a result, the product of them estimates M_{it} unbiasedly:

$$\begin{aligned}\widehat{M}_{it} = \widehat{\beta}_i'\widehat{F}_t &= \beta_i' H_2^{-1} H_2 F_t + \text{asymptotically normal term} + \text{higher order terms} \\ &= M_{it} + \text{asymptotically normal term} + \text{higher order terms},\end{aligned}$$

which allow us to conduct inference successfully.

This is how the two-step least squares procedure works. Since it is enough to estimate β and F unbiasedly "up to a rotation", the bias term can be absorbed by the rotation matrix. Hence, without the step which explicitly removes the bias, we can remove the bias using the two-step least square procedure.

3 Asymptotic Results

This section presents the inferential theory. We provide the asymptotic normality of the estimator of the group average of M_{it}. Before proceeding, we present some assumptions for the asymptotic normality of the estimator. Remind the following notation:

$$h_t(\zeta_i) = \sum_{r=1}^{K} \kappa_{t,r}\phi_r(\zeta_i) + M_{it}^R = \beta_i' F_t + M_{it}^R,$$

where $\beta_i = (\phi_1(\zeta_i), \ldots, \phi_K(\zeta_i))'$ and $F_t = (\kappa_{t,1}, \ldots, \kappa_{t,K})'$.

Assumption 3.1 (Sieve representation)**.** *(i) $\{h_t(\cdot)\}_{t\leq T}$ belong to ball $\mathcal{H}\left(\mathcal{Z}, \|\cdot\|_{L_2}, C\right)$ inside a Hilbert space spanned by the basis $\{\phi_r\}_{r\geq 1}$, with a uniform L_2-bound C:*

$$\sup_{h\in\mathcal{H}(\mathcal{Z},\|\cdot\|_{L_2})} \|h\| \leq C,$$

where $\mathcal{Z}$ is the support of ζ_i.

(ii) The sieve approximation error satisfies: For some $\nu > 0$, $\max_{i,t} |M_{it}^R| \leq CK^{-\nu}$.

(iii) For some $C > 0$, with probability converging to 1, $\max_i \frac{1}{K}\sum_{r=1}^{K} \phi_r^2(\zeta_i) \leq C$.

(iv) There is $c > 0$ such that for $\iota \in \{0, 1\}$, with probability converging to 1,

$$\psi_{min}\left(\frac{1}{N}\beta'\beta\right) > c, \quad \psi_{min}\left(\frac{1}{T}F'F\right) > c$$

where $\psi_{min}(\cdot)$ *denotes the smallest nonzero singular value of a matrix.*

(v) $\sum_{i,t} M_{it}^2 = \sum_{i,t} h_t^2(\zeta_i) \asymp NT$.

First, we present some assumptions for the sieve representation. Assumption 3.1 (ii) is well satisfied with a quite large ν if the functions $\{h_t(\cdot)\}$ are sufficiently smooth. For example, consider h_t belonging to a Hölder class: for some $a, b, C > 0$, uniform constants with respect to t,

$$\left\{h : \|D^b h(x_1) - D^b h(x_2)\| \leq C\|x_1 - x_2\|^a\right\},$$

and take usual basis like polynomials, trigonometric polynomials and B-splines, then

$$\max_{i,t} |M_{it}^R| \leq CK^{-\nu}, \qquad \nu = 2(a+b)/\dim(\zeta_i),$$

which can be arbitrary small for smooth functions even if K grows slowly. Assumptions 3.1 (i) and (iii) help us to bound $\max_i \|\beta_i\|$ and $\max_t \|F_t\|$ which are used to show the incoherence condition (Assumption A.1 in the supplement) because $\max_i \|\beta_i\|^2 = \max_i \sum_{r=1}^K \phi_r^2(\zeta_i)$ and

$$\max_t \|F_t\|^2 \leq \max_t \sum_{r=1}^{\infty} \kappa_{t,r}^2 = \max_t \|h_t\|_{L_2}^2 \leq \sup_{h \in \mathcal{H}\left(\mathcal{Z}, \|\cdot\|_{L_2}\right)} \|h\|^2 .$$

Assumptions 3.1 (iii) can be satisfied if the basis is a bounded basis like trigonometric basis or ζ_i has a compact support. In addition, Assumption 3.1 (v) controls the size of $\|M\|_F$ and it can be easily satisfied.

Assumption 3.2 (DGP for ε_{it} and ω_{it})**.** *(i) Conditioning on* $\mathcal{M}$*,* ε_{it} *is i.i.d. zero-mean, sub-gaussian random variable such that* $\mathbb{E}[\varepsilon_{it}|\mathcal{M}] = 0$*,* $\mathbb{E}[\varepsilon_{it}^2|\mathcal{M}] = \sigma^2$*,* $\mathbb{E}[\exp(s\varepsilon_{it})|\mathcal{M}] \leq \exp(Cs^2\sigma^2)$*,* $\forall s \in \mathbb{R}$*, for some constant* $C > 0$*.*

(ii) Ω *is independent of* $\mathcal{E}$*. Conditioning on* $\mathcal{M}$*,* ω_{it} *is independent across* t*. In addition,* $\mathbb{E}[\omega_{it}|\mathcal{M}] = \mathbb{E}[\omega_{it}] = p_i$ *and there is a constant* $\underline{p}$ *such that* $0 < \underline{p} \leq p_i$ *for all* i*.*

(iii) Let a_t *be the column of either* $\Omega - \Pi \mathbf{1}_N \mathbf{1}_T'$ *or* $\Omega \circ \mathcal{E}$*. Then,* $\{a_t\}_{t \leq T}$ *are independent*

sub-gaussian random vector with $\mathbb{E}[a_t] = 0$*; more specifically, there is* $C > 0$ *such that*

$$\max_{t \leq T} \sup_{\|x\|=1} \mathbb{E}[\exp(sa_t'x)] \leq \exp(s^2 C), \quad \forall s \in \mathbb{R}.$$

We assume the heterogeneous observation probability across i. It generalizes the homogeneous observation probability assumption which is a typical assumption in the matrix completion literature. The sub-gaussian assumption in Assumption 3.2 (iii) helps us to bound $\|\Omega \circ \mathcal{E}\|$ and $\|\Omega - \Pi \mathbf{1}_N \mathbf{1}_T'\|$.

While the serial independence of the missing data indicators ω_{it} is assumed, we allow they are cross-sectional dependence among i. In doing so, we assume a cluster structure in $\{1, \ldots, N\}$, i.e., there is a family of nonempty disjoint clusters, $\mathcal{C}_1, \ldots, \mathcal{C}_\rho$ such that $\cup_{g=1}^{\rho} \mathcal{C}_g = \{1, \ldots, N\}$. So we divide units $\{1, ..., N\}$ into ρ disjoint clusters.

Assumption 3.3 (Cross-sectional Dependence in ω_{it})**.** *(i) Cross sectional units* ω_{it} *are independent across clusters. Within the same cluster, arbitrary dependence is allowed, but overall, we require* $\max_t \max_i \sum_{j=1}^{N} |Cov(\omega_{it}, \omega_{jt}|\mathcal{M})| < C$.

(ii) The number of elements in clusters should satisfy: There is $\vartheta \geq 1$ *such that the maximum number of elements in one cluster is bounded by* ϑ*. That is,* $\max_g |\mathcal{C}_g|_o \leq \vartheta$*. Here,* ϑ *is allowed to increase as* N *and* T *increase.*

Due to the cluster structure in Assumption 3.3 (i), we can construct a "leave-cluster-out" estimator $\breve{\beta}^{\{-i\}}$ which is independent from the sample of unit i. Similarly to the idea of (2.4) and (2.5), we can rule out the samples of the cluster that includes unit i. The difference from (2.5) is that we identify all the units which are in the same cluster as unit i and replace their rows of the goodness of fit part by their conditional expectations.[11] Together with the leave-one-out estimator $\breve{\beta}^{(-t)}$, the leave-cluster-out estimator $\breve{\beta}^{\{-i\}}$ plays a pivotal role in showing the solution of (2.3) is close to that of (2.4).

[11] For the formal definitions of the estimators, please refer to Section A.6 of the Supplement and Remark A.1 in the section.

The parameter for the cluster size ϑ is bounded by Assumption 3.4 (ii). For instance, in the case where $N \asymp T$ and $\{h_t(\cdot)\}_{t\leq T}$ are smooth enough, if we estimate the cross-sectional average of a certain period, the assumption requires $\vartheta \approx o(\sqrt{N/\log N})$ since K is allowed to grow very slowly by setting a large ν.

We are interested in making inference about group averaged effects. Let $\mathcal{G}$ be a particular group; the object of interest is

$$\frac{1}{|\mathcal{G}|_o}\sum_{(i,t)\in\mathcal{G}} M_{it} = \frac{1}{|\mathcal{G}|_o}\sum_{(i,t)\in\mathcal{G}} h_t(\zeta_i).$$

Here the group of interest as $\mathcal{G} = \mathcal{I}\times\mathcal{T}$ where $\mathcal{I}\subset\{1,\dots,N\}$ and $\mathcal{T}\subset\{1,\dots,T\}$. We impose the following assumption on the rates of ϑ, K and $\min\{|\mathcal{I}|_o, |\mathcal{T}|_o\}$. Recall that K denotes the true rank of the low-rank matrix.

Assumption 3.4 (Slowly increasing ϑ, K and $\min\{|\mathcal{I}|_o, |\mathcal{T}|_o\}$)**.**

(i) $\min\{|\mathcal{I}|_o^{\frac{1}{2}}, |\mathcal{T}|_o^{\frac{1}{2}}\}K^{\frac{7}{2}}\max\{\sqrt{N}\log^2 N, \sqrt{T}\log^2 T\} = o(\min\{N,T\})$,

(ii) $\min\{|\mathcal{I}|_o^{\frac{1}{2}}, |\mathcal{T}|_o^{\frac{1}{2}}\}\vartheta K^{\frac{7}{2}}\max\{N\sqrt{\log N}, T\sqrt{\log T}\} = o(\min\{N^{\frac{3}{2}}, T^{\frac{3}{2}}\})$,

(iii) $\min\{|\mathcal{I}|_o, |\mathcal{T}|_o\}\max\{N,T\} = o(K^{2\nu-3})$.

Assumptions 3.4 (i), (ii) accommodate the case where the parameters ϑ, K, and $\min\{|\mathcal{I}|_o, |\mathcal{T}|_o\}$ increase slowly as N and T go to infinity. If $\min\{|\mathcal{I}|_o, |\mathcal{T}|_o\}$ and ϑ are finite (or increase slowly), it is easily satisfied since K grows slowly as long as $\{h_t(\cdot)\}_{t\leq T}$ are smooth enough. On the other hand, Assumption 3.4 (iii) together with Assumption 3.1 (ii) controls the size of the approximate error. It shows that K is allowed to grow slowly if ν is large.

Lastly, the following assumption requires that the nonzero singular values of $M^\star$ have the same order and proper gaps between each other. Let ψ_r be the r-th largest singular value of $M^\star$.

Assumption 3.5 (Eigengap)**.** *There are $c, C > 0$ such that with probability converging to 1, $\psi_1 \leq C\psi_K$ and $\psi_r - \psi_{r+1} \geq c\psi_K$, $r = 1,\dots,K$, where ψ_r is the r-th singular value of*

$M^\star$.

Then, under the above assumptions, the estimator for the group average of M_{it} has the asymptotic normality as follows.

Theorem 3.1. *Suppose Assumptions 3.1 - 3.5 hold. In addition, suppose that* $\|\beta\|_F = O_P(\sqrt{NK})$, $\|F\|_F = O_P(\sqrt{TK})$ *and* $\|\bar{\beta}_{\mathcal{I}}\|$, $\|\bar{F}_{\mathcal{T}}\|$ *are bounded away from zero, where* $\bar{\beta}_{\mathcal{I}} = \frac{1}{|\mathcal{I}|_o}\sum_{j\in\mathcal{I}}\beta_j$ *and* $\bar{F}_{\mathcal{T}} = \frac{1}{|\mathcal{T}|_o}\sum_{s\in\mathcal{T}}F_s$. *Then,*

$$\mathcal{V}_{\mathcal{G}}^{-\frac{1}{2}}\left(\frac{1}{|\mathcal{G}|_o}\sum_{(i,t)\in\mathcal{G}}\widehat{M}_{it} - \frac{1}{|\mathcal{G}|_o}\sum_{(i,t)\in\mathcal{G}}M_{it}\right) \xrightarrow{D} \mathcal{N}(0,1),$$

where $\mathcal{V}_{\mathcal{G}} = \sigma^2\left(\frac{1}{|\mathcal{T}|_o^2}\sum_{t\in\mathcal{T}}\bar{\beta}_{\mathcal{I}}'\left(\sum_{j=1}^N\omega_{jt}\beta_j\beta_j'\right)^{-1}\bar{\beta}_{\mathcal{I}} + \frac{1}{|\mathcal{I}|_o^2}\sum_{i\in\mathcal{I}}\bar{F}_{\mathcal{T}}'\left(\sum_{s=1}^T\omega_{is}F_sF_s'\right)^{-1}\bar{F}_{\mathcal{T}}\right)$.

Theorem 3.1 covers the cross-sectional average of a certain period t (one column of the matrix) or the time average of a certain unit i (one row of the matrix) as a special case. Indeed, it can be more general in the sense that $\mathcal{G}$ of multiple columns with multiple rows is also allowed. In addition, $\mathcal{G}$ can consist of solely a certain (i,t), implying that we can conduct inference for one specific element of the matrix. We present these results as corollaries of Theorem 3.1 in Section A.2 of the supplement.

Finally, we propose an estimator of the asymptotic variance. We simply change all quantities to their estimates. Although factors and loadings are estimated up to rotation matrices, it does not cause difficulties since the rotation matrices are multiplied by their inverse and removed.

Theorem 3.2 (Feasible CLT)**.** *Under the assumptions of Theorem 3.1, we have*

$$\widehat{\mathcal{V}}_{\mathcal{G}}^{-\frac{1}{2}}\left(\frac{1}{|\mathcal{G}|_o}\sum_{(i,t)\in\mathcal{G}}\widehat{M}_{it} - \frac{1}{|\mathcal{G}|_o}\sum_{(i,t)\in\mathcal{G}}M_{it}\right) \xrightarrow{D} \mathcal{N}(0,1),$$

where

$$\widehat{\mathcal{V}}_{\mathcal{G}} = \widehat{\sigma}^2 \left(\frac{1}{|\mathcal{T}|_o^2} \sum_{t \in \mathcal{T}} \widehat{\bar{\beta}}_{\mathcal{I}}' \left(\sum_{j=1}^{N} \omega_{jt} \widehat{\beta}_j \widehat{\beta}_j' \right)^{-1} \widehat{\bar{\beta}}_{\mathcal{I}} + \frac{1}{|\mathcal{I}|_o^2} \sum_{i \in \mathcal{I}} \widehat{\bar{F}}_{\mathcal{T}}' \left(\sum_{s=1}^{T} \omega_{is} \widehat{F}_s \widehat{F}_s' \right)^{-1} \widehat{\bar{F}}_{\mathcal{T}} \right),$$

$\widehat{\bar{\beta}}_{\mathcal{I}} = \frac{1}{|\mathcal{I}|_o} \sum_{a \in \mathcal{I}} \widehat{\beta}_a$, $\widehat{\bar{F}}_{\mathcal{T}} = \frac{1}{|\mathcal{T}|_o} \sum_{a \in \mathcal{T}} \widehat{F}_a$, $\widehat{\sigma}^2 = \frac{1}{|\mathcal{W}|_o} \sum_{(i,t) \in \mathcal{W}} \widehat{\varepsilon}_{it}^2$, $\mathcal{W} = \{(i,t) : \omega_{it} = 1\}$ *and* $\widehat{\varepsilon}_{it} = y_{it} - \widehat{\beta}_i' \widehat{F}_t$.

4 Applications to Heterogeneous Treatment Effect Estimation

In this section, we propose the inference procedure for treatment effects by utilizing the asymptotic results in Section 3. Following the causal potential outcome setting (e.g., Rubin (1974), Imbens and Rubin (2015)), we assume that for each of N units and T time periods, there exists a pair of potential outcomes, $y_{it}^{(0)}$ and $y_{it}^{(1)}$ where $y_{it}^{(0)}$ denotes the potential outcome of the untreated situation and $y_{it}^{(1)}$ denotes the potential outcome of the treated situation. Importantly, among potential outcomes $y_{it}^{(0)}$ and $y_{it}^{(1)}$, we can observe only one realized outcome $y_{it}^{(\Upsilon_{it})}$ where $\Upsilon_{it} = 1\{\text{unit } i \text{ is treated at period } t\}$. Hence, we have two incomplete potential outcome matrices, $Y^{(0)}$ and $Y^{(1)}$, having missing components, and the problem of estimating the treatment effects can be cast as a matrix completion problem because of the missing components in the two matrices.

Specifically, we consider the nonparametric model such that for each $\iota \in \{0, 1\}$,

$$y_{it}^{(\iota)} = M_{it}^{(\iota)} + \varepsilon_{it} = h_t^{(\iota)}(\zeta_i) + \varepsilon_{it},$$

where ε_{it} is the noise, ζ_i is a vector of unit specific latent state variables. We regard $h_t^{(\iota)}(\cdot)$ as a deterministic function while ζ_i is a random vector. In the model, the treatment effect comes from the difference between the time-varying treatment function $h_t^{(1)}(\cdot)$ and the

control function $h_t^{(0)}(\cdot)$. Let $\omega_{it}^{(\iota)} = 1\{y_{it}^{(\iota)} \text{ is observed}\}$. Then, $\omega_{it}^{(1)} = \Upsilon_{it}$ and $\omega_{it}^{(0)} = 1 - \Upsilon_{it}$ because we observe $y_{it}^{(1)}$ when there is a treatment on (i,t) and observe $y_{it}^{(0)}$ when there is no treatment on (i,t).

We suppose the following seive representation for $h_t^{(\iota)}$:

$$h_t^{(\iota)}(\zeta_i) = \sum_{r=1}^{K} \kappa_{t,r}^{(\iota)} \phi_r(\zeta_i) + M_{it}^{R(\iota)}, \qquad \iota \in \{0,1\}$$

where $\kappa_{t,r}^{(\iota)}$ is the sieve coefficient, $\phi_r(\zeta_i)$ is the sieve transformation of ζ_i using the basis function $\phi_r(\cdot)$ and $M_{it}^{R(\iota)}$ is the sieve approximation error. Then, by representing $\sum_{r=1}^{K} \kappa_{t,r}^{(\iota)} \phi_r(\zeta_i)$ as $\beta_i' F_t^{(\iota)}$ where $\beta_i = [\phi_1(\zeta_i), \ldots, \phi_K(\zeta_i)]'$ and $F_t^{(\iota)} = [\kappa_{t,1}^{(\iota)}, \ldots, \kappa_{t,K}^{(\iota)}]'$, $h_t^{(\iota)}(\zeta_i)$ can be successfully represented as the approximate factor structure.

We make inference about the average treatment effect for a particular group of interest $(i,t) \in \mathcal{G}$:

$$\frac{1}{|\mathcal{G}|_o} \sum_{(i,t)\in\mathcal{G}} \Gamma_{it}, \quad \text{where } \Gamma_{it} = M_{it}^{(1)} - M_{it}^{(0)}.$$

The individual treatment effect Γ_{it} is estimated by $\widehat{\Gamma}_{it} = \widehat{M}_{it}^{(1)} - \widehat{M}_{it}^{(0)}$ where $\widehat{M}_{it}^{(0)}$ and $\widehat{M}_{it}^{(1)}$ are estimators of $M_{it}^{(0)}$ and $M_{it}^{(1)}$, respectively. Hence, by implementing the estimation steps in Algorithm 1 for each $\iota \in \{0,1\}$, we can derive the estimators for the group average of $M_{it}^{(0)}$ and $M_{it}^{(1)}$, and construct the average treatment effect estimator.

The notations are essentially the same as those in Section 2, and we just put the superscript (ι) to all notations to distinguish the pair of potential realizations.[12] We introduce assumptions for the asymptotic normality of the average treatment estimator. Basically, they imply that each potential realization satisfies the assumptions in Section 3.

Assumption 4.1 (Sieve representation)**.** *(i) For all $\iota \in \{0,1\}$, $\{h_t^{(\iota)}(\cdot)\}_{t\le T}$ belong to ball $\mathcal{H}\left(\mathcal{Z}, \|\cdot\|_{L_2}, C\right)$ inside a Hilbert space spanned by the basis $\{\phi_r\}_{r\ge 1}$, with a uniform L_2-bound C, that is, $\sup_{h\in\mathcal{H}(\mathcal{Z},\|\cdot\|_{L_2})} \|h\| \le C$, where $\mathcal{Z}$ is the support of ζ_i.*

[12] Exceptionally, because the notations concerning the group $\mathcal{G}$ do not depend on the potential realizations, we do not put the superscript (ι) to the notations concerning the group. In addition, ϑ, ε_{it}, β_i and K are same across the potential realizations in our model, so we do not put the superscript (ι) to them also.

(ii) The sieve approximation error satisfies: For some $\nu > 0$, $\max_{i,t} |M_{it}^{R(\iota)}| \leq CK^{-\nu}$.

(iii) For some $C \geq 0$, *with probability converging to 1,* $\max_i \frac{1}{K} \sum_{r=1}^{K} \phi_r^2(\zeta_i) \leq C$.

(iv) There is $c > 0$ *such that for* $\iota \in \{0,1\}$, *with probability converging to 1,* $\psi_{min}\left(\frac{1}{N}\beta'\beta\right) > c$, $\psi_{min}\left(\frac{1}{T}F^{(\iota)\prime}F^{(\iota)}\right) > c$.

(v) For all $\iota \in \{0,1\}$, $\sum_{i,t}\left(h_t^{(\iota)}(\zeta_i)\right)^2 \asymp NT$.

Assumption 4.2 (DGP for ε_{it} and Υ_{it})**.** *(i) Let* $\zeta = \{\zeta_i\}_{1\leq i\leq N}$. *Conditioning on* ζ, ε_{it} *is i.i.d. zero-mean, sub-gaussian random variable such that* $\mathbb{E}[\varepsilon_{it}|\zeta] = 0$, $\mathbb{E}[\varepsilon_{it}^2|\zeta] = \sigma^2$, *and* $\mathbb{E}[\exp(s\varepsilon_{it})|\zeta] \leq \exp(Cs^2\sigma^2)$, $\forall s \in \mathbb{R}$, *for some constant* $C > 0$.

(ii) Let $\Upsilon = [\Upsilon_{it}]_{N\times T}$. Υ *is independent from* $\mathcal{E}$. *Conditioning on* ζ, Υ_{it} *is independent across* t.[13] *In addition,* $\mathbb{E}[\Upsilon_{it}|\zeta] = \mathbb{E}[\Upsilon_{it}] = p_i^{(1)}$ *and there are constants* $\underline{p}$ *and* $\overline{p}$ *such that* $0 < \underline{p} \leq p_i^{(1)} \leq \overline{p} < 1$ *for all* i.

(iii) Let $\Pi^{(1)} = diag(p_1^{(1)}, \ldots, p_N^{(1)})$. *Let* a_t *be the column of* $\Upsilon - \Pi^{(1)}\mathbf{1}_N\mathbf{1}_T'$, $\Upsilon \circ \mathcal{E}$, *or* $(\mathbf{1}_N\mathbf{1}_T' - \Upsilon)\circ\mathcal{E}$. *Then,* $\{a_t\}_{t\leq T}$ *are independent sub-gaussian random vectors with* $\mathbb{E}[a_t] = 0$.

Assumption 4.3 (Cross-sectional Dependence in Υ_{it})**.** *(i) Let* $\mathcal{C}_{g(i)}$ *be the cluster where the unit* i *is included in. Then, for any units* $j_1, \ldots, j_m$ *which are not in* $\mathcal{C}_{g(i)}$, $\{\Upsilon_{j_1t}, \ldots, \Upsilon_{j_mt}\}$ *is independent from* Υ_{it} *for all* t. *In addition,* $\max_g |\mathcal{C}_g|_o \leq \vartheta$ *where* ϑ *increases as* N *and* T *increase.*

(ii) We have $\max_t \max_i \sum_{j=1}^{N} |Cov(\Upsilon_{it}, \Upsilon_{jt}|\zeta)| < C$.

Since all randomness of $M^{(\iota)}$ comes from ζ and $\Upsilon_{it} = \omega_{it}^{(1)} = 1 - \omega_{it}^{(0)}$, Assumption 4.2 and 4.3 imply Assumption 3.2 and 3.3 for each $\iota \in \{0,1\}$. Here, we assume the heterogeneous treatment probability across i. Note that $p_i^{(0)}$ becomes zero if $p_i^{(1)} = 1$. Hence, we set the upper bound $\overline{p} < 1$ of $p_i^{(1)}$ to estimate $M^{(0)}$ successfully.

Assumption 4.4 (Eigengap)**.** *There are* $c, C > 0$ *such that with probability converging to 1, for all* $\iota \in \{0,1\}$, $\psi_1^{(\iota)} \leq C\psi_K^{(\iota)}$ *and* $\psi_r^{(\iota)} - \psi_{r+1}^{(\iota)} \geq c\psi_K^{(\iota)}$, $\quad r = 1, \ldots, K$, *where* $\psi_r^{(\iota)}$ *is the*

[13] By the symmetry, we can also consider the model where Υ_{it} is independent across i and weakly dependent across t.

r-th singular value of $M^{\star(\iota)}$.

Then, we present the asymptotic normality of the average treatment effect estimator.

Theorem 4.1. *Suppose Assumptions 3.4 and 4.1 - 4.4 hold. For each* $\iota \in \{0,1\}$*, suppose that* $\|\beta\|_F = O_P(\sqrt{NK})$*,* $\|F^{(\iota)}\|_F = O(\sqrt{TK})$ *and* $\|\bar{\beta}_{\mathcal{I}}\|$*,* $\|\bar{F}_{\mathcal{T}}^{(\iota)}\|$ *are bounded away from zero, where* $\bar{F}_{\mathcal{T}}^{(\iota)} = \frac{1}{|\mathcal{T}|_o}\sum_{s\in\mathcal{T}} F_s^{(\iota)}$*. Then, we have*

$$\left(\mathcal{V}_{\mathcal{G}}^{(0)} + \mathcal{V}_{\mathcal{G}}^{(1)}\right)^{-\frac{1}{2}} \left(\frac{1}{|\mathcal{G}|_o}\sum_{(i,t)\in\mathcal{G}} \widehat{\Gamma}_{it} - \frac{1}{|\mathcal{G}|_o}\sum_{(i,t)\in\mathcal{G}} \Gamma_{it}\right) \xrightarrow{D} \mathcal{N}(0,1),$$

where

$$\mathcal{V}_{\mathcal{G}}^{(\iota)} = \sigma^2 \left(\frac{1}{|\mathcal{T}|_o^2}\sum_{t\in\mathcal{T}} \bar{\beta}_{\mathcal{I}}' \left(\sum_{j=1}^{N} \omega_{jt}^{(\iota)} \beta_j \beta_j'\right)^{-1} \bar{\beta}_{\mathcal{I}} + \frac{1}{|\mathcal{I}|_o^2}\sum_{i\in\mathcal{I}} \bar{F}_{\mathcal{T}}^{(\iota)\prime} \left(\sum_{s=1}^{T} \omega_{is}^{(\iota)} F_s^{(\iota)} F_s^{(\iota)\prime}\right)^{-1} \bar{F}_{\mathcal{T}}^{(\iota)}\right).$$

Corollary 4.2 (Feasible CLT)**.** *Under the assumptions of Theorem 4.1, we have*

$$\left(\widehat{\mathcal{V}}_{\mathcal{G}}^{(0)} + \widehat{\mathcal{V}}_{\mathcal{G}}^{(1)}\right)^{-\frac{1}{2}} \left(\frac{1}{|\mathcal{G}|_o}\sum_{(i,t)\in\mathcal{G}} \widehat{\Gamma}_{it} - \frac{1}{|\mathcal{G}|_o}\sum_{(i,t)\in\mathcal{G}} \Gamma_{it}\right) \xrightarrow{D} \mathcal{N}(0,1),$$

where for each $\iota \in \{0,1\}$*,*

$$\widehat{\mathcal{V}}_{\mathcal{G}}^{(\iota)} = \left(\widehat{\sigma}^{(\iota)}\right)^2 \left(\frac{1}{|\mathcal{T}|_o^2}\sum_{t\in\mathcal{T}} \widehat{\bar{\beta}}_{\mathcal{I}}^{(\iota)\prime} \left(\sum_{j=1}^{N} \omega_{jt}^{(\iota)} \widehat{\beta}_j^{(\iota)} \widehat{\beta}_j^{(\iota)\prime}\right)^{-1} \widehat{\bar{\beta}}_{\mathcal{I}}^{(\iota)} + \frac{1}{|\mathcal{I}|_o^2}\sum_{i\in\mathcal{I}} \widehat{\bar{F}}_{\mathcal{T}}^{(\iota)\prime} \left(\sum_{s=1}^{T} \omega_{is}^{(\iota)} \widehat{F}_s^{(\iota)} \widehat{F}_s^{(\iota)\prime}\right)^{-1} \widehat{\bar{F}}_{\mathcal{T}}^{(\iota)}\right).$$

Here, $\widehat{\bar{\beta}}_{\mathcal{I}}^{(\iota)} = \frac{1}{|\mathcal{I}|_o}\sum_{a\in\mathcal{I}} \widehat{\beta}_a^{(\iota)}$*,* $\widehat{\bar{F}}_{\mathcal{T}}^{(\iota)} = \frac{1}{|\mathcal{T}|_o}\sum_{a\in\mathcal{T}} \widehat{F}_a^{(\iota)}$*,* $\left(\widehat{\sigma}^{(\iota)}\right)^2 = \frac{1}{|\mathcal{W}^{(\iota)}|_o}\sum_{(i,t)\in\mathcal{W}^{(\iota)}} \left(\widehat{\varepsilon}_{it}^{(\iota)}\right)^2$*,* $\mathcal{W}^{(\iota)} = \{(i,t) : \omega_{it}^{(\iota)} = 1\}$ *and* $\widehat{\varepsilon}_{it}^{(\iota)} = y_{it}^{(\iota)} - \widehat{\beta}_i^{(\iota)\prime} \widehat{F}_t^{(\iota)}$*.*

5 Empirical study: Impact of the president on allocating the U.S. federal budget to the states

To illustrate the use of our inferential theory, we present an empirical study about the impact of the president on allocating the U.S. federal budget to the states. The allocation of the federal budget in the U.S. is the outcome of a complicated process involving diverse institutional participants. However, the president plays a particularly important role among the participants. Ex ante, the president is responsible for composing a proposal, which is supposed to be submitted to Congress, and initiates the actual authorization and appropriations processes. Ex post, once the budget has been approved, the president has a veto power that can be overridden only by a qualified majority equal to two-thirds of Congress. In addition, the president exploits extra additional controls over agency administrators who distribute federal funds.

There is a vast theoretical and empirical literature about the impact of the president on allocating the federal budget to the states (e.g., Cox and McCubbins (1986), Anderson and Tollison (1991), McCarty (2000), Larcinese et al. (2006), Berry et al. (2010)). In particular, Cox and McCubbins (1986) provide a theoretical model which supports the idea that more funds are allocated where the president has larger support because of the ideological relationship between voters and the president, and Larcinese et al. (2006) have found that states which supported the incumbent president in past presidential elections tend to receive more funds empirically. In this section, we further investigate the impact using our inferential theory for the heterogeneous treatment effect with a wider set of data.

Here, the hypothesis we want to test is whether federal funds are disproportionately targeted to states where the incumbent president is supported in the past presidential election. We use data on federal outlays for the 50 U.S. states with the District of Columbia

from 1953 to 2018.[14] Following the model in Section 4, we set the treatment indicator as

$$\Upsilon_{it} = 1\{\text{the state } i \text{ supported the president of year } t \text{ in the presidential election}\}.$$

If the candidate whom the state i supported in the previous presidential election is same as the president at year t, we consider it as "treated" and otherwise, we consider it as "untreated". In addition, for the outcome variable y_{it}, we use the following ratio:

$$y_{it} = \frac{\tilde{y}_{it}}{\sum_i \tilde{y}_{it}} \times 100, \quad \text{where } \tilde{y}_{it} \text{ is the per-capita federal grant in state } i \text{ at year } t.$$

This is each state's (per-capita) portion of the federal grant at each year. In fact, the per-capita federal grant, $\tilde{y}_{it}$, increases a lot as time goes by. Even after converting to the real dollars using the GDP deflator, the real per-capita federal grant of 2018 is about 12 times bigger than that of 1953. Because of this tendency, if we use the real per-capita federal grant as our outcome variable, the time average of the treatment effect largely depends on the treatment effect of the more recent years and that of the early years will be factored less into the time average of the treatment effect. To avoid this problem, we use the above normalized outcome y_{it} instead.

First, we study the time average of the treatment effect of each state. For each state i, we compute $\frac{1}{T}\sum_{t=1}^{T} \Gamma_{it}$. Here, we consider the time average of all periods (1953 - 2018). Figure 1 presents the estimates of the average treatment effect and the corresponding t-statistics. We can check that there are statistically significant positive treatment effects in most states. However, while some states like Alaska, D.C., South Dakota, and Wyoming show the evidence of positive treatment effects even at very conservative significant level, some states like Arkansas, Georgia, Kentucky, Ohio, and West Virginia don't have statistically significant positive treatment effect.

To understand the reason of difference, we generate an indicator of long-term swing

[14] We get the data from the U.S. Census Bureau, NASBO (National Association of State Budget Officers), and SSA (Social Security Administration). Because of absence of data, the years, 1960, 1976∼1979, are excluded.

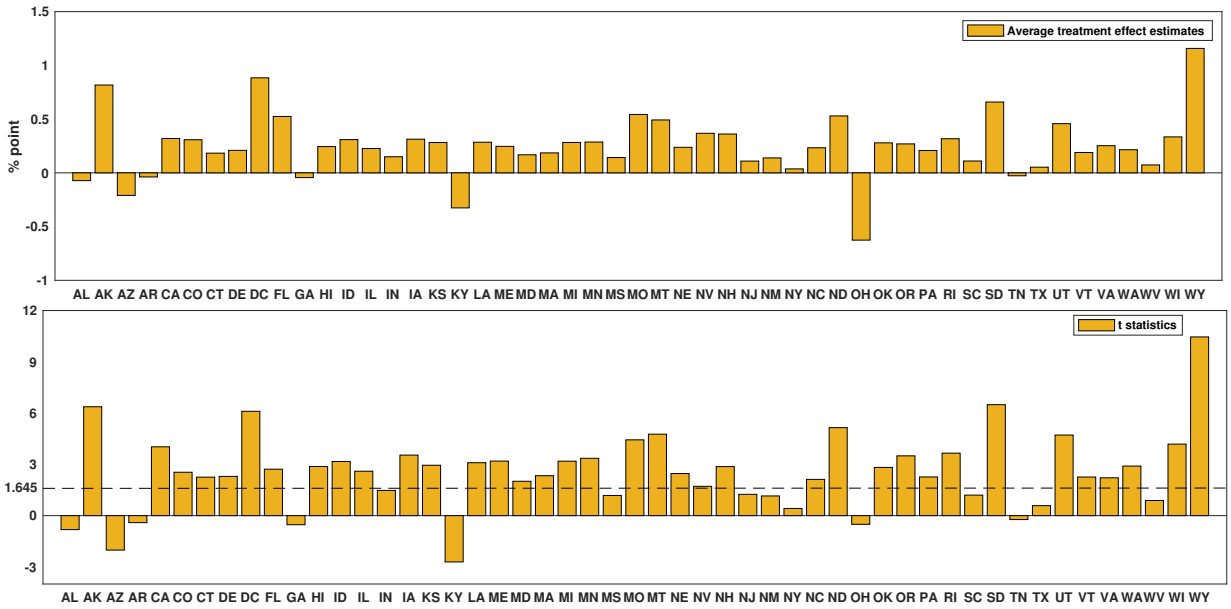

Figure 1: Time average treatment effect estimates of each state and corresponding t-statistics

NOTE: When we use the B-H procedure to control the size of FDR at 5%, the list of states with rejected decisions is unchanged.

which is based on the number of times a state swung its support from a party to another in the presidential elections in our data period. From Table 1, we can check that most states showing large t-statistics are in "Loyal states" while states having no statistically significant positive treatment effect are generally in "Swing state" or "Weak swing state". From this observation, we see that the treatment effect is closely related to the loyalty of states to parties.

Table 1: Number of swing of each state

Group	# of swing	States
Loyal states	0	DC
	2	AK, ID, KS, NE, ND, OK, SD, UT, WY
Weak loyal states	3	AZ, CA, CT, IL, ME, MA, MN, NJ, OR, SC, VT, VA, WA
	4	IN, MI, MT, TX
Weak swing states	5	AL, CO, DE, HI, MD, NV, NH, NM, NY, NC, RI
	6	IA, MS, MO, PA, TN, WI
Swing states	7	AR, GA, KY, WV
	8	FL, OH
	9	LA

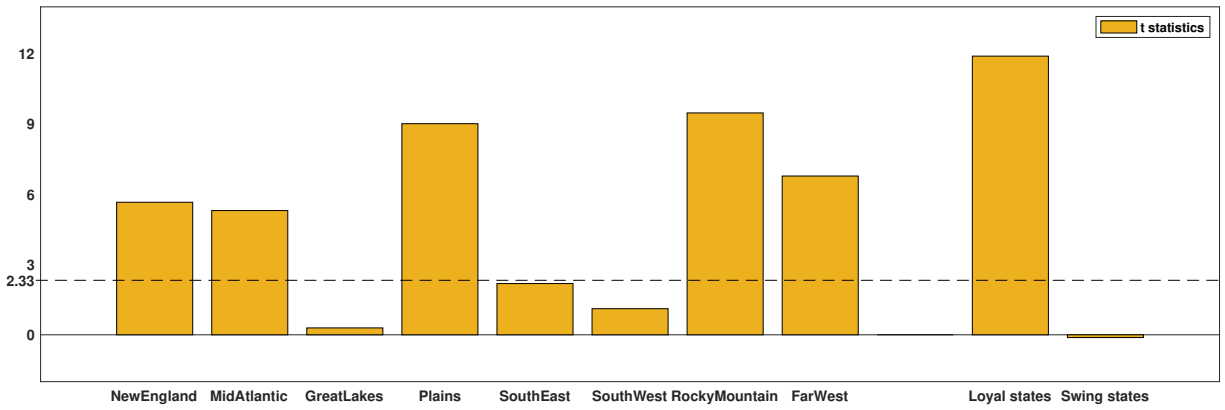

Figure 2: Test statistics for the time average treatment effect of each region

NOTE: "New England" includes CT, ME, MA, NH, RI,VT, "Mid Atlantic" includes DE, D.C., MD, NJ, NY, PA, "Great Lakes" includes IL, IN, MI, OH, WI, "Plains" includes IA, KS, MN, MO, NE, ND, SD, "South East" includes AL, AR, FL, GA, KY, LA, MS, NC, SC, TN, VI, WV, "South West" includes AZ, NM, OK, TX, "Rocky Mountain" includes CO, ID, MT, UT, WY, and "Far West" includes AK, CA,HI, NV, OR, WA.

In addition, Figure 2 shows the test statistics for the time average of the treatment effect of each region. At the 1% significant level, New England, Mid Atlantic, Plains, Rocky Mountain, and Far West have the positive treatment effects while Great Lakes, South East, and South West do not. Note that Many states in Great Lakes, South East, and South West are in "Swing states" or "Weak swing states". As we can see in Figure 2, "Swing states" do not have statistically significant positive treatment effects while "Loyal states" have significant positive treatment effects. This result is in line with the empirical study of Larcinese et al. (2006) finding that states with loyal supports tend to receive more funds, while swing states are not rewarded. In addition, it is aligned with the assertion of Cox and McCubbins (1986) that the targeting of loyal voters can be seen as a safer investment as compared to aiming for swing voters and risk-adverse political actors may allocate more funds to loyal states.

Figure 3 shows the test statistics for the average of the treatment effect of each president. For each president, let $\mathcal{T}$ denote his presidential period. We estimate the present treatment effect: ($i = 1, ..., N$ denote the states)

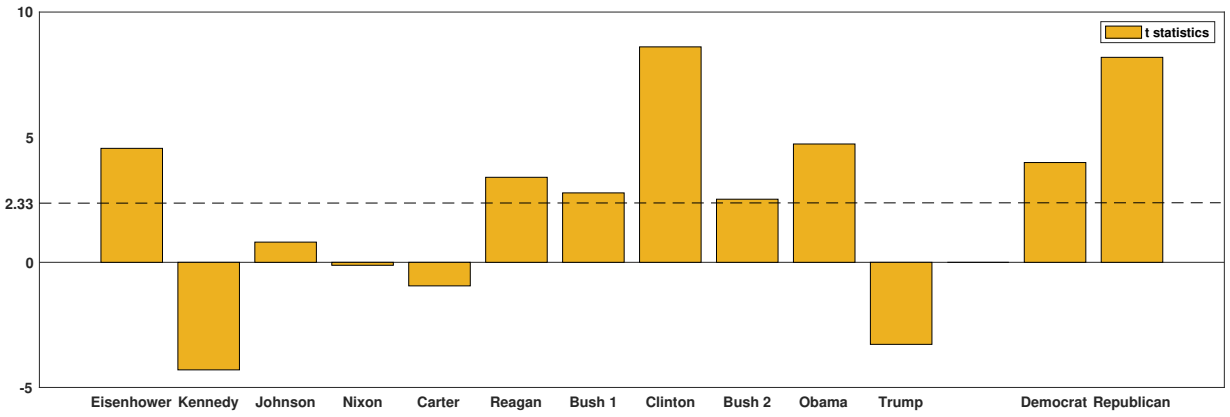

Figure 3: Test statistics for the average treatment effect of each president

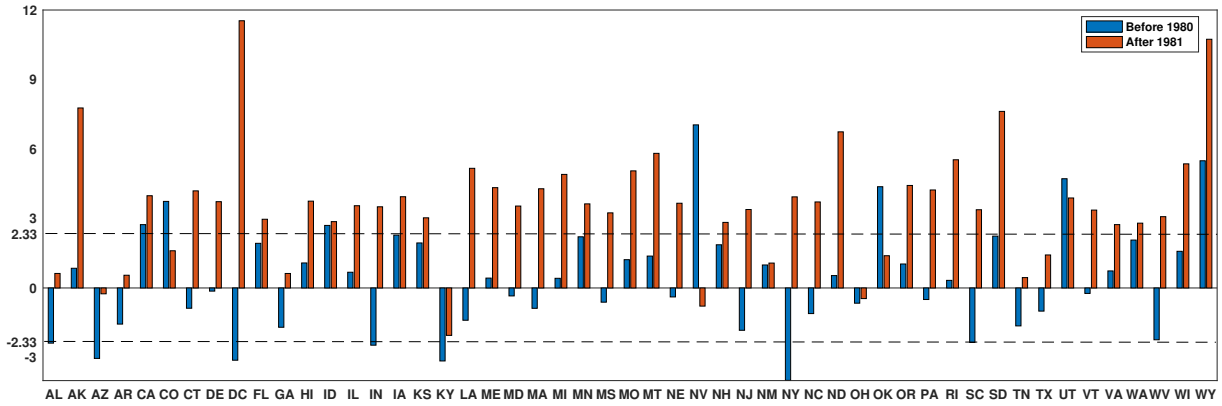

Figure 4: Test statistics for the average treatment effect before 1980 and after 1981

$$\frac{1}{|\mathcal{T}|_0 N} \sum_{t \in \mathcal{T}} \sum_{i=1}^{N} \Gamma_{it}.$$

Although there exist some exceptions, there are no statistically significant positive treatment effects before Carter, while there are significant positive treatment effects after Reagan. From Figure 4, we can check that before 1980, there is no significant positive treatment effect in most states, while there are significant positive treatment effects in most states after 1981. Hence, there is a big difference between 'before 1980' and 'after 1981' and the tendency that incumbent presidents reward states that showed their support in the president elections became significant after Reagan, that is, after the 1980s. It seems that after the 1980s, the presidents wanted to have more influence on the allocation of the federal funds to reward their supporters. One evidence is that starting from the 1980s, all pres-

idents have put forward proposals for the introduction of presidential line-item veto and tried to increase the power of the president to control federal spending.[15]

To summarize, we find the states that supported the incumbent president in past presidential elections tend to receive more federal funds and this tendency is stronger for the loyal states than the swing states. In addition, compared to before 1980, this tendency is stronger after the 1980s.

6 Simulation Study

In this section, we provide the finite sample performances of the estimators. We first study the performances of the estimators of M_{it} and $\frac{1}{|\mathcal{G}|_o}\sum_{(i,t)\in\mathcal{G}} M_{it}$, and then study performances of the average treatment effect estimators. To save space, some results are relegated to the supplement.

First of all, in order to check the estimation quality of our estimator, we compare the Frobenius norms of the estimation errors for several existing estimators of M. In addition to our two-step least squares (TLS) estimator, we consider the debiased nuclear norm penalized estimators from Xia and Yuan (2021) and Chen et al. (2019). The comparison also includes the estimators based on the inverse probability weighting method (e.g., Xiong and Pelger (2020))[16] and the EM algorithm method (e.g., Jin et al. (2021)). The plain nuclear norm penalized estimators and the TLS estimator using sample splitting are also considered for comparison. For the data-generating designs, we consider the following three models:

- Factor model: $y_{it} = \beta_{1,i}F_{1,t} + \beta_{2,i}F_{2,t} + \varepsilon_{it}$, where $\beta_{1,i}, F_{1,t}, \beta_{2,i}, F_{2,t} \sim \mathcal{N}\left(\frac{1}{\sqrt{2}}, 1\right)$, (6.1)

[15] For an overview on the proposals of line-item veto, please see Fisher (2004).

[16] Note that this method is different from the nuclear norm penalized estimation using inverse probability weighting in Section 2.1. This method does not use the nuclear norm penalization. For the details, please refer to Abbe et al. (2020), Xiong and Pelger (2020) and Fan et al. (2020).

- Nonparametric model 1: $y_{it} = h_t(\zeta_i) + \varepsilon_{it}$, where $h_t(\zeta) = h_t^{poly}(\zeta) := \sum_{r=1}^{\infty} \frac{|U_{t,r}|}{r^3} \cdot \zeta^r$,
- Nonparametric model 2: $y_{it} = h_t(\zeta_i) + \varepsilon_{it}$, where $h_t(\zeta) = h_t^{sine}(\zeta) := \sum_{r=1}^{\infty} \frac{|U_{t,r}|}{r^3} \sin(r\zeta)$.

Here, $U_{t,r}$ is generated from $\mathcal{N}(2,1)$ and ζ_i is generated from Uniform$[0,1]$. In addition, ε_{it} is generated from the standard normal distribution independently across i and t. The observation pattern follows a heterogeneous missing-at-random mechanism where $\omega_{it} \sim$ Bernoulli(p_i) and p_i is generated from Uniform$[0.3, 0.7]$.

Table 2: Frobenius norm of estimation errors for estimators of M

Sample size	N = 100, T =100			N = 200, T = 100			N = 100, T =200		
Model	Factor	Sine	Poly	Factor	Sine	Poly	Factor	Sine	Poly
TLS	0.3035	0.2129	0.2057	0.2613	0.1871	0.1777	0.2522	0.1831	0.1831
TLS with SS	0.3130	0.2152	0.2080	0.2699	0.1893	0.1805	0.2551	0.1835	0.1836
Plain Nuclear	0.5637	0.3869	0.3745	0.4827	0.3342	0.3334	0.4814	0.3418	0.3433
(Hetero) CFMY	0.3312	0.2230	0.2128	0.2798	0.1916	0.183	0.2740	0.1914	0.1917
(Hetero) XY	0.3870	0.2369	0.2275	0.3185	0.1984	0.1931	0.3104	0.2019	0.2033
IPW	0.5280	0.2446	0.2435	0.4994	0.2184	0.2117	0.4254	0.1997	0.2068
EM	0.3033	0.2134	0.206	0.2611	0.1872	0.1777	0.2517	0.1834	0.1832

NOTE: "TLS" denotes our two-step least squares estimator, and "TLS with SS" means the estimator which uses the two-step least squares estimation and the sample splitting together. "Plain Nuclear" refers to the weighted nuclear norm regularized estimator. Also, "(Hetero) CFMY" and "(Hetero) XY" denote the debiased estimators from Chen et al. (2019) and Xia and Yuan (2021), respectively. "(Hetero)" represents that they are modified to allow the heterogeneous observation probabilities. " IPW" and "EM" denote the inverse probability weighting method and the EM algorithm method respectively. In addition, " Sine" and " Poly" refer to the functions $h_t^{sine}(\zeta)$ and $h_t^{poly}(\zeta)$, respectively.

Table 2 reports $||\widehat{M} - M||_F/\sqrt{NT}$ averaged over 100 replications. We would like to highlight that our TLS estimator shows the best performance in almost all scenarios. Only the EM algorithm method is comparable to ours. Our estimator performs slightly better than the EM estimator in the nonparametric cases (Sine and Poly), while the EM estimator is slightly better than ours in the factor model case. Also, our method always outperforms the two-step least squares estimation using the sample splitting method. The debiased nuclear norm penalized estimators from Xia and Yuan (2021) and Chen et al. (2019) are

slightly worse than ours in this experiment. If we restrict our attention to the factor model case, the performance of our method is better than that of the IPW estimator with fairly large gaps. Lastly, the plain nuclear norm regularized estimator shows the worst performances uniformly.

Second, we study the finite sample distributions for standardized estimates defined as $\frac{\widehat{M}_{it}-M_{it}}{se(\widehat{M}_{it})}$. For comparison, we report the results of the nuclear norm regularized estimator and the two-step least squares (TLS) estimator using sample splitting, in addition to our TLS estimator which does not utilize the sample splitting method. For the nuclear norm regularized estimator, we use the sample standard deviation obtained from the simulations for $se(\widetilde{M}_{it})$ because the theoretical variance of this estimator is unknown. For the TLS estimator using sample splitting, we construct the standard error following Chernozhukov et al. (2019). Here, we consider the nonparametric models in (6.1). Hereinafter, the number of simulations is set to 1,000.

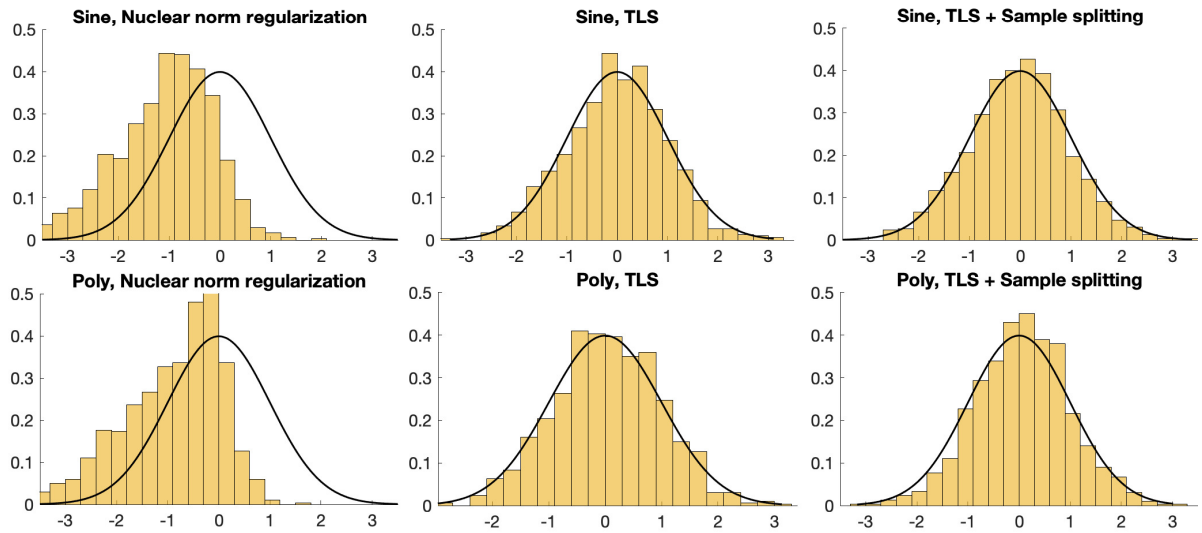

Figure 5: Histograms of standardized estimates, $(\widehat{M}_{it} - M_{it})/se(\widehat{M}_{it})$

NOTE: The sample size is $N = T = 200$. "Nuclear norm regularization" refers to the weighted nuclear norm regularized estimator and "TLS" denotes our TLS estimator which does not use the sample splitting method. "TLS + Sample splitting" refers to the TLS estimator using sample splitting. In addition, "Sine" and "Poly" refer to the functions $h_t^{sine}(\zeta)$ and $h_t^{poly}(\zeta)$, respectively.

Figure 5 plots the scaled histograms of the standardized estimates with the standard

normal density. As we expected in theory, it shows that the standardized estimates of our TLS estimator, which does not use the sample splitting method, have similar distributions to the standard normal distribution, while the distributions of the standardized estimates of the nuclear norm regularized estimator are biased and noticeably different from the standard normal distribution. In addition, it reveals that there is no big difference in the similarity to the normal distribution between the distributions of the TLS estimator "with sample splitting" and "without sample splitting". Without sample splitting, the TLS estimator itself provides a good approximation to the standard normal distribution so that it can be used for the inference successfully. The coverage probabilities of the confidence interval in the supplement also show similar results.

In addition, to check whether our inferential theory for the group average estimators works well, we present the scaled histograms of the standardized estimates of our TLS estimators (which does not use the sample splitting method) and the coverage probabilities of the confidence interval with various groups in Figure A.2 and Table A.2 of the supplement. They show that the standardized estimates of our TLS estimator have similar distributions to the standard normal distribution in all groups, and it seems that our inferential theories for diverse groups work well.

Next, we study the finite sample property of the average treatment effect estimator. Following Section 4, for each $\iota \in \{0, 1\}$, we generate the data from $y_{it}^{(\iota)} = h_t^{(\iota)}(\zeta_i) + \varepsilon_{it}$, where $h_t^{(0)}(\zeta) = \sum_{r=1}^{\infty} \frac{|U_{t,r}|}{r^a} \sin(r\zeta)$, $h_t^{(1)}(\zeta) = \sum_{r=1}^{\infty} \frac{|U_{t,r}|+2}{r^a} \sin(r\zeta)$. The power parameter $a > 1$ controls the decay speed of the sieve coefficients. The forms of the above functions and the treatment effect $\Gamma_{it} = h_t^{(1)}(\zeta_i) - h_t^{(0)}(\zeta_i)$ are in Figure 6.

Here, ε_{it} and $U_{t,r}$ are independently generated from the standard normal distribution and ζ_i is generated from Uniform$[0, 1]$. The treatment pattern follows $\Upsilon_{it} \sim$ Bernoulli$(p_i^{(1)})$ and $p_i^{(1)} \sim$ Uniform$[0.3, 0.7]$.

Figure 7 presents the scaled histograms of the standardized estimates of the average

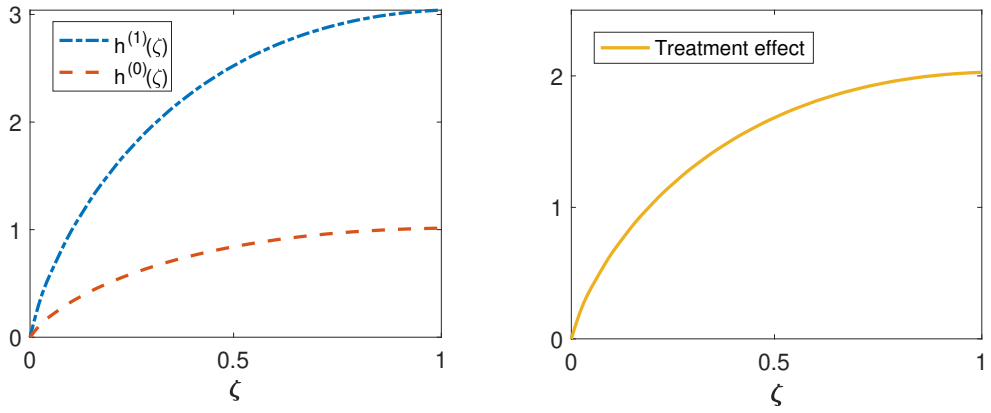

Figure 6: Shape of function $h_t^{(\iota)}(\zeta)$ and treatment effect function ($U_{t,r} = 1$, $a = 2$)

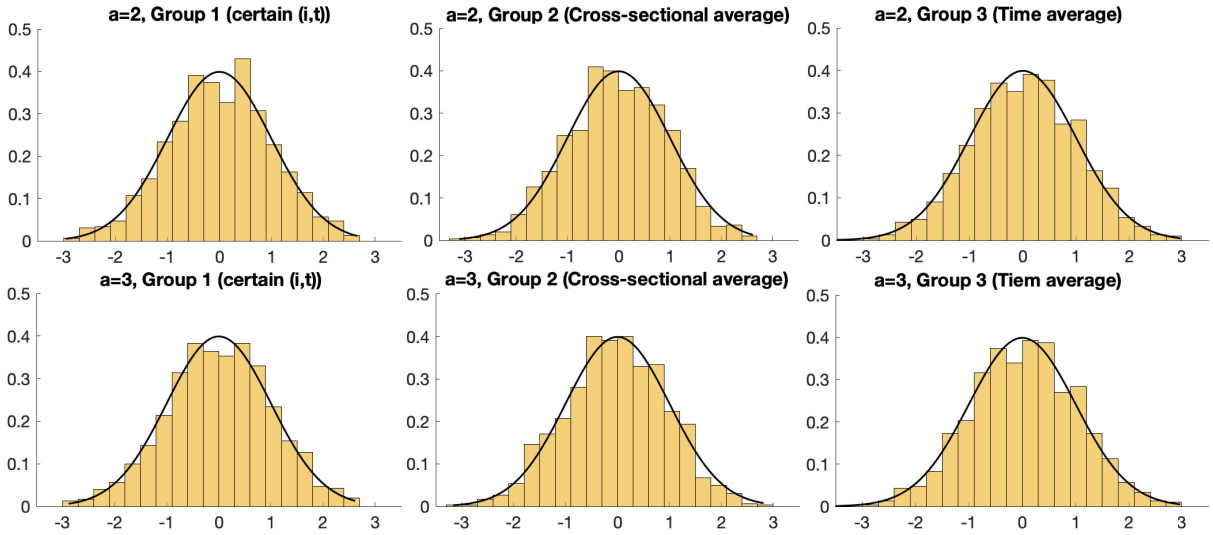

Figure 7: Histograms of standardized estimates, $\frac{\frac{1}{|\mathcal{G}|_o}\sum_{(i,t)\in\mathcal{G}}\widehat{\Gamma}_{it}-\frac{1}{|\mathcal{G}|_o}\sum_{(i,t)\in\mathcal{G}}\Gamma_{it}}{se\left(\frac{1}{|\mathcal{G}|_o}\sum_{(i,t)\in\mathcal{G}}\widehat{\Gamma}_{it}\right)}$

NOTE: Here, the sample size is $N = T = 300$. "Group 1" refers to $\mathcal{G}_1$, "Group 2" denotes $\mathcal{G}_2$ and "Group 3" refers to $\mathcal{G}_3$.

treatment effect estimators for the groups $\mathcal{G}_1$, $\mathcal{G}_2$ and $\mathcal{G}_3$ defined above. Here, the standard estimates are given as $\frac{\frac{1}{|\mathcal{G}|_o}\sum_{(i,t)\in\mathcal{G}}\widehat{\Gamma}_{it}-\frac{1}{|\mathcal{G}|_o}\sum_{(i,t)\in\mathcal{G}}\Gamma_{it}}{se\left(\frac{1}{|\mathcal{G}|_o}\sum_{(i,t)\in\mathcal{G}}\widehat{\Gamma}_{it}\right)}$. As we expected in the theory, it shows that the standardized estimates of the average treatment effect estimators of all groups have similar distributions to the standard normal distribution. In addition, the coverage probabilities of the confidence interval in the supplement also show similar results. Overall, the results are quite good, and it seems that our asymptotic theory for inference works well.

7 Conclusion

This paper studies the inferential theory for estimating low-rank matrices and provides an inference method for the average treatment effect as an application. Without the aid of sample splitting, our estimation procedure successfully resolves the problem of the shrinkage bias, and the resulting estimator attains the asymptotic normality. Unlike Chernozhukov et al. (2019, 2021) which exploit sample splitting, our estimation step is simple, and we can avoid some undesirable properties of sample splitting. In addition, this paper allows the heterogeneous observation probability and uses inverse probability weighting to control the effect of the heterogeneous observation probability. The simulation results show that our theory is valid in the finite sample.

8 Supplementary Materials

To save space, other additional contents are relegated to the supplement. All of the technical proofs are contained in the supplement.

References

Abbe, E., Fan, J., Wang, K., and Zhong, Y. (2020). Entrywise eigenvector analysis of random matrices with low expected rank. *Annals of statistics*, 48(3):1452.

Anderson, G. M. and Tollison, R. D. (1991). Congressional influence and patterns of new deal spending, 1933-1939. *The Journal of Law and Economics*, 34(1):161–175.

Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202.

Berry, C. R., Burden, B. C., and Howell, W. G. (2010). The president and the distribution of federal spending. *American Political Science Review*, 104(4):783–799.

Cai, J.-F., Candès, E. J., and Shen, Z. (2010). A singular value thresholding algorithm for matrix completion. *SIAM Journal on optimization*, 20(4):1956–1982.

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

Candès, E. J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717.

Chen, J., Liu, D., and Li, X. (2020a). Nonconvex rectangular matrix completion via gradient descent without $\ell_{2,\infty}$ regularization. *IEEE Transactions on Information Theory*, 66(9):5806–5841.

Chen, Y., Chi, Y., Fan, J., Ma, C., and Yan, Y. (2020b). Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. *SIAM journal on optimization*, 30(4):3098–3121.

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

Chernozhukov, V., Hansen, C., Liao, Y., and Zhu, Y. (2021). Inference for low-rank models. *arXiv preprint arXiv:2107.02602*.

Chernozhukov, V., Hansen, C. B., Liao, Y., and Zhu, Y. (2019). Inference for heterogeneous effects using low-rank estimations. Technical report, cemmap working paper.

Cox, G. W. and McCubbins, M. D. (1986). Electoral politics as a redistributive game. *The Journal of Politics*, 48(2):370–389.

Fan, J., Li, K., and Liao, Y. (2020). Recent developments on factor models and its applications in econometric learning. *arXiv preprint arXiv:2009.10103*.

Fisher, L. (2004). A presidential item veto. In *CRS Report for Congress*.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Jin, S., Miao, K., and Su, L. (2021). On factor models with random missing: Em estimation, inference, and cross validation. *Journal of Econometrics*, 222(1):745–777.

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

Larcinese, V., Rizzo, L., and Testa, C. (2006). Allocating the us federal budget to the states: The impact of the president. *The Journal of Politics*, 68(2):447–456.

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons.

Ma, C., Wang, K., Chi, Y., and Chen, Y. (2019). Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution. *Foundations of Computational Mathematics*, pages 1–182.

Ma, S., Goldfarb, D., and Chen, L. (2011). Fixed point and bregman iterative methods for matrix rank minimization. *Mathematical Programming*, 128(1-2):321–353.

Ma, W. and Chen, G. H. (2019). Missing not at random in matrix completion: The effectiveness of estimating missingness probabilities under a low nuclear norm assumption. *Advances in Neural Information Processing Systems*, 32.

Mazumder, R., Hastie, T., and Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *Journal of machine learning research*, 11(Aug):2287–2322.

McCarty, N. M. (2000). Presidential pork: Executive veto power and distributive politics. *American Political Science Review*, 94(1):117–129.

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

Parikh, N. and Boyd, S. (2014). Proximal algorithms. *Foundations and Trends in optimization*, 1(3):127–239.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Schnabel, T., Swaminathan, A., Singh, A., Chandak, N., and Joachims, T. (2016). Recommendations as treatments: Debiasing learning and evaluation. In *International Conference on Machine Learning*, pages 1670–1679.

Xia, D. and Yuan, M. (2021). Statistical inferences of linear forms for noisy matrix completion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(1):58–77.

Xiong, R. and Pelger, M. (2020). Large dimensional latent factor modeling with missing observations and applications to causal inference. arxiv eprint. *arXiv preprint arXiv:1910.08273*.