Inference for Low-rank Estimation with Application to Treatment Effect Estimation

Jungjun Choi Hyukjun Kwon

November 2021

Abstract

This paper studies the inferential theory for the (debiased) nuclear norm penalized estimator of the latent approximate low-rank matrix when the observation matrix is subject to missing. It also provides an inference method for the average treatment effect as an application. Although the nuclear norm penalization causes shrinkage bias which makes inference infeasible in general, our debiasing procedure successfully removes it, and the resulting debiased estimator attains the asymptotic normality. Unlike other debiasing schemes for the inference using the nuclear norm penalized estimator such as in Chernozhukov et al. (2019, 2021), our debiasing method does not resort to sample splitting. So our estimation step is simple, and we can avoid some undesirable properties of sample splitting. In addition, this paper allows dependent observation patterns and heterogeneous observation probabilities, and uses inverse probability weighting, which improves the estimation performance by treating units with different observation probabilities in a fair manner. We illustrate the proposed method in simulation experiments and the empirical study about the impact of the presidential vote on allocating the U.S. federal budget to the states.

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, localization, social networks recovery, magnetic resonance parameter mapping, and joint alignment; see, e.g., Rennie and Srebro (2005), Luo et al. (2010), Liu and Vandenberghe (2010), Zhang et al. (2015), and Chen and Chi (2018).

Although it may seem to be a problem that is very distinct from the type of problem studied in econometrics, the matrix completion problem spans a wide range of econometric applications. For instance, in settings where researchers are interested in causal effects of a binary treatment, one can consider the realized data as consisting of two incomplete potential outcome matrices, one for the outcomes of the treated situation and another for the outcomes of the untreated situation. Hence the task of estimating the treatment effects can be regarded as a matrix completion problem. In addition, the matrix completion method can be used to resolve the problems of the unbalanced panel like the reduction in sample size or bias problems by proposing estimation or inference methods that are robust to missing.

It is obvious that the recovery of missing entries is impossible in general if there is no further assumption on the matrix of interest. 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 the missing data problem:

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

where Y is the $N \times T$ observation matrix, M is the latent matrix of interest, and $\mathcal{E}$ represents the noise contamination. Importantly, M is assumed to be an approximate low-rank matrix and have an approximate factor structure $M \approx \beta F'$, where β is the factor loadings and F is the latent factors. In addition, we allow some entries of Y to be unobserved. 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 commonly used method 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) formulate the low-rank matrix completion problem in a simple setting (each entry of a low-rank

matrix is observed uniformly at random without noise), and provide the lower bound of the number of observations that are required to perfectly recover the matrix using the nuclear norm penalization. Candes and Plan (2010), Koltchinskii et al. (2011), Negahban and Wainwright (2012), and Chen et al. (2020) allow noise contaminations and provide convergence rates for the nuclear norm penalized estimator. In addition, 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) provides algorithms to compute the nuclear norm penalized estimator.

However, while there are plenty of works on the statistical rate of convergence for the nuclear norm penalized estimator, 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 debiasing method for the inference of the nuclear norm penalized estimator. Chen et al. (2019) and Xia and Yuan (2021) explicitly subtract the estimator of bias from the initial estimator and exploit the projection method to control the rank of the debiased estimator for reducing the variance. However, although they successfully obtain the asymptotic normality of the estimator, their methods have some limitations. Chen et al. (2019) assume normally distributed noise terms and derive the normality of the estimator from this assumption, and Xia and Yuan (2021) use the sample splitting method, which has several disadvantages. Moreover, both of them assume that missing occurs with the same probability across units and time. In addition, their debiasing methods do not allow any dependence structure in the observation pattern.

On the other hand, Chernozhukov et al. (2019, 2021) propose a different debiasing method. They use the two-step least square procedure with sample splitting. The two-step least square procedure is an implicit debiasing method in which we estimate the factors and the factor loadings successively using the least square estimations, starting at the singular vectors of the nuclear norm penalized estimator. This method takes advantage of the fact that it suffices to estimate the factors and loadings unbiasedly up to a rotation. The shrinkage bias caused by the nuclear norm penalization is absorbed by the rotation matrix, and thus the product of estimators for factors and loadings, which is the estimator for the common component M, is unbiased. One merit of this debiasing method is that it can allow the dependence structure in the observation pattern and heterogeneous observation probability. However, in addition to the two-step least square procedure, Chernozhukov et al. (2019, 2021) also exploit sample splitting to show some potential bias term is negligible.

We contribute to the literature by providing an inferential theory of the nuclear norm penalized estimation, which only uses the two-step least squares procedure for debiasing 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 full sample of observed Y onto the initial estimator for β.

3. To re-estimate β, regress the full sample of observed Y on the estimator for F.

4. The product of the estimators in Steps 2 and 3 is the final estimator for M.

In the estimation procedure of Chernozhukov et al. (2019, 2021), which is similar to ours except for using sample splitting, they randomly split the observed sample into two subsamples and use only one subsample to compute $\widetilde{M}$ and the initial estimator for β, and then use another subsample to conduct the two-step least square procedure in Steps 2-3.

The main merit of using sample splitting is that it makes the proof simple. It facilitates showing some potential bias term is negligible. Specifically, sample splitting artificially generates the independence between the initial estimator for β and the sample used in the first least squares step. This independence makes it simple to show the asymptotic negligibility of some residual term in the least squares step, which includes the shrinkage bias. Moreover, the two-step least square procedure using sample splitting can be used for the bias-correction of other initial estimators as long as the initial estimator has a sufficiently fast convergence rate. Hence, this method is not restricted to the nuclear-norm penalized estimator and it can be used for the debiasing method of diverse estimators for M.

However, sample splitting has several costs. First, sample splitting complicates the estimation procedure by its nature. In addition, sample splitting restricts the type of groups when conducting inference about the group averages. For instance, the inference procedure in Chernozhukov et al. (2019, 2021) can only conduct inference for the cross-sectional average of a fixed period t. It can be quite a strong restriction, especially in the average treatment effect estimation.[1] In addition, the estimator depends on the randomly chosen subsamples, and so, the estimated value is subject to this randomness. For the same target parameter, we may have different estimates depending on how the sample is split. Moreover, sample splitting can be computationally demanding in multiple tests. Chernozhukov et al. (2019, 2021) can estimate entries in only one time period, say t, for each implementation of their estimation procedure. Their procedure requires to re-split the sample if the period of interest t changes because the way of sample splitting varies with the time period of interest t. Therefore, one has to repeat their procedure T times to conduct inference for all time

[1] For example, conducting inference about the time average treatment effect of a certain i, or the cross-sectional average treatment effect of several periods is impossible.

periods. It can be very time-consuming, while only one execution of our procedure is enough for the same goal. Last but not least, because Chernozhukov et al. (2019, 2021) split the sample and use one half for the nuclear-norm penalized estimation and another half for the bias correction using the two-step least squares procedure, we need a relatively larger sample size for the finite sample case. Otherwise, the estimation quality would deteriorate in the finite sample.

Since we skip the sample splitting steps 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 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, namely, the sample in period t.[2] 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 correct the shrinkage bias.

The idea of the leave-one-out estimator has been employed in other recent works such as Chen et al. (2020) as well. We highlight that our leave-one-out estimator is different from theirs. In both Chen et al. (2020) 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, all samples in period t. However, it alone does not assure the independence between the leave-one-out estimator and samples of period t. Since the loss function for the leave-one-out problem is not convex, one cannot iterate until convergence. In fact, the gradient descent iteration must end when 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. Unlike Chen et al. (2020) who derive the stopping point from the problem using the full sample, we find the stopping point from the leave-one-out problem. While this change causes some nontrivial difficulties in the proof, we successfully resolve the problems. We leave more detailed discussions in Section 2.3.2.

Another important contribution of this paper is that our inference procedure allows more general data-observation patterns than the one commonly adopted in the matrix completion litera-

[2] 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 $t \leq T$.

ture and exploits a weighting method in the objective function to incorporate the heterogeneous observation probability. The aforementioned works such as Chen et al. (2019), Xia and Yuan (2021) assume that observation is uniformly at random, i.e., i) $\mathbb{E}[\omega_{it}] = p$ for all i and t, and ii) $\{\omega_{it}\}_{i\leq N,t\leq T}$ are independent across both i and t where $\omega_{it} = 1\{y_{it} \text{ is observed}\}$. These assumptions are quite restrictive in many applications. For example, the homogeneous observation probability assumption cannot accommodate the case where the movie-rating response rates are different across viewers in the online movie-providing platforms such as Netflix. Although generalizing these assumptions is important in applications, there are only a few studies on the nuclear norm penalized estimator that allows heterogeneous observation probabilities and correlated observation patterns. To the best of our knowledge, only Chernozhukov et al. (2019, 2021) consider the inferential theory of the nuclear norm penalized estimation with the generalized observation patterns.

In the paper, we allow the heterogeneous observation probabilities. Besides, we utilize 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)) to incorporate the heterogeneous observation probability. Intuitively, the inverse probability weighting is designed to treat units with different observation probabilities in a fair manner so that the estimation errors of units with high (low) observation probabilities are not factored more (less) into minimizing squared errors. The simulation result in Section 5 shows that the estimation performance of the nuclear norm penalized estimator can be improved by using inverse probability weighting in the presence of the heterogeneous observation probability. Furthermore, we accommodate the correlated observation pattern by assuming the cluster structure. Namely, ω_{it} and ω_{jt} are allowed to be correlated if units i and j are in the same cluster.[3] Compared to Chen et al. (2019), Xia and Yuan (2021) which do not allow any dependence in $\{\omega_{it}\}_{i\leq N,t\leq T}$, our inference procedure would be more relevant to the economic or other social science data.

Moreover, we contribute to the literature in the aspect of economic applications. Lately, economists have begun to utilize the nuclear norm penalized estimation in their research. Moon and Weidner (2018) study the inference for common parameters in a panel data model and estimate the low-rank matrix of the interactive fixed effects by using the nuclear norm penalized estimation. Chernozhukov et al. (2019) study a panel data model with heterogeneous effects where slopes are allowed to vary across both units and periods and estimate the low-rank matrix of slopes using the nuclear norm penalized estimation. Athey et al. (2021) exploit the matrix completion method

[3] The size of clusters can increase as the sample size increases.

to impute the missing potential outcomes for estimating treatment effects and provide rates of convergence for the estimated low-rank matrix. Chernozhukov et al. (2021) propose inferential results for the average treatment effect estimator using the nuclear norm penalized estimation. In addition, Giglio et al. (2020) develop a way to perform multiple testing on the alphas in the empirical asset pricing model, which is robust to missing data by using the nuclear norm penalized estimation.

In the paper, we provide the inferential theory for the average treatment effect estimator as an application.[4] Although Athey et al. (2021) propose a treatment effect estimator using the nuclear norm penalized estimation, they only present the convergence rate of the estimator without the inferential theory. In addition, unlike Chernozhukov et al. (2021) which also provide the inference procedure for the average treatment effect using the nuclear norm penalized estimation, we do not resort to sample splitting, and hence, we can avoid the several drawbacks of sample splitting described above. Especially, we can consider more diverse groups of (i, t) for the average treatment effect while Chernozhukov et al. (2021) can only consider the cross-sectional average of a certain time period t.

Last but not least, as a byproduct of showing the leave-one-out estimator is close to the initial estimate of loadings, this paper generalizes several results in Chen et al. (2020) which study the convergence rates of the nuclear-norm penalized estimator. Specifically, we generalize their results in the sense that i) the data matrices are nonsquare ($N \neq T$), ii) the matrix of interest $M := [M_{it}]_{N\times T}$ is random and consists of the low-rank matrix $M^\star$ with the low-rank approximation error $M^R := [M^R_{it}]_{N\times T}$, and iii) we assume the cross-sectionally correlated $\{\omega_{it}\}_{i\leq N, t\leq T}$ with heterogeneous observation probabilities.

This paper is organized as follows. Section 2 provides the model and the estimation procedure as well as our debiasing strategy. Section 3 gives the asymptotic results of our debiased estimator. Section 4 provides the inferential theory for the average treatment effect estimator as an application. Section 5 shows the simulation studies and Section 6 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 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 $O_p(1)$.

[4] Here, we consider heterogeneous treatment effects.

2 Model and Estimation

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

$$y_{it} = h_t(\zeta_i) + \varepsilon_{it}, \quad \text{if} \quad \omega_{it} = 1,$$

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 $\omega_{it} = 1\{y_{it} \text{ is observed}\}$.[5] 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.[6]

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 $\mathcal{M} := \{\beta, F, M^R\}$ 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 $\Pi = \text{diag}(p_1, \ldots, p_N)$. 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.

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

[6] 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 Remark 1.

2.1 Nuclear norm penalized estimation with inverse probability weighting

One of the commonly used method for the low-rank matrix completion is the nuclear norm penalization. A number of recent works, such as Beck and Teboulle (2009), Cai et al. (2010), Mazumder et al. (2010), Ma et al. (2011), Koltchinskii et al. (2011), Negahban and Wainwright (2011), Chen et al. (2020, 2019), have studied the nuclear norm penalization method for low-rank matrix compeletion and provide algorithms to compute the following convex program:

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

where $\Omega = [\omega_{it}]_{N\times T}$ and $\circ$ denotes the Hadamard product. However, the above papers only consider the case of the homogeneous observation probability. If there is heterogeneity in the observation probability, using the objective function (2.2) may not be optimal to estimate M. In fact, there are many cases where it is more reasonable to assume the heterogeneous observation probability. For instance, as noted earlier, the feedback probabilities can be different across viewers in the online movie-providing platform. Then, the observation probability of movie-rating data will be different across viewers. In this case, if we use the objective function (2.2), the estimation errors of people with low observation probability would be factored less into minimizing squared errors, and it may debase the estimation quality as explained in Section 5.

To avoid such a problem, in the case of the heterogeneous observation probability, we propose to use 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 $\widehat{\Pi} = \mathrm{diag}(\widehat{p}_1,\ldots,\widehat{p}_N)$, and $\widehat{p}_i = \frac{1}{T}\sum_{t=1}^{T}\omega_{it}$ for each $i\le N$.

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.[7] Hence, using the weighted objective function (2.3) is more suitable in the case of the heterogeneous observation probability. Besides, inverse probability weighting enhances the estimation quality by treating

[7] 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).

units equally. The units with high (low) observation probabilities would be factored more (less) into the unweighted objective function (2.2) minimization, resulting larger estimation errors. On the other hand, we may expect more equally distributed errors across units when we use the weighted objective function (2.3) since it compensates the effect of missing by putting the inverse of the observation probability.[8] Indeed, Figure 1 in Section 5 shows that using inverse probability weighting reduces the estimation error when there is heterogeneity in the observation probability.

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 debias it, we run the two-step least squares procedure which is from Chernozhukov et al. (2019, 2021).[9] However, 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 $t \leq T$, 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 $i \leq N$, 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.[10] 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.[11]

[8] Even if we leave other effects of inverse probability weighting aside, this equalization by itself tends to reduce the estimation error when it is measured in matrix norms, which are convex functions. As a simple example, let $A = \begin{bmatrix} 1 & 1 \\ 3 & 3 \end{bmatrix}$ and $B = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}$. Then any matrix norm, which we use in this paper, of A is larger than or equal to that of B.

[9] For a further discussion on the two-step least square method and the rotation debiasing, see Chernozhukov et al. (2019, 2021).

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

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

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 complicate statistical inference. To remove the bias, 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.[12] 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. Namely, they randomly split the full sample into two subsamples and use one subsample to derive $\widetilde{M}$ and $\widetilde{\beta} \coloneqq [\widetilde{\beta}_1, \ldots, \widetilde{\beta}_N]'$, and another subsample to run the least square for debiasing so that $\widetilde{\beta}$ can be independent of $\{\omega_{jt}\varepsilon_{jt}\}_{j \leq N}$. Given this independence, R_t can be easily bounded.

2.3.1 Disadvantages of sample splitting

However, though sample splitting makes the proof simple, it has multiple nontrivial disadvantages. First, sample splitting causes instability. Since the subsamples are randomly chosen in sample splitting, the estimated value is subject to this randomness. For the same target parameter, we have different estimated values depending on how the sample is split. It is well-known that this instability, when the sample is split across time, is substantial in empirical studies. Second, sample splitting strongly restricts the type of group of units and periods when researchers conduct inference for the group average of M_{it}. Specifically, if the observation probability is different across units, the method of Chernozhukov et al. (2019, 2021) can only consider the cross-sectional average of a fixed period t, such as $\frac{1}{|\mathcal{I}|_o} \sum_{i \in \mathcal{I}} M_{it}$ where $\mathcal{I} \subset \{1, \ldots, N\}$, and it is a quite strong constraint in the treatment effect application. In contrast, our method allows inference for more general groups of units and periods as noted in Section 3. Third, sample splitting can be computationally demanding in multiple tests. To make inference of the above cross-sectional average for all periods using the sample splitting method in Chernozhukov et al. (2019, 2021), we need to repeat Algorithm 1 T times because the way of sample splitting varies with the time period of interest. It can be very

[12] 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.

time-consuming, while our method runs Algorithm 1 only once for the same goal.[13] Last but not least, since Chernozhukov et al. (2019, 2021) use only a half of the full sample to compute the nuclear norm penalized estimator and another half to debias it, the sample size is required to be relatively larger. Otherwise, the estimation quality would deteriorate in the finite sample.

2.3.2 An alternative leave-one-out method

Motivated by Chen et al. (2020), 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:

$$\begin{aligned} L^{full}(B,f) &= \frac{1}{2}\|\Pi^{-\frac{1}{2}}\Omega\circ(Bf'-Y)\|_F^2+\frac{\lambda}{2}\|B\|_F^2+\frac{\lambda}{2}\|f\|_F^2 \\ &= \frac{1}{2}\|\Pi^{-\frac{1}{2}}\Omega\circ(Bf'-Y)\|_{F,(-t)}^2+\frac{1}{2}\|\Pi^{-\frac{1}{2}}\Omega\circ(Bf'-Y)\|_{F,t}^2+\frac{\lambda}{2}\|B\|_F^2+\frac{\lambda}{2}\|f\|_F^2 \quad (2.4) \\ L^{(-t)}(B,f) &= \frac{1}{2}\|\Pi^{-\frac{1}{2}}\Omega\circ(Bf'-Y)\|_{F,(-t)}^2+\frac{1}{2}\|Bf'-M^\star\|_{F,t}^2+\frac{\lambda}{2}\|B\|_F^2+\frac{\lambda}{2}\|f\|_F^2. \quad (2.5) \end{aligned}$$

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).[14]

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 important difference in the way of selecting the stopping point between Chen et al. (2020) and this paper. While Chen et al. (2020) derive the stopping point from the full sample problem (2.4), we derive the stopping point from

[13] Specifically, the reason for this is as follows. In the case of sample splitting in Chernozhukov et al. (2019, 2021), from one time execution of Algorithm 1 using matrix form, we can only estimate $\{M_{it}\}_{1\leq i\leq N}$ for one fixed t. On the other hand, in our method, we can estimate $\{M_{it}\}_{1\leq i\leq N,1\leq t\leq T}$ from one time execution of Algorithm 1 using matrix form.

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

the leave-one-out problem (2.5). Hence, it ensures that the estimator $\breve{\beta}^{(-t)}$ using this stopping point is truly independent of $\{\omega_{jt}, \varepsilon_{jt}\}_{j\leq N}$.

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 of it is derived from the full sample problem (2.4). While Chen et al. (2020) do not explicitly argue that the estimator derived from the leave-one-out problem (2.5) using their stopping point is independent of $\{\omega_{jt}, \varepsilon_{jt}\}_{j\leq N}$, for inference purposes, we require our leave-one-out estimator $\breve{\beta}^{(-t)}$ is independent of $\{\omega_{jt}, \varepsilon_{jt}\}_{j\leq N}$. Although this change causes some nontrivial difficulties in the proof, we successfully resolve the problems.

Being equipped with these two auxiliary non-convex estimators, we can bound R_t in 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. $\|b\| = o_P(1)$ can be shown using the independence between $\breve{\beta}^{(-t)}$ and $\{\omega_{jt}\varepsilon_{jt}\}_{j\leq N}$, which is along the same line as sample splitting.

3. In addition, $\|a\| = o_P(1)$ 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.

   (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.

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

2.3.3 Remaining bias absorbed by the rotation matrix

Once the asymptotic negligibility of R_t is guaranteed, we can proceed to show that the least square estimator of Step 3 of Algorithm 1, $\widehat{F}_t$, can estimate F_t unbiasedly up to a rotation. This is an interesting result because the regressor of the least square, $\widetilde{\beta}$, suffers from the shrinkage bias caused by the nuclear norm penalization.

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

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

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,

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

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 are negligible, 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 $H_2 := (I_K + \varphi)H_1^{-1}$. 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:

$$\begin{aligned}
\widehat{M}_{it} = \widehat{\beta}_i'\widehat{F}_t &= \beta_i'H_2^{-1}H_2F_t + \text{asymptotically normal term} + \text{higher order terms} \\
&= M_{it} + \text{asymptotically normal term} + \text{higher order terms},
\end{aligned}$$

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. For more discussion on the two-step least square debiasing, see Chernozhukov et al. (2019, 2021).

2.4 Choosing regularization parameter

In practice, we need to choose the regularization parameter λ. Following the idea in Chernozhukov et al. (2019), we use the condition for λ below

$$\|\widehat{\Pi}^{-1}\Omega \circ \mathcal{E}\| < \frac{7}{8}\lambda$$

which is introduced in Condition C.1 of the supplement. In the Gaussian case, we can compute the proper tuning parameter via simulation. Assume that $\varepsilon_{it} \sim \mathcal{N}(0, \sigma^2)$ and we generate the $N \times T$ matrix $\mathcal{U}$ whose elements u_{it} comes from i.i.d. $\mathcal{N}(0, \sigma^2)$ distribution. Then, $\|\widehat{\Pi}^{-1}\Omega \circ \mathcal{E}\|$ and $\|\widehat{\Pi}^{-1}\Omega \circ \mathcal{U}\|$ are identically distributed. Let $\bar{Q}(A; b)$ denote the b^{th} quantile of a random variable A. For $\delta_{NT} = o(1)$, we take

$$\lambda = (1 + c_1)\bar{Q}(\|\widehat{\Pi}^{-1}\Omega \circ \mathcal{E}\|; 1 - \delta_{NT}).$$

Then, we have

$$\|\widehat{\Pi}^{-1}\Omega \circ \mathcal{E}\| < \left(1 - \frac{c_1}{1 + c_1}\right)\lambda$$

with probability $1 - \delta_{NT}$. In the simulation study, we set $c_1 = 1/7$ and $\delta_{NT} = 0.05$.

However, to utilize the above method, we need the estimates of σ^2. We first estimate σ^2 using a more homogeneous model $y_{it} = m_t + \sigma^{-1}\epsilon_{it}$ (or $y_{it} = m + \sigma^{-1}\epsilon_{it}$) with $\text{Var}(\epsilon_{it}) = 1$. Here, m_t and m work as the homogeneous counterparts of M_{it}. By using this initial estimator of σ^2, we generate $\mathcal{U}$ and set the tuning parameter λ as above. Let $\widetilde{M}_{\text{init}}$ be the nuclear norm penalized estimator using this tuning parameter λ. By using $\tilde{\varepsilon}_{it} = y_{it} - \widetilde{M}_{\text{init}}$, we re-estimate σ^2 and update the value of the tuning parameter by re-simulating with the updated estimator of σ^2. We iterate this process until it converges.

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

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

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

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

Assumptions 3.1 (iii) can be satisfied if the basis is a bounded basis like trigonometic 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 $\mathbb{E}[\varepsilon_{it}|\mathcal{M}] = 0$, $\mathbb{E}[\varepsilon_{it}^2|\mathcal{M}] = \sigma^2$, and*

$$\mathbb{E}[\exp(s\varepsilon_{it})|\mathcal{M}] \leq \exp(Cs^2\sigma^2), \quad \forall s \in \mathbb{R},$$

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

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

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 ω_{it} is assumed, we allow the weak cross-sectional dependence of it. 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 $\cup_{g=1}^{\rho} \mathcal{C}_g = \{1, \ldots, N\}$.

Assumption 3.3 (Cross-sectional Dependence in ω_{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 $\mathcal{C}_{g(i)}$, $\{\omega_{j_1 t}, \ldots, \omega_{j_m t}\}$ is independent from ω_{it}*

for all t. In addition, 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, T increase.
(ii) We have $\max_t \max_i \sum_{j=1}^N |Cov(\omega_{it}, \omega_{jt}|\mathcal{M})| < C$.

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 is in the same cluster as unit i and replace their rows of the goodness of fit part by their conditional expectations.[15] 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).

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. Denoting the group of interest as $\mathcal{G} = \mathcal{I} \times \mathcal{T}$ where $\mathcal{I} \subset \{1, \ldots, N\}$ and $\mathcal{T} \subset \{1, \ldots, T\}$, we impose the following assumption on the rates of ϑ, K and $\min\{|\mathcal{I}|_o, |\mathcal{T}|_o\}$.

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

$$
\begin{aligned}
&(i) \quad \min\{|\mathcal{I}|_o^{\frac{1}{2}}, |\mathcal{T}|_o^{\frac{1}{2}}\} K^{\frac{7}{2}} \max\{\sqrt{N}\log^2 N, \sqrt{T}\log^2 T\} = o(\min\{N, T\}), \\
&(ii) \quad \min\{|\mathcal{I}|_o^{\frac{1}{2}}, |\mathcal{T}|_o^{\frac{1}{2}}\} \vartheta K^{\frac{7}{2}} \max\{N\sqrt{\log N}, T\sqrt{\log T}\} = o(\min\{N^{\frac{3}{2}}, T^{\frac{3}{2}}\}), \\
&(iii) \quad \min\{|\mathcal{I}|_o, |\mathcal{T}|_o\} \max\{N, T\} = o(K^{2\nu-3}).
\end{aligned}
$$

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, 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*

$$\psi_r - \psi_{r+1} \geq c\psi_K, \quad r = 1, \ldots, K,$$

where ψ_r is the r-th singular value of $M^\star$.

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

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* $\widehat{\varepsilon}_{it} = y_{it} - \widehat{\beta}_i'\widehat{F}_t$.

Remark 1. Although we consider the nonparametric panel model in the paper, our inferential theory can cover other approximate factor models having the form (1.1) also. We present the assumptions for the asymptotic normality of the estimator for the general approximated factor

model in Section A.3 of the supplement. Please refer to this section for details.

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 $\iota \in \{0, 1\}$,

$$y_{it}^{(\iota)} = M_{it}^{(\iota)} + \varepsilon_{it} = h_t^{(\iota)}(\zeta_i) + \varepsilon_{it}, \quad \text{if} \quad \omega_{it}^{(\iota)} = 1,$$

where $\omega_{it}^{(\iota)} = 1\{y_{it}^{(\iota)} \text{ is observed}\}$, ε_{it} is the noise and 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 $h_t^{(0)}(\cdot)$. Here, $\omega_{it}^{(1)} = \Upsilon_{it}$ 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 denote the individual treatment effect by $\Gamma_{it} = M_{it}^{(1)} - M_{it}^{(0)}$ and its estimator 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. Then, the average treatment effect for the group $\mathcal{G}$ can be represented as $\frac{1}{|\mathcal{G}|_o} \sum_{(i,t)\in\mathcal{G}} \Gamma_{it}$ and its estimator will be $\frac{1}{|\mathcal{G}|_o} \sum_{(i,t)\in\mathcal{G}} \widehat{\Gamma}_{it}$. 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 basically the same as those in Section 2, and we just put the superscript (ι) to all notations to distinguish the pair of potential realizations. 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.

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 $\iota \in \{0,1\}$, $\{h_t^{(\iota)}(\cdot)\}_{t\leq T}$ 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 $\nu > 0$, $\max_{i,t} |M_{it}^{R(\iota)}| \leq CK^{-\nu}$.
(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,

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

(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*

$$\mathbb{E}[\exp(s\varepsilon_{it})|\zeta] \leq \exp(Cs^2\sigma^2), \quad \forall s \in \mathbb{R},$$

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.[16] *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 one 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 $\mathbb{E}[a_t] = 0$; more specifically, there is

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

$C > 0$ such that

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

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 $\mathcal{C}_{g(i)}$, $\{\Upsilon_{j_1 t}, \ldots, \Upsilon_{j_m t}\}$ is independent from Υ_{it} for all t. In addition, there is $\vartheta \ge 1$ such that the maximum number of elements in one cluster is bounded by ϑ. That is, $\max_g |\mathcal{C}_g|_o \le \vartheta$. Here, ϑ is allowed to increase as N, 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 (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) $\min\{|\mathcal{I}|_o, |\mathcal{T}|_o\} \max\{N, T\} = o(K^{2\nu-3})$.

Assumption 4.5 (Eigengap). *There are $c, C > 0$ such that with probability converging to 1, for all $\iota \in \{0, 1\}$, $\psi_1^{(\iota)} \le C\psi_K^{(\iota)}$ and*

$$\psi_r^{(\iota)} - \psi_{r+1}^{(\iota)} \ge c\psi_K^{(\iota)}, \quad r = 1, \ldots, K,$$

where $\psi_r^{(\iota)}$ is the 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 4.1 - 4.5 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 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.

First of all, to check the estimation quality of our estimator, we compare the Frobenius norms of the estimation errors for several existing estimators of M. Here, we consider the inverse probability weighting method (e.g., Xiong and Pelger (2020)),[17] the EM algorithm method (e.g., Jin et al. (2021)), and the nuclear norm regularized estimator, in addition to our two-step least squares (TLS) debiased estimator. 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 $\beta_{1,i}, F_{1,t}, \beta_{2,i}, F_{2,t} \sim \mathcal{N}\left(\frac{1}{\sqrt{2}}, 1\right)$, (5.1)
- 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 $h_t(\zeta) = h_t^{sine}(\zeta) := \sum_{r=1}^{\infty} \frac{|U_{t,r}|}{r^3} \sin(r\zeta)$.

[17] 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).

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 1: Frobenius norm of estimation errors for estimators of M

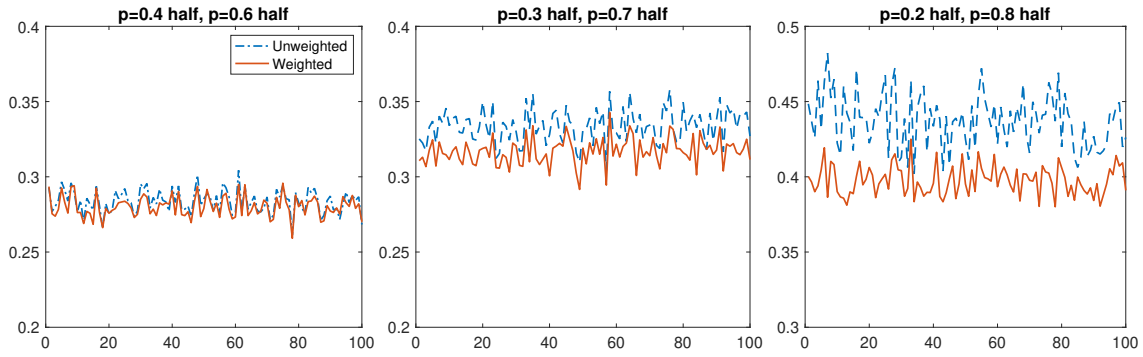
Sample size	N = 200, T =200			N = 200, T = 100			N = 100, T =200		
Model	Factor	Sine	Poly	Factor	Sine	Poly	Factor	Sine	Poly
Regularized (UW)	0.4108	0.2805	0.2789	0.4990	0.3353	0.3342	0.4998	0.3384	0.3406
Regularized (W)	0.3982	0.2780	0.2763	0.4843	0.3318	0.3324	0.4908	0.3380	0.3388
IPW	0.3776	0.1694	0.1692	0.4990	0.2154	0.2172	0.4039	0.2007	0.2013
EM algorithm	0.2052	0.1504	0.1465	0.2574	0.1833	0.1813	0.2541	0.1813	0.1788
TLS debiasing	0.2054	0.1503	0.1464	0.2577	0.1832	0.1812	0.2542	0.1811	0.1786

NOTE: "Regularized (UW)" refers to the unweighted nuclear norm regularized estimator, "Regularized (W)" means the weighted nuclear norm regularized estimator, "IPW" denotes the inverse probability weighting method, and "TLS debiasing" denotes our two-step least squares debiased estimator. In addition, "Sine" and "Poly" refer to the functions $h_t^{sine}(\zeta)$ and $h_t^{poly}(\zeta)$, respectively.

Table 1 reports $||\widehat{M} - M||_F/\sqrt{NT}$ averaged over 100 replications. In all scenarios, our TLS debiasing method and the EM algorithm method show the best results. Especially, the difference between these two methods (TLS, EM) and the other three methods (ReUW, ReW, IPW) are quite large in the factor model. If we compare our TLS debiasing method with the EM algorithm method, in the nonparametric models, our TLS debiasing method performs slightly better than the EM algorithm method, while the EM algorithm method is slightly better than our method in the factor model.

Second, to see the relative advantage of the weighted nuclear norm regularized estimator over the unweighted nuclear norm regularized estimator clearly, we compare the Frobenius norms of their estimation errors using diverse degree of heterogeneity in p_i. Here, we consider the first nonparametric model in (5.1) with the following three cases: (i) Half of the p_i is 0.6 and another half is 0.4, (ii) Half of the p_i is 0.7 and another half is 0.3, (iii) Half of the p_i is 0.8 and another half is 0.2. Figure 1 shows that the weighted nuclear norm regularized estimator performs better than the unweighted nuclear norm regularized estimator when there is heterogeneity in p_i. In addition, it reveals that the larger the degree of heterogeneity in p_i is, the better the relative performance of the weighted nuclear norm regularized estimator is. Hence, it is recommended to use the weighted nuclear norm regularized estimator, if there is heterogeneity in p_i.

Figure 1: Frobenius norm of estimation errors, $||\widetilde{M} - M||_F/\sqrt{NT}$



NOTE: The sample size is $N = T = 200$ and the number of simulation is set to 100. "Unweighted" refers to the unweighted nuclear norm regularized estimator and "Weighted" denotes the weighted nuclear norm regularized estimator using inverse probability weighting.

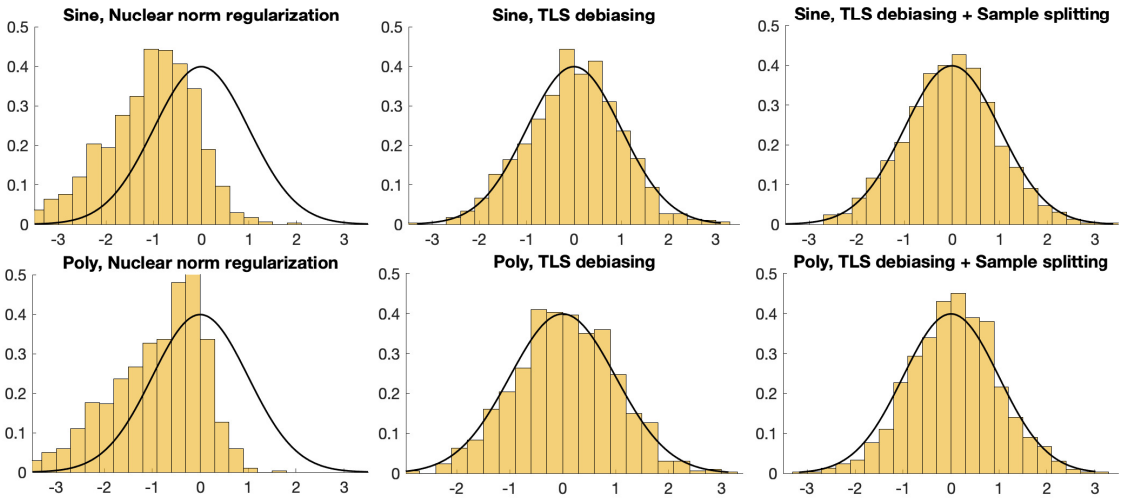
Third, we study the finite sample distributions for standardized estimates defined as

$$\frac{\widehat{M}_{it} - M_{it}}{se(\widehat{M}_{it})}.$$

For comparison, we report the results of the nuclear norm regularized estimator and the two-step least squares (TLS) debiased estimator with sample splitting, in addition to our TLS debiased estimator which does not utilize the sample splitting method. Here, the TLS debiased estimator means the debiased estimator using the TLS procedure. 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 debiased estimator with sample splitting, we construct the standard error following Chernozhukov et al. (2019). Here, we consider the nonparametric models in (5.1). Hereinafter, the number of simulations is set to 1,000.

Figure 2 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 debiased 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 debiased estimator "with sample splitting" and "without sample splitting". Without sample splitting, the TLS debiased estimator itself provides a good approximation to the standard normal distribution so that it can be used for the inference successfully.

Figure 2: 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 debiasing" denotes our TLS debiased estimator which does not use the sample splitting method. "TLS debiasing + Sample splitting" refers to the TLS debiased estimator with sample splitting. In addition, "Sine" and "Poly" refer to the functions $h_t^{sine}(\zeta)$ and $h_t^{poly}(\zeta)$, respectively.

Table 2: Coverage Probabilities of the Confidence Interval for M_{it}

Target Probability	Function, $h_t(\zeta)$	Sample size N	Sample size T	Regularized	TLS debiasing	TLS + SS
95%	Sine	150	150	79.2%	96.0%	96.5%
		200	200	80.9%	95.0%	95.5%
	Poly	150	150	83.0%	96.5%	96.4%
		200	200	83.4%	96.1%	96.2%
90%	Sine	150	150	68.9%	91.4%	91.2%
		200	200	72.6%	90.6%	90.7%
	Poly	150	150	74.5%	91.4%	92.8%
		200	200	77.8%	92.1%	91.6%

NOTE: "Regularized" refers to the weighted nuclear norm regularized estimator and "TLS debiasing" denotes our TLS debiased estimator which does not use the sample splitting method. "TLS + SS" refers to the TLS debiased estimator with sample splitting.

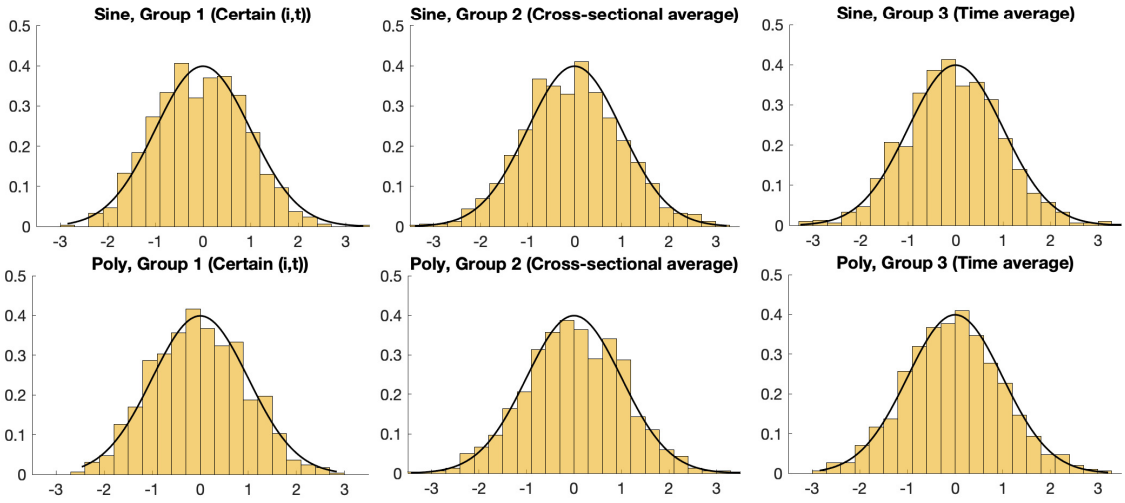
Table 2 presents the coverage probabilities of the confidence interval which is given by

$$[\widehat{M}_{it} - cv \cdot se(\widehat{M}_{it}), \widehat{M}_{it} + cv \cdot se(\widehat{M}_{it})]$$

where $cv = 1.645$ for the 90% confidence interval and $cv = 1.96$ for the 95% confidence interval. The result is similar to the above. The coverage probabilities of the TLS debiased estimators are close to

the target coverage probabilities, while those of the nuclear norm estimator are largely different from the target probabilities. There is no big difference in the coverage probabilities between the TLS debiased estimator "with sample splitting" and "without sample splitting", although the coverage probabilities of the TLS debiased estimators without sample splitting are slightly closer to the target probabilities compared to those with sample splitting in many cases.

Figure 3: Histograms of standardized estimates, $\frac{\frac{1}{|\mathcal{G}|_o}\sum_{(i,t)\in\mathcal{G}}\widehat{M}_{it}-\frac{1}{|\mathcal{G}|_o}\sum_{(i,t)\in\mathcal{G}}M_{it}}{se\left(\frac{1}{|\mathcal{G}|_o}\sum_{(i,t)\in\mathcal{G}}\widehat{M}_{it}\right)}$



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 $\mathcal{G}_3$.

In addition, to show that our asymptotic theory works well with various groups, Figure 3 and Table 3 present the scaled histograms of the standardized estimates of our TLS debiased estimators (which does not use the sample splitting method) and the coverage probabilities of the 95% confidence interval respectively with various groups. We generate the data using the same model as in the above. For the group, we consider the cross-sectional average of a certain t, the time average of a certain i, in addition to the certain (i,t). Specifically, we consider $\mathcal{G}_1 = \{(i,t)\}$, $\mathcal{G}_2 = \{(j,s) : 1 \leq j \leq N, s = t\}$ and $\mathcal{G}_3 = \{(j,s) : j = i, 1 \leq s \leq T\}$. We choose i and t randomly and fix them in the simulation replications. Here, the standard estimates are defined as

$$\frac{\frac{1}{|\mathcal{G}|_o}\sum_{(i,t)\in\mathcal{G}}\widehat{M}_{it}-\frac{1}{|\mathcal{G}|_o}\sum_{(i,t)\in\mathcal{G}}M_{it}}{se\left(\frac{1}{|\mathcal{G}|_o}\sum_{(i,t)\in\mathcal{G}}\widehat{M}_{it}\right)}$$

and the 95% confidence interval is given by

$$\left[\frac{1}{|\mathcal{G}|_o}\sum_{(i,t)\in\mathcal{G}}\widehat{M}_{it}-1.96se\left(\frac{1}{|\mathcal{G}|_o}\sum_{(i,t)\in\mathcal{G}}\widehat{M}_{it}\right),\frac{1}{|\mathcal{G}|_o}\sum_{(i,t)\in\mathcal{G}}\widehat{M}_{it}+1.96se\left(\frac{1}{|\mathcal{G}|_o}\sum_{(i,t)\in\mathcal{G}}\widehat{M}_{it}\right)\right].$$

Table 3: Coverage Probabilities of the Confidence Interval for $\frac{1}{|\mathcal{G}|_o}\sum_{(i,t)\in\mathcal{G}}M_{it}$

Function	Sample size		Group		
$h_t(\zeta)$	N	T	$\mathcal{G}_1$ (Certain (i,t))	$\mathcal{G}_2$ (Cross-sectional average)	$\mathcal{G}_3$ (Time average)
Poly	200	200	96.6%	92.5%	95.9%
	250	250	96.5%	93.4%	93.8%
	300	300	96.0%	93.9%	94.5%
Sine	200	200	95.5%	92.9%	95.7%
	250	250	96.5%	92.3%	93.0%
	300	300	97.0%	94.2%	94.5%

Figure 3 and Table 3 reveal that the standardized estimates of our TLS debiased 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

$$y_{it}^{(\iota)}=h_t^{(\iota)}(\zeta_i)+\varepsilon_{it},\quad \text{if } \Upsilon_{it}=\iota,$$

where

$$h_t^{(0)}(\zeta)=\sum_{r=1}^{\infty}\frac{|U_{t,r}|}{r^a}\sin(r\zeta),\quad h_t^{(1)}(\zeta)=\sum_{r=1}^{\infty}\frac{|U_{t,r}|+2}{r^a}\sin(r\zeta).$$

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 4.

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 5 presents the scaled histograms of the standardized estimates of the average 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

$$\frac{\frac{1}{|\mathcal{G}|_o}\sum_{(i,t)\in\mathcal{G}}\widehat{\Gamma}_{it}-\frac{1}{|\mathcal{G}|_o}\sum_{(i,t)\in\mathcal{G}}\Gamma_{it}}{se\left(\frac{1}{|\mathcal{G}|_o}\sum_{(i,t)\in\mathcal{G}}\widehat{\Gamma}_{it}\right)}.$$

Figure 4: Shape of function $h_t^{(\iota)}(\zeta)$ and treatment effect function ($U_{t,r} = 1$, $a = 2$)

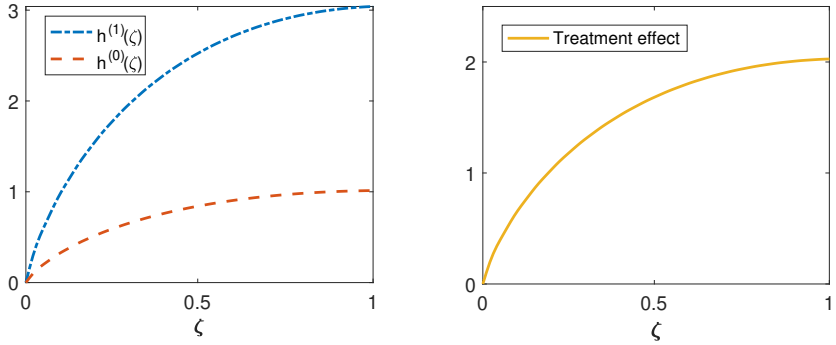
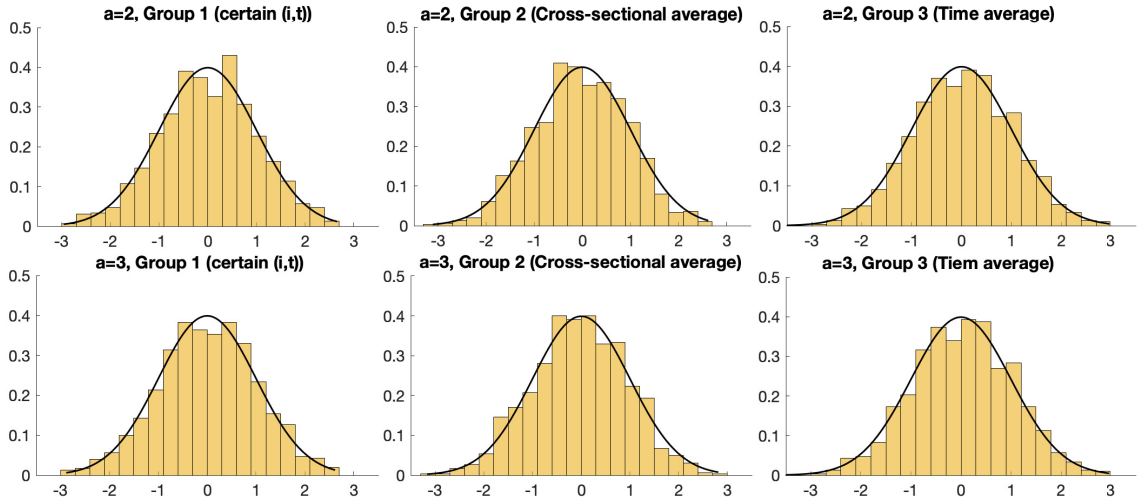


Figure 5: Histograms of standardized estimates, $\frac{\frac{1}{|\mathcal{G}|_o}\sum_{(i,t)\in\mathcal{G}}\widehat{\Gamma}_{it}-\frac{1}{|\mathcal{G}|_o}\sum_{(i,t)\in\mathcal{G}}\Gamma_{it}}{se\left(\frac{1}{|\mathcal{G}|_o}\sum_{(i,t)\in\mathcal{G}}\widehat{\Gamma}_{it}\right)}$



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 $\mathcal{G}_3$.

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, Table 4 provides the coverage probabilities of the 95% confidence interval defined in

$$\left[\frac{1}{|\mathcal{G}|_o}\sum_{(i,t)\in\mathcal{G}}\widehat{\Gamma}_{it}-1.96se\left(\frac{1}{|\mathcal{G}|_o}\sum_{(i,t)\in\mathcal{G}}\widehat{\Gamma}_{it}\right),\frac{1}{|\mathcal{G}|_o}\sum_{(i,t)\in\mathcal{G}}\widehat{\Gamma}_{it}+1.96se\left(\frac{1}{|\mathcal{G}|_o}\sum_{(i,t)\in\mathcal{G}}\widehat{\Gamma}_{it}\right)\right].$$

It also reveals that the coverage probabilities are quite close to the target probability 95% in all cases. Overall, the results are quite good, and it seems that our asymptotic theory for inference works well.

Table 4: Coverage Probabilities of the Confidence Interval for $\frac{1}{|\mathcal{G}|_o}\sum_{(i,t)\in\mathcal{G}}\Gamma_{it}$

Power	Sample size		Group		
a	N	T	$\mathcal{G}_1$ (Certain (i,t))	$\mathcal{G}_2$ (Cross-sectional average)	$\mathcal{G}_3$ (Time average)
2	200	200	95.3%	95.9%	96.1%
	300	300	94.8%	96.1%	94.9%
3	200	200	95.7%	95.8%	95.7%
	300	300	95.1%	94.1%	96.0%

6 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 which 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.[18] Following the model in Section 4, we set the treatment indicator as $\Upsilon_{it} =$

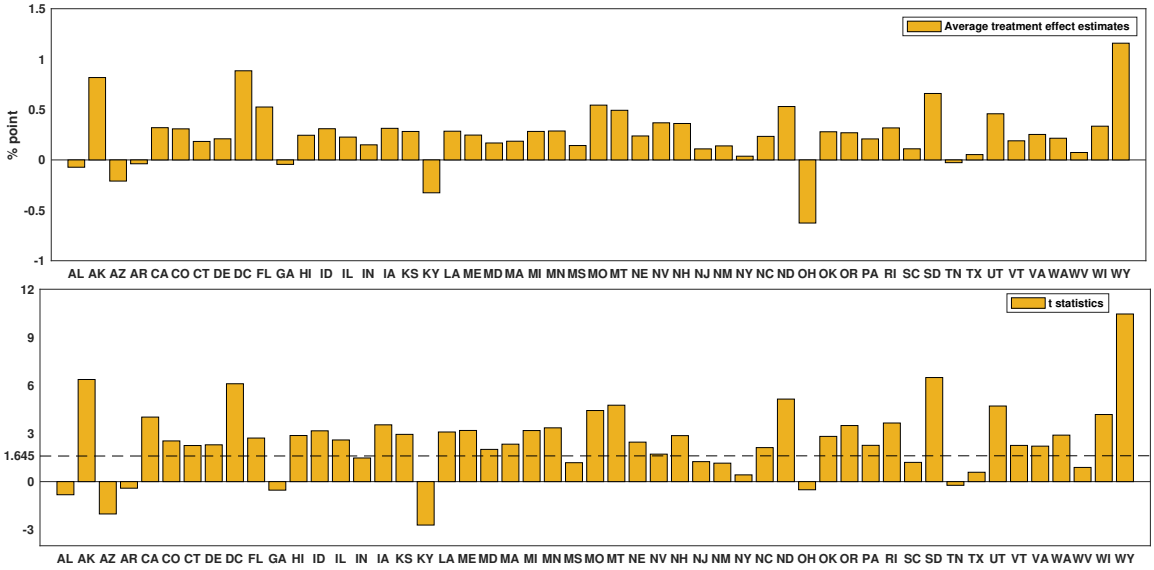
[18] We get the data from the U.S. Census Bureau, NASBO (National Association of State Budget Officers), and SSA (Social

$1\{$the state i supported the president of year t 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.

Figure 6: Time average treatment effect estimates of each state and corresponding t-statistics



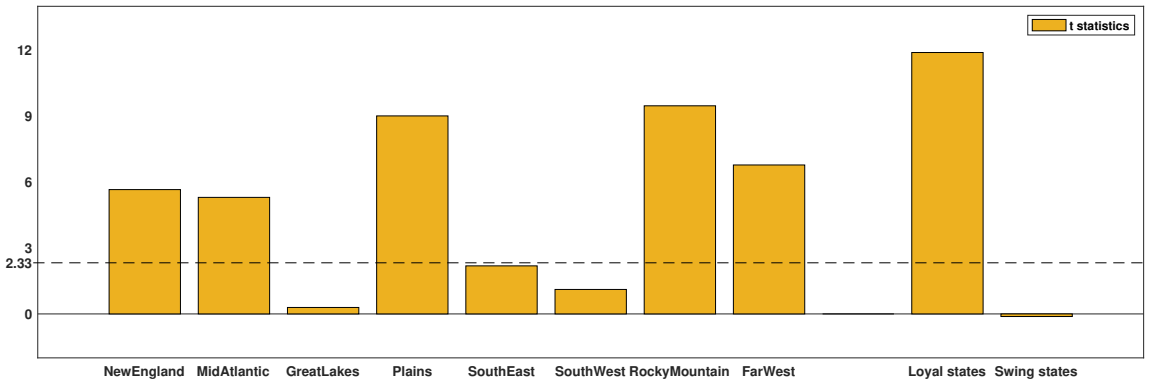
NOTE: When we use the B-H procedure to control the size of FDR at 5%, the list of states with rejected decisions is unchanged.

First, we study the time average of the treatment effect of each state. Here, we consider the time average of all periods (1953 - 2018). Figure 6 presents the estimates of the average treatment effect and the corresponding t-statistics. The first graph shows there is a positive treatment effect in most states and it can be seen as evidence that incumbent presidents tend to reward states that showed

Security Administration). Because of absence of data, the years, 1960, 1976∼1979, are excluded.

their support in elections. Alaska, D.C., South Dakota, and Wyoming show the largest treatment effects. Compared to the situation when the incumbent president is not the candidate whom the states supported in the latest presidential election, these states' portions of federal funds increase by 0.81, 0.88, 0.65, 1.15 percent points, respectively, if the incumbent president is the candidate whom the states supported. In addition, the second graph which presents the corresponding t-statistics shows that the result of existences of positive treatment effects in most states are statistically significant.

Figure 7: 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.

Table 5: Average number of times states in regions swung its support from a party to another

Region	NewEngland	MidAtlantic	Plains	RockyMountain	FarWest	GreatLakes	SouthEast	SouthWest
	3.7	4	3.3	3	3.5	5	6	3.5

In addition, Figure 7 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. This result may be related to the loyalty of states to parties. We generate an indicator of long-term swing which is based on the number of times a state swung its support from a party to another in the presidential elections from 1952 to 2016 and Table 5 reports the average of this indicator for each region. From the table, we can check that regions having statistically significant positive treatment effects usually have low number of swings. To check whether presidents reward states having loyalty rather than swing states, we make two groups (loyal states, swing states): states in "loyal states" have low number of swings (≤ 2) and states in "swing states" have high number of

swings (≥ 7),[19] and conduct tests for the average treatment effect of each group. As we can see in Figure 7, the swing states do not have statistically significant positive treatment effects while the 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 8: Test statistics for the average treatment effect of each president

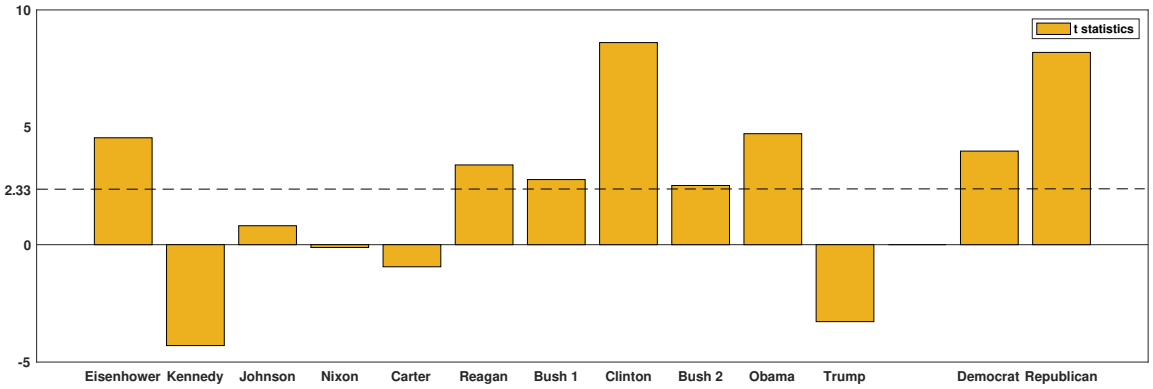


Figure 9: Test statistics for the average treatment effect before 1980 and after 1981

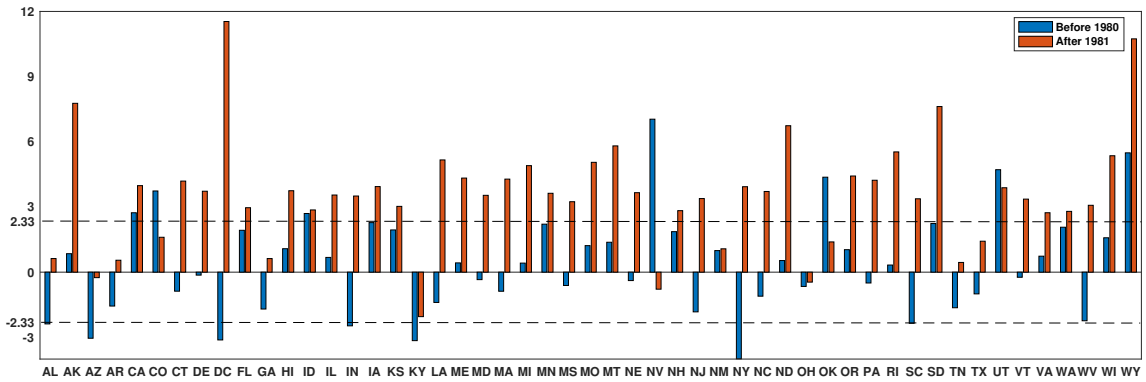


Figure 8 shows the test statistics for the average of the treatment effect of each president. Although there exist some exceptions, there are no statistically significant positive treatment effects before Carter, while there are significant positive treatment effects after Reagan. From Figure 9, we can 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, we can know that there is

[19] "Loyal states" include AK, D.C., ID, KS, NE, ND, OK, SD, UT, WY and "Swing states" include AR, FL, GA, KY, LA, OH, WV.

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 presidents 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.[20]

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.

7 Conclusion

This paper studies the inferential theory for the (debiased) nuclear norm penalized estimator of the latent approximate low-rank matrix when the observation matrix is subject to missing and provides an inference method for the average treatment effect as an application. Without the aid of sample splitting, our debiasing procedure successfully removes the shrinkage bias, and the debiased estimator attains the asymptotic normality. Unlike Chernozhukov et al. (2019, 2021) which exploit sample splitting to remove the bias, 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.

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

References

ABBE, E., J. FAN, K. WANG, AND Y. ZHONG (2020): "Entrywise eigenvector analysis of random matrices with low expected rank," *Annals of statistics*, 48, 1452.

ANDERSON, G. M. AND R. D. TOLLISON (1991): "Congressional influence and patterns of New Deal spending, 1933-1939," *The Journal of Law and Economics*, 34, 161–175.

ATHEY, S., M. BAYATI, N. DOUDCHENKO, G. IMBENS, AND K. KHOSRAVI (2021): "Matrix completion methods for causal panel data models," *Journal of the American Statistical Association*, 1–15.

BECK, A. AND M. TEBOULLE (2009): "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM journal on imaging sciences*, 2, 183–202.

BERRY, C. R., B. C. BURDEN, AND W. G. HOWELL (2010): "The president and the distribution of federal spending," *American Political Science Review*, 104, 783–799.

CAI, J.-F., E. J. CANDÈS, AND Z. SHEN (2010): "A singular value thresholding algorithm for matrix completion," *SIAM Journal on optimization*, 20, 1956–1982.

CANDES, E. J. AND Y. PLAN (2010): "Matrix completion with noise," *Proceedings of the IEEE*, 98, 925–936.

CANDÈS, E. J. AND B. RECHT (2009): "Exact matrix completion via convex optimization," *Foundations of Computational mathematics*, 9, 717.

CHEN, Y. AND Y. CHI (2018): "Harnessing structures in big data via guaranteed low-rank matrix estimation: Recent theory and fast algorithms via convex and nonconvex optimization," *IEEE Signal Processing Magazine*, 35, 14–31.

CHEN, Y., Y. CHI, J. FAN, C. MA, AND Y. YAN (2020): "Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization," *SIAM journal on optimization*, 30, 3098–3121.

CHEN, Y., J. FAN, C. MA, AND Y. YAN (2019): "Inference and uncertainty quantification for noisy matrix completion," *Proceedings of the National Academy of Sciences*, 116, 22931–22937.

CHERNOZHUKOV, V., C. HANSEN, Y. LIAO, AND Y. ZHU (2021): "Inference for low-rank models," *arXiv preprint arXiv:2107.02602*.

CHERNOZHUKOV, V., C. B. HANSEN, Y. LIAO, AND Y. ZHU (2019): "Inference for heterogeneous effects using low-rank estimations," Tech. rep., cemmap working paper.

COX, G. W. AND M. D. MCCUBBINS (1986): "Electoral politics as a redistributive game," *The Journal of Politics*, 48, 370–389.

FAN, J., K. LI, AND Y. LIAO (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*.

GIGLIO, S., Y. LIAO, AND D. XIU (2020): "Thousands of Alpha Tests," *The Review of Financial Studies*.

IMBENS, G. W. AND D. B. RUBIN (2015): *Causal inference in statistics, social, and biomedical sciences*, Cambridge University Press.

JIN, S., K. MIAO, AND L. SU (2021): "On factor models with random missing: EM estimation, inference, and cross validation," *Journal of Econometrics*, 222, 745–777.

KOLTCHINSKII, V., K. LOUNICI, A. B. TSYBAKOV, ET AL. (2011): "Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion," *The Annals of Statistics*, 39, 2302–2329.

LARCINESE, V., L. RIZZO, AND C. TESTA (2006): "Allocating the US federal budget to the states: The impact of the president," *The Journal of Politics*, 68, 447–456.

LITTLE, R. J. AND D. B. RUBIN (2019): *Statistical analysis with missing data*, vol. 793, John Wiley & Sons.

LIU, Z. AND L. VANDENBERGHE (2010): "Interior-point method for nuclear norm approximation with application to system identification," *SIAM Journal on Matrix Analysis and Applications*, 31, 1235–1256.

LUO, Z.-Q., W.-K. MA, A. M.-C. SO, Y. YE, AND S. ZHANG (2010): "Semidefinite relaxation of quadratic optimization problems," *IEEE Signal Processing Magazine*, 27, 20–34.

MA, S., D. GOLDFARB, AND L. CHEN (2011): "Fixed point and Bregman iterative methods for matrix rank minimization," *Mathematical Programming*, 128, 321–353.

MA, W. AND G. H. CHEN (2019): "Missing Not at Random in Matrix Completion: The Effectiveness of Estimating Missingness Probabilities Under a Low Nuclear Norm Assumption," in *Advances in Neural Information Processing Systems*, 14900–14909.

MAZUMDER, R., T. HASTIE, AND R. TIBSHIRANI (2010): "Spectral regularization algorithms for learning large incomplete matrices," *Journal of machine learning research*, 11, 2287–2322.

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

MOON, H. R. AND M. WEIDNER (2018): "Nuclear norm regularized estimation of panel regression models," *arXiv preprint arXiv:1810.10987*.

NEGAHBAN, S. AND M. J. WAINWRIGHT (2011): "Estimation of (near) low-rank matrices with noise and high-dimensional scaling," *The Annals of Statistics*, 1069–1097.

——— (2012): "Restricted strong convexity and weighted matrix completion: Optimal bounds with noise," *The Journal of Machine Learning Research*, 13, 1665–1697.

PARIKH, N. AND S. BOYD (2014): "Proximal algorithms," *Foundations and Trends in optimization*, 1, 127–239.

RENNIE, J. D. AND N. SREBRO (2005): "Fast maximum margin matrix factorization for collaborative prediction," in *Proceedings of the 22nd international conference on Machine learning*, 713–719.

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

SCHNABEL, T., A. SWAMINATHAN, A. SINGH, N. CHANDAK, AND T. JOACHIMS (2016): "Recommendations as Treatments: Debiasing Learning and Evaluation," in *International Conference on Machine Learning*, 1670–1679.

XIA, D. AND M. YUAN (2021): "Statistical inferences of linear forms for noisy matrix completion," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83, 58–77.

XIONG, R. AND M. PELGER (2020): "Large dimensional latent factor modeling with missing observations and applications to causal inference. arXiv eprint," *arXiv preprint arXiv:1910.08273*.

ZHANG, T., J. M. PAULY, AND I. R. LEVESQUE (2015): "Accelerating parameter mapping with a locally low rank constraint," *Magnetic resonance in medicine*, 73, 655–661.