Inference for Low-rank Estimation with Applications

Jungjun Choi Hyukjun Kwon

October 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 the alpha test in empirical asset pricing, which is robust to missing, and the average treatment effect estimator as applications. 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 heterogeneous observation probabilities and uses inverse probability weighting, which improves the estimation performance by treating units with different observation probabilities in an equal manner. We illustrate the proposed method in simulation experiments and the empirical study about the impact of the president on allocating the U.S. federal budget to the states.

1 Introduction

Consider the following approximate factor model subject to missing data problem:

$$z_{it} = M_{it} + \varepsilon_{it} = M^{\star}_{it} + M^{R}_{it} + \varepsilon_{it}, \quad \text{if } \omega_{it} = 1, \tag{1.1}$$

where z_{it} is the outcome for a unit i in a period t, M_{it} is the variable of interest, ε_{it} is the noise term. M_{it} consists of the factor structure component $M^{\star}_{it} \coloneqq \beta_i' F_t$ with the approximation error M^{R}_{it} where β_i and F_t are K-dimensional random vectors of factor loading and latent factor respectively. On top of the approximate factor model setting, we introduce a random variable $\omega_{it} \coloneqq \mathbf{1}\{z_{it} \text{ is observed}\}$ to accommodate the missing data problem.

In this practical setting, we provide the inferential theory for the group average of M_{it}'s, regardless of whether their corresponding z_{it}'s are observed or not. Therefore, at a high level, this problem can be understood as a recovery of full data from partially observed data. Recently, the *matrix completion* literature has been developed to solve this sort of problem. It is obvious that recovery is impossible in general if there is no further assumption. One common assumption is that the matrices of interest are of low-rank (or approximate low-rank) compared to their dimensions, which the (approximate) factor structure of (1.1) satisfies.[1]

The standard method for the *low-rank matrix completion* is the nuclear norm penalization, and it has been intensively developed in the last decade. Candès and Recht (2009) formulate the low-rank matrix completion problem in the possibly simplest setting (an entry of a low-rank matrix is observed uniformly at random, and observation is not contaminated by 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 the noise contamination and then provide convergence rates for the nuclear norm penalized estimator in terms of various norms. 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 for computing 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 rare. This is because the shrinkage bias caused by the penalization makes us hard to pin down the distribution of the estimator. Very recently, some studies propose the debiasing method for the inference of the nuclear norm

[1] Note that (1.1) can be written as a matrix form like (2.1).

penalized estimator. Chen et al. (2019) and Xia and Yuan (2021) use debiasing methods that subtract the estimator of bias term from the plain estimator and derive the normality of the estimator. However, Chen et al. (2019) assume the normality of noise 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 independently across units and time. On the other hand, Chernozhukov et al. (2019, 2021) exploit the two-step least squares method with sample splitting as a debiasing method under more general assumptions on the noise and the missing pattern. This method takes advantage of the fact that the low-rank matrix, $M^\star := [M^\star_{it}]_{N\times T}$, can be written as a product of factors and loadings. To put it briefly, it first obtains an estimate of $M^\star$ using the nuclear-norm penalization and derives the initial estimator of loadings from the left singular vectors of the estimate of $M^\star$. By regressing observed entries of $Z := [z_{it}]_{N\times T}$ onto this estimate of loadings, it estimates latent factors. Again, by regressing observed entries of Z onto the estimate of latent factors, it updates the estimate of loadings. The final estimator of $M^\star$ is then the product of the estimates for latent factors and loadings. Chernozhukov et al. (2019, 2021) show that, with the aid of sample splitting, this two-step least squares procedure successfully removes the shrinkage bias induced by the use of the nuclear-norm penalization. However, one drawback of these papers is that they still resort to the sample splitting method.

We contribute to the literature by providing the inferential theory of the nuclear norm penalized estimation, whose debiasing method only utilizes the two-step least squares procedure without sample splitting. Sample splitting has some undesirable properties. By nature, sample splitting complicates the estimation procedure. In addition, sample splitting restricts researchers from choosing the group of (i,t) freely when they conduct inference for the group average of M_{it}. For instance, when the observation probability is different across units, if we adopt the inference procedure in Chernozhukov et al. (2019, 2021), we can only conduct inference for the cross-sectional average of a fixed period t, i.e., $\frac{1}{|\mathcal{I}|_o}\sum_{i\in\mathcal{I}} M_{it}$ where $\mathcal{I} \subset \{1,\dots,N\}$. It can be quite a strong restriction in the average treatment effect estimator application.[2] 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 to split the sample. Moreover, sample splitting can be computationally demanding in multiple tests. To make inference of the above cross-sectional average for all periods using sample splitting in Chernozhukov et al. (2019, 2021), we need to repeat the estimation steps described in Algorithm 1 T times, because the way of sample splitting for the cross-sectional average estimation is different across each time. It

[2] 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.

can be very time-consuming while running Algorithm 1 one time is enough in our method for the same goal. Last but not least, it is well-known that sample splitting generally causes unnecessary loss of efficiency.

Chernozhukov et al. (2019, 2021) take advantage of sample splitting to artificially generate certain independence, which helps them to show some bias term in the least square estimation of the latent factors is negligible. To be specific, after splitting the data Z into two independent parts, say, Z_1 and Z_2, they derive the nuclear norm penalized estimator from Z_1 and use the initial estimate of loadings extracted from this nuclear norm penalized estimator together with the data of another part, Z_2, for the least square estimation of the latent factors.[3] Using this sample splitting procedure, they make the initial estimate of loadings be independent of the data which is used in the least square estimation of the latent factors and exploit this independence to show the bias term is negligible.

On the other hand, we show the bias term is negligible without sample splitting by using a (hypothetical) *leave-one-out* estimator. The leave-one-out estimator is an auxiliary estimator which is close to the initial estimate of loadings and independent of the data used in the least square estimation of the latent factors. Specifically, when we estimate the latent factors of a period t, the leave-one-out estimator is constructed from $\{z_{js}\}_{j\leq N, s\neq t}$ to be independent of $\{z_{jt}\}_{j\leq N}$ conditioning on M, because $\{z_{jt}\}_{j\leq N}$ is the data used in the least square estimation of the latent factors of the period t. Although the initial estimate of loadings itself is not independent of the data used in the estimation of the latent factors (since we use the full samples for the nuclear norm penalized estimation without sample splitting), we can replace the initial estimate of loadings by the leave-one-out estimator with some additional negligible term in the proof. Then, because the leave-one-out estimator is independent of the data used in the least square estimation of the latent factors of the period t, we can show that the bias term is negligible. Importantly, the leave-one-out estimator does not need to be computed in practice since we only need their existence and theoretical properties (such as its distances to the target parameter or other estimators) in the proof. Therefore, we can remove the sample splitting step without implementing additional steps in practice.

Although the idea of the leave-one-out estimator is originated from Chen et al. (2020), 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 calculated by using the gradient descent iteration from the leave-one-out problem which rules out the target data, for example $\{z_{jt}\}_{j\leq N}$. However, there is one more aspect we have to take care. Since the loss function for the leave-one-out problem

[3] In the two-step least squares method, the latent factors are estimated by regressing observed entries of Z onto the initial estimate of loadings.

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 target data, the leave-one-out estimator using this stopping point may not be truly independent of the target data. Unlike Chen et al. (2020) who derived the stopping point from the problem using the full dataset, 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 2.3.1.

The other important contribution of the paper is that our inference procedure allows more general data-missing patterns than typical matrix completion literature 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 missing 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. These assumptions are quite restrictive in many cases. For example, the homogeneous observation probability assumption cannot accommodate the fact that the movie rating response rates might be 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 and correlated missing. To our best knowledge, only Chernozhukov et al. (2019, 2021) consider the inferential theory of the nuclear norm penalized estimation with the generalized missing patterns.

In the paper, we allow the heterogeneous observation probabilities. Besides, unlike Chernozhukov et al. (2019, 2021), 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, inverse probability weighting is designed to treat units with different observation probabilities in an equal 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 missing 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. 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 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. (2018) exploit the matrix completion method 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[4] and the alpha test in empirical asset pricing as applications. Unlike the inference procedure for the average treatment effect in Chernozhukov et al. (2021), we do not resort to sample splitting, and hence, we can avoid several drawbacks of sample splitting described above. Besides, we generalize the alpha test in Giglio et al. (2020) by assuming a more realistic data-missing pattern. We allow the cross-sectional correlation and the heterogeneous observation probability for the missing pattern and use inverse probability weighting. In addition, we allow the number of latent factors to increase slowly.

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 and the alpha test in empirical asset pricing as applications. 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.

[4] Here, we consider heterogeneous treatment effects.

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

2 Model and Estimation

Specifically, we consider the following nonparametric panel model subject to missing data problem:

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

where z_{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 and ε_{it} is the noise.[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 $h_t(\cdot)$. Although this model looks different from the approximate factor structure in (1.1) at first glance, the 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 $1 \le r \le K$, $\phi_r(\zeta_i)$ 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

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

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

Denote the singular value decomposition (SVD) of $M^\star$ by $U_{M^\star} D_{M^\star} V_{M^\star}'$ where $U_{M^\star}' U_{M^\star} =$

[5] Trivially, our theory holds for the model of $z_{it} = h_i(\eta_t) + \varepsilon_{it}$ 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.

$V'_{M^\star}V_{M^\star} = I_K$. $D_{M^\star}$ is a $K \times K$ diagonal matrix with singular values in descending order, i.e., $D_{M^\star} = \text{diag}(\psi_1, \ldots, \psi_K)$ where $\psi_{\max} = \psi_1 > \cdots > \psi_K = \psi_{\min} > 0$. Let $\mathcal{M} \coloneqq \{\beta, F, M^R\}$ be a set of random matrices which 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.

2.1 Nuclear norm penalized estimation with inverse probability weighting

We first look at the following estimator of M:[7]

$$\underset{m \in \mathbb{R}^{N \times T}}{\arg\min} \frac{1}{2} \|\mathcal{P}_\Omega(m - Z)\|_F^2 + \lambda \cdot \text{rank}(m) \tag{2.2}$$

where $\Omega = [\omega_{it}]_{N \times T}$ and $\mathcal{P}_\Omega(X) = \Omega \circ X$ for any $N \times T$ matrix X, $\circ$ denotes the Hadamard product, $\lambda > 0$ is some regularization parameter, and $\text{rank}(\cdot)$ is the matrix rank function. This rank penalized estimator is seemingly a natural choice for the low-rank matrix estimation where we want to minimize the squared error under the restriction of the low-rank assumption. However, the matrix rank function is highly non-convex and it makes the estimation computationally intractable. So, researchers often resort to convex relaxations in order to obtain computationally feasible method. One typical example is the nuclear norm penalization. The nuclear norm is a convex surrogate for the matrix rank function and hence, it gives a great computational tractability in optimization. For this reason, a number of papers, 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), study the nuclear norm penalization method for low-rank matrix and provide algorithms to compute the following convex program:

$$\underset{m \in \mathbb{R}^{N \times T}}{\arg\min} \frac{1}{2} \|\mathcal{P}_\Omega(m - Z)\|_F^2 + \lambda \|m\|_* . \tag{2.3}$$

However, the above papers only consider the case of the homogeneous observation probability, and if there is heterogeneity in the observation probability, using the objective function (2.3) may not be the best way 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

[7] In the paper, the estimator of the low-rank matrix $M^\star$ works as the estimator of the matrix of interest M. This is because M can be well approximated by the low-rank matrix $M^\star$ as long as the approximation error M^R is sufficiently small. Hence, the estimator of M is same as the estimator of $M^\star$ in the paper.

observation probability of movie rating data will be different across viewers. In this case, if we use the objective function (2.3), the estimation errors of people with high observation probability would be factored more into minimizing squared errors, and it may debase the estimation quality as explained below.

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:

$$\widetilde{M} := \underset{m \in \mathbb{R}^{N \times T}}{\arg\min} \frac{1}{2} ||\widehat{\Pi}^{-\frac{1}{2}} \mathcal{P}_{\Omega} (m - Z) ||_F^2 + \lambda ||m||_* \tag{2.4}$$

where $\widehat{\Pi} = \text{diag}(\hat{p}_1, \ldots, \hat{p}_N)$, and $\hat{p}_i = \frac{1}{T} \sum_{t=1}^{T} \omega_{it}$ for each $i \leq N$.

As noted in Ma and Chen (2019), this inverse probability weighting debiases the objective function itself. If there is heterogeneity in the observation probability, $\left\| \Pi^{-\frac{1}{2}} \mathcal{P}_{\Omega} (m - Z) \right\|_F^2$ is an unbiased estimates of $\|m - Z\|_F^2$, which we would use if there is no missing entry, in the sense that $\mathbb{E}_{\Omega} \left[\left\| \Pi^{-\frac{1}{2}} \mathcal{P}_{\Omega} (m - Z) \right\|_F^2 \right] = \|m - Z\|_F^2$, while $\|\mathcal{P}_{\Omega}(m - Z)\|_F^2$ is biased.[8] Hence, using the weighted objective function (2.4) is more suitable in the case of the heterogeneous observation probability. Besides, inverse probability weighting enhances the estimation quality by treating units equally. The units with high (low) observation probabilities would be factored more (less) into the unweighted objective function (2.3) 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 (2.4) since it compensates the effect of missing by putting the inverse of the observation probability.[9] 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

Our estimation step is as follows.

Algorithm 1 Constructing the estimator for M.

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

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

[9] 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.

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 $\widehat{F}_t = \left(\sum_{j=1}^{N} \omega_{jt} \widetilde{\beta}_j \widetilde{\beta}_j'\right)^{-1} \sum_{j=1}^{N} \omega_{jt} \widetilde{\beta}_j z_{jt}$.

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

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]

2.3 Debiasing strategy

It is well-known that the nuclear-norm penalized estimator $\widetilde{M}$ is subject to shrinkage biases which complicate statistical inference. Therefore, for valid inference, removing the effect of the shrinkage biases is essential. Existing studies providing the inferential theory of the nuclear norm penalization like Chernozhukov et al. (2019, 2021) take advantage of the two-step least squares procedure with sample splitting to remove the shrinkage biases. However, sample splitting has several drawbacks as described below. One of the key contributions of this paper is that it provides an inferential theory whose debiasing method only utilizes the two-step least squares procedure without sample splitting.

Sample splitting does not only complicate the estimation procedure by its nature, but it also 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 to split the sample. 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

[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.

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 for the cross-sectional average estimation is different across each time. It can be very time-consuming, while our method runs Algorithm 1 only once for the same goal.[12] Last but not least, as noted in Chen et al. (2019), sample splitting generally causes unnecessary loss of efficiency.[13]

2.3.1 Unnecessariness of sample splitting

To understand how we can remove sample splitting in the debiasing procedure, note that we have the following maximization problem in Step 3

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

where $Q_t(f, b) = -\frac{1}{N}\sum_{j=1}^N \omega_{jt}(z_{jt} - f'b_j)^2$, $b = (b_1, \ldots, b_N)'$ and b_j are K dimensional vectors. Here, we can consider β as the nuisance parameter and F_t as the parameter of interest. The key of the debiasing strategy is to obtain an unbiased estimator of F_t (up to the rotation) while the estimator of the nuisance parameter, $\widetilde{\beta}$, suffers from the shrinkage bias. Let H_1 be a $K \times K$ rotation matrix such that $\frac{1}{\sqrt{N}}\beta H_1$ is the left singular vector of $M^\star$.[14] Then, by Taylor expansion with some simple algebra, we have

$$\begin{aligned}
&\widehat{F}_t - H_1^{-1}F_t \\
&\quad = -B^{-1}\frac{\partial Q_t(H_1^{-1}F_t, \beta H_1)}{\partial f} - B^{-1}\frac{\partial^2 Q_t(H_1^{-1}F_t, \beta H_1)}{\partial f \partial \mathrm{vec} b}\mathrm{vec}(\widetilde{\beta} - \beta H_1) + \text{higher order terms.}
\end{aligned}$$

where $B := \mathbb{E}\left[\frac{\partial^2 Q_t(H_1^{-1}F_t, \beta H_1)}{\partial f^2}\middle| \mathcal{M}\right]$. The first term is the score, which leads to the asymptotic normality. The convergence rate of this term is roughly $1/\sqrt{N}$. On the other hand, the second term represents the effect of β estimation which is subject to the shrinkage bias. Note that the second term can be divided into

$$-B^{-1}\frac{\partial^2 Q_t(H_1^{-1}F_t, \beta H_1)}{\partial f \partial \mathrm{vec} b}\mathrm{vec}(\widetilde{\beta} - \beta H_1) = \varphi H_1^{-1}F_t + R_{t,1} + R_{t,2} + R_{t,3}, \tag{2.5}$$

[12] 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 $\{M_{it}\}_{1\le i\le N}$ for one fixed t. On the other hand, in our method, we can estimate $\{M_{it}\}_{1\le i\le N, 1\le t\le T}$ from one time execution of Algorithm 1 using matrix form.

[13] However, Chernozhukov et al. (2019, 2021) use an averaging method to restore the asymptotic efficiency that would otherwise be lost because of the sample splitting.

[14] Note that $\frac{1}{\sqrt{N}}\widetilde{\beta}$ is the left singular vector of $\widetilde{M}$.

where

$$\varphi \coloneqq -B^{-1}H_1'\frac{1}{N}\sum_{j=1}^{N}p_j\beta_j(\widetilde{\beta}_j - H_1'\beta_j)', \quad R_{t,1} \coloneqq -B^{-1}H_1'\frac{1}{N}\sum_{j=1}^{N}(\omega_{jt}-p_j)\beta_j F_t'H_1'^{-1}\left(\widetilde{\beta}_j - H_1'\beta_j\right),$$

$$R_{t,2} \coloneqq B^{-1}\frac{1}{N}\sum_{j=1}^{N}\omega_{jt}\varepsilon_{jt}\left(\widetilde{\beta}_j - H_1'\beta_j\right), \quad R_{t,3} \coloneqq B^{-1}\frac{1}{N}\sum_{j=1}^{N}\omega_{jt}M_{jt}^{R}\left(\widetilde{\beta}_j - H_1'\beta_j\right).$$

Here, $R_{t,3}$ is of higher order as long as $\max_{i,t}|M_{it}^R|$ is sufficiently small. Hence, for $\widehat{F}_t$ to be an unbiased estimator of F_t (up to a rotation), showing $R_{t,1}$, $R_{t,2}$ are of higher order (roughly, $O_p(1/N)$) is the key task.[15] Chernozhukov et al. (2019, 2021) take advantage of sample splitting to show $R_{t,1}$, $R_{t,2}$ are of higher order. They rely on sample splitting to artificially generate the independence between $\{\omega_{jt},\varepsilon_{jt}\}_{j\leq N}$ and $\{\widetilde{\beta}_j\}_{j\leq N}$ which helps them to bound $R_{t,1}$, $R_{t,2}$ tightly. Sample splitting, however, has several drawbacks as noted above.

On the contrary, we use a different approach which exploits the hypothetical *leave-one-out* estimator to show that $R_{t,1}$, $R_{t,2}$ are of higher order without sample splitting. This leave-one-out estimator is close to $\widetilde{\beta}$ and independent of $\{\omega_{jt},\varepsilon_{jt}\}_{j\leq N}$. Although $\widetilde{\beta}$ itself is not independent of $\{\omega_{jt},\varepsilon_{jt}\}_{j\leq N}$ without sample splitting, we replace $\widetilde{\beta}$ in $R_{t,1}$, $R_{t,2}$ by this leave-one-out estimator with some additional negligible term and derive tight bounds of $R_{t,1}$, $R_{t,2}$ using the independence between $\{\omega_{jt},\varepsilon_{jt}\}_{j\leq N}$ and the leave-one-out estimator. It is worth noting that the leave-one-out estimator does not need to be computed in practice since we only need their existence and theoretical properties (such as their distances to the truth or other estimators) in the proof. Therefore, we can remove the sample splitting step without implementing alternative steps in practice.

Our technical arguments follow from the following rationales:

1. 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 two problems: one uses the full sample, and another uses the auxiliary leave-one-out:

$$f^{\text{infs}}(w,y) = \frac{1}{2}||\Pi^{-\frac{1}{2}}\mathcal{P}_{\Omega}\left(wy'-Z\right)||_F^2 + \frac{\lambda}{2}||w||_F^2 + \frac{\lambda}{2}||y||_F^2, \tag{2.6}$$

$$f^{\text{infs},(-t)}(w,y) = \frac{1}{2}\left\|\Pi^{-1/2}\mathcal{P}_{\Omega_{\cdot,-t}}(wy'-Z) + \mathcal{P}_{\cdot,t}(wy'-M^{\star})\right\|_F^2 + \frac{\lambda}{2}\left\|w\right\|_F^2 + \frac{\lambda}{2}\left\|y\right\|_F^2, \tag{2.7}$$

where w and y are $N\times K$ and $T\times K$ matrices, respectively.[16] Motivated by Chen et al. (2020),

[15] We introduce how we can deal with the term $\varphi H_1^{-1}F_t$ in Section 2.3.2.

[16] Here, $\mathcal{P}_{\Omega_{\cdot,-t}}(A) \coloneqq \Omega_{\cdot,-t}\circ A$ where $\Omega_{\cdot,-t} \coloneqq [\omega_{js}1\{s\neq t\}]_{N\times T}$ and $\mathcal{P}_{\cdot,t}(A) \coloneqq E_{\cdot,t}\circ A$ where $E_{\cdot,t} \coloneqq [1\{s=t\}]_{N\times T}$.

both problems are asymptotically equivalent to the original nuclear-norm penalization, and the estimator from the leave-one-out problem should be independent of the data left out, creating the same spirit of sample splitting for post-regularized inference.

2. Both problems should be iterated up to a stopping point, and the gradients of loss functions of non-convex problems should be sufficiently small at the stopping point. However, the gradients do not monotonically decrease as iteration increases because it is not a convex problem. So, one cannot let it iterate until convergence is reached, but has to stop at the point where the gradient is "small". Hence it is crucial to derive the "stopping point". We show that the iteration algorithms for the two non-convex problems have to stop after the same number of steps.

3. Unlike Chen et al. (2020) who derived the stopping point from the problem using the full dataset (2.6), we show that the stopping point should be derived using the leave-one-out problem (2.7) for inference purposes. This would then ensure that the estimator from leave-one-out problem using this stopping point is independent of the data left out. While this change causes some nontrivial difficulties in the proof, we successfully resolve the problems. Finally, this argument admits both homogeneous and heterogeneous missing patterns.

Let $W := U_{M^\star} D_{M^\star}^{1/2}$ and $Y := V_{M^\star} D_{M^\star}^{1/2}$. We denote the leave-one-out estimator of (W, Y) derived from the loss function $f^{\text{infs},(-t)}$ by $(\breve{W}^{(-t)}, \breve{Y}^{(-t)})$ and the corresponding rotation matrix by $\breve{H}^{(-t)}$.[17] Importantly, in the goodness of fit part, $\{p_j^{-1}\omega_{jt}, z_{jt}\}_{j\leq N}$ is replaced by its (approximate) mean $\{1, M^\star_{jt}\}_{j\leq N}$ so that $\{\omega_{jt}, \varepsilon_{jt}\}_{j\leq N}$ is excluded from $f^{\text{infs},(-t)}$. Hence, $(\breve{W}^{(-t)}, \breve{H}^{(-t)})$, which is derived from the loss function $f^{\text{infs},(-t)}$, is independent of $\{\omega_{jt}, \varepsilon_{jt}\}_{j\leq N}$ when $\{\omega_{js}, \varepsilon_{js}\}_{j\leq N, s\leq T}$ are independent across time.

We define the leave-one-out estimator of $\beta_l H_1$ as $\widehat{\beta}^{loo,(-t)} := \sqrt{N}\breve{W}^{(-t)}\breve{H}^{(-t)} D_{M^\star}^{-1/2}$ since $\beta_l H_1 = \sqrt{N} U_{M^\star}$ by the definition of H_1. By using this leave-one-out estimator, we can have the following decomposition and replace $\widetilde{\beta} - \beta H_1$ in $R_{t,1}$ and $R_{t,2}$ by $Term_1 + Term_2$:[18]

$$\widetilde{\beta} - \beta H_1 = \underbrace{(\widehat{\beta}^{loo,(-t)} - \beta H_1)}_{=Term_1,\ \text{independent from } \{\omega_{jt},\varepsilon_{jt}\}_{j\leq N}} + \underbrace{(\widetilde{\beta} - \widehat{\beta}^{loo,(-t)})}_{=Term_2,\ \text{not independent from } \{\omega_{jt},\varepsilon_{jt}\}_{j\leq N}}. \tag{2.8}$$

Note that $Term_1$ is independent of $\{\omega_{jt}, \varepsilon_{jt}\}$ in $R_{t,1}$ and $R_{t,2}$. In addition, we can show $||Term_1||$ has a similar bound to $||\widetilde{\beta} - \beta H_1||$ (roughly, $O_p(1)$). Hence, we easily show that the parts of $R_{t,1}$ and $R_{t,2}$ containing $Term_1$ are of higher-order by using this independence.[19]

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

[18] In the matrix form, $R_{t,1} = \frac{1}{N} B^{-1} H_1' \beta' (\Omega_t - \Pi) \left(\widetilde{\beta} - \beta H_1\right) H_1^{-1} F_t$ and $R_{t,2} = \frac{1}{N} B^{-1} \left(\widetilde{\beta} - \beta H_1\right)' \Omega_t \mathcal{E}_t$ where $\Omega_t = \text{diag}(\omega_{1t}, \ldots, \omega_{Nt})$, $\mathcal{E}_t = [\varepsilon_{1t}, \ldots, \varepsilon_{Nt}]'$.

[19] Because of the independence, if we condition on $Term_1$, the properties of $\{\omega_{jt}, \varepsilon_{jt}\}$ in $R_{t,1}$ and $R_{t,2}$ are unaffected by this conditioning. Hence, by conditioning on $Term_1$ and $\mathcal{M}$, we can treat $Term_1$, β, F and H_1 as constants and easily get the

In contrast, *Term*$_2$ is correlated with $\{\omega_{jt}, \varepsilon_{jt}\}_{1\leq j\leq N}$ because it contains $\widetilde{\beta}$ which is constructed from the full sample data without sample splitting. So, we cannot use the above method which exploits the independence and need to take a different approach to show the parts of $R_{t,1}$ and $R_{t,2}$ containing *Term*$_2$ are of higher order.

Parts of $R_{t,1}$ and $R_{t,2}$ containing *Term*$_2$

To show these parts are of higher order, we introduce another estimator of (W, Y) whose loss function f^{infs} is similar to that of the leave-one-out estimator $f^{\text{infs},(-t)}$. We denote this non-convex optimization estimator of (W, Y) derived from f^{infs} by $(\widetilde{W}, \widetilde{Y})$ and the corresponding rotation matrix by $\widetilde{H}$. Using the non-convex optimization estimator $(\widetilde{W}, \widetilde{Y})$, we construct two intermediate estimators of $\beta_l H_1$. First, by the same logic as the above leave-one-out estimator of $\beta_l H_1$, we define the first intermediate estimator of $\beta_l H_1$ as $\widehat{\beta}^{int_1} := \sqrt{N}\widetilde{W}\widetilde{H}D_{M^\star}^{-1/2}$. For the second intermediate estimator, we use a rotation matrix H_4 such that $\psi_{\min}^{-1/2}\widetilde{W}H_4$ is the left singular vector of $\widetilde{W}\widetilde{Y}'$. Because $\widetilde{W}\widetilde{Y}'$ is an estimator of $M^\star$ and $\frac{1}{\sqrt{N}}\beta H_1$ is the left singular vector of $M^\star$, we define the second intermediate estimator of βH_1 as $\widehat{\beta}^{int_2} := \sqrt{N}\psi_{\min}^{-1/2}\widetilde{W}H_4$.

Using these estimators, we can divide *Term*$_2$ in (2.8) into the following three terms:

$$\widetilde{\beta} - \widehat{\beta}^{loo,(-t)} = \underbrace{(\widehat{\beta}^{int_1} - \widehat{\beta}^{loo,(-t)})}_{Term_{2-1}} + \underbrace{(\widetilde{\beta} - \widehat{\beta}^{int_2})}_{Term_{2-2}} + \underbrace{(\widehat{\beta}^{int_2} - \widehat{\beta}^{int_1})}_{Term_{2-3}}.$$

Note that each term represents the distance between the estimators of $\beta_l H_1$.

For *Term*$_{2-1}$ and *Term*$_{2-2}$, we can have tighter bounds than that of $\widetilde{\beta} - \beta H_1$. First, *Term*$_{2-1}$ accounts for the proximity between the non-convex optimization estimator $\widetilde{W}$ and the leave-one-out estimator $\breve{W}^{(-t)}$ with some rotations. If we compare their loss functions f^{infs} and $f^{\text{infs},(-t)}$, the difference between two loss functions exists only in the time t. At the time t, $f^{\text{infs},(-t)}$ changes $\{p_j^{-1}\omega_{jt}, z_{jt}\}_{j\leq N}$ in f^{infs} to its (approximate) mean $\{1, M_{jt}^\star\}_{j\leq N}$ in the goodness of fit part. Hence, it is natural to anticipate that $\widetilde{W}\widetilde{H}$ and $\breve{W}^{(-t)}\breve{H}^{(-t)}$ are very close. Indeed, we can show that $||Term_{2-1}||$ has a much tighter bound than that of $||\widetilde{\beta} - \beta H_1||$.

In addition, *Term*$_{2-2}$ accounts for the distance between the nuclear norm penalized estimator $\widetilde{M}$ and the non-convex optimization estimator $\widetilde{W}\widetilde{Y}'$ because $\frac{1}{\sqrt{N}}\widetilde{\beta}$ and $\psi_{\min}^{-1/2}\widetilde{W}H_4$ are the left singular vectors of $\widetilde{M}$ and $\widetilde{W}\widetilde{Y}'$ respectively. Intuitively, from the elementary fact that

$$||m||_* = \inf_{w\in\mathbb{R}^{N\times K}, y\in\mathbb{R}^{T\times K}: wy'=m} \left\{\frac{1}{2}||w||_F^2 + \frac{1}{2}||y||_F^2\right\},$$

tight bounds for $R_{t,1}$ and $R_{t,2}$ using conditional concentration inequalities.

we can conjecture that the solution of (2.4) and that of (2.6) will be very close. Indeed, using the theory in Chen et al. (2020), we can formally show that $||\widetilde{M} - \widetilde{W}\widetilde{Y}'||$ has a sufficiently tight bound so that $||Term_{2-2}||$ can have a much smaller bound than the the bound of $||\widetilde{\beta} - \beta H_1||$. Since the orders of $||Term_{2-1}||$ and $||Term_{2-2}||$ are roughly $O_p(1/\sqrt{N})$, which is $\sqrt{N}$ times smaller than that of $||\widetilde{\beta} - \beta H_1||$, we can show that the parts of $R_{t,1}$, $R_{t,2}$ containing these terms are of higher order without the aid of the independence from $\{\omega_{jt}, \varepsilon_{jt}\}_{1\leq j\leq N}$.

On the other hand, in the case of $Term_{2-3}$, we cannot have a tight bound like the cases of $Term_{2-1}$ or $Term_{2-2}$ because there is no direct relation between $\widehat{\beta}^{int_1}$ and $\widehat{\beta}^{int_2}$. In addition, $Term_{2-3}$ is correlated with $\{\omega_{jt}, \varepsilon_{jt}\}_{1\leq j\leq N}$. Hence, we cannot use the above methods to show that the parts of $R_{t,1}$, $R_{t,2}$ containing $Term_{2-3}$ are of higher-order and need a new approach. Here, we exploit the relation

$$Term_{2-3} = \widetilde{W}\left(\widetilde{W}'\widetilde{W}\right)^{-1}\widetilde{W}'Term_{2-3} \tag{2.9}$$

to artificially generate the term in which we can use a concentration inequality to get a tight bound. For instance, using the above relation, the part of $R_{t,2}$ containing $Term_{2-3}$ can be bounded like

$$||R_{t,2,Term_{2-3}}|| = ||\frac{1}{N}B^{-1}Term'_{2-3}\Omega_t\mathcal{E}_t|| \leq \frac{C}{N}\left\|Term_{2-3}\right\|\left\|\widetilde{W}\left(\widetilde{W}'\widetilde{W}\right)^{-1}\right\|\left\|\widetilde{W}'\Omega_t\mathcal{E}_t\right\|,$$

where $\Omega_t = \text{diag}(\omega_{1t}, \ldots, \omega_{Nt})$, $\mathcal{E}_t = [\varepsilon_{1t}, \ldots, \varepsilon_{Nt}]'$. Here, we artificially generate the term $\widetilde{W}'\Omega_t\mathcal{E}_t$ in which we can derive $||\widetilde{W}'\Omega_t\mathcal{E}_t|| \approx ||W'\Omega_t\mathcal{E}_t|| = ||\sum_j \omega_{jt}\varepsilon_{jt}W_j|| \approx O_p(\sqrt{N})$ by Bernstein inequality. Then, since $||\widetilde{W}\left(\widetilde{W}'\widetilde{W}\right)^{-1}|| \approx O_p(1/\sqrt{N})$, we can show that the part of $R_{t,2}$ containing $Term_{2-3}$ are of higher order.[20]

In this way, we can show that $R_{t,1}$, $R_{t,2}$ are of higher order without using sample splitting. Deriving bounds of $Term_1$ and $Term_2$ relies on the theory in Chen et al. (2020) which study the convergence rates of the nuclear-norm penalized estimator. As a byproduct of finding the bounds of $Term_1$ and $Term_2$, we generalize their results in the sense that i) the data matrices are nonsquare ($N \neq T$), ii) the matrix M is random and consists of the low-rank matrix $M^\star$ and the approximation error M^R, importantly, iii) we assume the heterogeneous observation probability ($\mathbb{E}[\omega_{it}]$ is different across i) and use inverse probability weighting. iv) Last but not least, we allow the cross-sectional dependence of the missing patterns as described in Assumption 3.2.

[20] Note that, if we do not use the relation (2.9), we just have

$$||R_{t,2,Term_{2-3}}|| = ||\frac{1}{N}B^{-1}Term'_{2-3}\Omega_t\mathcal{E}_t|| \leq \frac{C}{N}||Term_{2-3}||||\Omega_t\mathcal{E}_t|| = O_p(1/\sqrt{N})$$

since $||\Omega_t\mathcal{E}_t|| = O_p(\sqrt{N})$. Hence, we cannot show that $R_{t,2,Term_{2-3}}$ is of higher order.

2.3.2 Remaining bias absorbed by the rotation matrix

Lastly, we need to deal with the remaining first order effect of $\widetilde{\beta} - \beta H_1$, that is, $\varphi H_1^{-1} F_t$ in (2.5). Since it is enough that F_t is well estimated up to a rotation, by defining $H_2 = (I_K + \varphi)H_1^{-1}$, this first order effect can be "absorbed" in this rotation matrix and we reach

$$\widehat{F}_t - H_2 F_t = -B^{-1} \frac{\partial Q_t(H_1^{-1} F_t, \beta H_1)}{\partial f} + \text{higher order terms.}$$

Therefore, we can successfully remove the shrinkage bias of the nuclear norm penalization and derive the unbiased estimator for F_t up to the rotation in Step 3. In addition, since it works as a fine support for running OLS to estimate β_i in Step 4, $\widehat{\beta}_i$ also unbiasedly estimates β_i up to the rotation. As a result, we have

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

which allows to conduct inference successfully.

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

$$||\widehat{\Pi}^{-1} \mathcal{P}_{\Omega}(\mathcal{E})|| < \frac{7}{8}\lambda$$

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, $||\widehat{\Pi}^{-1} \mathcal{P}_{\Omega}(\mathcal{E})||$ and $||\widehat{\Pi}^{-1} \mathcal{P}_{\Omega}(\mathcal{U})||$ 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

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

Then, we have

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

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 $z_{it} = m_t + \sigma^{-1}\epsilon_{it}$ (or $z_{it} = m + \sigma^{-1}\epsilon_{it}$) 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 $\tilde{\varepsilon}_{it} = z_{it} - \widetilde{M}_{\text{init}}$, 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}. Here, the group of interest $\mathcal{G}$ is defined as $\mathcal{G} = \mathcal{I} \times \mathcal{T}$ where $\mathcal{I}$ is a subset of $\{1, \ldots, N\}$ and $\mathcal{T}$ is a subset of $\{1, \ldots, T\}$. We define $\bar{\beta}_{\mathcal{I}}$ and $\bar{F}_{\mathcal{T}}$ as $\bar{\beta}_{\mathcal{I}} = \frac{1}{|\mathcal{I}|_o}\sum_{j\in\mathcal{I}}\beta_j$ and $\bar{F}_{\mathcal{T}} = \frac{1}{|\mathcal{T}|_o}\sum_{s\in\mathcal{T}}F_s$ respectively. Before proceeding, we present some assumptions for the asymptotic normality of the estimator.

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

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

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,

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

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

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,

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

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 incoherent 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.

Before we impose the next assumption, we present a cluster structure of $\{1, \ldots, N\}$ which helps us to allow the cross-sectional dependence in ω_{it}. We assume that there is a family of nonempty disjoint clusters, $\mathcal{C}_1, \ldots, \mathcal{C}_\rho$, such that $\cup_{g=1}^{\rho} \mathcal{C}_g = \{1, \ldots, N\}$ and allow the cross-sectional dependence of ω_{it} within a cluster.

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 $\|\varepsilon_{it}\|_{subG} \leq C\sigma$ 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 $\ddot{\omega}_{it} = \omega_{it} - p_i$ and $\ddot{\Omega} = [\ddot{\omega}_{it}]_{N \times T}$. Denote the columns of $\ddot{\Omega}$ and $\mathcal{P}_\Omega(\mathcal{E}) = \Omega \circ \mathcal{E}$ by $\ddot{\omega}_t$ and ς_t, respectively. Then, $\{\ddot{\omega}_t\}_{t \leq T}$ ($\{\varsigma_t\}_{t \leq T}$) are independent sub-gaussian random vectors with $\mathbb{E}[\ddot{\omega}_t] = 0$ ($\mathbb{E}[\varsigma_t] = 0$); more specifically, there is $C > 0$ such that for $a_t \in \{\ddot{\omega}_t, \varsigma_t\}$,

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

(iv) 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 $\mathcal{C}_{g(i)}$, $\{w_{j_1 t}, \ldots, w_{j_m t}\}$ is independent from w_{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.

(v) We have $\max_t \max_i \sum_{j=1}^N |Cov(\omega_{it}, \omega_{jt}|\mathcal{M})| < C$.

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 $\|\mathcal{P}_\Omega(\mathcal{E})\|$ and $\|\mathcal{P}_{\ddot{\Omega}}(\mathbf{1}\mathbf{1}')\|$. Assumptions 3.2 (iv), (v) allow ω_{it} to have a weak cross-sectional dependence.

The cluster assumption in 3.2 (iv) helps us to construct the leave-one-out estimator $(\breve{W}^{\{-i\}}, \breve{Y}^{\{-i\}})$ which is independent of $\{\omega_{is}, \varepsilon_{is}\}_{s\leq T}$ by excluding $\{z_{js}\}_{j\in\mathcal{C}_{g(i)}, s\leq T}$ from the loss function.[21] We exploit the leave-one-out estimators $(\breve{W}^{\{-i\}}, \breve{Y}^{\{-i\}})$ and $(\breve{W}^{(-t)}, \breve{Y}^{(-t)})$ to control the l_2/l_∞ error of the non-convex optimization estimator $(\widetilde{W}, \widetilde{Y})$. The proximity in terms of the l_2/l_∞ error plays a pivotal role in showing $||\widetilde{M} - \widetilde{W}\widetilde{Y}'||$ has a sufficiently tight bound.

The parameter for the cluster size ϑ is bounded by Assumption 3.3 (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 $\vartheta \approx o_p(\sqrt{N/\log N})$ since K is allowed to grow very slowly by setting a large ν.

Assumption 3.3 (Slowly increasing ϑ, K and $\min\{|\mathcal{I}|_o, |\mathcal{T}|_o\}$)**.**

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

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

$$(iii)\ \frac{\min\{|\mathcal{I}|_o, |\mathcal{T}|_o\}\max\{N,T\}}{K^{(2\nu-3)}} = o_p(1).$$

Assumptions 3.3 (i), (ii) accommodates 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 3.3 (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.

Assumption 3.4 (Eigengap)**.** *There are $c, C > 0$ such that with probability converging to 1, $\psi_1 \leq C\psi_K$*

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

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 $M^\star$.

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

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 or 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

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

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

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

4.1 Alpha test in empirical asset pricing

It is not uncommon in finance to deal with data having missing entries. So, it is crucial to devise inference tools that can work in the presence of missing. The alpha test in asset pricing models, which is used to detect assets having an abnormal return, also needs to work in the presence of missing because return data of assets are often missing due to infrequent transactions or the short lifespan of assets. In this section, as an application of the estimation method in Section 2, we introduce the alpha test procedure in asset pricing models, which is robust to missing data. The test procedure and corresponding inferential theory are somewhat advanced than those in Giglio et al. (2020) since we allow the heterogeneous observation probability and the cross-sectional correlation of missing patterns and utilize inverse probability weighting for the nuclear norm penalized estimation.

We start with a description of the model. We assume the $N \times 1$ vector of excess returns r_t follows a linear factor model:

$$r_t = \alpha + \beta\theta + \beta\left(f_t - \mathbb{E}f_t\right) + \varepsilon_t,$$

where $\{f_t\}_{t\leq T}$ are $K \times 1$ vectors of latent factors, $\beta = [\beta_1, \ldots, \beta_N]'$ is a $N \times K$ loading matrix, θ is a $K \times 1$ vector of factor risk premiums and ε_t is the idiosyncratic component. α is a $N \times 1$ vector and it is the parameter of interest. Here, the excess return $\{r_{it}\}_{i\leq N, t\leq T}$ is subject to missing and $\omega_{it} \coloneqq 1\{r_{it} \text{ is observed}\}$. $\mathcal{N}_t = \{1 \leq i \leq N : \omega_{it} = 1\}$ denote the set of assets whose returns are observed at the period t and $\mathcal{T}_i = \{1 \leq t \leq T : \omega_{it} = 1\}$ denote the collection of periods on which the i-th asset return is observed. In addition, we define $N_t = |\mathcal{N}_t|_o$ and $T_i = |\mathcal{T}_i|_o$.

Test procedure

The objective of the test is to find individual assets with truly positive alphas (or with nonzero alphas). To do so, we can formulate a collection of null hypotheses, one for each asset:

$$\mathbb{H}^i_{0,pos} : \alpha_i \leq 0 \ \ (\text{or } \mathbb{H}^i_{0,noz} : \alpha_i = 0), \qquad i = 1, \ldots, N.$$

First, we introduce the estimation step for α_i. Here, we use $\mathbb{M}_A = I_p - A(A'A)^{-1}A'$ to denote the annihilator matrix for any $p \times q$ matrix A. In addition, $1_a = (1, \ldots, 1)'$ is a $a \times 1$ vector of ones.

Algorithm 2 Constructing the estimator for α_i.

Step 1 (a) Obtain the serially demeaned return matrix $Z = [z_{it}]_{N\times T}$ such that $z_{it} = r_{it} - \bar{r}_i$ where $\bar{r}_i = \frac{1}{T_i}\sum_{t\in\mathcal{T}_i} r_{it}$ is the average return of the asset i at its observed periods.

(b) Derive the weighted nuclear norm penalized estimator $\widetilde{M}$ in (2.4) from the demeaned return matrix Z in (a). Estimate the latent factors and corresponding loadings using $\widetilde{M}$:

$$\widehat{\upsilon}_t = \left(\sum_{j\in\mathcal{N}_t} \widetilde{\beta}_j\widetilde{\beta}_j'\right)^{-1} \sum_{j\in\mathcal{N}_t} \widetilde{\beta}_j z_{jt}, \quad t = 1,\ldots,T,$$

$$\widehat{\beta}_i = \left(\sum_{s\in\mathcal{T}_i} \widehat{\upsilon}_s\widehat{\upsilon}_s'\right)^{-1} \sum_{s\in\mathcal{T}_i} \widehat{\upsilon}_s z_{is}, \quad i = 1,\ldots,N,$$

where $\widetilde{\beta} = [\widetilde{\beta}_1,\ldots,\widetilde{\beta}_N]'$ is the $N\times K$ matrix whose columns are $\sqrt{N}$ times the top K left singular vectors of $\widetilde{M}$.

Step 2 Run a cross-sectional regression of $\bar{r}$ on the estimated $\widehat{\beta}$ and a constant regressor 1_N to obtain the estimate $\widehat{\theta}$:

$$\widehat{\theta} = \left(\widehat{\beta}'\mathbb{M}_{1_N}\widehat{\beta}\right)^{-1}\left(\widehat{\beta}'\mathbb{M}_{1_N}\bar{r}\right),$$

where $\bar{r} = [\bar{r}_1,\ldots,\bar{r}_N]'$ and $\bar{r}_i = \frac{1}{T_i}\sum_{t\in\mathcal{T}_i} r_{it}$.

Step 3 Estimate α_i with a bias correction:

$$\widehat{\alpha}_i = \bar{r}_i - \widehat{\beta}_i'\widehat{\theta} - \widehat{A}_i, \quad \widehat{A}_i = \widehat{\xi}_i'\widehat{g}, \quad i = 1,\ldots,N,$$

where $\widehat{\xi}_i' = e_i' - \widehat{\beta}_i'(\widehat{\beta}'\mathbb{M}_{1_N}\widehat{\beta})^{-1}\widehat{\beta}'\mathbb{M}_{1_N}$, $e_i' = (0,\ldots,0,1,0,\ldots,0)$, $\widehat{g}_i = \frac{1}{T_i}\sum_{t\in\mathcal{T}_i}\widehat{\upsilon}_t'\widehat{\beta}_i$.

Roughly speaking, the estimation steps can be summarized as follows. First, we apply the matrix completion method to the serially demeaned return matrix to estimate the latent factors and loadings. Next, we implement cross-sectional regressions like Fama-MacBeth to estimate the risk premiums of the factors. And then, we estimate the alphas with some bias correction. Note that Step 1 (b) is basically same as Algorithm 1 in Section 2. Because the demeaned return matrix Z in Step 1 is subject to missing and has an approximate factor structure, we can exploit the matrix completion to estimate the loadings for latent factors.

Having described how we obtain the alpha estimates, we now turn to the construction of the test statistics. With the aid of Theorem 4.1, we can obtain the following test statistics for $\mathbb{H}^i_{0,pos}$ and

corresponding p-values :

$$t_i = \frac{\widehat{\alpha}_i}{se\,(\widehat{\alpha}_i)}, \quad pv_i = 1 - \Phi(t_i), \quad i = 1, \ldots, N, \tag{4.1}$$

where $se\,(\widehat{\alpha}_i) = \frac{1}{\sqrt{T_i}}\widehat{\mathcal{V}}_{\alpha_i}^{\frac{1}{2}}$, $\widehat{\mathcal{V}}_{\alpha_i} = \frac{1}{T_i}\sum_{t\in\mathcal{T}_i}\widehat{\varepsilon}_{it}^2(1-\widehat{\upsilon}_t'\widehat{\Sigma}_f^{-1}\widehat{\theta})^2$, $\widehat{\varepsilon}_{it} = r_{it} - \bar{r}_i - \widehat{\beta}_i'\widehat{\upsilon}_t$ is the residual, and $\widehat{\Sigma}_f = \frac{1}{T}\sum_{t=1}^{T}\widehat{\upsilon}_t\widehat{\upsilon}_t'$. For the null hypothesis $\mathbb{H}_{0,noz}^i$, we can simply replace the calculation of p-values in (4.1) by $pv_i = 2(1-\Phi(|t_i|))$.

Lastly, we apply the B-H procedures in Giglio et al. (2020) to control the size of the false discovery rate (FDR). Let $\gamma \in (0,1)$ be a predetermined level such that we want to ensure FDR $\leq \gamma$. Then, the B-H procedure is as follows.

Algorithm 3 B-H procedure.

Step 1 Sort in ascending order the collection of p-values, $\{pv_i : 1 \leq i \leq N\}$, of the individual test statistics $\{t_i\}$. Denote $pv_{(1)} \leq \cdots \leq pv_{(N)}$ as the sorted p-values.

Step 2 For $i = 1, \ldots, N$, reject $\mathbb{H}_0^i$ if $pv_i \leq pv_{(\widehat{k})}$, where $\widehat{k} = \max\{i \leq N : pv_{(i)} \leq \gamma i/N\}$.

Moreover, in the case of testing the null hypothesis $\mathbb{H}_{0,pos}^i$, we can exploit the alpha screening in Giglio et al. (2020) to improve the power of the B-H procedure.

Algorithm 4 Alpha screening B-H procedure.

Step 1 Define $\widehat{\mathcal{I}} = \left\{i \leq N : t_i > -\log(\log T)\sqrt{\log N}\right\}$.

Step 2 Sort the p-values, $pv_{(1)} \leq \cdots \leq pv_{(|\widehat{\mathcal{I}}|_o)}$ for $\{pv_i : i \in \widehat{\mathcal{I}}\}$.

Step 3 For $i \in \widehat{\mathcal{I}}$, reject $\mathbb{H}_{0,pos}^i$ if $pv_i \leq pv_{(\widehat{k})}$, where $\widehat{k} = \max\{i \in \widehat{\mathcal{I}} : pv_{(i)} \leq \gamma i/|\widehat{\mathcal{I}}|_o\}$. Accept all other $\mathbb{H}_{0,pos}^i$.

By using the above B-H procedures, we can successfully bound the size of the FDR. For the details about controlling the FDR, please refer to Section 2.2 of Giglio et al. (2020).

Asymptotic normality of the estimates of α_i

Now, we present the formal conditions for the asymptotic normality of the estimates of α_i which is crucial to construct the test statistics. We denote the $T\times K$ matrix of $\{f_t : t \leq T\}$ by F. In addition, $\Omega = [\omega_{it}]_{N\times T}$ and $\mathcal{E} = [\varepsilon_{it}]_{N\times T}$.

Assumption 4.1 (Factor and loading). *(i) $\{f_t\}_{t\leq T}$ are independent and identically distributed. $\{\alpha_i\}_{i\leq N}$ are independent across i, and α is independent of F and $\mathcal{E}$. We assume that $\{\beta_i\}_{i\leq N}$ and θ are nonrandom. (ii) There are constants $c, C > 0$ such that the following inequalities hold with probability converging to 1:*

$$c < \psi_K\left(\frac{1}{N}\sum_{i=1}^{N}\beta_i\beta_i'\right) \leq \psi_1\left(\frac{1}{N}\sum_{i=1}^{N}\beta_i\beta_i'\right) < C, \quad c < \psi_K\left(\frac{1}{T}\sum_{t=1}^{T}\upsilon_t\upsilon_t'\right) \leq \psi_1\left(\frac{1}{T}\sum_{t=1}^{T}\upsilon_t\upsilon_t'\right) < C,$$

where $\upsilon_t = f_t - \mathbb{E}f_t$.
(iii) With probability converging to 1, we have $\max_t ||f_t|| \leq C\mu^{\frac{1}{2}}$ *for some constant* $C > 0$. *Here,* μ *is allowed to increase as* N, T *increase. In addition, we have* $\max_i ||\beta_i|| < C$ *for some constant* $C > 0$.

Assumption 4.1 (i) imposes restrictions on the dependence structure of the DGP. Allowing for (serially) weakly dependent factors is possible by imposing extra mixing conditions for the time series. Assumption 4.1 (ii) is a standard assumption in the factor models (e.g., Stock and Watson (1998, 2002)). It ensures that the factors are asymptotically identified (up to a rotation) and that $\text{Cov}(r_t)$ has K growing singular values. In addition, the assumptions help us to bound operator norms of some matrices and their inverse matrices.

Assumption 4.2 (ε_{it} and ω_{it}). *(i) Conditioning on* F, ε_{it} *is i.i.d. zero-mean, sub-Gaussian random variable such that* $\mathbb{E}[\varepsilon_{it}|F] = 0$, $\mathbb{E}[\varepsilon_{it}^2|F] = \sigma^2$, *and* $\|\varepsilon_{it}\|_{subG} \leq C\sigma$ *for some constant* $C > 0$.
(ii) Ω *is independent of* $\mathcal{E}$. *Conditioning on* F, ω_{it} *is independent across* t. *In addition,* $\mathbb{E}[\omega_{it}|F] = \mathbb{E}[\omega_{it}] = p_i$ *and there is a constant* $\underline{p}$ *such that* $0 < \underline{p} \leq p_i$ *for all* i.
(iii) Let $\ddot{\omega}_{it} = \omega_{it} - p_i$ *and* $\ddot{\Omega} = [\ddot{\omega}_{it}]_{N\times T}$. *Denote the columns of* $\ddot{\Omega}$ *and* $\mathcal{P}_\Omega(\mathcal{E}) = \Omega \circ \mathcal{E}$ *by* $\ddot{\omega}_t$ *and* ς_t, *respectively. Then,* $\{\ddot{\omega}_t\}_{t\leq T}$ ($\{\varsigma_t\}_{t\leq T}$) *are independent sub-gaussian random vectors with* $\mathbb{E}[\ddot{\omega}_t] = 0$ ($\mathbb{E}[\varsigma_t] = 0$); *more specifically, there is* $C > 0$ *such that for* $a_t \in \{\ddot{\omega}_t, \varsigma_t\}$,

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

(iv) 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* $\mathcal{C}_{g(i)}$, $\{w_{j_1t}, \ldots, w_{j_mt}\}$ *is independent from* w_{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.*
(v) We have $\max_t \max_i \sum_{j=1}^N |Cov(\omega_{it}, \omega_{jt}|F)| < C$.

This is the assumption for the error and the missing pattern. Contrary to Giglio et al. (2020), we allow the cross-sectional correlation and the heterogeneous observation probability for the missing pattern and use inverse probability weighting in the nuclear norm penalized estimation.

Assumption 4.3 (Order for parameters). *We have*

$$\frac{\mu^{\frac{1}{2}}K^2\max\{\sqrt{N}\log^{\frac{5}{2}}N, \sqrt{T}\log^{\frac{5}{2}}T\}}{\min\{N,T\}} = o_p(1), \quad \frac{\vartheta\mu^{\frac{3}{2}}K^{\frac{3}{2}}\max\{\sqrt{N}\log N, \sqrt{T}\log T\}}{\min\{N,T\}} = o_p(1).$$

Assumption 4.4 (Eigengap). *Let* $M^\star = [M_{it}^\star]_{N\times T}$ *where* $M_{it}^\star = \beta_i'l_t$, $l_t = (\upsilon_t - \bar{\upsilon})$, $\bar{\upsilon} = \frac{1}{T}\sum_{t=1}^T \upsilon_t$.

Then, there is $c > 0$ such that with probability converging to 1,

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

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

Assumptions 4.3 and 4.4 work similarly to Assumptions 3.3 (i), (ii) and 3.4 in Section 3. The assumptions are required to estimate the latent factors and loadings via the matrix completion method in Step 1 (b) of Algorithm 2. Here, we allow the number of factors K to increase slowly.

Assumption 4.5 (Moment bounds). *There are $c, C > 0$, such that*
(i) $\mathbb{E}||f_t||^4 < C$.
(ii) We have

$$\frac{\mathbb{E}\max_{i,j,d\leq N,k,l\leq K,t\leq T}\xi_{i,j,d,k,l,t}^4}{\max_{i,j,d\leq N,k,l\leq K,t\leq T}\mathbb{E}\xi_{i,j,d,k,l,t}^4} \leq C(\log N)^2 T,$$

where $\xi_{i,j,d,k,l,t} \in \{\varepsilon_{it}\varepsilon_{jt}, \varepsilon_{it}, \varepsilon_{it}\varrho_t, \varepsilon_{it}^2\varrho_{kt}^2, \varepsilon_{it}f_{kt}, \varepsilon_{it}^2 f_{kt}, \varepsilon_{it}^2 f_{kt}f_{lt}, \varepsilon_{it}^2\varepsilon_{jt}\varepsilon_{dt}\}$ and $\varrho_t = \frac{1}{\sqrt{N}}\beta'\mathcal{E}_t$ where $\mathcal{E}_t = [\varepsilon_{1t}, \ldots, \varepsilon_{Nt}]'$.
(iii) $||\Sigma_f^{-1}|| < C$ *and* $\mathbb{E}\left[\varepsilon_{it}^2\left(1 - \upsilon_t'\Sigma_f^{-1}\theta\right)^2\right] > c$, *where* $\Sigma_f = \mathbb{E}\upsilon_t\upsilon_t'$ *and* $\upsilon_t = f_t - \mathbb{E}f_t$.

Lastly, we assume some moment bounds. Assumption 4.5 (ii) requires that interchanging "max" with "$\mathbb{E}$" on $\xi_{i,j,d,k,l,t}$ imposes an additional term no larger than $O(T\log^2 N)$. It is a technical condition for applying concentration inequalities in Chernozhukov et al. (2016) to establish

$$\max_{\substack{i,j,d\leq N,\\ k,l\leq K}}\left|\frac{1}{T}\sum_{t=1}^{T}\xi_{i,j,d,k,l,t} - \mathbb{E}\xi_{i,j,d,k,l,t}\right| = O_p\left(\sqrt{\frac{\log N}{T}}\right),$$

which is used in many parts of the proof for uniform bounds. Then, under the above assumptions, we can derive the asymptotic distribution of the estimates for α_i for all $i = 1, \ldots, N$.

Theorem 4.1. *Suppose that Assumptions 4.1 - 4.5 hold. Then, when $\max_i T_i \log N = o(N)$, we have uniformly in $1 \leq i \leq N$,*

$$(i) \quad \sqrt{T_i}\left(\widehat{\alpha}_i - \alpha_i\right) = \frac{1}{\sqrt{T_i}}\sum_{t\in\mathcal{T}_i}\varepsilon_{it}\left(1 - \upsilon_t'\Sigma_f^{-1}\theta\right) + o_p\left(\frac{1}{\sqrt{\log N}}\right),$$

$$(ii) \quad \frac{\widehat{\alpha}_i - \alpha_i}{se\left(\widehat{\alpha}_i\right)} \xrightarrow{D} \mathcal{N}(0,1),$$

where $se\left(\widehat{\alpha}_i\right) = \frac{1}{\sqrt{T_i}}\widehat{\mathcal{V}}_{\alpha_i}^{\frac{1}{2}}$, $\widehat{\mathcal{V}}_{\alpha_i} = \frac{1}{T_i}\sum_{t\in\mathcal{T}_i}\widehat{\varepsilon}_{it}^2(1 - \widehat{\upsilon}_t'\widehat{\Sigma}_f^{-1}\widehat{\theta})^2$, $\widehat{\varepsilon}_{it} = r_{it} - \bar{r}_i - \widehat{\beta}_i'\widehat{\upsilon}_t$ is the residual, and $\widehat{\Sigma}_f = \frac{1}{T}\sum_{t=1}^{T}\widehat{\upsilon}_t\widehat{\upsilon}_t'$.

Then, based on Theorem 4.1 (ii), we can construct the t-statistics and obtain corresponding p-values for individual alphas described in (4.1).

4.2 Heterogeneous treatment effect estimator

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, $z_{it}^{(0)}$ and $z_{it}^{(1)}$ where $z_{it}^{(0)}$ denotes the potential outcome of the untreated situation and $z_{it}^{(1)}$ denotes the potential outcome of the treated situation. Importantly, among potential outcomes $z_{it}^{(0)}$ and $z_{it}^{(1)}$, we can observe only one realized outcome $z_{it}^{(\Upsilon_{it})}$ where $\Upsilon_{it} = 1\{\text{unit } i \text{ is treated at period } t\}$. Hence, we have two incomplete potential outcome matrices, $Z^{(0)}$ and $Z^{(1)}$, 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 $\iota \in \{0, 1\}$,

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

where $\omega_{it}^{(\iota)} = 1\{z_{it}^{(\iota)} \text{ is observed}\}$, 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 $z_{it}^{(1)}$ when there is a treatment on (i, t) and observe $z_{it}^{(0)}$ 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 of $M_{it}^{(0)}$ and $M_{it}^{(1)}$ for the group $\mathcal{G}$ 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 4.6 (Sieve representation). *(i) For all $\iota \in \{0,1\}$, $\{h_t^{(\iota)}(\cdot)\}_{t\leq T}$ belong to ball $\mathcal{H}\left(\mathcal{Z}, \|\|_{L_2}, C\right)$ inside a Hilbert space spanned by the basis $\{\phi_r\}_{r\geq 1}$, with a uniform L_2-bound C:*

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

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 4.7 (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 $\|\varepsilon_{it}\|_{subG} \leq C\sigma$ 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.[22] *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 $\ddot{\Upsilon}_{it} = \Upsilon_{it} - \mathbb{E}[\Upsilon_{it}]$ and $\ddot{\Upsilon} = [\ddot{\Upsilon}_{it}]_{N\times T}$. Let $\mathcal{P}_{\Upsilon}(\mathcal{E}) = \Upsilon \circ \mathcal{E}$ and $\mathcal{P}_{1-\Upsilon}(\mathcal{E}) = (\mathbf{1}_N\mathbf{1}_T' - \Upsilon) \circ \mathcal{E}$. Denote the columns of $\ddot{\Upsilon}$, $\mathcal{P}_{\Upsilon}(\mathcal{E})$, $\mathcal{P}_{1-\Upsilon}(\mathcal{E})$ by $\ddot{\Upsilon}_t$, $\varsigma_t^{(1)}$, $\varsigma_t^{(0)}$, respectively. Then, $\{\ddot{\Upsilon}_t\}_{t\leq T}$ ($\{\varsigma_t^{(\iota)}\}_{t\leq T}$) are independent sub-gaussian random vectors with $\mathbb{E}[\ddot{\Upsilon}_t] = 0$ ($\mathbb{E}[\varsigma_t^{(\iota)}] = 0$ for $\iota \in \{0,1\}$); more specifically, there is $C > 0$ such that for $a_t \in \{\ddot{\Upsilon}_t, \varsigma_t^{(0)}, \varsigma_t^{(1)}\}$,

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

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

(iv) 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 $\vartheta \geq 1$ 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.
(v) 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 4.7 implies Assumption 3.2 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 4.8 (Slowly increasing ϑ, K and $\min\{|\mathcal{I}|_o, |\mathcal{T}|_o\}$)**.**

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

Assumption 4.9 (Eigengap)**.** *There are $c, C > 0$ such that with probability converging to 1, for all $\iota \in \{0, 1\}$, $\psi_1^{(\iota)} \leq C\psi_K^{(\iota)}$ and*

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

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

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

Theorem 4.2. *Suppose Assumptions 4.6 - 4.9 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. 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 4.3 (Feasible CLT). *Under the assumptions of Theorem 4.2, we have*

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

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

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

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

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

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

[23] 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

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.

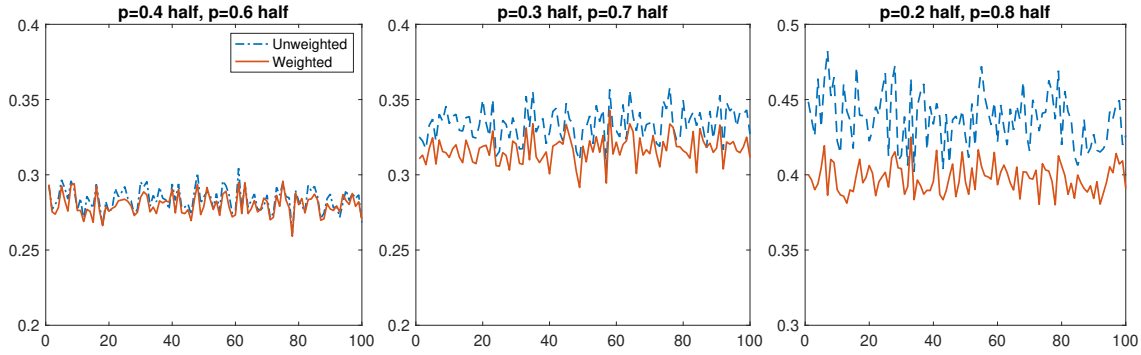
Third, we study the finite sample distributions for standardized estimates defined as

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

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

and Pelger (2020), and Fan et al. (2020).

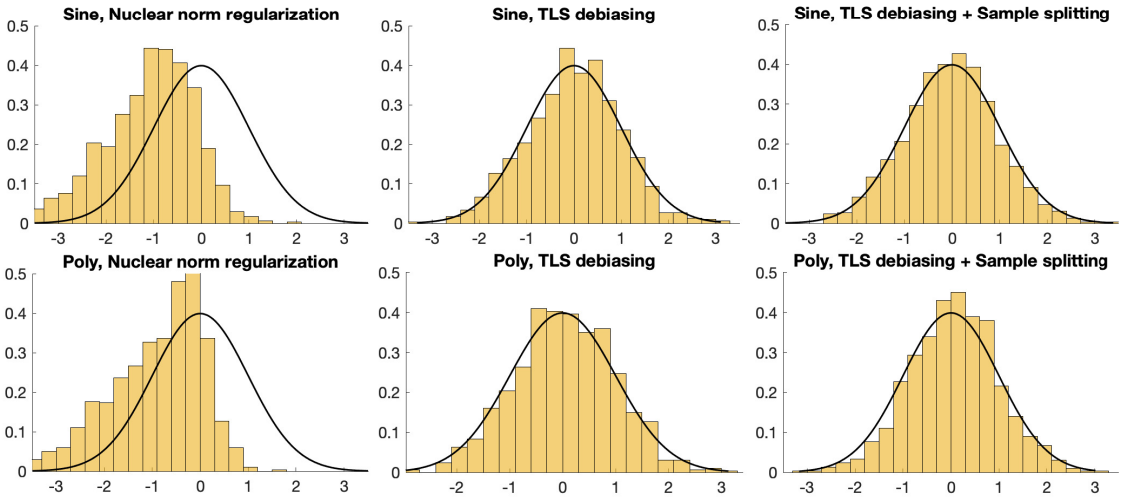
Figure 1: Frobenius norm of estimation errors, $||\widetilde{M} - M||_F/\sqrt{NT}$



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.

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: 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.

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.

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.

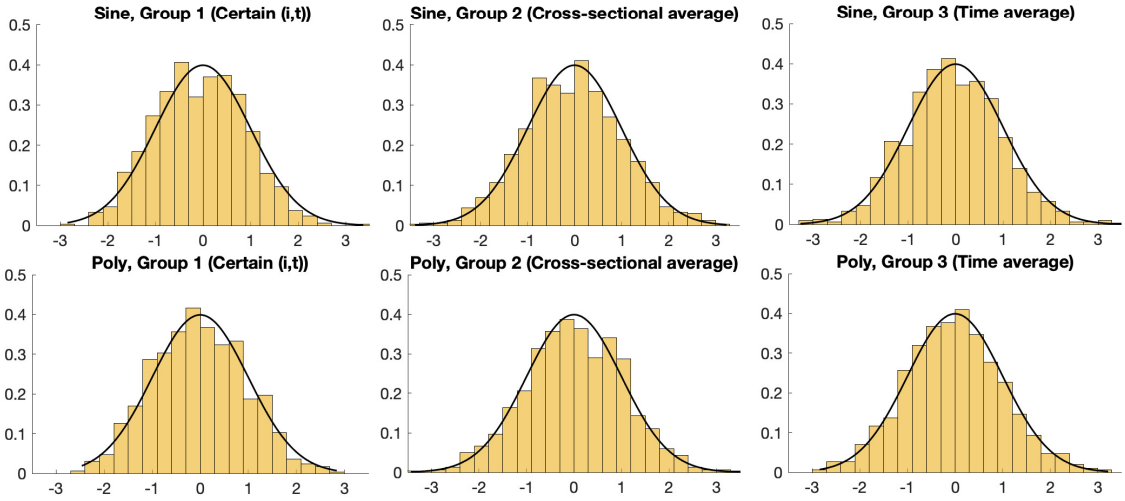
Table 2 presents the coverage probabilities of the confidence interval which is given by

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

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.

In addition, to show that our asymptotic theory works well with various groups, Figure 3

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 $\mathcal{G}_3$.

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

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

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

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

Function $h_t(\zeta)$	Sample size N	T	Group $\mathcal{G}_1$ (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%

Section 4.2, for each $\iota \in \{0,1\}$, we generate the data from

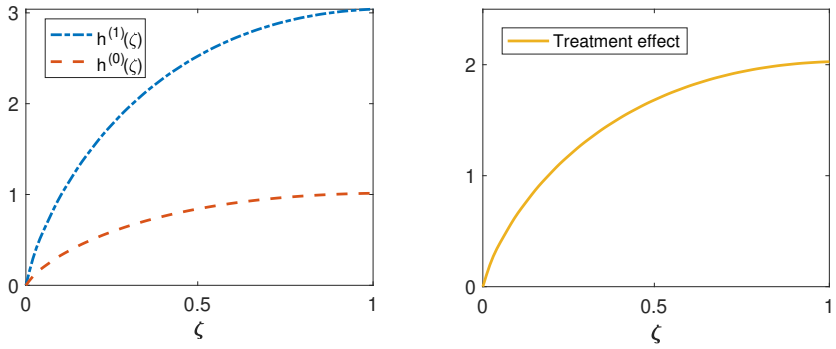
$$z_{it}^{(\iota)} = h_t^{(\iota)}(\zeta_i) + \varepsilon_{it}, \quad \text{if } \Upsilon_{it} = \iota,$$

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.

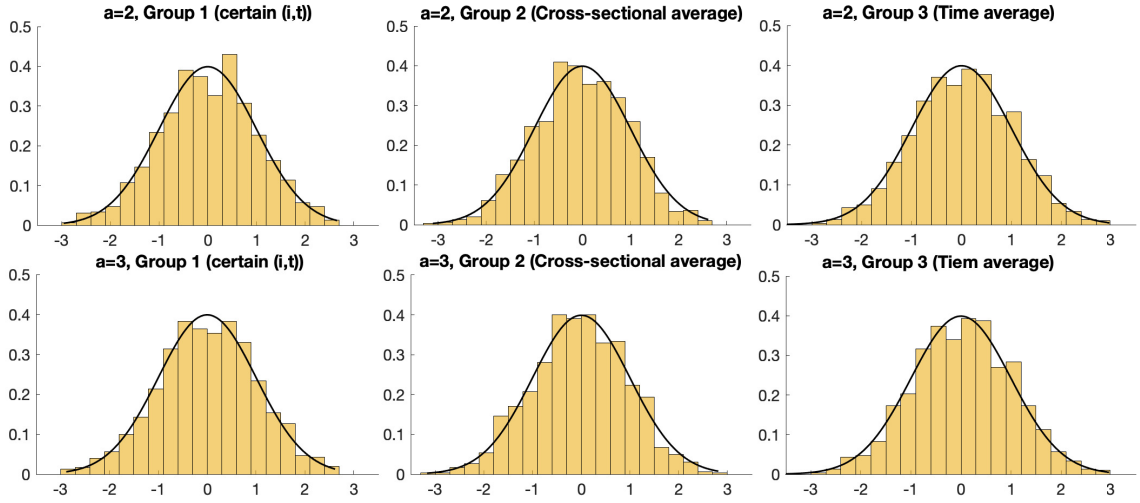
Figure 4: Shape of function $h_t^{(\iota)}(\zeta)$ and treatment effect function ($U_{t,r} = 1$, $a = 2$)



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

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

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

As we expected in the theory, it shows that the standardized estimates of the average treatment effect estimators of all groups have similar distributions to the standard normal distribution. In addition, 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.

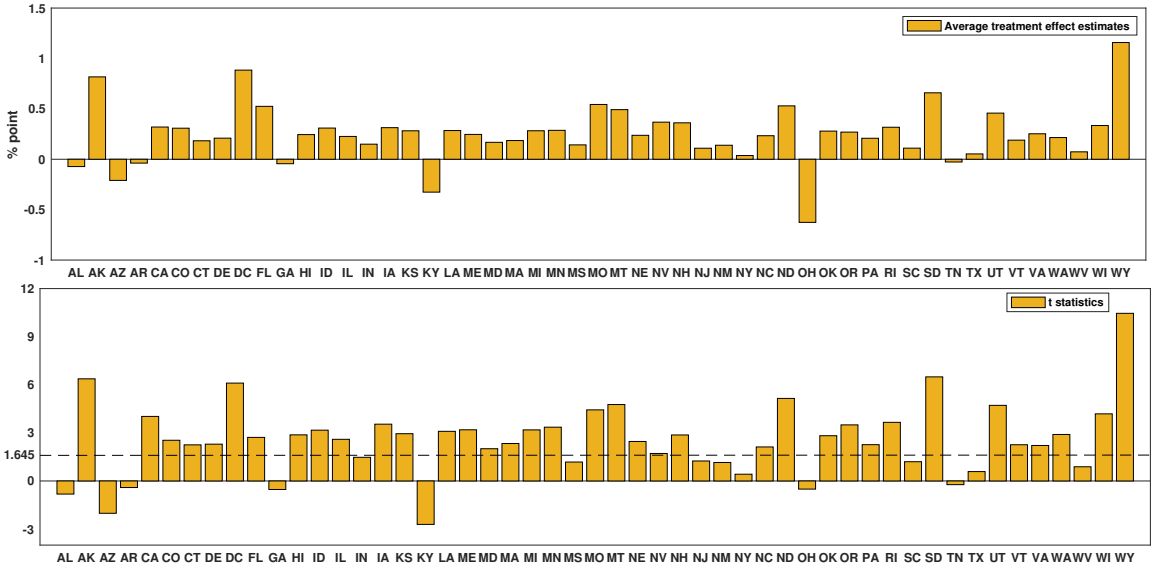
Here, the hypothesis we want to test is whether federal funds are disproportionately targeted to states where the incumbent president is supported in the past presidential election. We use data on federal outlays for the 50 U.S. states with the District of Columbia from 1953 to 2018.[24] Following the model in Section 4.2, we set the treatment indicator as $\Upsilon_{it} = 1\{\text{the state } i \text{ supported the president of year } t \text{ in the presidential election}\}$. If the candidate whom the state i supported in the previous presidential election is same as the president at year t, we consider it as "treated" and otherwise, we consider it as "untreated". In addition, for the outcome variable z_{it}, we use the following ratio:

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

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

This is each state's (per-capita) portion of the federal grant at each year. In fact, the per-capita federal grant, $\tilde{z}_{it}$, 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 z_{it} instead.

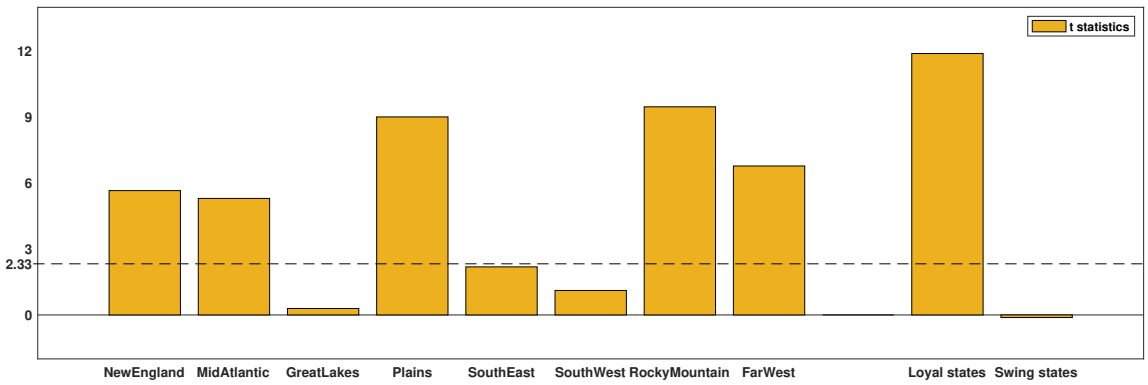
Figure 6: Time average treatment effect estimates of each state and corresponding t-statistics



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 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),[25] 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

[25] "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.

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

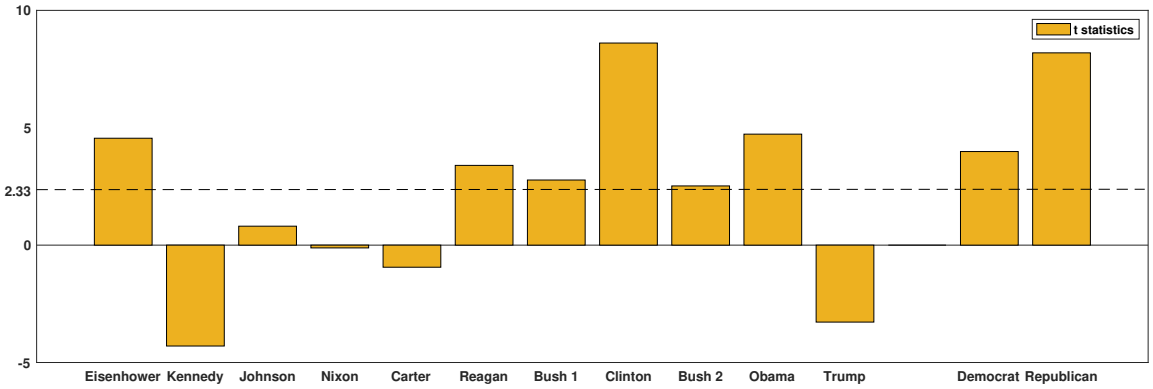


Figure 9: Test statistics for the average treatment effect before 1980 and after 1981

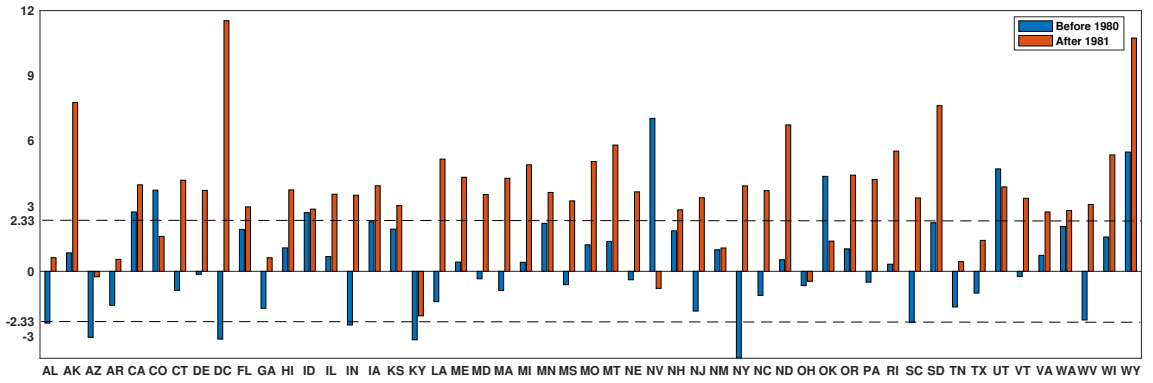


Figure 8 shows the test statistics for the average of the treatment effect of each president. Although there exist some exceptions, there are no statistically significant positive treatment effects before Carter, while there are significant positive treatment effects after Reagan. From Figure 9, we can 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 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.[26]

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

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 the alpha test in empirical asset pricing and the average treatment effect estimator as the applications. 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.

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 (2018): "Matrix completion methods for causal panel data models," Tech. rep., National Bureau of Economic Research.

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., 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., D. CHETVERIKOV, AND K. KATO (2016): "Testing many moment inequalities," Tech. rep., cemmap working paper.

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.

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.

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.

STOCK, J. H. AND M. W. WATSON (1998): "Diffusion indexes," *NBER working paper*.

——— (2002): "Forecasting using principal components from a large number of predictors," *Journal of the American statistical association*, 97, 1167–1179.

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*.