Inference for low-rank completion without sample splitting with application to treatment effect estimation

Jungjun Choi [a,*], Hyukjun Kwon [b], Yuan Liao [b]

[a] *Columbia University, 1255 Amsterdam Avenue, New York, NY 10027, United States*
[b] *Rutgers University, 75 Hamilton St, New Brunswick, NJ 08901, United States*

ARTICLE INFO

ABSTRACT

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

1. Introduction

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

In this paper, we focus on the following approximate low-rank model with a factor structure:

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

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

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

---

* Corresponding author.

E-mail addresses: jc5805@columbia.edu (J. Choi), hk731@rutgers.edu (H. Kwon), yuan.liao@rutgers.edu (Y. Liao).

(2014) provide algorithms to compute the nuclear norm penalized estimator. However, research on inference is still limited. This is because the shrinkage bias caused by the penalization, as well as the lack of the closed-form expression of the estimator, hinders the distributional characterization of the estimator.

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

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

Note that steps 2–3 are only conducted once without further iterations.

An important contribution is that we do not rely on the sample splitting to make inference, but simply use the full (observed) sample in every step of our procedure. There are at least three advantages to avoid sample splitting. First, the resulting estimator using sample splitting is unstable and random even conditioning on the data. Second, sample splitting requires relatively large T in practice, because it practically works with only $T/2$ observations. This is demanding in applied micro applications when T is just a few decades. In the simulation study, we show that the performance of the estimator using sample splitting is worse than that of the estimator without sample splitting when T is relatively small. Lastly, sample splitting increases computational costs in multiple tests because for each target time 't', we need to use different sample splitting.

Technically, we apply a new approach to showing the negligibility of the potential bias terms, by making use of a hypothetically defined *auxiliary leave-one-out* (ALOO) estimator. We emphasize the word "auxiliary" because it is only introduced in the technical argument, but *not* implemented in the estimation. So it is a hypothetical estimator, which is to be shown that it is

(i) asymptotically equivalent to the initial estimator for β in Step 1 and

(ii) independent of the sample used in the least squares estimation, namely, the sample in period t. Using the ALOO estimator, we can separate out the part in the initial estimator for β, which is correlated with the sample in period t. Once we separate out the correlated part, we can enjoy a similar effect to the sample splitting. And we show the separated correlated part is sufficiently small. Importantly, the leave-one-out estimator only appears in the proof as an auxiliary point of the initial estimator for β, so we do not need to compute it in the estimation procedure, which allows us to remove the sample splitting step without implementing any additional steps.

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

1.1. Relation to the literature

Very recently, some studies proposed the ways of achieving unbiased estimation for the inference of the nuclear norm penalized estimator. Chernozhukov et al. (2019, 2023) propose a two-step least square procedure with sample splitting, which estimates the factors and loadings successively using the least square estimations. As we discussed earlier, sampling splitting comes with several undesirable costs.

The idea of the ALOO estimator has been employed in other recent works such as Ma et al. (2019), Chen et al. (2019, 2020a,b) and Yan et al. (2021) as well. Among them, in particular, Chen et al. (2019) pioneered using this idea to convex relaxation of low-rank inference. This paper has some important contributions compared to Chen et al. (2019).

1. We consider a general nonparametric panel model which is an approximate low-rank model rather than an exact low-rank model.
2. This paper accommodates more general data-observation patterns: the heterogeneous observational probabilities and the correlated observation patterns by assuming the cluster structure and allowing dependence within a cluster.
3. The inferential theory for the average treatment effect estimation is provided as an application.
4. We formally address a technical issue concerning the ALOO estimator. The ALOO estimator is to be (hypothetically) calculated by using the gradient descent iteration from the leave-one-out problem, which rules out, for example, samples in period t. This exclusion is designed to guarantee the independence between the leave-one-out estimator and the period t sample. However, due to the non-convexity of the loss functions, the gradient descent iteration must stop where the gradient of the loss function is sufficiently "small". If this stopping point depends on the sample in period t, as in Chen et al. (2019) who derive the stopping point from the problem using the full sample, the leave-one-out estimator using this stopping point may not be truly independent of the sample in period t. This dependence frustrates the analysis of the bounds regarding the leave-one-out estimator. We provide two solutions for this potential dependence issue to be detailed in the paper.
5. Our method does not have an explicit debias step, but is based on refitting least squares. While we do not claim that this estimator is advantageous over the explicit debiasing method, we view our estimator as the natural extension of "post model selection methods" to the low rank framework.

Other related works on inference include Xia and Yuan (2021), Xiong and Pelger (2023), and Jin et al. (2021). We compare these methods with ours in simulations.

Lastly, a comparison with other literature that takes advantage of a low-rank model to estimate the treatment effect would be helpful. The close connection between low-rank completion and treatment effect estimation was first made formal by Athey et al. (2021) who showed that the nuclear norm regularization can be useful for causal panel data by presenting the convergence rate of the estimator. Another line of research develops inferential theories without sample splitting under different patterns of treatment assignments. For example, Arkhangelsky et al. (2021) introduce a new estimator for the overall average treatment effect, that is designed to be more robust than the traditional difference-in-differences and synthetic control estimators, under the block-structured treatment assignments. Farias et al. (2021) also focus on the overall average treatment effect under the assumption that the treatment assignment is sufficiently orthogonal to the tangent space of the underlying low-rank matrix parameter. They show that their assumption may incorporate the treatment patterns in the synthetic control literature. On the other hand, Agarwal et al. (2021) and Bai and Ng (2021) consider heterogeneous treatment effects. While they allow strongly dependent treatment assignments, they assume the availability of fully observed submatrices, where they apply PCA-based methods.

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

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

2. Model and estimation

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

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

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

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

where $\beta_i = (\phi_1(\zeta_i), \ldots, \phi_K(\zeta_i))'$ and $F_t = (\kappa_{t,1}, \ldots, \kappa_{t,K})'$. Here, M_{it}^R is the sieve approximation error and, for all $1 \le r \le K$, $\phi_r(\zeta_i)$ is the sieve transformation of ζ_i using the basis function $\phi_r(\cdot)$ and $\kappa_{t,r}$ is the sieve coefficient. Then,

$$M = [M_{it}]_{N \times T}, \quad M_{it} = h_t(\zeta_i)$$

can be successfully represented as the approximate factor structure. In matrix form, we can represent the model as

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

where we denote $Y = [y_{it}]_{N \times T}$, $M = [M_{it}]_{N \times T}$, $M^\star = [M_{it}^\star]_{N \times T}$, $M^R = [M_{it}^R]_{N \times T}$, $\beta = [\beta_1, \ldots, \beta_N]'$, $F = [F_1, \ldots, F_T]'$, and $\mathcal{E} = [\varepsilon_{it}]_{N \times T}$.

Let $\mathcal{M} := (\beta, F, M^R)$ be the triplet of random matrices that compose M. In the paper, we allow the heterogeneous observation probability, i.e., $P(\omega_{it} = 1) = p_i$ and denote $\Pi = \text{diag}(p_1, \ldots, p_N)$. Here, we shall assume the sieve dimension K is pre-specified by researchers and propose some data-driven ways of choosing K in Appendix A.

2.1. Nuclear norm penalized estimation with inverse probability weighting

To accommodate the heterogeneous observation probability, this paper uses 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), Schnabel et al. (2016) and Little and Rubin (2019)), in addition to the nuclear norm penalization:

$$\widetilde{M} := \arg\min_{A \in \mathbb{R}^{N \times T}} \frac{1}{2} \|\widehat{\Pi}^{-\frac{1}{2}} \Omega \circ (A - Y)\|_F^2 + \lambda \|A\|_* \tag{2.2}$$

where $\widehat{\Pi} = \text{diag}(\widehat{p}_1, \ldots, \widehat{p}_N)$, and $\widehat{p}_i = \frac{1}{T} \sum_{t=1}^{T} \omega_{it}$ for each $i \le N$, $\Omega = [\omega_{it}]_{N \times T}$ and $\circ$ denotes the Hadamard product. As noted in Ma and Chen (2019), this inverse probability weighting debiases the objective function itself. If there is heterogeneity in the observation probability, $\|\Pi^{-\frac{1}{2}} \Omega \circ (A - Y)\|_F^2$ is an unbiased estimate of $\|A - Y\|_F^2$, which we would use if there is no missing entry, in the sense that $\mathbb{E}_\Omega[\|\Pi^{-\frac{1}{2}} \Omega \circ (A - Y)\|_F^2] = \|A - Y\|_F^2$, while $\|\Omega \circ (A - Y)\|_F^2$ is biased.

2.2. Estimation procedure

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

Algorithm 1 Constructing the estimator for M.

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

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

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

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

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

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.

2.3. A general discussion of the main idea

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

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

where H_1 is some rotation matrix.[1] This term represents the effect of the bias of the nuclear-norm penalization since $\widetilde{\beta}_j$ is derived from the nuclear-norm penalized estimator. Chernozhukov et al. (2019, 2023) resort to sample splitting to show the asymptotic negligibility of R_t.

2.3.1. The auxiliary leave-one-out method

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

$$L^{\text{full}}(B, F) = \frac{1}{2} \| \Pi^{-\frac{1}{2}} \Omega \circ (BF' - Y) \|_F^2 + \frac{\lambda}{2} \|B\|_F^2 + \frac{\lambda}{2} \|F\|_F^2$$

$$= \frac{1}{2} \| \Pi^{-\frac{1}{2}} \Omega \circ (BF' - Y) \|_{F,(-t)}^2 + \frac{1}{2} \| \Pi^{-\frac{1}{2}} \Omega \circ (BF' - Y) \|_{F,t}^2 + \frac{\lambda}{2} \|B\|_F^2 + \frac{\lambda}{2} \|F\|_F^2 \qquad (2.3)$$

$$L^{(-t)}(B, F) = \frac{1}{2} \| \Pi^{-\frac{1}{2}} \Omega \circ (BF' - Y) \|_{F,(-t)}^2 + \frac{1}{2} \| BF' - M^\star \|_{F,t}^2 + \frac{\lambda}{2} \|B\|_F^2 + \frac{\lambda}{2} \|F\|_F^2. \qquad (2.4)$$

Here, $\| \cdot \|_{F,(-t)}$ denotes the Frobenius norm which is computed ignoring tth column and $\| \cdot \|_{F,t}$ is the Frobenius norm of only tth column. Note that the only difference between (2.3) and (2.4) is that the tth column of the goodness of fit part in (2.3) is replaced by its conditional expectation in (2.4). So, $\{\omega_{jt}, \varepsilon_{jt}\}_{j \leq N}$ is excluded from the problem (2.4).

We emphasize that (i) both problems defined above are non-convex; (ii) both problems are "auxiliary", meaning that they are introduced only for proofs, not actually implemented. (iii) Optimizing $L^{(-t)}(B, F)$ is an auxiliary leave-one-out (ALOO) problem, leading to the ALOO estimator $\breve{\beta}^{(-t)}$ to be discussed below.

Because of the non-convexity, both hypothetical problems should be computed iteratively until the gradients of the non-convex loss functions become "sufficiently small". However, the gradients do not monotonically decrease as iteration proceeds since the problem is non-convex. So, one cannot let it iterate until convergence is reached, but has to stop at the point where the gradient is small enough. The choice of this "stopping point" is crucial in the analysis of the residual terms. Chen et al. (2019) define the stopping point using the full sample problem (2.3), which potentially causes dependence problem of the leave-one-out estimators. We propose two approaches of addressing this issue.

---

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

Approach I First, we derive the stopping point from the leave-one-out problem (2.4). Let $B^{\text{full},\tau}$ and $B^{(-t),\tau}$ be τth iterates of the gradient decent for (2.3) and (2.4), respectively. Fix t of interest and suppose we iterate both problems τ_t times, where τ_t depends on t. Define the "solutions" at τ_tth iterations:

$$\check{\beta}^{\text{full},t} = B^{\text{full},\tau_t} \quad \text{and} \quad \check{\beta}^{(-t)} = B^{(-t),\tau_t}.$$

Hence, they share the same stopping point τ_t. Noticeably, although $\check{\beta}^{\text{full},t}$ is a solution for the full sample problem (2.3), it depends on t through τ_t. In this first approach, we derive the stopping point from the ALOO problem (2.4). Hence, it ensures that the estimator $\check{\beta}^{(-t)}$ using this stopping point is independent of the tth period sample, $\{\omega_{jt}\varepsilon_{jt}\}_{j\leq N}$. This introduces nontrivial technical challenges. Namely, τ_t, being derived from the problem $L^{(-t)}(B,F)$, depends on t, so the "full-problem" solution $\check{\beta}^{\text{full},t}$ would therefore also depend on t. We derive the uniform convergence of both $\check{\beta}^{\text{full},t}$ and $\check{\beta}^{(-t)}$ uniformly in $t = 1, \dots, T$.

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

1. First, decompose R_t into two terms:

$$\begin{aligned}
R_t &= \frac{1}{\sqrt{N}} \sum_{j=1}^{N} \omega_{jt}\varepsilon_{jt}(\widetilde{\beta}_j - H_1'\beta_j) \\
&= \underbrace{\frac{1}{\sqrt{N}} \sum_{j=1}^{N} \omega_{jt}\varepsilon_{jt}(\widetilde{\beta}_j - \check{\beta}_j^{(-t)})}_{:=a} + \underbrace{\frac{1}{\sqrt{N}} \sum_{j=1}^{N} \omega_{jt}\varepsilon_{jt}(\check{\beta}_j^{(-t)} - H_1'\beta_j)}_{:=b}.
\end{aligned} \tag{2.5}$$

2. $\max_t \|b\| = o_P(1)$ can be shown relatively easily due to the genuine independence between $\check{\beta}^{(-t)}$ and $\{\omega_{jt}\varepsilon_{jt}\}_{j\leq N}$, which is along the same line as sample splitting. Importantly, it is crucial to require that τ_t should not depend on observations of time t. So the stopping time should be defined carefully, which is one of the main technical contributions of the paper.

3. In addition, $\max_t \|a\| = o_P(1)$ comes from the following two rationales:

$$a = \frac{1}{\sqrt{N}} \sum_{j=1}^{N} \omega_{jt}\varepsilon_{jt}(\widetilde{\beta}_j - \check{\beta}_j^{\text{full},t}) + \frac{1}{\sqrt{N}} \sum_{j=1}^{N} \omega_{jt}\varepsilon_{jt}(\check{\beta}_j^{\text{full},t} - \check{\beta}_j^{(-t)}).$$

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

   Their loss functions (2.3) and (2.4) are very similar and they share the same stopping point τ_t. Therefore, $\max_t \|\check{\beta}^{\text{full},t} - \check{\beta}^{(-t)}\|$ is sufficiently small. Following the guidance of Chen et al. (2020b), we apply the mathematical induction.

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

   Note that $\check{\beta}^{\text{full},t}$ is derived from the non-convex problem (2.3) and $\widetilde{\beta}$ comes from the nuclear norm penalization (2.2). Although the loss functions (2.2) and (2.3) are seemingly distinct, their penalty terms are closely related in the sense that

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

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

   Hence, we have $\max_t \|R_t\| = o_P(1)$.

Approach II Alternatively, we can follow the definition of the stopping point in Chen et al. (2019), which uses the full sample. And then, we correct their proof by showing that, although the leave-one-out estimator is not independent of the sample data in period t, we can still obtain a uniform bound over iterations. Denote the stopping point from Chen et al. (2019) as τ^*. In lieu of $(B^{\text{full},\tau_t}, B^{(-t),\tau_t})$, we use $(B^{\text{full},\tau^*}, B^{(-t),\tau^*})$ as the solutions for (2.3) and (2.4), respectively.

Recall the decomposition (2.5). The analysis of term a is analogous to the previous case. Regarding term b, we highlight that $\check{\beta}^{(-t)}$, which is $B^{(-t),\tau^*}$, is not independence from the sample in period t, i.e., $\{\omega_{jt}\varepsilon_{jt}\}_{j\leq N}$, since the stopping point τ^* depends on it. We will provide a uniform bound over iteration τ and period t for term b:

$$\begin{aligned}
\max_t \|b\| &= \max_t \left\| \frac{1}{\sqrt{N}} \sum_{j=1}^{N} \omega_{jt}\varepsilon_{jt}(\check{\beta}_j^{(-t)} - H_1'\beta_j) \right\| = \max_t \left\| \frac{1}{\sqrt{N}} \sum_{j=1}^{N} \omega_{jt}\varepsilon_{jt}(B_j^{(-t),\tau^*} - H_1'\beta_j) \right\| \\
&\leq \max_t \max_\tau \left\| \frac{1}{\sqrt{N}} \sum_{j=1}^{N} \omega_{jt}\varepsilon_{jt}(B_j^{(-t),\tau} - H_1'\beta_j) \right\| = o_P(1).
\end{aligned}$$

In either way, we can successfully show the negligibility of R_t uniformly in t without resorting to sample splitting. We highlight that the first approach is more natural in the sense that it automatically ensures the independence that we need for term b. Our first approach, while technically more involved, is potentially more applicable to general machine learning inferences that rely on auxiliary leave-one-out estimators, because of the natural independence. In contrast, it is unclear whether the second approach is still applicable in other cases.

2.3.2. Why is the auxiliary leave-one-out problem defined in this way?

It is natural to ask why would not we define the ALOO estimator more naturally as the original estimator $\widetilde{\beta}$, but simply dropping the t th column from the data matrix in the optimization? One of the key differences between $L^{(-t)}(B, F)$ in (2.4) and the "more natural dropping-t" loss, is that the t th column in the least squares part of $L^{(-t)}(B, F)$ is not simply dropped, but is replaced by its expectation:

$$\mathbb{E}\|\Pi^{-\frac{1}{2}}\Omega \circ \left(BF' - Y\right)\|_{F,t}^2 = \|BF' - M^\star\|_{F,t}^2 + C$$

where the constant C does not depend on (B, F). The reason for defining the ALOO loss function in this way is to gain "hypothetical efficiency", so that the ALOO estimator would be closer to the full-sample estimator.

It is easier to understand the issue using a simple example. Consider estimating the mean $\mathbb{E}Y_t$ using iid data Y_t. The full-sample estimator $\widehat{\mu}$ is the solution to

$$\widehat{\mu} = \arg\min_\mu L(\mu), \quad \text{where } L(\mu) = \sum_{s=1}^T (Y_s - \mu)^2.$$

Now consider the ALOO version of this problem. Our definition of $L^{(-t)}(\mu)$ is *not* dropping Y_t, but replacing $(Y_t - \mu)^2$ with its expectation:

$$\breve{\mu}^{(-t)} = \arg\min_\mu L^{(-t)}(\mu), \quad \text{where } L^{(-t)}(\mu) = \sum_{s\neq t}(Y_s - \mu)^2 + \mathbb{E}(Y_t - \mu)^2.$$

The solution is then $\breve{\mu}^{(-t)} = \frac{1}{T}(\sum_{s\neq t} Y_s + \mathbb{E}Y_t)$. Then straightforward calculations can verify that $\breve{\mu}^{(-t)}$ (although infeasible) is more efficient and "closer" to the full-sample average $\widehat{\mu}$ than the naive dropping-t estimator $\bar{Y}_{-t} := \frac{1}{T-1}\sum_{s\neq t} Y_s$. For instance,

$$\frac{\mathrm{Var}(\breve{\mu}^{(-t)})}{\mathrm{Var}(\bar{Y}_{-t})} = \left(\frac{T-1}{T}\right)^2 < 1, \quad \frac{\mathbb{E}(\breve{\mu}^{(-t)} - \widehat{\mu})^2}{\mathbb{E}(\bar{Y}_{-t} - \widehat{\mu})^2} = \frac{T-1}{T} < 1.$$

The definitions of $L^{(-t)}(B, F)$ and $L^{(-t)}(\mu)$ also fulfill the intuition of the EM algorithm, which imputes the missing data in the loss function by its conditional expectations before optimizations, rather than simply dropping the missing values.

2.3.3. Singular vector estimation is unbiased

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

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

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

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

$$\sqrt{N}(\widehat{F}_t - H_1^{-1}F_t)$$
$$= -\sqrt{N}A^{-1}\frac{\partial Q_t(H_1^{-1}F_t, \beta H_1)}{\partial f} - \underbrace{\sqrt{N}A^{-1}\frac{\partial^2 Q_t(H_1^{-1}F_t, \beta H_1)}{\partial f\, \partial \mathrm{vec}(B)}\mathrm{vec}(\widetilde{\beta} - \beta H_1)}_{d} + o_P(1). \tag{2.6}$$

The first term is the score which leads to the asymptotic normality and the second term represents the effect of the β estimation which is subject to the shrinkage bias. The second term, while is the "usual bias" in a generic machine learning inference problem, can be shown to take the form:

$$d = \sqrt{N}\varphi H_1^{-1}F_t + o_P(1)$$

for some $\varphi = o_P(1)$. It has a useful feature of being on the space of F_t. Making use of this fact, (2.6) can be re-written as follows:

$$\sqrt{N}(\widehat{F}_t - H_2 F_t) = -\underbrace{\sqrt{N}A^{-1}\frac{\partial Q_t(H_1^{-1}F_t, \beta H_1)}{\partial f}}_{\text{asymptotically normal}} + o_P(1)$$

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

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

$$\widehat{M}_{it} = \hat{\beta}_i' \widehat{F}_t \approx \beta_i' H_2^{-1} H_2 F_t = M_{it}$$

which allows us to conduct inference successfully. This is how the two-step least squares procedure works.

3. Asymptotic results

3.1. Inferential theory

This section presents the inferential theory. We provide the asymptotic normality of the estimator of the group average of M_{it}. Our assumptions allow the rank K to grow, but slowly. Remind the following notation:

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

where $\beta_i = (\phi_1(\zeta_i), \dots, \phi_K(\zeta_i))'$ and $F_t = (\kappa_{t,1}, \dots, \kappa_{t,K})'$. Let $S_\beta = N^{-1} \sum_{i=1}^{N} \beta_i \beta_i'$, $S_F = T^{-1} \sum_{s=1}^{T} F_s F_s'$, and $Q = S_\beta^{1/2} S_F^{1/2}$.

Assumption 3.1 (*Sieve Representation*). (i) $\{h_t(\cdot)\}_{t \le T}$ belong to ball $\mathcal{H}\left(\mathcal{Z}, \|\cdot\|_{L_2}, C\right)$ inside a Hilbert space spanned by the basis $\{\phi_r\}_{r \ge 1}$, with a uniform L_2-bound C: $\sup_{h \in \mathcal{H}(\mathcal{Z}, \|\cdot\|_{L_2})} \|h\| \le C$, where $\mathcal{Z}$ is the support of ζ_i.

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

(iii) For some $C > 0$, $\max_{r \le K} \sup_\zeta |\phi_r(\zeta)| < C$. In addition, there is $\eta > 0$ such that $\psi_{\min}^{-1}(S_\beta) < \eta$ and $\psi_{\min}^{-1}(S_F) < \eta$ with probability converging to 1.

(iv) $(NT)^{-1} \sum_{i,t} h_t^2(\zeta_i) = O_P(1)$.

(v) There are constants $\delta, g \ge 0$ such that $\psi_1(Q)/\psi_K(Q) = O_P(K^\delta)$, $\min_{1 \le r \le K-1} \psi_r(Q) - \psi_{r+1}(Q) \ge cK^{-g}$ for some constant $c > 0$.

First, we present some assumptions for the sieve representation. Assumption 3.1(ii) is well satisfied with a large ν if the functions $\{h_t(\cdot)\}$ are sufficiently smooth. For example, consider h_t belonging to a Hölder class: for some $a, b, C > 0$, $\{h : \|D^b h(x_1) - D^b h(x_2)\| \le C\|x_1 - x_2\|^a\}$. In addition, suppose that we take a usual basis like polynomials, trigonometric polynomials, and B-splines. Then, $\max_{i,t} |M_{it}^R| \le CK^{-\nu}$, and $\nu = 2(a+b)/\dim(\zeta_i)$. So, Assumption 3.1(ii) is satisfied with very large ν if $\{h_t(\cdot)\}$ are smooth. In addition, the first part of Assumption 3.1(iii) can be satisfied if the basis is a bounded basis like trigonometric basis or ζ_i has a compact support. Assumption 3.1(iv) and (v) are not restrictive, and have been verified by Chernozhukov et al. (2023).

Assumption 3.2 (*DGP for ε_{it} and ω_{it}*). (i) Conditioning on $\mathcal{M}$, ε_{it} is zero-mean, sub-gaussian random variable such that $\mathbb{E}[\varepsilon_{it}|\mathcal{M}] = 0$, $\mathbb{E}[\varepsilon_{it}^2|\mathcal{M}] = \sigma_{it}^2 \le \sigma^2$, $\mathbb{E}[\exp(s\varepsilon_{it})|\mathcal{M}] \le \exp(Cs^2\sigma^2)$, $\forall s \in \mathbb{R}$ for some constant $C > 0$. We assume that σ^2 is bounded above and σ_{it}^2 are bounded away from zero. In addition, ε_{it} is indepedent across i and t.

(ii) Ω is independent of $\mathcal{E}$. Conditioning on $\mathcal{M}$, ω_{it} is independent across t. In addition, $\mathbb{E}[\omega_{it}|\mathcal{M}] = \mathbb{E}[\omega_{it}] = p_i$ where $0 < p_{\min} \le p_i \le p_{\max} \le 1$.

(iii) Let a_t be the column of either $\Omega - \Pi \mathbf{1}_N \mathbf{1}_T'$ or $\Omega \circ \mathcal{E}$. Then, $\{a_t\}_{t \le T}$ are independent sub-gaussian random vector with $\mathbb{E}[a_t] = 0$; more specifically, there is $C > 0$ such that

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

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

While the serial independence of the missing data indicators ω_{it} is assumed, we allow they are cross-sectional dependence among i. In doing so, we assume a cluster structure in $\{1, \dots, N\}$, i.e., there is a family of nonempty disjoint clusters, $C_1, \dots, C_\rho$ such that $\cup_{g=1}^{\rho} C_g = \{1, \dots, N\}$. So we divide units $\{1, \dots, N\}$ into ρ disjoint clusters. In addition, denote the size of the largest cluster by ϑ. That is, $\vartheta = \max_g |C_g|_o$. We highlight that ϑ is allowed to increase as N and T increase.

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

Due to the cluster structure in Assumption 3.3(i), we can construct a "leave-cluster-out" estimator $\breve{\beta}^{\{-i\}}$ which is independent of the sample of unit i. Similarly to the idea of (2.3) and (2.4), we can rule out the samples of the cluster that includes unit i. The difference from (2.4) is that we identify all the units which are in the same cluster as unit i and replace their rows of the goodness

of fit part by their conditional expectations.[2] Together with the leave-one-out estimator $\breve{\beta}^{(-t)}$, the leave-cluster-out estimator $\breve{\beta}^{\{-i\}}$ plays a pivotal role in showing the solution of (2.2) is close to that of (2.3).

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

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

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

Here the group of interest as $\mathcal{G} = \mathcal{I} \times \mathcal{T}$ where $\mathcal{I} \subseteq \{1, \ldots, N\}$ and $\mathcal{T} \subseteq \{1, \ldots, T\}$. We impose the following assumption on the rates of parameters. Define a sequence ψ_{NT} as $\psi_{NT} \asymp \sqrt{K^{-(2\delta+1)} \sum_{i=1}^{N} \sum_{t=1}^{T} h_t^2(\zeta_i)}$. It is a lower bound of $\psi_{\min}(\beta F')$ and works as the parameter for signal. Recall that K denotes the sieve dimension.

Assumption 3.4 (*Parameter Size and Signal-to-noise Ratio*). Let $\gamma = \frac{p_{\max}}{p_{\min}}$ and $\tilde{\vartheta} = \max\{\vartheta, \log N + \log T\}$. Then, we have

(i) $\min\{|\mathcal{I}|_o^{\frac{1}{2}}, |\mathcal{T}|_o^{\frac{1}{2}}\} \, \tilde{\vartheta} \eta^3 \gamma^4 K^{(4+2g+\frac{13}{2}\delta)} \max\{\sqrt{N \log N}, \sqrt{T \log T}\} = o(p_{\min}^{\frac{3}{2}} \min\{N, T\}),$

  $\min\{|\mathcal{I}|_o^{\frac{1}{2}}, |\mathcal{T}|_o^{\frac{1}{2}}\} \eta^{\frac{1}{2}} \gamma^3 K^{(1+g+\frac{7}{2}\delta)} \max\{N^{\frac{3}{2}}, T^{\frac{3}{2}}\} = o(p_{\min}^{\frac{3}{2}} \psi_{NT}^2),$

(ii) $\min\{|\mathcal{I}|_o^{\frac{1}{2}}, |\mathcal{T}|_o^{\frac{1}{2}}\} \eta^{\frac{3}{2}} \gamma^2 \max\{\sqrt{N}, \sqrt{T}\} = o(p_{\min}^{\frac{1}{2}} K^{(\nu-2\delta-\frac{3}{2})}),$

  $\min\{|\mathcal{I}|_o^{\frac{1}{2}}, |\mathcal{T}|_o^{\frac{1}{2}}\} \eta^{\frac{1}{2}} \gamma^{\frac{3}{2}} \max\{\sqrt{N}, \sqrt{T}\} \sqrt{NT} = o(\psi_{NT} p_{\min}^{\frac{1}{2}} K^{(\nu-\delta-\frac{1}{2})}).$

Assumption 3.4(ii) is used to bound the sieve approximation error. For this condition to be satisfied, the smoothness of $\{h_t(\cdot)\}_{t \leq T}$ is crucial. If $\{h_t(\cdot)\}_{t \leq T}$ are smooth enough, $\nu = 2(a+b)/\dim(\zeta_i)$ can be arbitrarily large. Hence, Assumption 3.4(ii) can be easily satisfied with a slowly increasing K as long as $\{h_t(\cdot)\}_{t \leq T}$ is smooth.

Assumption 3.4(i) is the conditions about sample complexity and signal-to-noise ratio. As long as K, η, γ are bounded or increase sufficiently slowly, it would be satisfied. Note that, in the cases like the cross-sectional average of a certain period t or the time average of a certain unit i, $\min\{|\mathcal{I}|_o^{\frac{1}{2}}, |\mathcal{T}|_o^{\frac{1}{2}}\} = 1$. In many interesting cases, $\min\{|\mathcal{I}|_o^{\frac{1}{2}}, |\mathcal{T}|_o^{\frac{1}{2}}\}$ is usually not that large. However, due to Assumption 3.4(i), we cannot derive the inferential theory in the case where both $|\mathcal{I}|_o$ and $|\mathcal{T}|_o$ are large like $|\mathcal{I}|_o = N$ and $|\mathcal{T}|_o = T$. In this case, the asymptotically normal distribution part cannot dominate other residual parts, since the convergence rate of the asymptotically normal distribution part is roughly $\frac{1}{\sqrt{N|\mathcal{T}|_o}} + \frac{1}{\sqrt{T|\mathcal{I}|_o}}$, while that of the residual term is similar to or greater than $\frac{1}{\sqrt{NT}}$ regardless of the group size. For inference, at least one part of the asymptotically normal term should dominate other residual terms. On the other hand, in terms of the convergence rate, the large sizes of $|\mathcal{I}|_o$ and $|\mathcal{T}|_o$ are beneficial, as noted in Appendix B. In addition, for comparison with the conditions of other low-rank literature, it would be helpful to refer to Assumption C.2 in Appendix C where we consider the general low-rank model.

Under the above assumptions, Theorem C.1 shows that the estimator for the group average of M_{it} has the asymptotic normality:

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

where the asymptotic variance $\mathcal{V}_{\mathcal{G}}$ is given in the statement of Theorem C.1, and needs to be estimated. In this result, $\mathcal{G}$ can consist of either multiple columns with multiple rows or solely a certain (i,t), implying that we can conduct inference for one specific element of the matrix. To make the estimation of $\mathcal{V}_{\mathcal{G}}$ feasible, we consider the case of $\mathbb{E}[\varepsilon_{it}^2 | \mathcal{M}] = \sigma_i^2$. Let U_i' is the ith row of the left singular vector of $\beta F'$ and V_t' is the tth row of the right singular vector of $\beta F'$. The following theorem gives the feasible asymptotic normality.

Theorem 3.1 (*Feasible CLT*). Suppose Assumptions 3.1–3.4 hold. In addition, suppose that $\left\| \frac{\sqrt{N}}{|\mathcal{I}|_o} \sum_{i \in \mathcal{I}} U_{M^*,i} \right\| \geq c$ and $\left\| \frac{\sqrt{T}}{|\mathcal{T}|_o} \sum_{t \in \mathcal{T}} V_{M^*,t} \right\| \geq c$ for some constant $c > 0$. Then 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

$$\widehat{\mathcal{V}}_{\mathcal{G}} = \frac{1}{|\mathcal{T}|_o^2} \sum_{t \in \mathcal{T}} \widehat{\overline{\beta}}_{\mathcal{I}}' \left(\sum_{j=1}^{N} \omega_{jt} \widehat{\beta}_j \widehat{\beta}_j' \right)^{-1} \left(\sum_{j=1}^{N} \omega_{jt} \widehat{\sigma}_j^2 \widehat{\beta}_j \widehat{\beta}_j' \right) \left(\sum_{j=1}^{N} \omega_{jt} \widehat{\beta}_j \widehat{\beta}_j' \right)^{-1} \widehat{\overline{\beta}}_{\mathcal{I}}$$

$$+ \frac{1}{|\mathcal{I}|_o^2} \sum_{i \in \mathcal{I}} \widehat{\sigma}_i^2 \widehat{\overline{F}}_{\mathcal{T}}' \left(\sum_{s=1}^{T} \omega_{is} \widehat{F}_s \widehat{F}_s' \right)^{-1} \widehat{\overline{F}}_{\mathcal{T}},$$

---

[2] For the formal definitions of the estimators, please refer to Appendix D and Remark 1.

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

3.2. Semiparametric efficiency

We now establish the semiparametric efficiency of our estimator, following a similar approach as in Jankova and Van De Geer (2018). In order to make calculation tractable, we suppose that $\omega_{it} \sim \text{Bernoulli}(p)$ and $\varepsilon_{it} \sim \mathcal{N}(0, \sigma^2)$ are independent across (i, t). We will focus on the case of block group, where both $|\mathcal{I}|_o$ and $|\mathcal{T}|_o$ are finite or growing slowly, satisfying $N|\mathcal{T}|_o \ll T^2|\mathcal{I}|_o^2$ and $T|\mathcal{I}|_o \ll N^2|\mathcal{T}|_o^2$. The other cases like cross-sectional and serial groups (e.g., $|\mathcal{I}|_o = N$ and $|\mathcal{T}|_o$ is finite or slowly growing, or vice versa) can also be attained, which are very similar to Theorem 4.2 in Chernozhukov et al. (2023). Hence, we omit them. The novelty of our efficiency theorem is that it is for estimating the general block group.

As specified in Theorem C.1, the asymptotic variance in this case is

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

$$= s_*^2(M, p, \sigma) + o(s_*^2(M, p, \sigma))$$

$$s_*^2(M, p, \sigma) := \frac{\sigma^2}{p} \frac{1}{|\mathcal{T}|_o} \bar{\beta}_{\mathcal{I}}' (\beta' \beta)^{-1} \bar{\beta}_{\mathcal{I}} + \frac{\sigma^2}{p} \frac{1}{|\mathcal{I}|_o} \bar{F}_{\mathcal{T}}' (F'F)^{-1} \bar{F}_{\mathcal{T}}.$$

The following theorem shows that $s_*^2(M, p, \sigma)$ is the asymptotic Cramér-Rao bound for asymptotically unbiased estimators.

Theorem 3.2. *Suppose* $\omega_{it} \sim \text{Bernoulli}(p)$ *and* $\varepsilon_{it} \sim \mathcal{N}(0, \sigma^2)$ *are independent across* (i, t). *Suppose also that* $N|\mathcal{T}|_o \ll T^2|\mathcal{I}|_o^2$ *and* $T|\mathcal{I}|_o \ll N^2|\mathcal{T}|_o^2$. *Define*

$$\mathcal{A} = \{(M, p, \sigma) : M = M^\star + M^R, M^\star = \beta F', \text{rank}(M^\star) \le K, \ and \ Assumptions \ 3.1\text{–}3.4 \ hold\}.$$

Let $U(Y, \Omega)$ *be an asymptotically unbiased estimator of* $|\mathcal{G}|^{-1} \sum_{(i,t) \in \mathcal{G}} M_{it}$ *in that*

$$\mathbb{E}_{M,p,\sigma} U(Y, \Omega) - |\mathcal{G}|^{-1} \sum_{(i,t) \in \mathcal{G}} M_{it} = o(s_*(M, p, \sigma))$$

where $\mathbb{E}_{M,p,\sigma}$ *denotes the expectation with respect to given* (M, p, σ). *Then for any sequence of* $(M, p, \sigma) \in \mathcal{A}$, *we have*

$$\liminf_{N,T \to \infty} \frac{\mathbb{E}_{M,p,\sigma} \left[U(Y, \Omega) - |\mathcal{G}|^{-1} \sum_{(i,t) \in \mathcal{G}} M_{it} \right]^2}{s_*^2(M, p, \sigma)} \ge 1,$$

with probability converging to 1.

4. Applications to heterogeneous treatment effect estimation

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

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

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

where ε_{it} is the noise, ζ_i is a vector of unit specific latent state variables. We regard $h_t^{(\iota)}(\cdot)$ as a deterministic function while ζ_i is a random vector. In the model, the treatment effect comes from the difference between the time-varying treatment function $h_t^{(1)}(\cdot)$ and the control function $h_t^{(0)}(\cdot)$. Let $\omega_{it}^{(\iota)} = 1\{y_{it}^{(\iota)} \text{ is observed}\}$. Then, $\omega_{it}^{(1)} = Y_{it}$ and $\omega_{it}^{(0)} = 1 - Y_{it}$ because we observe $y_{it}^{(1)}$ when there is a treatment on (i, t) and observe $y_{it}^{(0)}$ when there is no treatment on (i, t).

We suppose the following sieve 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 $\beta_i = [\phi_1(\zeta_i), \dots, \phi_K(\zeta_i)]'$ and $F_t^{(\iota)} = [\kappa_{t,1}^{(\iota)}, \dots, \kappa_{t,K}^{(\iota)}]'$, $h_t^{(\iota)}(\zeta_i)$ can be successfully represented as the approximate factor structure.

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

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

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

The notations are essentially the same as those in Section 2, and we just put the superscript (ι) to all notations to distinguish the pair of potential realizations.

Theorem 4.1 (*Feasible CLT*). *Suppose the assumptions of Theorem 3.1 hold for each $\iota \in \{0, 1\}$. With $\mathbb{E}[\varepsilon_{it}^2 | \mathcal{M}] = \sigma_i^2$, we have*

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

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

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

$$+ \frac{1}{|\mathcal{I}|_o^2} \sum_{i \in \mathcal{I}} \widehat{\sigma}_i^{(\iota)2} \widehat{\bar{F}}_{\mathcal{T}}^{(\iota)\prime} \left(\sum_{s=1}^T \omega_{is}^{(\iota)} \widehat{F}_s^{(\iota)} \widehat{F}_s^{(\iota)\prime} \right)^{-1} \widehat{\bar{F}}_{\mathcal{T}}^{(\iota)}.$$

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

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

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

There is a vast theoretical and empirical literature about the impact of the president on allocating the federal budget to the states (e.g., Cox and McCubbins (1986), Anderson and Tollison (1991), McCarty (2000), Larcinese et al. (2006) and 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. We contribute by showing 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. The data are obtained from websites of the U.S. Census Bureau, NASBO (National Association of State Budget Officers), and SSA (Social Security Administration).

Following Section 4, we set the treatment indicator as $Y_{it} = 1$ if the state i supported the president of year t in the presidential election, and $Y_{it} = 0$ otherwise. If the candidate whom the state i supported in the previous presidential election is the same as the president at year t, we consider it as "treated" and otherwise, we consider it as "untreated". While applying our inferential procedure, we adopt the assumption that the treatment (whether state i supported the resident in the election) is exogenously assigned, which is probably not practical, but we take our stand on this assumption in this study, and do not claim a causal interpretation of the treatment effect.

In addition, for the outcome variable y_{it}, we use the following ratio: $y_{it} = (\tilde{y}_{it} / \sum_i \tilde{y}_{it}) \times 100$ where $\tilde{y}_{it}$ is the per-capita federal grant in state i at year t. Note that the outcome variable, y_{it}, is a proportion so that $\sum_i y_{it} = 100$ for all t, which is to treat each period equally.

Our inferential theory allows novel approaches to study the following effects:

1. State Effects: the time average of the treatment effect of each state i, i.e., $T^{-1} \sum_{t=1}^T \Gamma_{it}$.
2. Region Effects: the time average of the treatment effect of each "Region", i.e.,

$$\frac{1}{|\text{Region}|_0} \sum_{i \in \text{Region}} \frac{1}{T} \sum_{t=1}^T \Gamma_{it}.$$
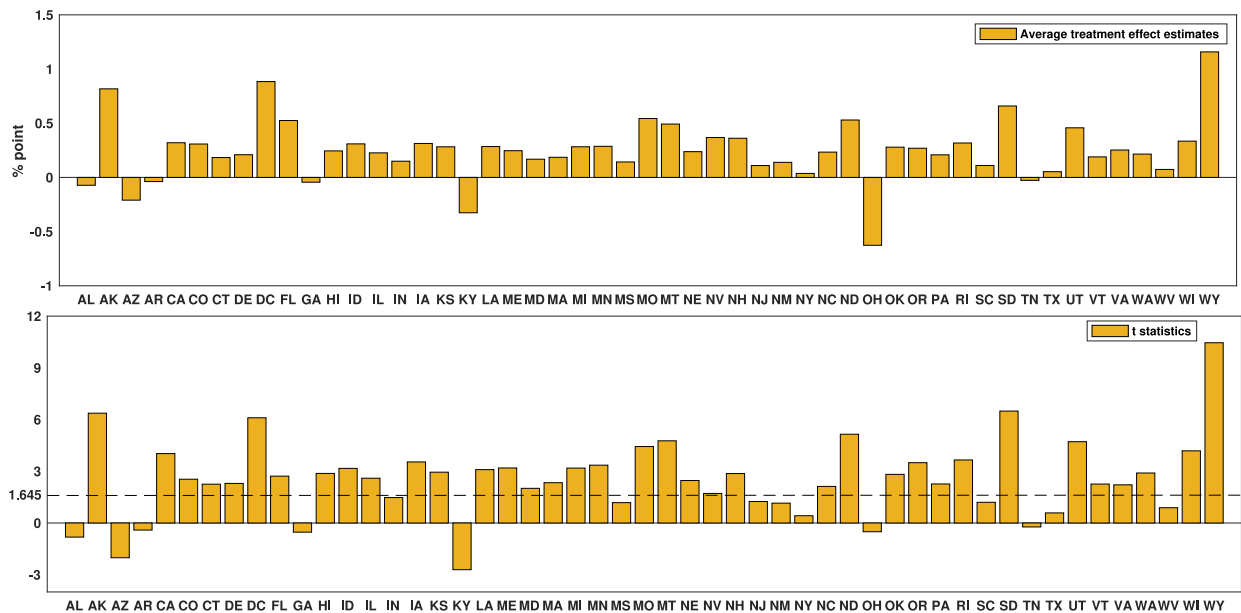
Fig. 1. State effects and corresponding t-statistics. NOTE: When we use the Benjamini and Hochberg (1995) procedure to control the size of the false discovery rate at 5%, the list of states with significant effects is unchanged.

Table 1
Number of swings of each state.

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

3. Loyal/Swing Effects: the time average of the treatment effect of "loyal" and "swing" states, e.g.,

$$\frac{1}{|\text{Loyal States}|_0} \sum_{i \in \text{Loyal States}} \frac{1}{T} \sum_{t=1}^{T} \Gamma_{it}. \quad \text{(see Table 1 for the definition of "Loyal States")}$$

4. President Effects: the average treatment effect of each president, i.e.,

$$\frac{1}{|\mathcal{T}|_0} \sum_{t \in \mathcal{T}} \frac{1}{N} \sum_{i=1}^{N} \Gamma_{it}. \quad (\mathcal{T} \text{ denotes the period of a given President in Office})$$

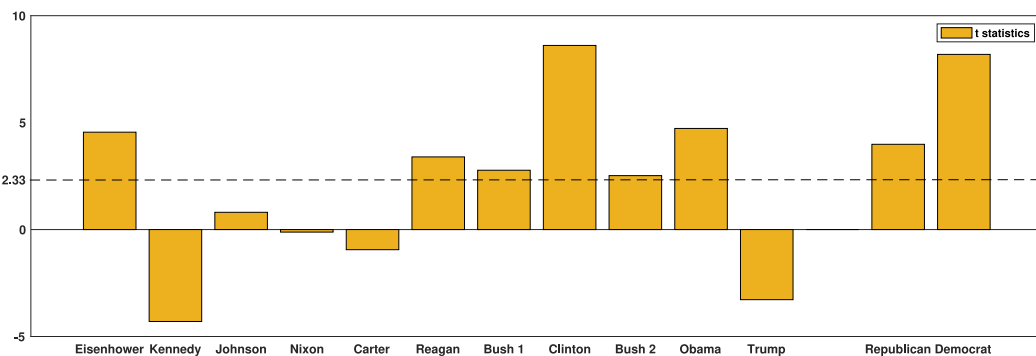5. Party Effects: the average treatment effect of each Party, i.e.,

$$\frac{1}{|S|_0} \sum_{t \in S} \frac{1}{N} \sum_{i=1}^{N} \Gamma_{it}. \quad (S \text{ denotes the period of a given Party to which the President belonged})$$

First, Fig. 1 presents the State Effects and the corresponding t-statistics. The results suggest significantly positive treatment effects in most states. To investigate the reason of differences, we categorize states according to the number of times a state swung the party it supports in the presidential elections as in Table 1. Together with Fig. 1, it shows that most states with large t-statistics are in "Loyal states" while the other states are generally in "Swing state" or "Weak swing state". It suggests that the treatment effect is closely related to the loyalty of states to parties.

In addition, the results for the Region Effects in Fig. 2 show that, at the 1% significant level, New England, Mid Atlantic, Plains, Rocky Mountain, and Far West have the positive treatment effects while Great Lakes, South East, and South West do not. Note that Many states in Great Lakes, South East, and South West are in "Swing states" or "Weak swing states". As we can see in Fig. 2, "Swing states" do not have statistically significant positive treatment effects while "Loyal states" have significant positive treatment effects. This result is in line with the empirical study of Larcinese et al. (2006) finding that states with loyal supports tend to receive more funds, while swing states are not rewarded. In addition, it is aligned with the assertion of Cox and McCubbins (1986) that the targeting of loyal voters can be seen as a safer investment as compared to aiming for swing voters and risk-averse political actors may allocate more funds to loyal states.
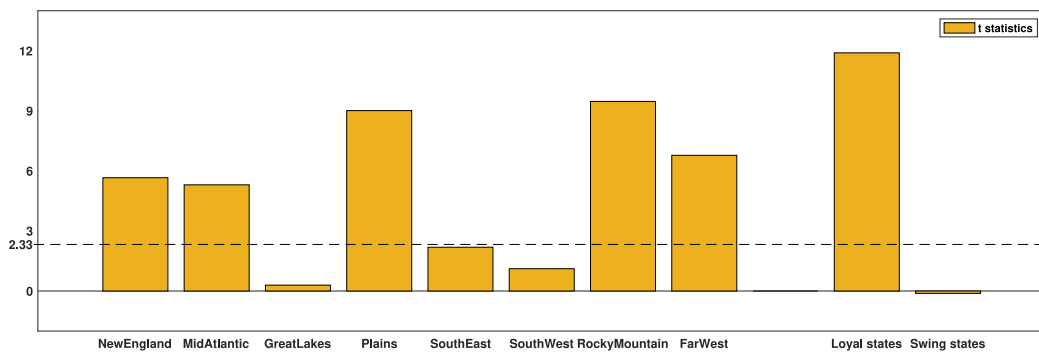
Fig. 2. Test statistics for the Region Effects and the Loyal/Swing Effects. NOTE: "New England" includes CT, ME, MA, NH, RI,VT, "Mid Atlantic" includes DE, D.C., MD, NJ, NY, PA, "Great Lakes" includes IL, IN, MI, OH, WI, "Plains" includes IA, KS, MN, MO, NE, ND, SD, "South East" includes AL, AR, FL, 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.

Fig. 3. Test statistics for the President Effects and the Party Effects.
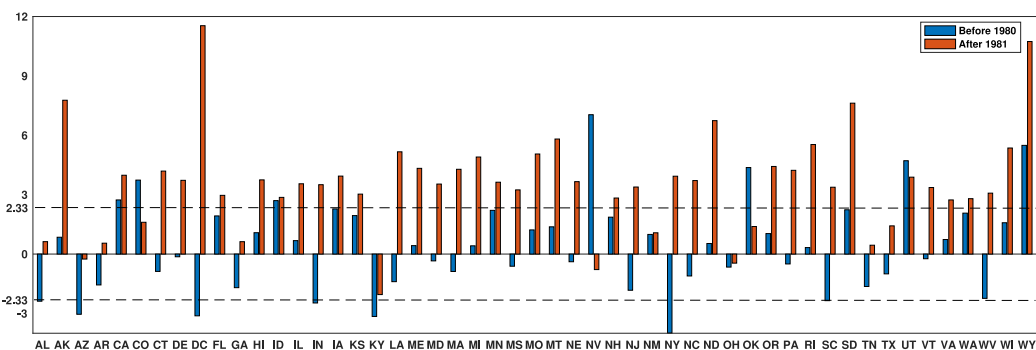
Fig. 4. Test statistics for the average treatment effect before 1980 and after 1981.

Fig. 3 shows the President Effects and the Party Effects. Despite some exceptions, there are no statistically significant positive treatment effects before Carter, while there are significant positive treatment effects after Reagan. Fig. 4 shows that before 1980, there is no significant positive treatment effect in most states, while there are significant positive treatment effects in most states after 1981. Hence, there is a substantial difference between 'before 1980' and 'after 1981' and the tendency that incumbent presidents reward states that showed their support in the presidential elections became significant after Reagan, that is, after the 1980s. It suggests that after the 1980s, the presidents show more influence on the allocation of federal funds to reward their supporters. 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.

Table 2
$\|\widehat{M} - M\|_F / \sqrt{NT}$.

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

NOTE: "Sine" and "Poly" refer to the functions $h_t^{\text{sine}}(\zeta)$ and $h_t^{\text{poly}}(\zeta)$, respectively.

Finally, when testing for the treatment effects of multiple states, the tests may subject to the issue of multiple testing problems, with undesirable false discovery rates (FDR). We also address this issue by adopting the procedure of Benjamini and Hochberg (1995) to control the FDR at 5%. We find that the list of states with significant treatment effects is unchanged.

6. Simulation study

This section provides the finite sample performances of the estimators. We first study the performances of the estimators of M_{it} and $|\mathcal{G}|_o^{-1} \sum_{(i,t) \in \mathcal{G}} M_{it}$, and then study performances of the average treatment effect estimators. To save space, some results are relegated to Appendix.

First of all, in order to check the estimation quality of our estimator, we compare the Frobenius norms of the estimation errors for several existing estimators of M. Our two-step least squares is labeled as "TLS". We also consider the debiased nuclear norm penalized estimators from Xia and Yuan (2021), "(Hetero) XY", and Chen et al. (2019), "(Hetero) CFMY". "(Hetero)" represents that they are modified to allow the heterogeneous observation probabilities. The comparison also includes the inverse probability weight based estimator, "IPW", from Xiong and Pelger (2023), and the EM algorithm based estimator, "EM", from Jin et al. (2021). The plain nuclear norm penalized estimator, "Plain Nuclear", and the TLS estimator using sample splitting, "TLS with SS", are also considered. For the data-generating designs, we consider the following three models:

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

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

- Nonparametric model 2: $y_{it} = h_t(\zeta_i) + \varepsilon_{it}$, where $h_t(\zeta) = h_t^{\text{sine}}(\zeta) := \sum_{r=1}^{\infty} \frac{|U_{t,r}|}{r^3} \sin(r\zeta)$. (6.1)

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

Table 2 reports $\|\widehat{M} - M\|_F / \sqrt{NT}$ averaged over 100 replications. We highlight that the TLS shows the best performance in almost all scenarios. Only the EM is comparable to ours, but it computes much slower since it requires multi-step iterations. In contrast, our proposed method does not iterate. Also, our method always outperforms the TLS with SS. The (Hetero) XY and (Hetero) CFMY are slightly worse than ours in this experiment. Lastly, both the IPW and the Plain Nuclear show the worst performances uniformly. The IPW, being non-statistically efficient, is only slightly better than the Plain Nuclear.

Additionally, to show the relative advantage of TLS over TLS with sample splitting, Table 3 reports $(\widehat{M}_{it} - M_{it})^2$ in the case where T is small. Here, we choose (i, t) randomly and fix it during replications. As we can check in the table, when T is relatively small, the performance of TLS with sample splitting is much worse than that of TLS without sample splitting. Especially, in the factor model, the difference in performance is quite large.

Second, we study the finite sample distributions for standardized estimates defined as $(\widehat{M}_{it} - M_{it})/se(\widehat{M}_{it})$. For comparison, we report the results of the Plain Nuclear and the TLS with SS, in addition to the TLS. For the Plain Nuclear, we use the sample standard deviation obtained from the simulations for $se(\widetilde{M}_{it})$ because the theoretical variance of it is unknown. For the TLS with SS, we construct the standard error following Chernozhukov et al. (2019). Here, we consider the nonparametric models in (6.1). Hereinafter, the number of replications is 1000, and the sample size is $N = T = 200$.

Fig. 5 plots the scaled histograms of the standardized estimates with the standard normal density. As we expected in theory, it shows that the standardized TLS and the standardized TLS with SS fit the standard normal distribution well, while the standardized Plain Nuclear is biased. Without sample splitting, the TLS itself provides a good approximation to the standard normal distribution so that it can be used for the inference successfully. The coverage probabilities of confidence interval in Appendix also show similar results.

Table 3
$(\widehat{M}_{it} - M_{it})^2$ comparison between TLS and TLS with SS.

Model	Factor			Sine			Poly		
Sample size	TLS	TLS w/ SS	Ratio	TLS	TLS w/ SS	Ratio	TLS	TLS w/ SS	Ratio
N = 100, T = 20	0.4665	2.8951	16.1%	0.1401	0.1702	82.3%	0.1272	0.1894	67.2%
N = 100, T = 40	0.2162	0.2685	80.5%	0.0736	0.0819	89.9%	0.0807	0.0865	93.3%
N = 100, T = 60	0.1111	0.1300	85.5%	0.0603	0.0637	94.7%	0.0538	0.0567	94.9%

NOTE: The values are the averaged $(\widehat{M}_{it} - M_{it})^2$ over 1000 replications. "Ratio" denotes the ratio between performances of TLS and TLS with SS. Here, we assume $\omega_{it} \sim \text{Bernoulli}(0.5)$. When $T = 20$, the working sample size for the sample splitting is only 10, which leads to singularity issues in the inverse covariance matrix estimation. As a result, the estimator performs badly in this case.
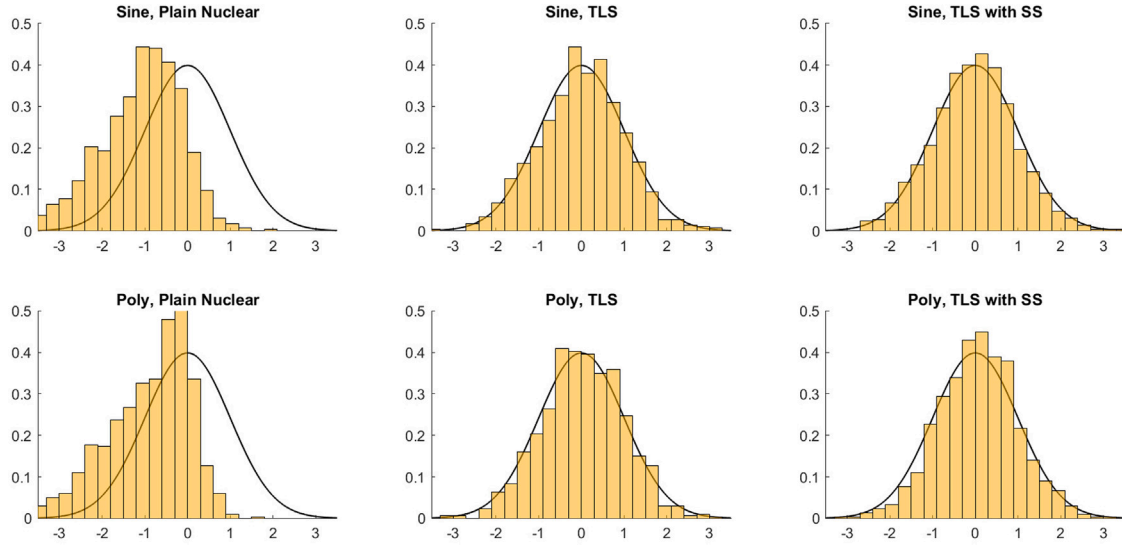
Fig. 5. Histograms of standardized estimates, $(\widehat{M}_{it} - M_{it})/se(\widehat{M}_{it})$.

Next, we study the finite sample performance of the average treatment effect estimator. Following Section 4, for each $\iota \in \{0, 1\}$, we generate the data from $y_{it}^{(\iota)} = h_t^{(\iota)}(\zeta_i) + \varepsilon_{it}$, where $h_t^{(0)}(\zeta) = \sum_{r=1}^{\infty} |U_{t,r}| r^{-a} \sin(r\zeta)$, $h_t^{(1)}(\zeta) = \sum_{r=1}^{\infty} (|U_{t,r}| + 2) r^{-a} \sin(r\zeta)$. 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 Fig. 6.

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

Fig. 7 presents the scaled histograms of the standardized estimates of the average treatment effect estimators for the groups $\mathcal{G}_1 = \{(i, t)\}$, $\mathcal{G}_2 = \{(j, t) : 1 \leq j \leq N\}$, and $\mathcal{G}_3 = \{(i, s) : 1 \leq s \leq T\}$. Here, the standard estimates are given as

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

As expected in theory, the standardized estimates of the average treatment effect estimators of all groups approximately show the standard normal distribution. In addition, the coverage probabilities of the confidence interval in Appendix also show similar results.

7. Conclusion

This paper studies the inferential theory for low-rank matrices and provides an inference method for the average treatment effect as an application. Without the aid of sample splitting, our estimation procedure successfully resolves the problem of the shrinkage bias, and the resulting estimator attains the asymptotic normality. Unlike Chernozhukov et al. (2019, 2023) which exploit sample splitting, our estimation step is simple, and we can avoid some undesirable properties of sample splitting. In addition, this paper allows the heterogeneous observation probability and uses inverse probability weighting to control the effect of the heterogeneous observation probability.
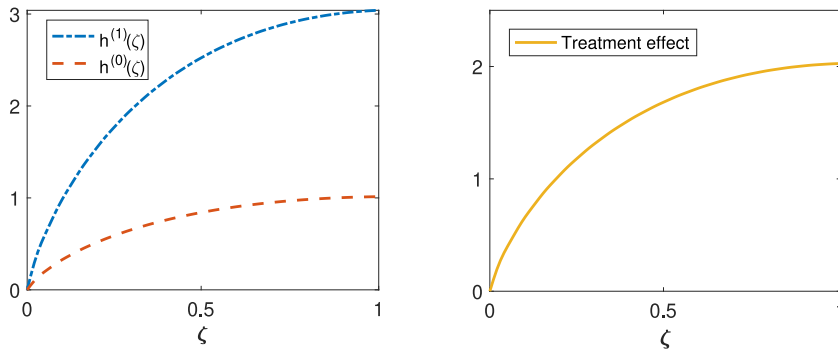
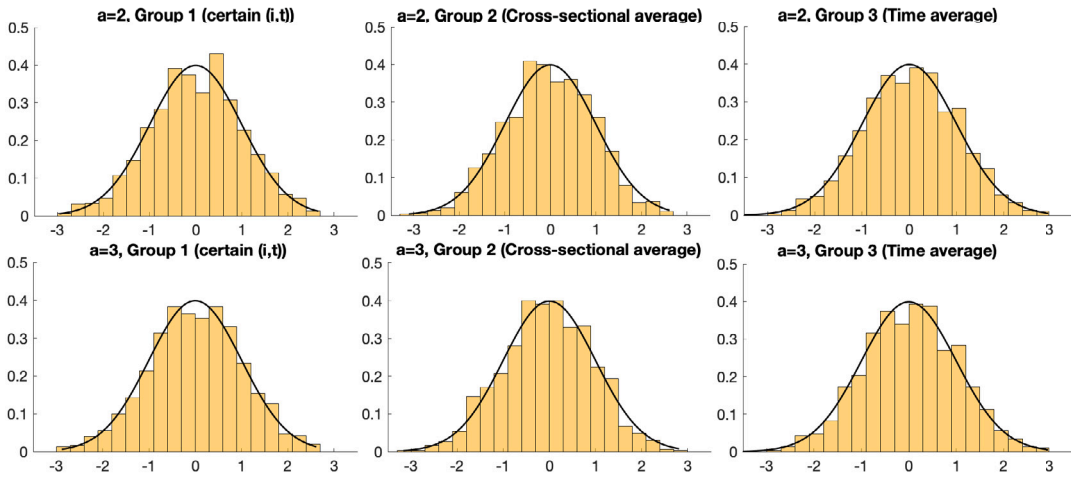**Fig. 6.** Shape of function $h_t^{(t)}(\zeta)$ and treatment effect function ($U_{t,r} = 1$, $a = 2$).

Fig. 7. Histograms of standardized estimates, $\frac{\frac{1}{|\mathcal{G}|_o}\sum_{(i,t)\in\mathcal{G}}\widehat{\Gamma}_{it} - \frac{1}{|\mathcal{G}|_o}\sum_{(i,t)\in\mathcal{G}}\Gamma_{it}}{se\left(\frac{1}{|\mathcal{G}|_o}\sum_{(i,t)\in\mathcal{G}}\widehat{\Gamma}_{it}\right)}$. NOTE: Here, the sample size is $N = T = 300$. "Group 1" refers to $\mathcal{G}_1$, "Group 2" denotes $\mathcal{G}_2$ and "Group 3" refers to $\mathcal{G}_3$.

Appendix A. Data-driven ways of choosing K

Using a consistent estimator of K

To choose the sieve dimension K, we can use the following rank estimator of $M^\star$ in the general approximate factor model $\widehat{K} = \sum_r 1\{\psi_r(\widetilde{M}) \geq ((N + T)/2)^{\frac{11}{20}}\|\widetilde{M}\|^{\frac{1}{4}}\}$ where $\psi_r(\widetilde{M})$ denotes the rth largest singular value of $\widetilde{M}$. As noted in Claim G.1(iii), it works as a consistent rank estimator for $M^\star$ in the general approximate factor model. By the same token in Footnote 5 of Bai (2003), our inferential theory for the general approximate factor model is not affected even if the rank K is unknown and estimated using this estimator since $P(\widehat{K} = K) \to 1$.

Cross-validation method

When the matrix of interest M is approximated by a low-rank structure via a sieve representation like our main model, we can treat the sieve dimension K as a tuning parameter. Hence, we introduce one data-driven way of selecting K which exploits the cross-validation which is similar to the idea in Athey et al. (2021). From the observed sample $\{(i,t) : \omega_{it} = 1\}$, we randomly create a subsample by using a Bernoulli process, namely the subsample is $\{(i,t) : \omega_{it}X_{it} = 1\}$ where $\{X_{it}\}_{i\leq N, t\leq T}$ are independent Bernoulli random variables of probability $\sum_{i,t}\omega_{it}/NT$, which is independent of $\{\omega_{it}\}_{i\leq N, t\leq T}$. This guarantees that we have $\sum_{i,t}\omega_{it}/NT \approx \sum_{i,t}\omega_{it}X_{it}/\sum_{i,t}\omega_{it}$. We then pre-specify the set of candidates of K as $\{K_1, K_2, \ldots\}$ and compute the estimates $\widehat{M}_{K_1}, \widehat{M}_{K_2}, \ldots$, respectively, using only the subsample. To compare their out-of-sample performance, we measure the mean squared error of them on $\{(i,t) : \omega_{it}(1 - X_{it}) = 1\}$. For robustness, we repeat this process five times, creating different independent subsamples each time, to obtain five mean squared errors for each $K \in \{K_1, K_2, \ldots\}$. The sieve dimension which minimizes the sum of five mean squared errors is chosen. In our simulation study, we use this method with $\{2, 4, 6, 8, 10\}$ as the set of candidates of K.

Appendix B. Finite sample convergence rate

For completeness, this section studies the finite sample convergence rate of our estimator. First, we provide several conditions. Here, $a \lesssim b$ means $|a|/|b| \leq C$ for some constant $C > 0$. $a \ll b$ indicates $|a| \leq c|b|$ for some sufficiently small constant $c > 0$.

Assumption B.1 (*Sieve Representation*). (i) $\{h_t(\cdot)\}_{t \leq T}$ belong to ball $\mathcal{H}\left(\mathcal{Z}, \|\cdot\|_{L_2}, C\right)$ inside a Hilbert space spanned by the basis $\{\phi_r\}_{r \geq 1}$, with a uniform L_2-bound C: $\sup_{h \in \mathcal{H}(\mathcal{Z}, \|\cdot\|_{L_2})} \|h\| \leq C$, where $\mathcal{Z}$ is the support of ζ_i.

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

(iii) For some $C > 0$, $\max_{r \leq K} \sup_\zeta |\phi_r(\zeta)| < C$. In addition, there is $\eta > 0$ such that $\psi_{\min}^{-1}\left(S_\beta\right) < \eta$ and $\psi_{\min}^{-1}\left(S_F\right) < \eta$.

(iv) $\sum_{i,t} h_t^2(\zeta_i) \lesssim NT$.

(v) There are constants $\delta, g \geq 0$ such that $\psi_1(Q)/\psi_K(Q) \lesssim K^\delta$, $\min_{1 \leq r \leq K-1} \psi_r(Q) - \psi_{r+1}(Q) \geq cK^{-g}$ for some constant $c > 0$.

This condition is basically the same as Assumption 3.1, and we modify some notation to be suitable for finite sample analysis.

Assumption B.2 (*Parameter Size and Signal-to-noise Ratio*). Let $\gamma = \frac{p_{\max}}{p_{\min}}$ and $\tilde{\vartheta} = \max\{\vartheta, \log N + \log T\}$. Then, we have

$$(i) \quad \tilde{\vartheta}\eta^{\frac{3}{2}}\gamma^{\frac{5}{2}}K^{(2+2g+\frac{9}{2}\delta)}\max\{\sqrt{N\log N}, \sqrt{T\log T}\} \ll p_{\min}^{\frac{1}{2}}\min\{N, T\},$$

$$\gamma^{\frac{3}{2}}K^{(g+\frac{3}{2}\delta)}\max\{N, T\} \ll p_{\min}^{\frac{1}{2}}\min\{\sqrt{N\log N}, \sqrt{T\log T}\}\psi_{NT},$$

$$(ii) \quad \min\{|\mathcal{I}|_o^{\frac{1}{2}}, |\mathcal{T}|_o^{\frac{1}{2}}\}\max\{\sqrt{N}, \sqrt{T}\} \ll p_{\min}^{\frac{1}{2}}K^{(\nu-\frac{1}{2}-2\delta)},$$

$$\min\{|\mathcal{I}|_o^{\frac{1}{2}}, |\mathcal{T}|_o^{\frac{1}{2}}\}\max\{\sqrt{N}, \sqrt{T}\}\sqrt{NT} \ll \gamma^{\frac{1}{2}}\psi_{NT}K^\nu.$$

The above condition is weaker than the condition for the asymptotic normality (Assumption 3.4). For example, Assumption B.2(i) does not restrict the size of the interesting group, $\min\{|\mathcal{I}|_o, |\mathcal{T}|_o\}$, unlike Assumption 3.4(i). Hence, we can deal with the case where $|\mathcal{I}|_o = N$ and $|\mathcal{T}|_o = T$. In addition, it allows for a weaker signal-to-noise ratio than that of Assumption 3.4.

Proposition B.1. *Suppose Assumptions 3.2, 3.3, B.1, and B.2. Then, with probability at least $1 - O(\min\{N^{-3}, T^{-3}\})$, we have*

$$\left\| \frac{1}{|\mathcal{G}|_o} \sum_{(i,t) \in \mathcal{G}} \widehat{M}_{it} - \frac{1}{|\mathcal{G}|_o} \sum_{(i,t) \in \mathcal{G}} M_{it} \right\| \leq C \left(\frac{\sigma\eta^{\frac{1}{2}}K^{\frac{1}{2}}\max\{\sqrt{\log N}, \sqrt{\log T}\}}{p_{\min}^{\frac{1}{2}}\sqrt{N|\mathcal{T}|_o}} + \frac{\sigma\eta^{\frac{1}{2}}K^{\frac{1}{2}}\max\{\sqrt{\log N}, \sqrt{\log T}\}}{p_{\min}^{\frac{1}{2}}\sqrt{T|\mathcal{I}|_o}} \right.$$

$$\left. + \frac{\sigma\tilde{\vartheta}\gamma^{\frac{7}{2}}K^{(4+2g+\frac{13}{2}\delta)}\eta^3\max\{\log N, \log T\}}{p_{\min}^{\frac{3}{2}}\min\{N, T\}} + \frac{\sigma^3\gamma^2 K^{(\frac{7}{2}\delta+g+1)}\eta^{\frac{1}{2}}\max\{N, T\}}{p_{\min}^2\psi_{NT}^2} \right)$$

for some constant $C > 0$.

The first two terms represent the asymptotically normal distribution parts, while the last two terms are the residual parts related to the estimation errors of β_i and f_t. If we ignore some small parameters and logarithmic terms, the convergence rate of the first two terms is reduced to

$$\frac{1}{\sqrt{N|\mathcal{T}|_o}} + \frac{1}{\sqrt{T|\mathcal{I}|_o}}.$$

However, if both $|\mathcal{I}|_o$ and $|\mathcal{T}|_o$ are large, as in the case where $|\mathcal{I}|_o = N$ and $|\mathcal{T}|_o = T$, the asymptotically normal parts cannot dominate the residual parts. Thus, we are unable to derive the inferential theory in this case. For inference, at least one part of the asymptotically normal terms should dominate other residual terms. On the other hand, in terms of the convergence rate, the large sizes of $|\mathcal{I}|_o$ and $|\mathcal{T}|_o$ are beneficial.

Appendix C. Inferential theory for the general approximated factor model

This section provides assumptions for the asymptotic normality of the estimator of the group average of M_{it} for the general approximated factor model having the form $Y = M + \mathcal{E}$ where $M = M^\star + M^R$, $rank(M^\star) = K$. For this, we define some additional notations. The condition number of $M^\star$ is defined as $q := \psi_{\max}(M^\star)/\psi_{\min}(M^\star)$. Define $\bar{c} = \min_{1 \leq r \leq K+1}\left|c_{r-1}^2 - c_r^2\right|$, where $c_r := \psi_r(M^\star)/\psi_{\min}(M^\star)$, and $c_{inv} := 1/\bar{c}$.[3]

---

[3] We set $c_0 := \infty$. Note that $\psi_r = 0$ for $r > K$, and that $c_1^2 = q^2 \geq c_2^2 \geq c_K^2 = 1$ for all $1 \leq r \leq K$. $\bar{c}$ is always smaller than 1 since $c_K^2 - c_{K+1}^2 = 1$. Hence, $c_{inv} \geq 1$. We allow c_{inv} to increase slowly as N and T increase.

Assumption C.1 (*Incoherence*). The matrix $M^\star$ satisfies μ-incoherence condition. That is, $\|U_{M^\star}\|_{2,\infty} \le \sqrt{\frac{\mu}{N}} \|U_{M^\star}\|_F = \sqrt{\frac{\mu K}{N}}$ and $\|V_{M^\star}\|_{2,\infty} \le \sqrt{\frac{\mu}{T}} \|V_{M^\star}\|_F = \sqrt{\frac{\mu K}{T}}$ with probability converging to 1. Here, μ is allowed to increase as N, T increase.

Assumption C.2 (*Parameters Size*). Let $\gamma = \frac{p_{\max}}{p_{\min}}$ and $\tilde{\vartheta} = \max\{\vartheta, \log N + \log T\}$. Then, we have

(i) $\min\{|\mathcal{I}|_o^{1/2}, |\mathcal{T}|_o^{1/2}\} \tilde{\vartheta} c_{\text{inv}}^{\frac{15}{2}} \mu^3 K^4 \gamma^{\frac{7}{2}} \max\{\sqrt{N \log N}, \sqrt{T \log T}\} = o_P(p_{\min} \min\{N, T\})$,

(ii) $\min\{|\mathcal{I}|_o^{1/2}, |\mathcal{T}|_o^{1/2}\} \tilde{\vartheta} c_{\text{inv}}^2 q^7 \mu^{\frac{5}{2}} K^{\frac{7}{2}} \gamma^4 \max\{N \sqrt{\log N}, T \sqrt{\log T}\} = o_P(\psi_{\min} p_{\min}^{\frac{3}{2}} \min\{\sqrt{N}, \sqrt{T}\})$,

(iii) $\min\{|\mathcal{I}|_o^{1/2}, |\mathcal{T}|_o^{1/2}\} \vartheta c_{\text{inv}}^2 q^6 \mu^2 K^{\frac{7}{2}} \gamma^{\frac{7}{2}} \max\{N^{\frac{3}{2}} \sqrt{\log N}, T^{\frac{3}{2}} \sqrt{\log T}\} = o_P(\psi_{\min}^2 p_{\min})$,

(iv) $\min\{|\mathcal{I}|_o^{1/2}, |\mathcal{T}|_o^{1/2}\} c_{\text{inv}} q^{\frac{5}{2}} \mu^{\frac{1}{2}} K \gamma^3 \max\{N^2, T^2\} \min\{\sqrt{N}, \sqrt{T}\} = o_P(\psi_{\min}^3 p_{\min}^{\frac{3}{2}})$.

Assumption C.3 (*Low-rank Approximation Error M^R*). The low-rank approximation error M^R satisfies the following condition:

$$\max_{i,t} |M_{it}^R| = o_P\left(\frac{p_{\min}^{\frac{5}{2}}}{\min\{|\mathcal{I}|_o^{1/2}, |\mathcal{T}|_o^{1/2}\} p_{\max}^2 q^2 \mu^{\frac{3}{2}} K^{\frac{3}{2}} \max\{\sqrt{N}, \sqrt{T}\}} + \frac{\psi_{\min} p_{\min}^2}{\min\{|\mathcal{I}|_o^{1/2}, |\mathcal{T}|_o^{1/2}\} p_{\max}^{\frac{3}{2}} q \mu^{\frac{1}{2}} K^{\frac{1}{2}} \max\{\sqrt{N}, \sqrt{T}\} \sqrt{NT}} \right)$$

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

Theorem C.1. *Suppose Assumptions 3.2, 3.3 and C.1–C.3 hold. In addition, suppose that $\left\| \frac{\sqrt{N}}{|\mathcal{I}|_o} \sum_{i \in \mathcal{I}} U_{M^*,i} \right\| \ge c$ and $\left\| \frac{\sqrt{T}}{|\mathcal{T}|_o} \sum_{t \in \mathcal{T}} V_{M^*,t} \right\| \ge c$ for some constant $c > 0$. 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 $\mathcal{V}_{\mathcal{G}} = \frac{1}{|\mathcal{T}|_o^2} \sum_{t \in \mathcal{T}} \bar{\beta}_{\mathcal{I}}' \left(\sum_{j=1}^N \omega_{jt} \beta_j \beta_j' \right)^{-1} \left(\sum_{j=1}^N \omega_{jt} \sigma_{jt}^2 \beta_j \beta_j' \right) \left(\sum_{j=1}^N \omega_{jt} \beta_j \beta_j' \right)^{-1} \bar{\beta}_{\mathcal{I}}$

$$+ \frac{1}{|\mathcal{I}|_o^2} \sum_{i \in \mathcal{I}} \bar{F}_{\mathcal{T}}' \left(\sum_{s=1}^T \omega_{is} F_s F_s' \right)^{-1} \left(\sum_{s=1}^T \omega_{is} \sigma_{is}^2 F_s F_s' \right) \left(\sum_{s=1}^T \omega_{is} F_s F_s' \right)^{-1} \bar{F}_{\mathcal{T}},$$

$\bar{\beta}_{\mathcal{I}} = \frac{1}{|\mathcal{I}|_o} \sum_{i \in \mathcal{I}} \beta_i$, $\bar{F}_{\mathcal{T}} = \frac{1}{|\mathcal{T}|_o} \sum_{s \in \mathcal{T}} F_s$. *In addition, Assumptions C.1–C.3 are satisfied under Assumptions 3.1–3.4 by setting $\mu = C\eta$ for some constant $C > 0$.*

In fact, Assumptions C.1–C.3 are verified by Lemma G.1.

Theorem C.2 (*Feasible CLT*). *Under the assumptions of Theorem C.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 $\widehat{\mathcal{V}}_{\mathcal{G}}$ is the same as the one in Theorem 3.1.

Appendix D. Formal definitions of the non-convex estimator and the leave-one-out estimator

Here, we introduce formal definitions of the non-convex optimization estimator $(\widetilde{W}^{[l]}, \widetilde{Z}^{[l]})$ and the leave-one-out estimator $(\breve{W}^{(l)}, \breve{Z}^{(l)})$ where $1 \le l \le N + T$. We start with defining the following two loss functions:

$$f^{\text{infs}}(w, z) := \frac{1}{2} \| \Pi^{-\frac{1}{2}} \mathcal{P}_{\Omega} (wz' - Y) \|_F^2 + \frac{\lambda}{2} \|w\|_F^2 + \frac{\lambda}{2} \|z\|_F^2, \tag{D.1}$$

$$f^{\text{infs},(l)}(w, z) \tag{D.2}$$

$$:= \begin{cases} \frac{1}{2} \left\| \Pi^{-1/2} \mathcal{P}_{\Omega_{-l,\cdot}}(wz' - Y) \right\|_F^2 + \frac{1}{2} \left\| \mathcal{P}_{l,\cdot}(wz' - M^\star) \right\|_F^2 + \frac{\lambda}{2} \|w\|_F^2 + \frac{\lambda}{2} \|z\|_F^2, & \text{if } 1 \le l \le N, \\ \frac{1}{2} \left\| \Pi^{-1/2} \mathcal{P}_{\Omega_{\cdot,-(l-N)}}(wz' - Y) \right\|_F^2 + \frac{1}{2} \left\| \mathcal{P}_{\cdot,(l-N)}(wz' - M^\star) \right\|_F^2 + \frac{\lambda}{2} \|w\|_F^2 + \frac{\lambda}{2} \|z\|_F^2, \\ & \text{if } N + 1 \le l \le N + T, \end{cases}$$

where w and z are $N \times K$ and $T \times K$ matrices, respectively.[4] The loss function (D.1) is for the non-convex optimization estimator $(\widetilde{W}^{[l]}, \widetilde{Z}^{[l]})$ and the loss function (D.2) is for the leave-one-out estimator $\breve{W}^{(l)}$. In the loss function (D.2), we use the following definitions. Let $C_{g(i)}$ be the cluster where the unit i is included in. For each $N \times T$ matrix D, let $\mathcal{P}_\Omega = \Omega \circ D$. Also, for each $N \times T$ matrix D and for each $1 \leq l \leq N$, let $\mathcal{P}_{\Omega_{-l,\cdot}}(D) := \Omega_{-l,\cdot} \circ D$ where $\Omega_{-l,\cdot} := [\omega_{js} 1\{j \notin C_{g(l)}\}]_{N \times T}$, and $\mathcal{P}_{l,\cdot}(D) := E_{l,\cdot} \circ D$ where $E_{l,\cdot} := [1\{j \in C_{g(l)}\}]_{N \times T}$. Roughly speaking, $f^{\mathrm{infs},(l)}$ changes $\{p_j^{-1} \omega_{js}, y_{js}\}_{j \in C_{g(l)}, s \leq T}$ in f^{infs} to its (approximate) population mean $\{1, M_{js}^\star\}_{j \in C_{g(l)}, s \leq T}$. Hence, the leave-one-out estimator constructed from the loss function $f^{\mathrm{infs},(l)}$ can be independent of $\{\omega_{ls}, \varepsilon_{ls}\}_{s \leq T}$ because $f^{\mathrm{infs},(l)}$ excludes $\{\omega_{js}, \varepsilon_{js}\}_{j \in C_{g(l)}, s \leq T}$ which is in the cluster where the unit l is included in.

On the other hand, for each $N + 1 \leq l \leq N + T$, we define $\mathcal{P}_{\Omega_{\cdot, -(l-N)}}(D) := \Omega_{\cdot, -(l-N)} \circ D$ where $\Omega_{\cdot, -(l-N)} := [\omega_{js} 1\{s \neq l - N\}]_{N \times T}$, and $\mathcal{P}_{\cdot, (l-N)}(D) := E_{\cdot, (l-N)} \circ D$ where $E_{\cdot, (l-N)} := [1\{s = l - N\}]_{N \times T}$. In this case, $f^{\mathrm{infs},(l)}$ changes $\{p_j^{-1} \omega_{js}, y_{js}\}_{j \leq N, s = l-N}$ in f^{infs} to $\{1, M_{js}^\star\}_{j \leq N, s = l-N}$. So, the leave-one-out estimator constructed from $f^{\mathrm{infs},(l)}$ is independent of $\{\omega_{j,(l-N)}, \varepsilon_{j,(l-N)}\}_{j \leq N}$ because $f^{\mathrm{infs},(l)}$ excludes $\{\omega_{j,(l-N)}, \varepsilon_{j,(l-N)}\}_{j \leq N}$ and $\omega_{js}, \varepsilon_{js}$ are independent across time.

To define the gradient descent iterates, we denote the singular value decomposition (SVD) of $M^\star$ by $U_{M^\star} D_{M^\star} V'_{M^\star}$ where $U'_{M^\star} U_{M^\star} = V'_{M^\star} V_{M^\star} = I_K$. $D_{M^\star}$ is a $K \times K$ diagonal matrix with singular values in descending order, i.e., $D_{M^\star} = \mathrm{diag}(\psi_1, \ldots, \psi_K)$ where $\psi_{\max} = \psi_1 > \cdots > \psi_K = \psi_{\min} > 0$. Then, based on (D.1), we define the following gradient descent iterates:

$$\begin{bmatrix} W^{\tau+1} \\ Z^{\tau+1} \end{bmatrix} = \begin{bmatrix} W^\tau - \eta \nabla_W f^{\mathrm{infs}}(W^\tau, Z^\tau) \\ Z^\tau - \eta \nabla_Z f^{\mathrm{infs}}(W^\tau, Z^\tau) \end{bmatrix} \tag{D.3}$$

where $W^0 = W := U_{M^\star} D_{M^\star}^{\frac{1}{2}}$, $Z^0 = Z := V_{M^\star} D_{M^\star}^{\frac{1}{2}}$, $\tau = 0, 1, \ldots, \tau_0 - 1$, and $\tau_0 = \max\{N^{18}, T^{18}\}$. Here, $\eta > 0$ is the step size. Similarly, for (D.2), we define

$$\begin{bmatrix} W^{\tau+1,(l)} \\ Z^{\tau+1,(l)} \end{bmatrix} = \begin{bmatrix} W^{\tau,(l)} - \eta \nabla_W f^{\mathrm{infs},(l)}(W^{\tau,(l)}, Z^{\tau,(l)}) \\ Z^{\tau,(l)} - \eta \nabla_Z f^{\mathrm{infs},(l)}(W^{\tau,(l)}, Z^{\tau,(l)}) \end{bmatrix} \tag{D.4}$$

where $W^{0,(l)} = W$, $Z^{0,(l)} = Z$. Note that the gradient descent iterates in (D.3) and (D.4) cannot be feasibly computed because the initial values (W, Z), the missing probability (Π), and the cluster structure are unknown. However, it does not cause any problem in the paper since we do not need to actually compute $W^\tau, Z^\tau, W^{\tau,(l)}$, and $Z^{\tau,(l)}$ and only use their existence and theoretical properties for the proof. We also define for each τ and l,

$$H^\tau := \underset{O \in \mathcal{O}^{K \times K}}{\arg\min} \|\mathcal{F}^\tau O - \mathcal{F}\|_F, \quad H^{\tau,(l)} := \underset{O \in \mathcal{O}^{K \times K}}{\arg\min} \left\|\mathcal{F}^{\tau,(l)} O - \mathcal{F}\right\|_F,$$

$$Q^{\tau,(l)} := \underset{O \in \mathcal{O}^{K \times K}}{\arg\min} \left\|\mathcal{F}^{\tau,(l)} O - \mathcal{F}^\tau H^\tau\right\|_F, \quad \text{where } \mathcal{F}^\tau := \begin{bmatrix} W^\tau \\ Z^\tau \end{bmatrix}, \quad \mathcal{F}^{\tau,(l)} := \begin{bmatrix} W^{\tau,(l)} \\ Z^{\tau,(l)} \end{bmatrix}, \quad \mathcal{F} := \begin{bmatrix} W \\ Z \end{bmatrix},$$

and $\mathcal{O}^{K \times K}$ is the set of $K \times K$ orthogonal matrix. Importantly, by the definition, $H^{\tau,(l)}$ is also independent to the observations in l.

In this paper, as emphasized in the main text, we consider the non-convex optimization estimator $(\widetilde{W}^{[l]}, \widetilde{Z}^{[l]})$ and the leave-one-out estimator $(\breve{W}^{(l)}, \breve{Z}^{(l)})$ at two different stopping points. Let $\tau_l^* := \arg\min_{0 \leq \tau < \tau_o} \left\|\nabla f^{\mathrm{infs},(l)}(W^{\tau,(l)}, Z^{\tau,(l)})\right\|_F$. First, we use the stopping point τ_l^*, i.e.,

$$(\widetilde{W}^{[l]}, \widetilde{Z}^{[l]}) := (W^{\tau_l^*}, Z^{\tau_l^*}) \quad \text{from (D.3)}, \quad (\breve{W}^{(l)}, \breve{Z}^{(l)}) := (W^{\tau_l^*,(l)}, Z^{\tau_l^*,(l)}) \quad \text{from (D.4)},$$

and $\widetilde{H}^{[l]} := H^{\tau_l^*}$, $\breve{H}^{(l)} := H^{\tau_l^*,(l)}$. For each l, we set the same iteration number τ_l^* for the non-convex optimization estimator $(\widetilde{W}^{[l]}, \widetilde{Z}^{[l]})$ and the leave-one-out estimator $(\breve{W}^{(l)}, \breve{Z}^{(l)})$ to ensure that they are close to each other. Note that, although the loss function (D.1) does not depend on l, due to τ_l^*, the non-convex optimization estimator $(\widetilde{W}^{[l]}, \widetilde{Z}^{[l]})$ depend on l. Namely, $(\widetilde{W}^{[l]}, \widetilde{Z}^{[l]})$ is selected to be close to the leave-one-out estimator $(\breve{W}^{(l)}, \breve{Z}^{(l)})$ among many gradient descent iterates in (D.3). At last, we choose $H_4^{[l]}$ so that $\psi_{\min}^{-1/2} \widetilde{W}^{[l]} H_4^{[l]}$ is the left singular vector of $\widetilde{W}^{[l]} \widetilde{Z}^{[l]'}$.

Secondly, we use the stopping point $\tau^* := \arg\min_{0 \leq \tau < \tau_o} \left\|\nabla f^{\mathrm{infs}}(W^\tau, Z^\tau)\right\|_F$. For brevity, we will use the same notations for the estimators. Namely,

$$(\widetilde{W}^{[l]}, \widetilde{Z}^{[l]}) := (W^{\tau^*}, Z^{\tau^*}) \quad \text{from (D.3)}, \quad (\breve{W}^{(l)}, \breve{Z}^{(l)}) := (W^{\tau^*,(l)}, Z^{\tau^*,(l)}) \quad \text{from (D.4)},$$

and $\widetilde{H}^{[l]} := H^{\tau^*}$, $\breve{H}^{(l)} := H^{\tau^*,(l)}$. Also, $H_4^{[l]}$ is defined similarly. Here, we are abusing notation in the sense that $(\widetilde{W}^{[l]}, \widetilde{Z}^{[l]})$, $\widetilde{H}^{[l]}$ and $H_4^{[l]}$ do not actually depend on l. However, this notational abuse is going to make the proofs more streamlined.

Remark 1. In the main text, to facilitate understanding and save space, we use simpler notations. Specifically, $(\breve{\beta}^{\mathrm{full},t}, \breve{\beta}^{(-t)}, \breve{\beta}^{\{-i\}})$ in the main text is the same as

$$\left(\breve{\beta}^{\mathrm{full},t}, \breve{\beta}^{(-t)}, \breve{\beta}^{\{-i\}}\right) := \left(\sqrt{N} \widetilde{W}^{[N+t]} \widetilde{H}^{[N+t]} D_{M^\star}^{-\frac{1}{2}}, \sqrt{N} \breve{W}^{(N+t)} \breve{H}^{(N+t)} D_{M^\star}^{-\frac{1}{2}}, \sqrt{N} \breve{W}^{(i)} \breve{H}^{(i)} D_{M^\star}^{-\frac{1}{2}}\right).$$

---

[4] In addition, $f^{\mathrm{fsbl}}(w, z)$ is defined by replacing Π with $\hat{\Pi}$ from $f^{\mathrm{infs}}(w, z)$.

Appendix E. Key part of proofs

As we mentioned in Section 2.3, the key for having an unbiased estimator for M_{it} is showing the following proposition:

Proposition E.1. *Suppose assumptions of Theorem C.1 hold.*[5] *Then, there is a $K \times K$ matrix H_2 so that*

$$\sqrt{N}(\hat{F}_t - H_2 F_t) = \sqrt{N} H_2 \left(\sum_{j=1}^{N} \omega_{jt} \beta_j \beta_j' \right)^{-1} \left(\sum_{j=1}^{N} \omega_{jt} \beta_j \varepsilon_{jt} \right) + \sqrt{N} R_t^F,$$

$$\max_t \| \sqrt{N} R_t^F \|$$

$$= O_P \left(\frac{\sigma p_{\max}^{\frac{3}{2}} \vartheta c_{\mathrm{inv}} q^{\frac{11}{2}} \mu^{\frac{3}{2}} K^{\frac{5}{2}} \sqrt{N} \max\{ \sqrt{\log N}, \sqrt{\log T} \}}{p_{\min}^3 \min\{N, T\}} + \frac{\sigma^2 p_{\max}^{\frac{5}{2}} \vartheta c_{\mathrm{inv}}^2 q^3 \mu K^2 \sqrt{N} \max\{ \sqrt{N \log N}, \sqrt{T \log T} \}}{\psi_{\min} p_{\min}^4 \min\{ \sqrt{N}, \sqrt{T} \}} \right.$$

$$\left. + \frac{\sigma^3 p_{\max}^{\frac{3}{2}} c_{\mathrm{inv}} q^{\frac{5}{2}} K^{\frac{1}{2}} \sqrt{N} \max\{N, T\}}{\psi_{\min}^2 p_{\min}^3} + \frac{p_{\max}^{\frac{1}{2}} \sqrt{N}}{p_{\min}} \max_{it} \left| M_{it}^R \right| \right) = o_P(1).$$

E.1. Important lemmas

An important step is to show that uniformly in t, the following two terms are negligible:

$$\frac{1}{\sqrt{N}} \sum_{j=1}^{N} \omega_{jt} \varepsilon_{jt} (\widetilde{\beta}_j - \breve{\beta}_j^{\mathrm{full},t}), \quad \frac{1}{\sqrt{N}} \sum_{j=1}^{N} \omega_{jt} \varepsilon_{jt} (\breve{\beta}_j^{\mathrm{full},t} - \breve{\beta}_j^{(-t)}). \tag{E.1}$$

The proof follows from Lemma E.2 below.

Lemma E.2. *Suppose assumptions of Theorem C.1 hold. Uniformly in $t \leq T$, the two terms in (E.1) are both $o_P(1)$. Specifically, their order is*

$$O_P \left(\frac{\sigma^2 p_{\max}^{\frac{3}{2}} \vartheta^{\frac{1}{2}} c_{\mathrm{inv}} q^{\frac{9}{2}} \mu^{\frac{1}{2}} K^{\frac{3}{2}} \sqrt{N} \max\{ \sqrt{N \log N}, \sqrt{T \log T} \}}{p_{\min}^2 \min\{ \sqrt{N}, \sqrt{T} \} \psi_{\min}} + \frac{\sigma^3 p_{\max}^{\frac{3}{2}} c_{\mathrm{inv}} q^{\frac{5}{2}} K^{\frac{1}{2}} \sqrt{N} \max\{N, T\}}{p_{\min}^2 \psi_{\min}^2} \right).$$

In addition, we have the following results:

(i) $\quad \max_t \| \widetilde{W}^{[t+N]} \widetilde{H}^{[t+N]} - \breve{W}^{(t+N)} \breve{H}^{(t+N)} \|_F = O_P \left(\dfrac{\sigma p_{\max}^{\frac{1}{2}} \vartheta^{\frac{1}{2}} q^{\frac{3}{2}} \mu^{\frac{1}{2}} K^{\frac{1}{2}} \max\{ \sqrt{N \log N}, \sqrt{T \log T} \}}{p_{\min} \psi_{\min}^{1/2} \min\{ \sqrt{N}, \sqrt{T} \}} \right),$

(ii) $\quad \max_t \| \widetilde{W}^{[t+N]} \widetilde{H}^{[t+N]} - W \| = O_P \left(\dfrac{\sigma p_{\max}^{\frac{1}{2}} q^{\frac{1}{2}} \max\{ \sqrt{N}, \sqrt{T} \}}{p_{\min} \psi_{\min}^{1/2}} \right),$

(iii) $\quad \max_t \| \widetilde{W}^{[t+N]} \widetilde{Z}^{[t+N]\prime} - \widetilde{M} \|_F = O_P \left(\dfrac{\sigma p_{\max} \vartheta^{\frac{1}{2}} q^{\frac{7}{2}} \mu^{\frac{1}{2}} K \max\{ \sqrt{N \log N}, \sqrt{T \log T} \}}{p_{\min}^2 \min\{ \sqrt{N}, \sqrt{T} \}} \right),$

(iv) $\quad \| \widetilde{M} - M^\star \| = O_P \left(\dfrac{\sigma p_{\max}^{\frac{1}{2}} q \max\{ \sqrt{N}, \sqrt{T} \}}{p_{\min}} \right),$

(v) $\quad \max_t \| \breve{W}^{(t+N)} \breve{H}^{(t+N)} - W \|_{2,\infty} = O_P \left(\dfrac{\sigma p_{\max}^{\frac{1}{2}} \vartheta^{\frac{1}{2}} q^{\frac{3}{2}} \mu^{\frac{1}{2}} K^{\frac{1}{2}} \max\{ \sqrt{N \log N}, \sqrt{T \log T} \}}{p_{\min} \psi_{\min}^{1/2} \min\{ \sqrt{N}, \sqrt{T} \}} \right).$

Proof of Lemma E.2. First of all, by Lemmas H.1–H.5, we have (H.1), (H.2), (H.3), (H.4) and (H.5). Hence, we have (i)–(v). Next, we prove terms in (E.1) are $o_P(1)$. By Remark 1, the first term is written as

$$\frac{1}{\sqrt{N}} \sum_{j=1}^{N} \omega_{jt} \varepsilon_{jt} (\widetilde{\beta}_j - \breve{\beta}_j^{\mathrm{full},t}) = N^{-\frac{1}{2}} (\widetilde{\beta} - \sqrt{N} \widetilde{W}^{[t+N]} \widetilde{H}^{[t+N]} D_{M^\star}^{-\frac{1}{2}})' \Omega_t \mathcal{E}_t$$

$$= N^{-\frac{1}{2}} (\widetilde{\beta} - \sqrt{N} \psi_{\min}^{-1/2} \widetilde{W}^{[t+N]} H_4^{[t+N]})' \Omega_t \mathcal{E}_t + \psi_{\min}^{-1/2} (H_4^{[t+N]} - \widetilde{H}^{[t+N]} D_{M^\star}^{-\frac{1}{2}} \psi_{\min}^{-1/2})' \widetilde{W}^{[t+N]\prime} \Omega_t \mathcal{E}_t \tag{E.2}$$

---

where $H_4^{[N+t]}$ is a $K \times K$ matrix introduced in Claim G.2, $\Omega_t = \text{diag}(\omega_{1t}, \ldots, \omega_{Nt})$, and $\mathcal{E}_t = [\varepsilon_{1t}, \ldots, \varepsilon_{Nt}]'$. As noted in Claim G.2(iii), we derive from Lemma E.2(iii) that

$$\max_{1 \le t \le T} \left\| \tilde{\beta} - \sqrt{N} \psi_{\min}^{-1/2} \widetilde{W}^{[t+N]} H_4^{[t+N]} \right\|_F = O_P \left(\frac{\sigma p_{\max} \vartheta^{\frac{1}{2}} c_{\text{inv}} q^{\frac{9}{2}} \mu^{\frac{1}{2}} K^{\frac{3}{2}} \sqrt{N} \max\{\sqrt{N \log N}, \sqrt{T \log T}\}}{p_{\min}^2 \min\{\sqrt{N}, \sqrt{T}\} \psi_{\min}} \right).$$

Hence, the first term of (E.2) is $O_P \left(\frac{\sigma^2 p_{\max}^{\frac{3}{2}} \vartheta^{\frac{1}{2}} c_{\text{inv}} q^{\frac{9}{2}} \mu^{\frac{1}{2}} K^{\frac{3}{2}} \sqrt{N} \max\{\sqrt{N \log N}, \sqrt{T \log T}\}}{p_{\min}^2 \min\{\sqrt{N}, \sqrt{T}\} \psi_{\min}} \right)$. For the second term of (E.2), note that

$$\max_t \| H_4^{[t+N]} - \widetilde{H}^{[t+N]} D_{M\star}^{-\frac{1}{2}} \psi_{\min}^{1/2} \|$$

$$\overset{(i)}{\le} \psi_{\min}^{1/2} O_P \left(\psi_{\min}^{-1/2} \right) \left[\max_t \| W - \widetilde{W}^{[t+N]} \widetilde{H}^{[t+N]} \| \| D_{M\star}^{-\frac{1}{2}} \| + \psi_{\min}^{-1/2} \max_t \| \widetilde{W}^{[t+N]} H_4^{[t+N]} - W D_{M\star}^{-\frac{1}{2}} \psi_{\min}^{1/2} \| \right]$$

$$\overset{(ii)}{=} O_P \left(\frac{\sigma p_{\max}^{\frac{1}{2}} c_{\text{inv}} q^2 K^{\frac{1}{2}} \max\{\sqrt{N}, \sqrt{T}\}}{p_{\min} \psi_{\min}} \right).$$

Here, (i) comes from Claim G.5(i), and (ii) comes from Lemma E.2(ii) and Claim G.5(ii). In addition,

$$\max_t \| \widetilde{W}^{[t+N]\prime} \Omega_t \mathcal{E}_t \| \le \max_t \| (\widetilde{W}^{[t+N]} \widetilde{H}^{[t+N]})' \Omega_t \mathcal{E}_t \| \le \max_t \| \widetilde{W}^{[t+N]} \widetilde{H}^{[t+N]} - W \| \| \Omega_t \mathcal{E}_t \| + \max_t \| W' \Omega_t \mathcal{E}_t \|.$$

From Lemma E.2(ii), we know $\max_t \| \widetilde{W}^{[t+N]} \widetilde{H}^{[t+N]} - W \| \| \Omega_t \mathcal{E}_t \| = O_P \left(\frac{\sigma^2 p_{\max} q^{\frac{1}{2}} \sqrt{N} \max\{\sqrt{N}, \sqrt{T}\}}{p_{\min} \psi_{\min}^{1/2}} \right)$. In addition, we have $\max_t \| W' \Omega_t \mathcal{E}_t \|$

$= O_P(\sigma q^{\frac{1}{2}} K^{\frac{1}{2}} \sqrt{\log T} \psi_{\min}^{1/2})$ from the matrix Bernstein inequality because $W = U_{M\star} D_{M\star}^{\frac{1}{2}}$. Hence, the second term of (E.2) is

$$O_P \left(\frac{\sigma^3 p_{\max}^{\frac{3}{2}} c_{\text{inv}} q^{\frac{5}{2}} K^{\frac{1}{2}} \sqrt{N} \max\{N, T\}}{p_{\min}^2 \psi_{\min}^2} + \frac{\sigma^2 p_{\max}^{\frac{1}{2}} c_{\text{inv}} q^{\frac{5}{2}} K \sqrt{\log T} \max\{\sqrt{N}, \sqrt{T}\}}{p_{\min} \psi_{\min}} \right).$$

Moreover, the second term of (E.1) can be written as

$$\frac{1}{\sqrt{N}} \sum_{j=1}^N \omega_{jt} \varepsilon_{jt} (\breve{\beta}_j^{\text{full},t} - \breve{\beta}_j^{(-t)}) = D_{M\star}^{-\frac{1}{2}} \left(\widetilde{W}^{[t+N]} \widetilde{H}^{[t+N]} - \breve{W}^{(t+N)} \breve{H}^{(t+N)} \right)' \Omega_t \mathcal{E}_t.$$

Then, we have from Lemma E.2(i) that

$$\max_t \| D_{M\star}^{-\frac{1}{2}} \left(\widetilde{W}^{[t+N]} \widetilde{H}^{[t+N]} - \breve{W}^{(t+N)} \breve{H}^{(t+N)} \right)' \Omega_t \mathcal{E}_t \| = O_P \left(\frac{\sigma^2 p_{\max} \vartheta^{\frac{1}{2}} q^{\frac{3}{2}} \mu^{\frac{1}{2}} K^{\frac{1}{2}} \sqrt{N} \max\{\sqrt{N \log N}, \sqrt{T \log T}\}}{p_{\min} \psi_{\min} \min\{\sqrt{N}, \sqrt{T}\}} \right).$$

This completes the proof. $\quad \square$

In addition, the following lemma shows the part in which the proofs are different depending on how we define the stopping point.

Lemma E.3. *Suppose assumptions of Theorem C.1 hold.*[6] *Then, we have*

(1) $\quad \max_t \| \frac{1}{\sqrt{N}} \sum_{j=1}^N \omega_{jt} \varepsilon_{jt} (\breve{\beta}_j^{(-t)} - H_1' \beta_j) \| = O_P \left(\frac{\sigma^2 p_{\max}^{\frac{1}{2}} \vartheta^{\frac{1}{2}} q^{\frac{1}{2}} K^{\frac{1}{2}} \sqrt{\log T} \max\{\sqrt{N}, \sqrt{T}\}}{p_{\min} \psi_{\min}} \right) = o_P(1),$

(2) $\quad \max_t \| \frac{1}{\sqrt{N}} \sum_{j=1}^N (\omega_{jt} - p_j) H_1' \beta_j (\breve{\beta}_j^{(-t)} - H_1' \beta_j) \| = O_P \left(\frac{\sigma p_{\max} \vartheta q^{\frac{1}{2}} \mu^{\frac{1}{2}} K \sqrt{\log T} \max\{\sqrt{N}, \sqrt{T}\}}{p_{\min} \psi_{\min}} \right) = o_P(1).$

Proof of Lemma E.3. (1)-i. Case of using τ_l^* as a stopping point:

Let $\xi_t := \breve{\beta}^{(-t)} - \beta H_1 = \sqrt{N} \breve{W}^{(t+N)} \breve{H}^{(t+N)} D_{M\star}^{-\frac{1}{2}} - \beta H_1$. To employ matrix Bernstein inequality, we first estimate $\max_t \| \xi_t \|_{2,\infty}$. Note $\| \xi_t \|_{2,\infty} \le \sqrt{N} \psi_{\min}^{-1/2} \| \breve{W}^{(t+N)} \breve{H}^{(t+N)} - W \|_{2,\infty}$. So, by Lemma E.2(v), we have

$$\max_t \| \xi_t \|_{2,\infty} = O_P \left(\frac{\sigma p_{\max}^{\frac{1}{2}} \vartheta^{\frac{1}{2}} q^{\frac{3}{2}} \mu^{\frac{1}{2}} K^{\frac{1}{2}} \sqrt{N} \max\{\sqrt{N \log N}, \sqrt{T \log T}\}}{p_{\min} \psi_{\min} \min\{\sqrt{N}, \sqrt{T}\}} \right).$$

---

[6] By Lemma G.1, it is enough to consider the assumptions of Theorem C.1.

Furthermore, we have

$$\max_t \|\xi_t\|_F \le \sqrt{N} \left(\max_t \|\widetilde{W}^{[t+N]} \widetilde{H}^{[t+N]} - \breve{W}^{(t+N)} \breve{H}^{(t+N)}\|_F + \|W - \widetilde{W}^{[t+N]} \widetilde{H}^{[t+N]}\|_F \right) \|D_{M^\star}^{-\frac{1}{2}}\|$$

$$= O_P \left(\frac{\sigma p_{\max}^{\frac{1}{2}} q^{\frac{1}{2}} K^{\frac{1}{2}} \sqrt{N} \max\{\sqrt{N}, \sqrt{T}\}}{p_{\min} \psi_{\min}} \right).$$

Because ξ_t only depends on $M^\star$ and Y excluding the tth column of Y, conditioning on $\{\mathcal{M}, \Omega\}$, $\{\varepsilon_{jt}\}_{j \le N}$ are independent of ξ_t. Hence, $\mathbb{E}\left[\varepsilon_{jt}|\mathcal{M}, \Omega, \xi_t\right] = \mathbb{E}\left[\varepsilon_{jt}|\mathcal{M}, \Omega\right] = 0$ and, conditioning on $\{\mathcal{M}, \Omega, \xi_t\}$, $\{\varepsilon_{jt}\}_{j \le N}$ are independent across j. Then, by matrix Bernstein inequality, we have

$$\|\xi_t' \Omega_t \mathcal{E}_t\| = \| \sum_{j=1}^N \omega_{jt} \varepsilon_{jt} \xi_{t,j}'\| \le C \left(\sigma \log T \log N \max_t \|\xi_t\|_{2,\infty} + \sigma \sqrt{\log T} \max_t \|\xi_t\|_F \right)$$

with probability exceeding $1 - O(T^{-100})$ and so, $\max_t \|\xi_t' \Omega_t \mathcal{E}_t\| = O_P \left(\frac{\sigma^2 p_{\max}^{\frac{1}{2}} \vartheta^{\frac{1}{2}} q^{\frac{1}{2}} K^{\frac{1}{2}} \sqrt{N \log T} \max\{\sqrt{N}, \sqrt{T}\}}{p_{\min} \psi_{\min}} \right).$

(1)-ii. Case of using τ^* as a stopping point:

In this case, we note that ξ_t is no longer independent of $\{\varepsilon_{jt}\}_{j \le N}$ conditioning on $\{\mathcal{M}, \Omega\}$, due to the fact that τ^* does depend on the full sample. Therefore, we cannot directly apply the Bernstein inequality as in the τ_l^* case. Instead, we apply Lemma H.10 and obtain the same bound for $\max_t \|\xi_t' \Omega_t \mathcal{E}_t\|$.

(2)-i. Case of using τ_l^* as a stopping point:

The proof is similar to that in (1-i). So, we omit it.

(2)-ii. Case of using τ^* as a stopping point:

The proof is the same as that in (1-ii) although we use Lemma H.11 instead. □

E.2. Proof of Proposition E.1

First of all, by Claim G.1(i), we can know that there is a $K \times K$ matrix H_1 such that $\frac{1}{\sqrt{N}} \beta H_1$ is the left singular vector of $M^\star$. That is, $\frac{1}{\sqrt{N}} \beta H_1 = U_{M^\star}$. Let $\widetilde{B}_t := \frac{1}{N} \sum_{j=1}^N \omega_{jt} \widetilde{\beta}_j \widetilde{\beta}_j'$, $B_t^* := \frac{1}{N} \sum_{j=1}^N \omega_{jt} H_1' \beta_j \beta_j' H_1$ and $B := \frac{1}{N} \sum_{j=1}^N p_j H_1' \beta_j \beta_j' H_1$. Then, we define $H_2 := (I_K + \varphi) H_1^{-1}$ where $\varphi := \frac{1}{N} B^{-1} H_1' \beta' \Pi \left(\beta H_1 - \widetilde{\beta} \right)$. Note that both B and H_2 do not depend on i or t. Because $\widehat{F}_t = \left(\sum_{j=1}^N \omega_{jt} \widetilde{\beta}_j \widetilde{\beta}_j' \right)^{-1} \sum_{j=1}^N \omega_{jt} \widetilde{\beta}_j y_{jt}$ by definition, basic algebras shows the following identity:

$$\widehat{F}_t - H_2 F_t = H_2 \left(\sum_{j=1}^N \omega_{jt} \beta_j \beta_j' \right)^{-1} \left(\sum_{j=1}^N \omega_{jt} \beta_j \varepsilon_{jt} \right) + \sum_{d=1}^6 \Delta_{d,t},$$

$$\Delta_{1,t} := \widetilde{B}_t^{-1} \frac{1}{N} \sum_{j=1}^N \omega_{jt} \varepsilon_{jt} \left(\widetilde{\beta}_j - H_1' \beta_j \right) - B^{-1} H_1' \frac{1}{N} \sum_{j=1}^N (\omega_{jt} - p_j) \beta_j F_t' H_1'^{-1} \left(\widetilde{\beta}_j - H_1' \beta_j \right),$$

$$\Delta_{2,t} := \left(\widetilde{B}_t^{-1} - B^{-1} \right) \frac{1}{N} \sum_{j=1}^N \omega_{jt} \widetilde{\beta}_j \left(\beta_j' H_1 - \widetilde{\beta}_j' \right) H_1^{-1} F_t,$$

$$\Delta_{3,t} := B^{-1} \frac{1}{N} \sum_{j=1}^N \omega_{jt} \left(\widetilde{\beta}_j - H_1' \beta_j \right) \left(\beta_j' H_1 - \widetilde{\beta}_j' \right) H_1^{-1} F_t,$$

$$\Delta_{4,t} := \left(\widetilde{B}_t^{-1} - B_t^{*-1} \right) H_1' \frac{1}{N} \sum_{j=1}^N \omega_{jt} \beta_j \varepsilon_{jt}, \quad \Delta_{5,t} := \left(H_1^{-1} - H_2 \right) \left(\sum_{j=1}^N \omega_{jt} \beta_j \beta_j' \right)^{-1} \left(\sum_{j=1}^N \omega_{jt} \beta_j \varepsilon_{jt} \right),$$

$$\Delta_{6,t} := \widetilde{B}_t^{-1} \frac{1}{N} \sum_{j=1}^N \omega_{jt} \widetilde{\beta}_j M_{jt}^R.$$

Step 1. We start from the first term of $\Delta_{1,t}$: $P_1 := \widetilde{B}_t^{-1} \frac{1}{N} \sum_{j=1}^N \omega_{jt} \varepsilon_{jt} \left(\widetilde{\beta}_j - H_1' \beta_j \right)$. We have $P_1 = P_{1,1} + P_{1,2}$ where

$$P_{1,1} := \widetilde{B}_t^{-1} \frac{1}{N} \sum_{j=1}^N \omega_{jt} \varepsilon_{jt} \left(\widetilde{\beta}_j - \breve{\beta}_j^{(-t)} \right) = \frac{1}{N} \widetilde{B}_t^{-1} \left(\widetilde{\beta} - \sqrt{N} \breve{W}^{(N+t)} \breve{H}^{(N+t)} D_{M^\star}^{-\frac{1}{2}} \right)' \Omega_t \mathcal{E}_t,$$

$$P_{1,2} := \widetilde{B}_t^{-1} \frac{1}{N} \sum_{j=1}^N \omega_{jt} \varepsilon_{jt} \left(\breve{\beta}_j^{(-t)} - H_1' \beta_j \right) = \frac{1}{N} \widetilde{B}_t^{-1} \left(\sqrt{N} \breve{W}^{(N+t)} \breve{H}^{(N+t)} D_{M^\star}^{-\frac{1}{2}} - \beta H_1 \right)' \Omega_t \mathcal{E}_t.$$

Note that $\max_t \|\widetilde{B}_t^{-1}\| = O_P(\frac{1}{p_{\min}})$ by Claim G.4(iii). Hence, we have by Lemma E.2,

$$\max_t \|P_{1,1}\| \le \max_t \|\widetilde{B}_t^{-1}\| N^{-\frac{1}{2}} \max_t \|N^{-\frac{1}{2}} (\widetilde{\beta} - \sqrt{N} \breve{W}^{(N+t)} \breve{H}^{(N+t)} D_{M^\star}^{-\frac{1}{2}})' \Omega_t \mathcal{E}_t\|$$

$$= O_P \left(\frac{\sigma^2 p_{\max}^{\frac{3}{2}} \vartheta^{\frac{1}{2}} c_{\text{inv}} q^{\frac{9}{2}} \mu^{\frac{1}{2}} K^{\frac{3}{2}} \max\{\sqrt{N \log N}, \sqrt{T \log T}\}}{p_{\min}^3 \min\{\sqrt{N}, \sqrt{T}\} \psi_{\min}} + \frac{\sigma^3 p_{\max}^{\frac{3}{2}} c_{\text{inv}} q^{\frac{5}{2}} K^{\frac{1}{2}} \max\{N, T\}}{p_{\min}^3 \psi_{\min}^2} \right).$$

Note that $\max_t \|P_{1,2}\| \le \frac{1}{N} \|\widetilde{B}_t^{-1}\| \max_t \|\xi_t' \Omega_t \mathcal{E}_t\|$. Then, using Lemma E.3, we have

$$\max_t \|P_{1,2}\| = O_P \left(\frac{\sigma^2 p_{\max}^{\frac{1}{2}} \vartheta^{\frac{1}{2}} q^{\frac{1}{2}} K^{\frac{1}{2}} \sqrt{\log T} \max\{\sqrt{N}, \sqrt{T}\}}{p_{\min}^2 \sqrt{N} \psi_{\min}} \right).$$

Step 2. By using the same logic in Step 1, we can bound the second term of $\Delta_{1,t}$,

$P_2 := B^{-1} H_1' \frac{1}{N} \sum_{j=1}^N (\omega_{jt} - p_j) \beta_j F_t' H_1'^{-1} (\widetilde{\beta}_j - H_1' \beta_j)$ similarly. The only difference is the part using the matrix Bernstein inequality since $\{\omega_{jt}\}_{j \le N}$ are dependent across j while $\{\varepsilon_{jt}\}_{j \le N}$ are independent across j. We split P_2 like $P_2 = P_{2,1} + P_{2,2}$ where

$$P_{2,1} := \frac{1}{N} B^{-1} H_1' \beta' (\Omega_t - \Pi) \left(\widetilde{\beta} - \sqrt{N} \breve{W}^{(t+N)} \breve{H}^{(t+N)} D_{M^*}^{-\frac{1}{2}} \right) H_1^{-1} F_t,$$

$$P_{2,2} := \frac{1}{N} B^{-1} H_1' \beta' (\Omega_t - \Pi) \left(\sqrt{N} \breve{W}^{(t+N)} \breve{H}^{(t+N)} D_{M^*}^{-\frac{1}{2}} - \beta H_1 \right) H_1^{-1} F_t.$$

By the same token as the part $P_{1,1}$ in Step 1 with the aids of Claims G.1–G.5, we can show that

$$P_{2,1} = O_P \left(\frac{\sigma^2 p_{\max}^{\frac{3}{2}} c_{\text{inv}} q^{\frac{7}{2}} \mu K \max\{\sqrt{N}, \sqrt{T}\}}{\psi_{\min} p_{\min}^3 \min\{\sqrt{N}, \sqrt{T}\}} + \frac{\sigma p_{\max}^{\frac{3}{2}} \vartheta q^{\frac{11}{2}} \mu^{\frac{3}{2}} K^{\frac{5}{2}} \max\{\sqrt{\log N}, \sqrt{\log T}\}}{p_{\min}^3 \min\{N, T\}} \right).$$

and so, we omit the proof. In addition, using Lemma E.3, the part $P_{2,2}$ can be bounded like

$$\max_t \|P_{2,2}\| \le \frac{1}{\sqrt{N}} \|B^{-1}\| \max_t \|\frac{1}{\sqrt{N}} H_1' \beta' (\Omega_t - \Pi) \xi_t\| \max_t \|H_1^{-1} F_t\| = O_P \left(\frac{\sigma p_{\max} \vartheta q^{\frac{3}{2}} \mu K^{\frac{3}{2}} \sqrt{\log T}}{p_{\min}^2 \sqrt{N} \min\{\sqrt{N}, \sqrt{T}\}} \right).$$

Step 3. We bound $\max_t \|\Delta_{2,t}\|$. By Claim G.1(iv) and Claim G.3(ii)

$$\max_t \|\Delta_{2,t}\| \le O_P(1) \max_t \|\widetilde{B}_t^{-1} - B^{-1}\| \max_j \|H_1 \beta_j\| p_{\max}^{\frac{1}{2}} \frac{1}{\sqrt{N}} \|\beta H_1 - \widetilde{\beta}\|_F \max_t \|H_1^{-1} F_t\|$$

$$= O_P \left(\frac{\sigma^2 p_{\max}^{\frac{5}{2}} c_{\text{inv}}^2 q^5 \mu K^2 \max\{\sqrt{N}, \sqrt{T}\}}{p_{\min}^4 \min\{\sqrt{N}, \sqrt{T}\} \psi_{\min}} + \frac{\sigma p_{\max}^{\frac{3}{2}} c_{\text{inv}} \vartheta q^3 \mu^{\frac{3}{2}} K^{\frac{5}{2}} \sqrt{\log T}}{\sqrt{N} \min\{\sqrt{N}, \sqrt{T}\}} \right).$$

Step 4. We now bound $\max_t \|\Delta_{3,t}\|$. By Claim G.1(iv) and Claim G.3(ii), we have

$$\max_t \|\Delta_{3,t}\| \le O_P(1) \|B^{-1}\| \frac{1}{\sqrt{N}} \|\widetilde{\beta} - \beta H_1\| \|\Pi\| \frac{1}{\sqrt{N}} \|\widetilde{\beta} - \beta H_1\| \max_t \|H_1^{-1} F_t\|$$

$$= O_P \left(\frac{\sigma^2 p_{\max}^2 c_{\text{inv}}^2 q^5 \mu^{\frac{1}{2}} K^{\frac{3}{2}} \max\{\sqrt{N}, \sqrt{T}\}}{p_{\min}^3 \min\{\sqrt{N}, \sqrt{T}\} \psi_{\min}} \right).$$

Step 5. We estimate $\max_t \|\Delta_{4,t}\|$. By Claims G.4(iv) and G.6(i), we have

$$\max_t \|\Delta_{4,t}\| \le \frac{1}{N} \max_t \|\widetilde{B}_t^{-1} - B_t^{*-1}\| \max_t \|(\beta H_1)' \Omega_t \mathcal{E}_t\| = O_P \left(\frac{\sigma^2 p_{\max}^2 \vartheta c_{\text{inv}} q^2 K \sqrt{\log T} \max\{\sqrt{N}, \sqrt{T}\}}{p_{\min}^3 \sqrt{N} \psi_{\min}} \right).$$

Step 6. We bound $\max_t \|\Delta_{5,t}\|$. First, note that $H_2 - H_1^{-1} = \varphi H_1^{-1}$ and $\|\varphi\| = O_P \left(\frac{\sigma p_{\max}^{\frac{1}{2}} c_{\text{inv}} q^2 K^{\frac{1}{2}} \max\{\sqrt{N}, \sqrt{T}\}}{p_{\min} \psi_{\min}} \right)$ as noted in the proof of Claim G.3. Moreover, by Claim G.4(iv), we have $\max_t \|H_1^{-1} (\sum_{j=1}^N \omega_{jt} \beta_j \beta_j')^{-1} H_1'^{-1}\| = \|(N B_t^*)^{-1}\| = O_P(\frac{1}{p_{\min} N})$. Hence, by Claim G.6(i),

$$\max_t \|\Delta_{5,t}\| \le \|\varphi\| \|H_1^{-1} (\sum_{j=1}^N \omega_{jt} \beta_j \beta_j')^{-1} H_1'^{-1}\| \max_t \|(\beta H_1)' \Omega_t \mathcal{E}_t\| = O_P \left(\frac{\sigma^2 p_{\max}^{\frac{1}{2}} c_{\text{inv}} q^2 K^{\frac{1}{2}} \max\{\sqrt{N}, \sqrt{T}\}}{p_{\min}^2 \psi_{\min}} \right).$$

Step 7. Lastly, we bound $\max_t \|\Delta_{6,t}\|$. Note that

$$\Delta_{6,t} = \left(\widetilde{B}_t^{-1} - B^{-1} \right) \frac{1}{N} \sum_{j=1}^N \omega_{jt} H_1' \beta_j M_{jt}^R + B^{-1} \frac{1}{N} \sum_{j=1}^N \omega_{jt} \left(\widetilde{\beta}_j - H_1' \beta_j \right) M_{jt}^R$$

$$+ \left(\widetilde{B}_t^{-1} - B^{-1} \right) \frac{1}{N} \sum_{j=1}^{N} \omega_{jt} \left(\widetilde{\beta}_j - H_1' \beta_j \right) M_{jt}^R + B^{-1} \frac{1}{N} \sum_{j=1}^{N} \omega_{jt} H_1' \beta_j M_{jt}^R.$$

By Claims G.1, G.3 and G.4, the last term dominates the first three terms. The last term is

$$\max_t \| B^{-1} \frac{1}{N} \sum_{j=1}^{N} \omega_{jt} H_1' \beta_j M_{jt}^R \| \le \frac{1}{\sqrt{N}} \| B^{-1} \| \| \beta H_1 \| p_{\max}^{\frac{1}{2}} \max_{it} |M_{it}^R| = O_P \left(\frac{p_{\max}^{\frac{1}{2}}}{p_{\min}} \right) \max_{it} |M_{it}^R|$$

by Claims G.3 and G.4. This completes the proof. $\square$

Appendix F. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2024.105682. For the sake of brevity, some of the technical proofs and simulation results are relegated to the Supplement.

References

Agarwal, A., Dahleh, M., Shah, D., Shen, D., 2021. Causal matrix completion. arXiv preprint arXiv:2109.15154.

Anderson, G.M., Tollison, R.D., 1991. Congressional influence and patterns of new deal spending, 1933–1939. J. Law Econom. 34 (1), 161–175.

Arkhangelsky, D., Athey, S., Hirshberg, D.A., Imbens, G.W., Wager, S., 2021. Synthetic difference-in-differences. Amer. Econ. Rev. 111 (12), 4088–4118.

Athey, S., Bayati, M., Doudchenko, N., Imbens, G., Khosravi, K., 2021. Matrix completion methods for causal panel data models. J. Amer. Statist. Assoc. 116 (536), 1716–1730.

Bai, J., 2003. Inferential theory for factor models of large dimensions. Econometrica 71 (1), 135–171.

Bai, J., Ng, S., 2021. Matrix completion, counterfactuals, and factor analysis of missing data. J. Amer. Statist. Assoc. 116 (536), 1746–1763.

Beck, A., Teboulle, M., 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. 2 (1), 183–202.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B Stat. Methodol. 57 (1), 289–300.

Berry, C.R., Burden, B.C., Howell, W.G., 2010. The president and the distribution of federal spending. Am. Polit. Sci. Rev. 104 (4), 783–799.

Cai, J.-F., Candès, E.J., Shen, Z., 2010. A singular value thresholding algorithm for matrix completion. SIAM J. Optim. 20 (4), 1956–1982.

Candes, E.J., Plan, Y., 2010. Matrix completion with noise. Proc. IEEE 98 (6), 925–936.

Candès, E.J., Recht, B., 2009. Exact matrix completion via convex optimization. Found. Comput. Math. 9 (6), 717–772.

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

Chen, Y., Fan, J., Ma, C., Yan, Y., 2019. Inference and uncertainty quantification for noisy matrix completion. Proc. Natl. Acad. Sci. 116 (46), 22931–22937.

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

Chernozhukov, V., Hansen, C.B., Liao, Y., Zhu, Y., 2019. Inference for Heterogeneous Effects Using Low-Rank Estimations. Technical Report, Cemmap Working Paper.

Chernozhukov, V., Hansen, C., Liao, Y., Zhu, Y., 2023. Inference for low-rank models. Ann. Statist. 51 (3), 1309–1330.

Cox, G.W., McCubbins, M.D., 1986. Electoral politics as a redistributive game. J. Polit. 48 (2), 370–389.

Farias, V., Li, A., Peng, T., 2021. Learning treatment effects in panels with general intervention patterns. Adv. Neural Inf. Process. Syst. 34, 14001–14013.

Imbens, G.W., Rubin, D.B., 2015. Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge University Press.

Jankova, J., Van De Geer, S., 2018. Semiparametric efficiency bounds for high-dimensional models. Ann. Statist. 46 (5), 2336–2359.

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

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

Larcinese, V., Rizzo, L., Testa, C., 2006. Allocating the US federal budget to the states: The impact of the president. J. Polit. 68 (2), 447–456.

Little, R.J., Rubin, D.B., 2019. Statistical Analysis with Missing Data, vol. 793, John Wiley & Sons.

Ma, W., Chen, G.H., 2019. Missing not at random in matrix completion: The effectiveness of estimating missingness probabilities under a low nuclear norm assumption. Adv. Neural Inf. Process. Syst. 32.

Ma, S., Goldfarb, D., Chen, L., 2011. Fixed point and bregman iterative methods for matrix rank minimization. Math. Program. 128 (1–2), 321–353.

Ma, C., Wang, K., Chi, Y., Chen, Y., 2019. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution. Found. Comput. Math. 20 (3), 451–632.

Mazumder, R., Hastie, T., Tibshirani, R., 2010. Spectral regularization algorithms for learning large incomplete matrices. J. Mach. Learn. Res. 11 (8), 2287–2322.

McCarty, N.M., 2000. Presidential pork: Executive veto power and distributive politics. Am. Polit. Sci. Rev. 94 (1), 117–129.

Negahban, S., Wainwright, M.J., 2012. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. J. Mach. Learn. Res. 13 (1), 1665–1697.

Parikh, N., Boyd, S., 2014. Proximal algorithms. Found. Trends Optim. 1 (3), 127–239.

Rubin, D.B., 1974. Estimating causal effects of treatments in randomized and nonrandomized studies.. J. Educ. Psychol. 66 (5), 688.

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

Xia, D., Yuan, M., 2021. Statistical inferences of linear forms for noisy matrix completion. J. R. Stat. Soc. Ser. B Stat. Methodol. 83 (1), 58–77.

Xiong, R., Pelger, M., 2023. Large dimensional latent factor modeling with missing observations and applications to causal inference. J. Econometrics 233 (1), 271–301.

Yan, Y., Chen, Y., Fan, J., 2021. Inference for heteroskedastic PCA with missing data. arXiv preprint arXiv:2107.12365.