INFERENCE FOR HETEROGENEOUS EFFECTS USING LOW-RANK ESTIMATION OF FACTOR SLOPES

VICTOR CHERNOZHUKOV, CHRISTIAN HANSEN, YUAN LIAO, AND YINCHU ZHU

Abstract. We study a panel data model with general heterogeneous effects where slopes are allowed to vary across both individuals and over time. The key dimension reduction assumption we employ is that the heterogeneous slopes can be expressed as having a factor structure so that the high-dimensional slope matrix is low-rank and can thus be estimated using low-rank regularized regression. We provide a multi-step estimation procedure for the heterogeneous effects. The procedure makes use of sample-splitting and partialing-out to accommodate inference following the use of penalized low-rank estimation. We formally verify that the resulting estimator is asymptotically normal allowing simple construction of inferential statements for the individual-time-specific effects and for cross-sectional averages of these effects. We illustrate the proposed method in simulation experiments.

Key words: nuclear norm penalization, singular value thresholding, sample splitting, interactive effects, post-selection inference

1. Introduction

This paper studies inference within the following panel data model:

$$y_{it} \quad = \quad x_{it}'\theta_{it} + \alpha_i' g_t + u_{it}, \quad i = 1, ..., N, \quad t = 1, ..., T, \tag{1.1}$$

where y_{it} is the scalar outcome of interest, x_{it} is a d-dimensional vector of observed covariates with heterogeneous slopes θ_{it}, α_i and g_t are unobserved fixed effects, and u_{it} is an unobserved error term. The model permits general heterogeneity in the sense that fixed effects appear in the model interactively and the slope θ_{it} is allowed to vary across both i and t. The main goal of this paper is to provide an asymptotically valid procedure for performing statistical inference for averages of the θ_{it} taken across subgroups in the population at specific time periods. The subgroups may consist of single individuals, in which case inference is for a specific θ_{it}; the entire cross-section; or any pre-specified subset of the cross-sectional units.

Date: First draft: December 25, 2018. This draft: June 30, 2020.

We are grateful to seminar participants at Iowa State, University of North Carolina - Chapel Hill, Duke, Connecticut, Syracuse, IAAE 2019, and University of Chile for helpful discussions.

The main dimension reduction assumption employed in this paper is that θ_{it} can be expressed as

$$\theta_{it} = \lambda_i' f_t.$$

That is, we assume the slopes θ_{it} can be represented by a factor structure where λ_i is a matrix of loadings and f_t is a vector of factors. We allow f_t and g_t to have overlapping components and allow λ_i and f_t to be constant across $i \leq N$ and $t \leq T$ respectively. Thus, the model accommodates homogeneous slopes as a special case. We note that θ_{it} is not subject to rotational indeterminacy and is well-identified and cleanly estimable.

It will be useful to represent the model in matrix form. Let Θ_r and X_r be the $N \times T$ matrices with r^{th} component $\theta_{it,r}$ and $x_{it,r}$ respectively. Let M, Y, and U be the $N \times T$ matrices of $\alpha_i' g_t$, y_{it}, and u_{it}. Finally, let $\odot$ denote the matrix element-wise product. Using this notation, the matrix form of (1.1) is

$$Y = \sum_{r=1}^{d} X_r \odot \Theta_r + M + U.$$

Under the maintained factor structure, the slope and fixed effect matrices, Θ_r and M, have rank at most equal to their associated numbers of factors. This structure motivates estimating the model parameters via low-rank estimation:

$$\begin{aligned} &\min_{\{\Theta_1,...,\Theta_d,M\}} \|Y - \sum_{r=1}^{d} X_r \odot \Theta_r - M\|_F^2 + P_0(\Theta_1, ..., \Theta_d, M) \\ &\text{where} \quad P_0(\Theta_1, ..., \Theta_d, M) = \sum_{r=1}^{d} \nu_r \|\Theta_r\|_n + \nu_0 \|M\|_n \end{aligned} \tag{1.2}$$

for some tuning parameters $\nu_0, \nu_1, ..., \nu_d > 0$ and $\|.\|_F$ and $\|.\|_n$ respectively denoting the matrix Frobenius norm and the nuclear norm. In particular, let $\psi_1(\Theta) \geq ... \geq \psi_{\min\{N,T\}}(\Theta)$ be the sorted singular values of an $N \times T$ matrix Θ, then

$$\|\Theta\|_n := \sum_{i=1}^{\min\{N,T\}} \psi_i(\Theta).$$

The low-rank estimators defined in (1.2) will be consistent with suitable choice of the tuning parameters. However, the use of regularization, which may result in large shrinkage bias in finite samples, complicates inference. This paper contributes to the literature on penalized low-rank estimation by providing an approach to obtaining valid inferential statements after applying singular value thresholding (SVT) type regularization. We use the solution of (1.2) as initial estimates and obtain their singular vectors as the preliminary estimates of λ_i and α_i. We then estimate the factors and loading iteratively via least squares. This procedure is integrated with partialing-out and sample-splitting to further alleviate the effect of regularized estimation.

The main contribution of our paper is in providing a method and accompanying theory for obtaining valid inference about heterogeneous effects within the factor-slope structure. By incorporating general covariates and allowing slopes to vary both with i and t, our model is different from those in the low-rank estimation and interactive fixed effects literature. The factor-slope structure captures a rich spectrum of heterogeneous effects and enables hypothesis tests that are important for economists and policy makers to understand effects of policies over the sample period. Leveraging the factor-slope structure also allows us to transform many fundamental high-dimensional inferential problems into more tractable low-dimensional problems. For example, under mild conditions, the problem of testing the hypothesis of no effect at time t, $H_0 : \theta_{it} = 0$ for all i at a given t, is equivalent to testing $f_t = 0$ almost surely. We illustrate the use of our results for testing hypotheses of immediate interest after presenting the main results in Section 3.

Nuclear norm penalization is now a standard technique for estimating low-rank models; see, e.g., Negahban and Wainwright (2011); Recht et al. (2010); Sun and Zhang (2012); Candès and Tao (2010); Koltchinskii et al. (2011). In recent work, Bai and Ng (2019) considered iterative ridge regressions with rank constraints in the setting of pure factor models. Athey et al. (2018) consider the use of matrix completion methods to impute missing potential outcomes for estimating causal effects with a binary treatment variable in a setting with general treatment effect heterogeneity, and provide rates of convergence for the estimated low-rank matrix. We contribute to this literature by considering non-binary covariates with factor slopes and providing distributional results for individual-time specific slope coefficients and their cross-sectional averages.

Our paper is also related to the extensive panel data literature with interactive fixed effects, e.g., Bai (2009), Moon and Weidner (2015), Su et al. (2015), and Ahn et al. (2013). In particular, Moon and Weidner (2018) use nuclear-norm regularization for estimating interactive fixed effects in a model with homogeneous slopes. Studying slope heterogeneity has also been an important topic for panel data. For example, Chamberlain and Hirano (1999) and Pesaran (2006) consider models with coefficients that are not time-varying but are heterogeneous across individuals. Bonhomme and Manresa (2015) and Su et al. (2016) study settings where coefficients are assumed to be homogeneous within latent groups and estimation proceeds by simultaneously estimating group-specific effects and the group membership of each observation. There are also other approaches that allow for heterogeneity in both i and t in the literature. For example, Feng et al. (2017) assume that slopes are functions of observed time varying categorical variables, and Su and Wang (2017) considers a model in which slope coefficients vary smoothly over time so are locally time invariant. One could also adopt hierarchical Bayesian inference for the heterogeneous coefficients by assuming they are drawn from hierarchical priors, e.g., Hsiao et al. (1999).

Relative to these approaches, we use a different dimension reduction strategy, which allows general variation in coefficients across individuals and over time without relying on additional structure over the factors or factor loadings. Within this structure, we are able to provide inference for a variety of effects of interest. Finally, we impose a strong factor assumption so that the ranks - number of factors - can be consistently estimated from estimated singular values of the low rank matrices. Thus, our paper is also connected with the literature on estimating the number of factors, e.g., Bai and Ng (2002), Onatski (2010), and Ahn and Horenstein (2013).

The rest of the paper is organized as follows. Section 2 introduces the post-SVT algorithms that define our estimators. Section 3 provides asymptotic theory. Section 4 presents simulation results. Proofs are given in a supplementary appendix.

2. The Model

We consider the model

$$\begin{aligned} &y_{it} = \theta_{it} x_{it} + \alpha_i' g_t + u_{it}, \quad i = 1, ..., N, \quad t = 1, ..., T, \text{ with} \\ &\theta_{it} = \lambda_i' f_t. \end{aligned} \tag{2.1}$$

We observe (y_{it}, x_{it}), and the goal is to make inference about θ_{it} or averages of the θ_{it} taken across groups formed from cross-sectional units. Here $\alpha_i' g_t$ are interactive fixed effects as in Bai (2009). For ease of presentation, we focus on the case where

x_{it} is univariate, that is, $\dim(x_{it}) = 1$. The extension to the multivariate case is straightforward and is provided in Appendix A in the supplementary material.

We assume that $\{\lambda_i, \alpha_i\}$ are deterministic sequences, while $\{f_t, g_t\}$ are random. We allow arbitrary dependence and similarity among $\{f_t : t \leq T\}$ and impose nearly no restrictions on the sequence for λ_i and f_t. Allowing arbitrary dependence and imposing weak structure is important for accommodating leading special cases such as homogeneous models where $\theta_{it} = \theta$ for all (i, t) which can be obtained by setting $\lambda_i = \lambda$ and $f_t = f$ for all (i, t).

Throughout, we assume the ranks $\dim(\lambda_i) = K_1$ and $\dim(\alpha_i) = K_2$ are fixed. We note that this differs from much of the matrix completion literature which explicitly considers model sequences which allow ranks to increase. We leave the extension to the increasing rank case to future work. We also initially assume that both K_1 and K_2 are known. In Section 3.5, we discuss consistent rank estimation and note that the use of consistently estimated ranks does not impact the asymptotic distribution of the heterogeneous slope estimator under our assumptions.

2.1. **Nuclear Norm Penalized Estimation.**

Let (Y, X, U) be $N \times T$ matrices of (y_{it}, x_{it}, u_{it}). Then (2.1) may be expressed in matrix form as

$$Y \;=\; M + X \odot \Theta + U$$

where Θ and M are matrices of $(\theta_{it}, \alpha_i' g_t)$. Motivated by their low-rank structures, we start with the following penalized nuclear-norm optimization problem:

$$\begin{aligned} (\widetilde{\Theta}, \widetilde{M}) &= \arg\min_{\Theta, M} F(\Theta, M), \\ F(\Theta, M) &:= \|Y - M - X \odot \Theta\|_F^2 + \nu_0 \|M\|_n + \nu_1 \|\Theta\|_n \end{aligned} \tag{2.2}$$

for some tuning parameters $\nu_0, \nu_1 > 0$.

For a fixed matrix Y, let $U_Y D_Y V_Y' = Y$ be its singular value decomposition. Define the singular value thresholding (SVT) operator

$$S_\lambda(Y) = U_Y D_\lambda V_Y',$$

where D_λ is defined by replacing the diagonal entry D_{ii} of D by $\max\{D_{ii} - \lambda, 0\}$. That is, $S_\lambda(Y)$ applies soft-thresholding on the singular values of Y.

The solution of (2.2) can be obtained by iteratively using SVT estimation. Given Θ, solving for M in (2.2) leads to the solution

$$S_{\nu_0/2}(Z_\Theta) = \arg\min_M \|Z_\Theta - M\|_F^2 + \nu_0\|M\|_n$$

for $Z_\Theta = Y - X \odot \Theta$. Similarly, given M, let $Z_M = Y - M$. We solve for Θ via

$$\Theta_M := \arg\min_\Theta \|Z_M - X \odot \Theta\|_F^2 + \nu_1\|\Theta\|_n,$$

which satisfies the following KKT condition (Ma et al., 2011): For any $\tau > 0$,

$$\Theta_M = S_{\tau\nu_1/2}(\Theta_M - \tau X \odot (X \odot \Theta_M - Z_M)).$$

As such, we employ the following algorithm to iteratively solve for $\widetilde{M}$ and $\widetilde{\Theta}$ as the global solution to (2.2).

Algorithm 2.1. Compute the nuclear-norm penalized regression as follows:

Step 1: Fix the "step size" $\tau \in (0, 1/\max_{it} x_{it}^2)$. Initialize Θ_0, M_0 and set $k = 0$.

Step 2: Let

$$\begin{aligned} \Theta_{k+1} &= S_{\tau\nu_1/2}(\Theta_k - \tau X \odot (X \odot \Theta_k - Y + M_k)), \\ M_{k+1} &= S_{\nu_0/2}(Y - X \odot \Theta_{k+1}). \end{aligned}$$

Set k to $k + 1$.

Step 3: Repeat step 2 until convergence.

The following proposition verifies that the evaluated objective function $F(\Theta_{k+1}, M_{k+1})$ is monotonically decreasing and converges to the global minimum at the rate $O(k^{-1})$. In practice, we set $\tau = (1 - \epsilon)/\max_{it} x_{it}^2$ for $\epsilon = 0.01$.

Proposition 2.1. *Let $(\widetilde{\Theta}, \widetilde{M})$ be a global minimum for $F(\Theta, M)$. Then for any $\tau \in (0, 1/\max_{it} x_{it}^2)$, and any initial Θ_0, M_0, we have*

$$F(\Theta_{k+1}, M_{k+1}) \leq F(\Theta_{k+1}, M_k) \leq F(\Theta_k, M_k),$$

for each $k \geq 0$. In addition, for all $k \geq 1$,

$$F(\Theta_{k+1}, M_{k+1}) - F(\widetilde{\Theta}, \widetilde{M}) \leq \frac{1}{k\tau}\|\Theta_1 - \widetilde{\Theta}\|_F^2. \tag{2.3}$$

2.2. **Intuition for Main Algorithm.** Nuclear-norm penalized estimators are generally inappropriate for inference because they may suffer from substantial shrinkage bias. We consider a debiased estimator for inference. In this section, we provide a heuristic argument that highlights the chief complications that arise in obtaining reliable asymptotic approximations in our setting. We provide intuition for how we use partialing-out, sample-splitting, and iterative OLS estimation to deal with these complications. Part of the argument deals with the rotation of the factors and to the best of our knowledge has not appeared previously in the post-regularization inference literature.

To understand the key issues, we work in a simplified setting where $\dim(f_t) = \dim(g_t) = 1$. Suppose we have preliminary estimates $\widetilde{\lambda}_i$ and $\widetilde{\alpha}_i$ obtained by extracting the first singular vector from $\widetilde{\Theta}$ and $\widetilde{M}$ obtained from (2.2). One might then attempt to estimate f_t as the coefficient on $\widetilde{\lambda}_i x_{it}$ from the regression of y_{it} on $\widetilde{\lambda}_i x_{it}$ and $\widetilde{\alpha}_i$. Letting $\check{f}_t$ denote this coefficient, $\check{f}_t$ would have the following expansion: For some $\widehat{Q}$ and rotation matrices H_1 and H_2,

$$\begin{aligned}\sqrt{N}(\check{f}_t - H_1^{-1} f_t) = \widehat{Q}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\lambda_i x_{it} u_{it} + \widehat{Q}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\lambda_i x_{it}(\widetilde{\alpha}_i - H_2\alpha_i)g_t \\ + \widehat{Q}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\lambda_i x_{it}^2(\widetilde{\lambda}_i - H_1\lambda_i) f_t + o_P(1).\end{aligned} \tag{2.4}$$

The usual asymptotic behavior will be driven by the term $\widehat{Q}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\lambda_i x_{it} u_{it}$, and we would like that the remaining terms are asymptotically negligible. Unfortunately, $\widetilde{\Delta}_\alpha = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\lambda_i x_{it}(\widetilde{\alpha}_i - H_2\alpha_i)$ and $\widetilde{\Delta}_\lambda = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\lambda_i x_{it}^2(\widetilde{\lambda}_i - H_1\lambda_i)$, which respectively capture the effect of estimation error in $\widetilde{\alpha}$ and $\widetilde{\lambda}$, will generally not vanish asymptotically. Focusing on the term $\widetilde{\Delta}_\alpha$, note that regularization bias will mean that $\widetilde{\alpha}_i - H_2\alpha_i$ is biased in general, resulting in the object inside the sum generally not being mean zero. Further, while consistent, $\widetilde{\alpha}_i$ will generally not converge fast enough to ensure that $\widetilde{\Delta}_\alpha = o_P(1)$. Thus, in general, $\widetilde{\Delta}_\alpha$ will result in a failure of asymptotic normality.

2.2.1. *Partialing-Out and Sample Splitting.* We deal with terms like $\widetilde{\Delta}_\alpha$ by leveraging additional structure on the observed variable x_{it} and making use of partialing-out and sample splitting. Write

$$x_{it} = \mu_{it} + e_{it}, \tag{2.5}$$

where μ_{it} is the mean process of x_{it} which is assumed to capture both time series dependence and strong sources of cross-sectional correlation. Specifically, we will maintain the assumption that e_{it} is a zero-mean process that is serially independent and cross-sectionally weakly dependent throughout our formal analysis. Using this decomposition of x_{it}, we can produce a "partialed-out" version of the model:

$$\dot{y}_{it} = \alpha_i' g_t + e_{it}\lambda_i' f_t + u_{it}, \qquad \text{where } \dot{y}_{it} = y_{it} - \mu_{it}\theta_{it}. \tag{2.6}$$

The transformed model now contains partialed-out regressors e_{it}. Using these, we can estimate f_t by regressing the estimated $\dot{y}_{it}$ onto $(\widetilde{\alpha}_i, \widetilde{\lambda}_i e_{it})$ which would lead to an expansion of estimated f_t as

$$\begin{aligned}
\sqrt{N}(\widehat{f}_t - H_1^{-1} f_t) &= \widehat{Q}\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \lambda_i e_{it} u_{it} + \Delta_\alpha g_t + \Delta_\lambda f_t + o_P(1), \\
\text{where} \qquad \Delta_\alpha &:= \widehat{Q}\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \lambda_i e_{it}(\widetilde{\alpha}_i - H_2\alpha_i)', \\
\Delta_\lambda &:= \widehat{Q}\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \lambda_i (\mathsf{E} e_{it}^2)(\widetilde{\lambda}_i - H_1\lambda_i)'.
\end{aligned} \tag{2.7}$$

It is immediate that $\Delta_\alpha = o_P(1)$ if e_{it} is mean zero, independent of $\lambda_i(\widetilde{\alpha}_i - H_2\alpha_i)$, and sufficient moments exist. The plausibility of these conditions relies on partialing-out the strongly dependent components in x_{it} and is tightly tied to the so-called *Neyman's orthogonality* that has been shown to be important in obtaining valid inference after high-dimensional estimation in a variety of contexts; see, e.g., Chernozhukov et al. (2018).

Operationalizing partialing-out in (2.6) requires a sufficiently high-quality estimate of e_{it}, which requires restricting the process μ_{it}. In our formal development, we assume $\mu_{it} = l_i' w_t$. Hence, x_{it} follows a factor model, where (l_i, w_t) respectively represent loadings and factors, and may thus be strongly intertemporally and cross-sectionally correlated. We allow w_t to overlap with (f_t, g_t) and take $\dim(w_t)$ to be fixed in our analysis. Other structures of μ_{it} could also be imposed with all results going through as long as μ_{it} is sufficiently well-estimable.

To complete the argument that $\Delta_\alpha = o_P(1)$, we also use sample-splitting. For a fixed t, let $I \subset \{1, ..., T\}\backslash t$ be a set of time indexes, and let

$$D_I = \{(y_{is}, x_{is}) : i \leq N, s \in I\}.$$

Rather than using the full sample to obtain initial estimates of $\widetilde{\lambda}_i$ and $\widetilde{\alpha}_i$, we run the nuclear-norm optimization using only data D_I. Assuming that e_{it} is serially independent and $t \notin I$, $\widetilde{\lambda}_i$ and $\widetilde{\alpha}_i$ are independent of e_{it} which allows us to easily verify that Δ_α and similar terms vanish asymptotically.

2.2.2. *A Final OLS Step: The effect of* $\widetilde{\lambda}_i - H_1\lambda_i$. We now consider the term Δ_λ in (2.7) which arises due to estimation error in $\widetilde{\lambda}_i$. The term Δ_λ results in issues that are analogous to the usual rotational indeterminacy issues in factor models but do not appear in the existing post-regularization inference literature.

Importantly, Δ_λ is *not* $o_P(1)$ since $\mathsf{E}e_{it}^2 \neq 0$. Returning to (2.7) and using that $\Delta_\alpha = o_P(1)$, we have

$$\sqrt{N}(\widehat{f}_t - H_1^{-1}f_t) = \widehat{Q}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\lambda_i e_{it}u_{it} + \Delta_\lambda f_t + o_P(1).$$

Thus, the use of the regularized estimator $\widetilde{\lambda}_i$ results in a non-vanishing asymptotic bias. Importantly, this bias manifests as an additional time-invariant rotation of the factors f_t. We can thus define $H_f := H_1^{-1} + \Delta_\lambda N^{-1/2}$ and establish that

$$\sqrt{N}(\widehat{f}_t - H_f f_t) = \widehat{Q}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\lambda_i e_{it}u_{it} + o_P(1).$$

Therefore, the effect of first-step estimation error $\widetilde{\lambda}_i - H_1'\lambda_i$ is "absorbed" by the adjusted rotation matrix. Once $\widehat{f}_t$ recovers the span of f_t, our formal algorithm will make use of a final least squares iteration to produce an estimator $\widehat{\lambda}_i$ that suitably recovers the appropriately rotated λ_i. Recovering these two compatibly rotated versions of f_t and λ_i is then sufficient for the inferential theory for $\widehat{\theta}_{it}$.

2.3. **Formal Estimation Algorithm.**

We now state the full estimation algorithm for θ_{it} for some fixed t. The algorithm is stated in the leading case where $x_{it} = l_i'w_t + e_{it}$ with $\mathsf{E}e_{it} = 0$. We then partial out the common component by subtracting the estimated μ_{it} from x_{it} and working with model (2.6).

Algorithm 2.2. Estimate θ_{it} as follows.

Step 1. Estimate the structure $x_{it} = \mu_{it} + e_{it}$. Use the principal components (PC) estimator to obtain $(\widehat{\mu}_{it}, \widehat{e}_{it})$ for all $i = 1, ..., N$, $t = 1, .., T$.

Step 2: Sample splitting. Randomly split the sample into $\{1, ..., T\}/\{t\} = I \cup I^c$, so that $|I|_0 = [(T-1)/2]$. Denote the $N \times |I|_0$ matrices of (y_{is}, x_{is}) for observations

$s \in I$ by Y_I, X_I. Run nuclear-norm penalized regression:

$$(\widetilde{M}_I, \widetilde{\Theta}_I) := \arg\min_{M,\Theta} \|Y_I - M - X_I \odot \Theta\|_F^2 + \nu_0 \|M\|_n + \nu_1 \|\Theta\|_n. \tag{2.8}$$

Let $\widetilde{\Lambda}_I = (\widetilde{\lambda}_1, ..., \widetilde{\lambda}_N)'$ be the $N \times K_1$ matrix whose columns are defined as $\sqrt{N}$ times the first K_1 eigenvectors of $\widetilde{\Theta}_I \widetilde{\Theta}_I'$. Let $\widetilde{A}_I = (\widetilde{\alpha}_1, ..., \widetilde{\alpha}_N)'$ be the $N \times K_2$ matrix whose columns are defined as $\sqrt{N}$ times the first K_2 eigenvectors of $\widetilde{M}_I \widetilde{M}_I'$.

Step 3. Estimate components for "partialing-out." Using $\widetilde{A}_I$ and $\widetilde{\Lambda}_I$, obtain

$$(\widetilde{f}_s, \widetilde{g}_s) := \arg\min_{f_s, g_s} \sum_{i=1}^N (y_{is} - \widetilde{\alpha}_i' g_s - x_{is} \widetilde{\lambda}_i' f_s)^2, \quad s \in I^c \cup \{t\}.$$

Update estimates of loadings as

$$(\dot{\lambda}_i, \dot{\alpha}_i) = \arg\min_{\lambda_i, \alpha_i} \sum_{s \in I^c \cup \{t\}} (y_{is} - \alpha_i' \widetilde{g}_s - x_{is} \lambda_i' \widetilde{f}_s)^2, \quad i = 1, ..., N.$$

Step 4. Estimate (f_t, λ_i) for use in inference about θ_{it}. Define $\widehat{y}_{is} = y_{is} - \widehat{\mu}_{is} \dot{\lambda}_i' \widetilde{f}_s$ and $\widehat{e}_{is} = x_{is} - \widehat{\mu}_{is}$. Let

$$\begin{aligned}
(\widehat{f}_{I,s}, \widehat{g}_{I,s}) &:= \arg\min_{f_s, g_s} \sum_{i=1}^N (\widehat{y}_{is} - \widetilde{\alpha}_i' g_s - \widehat{e}_{is} \widetilde{\lambda}_i' f_s)^2, \quad s \in I^c \cup \{t\} \\
(\widehat{\lambda}_{I,i}, \widehat{\alpha}_{I,i}) &:= \arg\min_{\lambda_i, \alpha_i} \sum_{s \in I^c \cup \{t\}} (\widehat{y}_{is} - \alpha_i' \widehat{g}_{I,s} - \widehat{e}_{is} \lambda_i' \widehat{f}_{I,s})^2, \quad i = 1, ..., N.
\end{aligned}$$

Step 5. Exchange I and I^c. Repeat steps 2-4 with I and I^c exchanged to obtain $(\widehat{\lambda}_{I^c,i}, \widehat{f}_{I^c,s} : s \in I \cup \{t\}, i \leq N)$.

Step 6. Estimate θ_{it}. Obtain the estimator of θ_{it}:

$$\widehat{\theta}_{it} := \frac{1}{2}[\widehat{\lambda}_{I,i}' \widehat{f}_{I,t} + \widehat{\lambda}_{I^c,i}' \widehat{f}_{I^c,t}].$$

Remark 2.1. We split the sample to $\{1, ..., T\} = I \cup I^c \cup \{t\}$ which ensures that e_{it} is independent of the data in both I and I^c for the given t of interest. Alternatively, we may split over individuals, which would allow for serial dependence in the e_{it} as long as they may be taken to be cross-sectionally independent.

Remark 2.2. Step 3 is needed to obtain a sufficiently high-quality estimate for $\dot{y}_{it} = y_{it} - \mu_{it} \lambda_i' f_t$ to allow application of the partialed-out equation (2.6). The estimators in Step 3 are still unsuitable for inference due to the use of the raw x_{it}, which is generally not zero-mean or serially independent. Appropriately centering and eliminating sources of serial dependence in x_{it} gives rise to Step 4.

2.4. **Choosing the tuning parameters.** We adopt a simple plug-in approach to choosing the tuning parameters ν_1 and ν_0. The "scores" of the penalized regression are given by $2U$ and $2X \odot U$. We then wish to choose tuning parameters (ν_0, ν_1) such that we achieve score domination in the sense that

$$2\|U\| < (1-c)\nu_0, \quad 2\|X \odot U\| < (1-c)\nu_1 \tag{2.9}$$

for some $c > 0$ with high probability where $\|.\|$ denotes the matrix operator norm.

As in the literature surrounding ℓ_1-penalized estimation of the high-dimensional linear model, achieving score domination will result in desirable rates of convergence and will produce consistent estimators of the ranks of the matrix parameters. To operationalize these choices, we assume that the columns of U and $X \odot U$, respectively $\{u_t\}$ and $\{x_t \odot u_t\}$, are sub-Gaussian vectors. Then, the eigenvalue-concentration inequality for sub-Gaussian random vectors (Theorem 5.39 of Vershynin (2010)) implies

$$\nu_0 \asymp \nu_1 \asymp \max\{\sqrt{N}, \sqrt{T}\}.$$

In the Gaussian case, further progress can be made. Suppose u_{it} is independent in both (i,t) and $u_{it} \sim \mathcal{N}(0, \sigma_{ui}^2)$. Let Z be an $N \times T$ matrix whose elements are independently generated from $\mathcal{N}(0, \sigma_{ui}^2)$. Then $\|X \odot U\| =^d \|X \odot Z\|$ and $\|U\| =^d \|Z\|$ where $=^d$ means "is identically distributed to". Let $Q(W; m)$ denote the m^{th} quantile of random variable W. For $\delta_{NT} = o(1)$, take

$$\nu_0 = 2(1+c_1)Q(\|Z\|; 1-\delta_{NT}), \quad \nu_1 = 2(1+c_1)Q(\|X \odot Z\|; 1-\delta_{NT})$$

which respectively denote $2(1+c_1)$ multiplied by the $1-\delta_{NT}$ quantile of $\|Z\|$ and $\|X \odot Z\|$. Then (2.9) holds with probability $1-\delta_{NT}$. In practice, we compute the quantiles by simulation replacing σ_{ui}^2 with an initial consistent estimator. In our simulation and empirical examples, we set $c_1 = 0.1$ and $\delta_{NT} = 0.05$.

3. Asymptotic Results

3.1. **Parameters of interest relevant to policy studies.** Our main inferential theory establishes the rate of convergence and asymptotic normality for a *group* *average effect.* Fix a cross-sectional subgroup

$$\mathcal{G} \subseteq \{1, 2, ..., N\}.$$

We are interested in inference for the group average effect at a fixed $t \leq T$:

$$\bar{\theta}_{\mathcal{G},t} := \frac{1}{|\mathcal{G}|_0} \sum_{i \in \mathcal{G}} \theta_{it},$$

where $|\mathcal{G}|_0$ denotes the group size. The group size can be either fixed or grow with N. This structure admits two interesting cases as extremes: (i) $\mathcal{G} = \{i\}$ for any fixed i, which allows inference for a fixed individual; and (ii) $\mathcal{G} = \{1, 2, ..., N\}$, which allows inference for the cross-sectional average effect $\bar{\theta}_t := \frac{1}{N}\sum_{i=1}^N \theta_{it}$.

The group average effect provides answers to various questions related to policy studies, e.g., the effect of the minimum wage in a state of interest or on average in the country at different points in time. Inference about $\bar{\theta}_{\mathcal{G},t}$ is also relevant to answering many questions which are important for understanding effects of policies over the sample period. For example, one may be interested in the following hypotheses related to policy effects during the sample period:

- **Tests of group homogeneity.** Given a finite number of groups of interest $\mathcal{G}_1, ..., \mathcal{G}_J$, one may wish to test the hypothesis of homogeneous average group effects:
$$H_0^1 : \bar{\theta}_{\mathcal{G}_1,t} = ... = \bar{\theta}_{\mathcal{G}_J,t}.$$
For instance, we may be interested in asking whether the average effect of the minimum wage in states in two different regions of the country, perhaps historically poorer and historically wealthier states, are the same.
- **Test of joint significance in a given time period.** One might be interested in testing that all effects in a given time period are zero:
$$H_0^2 : \theta_{it} = 0 \text{ for all } i.$$

We show how valid tests of these hypotheses result as simple extensions of our main results in Section 3.4. In the supplementary appendix, we also show how our results can be extended to testing homogeneity of effects over two time periods.

3.2. Assumptions.

We first introduce a key assumption about the nuclear-norm SVT procedure. We require some "invertibility" condition for the operator:

$$(\Delta_1, \Delta_2) :\rightarrow \Delta_1 + \Delta_2 \odot X,$$

when (Δ_1, Δ_2) is confined to a *restricted low-rank set* (e.g., Negahban and Wainwright (2011)). To describe this set, we first introduce some notation.

Let $U_1 D_1 V_1' = \Theta$ and $U_2 D_2 V_2' = M$ respectively be the singular value decompositions of the low-rank matrices Θ and M. Further decompose,

$$U_j = (U_{j,r}, U_{j,c}), \quad V_j = (V_{j,r}, V_{j,c}), \quad \text{for } j = 1, 2.$$

Here $(U_{j,r}, V_{j,r})$ are the singular vectors corresponding to nonzero singular values, while $(U_{j,c}, V_{j,c})$ are singular vectors corresponding to the zero singular values. In addition, for any $N \times T$ matrix Δ, let

$$\mathcal{P}_j(\Delta) = U_{j,c} U_{j,c}' \Delta V_{j,c} V_{j,c}', \text{ and } \mathcal{M}_j(\Delta) = \Delta - \mathcal{P}_j(\Delta).$$

Here $U_{j,c} U_{j,c}'$ and $V_{j,c} V_{j,c}'$ respectively are the projection matrices onto the columns of $U_{j,c}$ and $V_{j,c}$. Therefore, $\mathcal{M}_1(\cdot)$ and $\mathcal{M}_2(\cdot)$ can be considered as the projection matrices onto the "low-rank spaces" of Θ and M respectively, and $\mathcal{P}_1(\cdot)$ and $\mathcal{P}_2(\cdot)$ are projections onto their orthogonal spaces.

Define the *restricted low-rank set* as, for some $c_1, c_2 > 0$,

$$\mathcal{C}(c_1, c_2) = \Big\{ (\Delta_1, \Delta_2) : \|\mathcal{P}_1(\Delta_1)\|_n + \|\mathcal{P}_2(\Delta_2)\|_n \le c_1 \|\mathcal{M}_1(\Delta_1)\|_n + c_1 \|\mathcal{M}_2(\Delta_2)\|_n,$$

$$\|\Delta_1\|_F^2 + \|\Delta_2\|_F^2 \ge c_2 \sqrt{NT} \Big\}.$$

Assumption 3.1 (Restricted strong convexity)**.** *For any $c_1 > 0$, there are constants $c_2, \kappa, \eta > 0$, uniformly for $(\Delta_1, \Delta_2) \in \mathcal{C}(c_1, c_2)$, such that*

$$\|\Delta_1 + \Delta_2 \odot X\|_F^2 \ge \kappa \|\Delta_1\|_F^2 + \kappa \|\Delta_2\|_F^2 - (N + T)\eta \tag{3.1}$$

with probability approaching one. The same condition holds when (M, Θ) are replaced with $\Theta_I = (\lambda_i' f_t : i \le N, t \in I)$, $M_I = (\alpha_i' g_t : i \le N, t \in I)$, and with (Θ_{I^c}, M_{I^c}), which are defined similarly.

Remark 3.1. Restricted strong convexity has been well studied in the low-rank estimation literature in the case with a single matrix parameter, e.g., Klopp (2014) and Negahban and Wainwright (2012). Assumption 3.1 extends the concept to the multivariate case with general regressors. We verify this assumption in the supplementary material under primitive conditions. The key primitive condition requires that there is sufficient variability among the regressors x_{it}, which is analogous to the usual rank condition employed in regression. Specifically, for the decomposition $x_{it} = \mu_{it} + e_{it}$ for μ_{it} and e_{it} defined as in (2.5), a sufficient condition is that all eigenvalues of Σ_{it} are bounded away from zero almost surely, where

$$\Sigma_{it} = \begin{pmatrix} 1 & \mathsf{E}_\mu x_{it} \\ \mathsf{E}_\mu x_{it} & \mathsf{E}_\mu x_{it}^2 \end{pmatrix} = \begin{pmatrix} 1 & \mu_{it} \\ \mu_{it} & \mu_{it}^2 + \mathsf{E}_\mu e_{it}^2 \end{pmatrix}$$

and $\mathsf{E}_\mu(\cdot)$ denotes the conditional expectation given $\{\mu_{it} : i \leq N, t \leq T\}$. Satisfaction of this condition allows separately identifying θ_{it} and the interactive fixed effects in M. This condition is straightforward to verify. Note that the minimum eigenvalue of Σ_{it} equals

$$\psi_{\min}(\Sigma_{it}) := \frac{2\mathsf{E}_\mu e_{it}^2}{\gamma_{it} + \sqrt{\gamma_{it}^2 - 4\mathsf{E}_\mu e_{it}^2}} \geq \frac{\mathsf{E}_\mu e_{it}^2}{\gamma_{it}}$$

where $\gamma_{it} = 1 + \mu_{it}^2 + \mathsf{E}_\mu e_{it}^2$. Thus, $\psi_{\min}(\Sigma_{it})$ is bounded away from zero uniformly for all (i,t) as long as $\mathsf{E}_\mu e_{it}^2$ is bounded away from zero and μ_{it}^2 is bounded away from infinity.

The following assumption places mild restrictions on the latent factors. The assumed conditions hold for serially independent sequences and also allow for perfectly dependent sequences where $(f_t, g_t) = (f, g)$ for some time-invariant (f, g) by setting $\dim(f_t) = \dim(g_t) = 1$.

Assumption 3.2. *As $T \to \infty$, the sub-samples (I, I^c) satisfy:*

$$\frac{1}{|I|_0}\sum_{t\in I} f_t f_t' = \frac{1}{T}\sum_{t=1}^T f_t f_t' + O_P(T^{-1/2}) = \frac{1}{|I^c|_0}\sum_{t\in I^c} f_t f_t',$$
$$\frac{1}{|I|_0}\sum_{t\in I} g_t g_t' = \frac{1}{T}\sum_{t=1}^T g_t g_t' + O_P(T^{-1/2}) = \frac{1}{|I^c|_0}\sum_{t\in I^c} g_t g_t'.$$

In addition, there is a $c > 0$ such that all the eigenvalues of $\frac{1}{T}\sum_{t=1}^T f_t f_t'$ and $\frac{1}{T}\sum_{t=1}^T g_t g_t'$ are bounded from below by c almost surely.

The next assumption requires that the factors be strong. In addition, we require distinct eigenvalues in order to identify their corresponding eigenvectors, and therefore, (λ_i, α_i). These conditions are strong, though standard in the factor modeling literature. It may be interesting, but is beyond the scope of the present work, to consider estimation and inference in the presence of weak factors.

Assumption 3.3 (Valid factor structures with strong factors). *There are constants $c_1 > ... > c_{K_1} > c > 0$, and $c'_1 > ... > c'_{K_2} > c > 0$, so that up to a term $o_P(1)$,*

(i) c'_j equals the j^{th} largest eigenvalue of $(\frac{1}{T}\sum_t g_t g_t')^{1/2}\frac{1}{N}\sum_{i=1}^N \alpha_i\alpha_i'(\frac{1}{T}\sum_t g_t g_t')^{1/2}$ for all $j = 1, ..., K_1$, and

(ii) c_j equals the j^{th} largest eigenvalue of $(\frac{1}{T}\sum_t f_t f_t')^{1/2}\frac{1}{N}\sum_{i=1}^N \lambda_i\lambda_i'(\frac{1}{T}\sum_t f_t f_t')^{1/2}$ for all $j = 1, ..., K_2$.

Proposition D.1 in the supplementary appendix shows that under the aforementioned conditions, the nuclear-norm regularized matrix estimators are consistent under the Frobenius norm and provides the rate of convergence. It extends known results from the low-rank estimation literature to the multi-dimensional case.

To obtain the asymptotic distribution of our estimator, we make a number of additional assumptions. Throughout, let (F, G, W, E, U) be the $T \times K$ matrices of $(f_t, g_t, w_t, e_{it}, u_{it})$, where K differs for different quantities.

Assumption 3.4 (Dependence). *(i) $\{e_{it}, u_{it}\}$ are independent across t; $\{e_{it}\}$ are also conditionally independent across t given $\{F, G, W, U\}$; $\{u_{it}\}$ are also conditionally independent across t given $\{F, G, W, E\}$.*

(ii) $\mathsf{E}(e_{it}|u_{it}, w_t, g_t, f_t) = 0, \quad \mathsf{E}(u_{it}|e_{it}, w_t, g_t, f_t) = 0$. Also, $\mathsf{E}e_{it}^2$ does not vary across t.

(iii) The $N \times T$ matrix $X \odot U$ has the following decomposition:

$$X \odot U = \Omega_{NT}\Sigma_T^{1/2}, \qquad \textit{where}$$

1. *$\Omega_{NT} := (\omega_1, ..., \omega_T)$ is an $N \times T$ matrix whose columns $\{\omega_t\}_{t\le T}$ are independent sub-gaussian random vectors with $\mathsf{E}\omega_t = 0$. More specifically, there is $C > 0$ such that*

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

2. *Σ_T is a $T \times T$ deterministic matrix whose eigenvalues are bounded from both below and above by constants.*

(iv) Weak conditional cross-sectional dependence: let $\mathcal{W} = (F, G, W)$. Let $\omega_{it} = u_{it}e_{it}$, and let c_i be a bounded nonrandom sequence. Almost surely,

$$\max_{t\le T}\frac{1}{N^3}\sum_{i,j,k,l\le N}|\mathsf{Cov}(e_{it}e_{jt}, e_{kt}e_{lt}|\mathcal{W}, U)| < C, \quad \max_{t\le T}\|\mathsf{E}(u_t u_t'|\mathcal{W}, E)\| < C$$

$$\max_{t\leq T}\mathsf{E}(|\frac{1}{\sqrt{N}}\sum_{i=1}^N c_i\omega_{it}|^4|\mathcal{W})<C,\quad \max_{t\leq T}\mathsf{E}(|\frac{1}{\sqrt{N}}\sum_{i=1}^N c_ie_{it}|^4|\mathcal{W},U)<C$$
$$\max_{t\leq T}\mathsf{E}(|\frac{1}{\sqrt{N}}\sum_{i=1}^N c_iu_{it}|^4|\mathcal{W},E)<C,\quad \max_{t\leq T}\max_{i\leq N}\frac{1}{N}\sum_{k,j\leq N}|\mathsf{E}(e_{kt}e_{it}e_{jt}|\mathcal{W},U)|<C$$
$$\max_{t\leq T}\max_{i\leq N}\sum_{j=1}^N|\mathsf{Cov}(e_{it}^m,e_{jt}^r|\mathcal{W},U)|<C,\quad m,r\in\{1,2\}$$
$$\max_{t\leq T}\max_{i\leq N}\frac{1}{N}\sum_{k,j\leq N}|\mathsf{Cov}(\omega_{it}\omega_{jt},\omega_{it}\omega_{kt}|\mathcal{W})|<C,\quad \max_{t\leq T}\max_{i\leq N}\sum_{j=1}^N|\mathsf{Cov}(\omega_{it},\omega_{jt})|\mathcal{W})|<C.$$

Assumption 3.5 (Cross-sectional CLT)**.** *As $N\to\infty$,*

$$V_{\lambda2}^{-1/2}\frac{1}{\sqrt{N}}\sum_{i=1}^N\lambda_ie_{it}u_{it}\to^d\mathcal{N}(0,I),\quad \textit{where } V_{\lambda2}=\mathsf{Var}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^N\lambda_ie_{it}u_{it}\middle|F\right).$$

Before stating our final assumption, it is useful to define a number of objects.

$$\begin{aligned}
b_{NT,1} &= \max_{t\leq T}\|\frac{1}{NT}\sum_{i=1}^N\sum_{s=1}^T w_s(e_{is}e_{it}-\mathsf{E}e_{is}e_{it})\|\\
b_{NT,2} &= (\max_{t\leq T}\frac{1}{T}\sum_{s=1}^T(\frac{1}{N}\sum_{i=1}^N e_{is}e_{it}-\mathsf{E}e_{is}e_{it})^2)^{1/2}\\
b_{NT,3} &= \max_{t\leq T}\|\frac{1}{N}\sum_{i=1}^N l_ie_{it}\|,\qquad b_{NT,4}=\max_{i\leq N}\|\frac{1}{T}\sum_{s=1}^T e_{is}w_s\|\\
b_{NT,5} &= \max_{i\leq N}\|\frac{1}{NT}\sum_{j=1}^N\sum_{s=1}^T l_j(e_{js}e_{is}-\mathsf{E}e_{js}e_{is})\|
\end{aligned}$$

In addition, we introduce Hessian matrices that are involved when iteratively estimating λ_i and f_t.

$$\begin{aligned}
D_{ft} &= \frac{1}{N}\Lambda'(\mathsf{diag}(X_t)M_\alpha\mathsf{diag}(X_t)\Lambda,\qquad D_{\lambda i}=\frac{1}{T}F'(\mathsf{diag}(\underline{X}_i)M_g\mathsf{diag}(\underline{X}_i))F,\\
\bar{D}_{ft} &= \frac{1}{N}\Lambda'\mathsf{E}((\mathsf{diag}(e_t)M_\alpha\mathsf{diag}(e_t))\Lambda+\frac{1}{N}\Lambda'(\mathsf{diag}(Lw_t)M_\alpha\mathsf{diag}(Lw_t)\Lambda,\\
\bar{D}_{\lambda i} &= \frac{1}{T}F'\mathsf{E}(\mathsf{diag}(E_i)M_g\mathsf{diag}(E_i))F+\frac{1}{T}F'(\mathsf{diag}(Wl_i)M_g\mathsf{diag}(Wl_i))F.
\end{aligned}$$

The above matrices involve the following notation. Let X_t, e_t denote the $N\times 1$ vector of x_{it} and e_{it}, and let $\underline{X_i}$ and E_i denote the $T\times 1$ vectors of x_{it} and e_{it} for a given i. Let L denote the $N\times\dim(w_t)$ matrix of l_i, and let W denote the

$T\times \dim(w_t)$ matrix of w_t. Let $M_g = I - G(G'G)^{-1}G'$ be a $T\times T$ projection matrix, and let M_α be an $N\times N$ matrix defined similarly. Finally, let $\mathsf{diag}(e_t)$ denote the diagonal matrix whose entries are elements of e_t; all other $\mathsf{diag}(.)$ matrices are defined similarly.

Let $C_{NT} := \min\{\sqrt{N}, \sqrt{T}\}$.

Assumption 3.6 (Moment bounds). *(i)* $\max_i(\|\lambda_i\| + \|\alpha_i\| + \|l_i\|) < C$.

(ii) $\max_{t\leq T} \|\frac{1}{N}\sum_i e_{it}\alpha_i\lambda_i'\|_F = o_P(1)$ *and* $\delta_{NT}\max_{it}|e_{it}| = o_P(1)$, *where*

$$\delta_{NT} := (C_{NT}^{-1} + b_{NT,4} + b_{NT,5})\max_{t\leq T}\|w_t\| + b_{NT,1} + b_{NT,3} + C_{NT}^{-1}b_{NT,2} + C_{NT}^{-1/2}.$$

(iii) Let $\psi_j(H)$ *denote the* j^{th} *largest singular value of matrix* H. *Suppose there is* $c>0$, *so that almost surely, for all* $t\leq T$ *and* $i\leq N$, $\min_{j\leq K_2}\psi_j(D_{\lambda i}) > c$, $\min_{j\leq K_2}\psi_j(D_{ft}) > c$, $\min_{j\leq K_2}\psi_j(\bar{D}_{\lambda i}) > c$ *and* $\min_{j\leq K_2}\psi_j(\bar{D}_{ft}) > c$. *In addition,*

$$c < \min_j \psi_j(\frac{1}{N}\sum_i l_i l_i') \leq \max_j \psi_j(\frac{1}{N}\sum_i l_i l_i') < C,$$
$$c < \min_j \psi_j(\frac{1}{T}\sum_t w_t w_t') \leq \max_j \psi_j(\frac{1}{T}\sum_t w_t w_t') < C.$$

(iv) $\max_{it}\mathsf{E}(e_{it}^8|U,F) < C$, *and* $\mathsf{E}\|w_t\|^4 + \mathsf{E}\|g_t\|^4 + \mathsf{E}\|f_t\|^4 < C$, *and* $\mathsf{E}\|g_t\|^4\|f_t\|^4 + \mathsf{E}u_{it}^4\|f_t\|^4 + \mathsf{E}e_{jt}^4\|f_t\|^8 + \mathsf{E}e_{jt}^4\|f_t\|^4\|g_t\|^4 + \mathsf{E}\|w_t\|^4\|g_t\|^4 < C$.

Assumption 3.6 (ii) imposes tail conditions on (e_{it}, w_t). If (e_{it}, w_t) are sub-Gaussian and e_{it} is independent across i, it is straightforward to obtain bounds for $b_{NT,1}, ..., b_{NT,5}$ under which one can verify that condition (ii) holds as long as $\log^2 T = o(N)$ and $(\log^2 T)(\log^2 N) = o(T)$.

3.3. **Main Results.**

Assume that $\mathcal{G}$ is a known cross-sectional subset of interest.

Theorem 3.1. *Suppose Assumptions 3.1- 3.6 hold. Fix any* $t\leq T$ *and* $\mathcal{G}\subseteq\{1,...,N\}$. *Suppose* $N, T\to\infty$ *and either (i)* $|\mathcal{G}|_0 = o(N)$ *or (ii)* $N = o(T^2)$ *holds. In addition, assume* $f_t'V_f f_t$ *and* $\bar{\lambda}_{\mathcal{G}}'V_\lambda\bar{\lambda}_{\mathcal{G}}$ *are both bounded away from zero. Then*

$$\Sigma_{\mathcal{G}}^{-1/2}\left(\frac{1}{|\mathcal{G}|_0}\sum_{i\in\mathcal{G}}\widehat{\theta}_{it} - \bar{\theta}_{\mathcal{G},t}\right) \to^d \mathcal{N}(0,1)$$

where

$$\Sigma_{\mathcal{G}} := \frac{1}{T|\mathcal{G}|_0}f_t'V_{f,\mathcal{G}}f_t + \frac{1}{N}\bar{\lambda}_{\mathcal{G}}'V_\lambda\bar{\lambda}_{\mathcal{G}}$$

with

$$\begin{aligned} V_{f,\mathcal{G}} &= \frac{1}{T}\sum_{s=1}^{T} \mathsf{Var}\left(\frac{1}{\sqrt{|\mathcal{G}|_0}} \sum_{i\in\mathcal{G}} \Omega_i f_s e_{is} u_{is} \middle| F \right), \quad \Omega_i = (\frac{1}{T}\sum_{s=1}^{T} f_s f_s' \mathsf{E} e_{is}^2)^{-1}, \\ V_\lambda &= V_{\lambda 1}^{-1} V_{\lambda 2} V_{\lambda 1}^{-1}, \quad V_{\lambda 1} = \frac{1}{N}\sum_{i=1}^{N} \lambda_i \lambda_i' \mathsf{E} e_{it}^2, \quad \bar{\lambda}_{\mathcal{G}} = \frac{1}{|\mathcal{G}|_0}\sum_{i\in\mathcal{G}} \lambda_i. \end{aligned}$$

Remark 3.2. As an implication of Theorem 3.1, which can be achieved under weaker conditions, the group average effect estimator has rate of convergence

$$\frac{1}{|\mathcal{G}|_0}\sum_{i\in\mathcal{G}} \widehat{\theta}_{it} - \bar{\theta}_{\mathcal{G},t} = O_P\left(\frac{1}{\sqrt{T|\mathcal{G}|_0}} + \frac{1}{\sqrt{N}} \right).$$

It is useful to compare this to the rate that would be obtained, for a fixed group of interest $\mathcal{G}$, using a model that imposes homogeneous effects at the group level:

$$y_{it} = \theta_{\mathcal{G},t} x_{it} + \alpha_i' g_t + u_{it}, \quad i \in \mathcal{G}. \tag{3.2}$$

The resulting "group-homogeneous effect estimator" $\widetilde{\theta}_{G,t}$ would satisfy

$$\widetilde{\theta}_{G,t} - \theta_{\mathcal{G},t} = O_P\left(|\mathcal{G}|_0^{-1/2} \right).$$

In effect, the homogeneous effect estimator uses only cross-sectional information within $\mathcal{G}$. By leveraging the factor coefficients, we estimate a heterogeneous average effect $\bar{\theta}_{\mathcal{G},t} = \frac{1}{|\mathcal{G}|_0}\sum_{i\in\mathcal{G}} \theta_{it}$ making use of all cross-sectional information which will result in a faster rate of convergence than $\widetilde{\theta}_{G,t}$ when $|\mathcal{G}|_0$ is small and will remain consistent even if the group size is finite. We also note that the homogeneous effects estimator will generally be inconsistent for group average effects when effects are actually heterogeneous. Note that we estimate the average effect for a given *known* group. See, e.g., Bonhomme and Manresa (2015) and Su et al. (2016) for results on estimating group effects with estimated group memberships.

Theorem 3.1 immediately leads to two special cases.

Corollary 3.1. *Suppose Assumptions 3.1-3.6 hold. Fix $t \leq T$.*

(i) Individual effect: Fix any $i \leq N$, then

$$\Sigma_i^{-1/2}\left(\widehat{\theta}_{it} - \theta_{it} \right) \to^d \mathcal{N}(0,1)$$

where $\Sigma_i := \frac{1}{T} f_t' V_{f,i} f_t + \frac{1}{N}\lambda_i' V_\lambda \lambda_i$ with $V_{f,i} = \frac{1}{T}\sum_{s=1}^{T} \mathsf{Var}\left(\Omega_i f_s e_{is} u_{is} \middle| F \right)$.

(ii) Cross-sectional average effect: Suppose $N = o(T^2)$ and $\liminf_N \sigma_\lambda^2 > 0$, then

$$\sqrt{N}\sigma_\lambda^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\widehat{\theta}_{it} - \bar{\theta}_t\right) \to^d \mathcal{N}(0,1)$$

where $\bar{\theta}_t = \frac{1}{N}\sum_{i=1}^{N}\theta_{it}$, $\bar{\lambda} = \frac{1}{N}\sum_{i=1}^{N}\lambda_i$ and $\sigma_\lambda^2 := \bar{\lambda}' V_\lambda \bar{\lambda}$.

We now discuss estimating the asymptotic variance. To preserve the rotation invariance property of the asymptotic variance, we estimate relevant quantities separately within subsamples and produce the final asymptotic variance estimator by averaging the results. Specifically, assuming products $e_{it}u_{it}$ are cross-sectionally independent for simplicity, let

$$\begin{aligned}
\widehat{v}_\lambda &= \frac{1}{2}(\widehat{\lambda}'_{I,\mathcal{G}}\widehat{V}_{\lambda1,I}^{-1}\widehat{V}_{\lambda2,I}\widehat{V}_{\lambda1,I}^{-1}\widehat{\lambda}_{I,\mathcal{G}} + \widehat{\lambda}'_{I^c,\mathcal{G}}\widehat{V}_{\lambda1,I^c}^{-1}\widehat{V}_{\lambda2,I^c}\widehat{V}_{\lambda1,I^c}^{-1}\widehat{\lambda}_{I^c,\mathcal{G}}),\\
\widehat{v}_{f,\mathcal{G}} &= \frac{1}{2}(\widehat{f}'_{I,t}\widehat{V}_{I,f}\widehat{f}_{I,t} + \widehat{f}'_{I^c,t}\widehat{V}_{I^c,f}\widehat{f}_{I^c,t}),\\
\widehat{V}_{S,f} &= \frac{1}{|\mathcal{G}|_0|S|_0}\sum_{s\notin S}\sum_{i\in\mathcal{G}}\widehat{\Omega}_{S,i}\widehat{f}_{S,s}\widehat{f}'_{S,s}\widehat{\Omega}_{S,i}\widehat{e}_{is}^2\widehat{u}_{is}^2,\\
\widehat{V}_{\lambda1,S} &= \frac{1}{N}\sum_j\widehat{\lambda}_{S,j}\widehat{\lambda}'_{S,j}\widehat{e}_{jt}^2, \quad \widehat{V}_{\lambda2,S} = \frac{1}{N}\sum_j\widehat{\lambda}_{S,j}\widehat{\lambda}'_{S,j}\widehat{e}_{jt}^2\widehat{u}_{jt}^2,\\
\widehat{\lambda}_{S,\mathcal{G}} &= \frac{1}{|\mathcal{G}|_0}\sum_{i\in\mathcal{G}}\widehat{\lambda}_{S,i}, \text{ and } \widehat{\Omega}_{S,i} = (\frac{1}{|S|_0}\sum_{s\in S}\widehat{f}_{S,s}\widehat{f}'_{S,s})^{-1}(\frac{1}{T}\sum_{s=1}^{T}\widehat{e}_{is}^2)^{-1}.
\end{aligned}$$

Corollary 3.2. *In addition to the assumptions of Theorem 3.1, assume $e_{it}u_{it}$ are cross-sectionally independent conditionally on F. Then for any fixed $t \le T$ and $\mathcal{G}$,*

$$\left(\frac{1}{N}\widehat{v}_\lambda + \frac{1}{T|\mathcal{G}|_0}\widehat{v}_{f,\mathcal{G}}\right)^{-1/2}\left(\frac{1}{|\mathcal{G}|_0}\sum_{i\in\mathcal{G}}\widehat{\theta}_{it} - \bar{\theta}_{\mathcal{G},t}\right) \to^d \mathcal{N}(0,1).$$

3.4. **Policy relevant tests.** We now show how we can use our results to test the policy relevant hypotheses described in Section 3.1. As part of this discussion, we illustrate leveraging the structure of coefficients in the factor-slope model, $\bar{\theta}_{\mathcal{G},t} = \bar{\lambda}'_{\mathcal{G}}f_t$, to provide tractable tests.

Consider testing the null hypothesis of homogeneity of group average effects in a given time period t:

$$H_0^1 : \bar{\theta}_{\mathcal{G}_1,t} = ... = \bar{\theta}_{\mathcal{G}_J,t}.$$

Let $S := (\widehat{\theta}_{\mathcal{G}_1,t} - \widehat{\theta}_{\mathcal{G}_2,t}, \widehat{\theta}_{\mathcal{G}_2,t} - \widehat{\theta}_{\mathcal{G}_3,t}, ..., \widehat{\theta}_{\mathcal{G}_{J-1},t} - \widehat{\theta}_{\mathcal{G}_J,t})'$. We define the test statistic

$$TS'(\Xi\widehat{D}\Xi')^{-1}S \text{ for } \Xi = \begin{pmatrix} 1 & -1 & 0 & \cdots & & \\ 0 & 1 & -1 & 0 & \cdots & \\ \vdots & & & & & \\ 0 & \cdots & & & 1 & -1 \end{pmatrix}$$

where $\widehat{D} = \mathsf{diag}\{\frac{1}{|\mathcal{G}_j|_0}\widehat{v}_{f,\mathcal{G}_j} : j \leq J\}$ is the diagonal matrix of the estimated asymptotic variances of $\widehat{\theta}_{\mathcal{G}_j,t}$. It may be interesting to note that this test is essentially a test of group homogeneity of loadings in the sense that $\bar{\theta}_{\mathcal{G}_j,t} - \bar{\theta}_{\mathcal{G}_k,t} = (\bar{\lambda}_{\mathcal{G}_j} - \bar{\lambda}_{\mathcal{G}_k})'f_t$ for any two groups $\mathcal{G}_j, \mathcal{G}_k$. Therefore, H_0^1 is the same as the null hypothesis of homogeneity of group average loadings except when the inner product between $\bar{\lambda}_{\mathcal{G}_j} - \bar{\lambda}_{\mathcal{G}_k}$ and f_t happens to be zero.

Next, consider the test of joint significance at a given time t:

$$H_0^2 : \theta_{it} = 0 \text{ for all } i.$$

The factor slope structure brings us at least two benefits to implement this test. First, although this hypothesis is high-dimensional, the problem can be transformed to testing a much simpler low-dimensional hypothesis, $H_0^{2\prime} : f_t = 0$. Under H_0^2, the factor structure yields $\Lambda f_t = 0$ where Λ is the $N \times \dim(f_t)$ matrix of λ_i. Suppose $(\Lambda'\Lambda)^{-1}$ exists, then H_0^2 is equivalent to $f_t = 0$ almost surely, which follows from left multiplication by $(\Lambda'\Lambda)^{-1}\Lambda'$. A secondary benefit of focusing on f_t is that failing to reject $f_t = 0$ will also imply $\theta_{it} = 0$ for any out-of-sample cross-sectional unit i influenced by the same factors.

Focusing on $H_0^{2\prime}$, a simple test can be constructed from

$$\widehat{f}_t := \frac{1}{2}(\widehat{f}_{I,t} + \widehat{f}_{I^c,t})$$

and associated estimated asymptotic variance $\widehat{V}_\lambda := \frac{1}{2}\widehat{V}_{\lambda 1,I}^{-1}\widehat{V}_{\lambda 2,I}\widehat{V}_{\lambda 1,I}^{-1} + \frac{1}{2}\widehat{V}_{\lambda 1,I^c}^{-1}\widehat{V}_{\lambda 2,I^c}\widehat{V}_{\lambda 1,I^c}^{-1}$.

The following theorem presents the asymptotic null distribution of the two tests. Let $\chi^2(a)$ denote the centered chi-square distribution with degrees of freedom a.

Theorem 3.2. *Suppose the assumptions of Corollary 3.2 hold.*

(i) For H_0^1, suppose $\bar{\lambda}_{\mathcal{G}_1} = ... = \bar{\lambda}_{\mathcal{G}_J}$ under H_0^1. Assume $\max_{j \leq J} |\mathcal{G}_j|_0 = o(T)$, $T\max_{j \leq J} |\mathcal{G}_j|_0 = o(N^2)$, $\max_j |\mathcal{G}_j|_0 / \min_j |\mathcal{G}_j|_0 = O(1)$ *and J is fixed. Under H_0^1,*

$$TS'(\Xi\widehat{D}\Xi')^{-1}S \to^d \chi^2(J-1).$$

(ii) For H_0^2, suppose $f_t = 0$ almost surely under H_0^2. Then under H_0^2,

$$N\widehat{f}_t'\widehat{V}_\lambda^{-1}\widehat{f}_t \to^d \chi^2(\dim(f_t)).$$

3.5. **Consistent rank estimation.** This section considers consistent estimation of the ranks K_1 and K_2. Several methods based on information criteria or eigengaps are available for estimating the number of factors; see, e.g., Bai and Ng (2002), Onatski (2010), and Ahn and Horenstein (2013). While these methods could be adapted to the current context, here we provide a simple method that is a natural byproduct of the nuclear norm penalized estimator (2.2).

Recall that $(\widetilde{M}, \widetilde{\Theta})$ are the low-rank estimators obtained from solving (2.2) using the full data $i = 1, ..., N$ and $t = 1, ..., T$. We can estimate K_1, K_2 as

$$\widehat{K}_1 = \sum_i 1\{\psi_i(\widetilde{\Theta}) \geq (\nu_0\|\widetilde{\Theta}\|)^{1/2}\}, \quad \widehat{K}_2 = \sum_i 1\{\psi_i(\widetilde{M}) \geq (\nu_1\|\widetilde{M}\|)^{1/2}\},$$

where $\psi_i(W)$ denotes the i^{th} largest singular-value of a matrix W. Proposition D.1 in the supplement shows that $\widehat{K}_1$ and $\widehat{K}_2$ are consistent estimators of the rank of Θ and M. All formal results in the previous sections are unaffected by using $\widehat{K}_1$ and $\widehat{K}_2$ obtained in an initial step before applying Algorithm 2.2.

4. Monte Carlo Simulations

In this section, we provide some simulation results for inference about a specific θ_{it}. We provide additional simulation results regarding the hypothesis tests discussed in Section 3.4 and in a dynamic model in the supplementary appendix.

We generate outcomes as

$$y_{it} = \alpha_i' g_t + x_{it,1}\theta_{it} + x_{it,2}\beta_{it} + u_{it}$$

where $\theta_{it} = \lambda_{i,1}' f_{t,1}$ and $\beta_{it} = \lambda_{i,2}' f_{t,2}$. The observed regressors are generated as

$$x_{it,r} = l_{i,r}' w_{t,r} + \mu_x + e_{it,r}, \quad r = 1, 2,$$

where $(e_{it,r}, u_{it})$ are generated independently from the standard normal distribution across (i, t, r). We set $\mu_x = 2$, so $x_{it,r}$ follows a factor model with an intercept. We then estimate the structure of $x_{it,r}$ for partialing-out via the principal components estimator applied to the matrix $(s_{ij,r})_{N\times N}$:

$$s_{ij,r} = \frac{1}{T}\sum_{t=}^{T}(x_{it,r} - \bar{x}_{i,r})(x_{jt,r} - \bar{x}_{j,r}), \quad \bar{x}_{i,r} = \frac{1}{T}\sum_t x_{it,r}.$$

We set each number of factors to one and generate all factors and loadings independently as draws from $\mathcal{N}(2,1)$ random variables. The loadings $(\alpha_i, \lambda_{i,1}, \lambda_{i,2}, l_{i,r})$ are treated as fixed in the simulations while the factors $(g_t, f_{t,1}, f_{t,2}, w_{t,r})$ are treated as random and sampled across replications. Results are based on 1000 replications.

We compare four inferential methods:

(I) ("Partial-out") The proposed estimator that partials out μ_{it} from x_{it} and uses sample-splitting. Feasible standard errors given in Corollary 3.2 are used.

(II) ("Par-infeasible") The proposed estimator that partials out μ_{it} from x_{it} and uses sample-splitting. Infeasible simulation standard errors are used for inference.

(III) ("No Par-out") The estimator that uses sample splitting but does not make use of partialing-out. That is, we run Steps 2 and 3 of Algorithm 2.2 using data for $s \in I$, and obtain $(\widetilde{f}_{I,s}, \dot{\lambda}_{I,i})$ for $s \in I^c \cup \{t\}$ and all i. We then exchange I and I^c to obtain $(\widetilde{f}_{I^c,t}, \dot{\lambda}_{I,i})$. The estimator is then defined as $\frac{1}{2}(\dot{\lambda}'_{I,i}\widetilde{f}_{I,t} + \dot{\lambda}'_{I^c,i}\widetilde{f}_{I^c,t})$. Infeasible simulation standard errors are used for inference.

(IV) ("Regularized") The estimator that uses nuclear-norm regularization only. That is, θ_{it} is directly estimated as the $(i,t)^{\text{th}}$ element of $\widetilde{\Theta}$ in (2.2). Infeasible simulation standard errors are used for inference.

We report results only for θ_{it} with $i = t = 1$; results for other values of (i,t) and for β_{it} are similar. Table 4 reports the fraction of simulation draws where the true value for θ_{it} was contained in the 95% confidence interval:

$$[\widehat{\theta}_{it} - 1.96se(\widehat{\theta}_{it}), \widehat{\theta}_{it} - 1.96se(\widehat{\theta}_{it})].$$

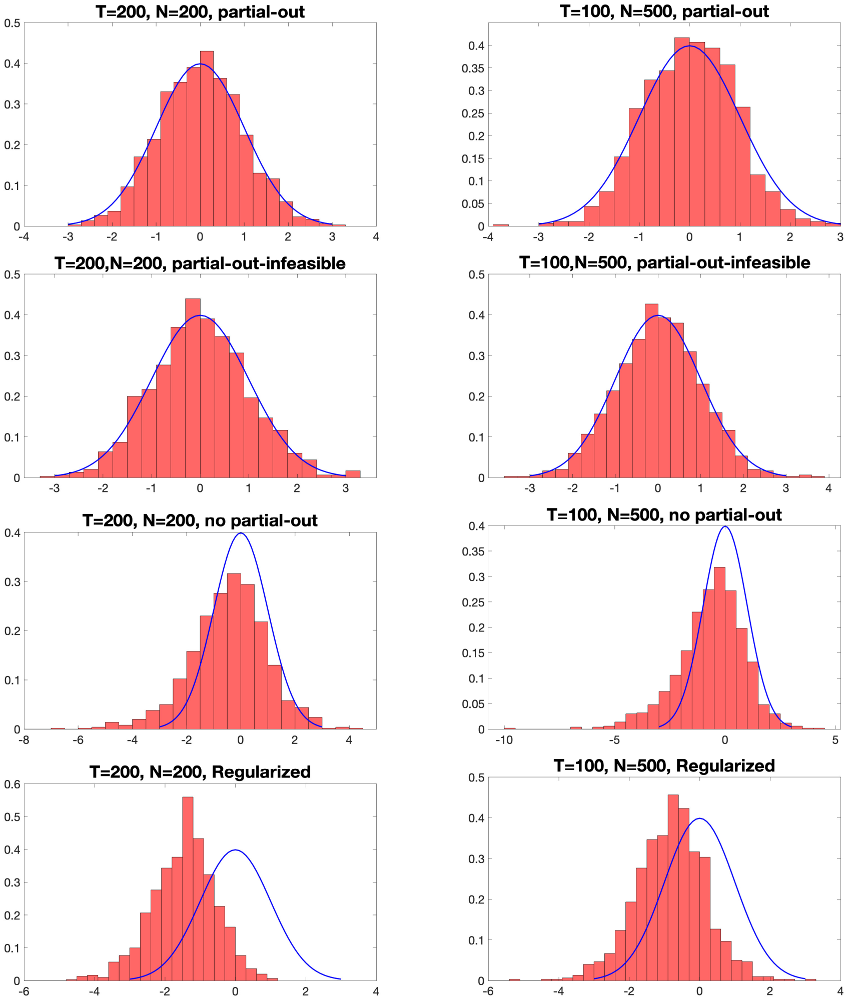
where $se(\widehat{\theta}_{it})$ is the respective standard error for (I)-(IV) defined above.

Figure 4 plots the (scaled) histogram of the standardized estimates, superimposed with the standard normal density. The top panels of Figure 4 are for the proposed estimation method; the middle panels are for estimation without partialing-out the mean structure of x_{it}, and the bottom panels are for the nuclear-norm regularized estimators without any debiasing. It appears that asymptotic theory provides a good approximation to the finite sample distributions for the proposed post-SVT method. In contrast, the estimated θ_{it} produced without partialing-out and produced by applying nuclear norm regularization without further debiasing noticeably deviate from the standard normal distribution.

Table 4.1. Coverage Probability in Static Simulation Design

N	T	Partial-out	Par-infeasible	No par-out	Regularized
100	100	0.943	0.953	0.794	0.778
200	200	0.952	0.944	0.835	0.911
500	100	0.945	0.946	0.804	0.883

Note: This table reports the simulated coverage probability of 95% confidence intervals. The "Partial-out" estimator uses feasible standard errors. All other methods use infeasible simulation standard errors. Results are based on 1000 simulation replications.



T=200, N=200, partial-out

T=100, N=500, partial-out

T=200,N=200, partial-out-infeasible

T=100,N=500, partial-out-infeasible

T=200, N=200, no partial-out

T=100, N=500, no partial-out

T=200, N=200, Regularized

T=100, N=500, Regularized

Figure 4.1. Histograms of standardized estimates $((\widehat{\theta}_{11} - \theta_{11})/se(\widehat{\theta}_{11}))$. The standard normal density function is superimposed on each histogram.

References

Ahn, S. C. and Horenstein, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica* **81** 1203–1227.

Ahn, S. C., Lee, Y. H. and Schmidt, P. (2013). Panel data models with multiple time-varying individual effects. *Journal of Econometrics* **174** 1–14.

Athey, S., Bayati, M., Doudchenko, N., Imbens, G. and Khosravi, K. (2018). Matrix completion methods for causal panel data models. Tech. rep., National Bureau of Economic Research.

Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica* **77** 1229–1279.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70** 191–221.

Bai, J. and Ng, S. (2019). Rank regularized estimation of approximate factor models. *Journal of Econometrics* **212** 78–96.

Bonhomme, S. and Manresa, E. (2015). Grouped patterns of heterogeneity in panel data. *Econometrica* **83** 1147–1184.

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

Chamberlain, G. and Hirano, K. (1999). Predictive distributions based on longitudinal earnings data. *Annales d'Economie et de Statistique* 211–242.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal* **21** C1–C68.

Feng, G., Gao, J., Peng, B. and Zhang, X. (2017). A varying-coefficient panel data model with fixed effects: Theory and an application to us commercial banks. *Journal of Econometrics* **196** 68–82.

Hsiao, C., Pesaran, M. and Tahmiscioglu, A. (1999). Bayes estimation of short-run coefficients in dynamic panel data models. In *Analysis of Panel Data and Limited Dependent Variable Models*.

Klopp, O. (2014). Noisy low-rank matrix completion with general sampling distribution. *Bernoulli* **20** 282–303.

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

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

Moon, H. R. and Weidner, M. (2018). Nuclear norm regularized estimation of panel regression models. *arXiv preprint arXiv:1810.10987* .

Moon, R. and Weidner, M. (2015). Linear regression for panel with unknown number of factors as interactive fixed effects. *Econometrica* **83** 1543–1579.

Negahban, S. and Wainwright, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics* **39** 1069–1097.

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

Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics* **92** 1004–1016.

Pesaran, H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica* **74** 967–1012.

Recht, B., Fazel, M. and Parrilo, P. A. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review* **52** 471–501.

Su, L., Jin, S. and Zhang, Y. (2015). Specification test for panel data models with interactive fixed effects. *Journal of Econometrics* **186** 222–244.

Su, L., Shi, Z. and Phillips, P. C. (2016). Identifying latent structures in panel data. *Econometrica* **84** 2215–2264.

Su, L. and Wang, X. (2017). On time-varying factor models: Estimation and testing. *Journal of Econometrics* **198** 84–101.

Sun, T. and Zhang, C.-H. (2012). Calibrated elastic regularization in matrix completion. In *Advances in Neural Information Processing Systems*.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027* .

Department of Economics, MIT, Cambridge, MA 02139
Email address: vchern@mit.edu

Booth School of Business, University of Chicago, Chicago, IL 60637
Email address: Christian.Hansen@chicagobooth.edu

Department of Economics, Rutgers University, New Brunswick, NJ 08901
Email address: yuan.liao@rutgers.edu

Department of Economics, Brandeis University, 415 South St, Waltham, MA 02453
Email address: yinchuzhu@brandeis.edu