A Unified Framework for Estimation of High-dimensional Conditional Factor Models

Qihui Chen
School of Management and Economics
The Chinese University of Hong Kong, Shenzhen
qihuichen@cuhk.edu.cn

This Draft: September 14, 2022
First Draft: September 1, 2022

Abstract

This paper develops a general framework for estimation of high-dimensional conditional factor models via nuclear norm regularization. We establish large sample properties of the estimators, and provide an efficient computing algorithm for finding the estimators as well as a cross validation procedure for choosing the regularization parameter. The general framework allows us to estimate a variety of conditional factor models in a unified way and quickly deliver new asymptotic results. We apply the method to analyze the cross section of individual US stock returns, and find that imposing homogeneity may improve the model's out-of-sample predictability.

KEYWORDS: Nuclear norm regularization, Factor models, Characteristics, Macro state variables, Factor zoo, Missing values

1 Introduction

A central question in empirical asset pricing is why different assets earn different average returns. Conditional factor models provide a general framework of utilizing conditional information in tackling the question (Gagliardini, Ossola, and Scaillet, 2020). This paper studies the following high-dimensional conditional factor model

$$y_{it} = \alpha_{it} + \beta_{it}' f_t + \varepsilon_{it} \text{ with } \alpha_{it} = a_i' x_{it} \text{ and } \beta_{it} = B_i' x_{it}, i = 1, \ldots, N, t = 1, \ldots, T, \quad (1)$$

where y_{it} is the excess return of asset i in time period t, f_t is a $K \times 1$ vector of *unobserved latent* factors, α_{it} is a pricing error, β_{it} a $K \times 1$ vector of risk exposures, ε_{it} is an error term, x_{it} is a $p \times 1$ vector of pre-specified explanatory variables known at the beginning of time period t (for example, constant, sieve transformations of asset's characteristics, sieve transformations of macro state variables, and their interactions), and a_i and B_i are unknown $p \times 1$ vector and $p \times K$ matrix of coefficients. The model captures time-variation in the risk exposures and the pricing error through their association with x_{it}. Meanwhile, it allows for distinguishing between risk and mispricing explanations of the role of x_{it} in predicting asset returns. In addition, since K can be much smaller than p, the model enables us to summarize the information in a large dimension of x_{it} into a small number of factors, that is, tame the "factor zoo" (Cochrane, 2011). However, estimating the model has at least two challenges: (i) the factors f_t are unknown/unobservable; (ii) the dimension of the unknown parameters $\{a_i, B_i, f_t\}_{i \leq N, t \leq T}$ is high.

The model nests many factor models in the literature. In contrast to homogeneous versions of conditional factor models (Kelly, Pruitt, and Su, 2017, 2019; Chen, Roussanov, and Wang, 2021), our model allows a_i and B_i to be heterogenous across i. This in turn enables our model to nest classical factor models (Ross, 1976; Chamberlain and Rothschild, 1982) where $x_{it} = 1$ and $a_i = 0$; semiparametric factor models (Connor, Hagmann, and Linton, 2012; Fan, Liao, and Wang, 2016; Kim, Korajczyk, and Neuhierl, 2020) where x_{it} consists of constant and sieve transformations of asset's time-invariant characteristics, and the rows of a_i and B_i corresponding to nonconstant explanatory variables are homogenous across i; and state-varying factor models (Pelger and Xiong, 2021) where x_{it} consists of constant and sieve transformations of macro state variables, and $a_i = 0$. In contrast to Gagliardini, Ossola, and Scaillet (2016), our model does not require observable f_t and allows for the presence of arbitrage and large p.

We provide a general framework for estimation of high-dimensional conditional factor models. Specifically, we develop a nuclear norm regularized estimation of the model in (1) with constraints on $\{a_i, B_i\}_{i \leq N}$. The estimation procedure consists of two steps: first estimating an $Np \times T$ reduced rank matrix composed of block matrices $\{a_i + B_i f_t\}_{i \leq N, t \leq T}$ using the nuclear norm regularization under the constraints; then extracting estimators of K, $\{a_i\}_{i \leq N}$, $\{B_i\}_{i \leq N}$ and $\{f_t\}_{t \leq T}$ from the estimated matrix using eigenvalue decom-

position. We establish asymptotic properties of the estimators under a restricted strong convexity condition. Specifically, we establish a rate of convergence of the estimators of the reduced rank matrix, $\{a_i\}_{i\leq N}$, $\{B_i\}_{i\leq N}$ and $\{f_t\}_{t\leq T}$, and consistency of the estimator of K. Our framework allows both $p \to \infty$ and $K \to \infty$, and may accommodate the presence of missing values, which are prevalent in stock return data sets.

The general framework allows us to estimate the aforementioned nested models in a unified way, while existing methods are model-specific or restrictive.[1] We specialize the general theory for each model by providing simple primitive conditions. Our contributions are four-fold. First, we provide a novel estimation of the homogenous conditional factor model, which allows p to grow as fast as N. Second, we provide an estimation of the semiparametric factor model, which allows for time-varying characteristics, nonzero pricing errors, and non-noisy intercepts in pricing errors and risk exposures. Third, we provide an estimator that can consistently estimate the factor space in the state-varying factor model. Fourth, to the best of our knowledge, our paper is the first one that provides an estimation of the unconstrained conditional factor model.

To facilitate the use of our estimation procedure in practice, we make two contributions. First, we provide an efficient computing algorithm for finding the constrained nuclear norm regularized estimator of the reduced rank matrix in each of the nested models. This is practically important since the constrained nuclear norm regularized estimation involves a high-dimensional constrained nonsmooth convex minimization. Second, we propose a cross validation (CV) procedure to choose the regularization parameter, and demonstrate its validity through a set of Monte Carlo simulations. This is useful since the estimates are usually sensitive to the choice of the regularization parameter. Our simulation studies show that the finite sample performance of our estimators by using the CV chosen regularization parameter is satisfactory and encouraging. We apply the unified framework to analyze the cross section of individual stock returns in the US market, and find that imposing homogeneity of a_i and B_i across i may improve the model's out-of-sample predictability.

Nuclear norm regularization has been widely used for reduced rank matrix estimation in the statistical literature. The parameter of interest there is usually the reduced rank matrix per se. For example, Negahban and Wainwright (2011) study an unconstrained nuclear norm regularized estimation of trace linear regression models under a restricted strong convexity condition; Rohde and Tsybakov (2011) consider the same problem

[1]For example, Connor and Korajczyk (1986), Stock and Watson (2002), Bai and Ng (2002) and Bai (2003) estimate the classical factor model by principal component analysis (PCA), while Fan, Liao, and Mincheva (2013) use a principal orthogonal complement thresholding method. Fan et al. (2016) propose a projected-PCA for the semiparametric factor model. Pelger and Xiong (2021) estimate the state-varying factor by a local version of PCA based on kernel smoothing. Chen et al. (2021) develop a regressed-PCA for the homogenous conditional factor model, while Gu, Kelly, and Xiu (2021) propose an autoencoder method. Gagliardini et al. (2016) require observable factors for estimating a conditional factor model with no arbitrage.

under a restricted isometry condition. Our work differs from these studies in at least two aspects. First, our work requires a constrained nuclear norm regularization, thus we need to extend the results to allow for enforcing constraints. Second, our parameters of interest are K, $\{a_i\}_{i\leq N}$, $\{B_i\}_{i\leq N}$ and $\{f_t\}_{t\leq T}$, rather than the reduced rank matrix.

There are several recent studies of unconstrained nuclear norm regularization in the econometric literature. For example, Bai and Ng (2019) use it to improve estimation of the classical factor model; Moon and Weidner (2018) leverage it to improve estimation of panel data models with interactive fixed effects; Athey, Bayati, Doudchenko, Imbens, and Khosravi (2021) adopt it in treatment effect estimation; Chernozhukov, Hansen, Liao, and Zhu (2018) employ it to estimate panel data models with heterogenous coefficients. To the best of our knowledge, the use of constrained nuclear norm regularization in estimating conditional factor models is not studied previously.

The literature on the cross section of asset returns is vast; here we focus on conditional factor models. While the focus of our paper is on models with latent factors (see previous paragraphs for a literature review), a large number of works in empirical asset pricing vastly relies on pre-specified observable factors (e.g, constructed via (Fama and French, 1993)'s portfolio-sorting approach utilizing asset's characteristics). These include Shanken (1990), Ferson and Harvey (1991, 1999), Lettau and Ludvigson (2001), Nagel and Singleton (2011) and Gagliardini et al. (2016), to name a few; see Gagliardini et al. (2020) for a comprehensive review. This line of works may suffer from the "characteristics versus covariances" debate (Daniel and Titman, 1997), and the "factor zoo" problem. We complement the literature by providing a unified method for estimating conditional factor models without the need to pre-specify the factors, which are well-suited for resolving both the debate and the problem (Chen et al., 2021).

The remainder of the paper is organized as follows. Section 2 presents several nested models. Section 3 develops the general estimation framework. Section 4 establishes the asymptotic properties of the estimators. Section 5 specializes the general theory for each of the nested models. Section 6 contains some simulation studies. Section 7 analyzes the cross section of individual US stock returns. Section 8 briefly concludes. Proofs of the main results are collected in Appendix A. The computing algorithms are presented in Appendix B, while useful lemmas are collected in Appendix C.

For convenience of the reader, we collect standard pieces of notation here, which will be used throughout the paper. We use I_k to denote a $k \times k$ identity matrix. We use $\|x\|$ to denote the Euclidian norm of a column vector x. For a symmetric matrix A, we denote its trace by $\text{tr}(A)$, its smallest and largest eigenvalues by $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$. For a matrix A, we denote its operator norm by $\|A\|_2$, its Frobenius norm by $\|A\|_F$, and its vectorization by $\text{vec}(A)$. We use $C \otimes D$ to denote the Kronecker product of matrices C and D. Unless specified, asymptotic statements in the paper shall be understood to hold as $N \to \infty$ with fixed T or as $(N, T) \to \infty$, whenever appropriate.

2 Related Examples

Our model in (1) nests many factor models in the literature.

Example 2.1 (Classical Factor Models)**.** The arbitrage pricing theory by Ross (1976) and Chamberlain and Rothschild (1982) leads to the following model

$$y_{it} = \lambda_i' f_t + e_{it}, \tag{2}$$

where λ_i is an unknown vector of risk exposures and e_{it} is the idiosyncratic component. Our model nests (2) where $x_{it} = 1$, $a_i = 0$, $B_i = \lambda_i'$, and $\varepsilon_{it} = e_{it}$.

Example 2.2 (Semiparametric Factor Models)**.** Connor et al. (2012), Fan et al. (2016) and Kim et al. (2020) consider the following model[2]

$$y_{it} = \phi(z_i) + \mu_i + (\Phi(z_i) + \lambda_i)' f_t + e_{it}, \tag{3}$$

where z_i is a vector of asset's time-invariant characteristics, $\phi(\cdot)$ and $\Phi(\cdot)$ are unknown functions, μ_i and λ_i are unknown scalar and vector (intercepts in pricing errors and risk exposures), and e_{it} is the idiosyncratic component. Following sieve methods, $\phi(z_i) = \phi' h(z_i) + \delta(z_i)$ and $\Phi(z_i) = \Phi' h(z_i) + \Delta(z_i)$, where $h(z_i)$ is a vector of basis functions of z_i (which does not consist of constant), ϕ and Φ are unknown vector and matrix of coefficients, and $\delta(z_i)$ and $\Delta(z_i)$ are negligible sieve approximation errors. Our model nests (3) where $x_{it} = (1, h(z_i)')'$, $a_i = (\mu_i, \phi')'$, $B_i = (\lambda_i, \Phi')'$, and $\varepsilon_{it} = e_{it} + \delta(z_i) + \Delta(z_i)' f_t$. Thus, the rows of a_i and B_i corresponding to $h(z_i)$ are homogenous across i.

Example 2.3 (State-varying Factor Models)**.** Pelger and Xiong (2021) study the following model

$$y_{it} = \Phi_i(z_t)' f_t + e_{it}, \tag{4}$$

where z_t is a vector of constant and macro state variables known at the beginning of time period t, $\Phi_i(\cdot)$ is a vector of unknown functions, and e_{it} is the idiosyncratic component. Following sieve methods, $\Phi_i(z_t) = \Phi_i' h(z_t) + \Delta_i(z_t)$, where $h(z_t)$ is a vector of basis functions of z_t (which may consist of constant), Φ_i is an unknown matrix of coefficients, and $\Delta_i(z_t)$ is a vector of negligible sieve approximation errors. Our model nests (4) where $x_{it} = h(z_t)$, $a_i = 0$, $B_i = \Phi_i$, and $\varepsilon_{it} = e_{it} + \Delta_i(z_t)' f_t$.

Example 2.4 (Homogeneous Conditional Factor Models)**.** Kelly et al. (2017, 2019) and Chen et al. (2021) develop the following model[3]

$$y_{it} = \phi(z_{it}) + \Phi(z_{it})' f_t + e_{it}, \tag{5}$$

[2] Fan et al. (2016) assume that $\phi(\cdot) = 0$ and $\mu_i = 0$, Connor et al. (2012) additionally assume that $\Phi(\cdot)$ are univariate functions and $\lambda_i = 0$, and Kim et al. (2020) assume that $\mu_i = 0$ and $\lambda_i = 0$.

[3] Kelly et al. (2017, 2019) assume that $\phi(\cdot)$ and $\Phi(\cdot)$ are linear functions.

where z_{it} is a vector of constant and asset's characteristics known at the beginning of time period t, $\phi(\cdot)$ and $\Phi(\cdot)$ are unknown functions, and e_{it} is the idiosyncratic component. Following sieve methods, $\phi(z_{it}) = \phi' h(z_{it}) + \delta(z_{it})$ and $\Phi(z_{it}) = \Phi' h(z_{it}) + \Delta(z_{it})$, where $h(z_{it})$ is a vector of basis functions of z_{it} (which may consist of constant), ϕ and Φ are unknown vector and matrix of coefficients, and $\delta(z_{it})$ and $\Delta(z_{it})$ are negligible sieve approximation errors. Our model nests (5) where $x_{it} = h(z_{it})$, $a_i = \phi$, $B_i = \Phi$, and $\varepsilon_{it} = e_{it} + \delta(z_{it}) + \Delta(z_{it})' f_t$. Thus, a_i and B_i are homogenous across i.

Example 2.5 (Unconstrained Conditional Factor Models)**.** In the absence of arbitrage opportunities, Gagliardini et al. (2016) propose the following model

$$y_{it} = z_t' \Phi_i z_t + z_{it}' \Psi_i z_t + z_t' \Upsilon_i f_t + z_{it}' \Lambda_i f_t + e_{it}, \tag{6}$$

where z_t is a vector of constant and macro state variables known at the beginning of time period t, z_{it} is a vector of asset's characteristics known at the beginning of time period t, Φ_i, Ψ_i, Υ_i and Λ_i are unknown matrices of coefficients satisfying certain no arbitrage constraints, and e_{it} is the idiosyncratic component. Our model nests (6) without no arbitrage constraints where x_{it} consists of quadratic transformations of z_t and z_{it}, a_i and B_i are transformations of Φ_i, Ψ_i, Υ_i and Λ_i, and $\varepsilon_{it} = e_{it}$. In contrast to their estimation method which relies on observable f_t, our estimation procedure treats f_t as latent factors, and allows for the presence of arbitrage and large p.

3 Estimation Strategy

We begin by rewriting the model in (1) using vectors/matrices. Let Π be an $Np \times T$ unknown parameter matrix that collects the product of (a_i, B_i) and $(1, f_t')'$, that is,

$$\Pi \equiv \begin{pmatrix} (a_1, B_1) \\ (a_2, B_2) \\ \vdots \\ (a_N, B_N) \end{pmatrix} \left(\begin{pmatrix} 1 \\ f_1 \end{pmatrix}, \begin{pmatrix} 1 \\ f_2 \end{pmatrix}, \cdots, \begin{pmatrix} 1 \\ f_T \end{pmatrix} \right) \equiv a 1_T' + B F', \tag{7}$$

where 1_T denotes a $T \times 1$ vector of ones, $a \equiv (a_1', a_2', \ldots, a_N')'$, $B \equiv (B_1', B_2', \ldots, B_N')'$, and $F \equiv (f_1, f_2, \ldots, f_T)'$. Let $X_{it} \equiv (e_{N,i} \otimes x_{it}) e_{T,t}'$ be an $Np \times T$ observed data matrix of x_{it}, where $e_{N,i}$ is the ith column of I_N and $e_{T,t}$ is the tth column of I_T. Then $x_{it}' a_i + x_{it}' B_i f_t = \mathrm{tr}(X_{it}' \Pi)$, so (1) can be compactly written as

$$y_{it} = \mathrm{tr}(X_{it}' \Pi) + \varepsilon_{it}. \tag{8}$$

Since Π has at most rank $K + 1$, (8) can be viewed as a trace linear regression model with reduced rank coefficient matrix Π (Negahban and Wainwright, 2011; Rohde and

Tsybakov, 2011). Thus, we first estimate Π by using the nuclear norm regularization (Fazel, 2002), which uses the nuclear norm penalty as a surrogate function to enforce the reduced rank constraint. Our estimator of Π is given by

$$\hat{\Pi} = \underset{\Gamma \in \mathcal{S} \subset \mathbf{R}^{Np \times T}}{\arg\min} \frac{1}{2} \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - \mathrm{tr}(X_{it}'\Gamma))^2 + \lambda_{NT} \|\Gamma\|_*, \tag{9}$$

where $\mathcal{S} \subset \mathbf{R}^{Np \times T}$ is convex, $\|\Gamma\|_*$ is the nuclear norm of Γ, and $\lambda_{NT} > 0$ is a regularization parameter.[4] In particular, introducing $\mathcal{S}$ allows us to enforce the constraints of Π induced by those of a and B, which has not been studied in the literature. We may set $\mathcal{S} = \mathbf{R}^{Np \times T}$ in Examples 2.1, 2.3 and 2.5, $\mathcal{S} = \mathcal{D}_M$ for $0 < M < \infty$ ($\mathcal{D}_M$ is given in (15)) in Example 2.2, and $\mathcal{S} = \{1_N \otimes \Gamma : \Gamma \in \mathbf{R}^{p \times T}\}$ in Example 2.4; see Section 5 for details. Since (9) involves a constrained nonsmooth convex minimization, $\hat{\Pi}$ does not have an analytical closed form in general. There are several algorithms for solving convex minimization problems with nuclear norm in the literature (Vandenberghe and Boyd, 1996; Bertsekas, 1999; Liu and Vandenberghe, 2010; Ma, Goldfarb, and Chen, 2011); however they may not be favored for high-dimensional settings with constraints. In Appendix B, we provide an efficient computing algorithm for each of Examples 2.1-2.5.

We next proceed to extract estimators for K, a, B and F from the nuclear norm regularized estimator $\hat{\Pi}$. Denote the estimators by $\hat{K}$, $\hat{a}$, $\hat{B}$ and $\hat{F}$. Let $M_T \equiv I_T - 1_T 1_T'/T$. Since $\Pi M_T = BF'M_T$, we may obtain $\hat{K}$ and $\hat{B}$ from the eigenvalues and eigenvectors of $\hat{\Pi} M_T \hat{\Pi}'$. Specifically, $\hat{K}$ is given by

$$\hat{K} = \sum_{j=1}^{Np} 1\{\lambda_j(\hat{\Pi} M_T \hat{\Pi}') \geq \delta_{NT}\}, \tag{10}$$

where $\lambda_j(A)$ denotes the jth largest eigenvalue of A and $\delta_{NT} > 0$ is a threshold value. If $\hat{K} = 0$, $\hat{a} = \hat{\Pi} 1_T/T$, $\hat{B} = 0$ and $\hat{F} = 0$; otherwise we proceed as follows. To estimate B, we use the following normalization: $B'B/N = I_K$ and $F'M_TF/T$ being diagonal with diagonal entries in descending order. Then the columns of $\hat{B}/\sqrt{N}$ are given by the eigenvectors of $\hat{\Pi} M_T \hat{\Pi}'$ corresponding to its largest $\hat{K}$ eigenvalues. To estimate a and F, we impose the following condition: $a'B = 0$. Since $a = (I_{Np} - B(B'B)^{-1}B')\Pi 1_T/T$ and $F = \Pi' B(B'B)^{-1}$, we thus obtain

$$\hat{a} = \left(I_{Np} - \frac{\hat{B}\hat{B}'}{N}\right) \frac{\hat{\Pi} 1_T}{T} \text{ and } \hat{F} = \frac{\hat{\Pi}'\hat{B}}{N}. \tag{11}$$

Remark 3.1. In the presence of $a = 0$, we may enforce the information to extract

[4] The nuclear norm of Γ is given by $\|\Gamma\|_* = \sum_{j=1}^{\min\{Np,T\}} \sigma_j(\Gamma)$, corresponding to the sum of its singular values, where $\sigma_j(\Gamma)$'s are the singular values of Γ. The nuclear norm of Γ is the convex envelope of the rank of Γ over the set of matrices with spectral norm no greater than one; see, for example, Recht, Fazel, and Parrilo (2010).

estimators for K, B and F from $\hat{\Pi}$ in a similar manner. Denote the estimators by $\tilde{K}$, $\tilde{B}$ and $\tilde{F}$. Since $\Pi = BF'$, we may obtain $\tilde{K}$ and $\tilde{B}$ from the eigenvalues and eigenvectors of $\hat{\Pi}\hat{\Pi}'$. Specifically, $\tilde{K}$ is given by

$$\tilde{K} = \sum_{j=1}^{Np} 1\{\lambda_j(\hat{\Pi}\hat{\Pi}') \geq \delta_{NT}\}.$$

If $\tilde{K} = 0$, $\tilde{B} = 0$ and $\tilde{F} = 0$; otherwise we proceed as follows. To estimate B, we use the following normalization: $B'B/N = I_K$ and $F'F/T$ being diagonal with diagonal entries in descending order. Then the columns of $\tilde{B}/\sqrt{N}$ are given by the eigenvectors of $\hat{\Pi}\hat{\Pi}'$ corresponding to its largest $\tilde{K}$ eigenvalues. Since $F = \Pi' B(B'B)^{-1}$, we thus obtain

$$\tilde{F} = \frac{\hat{\Pi}'\tilde{B}}{N}.$$

Perhaps the most natural approach to incorporate the reduced rank structure is to enforce the rank constraint directly. This leads to the following minimization problem

$$\min_{c_i \in \mathbf{R}^p, D_i \in \mathbf{R}^{p\times K}, g_t \in \mathbf{R}^K} \frac{1}{2}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - x_{it}'c_i - x_{it}'D_i g_t)^2. \tag{12}$$

However, solving (12) has at least two challenges.[5] First, it requires the knowledge of K, which has to be estimated prior to solving the problem. Second, (12) is nonconvex, and the solutions do not have an analytical closed form. These make it difficult not only to design a computing algorithm to find the solutions, but also to derive their asymptotic properties. One potential fix to the second challenge is alternating least squares, however it may suffer from non-convergence issues since the problem in (12) is nonconvex (Golub and Van Loan, 2013). In addition to the asymptotic properties that we derive in Sections 4 and 5, our estimators can be numerically solved for in an efficient way without the knowledge of K.

Remark 3.2. Our estimation procedure may accommodate the presence of missing values. In such cases, the double summations in (9) need to be replaced with summations over non-missing data. This amounts to redefining the observations as $y_{it}m_{it}$ and $x_{it}m_{it}$ and the error term as $\varepsilon_{it}m_{it}$, where $m_{it} = 0$ when y_{it} or x_{it} are missing and 1 otherwise.

4 Asymptotic Analysis

In this section, we conduct asymptotic analysis for our estimators in a general setup. Specifically, we establish consistency of $\hat{K}$ and a rate of convergence of $\hat{\Pi}$, $\hat{a}$, $\hat{B}$ and $\hat{F}$.

[5] Enforcing the constraints of a and B in Examples 2.2 and 2.4 does not resolve the challenges.

We begin by introducing the so-called restricted strong convexity condition (Negahban, Ravikumar, Wainwright, and Yu, 2012). It guarantees that the quadratic loss function in (9) is strictly convex over a restricted set of "low-rank" matrices. To describe the set, we define some notation. Let $\Pi = U\Sigma V'$ be a singular value decomposition of Π, where U and V are $Np \times Np$ and $T \times T$ orthonormal matrices and Σ is a diagonal matrix with singular values of Π in the diagonal in descending order. Write $U = (U_1, U_2)$ and $V = (V_1, V_2)$, where the columns of U_2 and V_2 are singular vectors corresponding to the zero singular values of Π. For any $Np \times T$ matrix Δ, let $\mathcal{P}(\Delta) \equiv U_2 U_2' \Delta V_2 V_2'$ and $\mathcal{M}(\Delta) \equiv \Delta - \mathcal{P}(\Delta)$. Heuristically, $\mathcal{M}(\Delta)$ can be thought of as the projection of Δ onto the "low-rank" space of Π, and $\mathcal{P}(\Delta)$ is the projection of Δ onto its orthogonal space. The restricted set of "low-rank" matrices is given by

$$\mathcal{C} \equiv \{\Delta \in \mathcal{S} \ominus \mathcal{S} : \|\mathcal{P}(\Delta)\|_* \leq 3\|\mathcal{M}(\Delta)\|_*\}, \tag{13}$$

where $\mathcal{S} \ominus \mathcal{S}$ is the Minkowski difference between $\mathcal{S}$ and $\mathcal{S}$, that is, $\mathcal{S} \ominus \mathcal{S} = \{\Gamma_1 - \Gamma_2 : \Gamma_1, \Gamma_2 \in \mathcal{S}\}$. We impose the restricted strong convexity condition as follows.

Assumption 4.1. *(i) Assume that $\Pi \in \mathcal{S} \subset \mathbf{R}^{Np \times T}$. For any $\Delta \in \mathcal{S} \ominus \mathcal{S}$, the following decomposition holds:*

$$\sum_{i=1}^{N} \sum_{t=1}^{T} |\mathrm{tr}(X_{it}' \Delta)|^2 = \mathcal{Q}_{NT}(\Delta) + \mathcal{L}_{NT}(\Delta)$$

such that for some constant $0 < \kappa < \infty$,

$$\mathcal{Q}_{NT}(\Delta) \geq \kappa \|\Delta\|_F^2 \text{ for all } \Delta \in \mathcal{C},$$

and for some $r_{NT} > 0$,

$$|\mathcal{L}_{NT}(\Delta)| \leq r_{NT} \|\Delta\|_* \text{ for all } \Delta \in \mathcal{S} \ominus \mathcal{S}.$$

(ii) The following condition holds:

$$\left| \sum_{i=1}^{N} \sum_{t=1}^{T} \mathrm{tr}(\varepsilon_{it} X_{it}' \Delta) \right| \leq \frac{1}{2} r_{NT} \|\Delta\|_* \text{ for all } \Delta \in \mathcal{S} \ominus \mathcal{S}.$$

Assumption 4.1 is weaker than the conditions of Corollary 1 in Negahban and Wainwright (2011), which require $\mathcal{S} = \mathbf{R}^{Np \times T}$ and $\mathcal{L}_{NT}(\cdot) = 0$, and are too restrictive in Examples 2.2 and 2.4. We refer to the condition: "$\mathcal{Q}_{NT}(\Delta) \geq \kappa \|\Delta\|_F^2$ for all $\Delta \in \mathcal{C}$" as the restricted strong convexity condition. Allowing $\mathcal{L}_{NT}(\cdot) \neq 0$ facilitates providing easy-to-verify primitive conditions for the restricted strong convexity condition. The rate r_{NT} plays an important role in determining the convergence rate of $\hat{\Pi}$, and thus determines how fast p and K can grow.

Assumption 4.2. *There exist some constants* $0 < d_{\min} \leq d_{\max} < \infty$ *such that: (i)* $d_{\min} < \lambda_{\min}(B'B/N) \leq \lambda_{\max}(B'B/N) < d_{\max}$ *for large* N*; (ii)* $\max_{t\leq T} \|f_t\| < d_{\max}$*; (iii)* $\lambda_{\min}(F'M_TF/T) > d_{\min}$*; (iv)* $a'a/N < d_{\max}$*; (v)* $a'B = 0$*.*

For simplicity of presentation, we assume that $\{a_i, B_i\}_{i\leq N}$ and $\{f_t\}_{t\leq T}$ are nonrandom. That is, all stochastic statements are implicitly conditional on their realization. Assumption 4.2(i) is similar to the pervasive condition in Stock and Watson (2002) and Bai and Ng (2002), which requires that f_t are strong factors. Assumptions 4.2(iv) and (v) are standard in the literature; see, for example, Chen et al. (2021). Assumptions 4.1 and 4.2 consist of high-level conditions; in Section 5, we provide simple primitive conditions for each of Examples 2.1-2.5.

Theorem 4.1. *Suppose Assumption 4.1 holds. Let* $\hat{\Pi}$*,* $\hat{K}$*,* $\hat{a}, \hat{B}$ *and* $\hat{F}$ *be given in* (9)-(11)*. Assume that* $0 < K < \min\{Np, T\} - 1$ *and* $\lambda_{NT} \geq 2r_{NT}$*. (i) Then*

$$\|\hat{\Pi} - \Pi\|_F \leq \frac{3\sqrt{2(K+1)}\lambda_{NT}}{\kappa}.$$

(ii) Suppose Assumption 4.2 additionally holds. Assume that $\delta_{NT}/(NT) \to 0$ *and* $\delta_{NT}/(K\lambda_{NT}^2) \to \infty$*. Let* $H \equiv (F'M_T\hat{F})(\hat{F}'M_T\hat{F})^{-1}$*. Then*

$$\begin{aligned} P(\hat{K} = K) &\to 1, \\ \|\hat{a} - a\| &= O_p\left(\frac{\sqrt{K}\lambda_{NT}}{\sqrt{T}}\right), \\ \|\hat{B} - BH\|_F &= O_p\left(\frac{\sqrt{K}\lambda_{NT}}{\sqrt{T}}\right), \\ \|\hat{F} - F(H')^{-1}\|_F &= O_p\left(\frac{\sqrt{K}\lambda_{NT}}{\sqrt{N}}\right). \end{aligned}$$

Theorem 4.1(i) is a deterministic statement on the estimation error of $\hat{\Pi}$; it extends Corollary 1 of Negahban and Wainwright (2011) by allowing for enforcing constraints of Π (i.e., $\mathcal{S} \neq \mathbf{R}^{Np\times T}$) in addition to the reduced rank constraint and $\mathcal{L}_{NT}(\cdot) \neq 0$. In some scenarios, Assumption 4.1(i) only holds with probability approaching one. In such cases, the result of Theorem 4.1(i) holds with probability approaching one, and the results of Theorem 4.1(ii) continues to hold. Due to the lack of identification, B and F can only be consistently estimated up to a rotational transformation, as usually occurred in high-dimensional factor analyses. The asymptotic results hold as $N \to \infty$ with fixed T or as $(N, T) \to \infty$, whenever appropriate. The rate r_{NT} determines the fastest convergence rate of the estimators. In Section 5, we specialize Theorem 4.1 for each of Examples 2.1-2.5. In all cases, p and K are allowed to grow with N or (N, T) for the consistency of the estimators, and the presence of missing values is allowed.

Remark 4.1. When $a = 0$, we can establish the same convergence rate for the restricted

estimators $\tilde{K}$, $\tilde{B}$ and $\tilde{F}$ in Remark 3.1 as in Theorem 4.1(ii). Let $G \equiv (F'\tilde{F})(\tilde{F}'\tilde{F})^{-1}$. If $a = 0$, under the same conditions of Theorem 4.1(ii), following the arguments in the proof of Theorem 4.1(ii), we can establish the following:

$$P(\tilde{K} = K) \to 1,$$
$$\|\tilde{B} - BG\|_F = O_p\left(\frac{\sqrt{K}\lambda_{NT}}{\sqrt{T}}\right),$$
$$\|\tilde{F} - F(G')^{-1}\|_F = O_p\left(\frac{\sqrt{K}\lambda_{NT}}{\sqrt{N}}\right).$$

Remark 4.2. Since $\hat{\Pi}$ allows for missing values, we may use a CV approach to choose the regularization parameter λ_{NT}. Specifically, we may randomly divide the observations into L folds with observations indexed by $\{\mathcal{I}_\ell\}_{\ell \le L}$, where $\mathcal{I}_\ell$ consists of observation indices in the ℓth fold, $\{\mathcal{I}_\ell\}_{\ell \le L}$ are mutually exclusive, and $\cup_{\ell \le L}\mathcal{I}_\ell = \mathcal{I} \equiv \{1, 2, \cdots, N\} \times \{1, 2, \cdots, T\}$. Rolling ℓ from 1 to L, we may leave observations $\{(y_{it}, x_{it}) : (i,t) \in \mathcal{I}_\ell\}$ out, use observations $\{(y_{it}, x_{it}) : (i,t) \in \mathcal{I}/\mathcal{I}_\ell\}$ for training, and calculate the out-of-sample mean square error MSE_ℓ for observations $\{(y_{it}, x_{it}) : (i,t) \in \mathcal{I}_\ell\}$. We may choose λ_{NT} by minimizing $\sum_{\ell=1}^{L} \text{MSE}_\ell / L$.

5 Revisiting Examples

In this section, we specialize Theorem 4.1 for each of Examples 2.1-2.5.

5.1 Examples 2.1, 2.3 and 2.5

In these examples, our goal is to estimate K, a, B and F. There is no restriction on a or B for us to impose. Thus, we may set $\mathcal{S} = \mathbf{R}^{Np \times T}$ in (9). Hence, $\mathcal{S} \ominus \mathcal{S} = \mathbf{R}^{Np \times T}$. By the fact that $|\text{tr}(C'D)| \le \|C\|_2 \|D\|_*$[6], for any $\Delta \in \mathcal{S} \ominus \mathcal{S}$,

$$\left|\sum_{i=1}^{N}\sum_{t=1}^{T} \text{tr}(\varepsilon_{it} X_{it}' \Delta)\right| \le \left\|\sum_{i=1}^{N}\sum_{t=1}^{T} X_{it}\varepsilon_{it}\right\|_2 \|\Delta\|_*. \tag{14}$$

Thus, Assumption 4.1(ii) is satisfied with $r_{NT} = O_p(\max\{\sqrt{Np}, \sqrt{T}\})$ as $(N,T) \to \infty$, if $\{(x_{1t}'\varepsilon_{1t}, x_{2t}'\varepsilon_{2t}, \ldots, x_{Nt}'\varepsilon_{Nt})'\}_{t \le T}$ is a sequence of independent sub-Gaussian vectors; see Lemma C.1(i). When $x_{it} = 1$, Assumption 4.1(i) is trivially satisfied with $\mathcal{L}_{NT}(\cdot) = 0$ and $\kappa = 1$, and Assumption 4.2(i) reduces to the pervasive condition in Stock and Watson (2002). When $a = 0$, Assumptions 4.2(iv) and (v) are trivially satisfied.

We summarize the conditions in the following assumptions.

[6] See, for example, Fact 11.14.1 in Bernstein (2018).

Assumption 5.1. *(i) There exists some constant* $0 < \kappa < \infty$ *such that*

$$\sum_{i=1}^{N}\sum_{t=1}^{T}|\mathrm{tr}(X_{it}'\Delta)|^2 \geq \kappa\|\Delta\|_F^2 \text{ for all } \Delta \in \mathcal{D},$$

where $\mathcal{D} \equiv \{\Delta \in \mathbf{R}^{Np\times T} : \|\mathcal{P}(\Delta)\|_* \leq 3\|\mathcal{M}(\Delta)\|_*\}$. *(ii)* $\{(x_{1t}'\varepsilon_{1t}, x_{2t}'\varepsilon_{2t}, \ldots, x_{Nt}'\varepsilon_{Nt})'\}_{t\leq T}$ *is a sequence of independent sub-Gaussian vectors.*[7]

In the case when $x_{it} = 1$, we may have a simple analytical closed form for $\hat{\Pi}$. Let Y be an $N \times T$ matrix with the itth entry y_{it}. Let $Y = U\Sigma V'$ be a singular value decomposition of Y, where U and V are $N \times N$ and $T \times T$ orthonormal matrices and Σ is an $N \times T$ diagonal matrix with singular values $\sigma_j(Y)$'s in the diagonal in descending order. For $x > 0$, let Σ_x be an $N \times T$ diagonal matrix with $\max\{0, \sigma_j(Y) - x\}$ in descending order. Then $\hat{\Pi} = U\Sigma_{\lambda_{NT}/2}V'$; see, for example, Cai, Candés, and Shen (2010) and Ma et al. (2011). In general cases, an analytical closed form is not available; in Appendix B, we provide an efficient algorithm for finding $\hat{\Pi}$.

We may apply Theorem 4.1 to conclude the following corollary.

Corollary 5.1. *Suppose Assumption 5.1(ii) holds. Let* $\hat{\Pi}$, $\hat{K}$, $\hat{a}, \hat{B}$ *and* $\hat{F}$ *be given in (9)-(11) with* $\mathcal{S} = \mathbf{R}^{Np\times T}$ *and* $\lambda_{NT} = \sqrt{(Np+T)\log N}$. *Assume that* $0 < K < \min\{Np, T\} - 1$. *(i) If* $x_{it} = 1$ *or Assumption 5.1(i) holds, then as* $(N,T) \to \infty$,

$$\frac{1}{\sqrt{NT}}\|\hat{\Pi} - \Pi\|_F = O_p\left(\sqrt{\frac{K(Np+T)\log N}{NT}}\right).$$

(ii) Suppose Assumptions 4.2(i)-(iii) additionally hold. Assume that as $(N,T) \to \infty$, $\delta_{NT}/(NT) \to 0$ *and* $\delta_{NT}/[K(Np+T)\log N] \to \infty$. *Let* $H \equiv (F'M_T\hat{F})(\hat{F}'M_T\hat{F})^{-1}$. *If* $a = 0$ *or Assumptions 4.2(iv)-(v) hold, then as* $(N,T) \to \infty$,

$$\begin{aligned}
P(\hat{K} = K) &\to 1, \\
\frac{1}{\sqrt{N}}\|\hat{a} - a\| &= O_p\left(\sqrt{\frac{K(Np+T)\log N}{NT}}\right), \\
\frac{1}{\sqrt{N}}\|\hat{B} - BH\|_F &= O_p\left(\sqrt{\frac{K(Np+T)\log N}{NT}}\right), \\
\frac{1}{\sqrt{T}}\|\hat{F} - F(H')^{-1}\|_F &= O_p\left(\sqrt{\frac{K(Np+T)\log N}{NT}}\right).
\end{aligned}$$

Corollary 5.1 requires large N and large T. In particular, $K(Np+T)\log N = o(NT)$ is required for the consistency of the estimators. This implies that p is allowed to grow

[7] Independence is not necessary here and also in Assumptions 5.2(iv), (v) and 5.4(iii). We may allow for weak dependence over t; see Lemma C.1.

as $(N,T) \to \infty$. In the presence of $a = 0$, we can establish the same convergence rate for the restricted estimators $\tilde{K}$, $\tilde{B}$ and $\tilde{F}$ in Remark 3.1; see Remark 4.1. While the result for Example 2.1 is known in the literature, the results for Examples 2.3 and 2.5 are new. Distinct from Pelger and Xiong (2021), we provide an estimator that can consistently estimate F up to a common rotational transformation, which is not state-specific. That is, we are able to consistently estimate the factor space. Moreover, large p is allowed. In contrast to Gagliardini et al. (2016), we provide an estimation method that does not require observable f_t, and allows for the presence of arbitrage and large p. No estimation method for the unconstrained conditional factor model is available in the literature.

Remark 5.1. In the presence of missing values, when $x_{it} = 1$ and m_{it}'s are i.i.d. random variables across both i and t (i.e., missing at random), Assumption 5.1(i) requires $P(m_{it} = 1) > 0$ by the law of large numbers. Thus, the proportion of missing values cannot be too large. This is analogous to the requirement in the matrix completion literature (Candés and Recht, 2009; Recht et al., 2010; Candés and Plan, 2010; Negahban and Wainwright, 2012) that the missing pattern is not too systematic and the proportion of missing values is not too large.

5.2 Example 2.2

Our goal is to estimate K, $\mu \equiv (\mu_1, \mu_2, \ldots, \mu_N)'$, $\Lambda \equiv (\lambda_1, \lambda_2, \ldots, \lambda_N)'$, ϕ, Φ and F. To apply the general framework, we fist observe that $a_i = (\mu_i, \phi')'$ and $B_i = (\lambda_i, \Phi')'$ for all i, that is, the rows of a_i and B_i corresponding to the nonconstant part of x_{it} are homogenous across i. Let $\pi_i \equiv \mu_i 1_T + F\lambda_i$ and $\Pi^* \equiv \phi 1_T' + \Phi F'$, which are $T \times 1$ vector and $(p-1) \times T$ matrix, then $\Pi = ((\pi_1, \Pi^{*\prime}), (\pi_2, \Pi^{*\prime}), \ldots, (\pi_N, \Pi^{*\prime}))'$. Thus, we may set

$$\mathcal{S} = \mathcal{D}_M \equiv \left\{ \begin{pmatrix} \gamma_1' \\ \Gamma^* \\ \gamma_2' \\ \Gamma^* \\ \vdots \\ \gamma_N' \\ \Gamma^* \end{pmatrix} : \begin{pmatrix} \gamma_1' \\ \gamma_2' \\ \vdots \\ \gamma_N' \end{pmatrix} \in \mathbf{R}^{N\times T}, \Gamma^* \in \mathbf{R}^{(p-1)\times T} \text{ and } \|\Gamma^*\|_{\max} \le M \right\} \tag{15}$$

for $0 < M < \infty$ in (9), where $\|\Gamma^*\|_{\max}$ denotes the largest absolute value of the entries of Γ^*.[8] Clearly, $\mathcal{D}_M$ is convex in $\mathbf{R}^{Np\times T}$, and $\mathcal{S} \ominus \mathcal{S} = \mathcal{D}_{2M}$.

Next, we verify Assumptions 4.1 and 4.2. Write $x_{it} = (1, x_{it}^{*\prime})'$. By Lemma C.4, for

[8] Imposing $\|\Gamma^*\|_\infty \le M$ facilitates providing easy-to-verify primitive conditions for Assumption 4.1(i).

any $\Delta \in \mathcal{S} \ominus \mathcal{S}$, there exists $\mathcal{R}_{NT}(\cdot)$ such that

$$\sum_{i=1}^{N}\sum_{t=1}^{T} |\mathrm{tr}(X_{it}'\Delta)|^2 \geq \min\left\{1, \min_{t\leq T} \lambda_{\min}\left(\frac{\sum_{i=1}^{N} x_{it}^* x_{it}^{*\prime}}{N}\right)\right\} \|\Delta\|_F^2 + 2\mathcal{R}_{NT}(\Delta), \tag{16}$$

$$|\mathcal{R}_{NT}(\Delta)| \leq 2M\sqrt{p-1}\left\|\begin{pmatrix} x_{11}^* & x_{12}^* & \cdots & x_{1T}^* \\ x_{21}^* & x_{22}^* & \cdots & x_{2T}^* \\ \vdots & \vdots & \vdots & \vdots \\ x_{N1}^* & x_{N2}^* & \cdots & x_{NT}^* \end{pmatrix}\right\|_2 \|\Delta\|_*, \tag{17}$$

and

$$\left|\sum_{i=1}^{N}\sum_{t=1}^{T} \mathrm{tr}(\varepsilon_{it}X_{it}'\Delta)\right| \leq \left(\begin{array}{l} \left\|\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} x_{i1}^*\varepsilon_{i1}, \frac{1}{\sqrt{N}}\sum_{i=1}^{N} x_{i2}^*\varepsilon_{i2}, \ldots, \frac{1}{\sqrt{N}}\sum_{i=1}^{N} x_{iT}^*\varepsilon_{iT}\right)\right\|_2 \\ \quad + \left\|\begin{pmatrix} \varepsilon_{11} & \varepsilon_{12} & \cdots & \varepsilon_{1T} \\ \varepsilon_{21} & \varepsilon_{22} & \cdots & \varepsilon_{2T} \\ \vdots & \vdots & \vdots & \vdots \\ \varepsilon_{N1} & \varepsilon_{N2} & \cdots & \varepsilon_{NT} \end{pmatrix}\right\|_2 \end{array}\right) \|\Delta\|_*. \tag{18}$$

In view of (16), (17) and Lemma C.1(ii), if $\min_{t\leq T} \lambda_{\min}(\sum_{i=1}^{N} x_{it}^* x_{it}^{*\prime}/N) \geq c_{\min}$ for some constant $0 < c_{\min} < \infty$ and $\{(x_{1t}^{*\prime}, x_{2t}^{*\prime}, \ldots, x_{Nt}^{*\prime})'\}_{t\leq T}$ is a sequence of independent sub-Gaussian vectors, then Assumption 4.1(i) is satisfied with $\mathcal{L}_{NT}(\cdot) = 2\mathcal{R}_{NT}(\cdot)$, $\kappa = \min\{1, c_{\min}\}$, and $r_{NT} = O_p(M\sqrt{p}\max\{\sqrt{Np}, \sqrt{T}\})$ as $(N,T) \to \infty$. The first condition, which can be verified using the law of large numbers, often holds with probability approaching one as $N \to \infty$. As discussed below Theorem 4.1, this is sufficient for us to establish a rate of convergence of $\hat{\Pi}$. In view of (18) and Lemmas C.1(i), Assumption 4.1(ii) is satisfied if $\{(\varepsilon_{1t}, \varepsilon_{2t}, \ldots, \varepsilon_{Nt})'\}_{t\leq T}$ is a sequence of independent sub-Gaussian vectors and $\|\sum_{i=1}^{N} x_{it}^*\varepsilon_{it}/\sqrt{Np}\|$ has bounded second moment. In addition, it is easy to see that Assumptions 4.2(i), (iv) and (v) hold, if $\lambda_{\min}(\Phi'\Phi) + \lambda_{\min}(\Lambda'\Lambda/N) > d_{\min}$, $\lambda_{\max}(\Phi'\Phi) < d_{\max}/2$, $\lambda_{\max}(\Lambda'\Lambda/N) < d_{\max}/2$, $\|\phi\|^2 < d_{\max}/2$ and $\|\mu\|^2/N < d_{\max}/2$ for some constants $0 < d_{\min} \leq d_{\max} < \infty$, $\phi'\Phi = 0$, and $\mu'\Lambda = 0$.

We summarize the conditions in the following assumptions.

Assumption 5.2. *(i) There is a constant $0 < c_{\min} < \infty$ such that: with probability approaching one as $(N,T) \to \infty$,*

$$\min_{t\leq T} \lambda_{\min}\left(\frac{1}{N}\sum_{i=1}^{N} x_{it}^* x_{it}^{*\prime}\right) \geq c_{\min}.$$

(ii) Assume that $\max_{t\leq T}\|\phi+\Phi f_t\|_\infty$ *is bounded. (iii)* $\max_{t\leq T}E[\|\sum_{i=1}^N x_{it}^*\varepsilon_{it}/\sqrt{Np}\|^2]$ *is bounded. (iv)* $\{(x_{1t}^{*\prime},x_{2t}^{*\prime},\ldots,x_{Nt}^{*\prime})'\}_{t\leq T}$ *is a sequence of independent sub-Gaussian vectors. (v)* $\{(\varepsilon_{1t},\varepsilon_{2t},\ldots,\varepsilon_{Nt})'\}_{t\leq T}$ *is a sequence of independent sub-Gaussian vectors.*

Assumption 5.3. *There are constants* $0<d_{\min}\leq d_{\max}<\infty$ *such that: (i)* $\lambda_{\min}(\Phi'\Phi+\Lambda'\Lambda/N)>d_{\min}$*; (ii)* $\lambda_{\max}(\Phi'\Phi)<d_{\max}/2$*; (iii)* $\lambda_{\max}(\Lambda'\Lambda/N)<d_{\max}/2$*; (iv)* $\max_{t\leq T}\|f_t\|<d_{\max}$*; (v)* $\lambda_{\min}(F'M_TF/T)>d_{\min}$*; (vi)* $\|\phi\|^2<d_{\max}/2$*; (vii)* $\|\mu\|^2/N<d_{\max}/2$*; (viii)* $\phi'\Phi=0$*; (ix)* $\mu'\Lambda=0$*.*

Since $\hat{\Pi}\in\mathcal{S}$, we may write $\hat{\Pi}=((\hat{\pi}_1,\hat{\Pi}^{*\prime}),(\hat{\pi}_2,\hat{\Pi}^{*\prime}),\ldots,(\hat{\pi}_N,\hat{\Pi}^{*\prime}))'$, where $\hat{\pi}_i$ is a $T\times 1$ vector and $\hat{\Pi}^*$ is a $(p-1)\times T$ matrix. By Lemma C.7 (iv), we may write $\hat{B}=((\hat{\lambda}_1,\hat{\Phi}'),(\hat{\lambda}_2,\hat{\Phi}'),\ldots,(\hat{\lambda}_N,\hat{\Phi}'))'$, where $\hat{\lambda}$ is a $\hat{K}\times 1$ vector and $\hat{\Phi}$ is a $(p-1)\times\hat{K}$ matrix. Given these, by simple algebra we may also write $\hat{a}=((\hat{\mu}_1,\hat{\phi}'),(\hat{\mu}_2,\hat{\phi}'),\ldots,(\hat{\mu}_N,\hat{\phi}'))'$, where $\hat{\mu}_i$ is a scalar and $\hat{\phi}$ is a $(p-1)\times 1$ vector. Thus, we may define the estimators of $\Pi^\diamond\equiv(\pi_1,\pi_2,\ldots,\pi_N)'$, Π^*, μ, Λ, ϕ and Φ as $\hat{\Pi}^\diamond\equiv(\hat{\pi}_1,\hat{\pi}_2,\ldots,\hat{\pi}_N)'$, $\hat{\Pi}^*$, $\hat{\mu}\equiv(\hat{\mu}_1,\hat{\mu}_2,\ldots,\hat{\mu}_N)'$, $\hat{\Lambda}\equiv(\hat{\lambda}_1,\hat{\lambda}_2,\ldots,\hat{\lambda}_N)'$, $\hat{\phi}$ and $\hat{\Phi}$. This implies that there is no need to enforce the homogeneity restriction of a and B in extracting $\hat{a}$ and $\hat{B}$ from $\hat{\Pi}$ to ensure the same homogeneity structure of $\hat{a}$ and $\hat{B}$.

We may simplify the computation of the estimators by plugging in the homogeneity restriction. By Lemma C.5, $\hat{\Pi}^\diamond$ and $\hat{\Pi}^*$ can be equivalently obtained as follows

$$\{\hat{\Pi}^\diamond,\hat{\Pi}^*\}=\underset{\substack{\Gamma^\diamond=(\gamma_{it})_{i\leq N,t\leq T}\in\mathbf{R}^{N\times T}\\ \Gamma^*=(\gamma_1^*,\ldots,\gamma_T^*)\in\mathbf{R}^{(p-1)\times T}\\ \|\Gamma^*\|_{\max}\leq M}}{\arg\min}\frac{1}{2}\sum_{i=1}^N\sum_{t=1}^T(y_{it}-\gamma_{it}-x_{it}^{*\prime}\gamma_t^*)^2+\lambda_{NT}\left\|\begin{pmatrix}\Gamma^\diamond\\ \sqrt{N}\Gamma^*\end{pmatrix}\right\|_*. \quad (19)$$

This equivalence has greatly simplified the computation of $\hat{\Pi}$, since (9) involves an $Np\times T$ unknown matrix while (19) involves two unknown matrices with relatively smaller sizes. In Appendix B, we provide an efficient algorithm for finding $\hat{\Pi}^\diamond$ and $\hat{\Pi}^*$. By Lemma C.7(ii) and (iv), $\hat{K}$ can be equivalently obtained as follows

$$\hat{K}=\sum_{j=1}^T 1\{\lambda_j(M_T(\hat{\Pi}^{\diamond\prime}\hat{\Pi}^\diamond+N\hat{\Pi}^{*\prime}\hat{\Pi}^*)M_T)\geq\delta_{NT}\}, \quad (20)$$

and $(\hat{\Lambda}'/\sqrt{N},\hat{\Phi}')'$ as the left singular vector of $(\hat{\Pi}^{\diamond\prime},\sqrt{N}\hat{\Pi}^{*\prime})'M_T$ corresponding to its largest $\hat{K}$ singular values. Moreover, it is straightforward to show that

$$\begin{aligned}\hat{\mu}&=\left(I_N-\frac{\hat{\Lambda}\hat{\Lambda}'}{N}\right)\frac{\hat{\Pi}^{\diamond\prime}1_T}{T}-\hat{\Lambda}\hat{\Phi}'\frac{\hat{\Pi}^{*\prime}1_T}{T},\\ \hat{\phi}&=(I_{p-1}-\hat{\Phi}\hat{\Phi}')\frac{\hat{\Pi}^{*\prime}1_T}{T}-\frac{\hat{\Phi}\hat{\Lambda}'}{N}\frac{\hat{\Pi}^\diamond 1_T}{T},\\ \hat{F}&=\frac{\hat{\Pi}^{\diamond\prime}\hat{\Lambda}}{N}+\hat{\Pi}^{*\prime}\hat{\Phi}.\end{aligned} \quad (21)$$

Finally, we apply Theorem 4.1 to obtain the consistency of $\hat{K}$ and a rate of convergence of $\hat{\Pi}^{\diamond}$, $\hat{\Pi}^{*}$,$\hat{\mu}$, $\hat{\Lambda}$, $\hat{\phi}$, $\hat{\Phi}$ and $\hat{F}$, as summarized in the following corollary.

Corollary 5.2. *Suppose Assumption 5.2 holds. Let* $\hat{\Pi}^{\diamond}$, $\hat{\Pi}^{*}$, $\hat{K}$, $\hat{\mu}$, $\hat{\Lambda}$, $\hat{\phi}, \hat{\Phi}$ *and* $\hat{F}$ *be given in* (19)-(21) *with* $\lambda_{NT} = M[\sqrt{(Np^2+Tp)\log N}]$. *Assume that* $0 < K < \min\{N+p-1, T\}-1$. *(i) Then as* $(N,T) \to \infty$,

$$\frac{1}{\sqrt{NT}}\|\hat{\Pi}^{\diamond} - \Pi^{\diamond}\|_F = O_p\left(M\sqrt{\frac{K(Np^2+Tp)\log N}{NT}}\right),$$
$$\frac{1}{\sqrt{T}}\|\hat{\Pi}^{*} - \Pi^{*}\|_F = O_p\left(M\sqrt{\frac{K(Np^2+Tp)\log N}{NT}}\right).$$

(ii) Suppose Assumption 5.3 additionally holds. Assume that as $(N,T) \to \infty$, $\delta_{NT}/(NT) \to 0$ *and* $\delta_{NT}/[M^2K(Np^2+Tp)\log N] \to \infty$. *Let* $H \equiv (F'M_T\hat{F})(\hat{F}'M_T\hat{F})^{-1}$. *Then as* $(N,T) \to \infty$,

$$P(\hat{K} = K) \to 1,$$
$$\frac{1}{\sqrt{N}}\|\hat{\mu} - \mu\| = O_p\left(M\sqrt{\frac{K(Np^2+Tp)\log N}{NT}}\right),$$
$$\frac{1}{\sqrt{N}}\|\hat{\Lambda} - \Lambda H\|_F = O_p\left(M\sqrt{\frac{K(Np^2+Tp)\log N}{NT}}\right),$$
$$\|\hat{\phi} - \phi\| = O_p\left(M\sqrt{\frac{K(Np^2+Tp)\log N}{NT}}\right),$$
$$\|\hat{\Phi} - \Phi H\|_F = O_p\left(M\sqrt{\frac{K(Np^2+Tp)\log N}{NT}}\right),$$
$$\frac{1}{\sqrt{T}}\|\hat{F} - F(H')^{-1}\|_F = O_p\left(M\sqrt{\frac{K(Np^2+Tp)\log N}{NT}}\right).$$

The rate in Corollary 5.2 is slower than that in Corollary 5.1. This is because the former relies on a set of easy-to-verify conditions—Assumption 5.2—rather than a version of Assumption 5.1. The rate can be improved to $O_p(\sqrt{K(N+p+T)\log N/(NT)})$ under Assumption 5.1. Our result is distinct from Fan et al. (2016) in several aspects. First, we allow for $\mu_i \neq 0$ and $\phi \neq 0$. This appears important in asset pricing, as they allow us to learn pricing errors. Second, we allow x_{it} to vary over t. This appears crucial in asset pricing, since many stock characteristics (e.g., book to market ratio and momentum) change from month to month. Third, we do not require that λ_i has zero mean and weak cross-sectional dependence (so λ_i can be interpreted as a vector of noises), which is barely justified in practice. We allow for non-noisy intercepts μ_i and λ_i in pricing errors and risk exposures. Fourth, we allow $K \to \infty$. In addition, our result extends

Chen et al. (2021) by allowing for the heterogeneity of μ_i and λ_i across i.

Remark 5.2. In addition, we may extract additional estimators for K, μ, Λ, ϕ, Φ and F from $\hat{\Pi}^{\diamond}$ and $\hat{\Pi}^{*}$ separately. First, since $\Pi^{\diamond}M_T = \Lambda F' M_T$, we may extract estimators for K, μ, Λ and F from $\hat{\Pi}^{\diamond}$ in a similar way to (10) and (11). Second, similarly, since $\Pi^{*}M_T = \Phi F' M_T$, we may extract estimators for K, ϕ, Φ and F from $\hat{\Pi}^{*}$. These estimators are different from $\hat{K}$, $\hat{\mu}$, $\hat{\Lambda}$, $\hat{\phi}$, $\hat{\Phi}$ and $\hat{F}$ in Corollary 5.2. However, following the arguments in the proof of Theorem 4.1(ii), we can establish the consistency and the same convergence rate for the estimators; the details are omitted.

Remark 5.3. In the presence of missing values, we need to modify Assumption 5.2. In particular, (16) is not true, but may hold with probability approaching one by replacing $\min\{1, \min_{t\le T}\lambda_{\min}(\sum_{i=1}^{N} x_{it}^{*}x_{it}^{*\prime}/N)\}$ with $\min\{q, \min_{t\le T}\lambda_{\min}(\sum_{i=1}^{N} m_{it}x_{it}^{*}x_{it}^{*\prime}/N)\}$ under a missing-at-random mechanism where q is the probability of nonmissing (i.e., m_{it}'s are i.i.d random variables across both i and t with $P(m_{it}=1)=q$, independent of $\{x_{it}^{*}, \varepsilon_{it}\}_{i\le N, t\le T}$). Thus, the proportion of missing values cannot be too large.

5.3 Example 2.4

Our goal is to estimate K, ϕ, Φ and F. To apply the general framework, we first observe that $a_i = \phi$ and $B_i = \Phi$ for all i, that is, a_i and B_i are homogenous across i. Let $\Pi_0 \equiv \phi 1_T' + \Phi F'$, which is a $p \times T$ matrix, then $\Pi = 1_N \otimes \Pi_0$. Thus, we may set $\mathcal{S} = \{1_N \otimes \Gamma : \Gamma \in \mathbf{R}^{p\times T}\}$ in (9). Clearly, $\mathcal{S}$ is convex in $\mathbf{R}^{Np\times T}$, and $\mathcal{S} \ominus \mathcal{S} = \mathcal{S}$.

Next, we verify Assumptions 4.1 and 4.2. By Lemma C.2, for any $\Delta \in \mathcal{S} \ominus \mathcal{S}$,

$$\sum_{i=1}^{N}\sum_{t=1}^{T} |\mathrm{tr}(X_{it}'\Delta)|^2 \le \min_{t\le T}\lambda_{\min}\left(\frac{\sum_{i=1}^{N} x_{it}x_{it}'}{N}\right)\|\Delta\|_F^2 \tag{22}$$

and

$$\left|\sum_{i=1}^{N}\sum_{t=1}^{T}\mathrm{tr}(\varepsilon_{it}X_{it}'\Delta)\right| \le \left\|\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}x_{i1}\varepsilon_{i1}, \frac{1}{\sqrt{N}}\sum_{i=1}^{N}x_{i2}\varepsilon_{i2}, \ldots, \frac{1}{\sqrt{N}}\sum_{i=1}^{N}x_{iT}\varepsilon_{iT}\right)\right\|_2 \|\Delta\|_{*}. \tag{23}$$

In view of (22), if $\min_{t\le T}\lambda_{\min}(\sum_{i=1}^{N} x_{it}x_{it}'/N) \ge c_{\min}$ for some constant $0 < c_{\min} < \infty$, then Assumption 4.1(i) is satisfied with $\mathcal{L}_{NT}(\cdot) = 0$ and $\kappa = c_{\min}$. As mentioned in Section 5.2, the condition usually holds with probability approaching one as $N \to \infty$, and this is sufficient for us to establish a rate of convergence of $\hat{\Pi}$. In view of (23), Assumption 4.1(ii) is trivially satisfied with $r_{NT} = O_p(\sqrt{p})$ as $N \to \infty$ with fixed T, if $\|\sum_{i=1}^{N} x_{it}\varepsilon_{it}/\sqrt{Np}\|$ has bounded second moment. Alternatively, if $\{\sum_{i=1}^{N} x_{it}\varepsilon_{it}/\sqrt{N}\}_{t\le T}$ is a sequence of independent sub-Gaussian vectors, then Assumption 4.1(ii) is satisfied with $r_{NT} = O_p(\max\{\sqrt{p}, \sqrt{T}\})$ as $(N,T) \to \infty$; see Lemma C.1(iii). In addition, Assumptions 4.2(i), (iv) and (v) reduce to $d_{\min} < \lambda_{\min}(\Phi'\Phi) \le$

$\lambda_{\max}(\Phi'\Phi) < d_{\max}$, $\|\phi\|^2 < d_{\max}$ for some constants $0 < d_{\min} \leq d_{\max} < \infty$, and $\phi'\Phi = 0$. The same assumptions have been imposed in Chen et al. (2021).

We summarize the conditions in the following assumptions.

Assumption 5.4. *(i) There is a constant $0 < c_{\min} < \infty$ such that: with probability approaching one as $N \to \infty$ with fixed T or as $(N,T) \to \infty$,*

$$\min_{t \leq T} \lambda_{\min}\left(\frac{1}{N}\sum_{i=1}^{N} x_{it}x_{it}'\right) \geq c_{\min}.$$

(ii) $E[\|\sum_{i=1}^{N} x_{it}\varepsilon_{it}/\sqrt{Np}\|^2]$ is bounded for each $t \leq T$. (iii) $\{\sum_{i=1}^{N} x_{it}\varepsilon_{it}/\sqrt{N}\}_{t\leq T}$ is a sequence of independent sub-Gaussian vectors.

Assumption 5.5. *There are constants $0 < d_{\min} \leq d_{\max} < \infty$ such that: (i) $d_{\min} < \lambda_{\min}(\Phi'\Phi) \leq \lambda_{\max}(\Phi'\Phi) < d_{\max}$; (ii) $\max_{t\leq T}\|f_t\| < d_{\max}$; (iii) $\lambda_{\min}(F'M_TF/T) > d_{\min}$; (iv) $\|\phi\|^2 < d_{\max}$; (v) $\phi'\Phi = 0$.*

In fact, the model in (8) with $\Pi = 1_N \otimes \Pi_0$ can be alternatively viewed as a multivariate linear regression model with reduced rank coefficient matrix Π_0, which has at most rank $K+1$. Following Example 1 of Negahban and Wainwright (2011),[9] we may come up with the following nuclear norm regularized estimator

$$\hat{\Pi}_0 = \underset{\Gamma=(\gamma_1,\ldots,\gamma_T)\in\mathbf{R}^{p\times T}}{\arg\min} \frac{1}{2}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - x_{it}'\gamma_t)^2 + \lambda_{0,NT}\|\Gamma\|_*, \tag{24}$$

where $\lambda_{0,NT} > 0$ is a regularization parameter. It turns out that $\hat{\Pi} = 1_N \otimes \hat{\Pi}_0$ when $\lambda_{0,NT} = \sqrt{N}\lambda_{NT}$. This is because the objective function in (9) under $\mathcal{S} = \{1_N \otimes \Gamma : \Gamma \in \mathbf{R}^{p\times T}\}$ reduces to the one in (24) when $\lambda_{0,NT} = \sqrt{N}\lambda_{NT}$; see Lemma C.3. Thus, $\hat{\Pi}_0$ does not yield a different estimator. The equivalence between $\hat{\Pi}$ and $\hat{\Pi}_0$ has greatly simplified the computation of $\hat{\Pi}$, given that (9) involves an $Np \times T$ unknown matrix with constraints while (24) is an unconstrained problem with an unknown matrix of smaller size. In Appendix B, we provide an efficient computing algorithm for finding $\hat{\Pi}_0$.

Given $\hat{\Pi}_0$, we may alternatively estimate K, ϕ, Φ and F as follows. Denote the estimators by $\hat{K}_0$, $\hat{\phi}_0$, $\hat{\Phi}_0$ and $\hat{F}_0$. Since $\Pi_0 M_T = \Phi F' M_T$, we may obtain $\hat{K}_0$ and $\hat{\Phi}_0$ from the eigenvalues and eigenvectors of $\hat{\Pi}_0 M_T \hat{\Pi}_0'$. Specifically, $\hat{K}_0$ is given by

$$\hat{K}_0 = \sum_{j=1}^{p} 1\{\lambda_j(\hat{\Pi}_0 M_T \hat{\Pi}_0') \geq \delta_{0,NT}\}, \tag{25}$$

where $\delta_{0,NT} > 0$ is a threshold value. If $\hat{K}_0 = 0$, $\hat{\phi}_0 = \hat{\Pi}_0 1_T/T$, $\hat{\Phi}_0 = 0$ and $\hat{F}_0 = 0$; otherwise we proceed as follows. To estimate Φ, we use the following normalization:

[9] In fact, the model is different from their model since x_{it} is changing over t. We find that the nuclear norm regularization method continues to work.

$\Phi'\Phi = I_K$ and $F'M_TF/T$ being diagonal with diagonal entries in descending order. Then the columns of $\hat{\Phi}_0$ are given by the eigenvectors of $\hat{\Pi}_0 M_T \hat{\Pi}_0'$ corresponding to its largest $\hat{K}_0$ eigenvalues. Since $\phi = (I_p - \Phi(\Phi'\Phi)^{-1}\Phi')\Pi_0 1_T/T$ and $F = \Pi_0'\Phi(\Phi'\Phi)^{-1}$, we thus obtain

$$\hat{\phi}_0 = (I_p - \hat{\Phi}_0\hat{\Phi}_0')\frac{\hat{\Pi}_0 1_T}{T} \text{ and } \hat{F}_0 = \hat{\Pi}_0'\hat{\Phi}_0. \tag{26}$$

It turns out that $\hat{K} = \hat{K}_0$, $\hat{B} = 1_N \otimes \hat{\Phi}_0$, $\hat{a} = 1_N \otimes \hat{\phi}_0$, and $\hat{F} = \hat{F}_0$ when $\lambda_{0,NT} = \sqrt{N}\lambda_{NT}$ and $\delta_{0,NT} = \delta_{NT}/N$. The first result follows by Lemma C.6(ii), since $\hat{K}$ is equal to the number of "large" singular values of $1_N \otimes \hat{\Pi}_0 M_T$ while $\hat{K}_0$ is equal to the number of "large" singular values of $\hat{\Pi}_0 M_T$. The second result follows by Lemma C.6(iv), since $\hat{B}/\sqrt{N}$ is the left singular vector matrix of $1_N \otimes \hat{\Pi}_0 M_T$ corresponding to its largest $\hat{K}$ singular values while $\hat{\Phi}_0$ is the left singular vector matrix of $\hat{\Pi}_0 M_T$ corresponding to its largest $\hat{K}_0$ singular values. The remaining two follow trivially by simple algebras. Thus, $\hat{K}_0$, $\hat{\phi}_0$, $\hat{\Phi}_0$ and $\hat{F}_0$ do not yield a different estimator either. This implies that there is no need to enforce the homogeneity restriction of a and B in extracting $\hat{a}$ and $\hat{B}$ from $\hat{\Pi}$ to ensure the same homogeneity structure of $\hat{a}$ and $\hat{B}$.

Finally, we may immediately obtain the consistency of $\hat{K}_0$ and a rate of convergence of $\hat{\Pi}_0$, $\hat{\phi}_0, \hat{\Phi}_0$ and $\hat{F}_0$ from Theorem 4.1, as summarized in the following corollary.

Corollary 5.3. *Suppose Assumptions 5.4(i) and (ii) hold. Let $\hat{\Pi}_0$, $\hat{K}_0$, $\hat{\phi}_0, \hat{\Phi}_0$ and $\hat{F}_0$ be given in (24)-(26) with $\lambda_{0,NT} = \sqrt{N(p+T)\log N}$. Assume $0 < K < \min\{p, T\} - 1$. (i) Then as $N \to \infty$ with fixed T,*

$$\frac{1}{\sqrt{T}}\|\hat{\Pi}_0 - \Pi_0\|_F = O_p\left(\sqrt{\frac{K(p+T)\log N}{NT}}\right).$$

(ii) Suppose Assumption 5.5 additionally holds. Assume that as $N \to \infty$ with fixed T, $\delta_{0,NT}/T \to 0$ and $N\delta_{0,NT}/[K(p+T)\log N] \to \infty$. Let $H \equiv (F'M_T\hat{F}_0)(\hat{F}_0'M_T\hat{F}_0)^{-1}$. Then as $N \to \infty$ with fixed T,

$$\begin{aligned} P(\hat{K}_0 = K) &\to 1, \\ \|\hat{\phi}_0 - \phi\| &= O_p\left(\sqrt{\frac{K(p+T)\log N}{NT}}\right), \\ \|\hat{\Phi}_0 - \Phi H\|_F &= O_p\left(\sqrt{\frac{K(p+T)\log N}{NT}}\right), \\ \frac{1}{\sqrt{T}}\|\hat{F}_0 - F(H')^{-1}\|_F &= O_p\left(\sqrt{\frac{K(p+T)\log N}{NT}}\right). \end{aligned}$$

(iii) If Assumption 5.4(ii) is replaced with Assumption 5.4(iii), then (i) and (ii) continue

to hold by replacing "as $N \to \infty$ with fixed T" with "as $(N,T) \to \infty$" in all places.

Corollary 5.3 shows that $\hat{\Pi}_0$, $\hat{K}_0$, $\hat{\phi}_0, \hat{\Phi}_0$ and $\hat{F}_0$ are consistent either under large N with fixed T or large N and large T. In particular, $K(p+T)\log N = o(NT)$ is required for the consistency. This implies that p can be as large as N. This is a significant improvement from the similar results in Theorems 4.2 and 6.1 of Chen et al. (2021), which require that p grows at a rate slower than $N^{1/3}$. In addition, we allow for $K \to \infty$ and weak cross-sectional dependence of x_{it}.

Remark 5.4. Corollary 5.3 allows for arbitrage exogenous missing pattern in the presence of missing values (i.e., $\{m_{it}\}_{i\leq N,t\leq T}$, independent of $\{x_{it},\varepsilon_{it}\}_{i\leq N,t\leq T}$, follow an arbitrary distribution). Let N_t be the number of available observations for each $t \leq T$. Assumption 5.4 reduces to the one by replacing summation/average over all i with summation/average over i with observations, which may hold as $\min_{t\leq T} N_t \to \infty$ with fixed T or as $(\min_{t\leq T} N_t, T) \to \infty$ under arbitrage exogenous missing pattern.

6 Simulation Studies

In this section, we conduct Monte Carlo simulations to investigate the finite sample performance of our estimators in Section 5.

We consider three different data generating processes (DGPs), which correspond to the settings in Examples 2.2, 2.4 and 2.5. In all three DGPs, we let

$$x_{it,1} = \sigma_t u_{it,1}, x_{it,2} = 0.3x_{i(t-1),2} + u_{it,2}, x_{it,3} = u_{it,3} \text{ and } x_{it,4} = 1, \tag{27}$$

where $u_{it} = (u_{it,1}, u_{it,2}, u_{it,3})'$ are i.i.d. $N(0, I_3)$ across both i and t, σ_t's are i.i.d. $U(1,2)$ over t, and $x_{i0,2}$'s are i.i.d. $N(0,1)$ across i. Let $x_{it} = (x_{it,1}, x_{it,2}, x_{it,3}, x_{it,4})'$, so $p = 4$. Let $f_t = 0.3f_{t-1} + \eta_t$, where η_t's are i.i.d. $N(1_2, I_2)$ over t and $f_0 \sim N(1_2/0.7, I_2/0.91)$, so $K = 2$. Let ε_{it}'s be i.i.d. $N(0,4)$ across both i and t. In the first DGP (denoted DGP1), we assume

$$a_i = \begin{pmatrix} 1 \\ \theta_i \\ 0 \\ 0 \end{pmatrix} \text{ and } B_i = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 2 & 0 \\ 0 & \delta_i \end{pmatrix}, \tag{28}$$

where θ_i's are i.i.d. $N(0,1)$ across i and δ_i's are i.i.d. $U(1,3)$ across i. In this DGP, a_i and B_i are heterogenous across i, which is the setting in Example 2.5. In the second

DGP (denoted DGP2), we assume

$$a_i = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix} \text{ and } B_i = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 2 & 0 \\ 0 & \delta_i \end{pmatrix}, \tag{29}$$

where δ_i's are i.i.d. $U(1,3)$ across i. In this DGP, the rows of a_i and B_i corresponding to the nonconstant part of x_{it} are homogenous across i, which is the setting in Example 2.2. In particular, $\mu = (0, 0, \ldots, 0)'$, $\phi = (1, 1, 0)'$,

$$\Lambda = \begin{pmatrix} 0 & \delta_1 \\ 0 & \delta_2 \\ \vdots & \vdots \\ 0 & \delta_N \end{pmatrix} \text{ and } \Phi = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 2 & 0 \end{pmatrix}. \tag{30}$$

In the third DGP (denoted DGP3), we assume

$$a_i = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix} \text{ and } B_i = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 2 & 0 \\ 0 & 2 \end{pmatrix}. \tag{31}$$

In this DGP, a_i and B_i are homogenous across i, which is the setting in Example 2.4. In particular, $\phi = a_1$ and $\Phi = B_1$. Here, u_{it}'s, σ_t's, $x_{i0,2}$'s, θ_i's, δ_i's, η_t's, f_0 and ε_{it}'s are mutually independent. We generate y_{it} according to the model in (1).

For DGP1, we implement the estimators in (9)-(11) with $\mathcal{S} = \mathbf{R}^{Np \times T}$. We evaluate the performance of $\hat{\Pi}$, $\hat{K}$, $\hat{a}$, $\hat{B}$ and $\hat{F}$. By Corollary 5.1, we let $\lambda_{NT} = c\sqrt{(Np+T)\log N}$ and $\delta_{NT} = 2(Np+T)\log N$ for some $c > 0$. For DGP2, we implement the estimators in (19)-(21). We evaluate the performance of $\hat{\Pi}^{\diamond}$, $\hat{\Pi}^{*}$, $\hat{K}$, $\hat{\mu}$, $\hat{\Lambda}$, $\hat{\phi}$, $\hat{\Phi}$ and $\hat{F}$. By Corollary 5.2, we let $\lambda_{NT} = c\sqrt{(Np+T)\log N}$ and $\delta_{NT} = 2(Np+T)\log N$ for some $c > 0$. For DGP3, implement the estimators in (24)-(26). We evaluate the performance of $\hat{\Pi}_0$, $\hat{K}_0$, $\hat{\phi}_0$, $\hat{\Phi}_0$ and $\hat{F}_0$. By Corollary 5.3, we let $\lambda_{0,NT} = c\sqrt{N(p+T)\log N}$ and $\delta_{0,NT} = 2(p+T)\log(N)/\sqrt{N}$ for some $c > 0$.

In order to choose the value of c, we first evaluate the performance of the 5-fold CV approach as described in Remark 4.2 (that is, $L = 5$). Figures 1-3 report the mean square errors of the regularized estimators ($\hat{\Pi}$, $(\hat{\Pi}^{\diamond\prime}, \sqrt{N}\hat{\Pi}^{*\prime})$, and $\hat{\Pi}_0$) by using fixed values of c and the CV method, where c is confined to $[0, 2]$.[10] All simulation results of this section are based on 200 simulation replications. The main findings are summarized as follows. First, the nuclear norm regularization can significantly improve

[10] Specifically, we consider the grid set $\{0, 0.05, 0.1, 0.2, \ldots, 0.9, 1, 1.5, 2\}$.

the performance of the estimators. As shown in Figures 1 and 2, the mean square error of the unregularized estimator (i.e., $c = 0$) does not decrease as both N and T increase (the value stays constantly around 20 in DGP1 and 10 in DGP2). In other words, the unregularized estimators may not be consistent. Using the nuclear norm regularization with a proper value of c (e.g., $c = 1$) not only reduces the mean square error for each combination of (N, T), but also shrinks the value towards zero as both N and T increase (e.g., the value for $c = 1$ is getting closer to zero as N and T increase). This suggests that the regularized estimators with a properly chosen value of c are consistent, which is in accordance with Corollaries 5.1 and 5.2. As shown in Figure 3, although the mean square error of the unregularized estimator is shrinking toward zero as N increases (note that the scale of the vertical axis changes across rows of graphs), the regularized estimator with a proper value of c (e.g., $c = 0.3$ or 0.4) enjoys a smaller mean square error. In sum, our simulations have demonstrated the important role played by the nuclear norm regularization. Second, in all three DGPs, the regularized estimators are very sensitive to the value of c. For example, as shown in Figure 3, choosing $c = 2$ can lead to a larger mean square error than that of the unregularized estimator in all combinations of (N, T). Therefore, it is important to choose the value of c in practice. Third, the CV approach works well in choosing the value of c toward minimizing the mean square error. In all three DGPs, the mean square error of the regularized estimator by using the CV chosen value of c is close to the smallest mean square error when using fixed values of c (the blue line almost hit the lowest point of the dash-dotted line in all graphs of Figures 1-3), regardless of the combination of (N, T).

We then investigate the performance of the estimators other than $\hat{\Pi}$, $\hat{\Pi}^{\diamond}$, $\hat{\Pi}^{*}$ and $\hat{\Pi}_0$ by using the CV chosen value of c. Tables I-III report their mean square errors or correct rates. The main findings are summarized as follows. First, the number factor estimators ($\hat{K}$ in DGP1 and DGP2 and $\hat{K}_0$ in DGP3) perform well in all cases (only two correct rates are below 100%). Second, all mean square errors in DGP1 and DGP2 decrease as both N and T increase, and all mean square errors in DGP3 decrease as N increases. This suggests that the estimators in DGP1 and DGP2 are consistent as $(N, T) \to \infty$, which is in accordance with Corollaries 5.1 and 5.2; the estimators in DGP3 are consistent as $N \to \infty$, which is in accordance with Corollary 5.3. Third, increasing N reduces the mean square errors of the factor estimators ($\hat{F}$ in DGP1 and DGP2 and $\hat{F}_0$ in DGP3) in all cases, while increasing T may not. In addition, increasing either N or T can reduce the mean square errors of $\hat{\Pi}^{*}$, $\hat{\phi}$, $\hat{\Phi}$, $\hat{\phi}_0$ and $\hat{\Phi}_0$. This is beyond what Corollaries 5.2 and 5.3 can explain, but may be due to the fact that the estimands (i.e., Π^{*}, ϕ, Φ, ϕ_0 and Φ_0) are homogenous parameters. In sum, our estimators have encouraging finite sample performance.

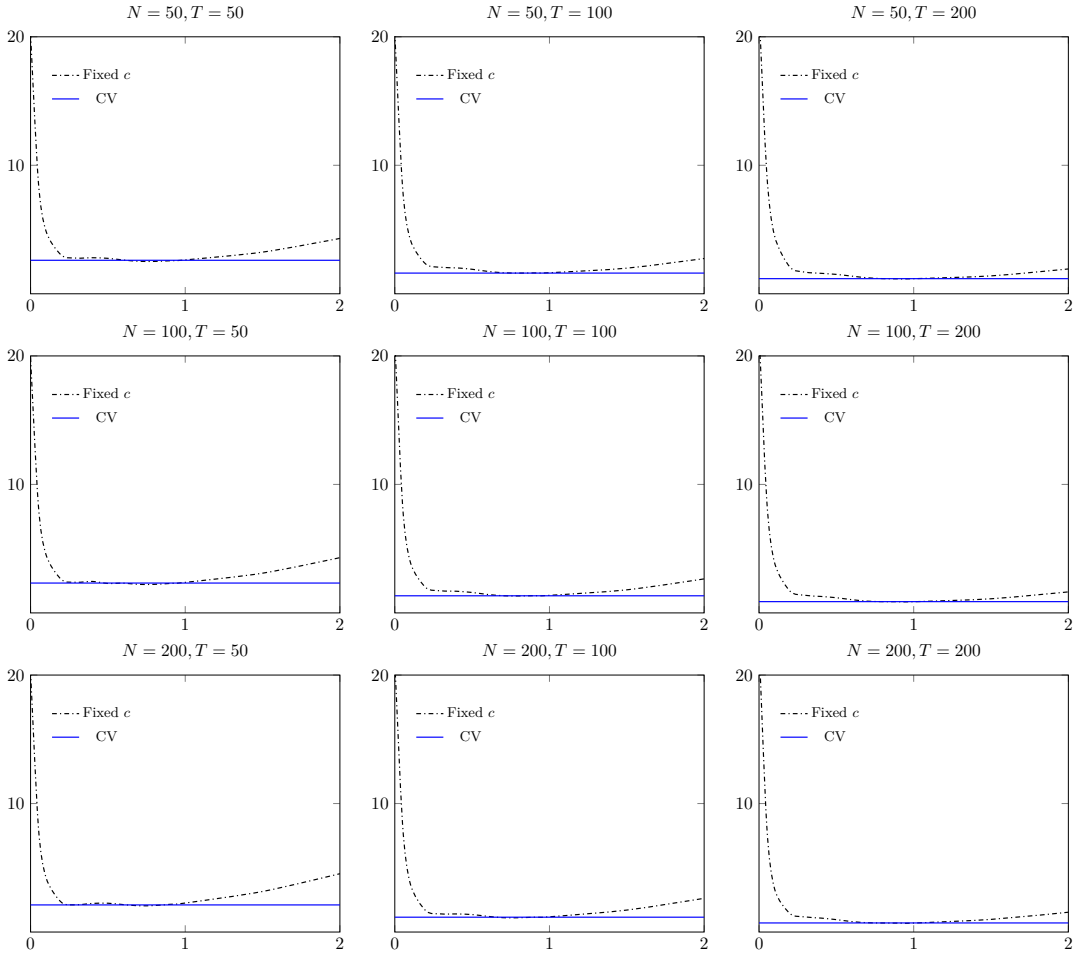

Figure 1. Mean square errors of $\hat{\Pi}$ when using fixed c and CV: DGP1

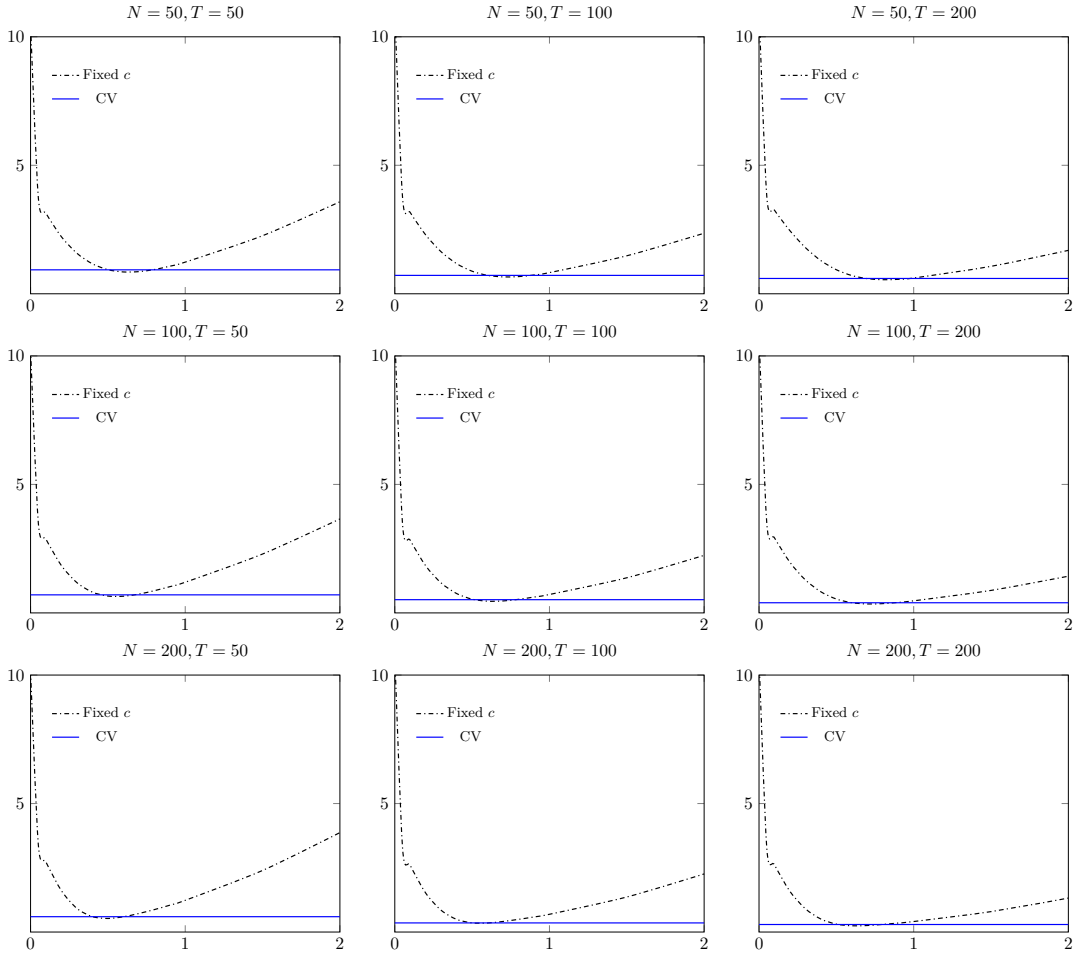

Figure 2. Mean square errors of $(\hat{\Pi}^{\diamond\prime}, \sqrt{N}\hat{\Pi}^{*\prime})$ when using fixed c and CV: DGP2

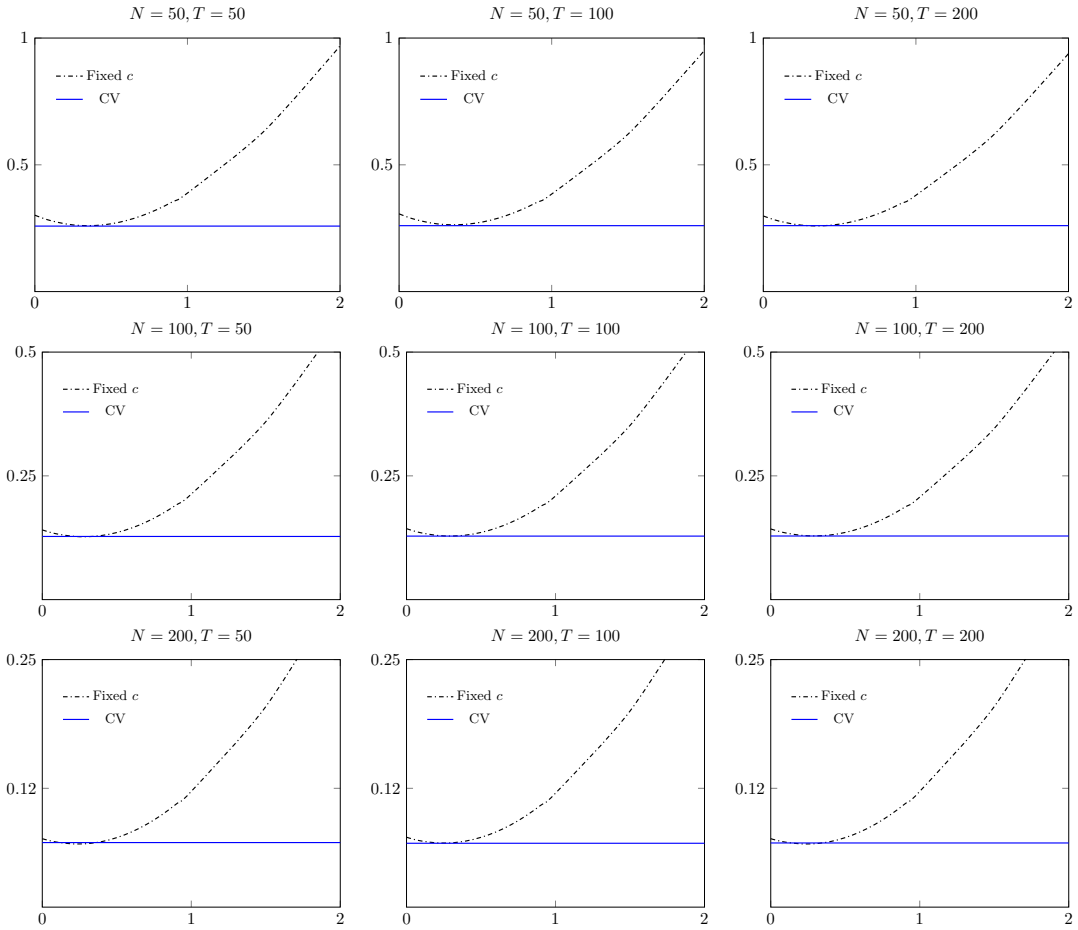

Figure 3. Mean square errors of $\hat{\Pi}_0$ when using fixed c and CV: DGP3

Table I. Mean square errors of $\hat{\Pi}$, $\hat{a}$, $\hat{B}$ and $\hat{F}$, and correct rates of $\hat{K}$: DGP1[†]

(N,T)	$\hat{\Pi}$	$\hat{a}$	$\hat{B}$	$\hat{F}$	$\hat{K}$
$(50, 50)$	2.607	1.127	0.820	0.217	0.955
$(50, 100)$	1.610	0.962	0.355	0.203	1.000
$(50, 200)$	1.176	0.720	0.157	0.201	1.000
$(100, 50)$	2.323	1.187	0.667	0.133	0.990
$(100, 100)$	1.332	1.051	0.306	0.171	1.000
$(100, 200)$	0.877	0.577	0.124	0.132	1.000
$(200, 50)$	2.111	1.240	0.570	0.095	1.000
$(200, 100)$	1.155	0.849	0.254	0.114	1.000
$(200, 200)$	0.707	0.506	0.103	0.091	1.000

[†] The mean square errors of $\hat{\Pi}$, $\hat{a}$, $\hat{B}$ and $\hat{F}$ are given by $\sum_{\ell=1}^{200} \|\hat{\Pi}^{(\ell)} - \Pi\|_F^2/200NT$, $\sum_{\ell=1}^{200} \|\hat{a}^{(\ell)} - a\|^2/200N$, $\sum_{\ell=1}^{200} \|\hat{B}^{(\ell)} - BH^{(\ell)}\|_F^2/200N$ and $\sum_{\ell=1}^{200} \|\hat{F}^{(\ell)} - F(H^{(\ell)\prime})^{-1}\|_F^2/200T$, where $\hat{\Pi}^{(\ell)}$, $\hat{a}^{(\ell)}$, $\hat{B}^{(\ell)}$ and $\hat{F}^{(\ell)}$ are estimates in the ℓth simulation replication, and $H^{(\ell)} \equiv (F'M_T\hat{F}^{(\ell)})(\hat{F}^{(\ell)\prime}M_T\hat{F}^{(\ell)})^{-1}$ is a rotational transformation matrix. The value of c is chosen from $\{0, 0.05, 0.1, 0.2, \ldots, 0.9, 1, 1.5, 2\}$ by using the 5-fold CV method as described in Remark 4.2.

Table II. Mean square errors of $\hat{\Pi}^{\diamond}$, $\hat{\Pi}^{*}$, $\hat{\mu}$, $\hat{\Lambda}$, $\hat{\phi}, \hat{\Phi}$ and $\hat{F}$, and correct rates of $\hat{K}$: DGP2[†]

(N,T)	$\hat{\Pi}^{\diamond}$	$\hat{\Pi}^{*}$	$\hat{\mu}$	$\hat{\Lambda}$	$\hat{\phi}$	$\hat{\Phi}$	$\hat{F}$	$\hat{K}$
$(50, 50)$	0.575	0.357	0.216	0.072	0.421	0.078	0.176	1.000
$(50, 100)$	0.418	0.298	0.108	0.037	0.373	0.045	0.172	1.000
$(50, 200)$	0.322	0.274	0.053	0.019	0.287	0.030	0.170	1.000
$(100, 50)$	0.450	0.253	0.206	0.070	0.399	0.065	0.111	1.000
$(100, 100)$	0.328	0.186	0.119	0.034	0.268	0.031	0.105	1.000
$(100, 200)$	0.241	0.154	0.059	0.015	0.173	0.019	0.096	1.000
$(200, 50)$	0.407	0.190	0.225	0.066	0.339	0.051	0.077	1.000
$(200, 100)$	0.219	0.136	0.119	0.033	0.222	0.026	0.073	1.000
$(200, 200)$	0.197	0.096	0.064	0.014	0.114	0.013	0.057	1.000

[†] The mean square errors of $\hat{\Pi}^{\diamond}$, $\hat{\Pi}^{*}$, $\hat{\mu}$, $\hat{\Lambda}$, $\hat{\phi}, \hat{\Phi}$ and $\hat{F}$ are given by $\sum_{\ell=1}^{200} \|\hat{\Pi}^{\diamond(\ell)} - \Pi^{\diamond}\|_F^2/200NT$, $\sum_{\ell=1}^{200} \|\hat{\Pi}^{*(\ell)} - \Pi^{*}\|_F^2/200T$,$\sum_{\ell=1}^{200} \|\hat{\mu}^{(\ell)} - \mu\|^2/200N$, $\sum_{\ell=1}^{200} \|\hat{\Lambda}^{(\ell)} - \Lambda H^{(\ell)}\|_F^2/200N$, $\sum_{\ell=1}^{200} \|\hat{\phi}^{(\ell)} - \phi\|^2/200$, $\sum_{\ell=1}^{200} \|\hat{\Phi}^{(\ell)} - \Phi H^{(\ell)}\|^2/200$ and $\sum_{\ell=1}^{200} \|\hat{F}^{(\ell)} - F(H^{(\ell)\prime})^{-1}\|_F^2/200T$, where $\hat{\Pi}^{\diamond(\ell)}$, $\hat{\Pi}^{*(\ell)}$, $\hat{\mu}^{(\ell)}$, $\hat{\Lambda}^{(\ell)}$, $\hat{\phi}^{(\ell)}$, $\hat{\Phi}^{(\ell)}$ and $\hat{F}^{(\ell)}$ are estimates in the ℓth simulation replication, and $H^{(\ell)} \equiv (F'M_T\hat{F}^{(\ell)})(\hat{F}^{(\ell)\prime}M_T\hat{F}^{(\ell)})^{-1}$ is a rotational transformation matrix. The value of c is chosen from $\{0, 0.05, 0.1, 0.2, \ldots, 0.9, 1, 1.5, 2\}$ by using the 5-fold CV method as described in Remark 4.2.

Table III. Mean square errors of $\hat{\Pi}_0$, $\hat{\phi}_0$, $\hat{\Phi}_0$ and $\hat{F}_0$ ($\times 10^{-1}$), and correct rates of $\hat{K}_0$: DGP3[†]

(N,T)	$\hat{\Pi}_0$	$\hat{\phi}_0$	$\hat{\Phi}_0$	$\hat{F}_0$	$\hat{K}_0$
$(50, 50)$	2.583	0.615	0.081	1.731	1.000
$(50, 100)$	2.600	0.486	0.050	1.697	1.000
$(50, 200)$	2.601	0.328	0.030	1.643	1.000
$(100, 50)$	1.276	0.248	0.036	0.862	1.000
$(100, 100)$	1.283	0.196	0.022	0.832	1.000
$(100, 200)$	1.285	0.127	0.014	0.804	1.000
$(200, 50)$	0.652	0.131	0.019	0.447	1.000
$(200, 100)$	0.645	0.083	0.011	0.415	1.000
$(200, 200)$	0.648	0.056	0.007	0.408	1.000

[†] The mean square errors of $\hat{\Pi}_0$, $\hat{\phi}_0$, $\hat{\Phi}_0$ and $\hat{F}_0$ are given by $\sum_{\ell=1}^{200} \|\hat{\Pi}_0^{(\ell)} - \Pi_0\|_F^2/200T$, $\sum_{\ell=1}^{200} \|\hat{\phi}_0^{(\ell)} - \phi\|^2/200$, $\sum_{\ell=1}^{200} \|\hat{\Phi}_0^{(\ell)} - \Phi H^{(\ell)}\|_F^2/200$ and $\sum_{\ell=1}^{200} \|\hat{F}_0^{(\ell)} - F(H^{(\ell)\prime})^{-1}\|_F^2/200T$, where $\hat{\Pi}_0^{(\ell)}$, $\hat{\phi}_0^{(\ell)}$, $\hat{\Phi}_0^{(\ell)}$ and $\hat{F}_0^{(\ell)}$ are estimates in the ℓth simulation replication, and $H^{(\ell)} \equiv (F'M_T\hat{F}_0^{(\ell)})(\hat{F}_0^{(\ell)\prime}M_T\hat{F}_0^{(\ell)})^{-1}$ is a rotational transformation matrix. The value of c is chosen from $\{0, 0.05, 0.1, 0.2, \ldots, 0.9, 1, 1.5, 2\}$ by using the 5-fold CV method as described in Remark 4.2.

7 Empirical Analysis

In this section, we analyze the cross section of individual stock returns in the US market. We use the same data set as used in Chen et al. (2021), which is originally from Freyberger, Neuhierl, and Weber (2020). The data set contains monthly returns and 36 characteristics of 12,813 individual US stocks from September, 1968 to May, 2014. There are many stocks which have a large proportion of missing values. To satisfy the requirement that the proportion of missing values cannot be too large, we choose to discard stocks with sample length less than 200. This yields an unbalanced panel with $N = 2,121$ and $T = 549$, where each time period consists of at least 580 stocks that have observations on both returns and the 36 characteristics, and each stock has observations in at least 200 time periods. We also transform the values of each characteristic to relative ranking values with range $[-0.5, 0.5]$ in each time period.

We consider six different model specifications. In the first three specifications (denoted S1, S2 and S3), x_{it} consists of constant and the 36 characteristics. In the other three specifications (denoted S4, S5 and S6), x_{it} consists of constant and linear B-splines of 18 characteristics with one internal knot as studied in Chen et al. (2021).[11] In S1 and S4, we consider an unconstrained conditional factor model (as in Example 2.5), where a_i and B_i are allowed to be heterogeneous across i. In S2 and S5, we consider a semiparametric conditional factor model (as in Example 2.2), where the rows of a_i and B_i corresponding to the nonconstant explanatory variables in x_{it} are restricted to be homogenous. In S3 and S6, we consider a homogenous conditional factor model (as in Example 2.4), where a_i and B_i are restricted to be homogenous. We estimate the models for $K = 1, 2, \ldots, 10$, and choose the regularization parameter by the 5-fold CV approach as described in Remark 4.2. Specifically, we let $\lambda_{NT} = c\sqrt{(Np+T)\log N}$ in S1 and S4, $\lambda_{NT} = c\sqrt{(Np+T)\log N}$ in S2 and S5, $\lambda_{0,NT} = c\sqrt{N(p+T)\log N}$, and choose c from $\{0, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2, 5\}/100$.

Let us write $\hat{a} \equiv (\hat{a}_1', \hat{a}_2', \ldots, \hat{a}_N')'$, $\hat{B} \equiv (\hat{B}_1', \hat{B}_2', \ldots, \hat{B}_N')'$, and $\hat{F} \equiv (\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_T)'$. To evaluate the performance of the models, we follow Chen et al. (2021) to consider various goodness-of-fit measures. First, we consider the following types of in-sample R^2:

$$R^2 = 1 - \frac{\sum_{i,t}(y_{it} - x_{it}'\hat{a}_i - x_{it}'\hat{B}_i\hat{f}_t)^2}{\sum_{i,t} y_{it}^2}, \tag{32}$$

$$R^2_{T,N} = 1 - \frac{1}{N}\sum_i \frac{\sum_t(y_{it} - x_{it}'\hat{a}_i - x_{it}'\hat{B}_i\hat{f}_t)^2}{\sum_t y_{it}^2}, \tag{33}$$

$$R^2_{N,T} = 1 - \frac{1}{T}\sum_t \frac{\sum_i(y_{it} - x_{it}'\hat{a}_i - x_{it}'\hat{B}_i\hat{f}_t)^2}{\sum_i y_{it}^2}. \tag{34}$$

The first one is total R^2. The second one measures the cross-sectional average of time

[11] See their paper for the list of the 18 characteristics and how they are selected.

series R^2 across all stocks, which reflects the ability of the extracted factors to capture common variation in asset returns. The third one measures the time series average of cross-sectional R^2, which is the one of interest for evaluating models' ability to explain the cross-section of average returns. Second, we assess the out-of-sample prediction. For $t \geq 300$, we use the data through $t-1$ to implement our estimation procedure and obtain estimators, say $\hat{a}_{it}$, $\hat{B}_{it}$, $\hat{F}_t \equiv (\hat{f}_1^{(t)}, \hat{f}_2^{(t)}, \ldots, \hat{f}_{t-1}^{(t)})'$; and then compute the out-of-sample prediction of y_{it} as $x_{it}'\hat{a}_{it} - x_{it}'\hat{B}_{it}\hat{\lambda}_t$, where $\hat{\lambda}_t = \sum_{s\leq t-1} \hat{f}_s^{(t)}/(t-1)$. We can define analogously three types of out-of-sample predictive R^2's by replacing $\hat{a}_i$, $\hat{B}_i$ and $\hat{f}_t$ with $\hat{a}_{it}$, $\hat{B}_{it}$ and $\hat{\lambda}_t$:

$$R_O^2 = 1 - \frac{\sum_{i,t\geq 300}(y_{it} - x_{it}'\hat{a}_{it} - x_{it}'\hat{B}_{it}\hat{\lambda}_t)^2}{\sum_{i,t\geq 300} y_{it}^2}, \tag{35}$$

$$R_{T,N,O}^2 = 1 - \frac{1}{N}\sum_i \frac{\sum_{t\geq 300}(y_{it} - x_{it}'\hat{a}_{it} - x_{it}'\hat{B}_{it}\hat{\lambda}_t)^2}{\sum_{t\geq 300} y_{it}^2}, \tag{36}$$

$$R_{N,T,O}^2 = 1 - \frac{1}{T-299}\sum_{t\geq 300} \frac{\sum_i(y_{it} - x_{it}'\hat{a}_{it} - x_{it}'\hat{B}_{it}\hat{\lambda}_t)^2}{\sum_i y_{it}^2}. \tag{37}$$

The results are reported in Figure 4. The main findings are summarized as follows. First, the in-sample R^2's increase as K increases, while the out-of-sample R^2's are invariant to the change of K. The invariance property follows because $\hat{\lambda} = \sum_{t\leq T} \hat{f}_t/T = \hat{F}'1_T/T = \hat{B}'\hat{\Pi}1_T/(NT)$, $\hat{a} + \hat{B}\hat{\lambda} = \hat{\Pi}1_T/T$, and the out-of-sample prediction of y_{it} does not depend on K. Second, among the linear models (i.e., S1,S2 and S3), S1 has the best in-sample performance in all three in-sample R^2's regardless of the value of K, while S3 has the best out-of-sample performance in all three out-of-sample R^2's. This suggests that imposing homogeneity of a_i and B_i across i may improve the model's out-of-sample predictability, though the in-sample fit is worsened. Similarly, for the spline models (i.e., S4,S5 and S6), imposing homgogeneity of a_i and B_i across i may improve the model's out-of-sample predictability. Third, S5 and S6 have better out-of-sample performance than S2 and S3, respectively. This implies that using spline transformation of characteristics may improve the model's out-of-sample predictability, suggesting the importance of nonlinearity. Overall, S1 has the best in-sample performance, while S6 has the best out-of-sample performance.

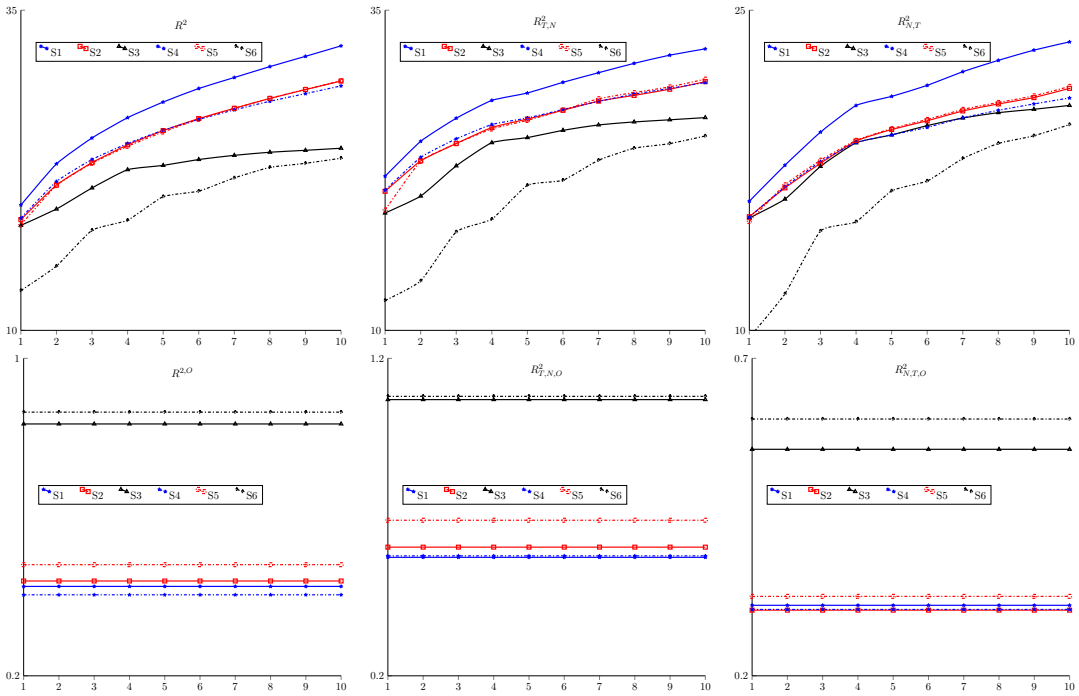

Figure 4. In-sample and out-of-sample R^2's

8 Conclusion

This paper developed a nuclear norm regularized estimation of high-dimensional conditional factor models, and established large sample properties of the estimators. The method allows us to estimate a variety of conditional factor models in a unified framework and quickly deliver new asymptotic results. We applied the method to study the cross section of individual US stock returns, and found that imposing homogeneity of a_i and B_i across i might improve the model's out-of-sample performance. In asset pricing, several inference problems, including testing whether pricing errors are zero and specification test of risk exposure functions, are important for evaluating and comparing factor models. Chernozhukov et al. (2018) provide an asymptotically valid inferential procedure for a special case when $a_i = 0$, $p = 1$ and x_{it} exhibits a factor structure, which nevertheless is not suited for asset pricing. Developing a general inferential method within our framework is an interesting future research question.

Appendix A Proofs of Main Results

Proof of Theorem 4.1: (i) The result in fact follows as a modification of the proof of Corollary 1 in Negahban and Wainwright (2011). By the definition of $\hat{\Pi}$,

$$\frac{1}{2}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - \mathrm{tr}(X_{it}'\hat{\Pi}))^2 + \lambda_{NT}\|\hat{\Pi}\|_* \leq \frac{1}{2}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - \mathrm{tr}(X_{it}'\Pi))^2 + \lambda_{NT}\|\Pi\|_*. \quad \text{(A.1)}$$

Let $\Delta \equiv \hat{\Pi} - \Pi \in \mathcal{S} \ominus \mathcal{S}$. Noting that $\sum_{i=1}^{N}\sum_{t=1}^{T}|\mathrm{tr}(X_{it}'\Delta)|^2 = \mathcal{Q}_{NT}(\Delta) + \mathcal{L}_{NT}(\Delta)$, we may rearrange (A.1) to obtain

$$\begin{aligned}\frac{1}{2}\mathcal{Q}_{NT}(\Delta) &\leq -\frac{1}{2}\mathcal{L}_{NT}(\Delta) + \sum_{i=1}^{N}\sum_{t=1}^{T}\mathrm{tr}(\varepsilon_{it}X_{it}'\Delta) + \lambda_{NT}\|\Pi\|_* - \lambda_{NT}\|\Pi + \Delta\|_* \\ &\leq r_{NT}\|\Delta\|_* + \lambda_{NT}\|\Pi\|_* - \lambda_{NT}\|\Pi + \Delta\|_* \\ &\leq \lambda_{NT}\left(\frac{1}{2}\|\Delta\|_* + \|\Pi\|_* - \|\Pi + \Delta\|_*\right), \end{aligned} \quad \text{(A.2)}$$

where the first inequality follows by Assumption 4.1 and the second inequality follows since $\lambda_{NT} \geq 2r_{NT}$. Since $\Delta = \mathcal{P}(\Delta) + \mathcal{M}(\Delta)$, it follows that

$$\begin{aligned}\|\Pi\|_* - \|\Pi + \Delta\|_* &= \|\Pi\|_* - \|\Pi + \mathcal{P}(\Delta) + \mathcal{M}(\Delta)\|_* \\ &\leq \|\Pi\|_* - \|\Pi + \mathcal{P}(\Delta)\|_* + \|\mathcal{M}(\Delta)\|_* \\ &= \|\mathcal{M}(\Delta)\|_* - \|\mathcal{P}(\Delta)\|_* \end{aligned} \quad \text{(A.3)}$$

where the inequality follows by the triangle inequality and the second equality follows by Lemma A.1(i). Since $\|\Delta\|_* \leq \|\mathcal{P}(\Delta)\|_* + \|\mathcal{M}(\Delta)\|_*$, combining (A.2) and (A.3) gives

$$0 \leq \frac{1}{2}\mathcal{Q}_{NT}(\Delta) \leq \lambda_{NT}\left(\frac{3}{2}\|\mathcal{M}(\Delta)\|_* - \frac{1}{2}\|\mathcal{P}(\Delta)\|_*\right). \tag{A.4}$$

Therefore, $\|\mathcal{P}(\Delta)\|_* \leq 3\|\mathcal{M}(\Delta)\|_*$ and $\Delta \in \mathcal{C}$. This in turn together with (A.4) and Assumption 4.1(i) implies that

$$\begin{aligned}\frac{1}{2}\kappa\|\Delta\|_F^2 &\leq \lambda_{NT}\left(\frac{3}{2}\|\mathcal{M}(\Delta)\|_* - \frac{1}{2}\|\mathcal{P}(\Delta)\|_*\right) \leq \frac{3}{2}\lambda_{NT}\|\mathcal{M}(\Delta)\|_* \\ &\leq \frac{3}{2}\lambda_{NT}\sqrt{2(K+1)}\|\mathcal{M}(\Delta)\|_F \leq \frac{3}{2}\lambda_{NT}\sqrt{2(K+1)}\|\Delta\|_F,\end{aligned} \tag{A.5}$$

where the second inequality follows since $\|\mathcal{P}(\Delta)\|_* \geq 0$, the third inequality follows by the Cauchy-Schwartz inequality (i.e., $\|A\|_* \leq \sqrt{\text{rank}(A)}\|A\|_F$) and Lemma A.1(ii), and the last inequality follows by Lemma A.1(iii). Thus, the result follows by (A.5).

(ii) Let $\sigma_j(A)$ denote the jth largest singular value of A, so $\lambda_j(\hat{\Pi}M_T\hat{\Pi}') = \sigma_j^2(\hat{\Pi}M_T)$. If $\hat{K} \neq K$, then $\lambda_K(\hat{\Pi}M_T\hat{\Pi}') < \delta_{NT}$ or $\lambda_{K+1}(\hat{\Pi}M_T\hat{\Pi}') \geq \delta_{NT}$, equivalently, $\sigma_K(\hat{\Pi}M_T) < \sqrt{\delta_{NT}}$ or $\sigma_{K+1}(\hat{\Pi}M_T) \geq \sqrt{\delta_{NT}}$. Thus, we obtain

$$P(\hat{K} \neq K) \leq P(\sigma_K(\hat{\Pi}M_T) < \sqrt{\delta_{NT}}) + P(\sigma_{K+1}(\hat{\Pi}M_T) \geq \sqrt{\delta_{NT}}). \tag{A.6}$$

By the Weyl's inequality, we have

$$\sup_{j \leq \min\{Np,T\}} |\sigma_j(\hat{\Pi}M_T) - \sigma_j(\Pi M_T)| \leq \|\hat{\Pi}M_T - \Pi M_T\|_F \leq \|\hat{\Pi} - \Pi\|_F, \tag{A.7}$$

where the second inequality follows since $\|CD\|_F \leq \|C\|_F\|D\|_2$ and $\|M_T\|_2 = 1$. It then follows from (A.7) and Theorem 4.1(i) that with probability approaching one,

$$\sigma_K(\hat{\Pi}M_T) \geq \sigma_K(\Pi M_T) - O_p(\sqrt{K}\lambda_{NT}) \geq \sqrt{\delta_{NT}} \tag{A.8}$$

and

$$\sigma_{K+1}(\hat{\Pi}M_T) \leq \sigma_{K+1}(\Pi M_T) + O_p(\sqrt{K}\lambda_{NT}) < \sqrt{\delta_{NT}}, \tag{A.9}$$

where the second equality in (A.8) follows since $\delta_{NT}/(K\lambda_{NT}^2) \to \infty$, $\delta_{NT}/(NT) \to 0$ and $\sigma_K^2(\Pi M_T/\sqrt{NT}) = \lambda_{\min}((B'B/N)(F'M_TF/T)) > d_{\min}^2$, and the second equality in (A.9) follows since $\sigma_{K+1}(\Pi M_T) = 0$ and $\delta_{NT}/(K\lambda_{NT}^2) \to \infty$. Thus, the first result follows from (A.6), (A.8) and (A.9).

It is without loss of generality to assume that $\hat{K} = K$. Let V be a $K \times K$ diagonal

matrix of the first K largest eigenvalues of $\hat{\Pi}M_T\hat{\Pi}'/(NT)$. By the definitions of $\hat{B}$,

$$\hat{B} = \frac{1}{NT}\hat{\Pi}M_T\hat{\Pi}'\hat{B}V^{-1} = BH + \frac{1}{NT}(\hat{\Pi} - \Pi)M_T\hat{\Pi}'\hat{B}V^{-1}, \tag{A.10}$$

where the second equality follows since $\hat{F}'M_T\hat{F}/T = V$, $\Pi M_T = BF'M_T$ and $\hat{F} = \hat{\Pi}'\hat{B}$. By Assumptions 4.2(i), (ii) and (iv), $\|\Pi/\sqrt{NT}\|_F$ is bounded. Since $\sqrt{K}\lambda_{NT}/\sqrt{NT} = o(1)$, $\|\hat{\Pi}/\sqrt{NT}\|_F = O_p(1)$ by Theorem 4.1(i). Thus, the thid result follows from (A.10), Lemma A.1(i) and Theorem 4.1(i). By the definition of $\hat{a}$,

$$\begin{aligned}\hat{a} = a - \frac{1}{N}\hat{B}(\hat{B} - BH)'a - \left(I_{Np} - \frac{\hat{B}\hat{B}'}{N}\right)(\hat{B} - BH)H^{-1}\frac{1}{T}F'1_T \\ + \left(I_{Np} - \frac{\hat{B}\hat{B}'}{N}\right)\frac{1}{T}(\hat{\Pi} - \Pi)1_T,\end{aligned} \tag{A.11}$$

where we have used $a'B = 0$ and $\Pi = a1_T' + BF'$. By Assumptions 4.2(ii) and (iv), $\|F'1_T/T\|$ and $\|a/\sqrt{N}\|$ are bounded. Thus, the second result follows from (A.11), the second result, Lemma A.1(ii) and Theorem 4.1(i). By the definition of $\hat{F}$,

$$\hat{F} = F(H')^{-1} - F(H')^{-1}\frac{1}{N}(\hat{B} - BH)'\hat{B} + \frac{1}{N}1_Ta'(\hat{B} - BH) + \frac{1}{N}(\hat{\Pi} - \Pi)'\hat{B}, \tag{A.12}$$

where we have used $a'B = 0$ and $\Pi = a1_T' + BF'$. Thus, the last result follows from (A.12), the second result, Lemma A.1(ii) and Theorem 4.1(i). ■

Appendix A.1 Technical Lemmas

Lemma A.1. *For any $Np \times T$ matrix Δ, let $\mathcal{P}(\Delta)$ and $\mathcal{M}(\Delta)$ be given in Section 4. Assume $0 < K < \min\{Np, T\} - 1$. For any $Np \times T$ matrix Δ, the followings are true.*
(i) $\|\Pi + \mathcal{P}(\Delta)\|_* = \|\Pi\|_* + \|\mathcal{P}(\Delta)\|_*$.
(ii) The rank of $\mathcal{M}(\Delta)$ is no greater than $2(K+1)$.
(iii) $\|\Delta\|_F^2 = \|\mathcal{P}(\Delta)\|_F^2 + \|\mathcal{M}(\Delta)\|_F^2$.

Proof: (i) Since $\mathcal{P}(\Delta) = U_2U_2'\Delta V_2V_2'$ and $\Pi = U_1\Sigma_{11}V_1'$ where Σ_{11} is square diagonal matrix with nonzero singular values of Π in the diagonal in descending order, the result follows by Lemma 2.3 of Recht et al. (2010).

(ii) We have the following decomposition:

$$\begin{aligned}\Delta &= U(U_1, U_2)'\Delta(V_1, V_2)V' \\ &= U\begin{pmatrix} U_1'\Delta V_1 & U_1'\Delta V_2 \\ U_2'\Delta V_1 & U_2'\Delta V_2 \end{pmatrix}V' \\ &= U\begin{pmatrix} 0 & 0 \\ 0 & U_2'\Delta V_2 \end{pmatrix}V' + U\begin{pmatrix} U_1'\Delta V_1 & U_1'\Delta V_2 \\ U_2'\Delta V_1 & 0 \end{pmatrix}V'\end{aligned}$$

$$= \mathcal{P}(\Delta) + U \begin{pmatrix} U_1'\Delta V_1 & U_1'\Delta V_2 \\ U_2'\Delta V_1 & 0 \end{pmatrix} V'. \tag{A.13}$$

Therefore, by (A.13) we obtain

$$\mathcal{M}(\Delta) = U \begin{pmatrix} U_1'\Delta V_1 & U_1'\Delta V_2 \\ U_2'\Delta V_1 & 0 \end{pmatrix} V'. \tag{A.14}$$

Thus, by (A.14) it follows that

$$\begin{aligned} \text{rank}(\mathcal{M}(\Delta)) &= \text{rank}\left(\begin{pmatrix} U_1'\Delta V_1 & U_1'\Delta V_2 \\ U_2'\Delta V_1 & 0 \end{pmatrix}\right) \\ &\leq \text{rank}\left(\begin{pmatrix} U_1'\Delta V_1 & U_1'\Delta V_2 \\ 0 & 0 \end{pmatrix}\right) + \text{rank}\left(\begin{pmatrix} 0 & 0 \\ U_2'\Delta V_1 & 0 \end{pmatrix}\right) \\ &\leq 2(K+1), \end{aligned} \tag{A.15}$$

where the first inequality follows by the fact that $\text{rank}(C+D) \leq \text{rank}(C) + \text{rank}(D)$ (see, for example, Fact 2.10.17 in Bernstein (2018)) and the second inequality follows since Π has at most rank $K+1$.

(iii) By (A.13) and (A.14), we obtain

$$\begin{aligned} \|\mathcal{P}(\Delta)\|_F^2 + \|\mathcal{M}(\Delta)\|_F^2 &= \left\| \begin{pmatrix} 0 & 0 \\ 0 & U_2'\Delta V_2 \end{pmatrix} \right\|_F^2 + \left\| \begin{pmatrix} U_1'\Delta V_1 & U_1'\Delta V_2 \\ U_2'\Delta V_1 & 0 \end{pmatrix} \right\|_F^2 \\ &= \|\Delta\|_F^2, \end{aligned} \tag{A.16}$$

where the second equality follows by the first two equalities in (A.13). ■

Lemma A.2. *Suppose Assumption 4.2 holds. Let V be a $K \times K$ diagonal matrix of the first K largest eigenvalues of $\hat{\Pi} M_T \hat{\Pi}'/(NT)$. Assume that $\|\hat{\Pi} - \Pi\|_F = o_p(\sqrt{NT})$ and $P(\hat{K} = K) \to 1$. Then (i) $\|V\|_2 = O_p(1)$, $\|V^{-1}\|_2 = O_p(1)$, and $\|H\|_2 = O_p(1)$; (ii) $\|H^{-1}\|_2 = O_p(1)$, if $\|\hat{B} - BH\|_F = o_p(\sqrt{N})$.*

Proof: (i) Let $\sigma_j(A)$ denote the jth largest singular value of A, so $\lambda_j(\hat{\Pi} M_T \hat{\Pi}'/(NT)) = \sigma_j^2(\hat{\Pi} M_T/\sqrt{NT})$ and $\lambda_j(\Pi M_T \Pi'/(NT)) = \sigma_j^2(\Pi M_T/\sqrt{NT})$. By the triangle inequality, it follows from (A.7) that

$$\sqrt{\|V\|_2} = \sigma_1(\hat{\Pi} M_T/\sqrt{NT}) \leq \sigma_1(\Pi M_T/\sqrt{NT}) + \|\hat{\Pi} - \Pi\|_F/\sqrt{NT} = O_p(1). \tag{A.17}$$

where the last equality follows since $\sigma_1(\Pi M_T/\sqrt{NT})$ is bounded. Similarly,

$$\sqrt{\|V^{-1}\|_2} = \sigma_K^{-1}(\hat{\Pi} M_T/\sqrt{NT}) \leq \sigma_K^{-1}(\Pi M_T/\sqrt{NT}) + o_p(1) = O_p(1), \tag{A.18}$$

where the last equality follows since $\sigma_K^2(\Pi M_T/\sqrt{NT}) = \lambda_{\min}((B'B/N)(F'M_TF/T)) >$

$d^2_{\min}$. Let $H^\diamond \equiv (F'M_TF/T)(B'\hat{B}/N)V^{-1}$. Recall that $H = (F'M_T\hat{\Pi}'\hat{B}/T)V^{-1}$. Then,

$$\|H - H^\diamond\|_2 \leq \frac{1}{NT}\|F\|_2\|\hat{\Pi} - \Pi\|_F\|\hat{B}\|_2\|V^{-1}\|_2 = o_p(1), \tag{A.19}$$

where the equality follows by Assumption 4.2(ii). Since $\|H^\diamond\|_2 = O_p(1)$, it follows from (A.19) that $\|H\|_2 = O_p(1)$.

(ii) Since $\|\hat{B} - BH\|_F = o_p(\sqrt{N})$, we have $\|\hat{B}'\hat{B}/(N) - H'(B'B/N)H\|_F = o_p(1)$ by the triangle inequality. This implies that $I_K - \lambda_{\max}(B'B/N)H'H$ is negative semidefinite with probability approaching one. Therefore, the eigenvalues of $H'H$ are no smaller than $\lambda^{-1}_{\max}(B'B/N)$ with probability approaching one. Thus, the result of the lemma follows from Assumption 4.2(i). ■

Appendix B Computing Algorithms

In this appendix, we present computing algorithms for finding the nuclear norm regularized estimators in Examples 2.1-2.5. Specifically, we use the accelerated proximal gradient algorithm by Ji and Ye (2009) and Toh and Yun (2010). The algorithm solves the following general nonsmooth convex minimization problem:

$$\min_{\Gamma\in\mathbf{R}^{m\times T}} F(\Gamma) \equiv f(\Gamma) + \varphi_{NT}\|\Gamma\|_*, \tag{B.1}$$

where $\Gamma \in \mathbf{R}^{m\times T}$ is the decision matrix, $f : \mathbf{R}^{m\times T} \mapsto [0,\infty)$ is a smooth loss function with the gradient $\nabla f(\Gamma)$ being Lipschitz continuous with constant L_f (namely, $\|\nabla f(\Gamma^{(1)}) - \nabla f(\Gamma^{(2)})\|_F \leq L_f\|\Gamma^{(1)} - \Gamma^{(2)}\|_F$ for any $\Gamma^{(1)}, \Gamma^{(2)} \in \mathbf{R}^{m\times T}$), $\|\Gamma\|_*$ is the nuclear norm of Γ, $\varphi_{NT} > 0$ is a regularization parameter. The algorithm consists of recursively solving a sequence of minimizations of linear approximations of $f(\Gamma)$ regularized by a quadratic proximal term and the nuclear norm, which is given by

$$\begin{aligned}&\min_{\Gamma\in\mathbf{R}^{m\times T}} Q_{\tau_k}(\Gamma,\Gamma_k) \equiv f(\Gamma_k) + \mathrm{tr}((\Gamma - \Gamma_k)'\nabla f(\Gamma_k)) + \frac{\tau_k}{2}\|\Gamma - \Gamma_k\|^2_F + \varphi_{NT}\|\Gamma\|_*,\\ :=&\min_{\Gamma\in\mathbf{R}^{m\times T}} \frac{\tau_k}{2}\left\|\Gamma - \left(\Gamma_k - \frac{1}{\tau_k}\nabla f(\Gamma_k)\right)\right\|^2_F + \varphi_{NT}\|\Gamma\|_* + f(\Gamma_k) - \frac{1}{2\tau_k}\|\nabla f(\Gamma_k)\|^2_F \end{aligned} \tag{B.2}$$

for $k \in \mathbf{Z}^+$, where $\tau_k > 0$ and Γ_k are recursively updated. The algorithm is attractive in two aspects. First, the problem in (B.2) can be explicitly solved via the singular value decomposition of $\Gamma_k - \frac{1}{\tau_k}\nabla f(\Gamma_k)$ and then applying some soft-thresholding on the singular values. This is because $f(\Gamma_k) - \frac{1}{2\tau_k}\|\nabla f(\Gamma_k)\|^2_F$ does not depend on Γ and $\min_{\Gamma\in\mathbf{R}^{m\times T}} \frac{\tau_k}{2}\|\Gamma - [\Gamma_k - \frac{1}{\tau_k}\nabla f(\Gamma_k)]\|^2_F + \varphi_{NT}\|\Gamma\|_*$ can be explicitly solved by the technique; see, for example, Cai et al. (2010) and Ma et al. (2011). For $A \in \mathbf{R}^{m\times T}$, let $A = U\Sigma V'$ be a singular value decomposition of A, where $U \in \mathbf{R}^{m\times m}$ with $U'U = I_m$, $V \in \mathbf{R}^{T\times T}$ with $V'V = I_T$, and $\Sigma \in \mathbf{R}^{m\times T}$ is a diagonal matrix with singular values

in the diagonal in descending order. For $x > 0$, define $\mathcal{S}_x(A) \equiv U\Sigma_x V'$, where Σ_x is diagonal with the jjth entry equal to $\max\{0, \Sigma_{jj} - x\}$ for all j and Σ_{jj} denotes the jjth entry of Σ. The solution to (B.2) is given by

$$\mathcal{S}_{\tau_k^{-1}\varphi_{NT}}\left(\Gamma_k - \frac{1}{\tau_k}\nabla f(\Gamma_k)\right). \tag{B.3}$$

Second, Ji and Ye (2009) and Toh and Yun (2010) show that if $\tau_k > 0$ and Γ_k are updated properly, the algorithm can achieve the optimal convergence rate of $O(1/k^2)$.

Let $\eta \in (0,1)$ be a given constant. Choose $\Gamma_0^* = \Gamma_1^* \in \mathbf{R}^{m\times T}$. Set $w_0 = w_1 = 1$ and $\tau_0 = L_f$. Set $k = 1$. The algorithm is given as follows.

Step 1. Set $\Gamma_k = \Gamma_k^* + \frac{w_{k-1}-1}{w_k}(\Gamma_k^* - \Gamma_{k-1}^*)$.

Step 2. Set $\hat{\tau}_0 = \eta\tau_{k-1}$. Set $j = 0$ and execute the following step:

- Compute $A_j = \mathcal{S}_{\hat{\tau}_j^{-1}\varphi_{NT}}(\Gamma_k - \hat{\tau}_j^{-1}\nabla f(\Gamma_k))$. If $F(A_j) \le Q_{\hat{\tau}_j}(A_j, \Gamma_k)$, set $\tau_k = \hat{\tau}_j$ and proceed to **Step 3**; Otherwise, set $\hat{\tau}_{j+1} = \min\{\eta^{-1}\hat{\tau}_j, \tau_0\}$ and $j = j + 1$, and return to the beginning of this step.

Step 3. Set $\Gamma_{k+1}^* = \mathcal{S}_{\tau_k^{-1}\varphi_{NT}}(\Gamma_k - \tau_k^{-1}\nabla f(\Gamma_k))$.

Step 4. Set $w_{k+1} = (1 + \sqrt{1 + 4w_k^2})/2$.

Step 5. Compute $D_{k+1} = \tau_k(\Gamma_k - \Gamma_{k+1}^*) + \nabla f(\Gamma_{k+1}^*) - \nabla f(\Gamma_k)$. If $\|D_{k+1}\|_F / [\tau_k \max\{1, \|\Gamma_{k+1}^*\|_F\}] \le \epsilon$ where ϵ is a pre-specified tolerance level, set the output $\hat{\Pi} = \Gamma_{k+1}^*$. Otherwise, set $k = k + 1$ and return to **Step 1**.

Step 2 is to ensure that the objective value generated at the kth iteration is bounded by the minimum of the approximating function, that is, $F(\Gamma_{k+1}^*) \le Q_{\tau_k}(\Gamma_{k+1}^*, \Gamma_k)$, which is crucial to the algorithm. Alternatively, we may fix $\tau_k = L_f$ to meet the requirement; see, for example, Lemma 1.2.3 of Nesterov (2003). By shrinking τ_k, the resulting solution tends to have lower rank than the one generated by setting $\tau_k = L_f$, since smaller value of τ_k may lead to fewer nonzero singular values in $\mathcal{S}_{\tau_k^{-1}\varphi_{NT}}(\Gamma_k - \tau_k^{-1}\nabla f(\Gamma_k))$. Steps 1 and 4 are key steps for the convergence rate of $O(1/k^2)$. Rather than fixing the search point (i.e.,Γ_k) at the solution from the previous iteration (i.e., Γ_k^*), the algorithm constructs the search point as a linear combination of the solutions from the latest two iterations. This may accelerate the convergence rate from $O(1/k)$ to $O(1/k^2)$ (Nesterov, 1983, 2003); see Ji and Ye (2009) and Toh and Yun (2010) for the proofs. The sequence w_k is generated in the manner in Step 4 to satisfy the constraint $w_{k+1}^2 - w_{k+1} \le w_k^2$. In Step 5, D_{k+1} is a subgradient of $F(\Gamma)$ at $\Gamma = \Gamma_{k+1}^*$, see Toh and Yun (2010). In simulations and real data applications, we set $\eta = 0.8$, $\Gamma_0^* = \Gamma_1^* = 0$ and $\epsilon = 10^{-5}$.

We next show how the problems in (9) with $\mathcal{S} = \mathbf{R}^{Np\times T}$, (19) and (24), which respectively define our estimators in Examples 2.1, 2.3 and 2.5, Example 2.2, and Example

2.4, can fit into the general framework in (B.1). In all cases, the algorithms can be easily adapted to allow for the presence of missing values. In both Examples 2.1, 2.3 and 2.5 and Example 2.4, we can simply replace the observations with $y_{it}m_{it}$ and $x_{it}m_{it}$, where m_{it} is a dummy variable of missing status defined in Remark 3.2. It is straightforward to modify the algorithm to accommodate the presence of missing values in Example 2.2. Below we focus on the case without missing values.

Appendix B.1 Examples 2.1, 2.3 and 2.5

For (9) with $\mathcal{S} = \mathbf{R}^{Np\times T}$, we may set $m = Np$, $\varphi_{NT} = \lambda_{NT}$ and

$$f(\Gamma) = \frac{1}{2}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - x_{it}'\gamma_{it})^2 \text{ for } \Gamma \equiv \begin{pmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1T} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2T} \\ \vdots & \vdots & \vdots & \vdots \\ \gamma_{N1} & \gamma_{N2} & \cdots & \gamma_{NT} \end{pmatrix} \in \mathbf{R}^{Np\times T}. \tag{B.4}$$

We need to show that the gradient $\nabla f(\Gamma)$ is Lipschitz continuous. It follows that

$$\nabla f(\Gamma) = \begin{pmatrix} x_{11}(x_{11}'\gamma_{11} - y_{11}) & x_{12}(x_{12}'\gamma_{11} - y_{12}) & \cdots & x_{1T}(x_{1T}'\gamma_{1T} - y_{1T}) \\ x_{21}(x_{21}'\gamma_{21} - y_{21}) & x_{22}(x_{22}'\gamma_{22} - y_{22}) & \cdots & x_{2T}(x_{2T}'\gamma_{2T} - y_{2T}) \\ \vdots & \vdots & \vdots & \vdots \\ x_{N1}(x_{N1}'\gamma_{N1} - y_{N1}) & x_{N2}(x_{N2}'\gamma_{N2} - y_{N2}) & \cdots & x_{NT}(x_{NT}'\gamma_{NT} - y_{NT}) \end{pmatrix}. \tag{B.5}$$

Indeed,$\nabla f(\Gamma)$ is Lipschitz continuous with constant $L_f = \max_{i\le N, t\le T}\|x_{it}\|^2$, because for $\Gamma^{(1)} \equiv (\gamma_{it}^{(1)}) \in \mathbf{R}^{Np\times T}$ and $\Gamma^{(2)} \equiv (\gamma_{it}^{(2)}) \in \mathbf{R}^{Np\times T}$,

$$\begin{aligned} &\|\nabla f(\Gamma^{(1)}) - \nabla f(\Gamma^{(2)})\|_F^2 \\ &= \left\| \begin{pmatrix} x_{11}x_{11}'(\gamma_{11}^{(1)} - \gamma_{11}^{(2)}) & x_{12}x_{12}'(\gamma_{12}^{(1)} - \gamma_{12}^{(2)}) & \cdots & x_{1T}x_{1T}'(\gamma_{1T}^{(1)} - \gamma_{1T}^{(2)}) \\ x_{21}x_{21}'(\gamma_{21}^{(1)} - \gamma_{21}^{(2)}) & x_{22}x_{22}'(\gamma_{22}^{(1)} - \gamma_{22}^{(2)}) & \cdots & x_{2T}x_{2T}'(\gamma_{2T}^{(1)} - \gamma_{2T}^{(2)}) \\ \vdots & \vdots & \vdots & \vdots \\ x_{N1}x_{N1}'(\gamma_{N1}^{(1)} - \gamma_{N1}^{(2)}) & x_{N2}x_{N2}'(\gamma_{N2}^{(1)} - \gamma_{N2}^{(2)}) & \cdots & x_{NT}x_{NT}'(\gamma_{NT}^{(1)} - \gamma_{NT}^{(2)}) \end{pmatrix} \right\|_F^2 \\ &= \sum_{i=1}^{N}\sum_{t=1}^{T}\|x_{it}x_{it}'(\gamma_{it}^{(1)} - \gamma_{it}^{(2)})\|^2 \\ &\le \max_{i\le N, t\le T}\|x_{it}\|^4\|\Gamma^{(1)} - \Gamma^{(2)}\|_F^2. \end{aligned} \tag{B.6}$$

Appendix B.2 Example 2.2

By changing values, we may equivalently rewrite (19) as

$$\begin{pmatrix} \hat{\Pi}^{\diamond} \\ \sqrt{N}\hat{\Pi}^{*} \end{pmatrix} = \underset{\substack{\Gamma^{\diamond}=(\gamma_{it})_{i\leq N,t\leq T}\in\mathbf{R}^{N\times T} \\ \Gamma^{*}=(\gamma_1^*,\ldots,\gamma_T^*)\in\mathbf{R}^{(p-1)\times T} \\ \|\Gamma^*\|_{\max}\leq\sqrt{N}M}}{\arg\min} \frac{1}{2}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it}-\gamma_{it}-w_{it}^{*\prime}\gamma_t^*)^2 + \lambda_{NT}\left\|\begin{pmatrix}\Gamma^{\diamond}\\ \Gamma^{*}\end{pmatrix}\right\|_{*}, \tag{B.7}$$

where $w_{it}^* = x_{it}^*/\sqrt{N}$. Here, we consider the problem by dropping the constraint that $\|\Gamma^*\| \leq \sqrt{N}M$. First, as noted in Footnote 8, the constraint is only a technical condition that simplifies the proof, so may not be necessary. Second, in practice, the constraint is not binding for a sufficiently large value of M, thus can be dropped. Thus, we may set $m = N + p - 1$, $\varphi_{NT} = \lambda_{NT}$ and

$$f(\Gamma) = \frac{1}{2}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it}-\gamma_{it}-w_{it}^{*\prime}\gamma_t^*\right)^2 \text{ for } \Gamma \equiv \begin{pmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1T} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2T} \\ \vdots & \vdots & \vdots & \vdots \\ \gamma_{N1} & \gamma_{N2} & \cdots & \gamma_{NT} \\ \gamma_1^* & \gamma_2^* & \cdots & \gamma_T^* \end{pmatrix} \in \mathbf{R}^{(N+p-1)\times T}. \tag{B.8}$$

We need to show that the gradient $\nabla f(\Gamma)$ is Lipschitz continuous. It follows that

$$\nabla f(\Gamma) = \begin{pmatrix} \gamma_{11}+w_{11}^{*\prime}\gamma_1^*-y_{11} & \gamma_{12}+w_{12}^{*\prime}\gamma_2^*-y_{12} & \cdots & (\gamma_{1T}+w_{1T}^{*\prime}\gamma_T^*-y_{1T}) \\ \gamma_{21}+w_{21}^{*\prime}\gamma_1^*-y_{21} & \gamma_{22}+w_{22}^{*\prime}\gamma_2^*-y_{22} & \cdots & (\gamma_{2T}+w_{2T}^{*\prime}\gamma_T^*-y_{2T}) \\ \vdots & \vdots & \vdots & \vdots \\ \gamma_{N1}+w_{N1}^{*\prime}\gamma_1^*-y_{N1} & \gamma_{N2}+w_{N2}^{*\prime}\gamma_2^*-y_{N2} & \cdots & (\gamma_{NT}+w_{NT}^{*\prime}\gamma_T^*-y_{NT}) \\ \sum_{i=1}^N w_{i1}^*(\gamma_{i1}+w_{i1}^{*\prime}\gamma_1^*-y_{i1}) & \sum_{i=1}^N w_{i2}^*(\gamma_{i2}+w_{i2}^{*\prime}\gamma_2^*-y_{i2}) & \cdots & \sum_{i=1}^N w_{iT}^*(\gamma_{iT}+w_{iT}^{*\prime}\gamma_T^*-y_{iT}) \end{pmatrix}, \tag{B.9}$$

and for $\Gamma^{(1)} \equiv (\gamma_{it}^{(1)}, \gamma_t^{*(1)}) \in \mathbf{R}^{(N+p-1)\times T}$ and $\Gamma^{(2)} \equiv (\gamma_{it}^{(2)}, \gamma_t^{*(2)}) \in \mathbf{R}^{(N+p-1)\times T}$,

$$\|\nabla f(\Gamma^{(1)}) - \nabla f(\Gamma^{(2)})\|_F^2$$

$$
= \left\| \begin{pmatrix}
\gamma_{11}^{(1)} - \gamma_{11}^{(2)} + w_{11}^{*\prime}(\gamma_1^{*(1)} - \gamma_1^{*(2)}) \\
\gamma_{21}^{(1)} - \gamma_{21}^{(2)} + w_{21}^{*\prime}(\gamma_1^{*(1)} - \gamma_1^{*(2)}) \\
\vdots \\
\gamma_{N1}^{(1)} - \gamma_{N1}^{(2)} + w_{N1}^{*\prime}(\gamma_1^{*(1)} - \gamma_1^{*(2)}) \\
\sum_{i=1}^N w_{i1}^*(\gamma_{i1}^{(1)} - \gamma_{i1}^{(2)}) + \sum_{i=1}^N w_{i1}^* w_{i1}^{*\prime}(\gamma_1^{*(1)} - \gamma_1^{*(2)})
\end{pmatrix} \right.
$$

$$
\begin{pmatrix}
\gamma_{12}^{(1)} - \gamma_{12}^{(2)} + w_{12}^{*\prime}(\gamma_2^{*(1)} - \gamma_2^{*(2)}) \\
\gamma_{22}^{(1)} - \gamma_{22}^{(2)} + w_{22}^{*\prime}(\gamma_2^{*(1)} - \gamma_2^{*(2)}) \\
\vdots \\
\gamma_{N2}^{(1)} - \gamma_{N2}^{(2)} + w_{N2}^{*\prime}(\gamma_2^{*(1)} - \gamma_2^{*(2)}) \\
\sum_{i=1}^N w_{i2}^*(\gamma_{i2}^{(1)} - \gamma_{i2}^{(2)}) + \sum_{i=1}^N w_{i2}^* w_{i2}^{*\prime}(\gamma_2^{*(1)} - \gamma_2^{*(2)})
\end{pmatrix}
$$

$$
\left. \begin{matrix}
\cdots & \gamma_{1T}^{(1)} - \gamma_{1T}^{(2)} + w_{1T}^{*\prime}(\gamma_T^{*(1)} - \gamma_T^{*(2)}) \\
\cdots & \gamma_{2T}^{(1)} - \gamma_{2T}^{(2)} + w_{2T}^{*\prime}(\gamma_T^{*(1)} - \gamma_T^{*(2)}) \\
\vdots & \vdots \\
\cdots & \gamma_{NT}^{(1)} - \gamma_{NT}^{(2)} + w_{NT}^{*\prime}(\gamma_T^{*(1)} - \gamma_T^{*(2)}) \\
\cdots & \sum_{i=1}^N w_{iT}^*(\gamma_{iT}^{(1)} - \gamma_{iT}^{(2)}) + \sum_{i=1}^N w_{iT}^* w_{iT}^{*\prime}(\gamma_T^{*(1)} - \gamma_T^{*(2)})
\end{matrix} \right) \right\|_F^2
$$

$$
\begin{aligned}
&= \sum_{i=1}^N \sum_{t=1}^T \left[\gamma_{it}^{(1)} - \gamma_{it}^{(2)} + w_{it}^{*\prime}(\gamma_t^{*(1)} - \gamma_t^{*(2)})\right]^2 \\
&\quad + \sum_{t=1}^T \left\| \sum_{i=1}^N w_{it}^*(\gamma_{it}^{(1)} - \gamma_{it}^{(2)}) + \sum_{i=1}^N w_{iT}^* w_{it}^{*\prime}(\gamma_t^{*(1)} - \gamma_t^{*(2)}) \right\|^2 \\
&\leq 2\sum_{i=1}^N \sum_{t=1}^T (\gamma_{it}^{(1)} - \gamma_{it}^{(2)})^2 + 2\max_{t\leq T} \lambda_{\max}\left(\sum_{i=1}^N w_{it}^* w_{it}^{*\prime}\right) \sum_{t=1}^T \|\gamma_t^{*(1)} - \gamma_t^{*(2)}\|^2 \\
&\quad + 2N \max_{i\leq N, t\leq N} \|w_{it}^*\|^2 \sum_{i=1}^N \sum_{t=1}^T (\gamma_{it}^{(1)} - \gamma_{it}^{(2)})^2 + 2\max_{t\leq T} \lambda_{\max}^2\left(\sum_{i=1}^N w_{it}^* w_{it}^{*\prime}\right) \sum_{t=1}^T \|\gamma_t^{*(1)} - \gamma_t^{*(2)}\|^2 \\
&\leq 2\max\left\{1 + N \max_{i\leq N, t\leq N} \|w_{it}^*\|^2, \max_{t\leq T} \lambda_{\max}\left(\sum_{i=1}^N w_{it}^* w_{it}^{*\prime}\right) + \max_{t\leq T} \lambda_{\max}^2\left(\sum_{i=1}^N w_{it}^* w_{it}^{*\prime}\right)\right\} \\
&\qquad \times \|\Gamma^{(1)} - \Gamma^{(2)}\|_F^2 \\
&= 2\max\left\{1 + \max_{i\leq N, t\leq N} \|x_{it}^*\|^2, \max_{t\leq T} \lambda_{\max}\left(\frac{1}{N}\sum_{i=1}^N x_{it}^* x_{it}^{*\prime}\right) + \max_{t\leq T} \lambda_{\max}^2\left(\frac{1}{N}\sum_{i=1}^N x_{it}^* x_{it}^{*\prime}\right)\right\} \\
&\qquad \times \|\Gamma^{(1)} - \Gamma^{(2)}\|_F^2, \qquad \text{(B.10)}
\end{aligned}
$$

where the first inequality follows by the Cauchy Schwartz inequality and the triangle inequality. Thus, $\nabla f(\Gamma)$ is Lipschitz continuous with constant $L_f = \sqrt{2}[\max\{1 + \max_{i\leq N, t\leq N} \|x_{it}^*\|^2, \max_{t\leq T} \lambda_{\max}(\sum_{i=1}^N x_{it}^* x_{it}^{*\prime}/N) + \max_{t\leq T} \lambda_{\max}^2(\sum_{i=1}^N x_{it}^* x_{it}^{*\prime}/N)\}]^{1/2}$.

Appendix B.3 Example 2.4

For (24), we may set $m = p$, $\varphi_{NT} = \lambda_{0,NT}$ and

$$f(\Gamma) = \frac{1}{2}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - x_{it}'\gamma_t)^2 \text{ for } \Gamma \equiv (\gamma_1, \gamma_2, \ldots, \gamma_T) \in \mathbf{R}^{p\times T}. \tag{B.11}$$

We need to show that the gradient $\nabla f(\Gamma)$ is Lipschitz continuous. It follows that

$$\nabla f(\Gamma) = \left(\sum_{i=1}^{N} x_{i1}(x_{i1}'\gamma_1 - y_{i1}), \sum_{i=1}^{N} x_{i2}(x_{i2}'\gamma_2 - y_{i2}), \ldots, \sum_{i=1}^{N} x_{iT}(x_{iT}'\gamma_T - y_{iT})\right). \tag{B.12}$$

Indeed, $\nabla f(\Gamma)$ is Lipschitz continuous with constant $L_f = \max_{t\le T}\lambda_{\max}(\sum_{i=1}^{N} x_{it}x_{it}')$, because for $\Gamma^{(1)} \equiv (\gamma_1^{(1)}, \gamma_2^{(1)}, \ldots, \gamma_T^{(1)}) \in \mathbf{R}^{p\times T}$ and $\Gamma^{(2)} \equiv (\gamma_1^{(2)}, \gamma_2^{(2)}, \ldots, \gamma_T^{(2)}) \in \mathbf{R}^{p\times T}$,

$$\begin{aligned}
&\|\nabla f(\Gamma^{(1)}) - \nabla f(\Gamma^{(2)})\|_F^2 \\
&= \left\|\sum_{i=1}^{N} x_{i1}x_{i1}'(\gamma_1^{(1)} - \gamma_1^{(2)}), \sum_{i=1}^{N} x_{i2}x_{i2}'(\gamma_2^{(1)} - \gamma_2^{(2)}) \ldots, \sum_{i=1}^{N} x_{iT}x_{iT}'(\gamma_T^{(1)} - \gamma_T^{(2)})\right\|_F^2 \\
&= \sum_{t=1}^{T}\left\|\sum_{i=1}^{N} x_{it}x_{it}'(\gamma_t^{(1)} - \gamma_t^{(2)})\right\|^2 \\
&\le \max_{t\le T}\lambda_{\max}^2\left(\sum_{i=1}^{N} x_{it}x_{it}'\right)\|\Gamma^{(1)} - \Gamma^{(2)}\|_F^2.
\end{aligned} \tag{B.13}$$

Appendix C Useful Lemmas

Lemma C.1. *(i) Let $\{\xi_{Nt}\}_{t\le T}$ be a sequence of independent $Np\times 1$ sub-Gaussian vectors with $\lambda_{\max}(E[\xi_{Nt}\xi_{Nt}'])$ bounded. Assume that $(x_{1t}'\varepsilon_{1t}, x_{2t}'\varepsilon_{2t}, \ldots, x_{Nt}'\varepsilon_{Nt})'$ is the tth column of $\Xi_{NT}\Omega_{NT}$, where $\Xi_{NT} = (\xi_{N1}, \xi_{N2}, \ldots, \xi_{NT})$ and Ω_{NT} is a $T\times T$ deterministic (possibly non-diagonal) matrix with $\|\Omega_{NT}\|_2$ bounded. Then as $(N,T)\to\infty$,*

$$\left\|\sum_{i=1}^{N}\sum_{t=1}^{T} X_{it}\varepsilon_{it}\right\|_2 = O_p(\max\{\sqrt{Np}, \sqrt{T}\}).$$

(ii) Let $\{\nu_{Nt}\}_{t\le T}$ be a sequence of independent $Np\times 1$ sub-Gaussian vectors with bounded $\lambda_{\max}(E[\nu_{Nt}\nu_{Nt}'])$. Assume that $(x_{1t}', x_{2t}', \ldots, x_{Nt}')'$ is the tth column of $\mathcal{V}_{NT}\Omega_{NT}$, where $\mathcal{V}_{NT} = (\nu_{N1}, \nu_{N2}, \ldots, \nu_{NT})$ and Ω_{NT} is a $T\times T$ deterministic (possibly non-diagonal) matrix with $\|\Omega_{NT}\|_2$ bounded. Then as $(N,T)\to\infty$,

$$\left\|\sum_{i=1}^{N}\sum_{t=1}^{T} X_{it}\right\|_2 = O_p(\max\{\sqrt{Np}, \sqrt{T}\}).$$

(iii) Let $\{\eta_{Nt}\}_{t\leq T}$ be a sequence of independent $p\times 1$ sub-Gaussian vectors with bounded $\lambda_{\max}(E[\eta_{Nt}\eta_{Nt}'])$. Assume that $\sum_{i=1}^N x_{it}\varepsilon_{it}/\sqrt{N}$ is the tth column of $\Upsilon_{NT}\Omega_{NT}$, where $\Upsilon_{NT}\equiv(\eta_{N1},\eta_{N2},\ldots,\eta_{NT})$ and Ω_{NT} is a $T\times T$ deterministic (possibly non-diagonal) matrix with $\|\Omega_{NT}\|_2$ bounded. Then as $(N,T)\to\infty$,

$$\left\|\left(\frac{1}{\sqrt{N}}\sum_{i=1}^N x_{i1}\varepsilon_{i1},\frac{1}{\sqrt{N}}\sum_{i=1}^N x_{i2}\varepsilon_{i2},\ldots,\frac{1}{\sqrt{N}}\sum_{i=1}^N x_{iT}\varepsilon_{iT}\right)\right\|_2=O_p(\max\{\sqrt{p},\sqrt{T}\}).$$

PROOF: (i) Since $(x_{1t}'\varepsilon_{1t},x_{2t}'\varepsilon_{2t},\ldots,x_{Nt}'\varepsilon_{Nt})'$ is the tth column of $\sum_{i=1}^N\sum_{t=1}^T X_{it}\varepsilon_{it}$,

$$\sum_{i=1}^N\sum_{t=1}^T X_{it}\varepsilon_{it}=\Xi_{NT}\Omega_{NT}. \tag{C.1}$$

Applying Theorem 5.39 and Remark 5.40 in Vershynin (2010) on Ξ_{NT}', we obtain $\|\Xi_{NT}\|_2=O_p(\max\{\sqrt{Np},\sqrt{T}\})$ as $(N,T)\to\infty$. Thus, the result follows by (C.1) since $\|\Omega_{NT}\|_2$ is bounded and $\|CD\|_2\leq\|C\|_2\|D\|_2$.

(ii) and (iii) The proof is similar to the proof of (i), thus omitted. ■

Lemma C.2. *Let $\mathcal{F}\equiv(\sum_{i=1}^N x_{i1}\varepsilon_{i1}/\sqrt{N},\sum_{i=1}^N x_{i2}\varepsilon_{i2}/\sqrt{N},\ldots,\sum_{i=1}^N x_{iT}\varepsilon_{iT}/\sqrt{N})$. For any $\Delta\in\{1_N\otimes\Gamma:\Gamma\in\mathbf{R}^{p\times T}\}$, we have*

$$\sum_{i=1}^N\sum_{t=1}^T|\mathrm{tr}(X_{it}'\Delta)|^2\geq\min_{t\leq T}\lambda_{\min}\left(\frac{\sum_{i=1}^N x_{it}x_{it}'}{N}\right)\|\Delta\|_F^2$$

and

$$\left|\sum_{i=1}^N\sum_{t=1}^T\mathrm{tr}(\varepsilon_{it}X_{it}'\Delta)\right|\leq\|\mathcal{F}\|_2\|\Delta\|_*.$$

PROOF: Fix $\Delta=1_N\otimes\Gamma$ for some $\Gamma\in\mathbf{R}^{p\times T}$. Write $\Gamma=(\gamma_1,\gamma_2,\ldots,\gamma_T)$, where γ_t is a $p\times 1$ vector. Since $\mathrm{tr}(X_{it}'\Delta)=x_{it}'\gamma_t$, it follows that

$$\begin{aligned}\sum_{i=1}^N\sum_{t=1}^T|\mathrm{tr}(X_{it}'\Delta)|^2&=\sum_{i=1}^N\sum_{t=1}^T|x_{it}'\gamma_t|^2\\&=N\sum_{t=1}^T\gamma_t'\left(\frac{\sum_{i=1}^N x_{it}x_{it}'}{N}\right)\gamma_t\\&\geq\min_{t\leq T}\lambda_{\min}\left(\frac{\sum_{i=1}^N x_{it}x_{it}'}{N}\right)N\|\Gamma\|_F^2\\&=\min_{t\leq T}\lambda_{\min}\left(\frac{\sum_{i=1}^N x_{it}x_{it}'}{N}\right)\|\Delta\|_F^2,\end{aligned} \tag{C.2}$$

where the last equality holds since $\|\Delta\|_F^2 = N\|\Gamma\|_F^2$. For the same reason, we have

$$\begin{aligned}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathrm{tr}(\varepsilon_{it}X_{it}'\Delta) &= \sum_{i=1}^{N}\sum_{t=1}^{T}\varepsilon_{it}x_{it}'\gamma_t \\ &= \mathrm{tr}(\mathcal{F}'\sqrt{N}\Gamma) \\ &\leq \|\mathcal{F}\|_2\sqrt{N}\|\Gamma\|_* \\ &= \|\mathcal{F}\|_2\|\Delta\|_*, \end{aligned} \tag{C.3}$$

where the inequality holds by the fact that $|\mathrm{tr}(C'D)| \leq \|C\|_2\|D\|_*$, and the last equality follows by Lemma C.6(iii). This completes the proof of the lemma. ■

Lemma C.3. *For any $\Gamma = (\gamma_1, \gamma_2, \ldots, \gamma_T) \in \mathbf{R}^{p\times T}$, we have*

$$\frac{1}{2}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - \mathrm{tr}(X_{it}'(1_N \otimes \Gamma)))^2 + \lambda_{NT}\|1_N \otimes \Gamma\|_* = \frac{1}{2}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - x_{it}'\gamma_t)^2 + \sqrt{N}\lambda_{NT}\|\Gamma\|_*.$$

Proof: Fix $\Gamma = (\gamma_1, \gamma_2, \ldots, \gamma_T) \in \mathbf{R}^{p\times T}$. It is easy to see that $\mathrm{tr}(X_{it}'(1_N \otimes \Gamma)) = x_{it}'\gamma_t$. By Lemma C.6(iii), $\|1_N \otimes \Gamma\|_* = \sqrt{N}\|\Gamma\|_*$. Thus, the result follows. ■

Lemma C.4. *Recall $x_{it} = (1, x_{it}^{*\prime})'$. Let $\mathcal{X}^*$ be an $N \times T$ block matrix with the itth block x_{it}^*, $\mathcal{E}$ be an $N \times T$ matrix with the itth entry ε_{it}, and $\mathcal{F}^* \equiv (\sum_{i=1}^N x_{i1}^*\varepsilon_{i1}/\sqrt{N}$, $\sum_{i=1}^N x_{i2}^*\varepsilon_{i2}/\sqrt{N}, \ldots, \sum_{i=1}^N x_{iT}^*\varepsilon_{iT}/\sqrt{N})$. For any $\Delta \in \mathcal{D}_M$ given in (15), we have*

$$\sum_{i=1}^{N}\sum_{t=1}^{T}|\mathrm{tr}(X_{it}'\Delta)|^2 \geq \min\left\{1, \min_{t\leq T}\lambda_{\min}\left(\frac{\sum_{i=1}^N x_{it}^*x_{it}^{*\prime}}{N}\right)\right\}\|\Delta\|_F^2 + 2\mathcal{R}_{NT}(\Delta)$$

for some $\mathcal{R}_{NT}(\Delta)$ such that $|\mathcal{R}_{NT}(\Delta)| \leq M\sqrt{p-1}\|\mathcal{X}^*\|_2\|\Delta\|_*$, and

$$\left|\sum_{i=1}^{N}\sum_{t=1}^{T}\mathrm{tr}(\varepsilon_{it}X_{it}'\Delta)\right| \leq (\|\mathcal{E}\|_2 + \|\mathcal{F}^*\|_2)\|\Delta\|_*.$$

Proof: Fix $\Delta = ((\gamma_1, \Gamma^{*\prime}), (\gamma_2, \Gamma^{*\prime}), (\gamma_N, \Gamma^{*\prime}))' \in \mathcal{D}_M$ for some $(\gamma_1, \gamma_2, \ldots, \gamma_N)' \in \mathbf{R}^{N\times T}$ and $\Gamma^* \in \mathbf{R}^{(p-1)\times T}$. Write $\gamma_i = (\gamma_{i1}, \gamma_{i2}, \ldots, \gamma_{iT})'$ and $\Gamma^* = (\gamma_1^*, \gamma_2^*, \ldots, \gamma_T^*)$, where γ_{it} is a scalar and γ_t^* is a $(p-1)\times 1$ vector. Since $x_{it} = (1, x_{it}^*)'$, it follows that $\mathrm{tr}(X_{it}'\Delta) = \gamma_{it} + x_{it}^{*\prime}\gamma_t^*$ and then

$$\begin{aligned}&\sum_{i=1}^{N}\sum_{t=1}^{T}|\mathrm{tr}(X_{it}'\Delta)|^2 = \sum_{i=1}^{N}\sum_{t=1}^{T}(\gamma_{it} + x_{it}^{*\prime}\gamma_t^*)^2 \\ &= \sum_{i=1}^{N}\sum_{t=1}^{T}\gamma_{it}^2 + N\sum_{t=1}^{T}\gamma_t^{*\prime}\left(\frac{\sum_{i=1}^N x_{it}^*x_{it}^{*\prime}}{N}\right)\gamma_t^* + 2\sum_{i=1}^{N}\sum_{t=1}^{T}\gamma_{it}x_{it}^{*\prime}\gamma_t^* \\ &\geq \min\left\{1, \min_{t\leq T}\lambda_{\min}\left(\frac{\sum_{i=1}^N x_{it}^*x_{it}^{*\prime}}{N}\right)\right\}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\gamma_{it}^2 + N\|\Gamma^*\|_F^2\right) + 2\sum_{i=1}^{N}\sum_{t=1}^{T}\gamma_{it}x_{it}^{*\prime}\gamma_t^*\end{aligned}$$

$$= \min\left\{1, \min_{t\leq T} \lambda_{\min}\left(\frac{\sum_{i=1}^N x_{it}^* x_{it}^{*\prime}}{N}\right)\right\} \|\Delta\|_F^2 + 2\sum_{i=1}^N\sum_{t=1}^T \gamma_{it} x_{it}^{*\prime}\gamma_t^*, \tag{C.4}$$

where the last equality holds since $\|\Delta\|_F^2 = \sum_{i=1}^N \sum_{t=1}^T \gamma_{it}^2 + N\|\Gamma^*\|_F^2$. Write $x_{it}^* = (x_{it,1}^*, x_{it,2}^*, \ldots, x_{it,p-1}^*)'$ and $\gamma_t^* = (\gamma_{1t}^*, \gamma_{2t}^*, \ldots, \gamma_{(p-1)t}^*)$. Let $\Gamma^\diamond \equiv (\gamma_1, \gamma_2, \ldots, \gamma_N)'$, $\Gamma_j^\dagger \equiv \Gamma^\diamond \text{diag}(\gamma_{j1}^*, \gamma_{j2}^*, \ldots, \gamma_{jT}^*)$, and X_j^* be an $N \times T$ matrix with the itth entry $x_{it,j}^*$. Write $\Gamma^\diamond = (\zeta_1, \zeta_2, \ldots, \zeta_T)$. It follows that

$$\begin{aligned}
\sum_{i=1}^N\sum_{t=1}^T \gamma_{it} x_{it}^{*\prime}\gamma_t^* &= \sum_{j=1}^{p-1}\sum_{i=1}^N\sum_{t=1}^T \gamma_{it} x_{it,j}^*\gamma_{jt}^* \\
&= \sum_{j=1}^{p-1} \text{tr}(X_j^{*\prime}\Gamma_j^\dagger) \\
&= \text{tr}\left(\begin{pmatrix} X_1^* \\ X_2^* \\ \vdots \\ X_{p-1}^* \end{pmatrix}' \begin{pmatrix} \Gamma_1^\dagger \\ \Gamma_2^\dagger \\ \vdots \\ \Gamma_{p-1}^\dagger \end{pmatrix}\right) \\
&\leq \left\|\begin{pmatrix} X_1^* \\ X_2^* \\ \vdots \\ X_{p-1}^* \end{pmatrix}\right\|_2 \left\|\begin{pmatrix} \Gamma_1^\dagger \\ \Gamma_2^\dagger \\ \vdots \\ \Gamma_{p-1}^\dagger \end{pmatrix}\right\|_* \\
&= \|\mathcal{X}^*\|_2 \left\|\begin{pmatrix} \Gamma_1^\dagger \\ \Gamma_2^\dagger \\ \vdots \\ \Gamma_{p-1}^\dagger \end{pmatrix}\right\|_* \\
&\leq \max_{j\leq p-1, t\leq T} |\gamma_{jt}^*| \sum_{j=1}^{p-1} \sqrt{p-1}\|\mathcal{X}^*\|_2 \|\Gamma_j^\dagger\|_*,
\end{aligned} \tag{C.5}$$

where the first inequality holds by the fact that $|\text{tr}(C'D)| \leq \|C\|_2\|D\|_*$, the fourth equality holds since $\mathcal{X}^*$ and $(X_1^{*\prime}, X_2^{*\prime}, \ldots, X_{p-1}^{*\prime})'$ share a common set of nonzero singular values, the last inequality follows since the nonzero singular values of $(\Gamma_1^{\dagger\prime}, \Gamma_2^{\dagger\prime}, \ldots, \Gamma_{p-1}^{\dagger\prime})'$ are given by the square root of the nonzero eigenvalues of

$$(\Gamma_1^{\dagger\prime}, \Gamma_2^{\dagger\prime}, \ldots, \Gamma_{p-1}^{\dagger\prime}) \begin{pmatrix} \Gamma_1^\dagger \\ \Gamma_2^\dagger \\ \vdots \\ \Gamma_{p-1}^\dagger \end{pmatrix} = \sum_{j=1}^{p-1} \Gamma_j^{\dagger\prime}\Gamma_j^\dagger$$

$$= \sum_{j=1}^{p-1} \begin{pmatrix} \gamma_{j1}^* & 0 & \cdots & 0 \\ 0 & \gamma_{j2}^* & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \gamma_{jT}^* \end{pmatrix} \Gamma^{\diamond\prime}\Gamma^{\diamond} \begin{pmatrix} \gamma_{j1}^* & 0 & \cdots & 0 \\ 0 & \gamma_{j2}^* & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \gamma_{jT}^* \end{pmatrix}$$

$$= \sum_{j=1}^{p-1}\sum_{t=1}^{T} \gamma_{jt}^{*2}\zeta_t\zeta_t' \preceq \max_{j\le p-1, t\le T} |\gamma_{jt}^*|^2 \sum_{j=1}^{p-1}\sum_{t=1}^{T} \zeta_t\zeta_t' = (p-1) \max_{j\le p-1, t\le T} |\gamma_{jt}^*|^2 \Gamma^{\diamond\prime}\Gamma^{\diamond}, \quad \text{(C.6)}$$

and "$C \preceq D$" means that $D - C$ is positive semi-definite. Thus, the first result of the lemma follows from (C.4) and (C.5) by letting $\mathcal{R}_{NT}(\Delta) = \sum_{i=1}^{N}\sum_{t=1}^{T} \gamma_{it} x_{it}^{*\prime}\gamma_t^*$. Since $\mathrm{tr}(X_{it}'\Delta) = \gamma_{it} + x_{it}^{*\prime}\gamma_t^*$,

$$\begin{aligned} \sum_{i=1}^{N}\sum_{t=1}^{T} \mathrm{tr}(\varepsilon_{it} X_{it}'\Delta) &= \sum_{i=1}^{N}\sum_{t=1}^{T} \varepsilon_{it}\gamma_{it} + \sum_{i=1}^{N}\sum_{t=1}^{T} \varepsilon_{it} x_{it}^{*\prime}\gamma_t^* \\ &= \mathrm{tr}(\mathcal{E}'\Gamma) + \mathrm{tr}\left(\mathcal{F}^{*\prime}\sqrt{N}\Gamma^*\right) \\ &\le \|\mathcal{E}\|_2 \|\Gamma\|_* + \|\mathcal{F}^*\|_2 \sqrt{N}\|\Gamma^*\|_* \\ &\le (\|\mathcal{E}\|_2 + \|\mathcal{F}^*\|_2) \left\| \begin{pmatrix} \Gamma \\ \sqrt{N}\Gamma^* \end{pmatrix} \right\| \\ &= (\|\mathcal{E}\|_2 + \|\mathcal{F}^*\|_2) \|\Delta\|_*, \end{aligned} \quad \text{(C.7)}$$

where the first inequality holds by the fact that $|\mathrm{tr}(C'D)| \le \|C\|_2 \|D\|_*$, the second inequality follows since $\|\Gamma\|_* \le \|(\Gamma', \sqrt{N}\Gamma^{*\prime})'\|_*$ and $\sqrt{N}\|\Gamma^*\|_* \le \|(\Gamma', \sqrt{N}\Gamma^{*\prime})'\|_*$, and the last equality follows by Lemma C.7(iii). This completes the proof of the lemma. ■

Lemma C.5. *Recall $x_{it} = (1, x_{it}^{*\prime})'$. For any $\Gamma^{\diamond} = (\gamma_1, \gamma_2, \ldots, \gamma_N)' \in \mathbf{R}^{N\times T}$ and $\Gamma^* = (\gamma_1^*, \gamma_2^*, \ldots, \gamma_T^*)' \in \mathbf{R}^{(p-1)\times T}$, we have*

$$\begin{aligned} &\frac{1}{2}\sum_{i=1}^{N}\sum_{t=1}^{T} \left(y_{it} - \mathrm{tr}\left(X_{it}' \begin{pmatrix} \gamma_1' \\ \Gamma^* \\ \gamma_2' \\ \Gamma^* \\ \vdots \\ \gamma_N' \\ \Gamma^* \end{pmatrix} \right) \right)^2 + \lambda_{NT} \left\| \begin{pmatrix} \gamma_1' \\ \Gamma^* \\ \gamma_2' \\ \Gamma^* \\ \vdots \\ \gamma_N' \\ \Gamma^* \end{pmatrix} \right\|_* \\ &= \frac{1}{2}\sum_{i=1}^{N}\sum_{t=1}^{T} (y_{it} - \gamma_{it} - x_{it}^{*\prime}\gamma_t^*)^2 + \lambda_{NT} \left\| \begin{pmatrix} \Gamma^{\diamond} \\ \sqrt{N}\Gamma^* \end{pmatrix} \right\|_*, \end{aligned}$$

where $\gamma_i = (\gamma_{i1}, \gamma_{i2}, \ldots, \gamma_{iT})'$.

Proof: Fix $\Gamma^{\diamond} = (\gamma_1, \gamma_2, \ldots, \gamma_N)' \in \mathbf{R}^{N\times T}$ and $\Gamma^* = (\gamma_1^*, \gamma_2^*, \ldots, \gamma_T^*)' \in \mathbf{R}^{(p-1)\times T}$. It is easy to see that $\mathrm{tr}(X_{it}'((\gamma_1, \Gamma^{*\prime}), (\gamma_2, \Gamma^{*\prime}), (\gamma_N, \Gamma^{*\prime}))') = \gamma_{it} + x_{it}^{*\prime}\gamma_t^*$. By Lemma C.7(iii),

$\|((\gamma_1, \Gamma^{*\prime}), (\gamma_2, \Gamma^{*\prime}), (\gamma_N, \Gamma^{*\prime}))'\|_* = \|(\Gamma^{\diamond\prime}, \sqrt{N}\Gamma^*)'\|_*$. Thus, the result follows. ■

Appendix C.1 Technical Lemmas

Lemma C.6. *For any matrix A, (i) the rank of $1_k \otimes A$ is equal to the rank of A; (ii) the nonzero singular values of $1_k \otimes A$ are equal to the nonzero singular values of A multiplied by $\sqrt{k}$; (iii) $\|1_k \otimes A\|_* = \sqrt{k}\|A\|_*$; (iv) the left singular vector matrix of nonzero matrix $1_k \otimes A$ corresponding to its nonzero singular values are given by $1_k \otimes U/\sqrt{k}$, where U is the left singular vector matrix of A corresponding to its nonzero singular values.*

Proof: It is without loss of generality to assume that A is nonzero. Let $d > 0$ be the rank of A and $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_d > 0$ be nonzero singular values of A. Let $A = U\Sigma V'$ be a singular value decomposition of A, where Σ is a $d \times d$ diagonal matrix with σ_j's in the diagonal in descending order. It follows that

$$1_k \otimes A = \frac{1}{\sqrt{k}}(1_k \otimes U)\sqrt{k}\Sigma V', \tag{C.8}$$

which gives a singular value decomposition of $1_k \otimes A$. Thus, the rank of $1_k \otimes A$ is equal to d, the nonzero singular values of $1_k \otimes A$ given by $\sqrt{k}\sigma_1 \geq \sqrt{k}\sigma_2 \geq \ldots \geq \sqrt{k}\sigma_d > 0$, and the left singular vector matrix of $1_k \otimes A$ corresponding to its nonzero singular values is $1_k \otimes U/\sqrt{k}$. This completes the proof of the lemma. ■

Lemma C.7. *For any matrices $C = (c_1, c_2, \ldots, c_k)'$ and D with the same number of columns where c_j's are column vectors, (i) the rank of $(c_1, D', c_2, D', \ldots, c_k, D')$ is equal to the rank of $(C', \sqrt{k}D')$; (ii) the nonzero singular values of $(c_1, D', c_2, D', \ldots, c_k, D')$ are equal to the nonzero singular values of $(C', \sqrt{k}D')$; (iii) $\|(c_1, D', c_2, D', \ldots, c_k, D')\|_* = \|(C', \sqrt{k}D')\|_*$; (iv) the left singular vector matrix of nonzero matrix $(c_1, D', c_2, D', \ldots, c_k, D')'$ corresponding to its nonzero singular values have the form of $(u_1, V', u_2, V', \ldots, u_k, V')'$, where $U = (u_1, u_2, \ldots, u_k)'$ and V have the same number of rows with C and D, respectively. Moreover, $(U', \sqrt{k}V')'$ is the left singular vector matrix of $(C', \sqrt{k}D')'$ corresponding to its nonzero singular values.*

Proof: It is without loss of generality to assume that C or D is nonzero. Let $d > 0$ be the rank of $(c_1, D', c_2, D', \ldots, c_k, D')$ and $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_d > 0$ be the nonzero singular values of $(c_1, D', c_2, D', \ldots, c_k, D')$. It follows that $\sigma_1^2 \geq \sigma_2^2 \geq \ldots \geq \sigma_d^2 > 0$ are

the nonzero eigenvalues of

$$(c_1, D', c_2, D', \ldots, c_k, D') \begin{pmatrix} c_1' \\ D \\ c_2' \\ D \\ \vdots \\ c_k' \\ D \end{pmatrix} = C'C + kD'D = (C', \sqrt{k}D') \begin{pmatrix} C \\ \sqrt{k}D \end{pmatrix}. \tag{C.9}$$

Thus, the nonzero singular values of $(C', \sqrt{k}D')$ are $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_d > 0$ and the rank of $(C', \sqrt{k}D')$ is equal to d. Let $(c_1, D', c_2, D', \ldots, c_k, D')' = U^* \Sigma V^{*\prime}$ be a singular value decomposition of $(c_1, D', c_2, D', \ldots, c_k, D')'$, where Σ is a $d \times d$ diagonal matrix with σ_j's in the diagonal in descending order. It follows that

$$U^* = \begin{pmatrix} c_1' \\ D \\ c_2' \\ D \\ \vdots \\ c_k' \\ D \end{pmatrix} V^* \Sigma^{-1} = \begin{pmatrix} c_1' V^* \Sigma^{-1} \\ DV^* \Sigma^{-1} \\ c_2' V^* \Sigma^{-1} \\ DV^* \Sigma^{-1} \\ \vdots \\ c_k' V^* \Sigma^{-1} \\ DV^* \Sigma^{-1} \end{pmatrix} = \begin{pmatrix} u_1' \\ V \\ u_2' \\ V \\ \vdots \\ u_k' \\ V \end{pmatrix}, \tag{C.10}$$

where $u_j = \Sigma^{-1} V^{*\prime} c_j$ and $V = DV^* \Sigma^{-1}$. In view of (C.9), V^* is also the right singular vector matrix of $(C', \sqrt{k}D')'$. Thus, the left singular vector matrix of $(C', \sqrt{k}D')'$ corresponding to its nonzero singular values is given by

$$\begin{pmatrix} C \\ \sqrt{k}D \end{pmatrix} V^* \Sigma^{-1} = \begin{pmatrix} CV^* \Sigma^{-1} \\ \sqrt{k}DV^* \Sigma^{-1} \end{pmatrix} = \begin{pmatrix} U \\ \sqrt{k}V \end{pmatrix}. \tag{C.11}$$

This completes the proof of the lemma. ■

References

Athey, S., M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi (2021): "Matrix completion methods for causal panel data models," *Journal of the American Statistical Association*, 116, 1716–1730.

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

Bai, J. and S. Ng (2002): "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70, 191–221.

——— (2019): "Rank regularized estimation of approximate factor models," *Journal of Econometrics*, 212, 78–96.

Bernstein, D. S. (2018): *Scalar, Vector, and Matrix Mathematics: Theory, Facts, and Formulas*, Princeton University Press, revised and expanded edition ed.

Bertsekas, D. P. (1999): *Nonlinear Programming*, Athena Scientific.

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

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

Candés, E. J. and B. Recht (2009): "Exact matrix completion via convex optimization," *Foundations of Computational mathematics*, 9, 717–772.

Chamberlain, G. and M. Rothschild (1982): "Arbitrage, factor structure, and mean-variance analysis on large asset markets," *Econometrica*, 51, 1281–1304.

Chen, Q., N. Roussanov, and X. Wang (2021): "Semiparametric Conditional Factor Models: Estimation and Inference," Tech. rep., arXiv preprint arXiv:2112.07121.

Chernozhukov, V., C. Hansen, Y. Liao, and Y. Zhu (2018): "Inference for Heterogeneous Effects using Low-Rank Estimation of Factor Slopes," Tech. rep., MIT.

Cochrane, J. H. (2011): "Presidential address: Discount rates," *The Journal of finance*, 66, 1047–1108.

Connor, G., M. Hagmann, and O. Linton (2012): "Efficient semiparametric estimation of the Fama–French model and extensions," *Econometrica*, 80, 713–754.

Connor, G. and R. A. Korajczyk (1986): "Performance measurement with the arbitrage pricing theory: A new framework for analysis," *Journal of financial economics*, 15, 373–394.

Daniel, K. and S. Titman (1997): "Evidence on the characteristics of cross sectional variation in stock returns," *Journal of Finance*, 52, 1–33.

Fama, E. F. and K. R. French (1993): "Common risk factors in the returns on stocks and bonds," *Journal of financial economics*, 33, 3–56.

Fan, J., Y. Liao, and M. Mincheva (2013): "Large covariance estimation by thresholding principal orthogonal complements," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75, 603–680.

Fan, J., Y. Liao, and W. Wang (2016): "Projected principal component analysis in factor models," *The Annals of Statistics*, 44, 219–254.

Fazel, M. (2002): "Matrix rank minimization with applications," Ph.D. thesis, Stanford University.

Ferson, W. and C. Harvey (1999): "Conditioning variables and the cross section of stock returns," *The Journal of Finance*, 4, 1325–1360.

Ferson, W. E. and C. R. Harvey (1991): "The variation of economic risk premiums," *Journal of Political Economy*, 99, 385–415.

Freyberger, J., A. Neuhierl, and M. Weber (2020): "Dissecting characteristics

nonparametrically," *The Review of Financial Studies*, 33, 2326–2377.

Gagliardini, P., E. Ossola, and O. Scaillet (2016): "Time-varying risk premium in large cross-sectional equity data sets," *Econometrica*, 84, 985–1046.

——— (2020): "Estimation of large dimensional conditional factor models in finance," in *Handbook of Econometrics*, vol. 7A, chap. 3.

Golub, G. H. and C. F. Van Loan (2013): *Matrix computations*, Johns Hopkins Universtiy Press.

Gu, S., B. Kelly, and D. Xiu (2021): "Autoencoder asset pricing models," *Journal of Econometrics*, 222, 429–450.

Ji, S. and J. Ye (2009): "An accelerated gradient method for trace norm minimization," in *Proceedings of the 26th annual international conference on machine learning*, 457–464.

Kelly, B. T., S. Pruitt, and Y. Su (2017): "Instrumented principal component analysis," Tech. rep., Available at SSRN 2983919.

——— (2019): "Characteristics are covariances: A unified model of risk and return," *Journal of Financial Economics*, 134, 501–524.

Kim, S., R. A. Korajczyk, and A. Neuhierl (2020): "Arbitrage portfolios," *Review of Financial Studies*, Forthcoming.

Lettau, M. and S. Ludvigson (2001): "Consumption, aggregate wealth, and expected stock returns," *Journal of Finance*, 56, 815–849.

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

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

Moon, H. R. and M. Weidner (2018): "Nuclear Norm Regularized Estimation of Panel Regression Models," Tech. rep., arXiv preprint arXiv:1810.10987.

Nagel, S. and K. J. Singleton (2011): "Estimation and evaluation of conditional asset pricing models," *The Journal of Finance*, 66, 873–909.

Negahban, S., P. Ravikumar, M. Wainwright, and B. Yu (2012): "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers," *Statistical Science*, 27, 538–557.

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

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

Nesterov, Y. (1983): "A method for solving the convex programming problem with convergence rate $O(1/k^2)$," *Soviet Mathematics Doklady*, 27, 372–376.

——— (2003): *Introductory lectures on convex optimization: A basic course*, Springer.

Pelger, M. and R. Xiong (2021): "State-varying factor models of large dimensions,"

Journal of Business Economics & Statistics, Accpted.

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

Rohde, A. and A. B. Tsybakov (2011): "Estimation of high-dimensional low-rank matrices," *The Annals of Statistics*, 39, 887–930.

Ross, S. A. (1976): "The Arbitrage Theory of Capital Asset Pricing," *Journal of Economic Theory*, 13, 341–360.

Shanken, J. (1990): "Intertemporal asset pricing: An empirical investigation," *Journal of Econometrics*, 45, 99–120.

Stock, J. H. and M. W. Watson (2002): "Forecasting using principal components from a large number of predictors," *Journal of the American Statistical Association*, 97, 1167–1179.

Toh, K. C. and S. Yun (2010): "An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems," *Pacific Journal of optimization*, 15, 615–640.

Vandenberghe, L. and S. Boyd (1996): "Semidefinite programming," *SIAM review*, 38, 49–95.

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