Low-Rank Estimation of Nonlinear Panel Data Models*

Kan Yao†

November 24, 2025

Abstract

This paper investigates nonlinear panel regression models with interactive fixed effects and introduces a general framework for parameter estimation under potentially non-convex objective functions. We propose a computationally feasible two-step estimation procedure. In the first step, nuclear-norm regularization (NNR) is used to obtain preliminary estimators of the coefficients of interest, factors, and factor loadings. The second step involves an iterative procedure for post-NNR inference, improving the convergence rate of the coefficient estimator. We establish the asymptotic properties of both the preliminary and iterative estimators. We also study the determination of the number of factors. Monte Carlo simulations demonstrate the effectiveness of the proposed methods in determining the number of factors and estimating the model parameters. In our empirical application, we apply the proposed approach to study the cross-market arbitrage behavior of U.S. nonfinancial firms.

Keywords: Nonlinear panel data models, fixed effects, nuclear-norm regularization, bias correction, time-varying heterogeneity

*I am deeply grateful to my advisors, Yuichi Kitamura, Donald Andrews, and Xiaohong Chen, for their invaluable support and guidance. I also thank Timothy Christensen, Max Cytrynbaum, Kengo Kato, Ed Vytlacil, and participants of the Econometrics Prospectus Workshops at Yale for helpful comments. All errors are my own.

†Department of Economics, Yale University. Email: kan.yao@yale.edu.

1 Introduction

Unobserved heterogeneity is of broad interest in both reduced-form and structural work in economics and other social sciences. Empirical data often include individual units sampled over time from diverse backgrounds, with factors unobserved by econometricians. Accommodating this heterogeneity in a flexible, yet parsimonious manner is challenging but essential in practice.

Panel data, which involve observations on individual units over time, provide a valuable framework to model latent structures within low-dimensional manifolds. This is often achieved by incorporating unobserved individual and time effects into the model, controlling for unobserved covariates that remain either time or cross-sectionally invariant. A fixed effects approach imposes no distributional assumptions on these unobserved effects, allowing them to be arbitrarily related to observed covariates. A notable example is two-way fixed effects estimation for difference-in-differences, which has become a leading method in applied economics.

The additive structure, however, assumes that the influence of unobserved factors on individual units remains constant over time, failing to capture more complex dynamics. Alternatively, fixed effects can interact multiplicatively, giving rise to interactive fixed effects (IFEs) or factor structures. This multiplicative form provides a more flexible representation of heterogeneity, as it allows common time-varying shocks (factors) to affect cross-sectional units with individual-specific sensitivities (factor loadings). For example, they can account for aggregate shocks with heterogeneous impacts on agents in macroeconomic models or capture multidimensional individual heterogeneity with time-varying effects in microeconomic models. This flexibility motivated the discussion of interactive effects in the econometrics literature. Bai (2009) and Moon and Weidner (2015, 2017) study the linear regression model, treating individual and time effects as nuisance parameters estimated by least squares. Chen, Fernández-Val, and Weidner (2021) and Wang (2022) extend the least squares estimator to nonlinear panel data models with convex log-likelihood functions.

However, many important and widely used economic models, such as random coefficient logit panel models, do not have convex objective functions. The problem becomes even more challenging when interactive fixed effects are introduced to capture unobserved heterogeneity. Interactive fixed effect panel models with non-convex objective functions remain largely unexplored in the literature. The current approaches in the literature are not applicable as they rely on closed-form solutions or global convexity.

The present paper proposes a new method for econometric estimation and inference in panel models with interactive fixed effects. The main estimation procedure consists of

two steps. In the first step, we obtain preliminary consistent estimators for all coefficients via a nuclear-norm regularization (NNR) procedure. The second step uses an iterative procedure to conduct post-nuclear-norm regularization estimation for the parameters of interest, treating our consistent regularized estimator as the initial value. We demonstrate a faster convergence of our iterative estimator compared to the preliminary estimator. Furthermore, we derive the asymptotic distribution of our iterative estimator.

In this context, I make two contributions. First, our general M-estimation framework accommodates models defined by potentially non-convex objective functions with respect to the unknown parameters of interest, while the results in Bai (2009) and Chen, Fernández-Val, and Weidner (2021) rely on closed-form solutions or global convexity. Our theoretical findings, therefore, broaden the applicability of IFEs in econometrics, complementing the existing toolbox for applied economists when a factor structure is central to the analysis. Additionally, when the objective function is convex in the index, the proposed estimator formulates and solves a convex optimization problem, avoiding the issue of multiple local minima. Moreover, it does not require prior knowledge of the number of factors, and we also propose a consistent estimator for the number of factors.

The second contribution is to establish the large sample properties of the second-step iterative estimator. Specifically, this estimator exhibits a contraction mapping property, and thus converges to the true parameter value in probability at a rapid rate. Demonstrating the numerical convergence properties is technically challenging, as it requires controlling both the asymptotic bias and variance of the iterative estimator at each iteration. The iterative procedure in our second-step estimation is conceptually similar to that of Moon and Weidner (2019) and Hong, Su, and Jiang (2023), but we do not have closed-form solutions as they do in linear models. We show that the estimator obtained through the second step exhibits the same asymptotic distribution as found in Chen, Fernández-Val, and Weidner (2021) in the case of convex objective functions.

This paper relates to three branches of the literature. First, it adds to the extensive literature on panel data models with IFEs. Bai (2009) and Moon and Weidner (2015, 2017) propose estimators based on quasi-maximum likelihood estimation and principal component analysis. Much previous work has focused on linear models, while nonlinear models have recently attracted attention. Chen, Fernández-Val, and Weidner (2021) extend Fernández-Val and Weidner (2016)'s results to nonlinear panel data models with IFEs. Chen, Dolado, and Gonzalo (2021) study quantile factor models, and Ando, Bai, and Li (2022), Gao, Liu, Peng, and Yan (2023) study binary panel choice models. Relatedly, Boneva and Linton (2017) and Chen and Zhang (2025) extend the common correlated effects (CCE) framework of Pesaran (2006) to nonlinear settings. We refer to Fernández-Val and Weidner (2018) for

a recent survey. We complement and extend the literature by generalizing previous results to a broad range of M-estimators.

Second, our work contributes to the burgeoning literature on nuclear norm penalization, a technique that has gained popularity in estimating low-rank matrices in statistics and econometrics. This approach has been explored in various works such as Agarwal, Negahban, and Wainwright (2012), Agarwal, Dahleh, Shah, and Shen (2021), Athey, Bayati, Doudchenko, Imbens, and Khosravi (2021), Belloni, Chen, Madrid Padilla, and Wang (2023), Beyhum and Gautier (2019), Candès and Recht (2009), Chernozhukov, Hansen, Liao, and Zhu (2019, 2023), Fan, Gong, and Zhu (2017), Feng (2023), Hong, Su, and Jiang (2023), Huang and Wolkowicz (2018), Ma, Su, and Zhang (2021), Miao, Phillips, and Su (2022), Negahban and Wainwright (2011), Negahban, Ravikumar, Wainwright, and Yu (2012), among others. Most of these works focus on establishing estimation error bounds, specifically in the Frobenius norm, for the NNR estimators. The convergence rate of our estimator is consistent with the existing results of the mean and quantile regressions, indicating that NNR can be successfully extended to more complex models without compromising performance.

More recent studies have advanced statistical inference for NNR-based estimators. For instance, linear models have been studied by Armstrong, Weidner, and Zeleneev (2023), Chernozhukov, Hansen, Liao, and Zhu (2023), Hong, Su, and Jiang (2023), Miao, Phillips, and Su (2022), Moon and Weidner (2019), among others. Two main debiasing strategies have emerged to address shrinkage bias. The first, inspired by the debiased-Lasso framework in the statistics literature, typically requires relatively few iterations. Chernozhukov, Hansen, Liao, and Zhu (2023) combine a rotation-based debiasing step with sample splitting, while Choi, Kwon, and Liao (2024) show that sample splitting is not essential. Armstrong, Weidner, and Zeleneev (2023) apply minimax linear estimation theory to construct robust debiased estimators. The second line of research—where our paper contributes—employs iterative bias-correction procedures, as exemplified by Hong, Su, and Jiang (2023), Miao, Phillips, and Su (2022), and Moon and Weidner (2019). Two very recent papers developed independently and in parallel with ours examine nonlinear panel data models with interactive fixed effects: Chen, Miao, and Su (2025) derive the asymptotic distribution of factor and loading estimators in logistic panel models without covariates, and Zeleneev and Zhang (2025) study single-index models under global convexity of the transformation function. In contrast, our results do not require global convexity, which allows us to accommodate models such as random-coefficients specifications.

Lastly, this paper is also related to the literature on estimating the parameters of high-dimensional models using ℓ_1 regularization. Rather than listing all relevant papers, we refer

the interested reader to the comprehensive textbook by Hastie, Tibshirani, and Wainwright (2015) and the recent survey by Belloni, Chernozhukov, Chetverikov, Hansen, and Kato (2018), and instead focus on a few key references. Specifically, Chetverikov and Sørensen (2025) develop a method called bootstrapping after cross-validation for selecting the penalty parameter in ℓ_1-penalized M-estimators in high dimensions. Städler, Bühlmann, and van de Geer (2010) derive an oracle inequality for the ℓ_1-penalized approach to estimating mixture regression models, while Beyhum and Portier (2024) study high-dimensional nonconvex Lasso-type M-estimators and establish their convergence rates. However, none of these studies consider panel data settings or nuclear norm regularization.

An empirical illustration of our procedure is also provided, where we use multiple firm-level financial data sets to study US non-financial corporation joint financing decisions between debt issuance and equity repurchase, revisiting Ma (2019). The results show that our estimator produces economically sensible outcomes that align with financial intuition. For example, firms are more likely to issue equity and retire debt when the cost of debt is high and the cost of equity is low, consistent with the view that corporations act as cross-market arbitrageurs in their own securities. In addition, we document substantial heterogeneity in firm sensitivities to market valuation measures.

The remainder of the paper is organized as follows. Section 2 introduces the class of IFE models. Section 3 proposes a general M-estimation framework and presents the main estimation procedure. In Section 4, we examine the asymptotic properties of the estimators, including the consistency and rate of convergence of the initial estimator and the rank estimator, as well as the convergence and asymptotic distribution of the second-step estimator. Section 5 outlines the implementation algorithms and provides Monte Carlo results. Section 6 contains the empirical applications. Finally, Section 7 concludes. All proofs and additional results are provided in the Appendix.

Notation. For a natural number $m \in \mathbb{N}$, we introduce the notation $[m] = \{1, \ldots, m\}$. For a vector $v = (v_1, \ldots, v_p)' \in \mathbb{R}^p$, we define its ℓ_1-norm as $\|v\|_1 = \sum_{j=1}^p |v_j|$, and its ℓ_2-norm norm as $\|v\| = \sqrt{\sum_{j=1}^p v_j^2}$. For a matrix $A = (a_{it})_{i,t} \in \mathbb{R}^{N\times T}$, we define its entry-wise ℓ_1-norm as $\|A\|_1 = \sum_{i=1}^N \sum_{t=1}^T |a_{it}|$, its Frobenius norm as $\|A\|_F = \sqrt{\sum_{i=1}^N \sum_{t=1}^T a_{it}^2}$, its infinity norm as $\|A\|_\infty = \max\{|a_{it}| : i \in [N], t \in [T]\}$, its nuclear norm as $\|A\|_* = \text{trace}\left(\sqrt{A'A}\right)$, its spectral norm as $\|A\| = \sup_{x:\|x\|=1} \sqrt{x'A'Ax}$, and its rank by $\text{rank}(A)$. Moreover, for a real matrix, we use $\sigma_s(\cdot)$ to denote its s-th largest singular value. For a real and symmetric matrix, we use $\mu_s(\cdot), \mu_{\max}(\cdot)$ and $\mu_{\min}(\cdot)$ to denote its s-th largest eigenvalue, the largest and smallest eigenvalues, respectively.

For a real number a, let $\text{sgn}(a) = 1$ if $a \geq 0$ and $\text{sgn}(a) = -1$ if $a < 0$. For a square matrix A whose j-th diagonal element is denoted as A_{jj}, define $\text{sgn}(A)$ as a diagonal matrix whose j-th diagonal element is equal to $\text{sgn}(A_{jj})$. We also define the operations $\vee$ and $\wedge$ as $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. Finally, for sequences $\{a_m\}_{m=1}^{\infty}$ and $\{b_m\}_{m=1}^{\infty}$ we use $a_m \lesssim b_m$ as the shorthand for the inequality $a_m \leq \bar{c} b_m$ for some finite positive $\bar{c}$ independent of a_m and b_m for sufficiently large m. $a_m \asymp b_m$ means that $a_m \lesssim b_m$ and $b_m \lesssim a_m$.

2 Model

Let the data observations be denoted as $\{(Y_{it}, X_{it}) : i \in [N], t \in [T]\}$, where $Y_{it} \in \mathcal{Y} \subset \mathbb{R}$ is the outcome variable and $X_{it} \in \mathcal{X} \subset \mathbb{R}^{d_x}$ is a vector of exogenous covariates for a fixed d_x. The indices i and t denote individuals and time periods, respectively. In matrix notation, let Y and X_j as $N \times T$ matrices representing the outcome and the j-th covariate for $j = 1, ..., d_x$. Let $X = (X_j)_{j \in [d_x]}$ collect all covariates. The support of (Y_{it}, X_{it}) is given by $\mathcal{Y} \times \mathcal{X}$.

We assume that for each individual i and time t, the outcome Y_{it} is allowed to depend on both the observed covariates X_{it} and the latent interactive effects given by individual-specific loadings $\lambda_i \in \mathbb{R}^r$ and time factors $f_t \in \mathbb{R}^r$, where r is a fixed rank. The effects λ_i and f_t, although unobserved by the econometrician, may confound the effect of X_{it} on Y_{it}. Following the fixed effects approach, we treat the realizations of $\{\lambda_i\}_{i=1}^N$ and $\{f_t\}_{t=1}^T$ as unrestricted parameters to be estimated.

Formally, we consider a class of nonlinear panel data models in which the true values of the common parameter vector $\theta \in \mathbb{R}^p$ and the $N \times r$ and $T \times r$ matrices of fixed effects, $\Lambda = (\lambda_1, ..., \lambda_N)'$ and $F = (f_1, ..., f_T)'$, are defined by the solution to the population optimization problem,[1]

$$(\theta_0, \Lambda_0, F_0) = \underset{\theta \in \Theta, \Lambda \in \Phi_\lambda^{N \times r}, F \in \Phi_f^{T \times r}}{\arg\min} \mathbb{E}\left[\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \ell\left(W_{it}; \theta, \lambda_i' f_t\right)\right] \tag{2.1}$$

where $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is a known loss function, which may be nonconvex with respect to θ and $\lambda_i' f_t$, and where $W_{it} = (Y_{it}, X_{it})$. The sets $\Theta \subset \mathbb{R}^p$, $\Phi_\lambda \subset \mathbb{R}$, and $\Phi_f \subset \mathbb{R}$ are parameter spaces. The expectation $\mathbb{E}[\cdot]$ is taken with respect to the distribution of the data, conditional on the fixed realizations of unobserved individual and time effects for all

[1] The model is identified only up to a normalization; see Remark 1. A formal discussion of the identification assumptions is provided in Assumption 3.

N and T. It is also referred to as a factor model with factor loadings λ_i and common factors f_t, and we will use the terms "factor" and "interactive fixed effect" synonymously. The conventional additive structure is a special case of the factor structure with $r = 2$, $\lambda_i = (\lambda_{i1}, 1)'$, and $f_t = (1, f_{1t})'$.

Let $\pi_{0,it} = \lambda_{0i}' f_{0t}$ denote the true value of π_{it} that generates the data, where $\pi_{0,it} \in \Phi \subset \mathbb{R}$ for each i and t. The matrix $\Pi_0 \in \Phi^{N \times T}$ collects all the fixed effects. Similarly, the matrix $\Pi = (\pi_{it})$ is treated as a parameter to be estimated. Thus, we can rewrite Model (2.1) as

$$(\theta_0, \Pi_0) = \underset{\theta \in \Theta, \Pi \in \Phi^{N \times T}}{\arg\min} \mathbb{E}\left[\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \ell\left(W_{it}; \theta, \pi_{it}\right)\right] \tag{2.2}$$
$$\text{s.t. } \pi_{it} = \lambda_i' f_t = \sum_{s=1}^{r} \lambda_{is} f_{ts}$$

Remark 1. Similar to linear factor models, the factor loadings Λ_0 and common factors F_0 in (2.1) cannot be separately identified without imposing normalization. The loss function is invariant under the transformation $\lambda_i \mapsto A'\lambda_i$ and $f_t \mapsto A^{-1} f_t$ for any non-singular $r \times r$ matrix A. On the other hand, Π_0 in (2.2) can be uniquely identified, as it is invariant to the normalization used to eliminate this indeterminacy. For further discussion on normalization in the context of linear factor models, see Robertson and Sarafidis (2015).

2.1 Examples

Some examples of models that fall within this framework and the scope of our methodology are as follows.

Example 1 (Linear Panel)**.** *The linear panel model is generated according to*

$$y_{it} = X_{it}' \theta_0 + \pi_{0,it} + \varepsilon_{it} \tag{2.3}$$

where ε_{it} is the idiosyncratic error term with $\mathbb{E}\left[\varepsilon_{it} \mid X_{it}\right] = 0$. We can estimate θ and π_{it} using $\ell\left(W_{it}; \theta, \pi_{it}\right) = \frac{1}{2}\left(y_{it} - X_{it}'\theta - \pi_{it}\right)^2$ and $W_{it} = (y_{it}, X_{it})$.

Example 2 (Linear Quantile Panel)**.** *Consider the linear model in (2.3) with the quantile restriction: $\mathrm{P}\left(\varepsilon_{it} \leq 0 \mid X_{it}; \theta_0, \pi_{0,it}\right) = \tau$. We estimate θ and π_{it} by using $\ell\left(y_{it}, X_{it}'\theta + \pi_{it}\right) = \rho_\tau\left(y_{it} - X_{it}'\theta - \pi_{it}\right)$, where $\rho_\tau(t) = t\left(\tau - K\left(-t/h\right)\right)$ represents a smoothed version of the usual check function. Here, $\tau \in (0,1)$ denotes a specified quantile of interest, K is a CDF-type kernel function, and h is the bandwidth parameter.*

Example 3 (Binary Choice Panel)**.** *The binary choice panel model is characterized as*

$$y_{it} = \mathbb{I}\left\{X_{it}'\theta_0 + \pi_{0,it} - \varepsilon_{it} \geq 0\right\} \tag{2.4}$$

where X_{it} and ε_{it} are independent and ε_{it} is distributed according to a known cumulative distribution function F. Since $\mathrm{P}\left(y_{it} = 1 \mid z_{it}\right) = F\left(z_{it}\right)$, we define the objective function as $-\ell\left(W_{it}; \theta, \pi_{it}\right) = y_{it} \log F\left(X_{it}'\theta + \pi_{it}\right) + \left(1 - y_{it}\right) \log\left[1 - F\left(X_{it}'\theta + \pi_{it}\right)\right]$, where $W_{it} = \left(y_{it}, X_{it}\right)$.

Example 4 (Random Coefficients Logit Panel)**.** *Consider a binary choice model such that*

$$y_{it} = \mathbb{I}\left\{X_{it}'\beta_{0,it} + \pi_{0,it} - \varepsilon_{it} \geq 0\right\}$$

where ε_{it} is a random variable with the Type I extreme value distribution, independent across individuals i over time t, and $\beta_{0,it}$ is a p-dimensional vector of random coefficients, assumed to be i.i.d. across individuals over time.[2] *Specifically, we assume $\beta_{0,it} \sim \mathcal{N}\left(\bar{\beta}_0, \Sigma_0\right)$, where $\bar{\beta}_0$ is a p-dimensional mean vector, and Σ_0 is a positive-definite $p \times p$ covariance matrix. The theoretical results presented in this paper also hold for other distributions of $\beta_{0,it}$. Let $f\left(\beta\right)$ denote the probability density function of the random coefficient β_{it} evaluated at β. We define the parameter set $\theta = \left(\bar{\beta}, \Sigma\right)$. The objective function is given by*

$$\ell\left(W_{it}; \theta, \pi_{it}\right) = -\log\left[\int L_{it}(\beta, \pi_{it}) f\left(\beta\right) d\beta\right]$$

where the integrand is given by $L_{it}(\beta, \pi_{it}) = p_{it}(\beta, \pi_{it})^{y_{it}}(1 - p_{it}(\beta, \pi_{it}))^{1-y_{it}}$, with $p_{it}\left(\beta, \pi_{it}\right) = \frac{\exp\left(X_{it}'\beta + \pi_{it}\right)}{1+\exp\left(X_{it}'\beta + \pi_{it}\right)}$, and $W_{it} = \left(y_{it}, X_{it}\right)$.

3 Estimation

In this subsection, we describe our two-step estimation procedure for the model: the initial estimator is based on nuclear norm regularization (NNR). It is followed by a second step iterative estimator. As we will show in Sections 4.1 and 4.2, under a set of regularity conditions, the first step estimation provides consistent estimators for θ_0, Λ_0 and F_0. The second step uses the estimators from the first step as the initial values, and iteratively updates estimators to enhance their properties. The asymptotic normality of the iterative estimator is established in Section 4.3.

[2] We could relax the i.i.d. assumption by imposing a parametric model upon $\beta_{0,it}$ over time t, e.g., an AR(1) process, but we do not pursue this approach here.

3.1 First Step: Nuclear Norm Regularization

We define the loss function as

$$\mathcal{L}_{NT}(\theta, \Pi) \coloneqq \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \ell(W_{it}; \theta, \pi_{it}) \tag{3.1}$$

Motivated by the literature on nuclear norm regularization, we propose estimating (θ_0, Π_0) by minimizing the following penalized criterion function:

$$(\hat{\theta}, \hat{\Pi}) = \underset{\theta \in \Theta, \Pi \in \Phi^{N \times T}}{\arg\min} \mathcal{L}_{NT}(\theta, \Pi) + \nu \|\Pi\|_* \tag{3.2}$$

where ν is a tuning parameter,[3] and $\|\cdot\|_*$ denotes the nuclear norm, which regularizes the singular values of the matrix Π. The estimation problem at hand is inherently high-dimensional, as it involves estimating a total of $p + r(N + T)$ parameters, where the number of parameters grows linearly with N and T. Specifically, p represents the number of parameters associated with θ, and $r(N + T)$ corresponds to those related to the low-rank matrix Π.

Our estimator can be viewed as a generalization of Lasso, with a key difference being that we impose sparsity not directly on individual parameters but on the singular values of the matrix Π. This regularization ensures that the number of non-zero singular values of Π is relatively sparse compared to N and T. Since we assume that the rank r of the matrix Π_0 is fixed and much smaller than both N and T, the nuclear norm regularization reduces the complexity by encouraging a low-rank structure in Π, similar to how Lasso enforces sparsity in individual coefficients.

Similar to Lasso, (3.2) can be seen as a convex relaxation of the following problem:

$$\begin{aligned} &\min_{\theta \in \Theta, \Pi \in \Phi^{N \times T}} \mathcal{L}_{NT}(\theta, \Pi) \\ &\text{s.t.} \quad \operatorname{rank}(\Pi) \leq r \end{aligned} \tag{3.3}$$

where the rank r is predetermined, and θ, λ_i and f_t are jointly estimated, as studied by Bai (2009) and Chen, Fernández-Val, and Weidner (2021), among others. While the formulation in (3.3) seems appealing, it presents two challenges.

First, the rank constraint makes (3.3) a non-convex optimization problem, even when the loss function is convex. In contrast, our proposed estimator, defined in (3.2), avoids

[3] The tuning parameter ν depends on N and T, but we omit the subscript for simplicity.

directly penalizing or constraining the rank of the estimated interactive fixed effects matrix. Instead, it seeks a $\hat{\Pi}$ with a small nuclear norm, serving as a convex relaxation of (3.3). Just as ℓ_1 minimization is the tightest convex relaxation of the combinatorial ℓ_0 minimization problem, nuclear norm minimization offers the tightest convex relaxation of the NP-hard rank minimization problem (Candes and Plan, 2010).

Second, when the loss function $\mathcal{L}_{NT}$ is non-convex, as in random coefficients logit panel models, there are no guarantees of convergence for the estimators defined in (3.3). Global convexity plays a crucial role in the literature on interactive fixed effects. For instance, Chen, Fernández-Val, and Weidner (2021) rely on global convexity to establish the consistency of θ and to bound the remainder terms in their stochastic expansions. In contrast, our penalized estimator accommodates a broader class of important economic models without global convexity.

Note that the NNR-based initial estimation does not require the number of factors r to be known beforehand. When r is unknown, we propose estimating r using singular value thresholding (SVT) as follows:

$$\hat{r} = \sum_{s=1}^{N \wedge T} \mathbb{I}\left\{\sigma_s\left(\hat{\Pi}\right) \geq NT\nu\right\} \tag{3.4}$$

Alternatively, r can be estimated using other methods, such as the panel information criteria (IC) and panel criteria (PC) methods of Bai and Ng (2002), or the eigenvalue ratio (ER) and growth ratio (GR) methods of Ahn and Horenstein (2013), among others. These methods remain valid as long as $\hat{\theta}$ is consistent. In Section 4.2, we first establish the consistency of $\hat{\theta}$ and $\hat{\Pi}$, then we prove that $\hat{r} = r$ with probability approaching one (w.p.a.1). Consequently, for the second-step estimation, we assume r is known.

3.2 Second Step: Iterative Estimation

It is widely recognized that NNR estimators are subject to shrinkage bias, which complicates statistical inference. Theorem 1 establishes the convergence rate of the initial estimator, which, although consistent, is slower than the $\sqrt{NT}$ rate. To address this issue, we propose a post nuclear-norm-regularization estimator to improve the convergence rate.

For simplicity, we focus primarily on maximum likelihood estimation. We assume that the outcome is generated by

$$Y_{it} \mid X_{it}; \theta, \lambda_i, f_t \sim g\left(\cdot | X_{it}; \theta, \lambda_i, f_t\right)$$

where $g(\cdot)$ is a known probability density function with respect to some dominating measure. The log-likelihood function is then given by

$$\ell\left(W_{it};\theta,\lambda_i'f_t\right) = -\log g\left(Y_{it} \mid X_{it};\theta,\lambda_i,f_t\right)$$

Since the common factors F and factor loadings Λ cannot be separately identified without imposing normalization (see Remark 1), we impose the same normalization on both F and Λ as in Bai (2009), without loss of generality. Specifically, we normalize such that $F'F/T = \mathbb{I}_r$ and $\Lambda'\Lambda/N$ is diagonal with non-increasing diagonal elements. It is important to note that the ultimate asymptotic results remain unchanged regardless of the specific normalization chosen for F and Λ.

To refine the initial estimators, we employ a localized iterative procedure. Starting with the initial estimators $\left(\hat{\theta},\hat{\Lambda},\hat{F}\right)$, the procedure iteratively updates these estimators by searching within a local region of radius $d_{NT} = c\log\left(N\wedge T\right)\gamma_{NT}$, where c is a positive constant, and γ_{NT} represents the convergence rate of the first-step estimator, as specified in Corollary 1 below. The procedure is defined as follows:

Step 1. For $m=0$, set $\left(\hat{\theta}^{(0)},\hat{\Lambda}^{(0)},\hat{F}^{(0)}\right) = \left(\hat{\theta},\hat{\Lambda},\hat{F}\right)$, the preliminary consistent estimator of (θ_0,Λ_0,F_0).

Step 2. Given $\hat{\theta}^{(m)}$, update the estimators of $\{\Lambda_0,F_0\}$ to $\left\{\hat{\Lambda}^{(m)},\hat{F}^{(m)}\right\}$ by solving

$$\left\{\hat{\Lambda}^{(m+1)},\hat{F}^{(m+1)}\right\} \in \underset{\Lambda\in\mathcal{B}\left(\hat{\Lambda}^{(m)},\sqrt{N}d_{NT}\right),F\in\mathcal{B}\left(\hat{F}^{(m)},\sqrt{T}d_{NT}\right)}{\operatorname{argmin}} \mathcal{L}_{NT}\left(\hat{\theta}^{(m)};\Lambda,F\right)$$

Step 3. Given $\left\{\hat{\Lambda}^{(m)},\hat{F}^{(m)}\right\}$, update the estimator of θ_0 to $\hat{\theta}^{(m+1)}$ according to

$$\hat{\theta}^{(m+1)} = \underset{\theta\in\mathcal{B}\left(\hat{\theta}^{(m)},d_{NT}\right)}{\operatorname{argmin}} \mathcal{L}_{NT}\left(\theta;\hat{\Lambda}^{(m+1)},\hat{F}^{(m+1)}\right)$$

Step 4. Iterate Steps 2-3 until a convergence criterion is met.

4 Theoretical Results

In this section, we examine the asymptotic properties of both the initial and iterative estimators. First, we study the consistency and convergence rate of the initial regularized estimator. We then establish the convergence and asymptotic distribution of the second-step iterative estimator. We consider an asymptotic framework where N and T tend to

infinity jointly. To keep the exposition simple, we focus on the single-index model with interactive fixed effects as the leading case. Similar results can be derived for other models with multiple indices.

Recall the definition of $\mathcal{L}_{NT}$ in (3.1). We further define

$$\bar{\mathcal{L}}(\theta, \Pi) = \mathbb{E}\left[\mathcal{L}_{NT}(\theta, \Pi)\right], \quad \tilde{\mathcal{L}}_{NT}(\theta, \Pi) = \mathcal{L}_{NT}(\theta, \Pi) - \mathbb{E}\left[\mathcal{L}_{NT}(\theta, \Pi)\right]$$

and the excess risk function as

$$\mathcal{E}(\theta, \Pi) = \bar{\mathcal{L}}(\theta, \Pi) - \bar{\mathcal{L}}(\theta_0, \Pi_0)$$

Let $\underline{c}$ and $\overline{c}$ denote generic positive constants that may vary in their occurrences.

Assumption 1 (Sampling and Covariates)**.** *(i) The data sequences $\{(W_{it})\}_{i\in[N],t\in[T]}$ obey the model (2.1);*
(ii) The data on units $(W_{it})_{t\in[T]}$ are independent across i, and for each i, the sequence $(W_{it})_{t\in[T]}$ is stationary and exponentially β-mixing with respect to t, conditional on Π. Specifically, there exist constants $C > 0$ and $\mu > 0$ such that $\forall q \geqslant 1$,

$$\sup_{1\leqslant i\leqslant N} \gamma_i(q; W) \leqslant C\exp(-\mu q)$$

where $\gamma_i(q; W) := \frac{1}{2}\sup_{\bar{t}\geq 1}\sup\sum_{l=1}^{L}\sum_{l'=1}^{L'}\left|\mathrm{P}\left(A_l \cap B_{l'}\right) - P\left(A_l\right)P\left(B_{l'}\right)\right|$ with the second supremum over all finite partitions $\{A_l\}$ in the σ-field generated by $(\{W_{it}\}_{t\leq\bar{t}})$ and $\{B_{l'}\}$ in the σ-field generated by $(\{W_{it}\}_{t\geq\bar{t}+q})$;
(iii) As $(N, T) \to \infty, N/T \to \kappa^2, 0 < \kappa < \infty$;
(iv) The covariate X_{it} has bounded support;
(v) The rank of Π_0 is fixed, denoted as r.

Assumption 2 (Lipschitz continuity)**.** *Suppose that $\ell(w; \theta, \pi)$ is Lipschitz continuous in (θ, π) in the sense that*

$$\left|\ell(w; \theta_1, \pi_1) - \ell(w; \theta_2, \pi_2)\right| \leqslant L(w)\left[\|\theta_1 - \theta_2\| + |\pi_1 - \pi_2|\right]$$

for a measurable function $L(\cdot)$ with $\max_{i\in[N],t\in[T]}\mathbb{E}\left[L\left(W_{it}\right)^4\right] < \infty$, where the data sequences $\{(W_{it})\}_{i\in[N],t\in[T]}$ satisfies Assumption 1.

Assumption 1 imposes basic regularity on the data generating process. Assumption 1(ii) limits inter-temporal dependence within the data. Specifically, μ controls the strength

of this temporal dependence. We adopt the assumption of exponentially beta-mixing for simplicity, but similar results can be achieved with polynomially beta-mixing data. If the observations $\{(W_{it})\}_{i\in[N],t\in[T]}$ are i.i.d. across i and over t, our theoretical results will hold without imposing Assumption 1(ii). Assumption 1(iii) imposes constraints on the relative rates at which N and T tend to infinity.[4] It is the standard assumption in the large-T panel data literature (Bai, 2009, Chen, Fernández-Val, and Weidner, 2021, Hahn and Kuersteiner, 2002). Assumption 1(iv) is included for convenience, but could be relaxed to allow covariates with sufficiently light tails. For instance, if the covariates are standard Gaussian, c_x can grow like $\sqrt{\log(pNT)}$ with probability approaching one. This avenue is not pursued here to maintain a clear and concise exposition. Additionally, we focus only on the case of fixed r; the analysis of approximately low-rank Π_0, analogous to approximate sparsity in the Lasso literature, is beyond the scope of this paper.

Assumption 2 imposes a Lipschitz continuity condition on the loss function with respect to the index. In the specific case where the loss function is globally Lipschitz with a constant independent of y or x, this assumption holds trivially.

4.1 Consistency

We show that the first-step regularized estimator is consistent, which is the starting point for establishing its rate of convergence. This section states the high-level conditions required for consistency, while Proposition 1 and Appendix B.3 present primitive sufficient conditions.

Assumption 3 (Identification)**.** *(i) θ_0 lies in the interior of Θ, and for each i and t, $\pi_{0,it}$ lies in the interior of Φ, where Θ and Φ are compact convex subsets of $\mathbb{R}^p$ and $\mathbb{R}$, respectively. Furthermore, Φ_λ and Φ_f are also compact convex subsets of $\mathbb{R}$;*
(ii) For all $\eta > 0$, there exists an $\varepsilon > 0$ such that for all $N \geq 1$ and $T \geq 1$,

$$\inf_{\substack{(\theta,\Pi)\in\Theta\times\Phi^{N\times T} \\ \|\theta-\theta_0\|^2+\frac{1}{NT}\|\Pi-\Pi_0\|_F^2\geq\eta^2}} \mathcal{E}(\theta,\Pi) \geq \varepsilon$$

This is an identification assumption that imposes restrictions on the shape of the excess risk function. When N and T are fixed, the condition is satisfied if the loss function is continuous and (θ_0, Π_0) is its unique minimizer. In the context of panel data, this condition needs to be satisfied uniformly in N and T.

[4]For the first-step estimator, it is sufficient that $T^{-1}\log N = o(1)$, which requires only that T increase sufficiently fast relative to $\log N$. This weaker condition guarantees the convergence rate of the first-step estimates and is implied by the joint asymptotic regime $N/T \to \kappa^2$ adopted here.

Thanks to the penalty term in (3.2), we can show that with probability approaching one, $\left(\hat{\theta}, \hat{\Pi}\right)$ belongs to a restricted set

$$\mathcal{B} = \left\{(\theta, \Pi) \in \Theta \times \Phi^{N \times T} : \|\Pi\|_* \leq c_{\mathcal{L}} \nu^{-1} + \|\Pi_0\|_*\right\}$$

where $c_{\mathcal{L}} = \bar{\mathcal{L}}(\theta_0, \Pi_0) + 1$. It is formally established in the Appendix. This result is important because it allows in the mathematical development to restrict the attention to a smaller set $\mathcal{B}$ included in the parameter space.

Assumption UC. $\sup_{(\theta,\Pi)\in\mathcal{B}} \left|\mathcal{L}_{NT}(\theta, \Pi) - \bar{\mathcal{L}}(\theta, \Pi)\right| = o_p(1)$

Assumption UC assumes uniform convergence. In the low-dimensional context, a similar condition is usually required on a compact set which does not depend on N and T. The main difference here is that the radius of $\mathcal{B}$ grows with N and T. We provide sufficient conditions for Assumption UC to hold in Proposition 1.

Lemma 1 (Consistency)**.** *Under Assumptions 3 and UC, if* $\nu \|\Pi_0\|_* = o(1)$*, then we have* $\left\|\hat{\theta} - \theta_0\right\|^2 + \frac{1}{NT}\left\|\hat{\Pi} - \Pi_0\right\|_F^2 = o_p(1)$.

The result depends on the condition that $\nu \|\Pi_0\|_* = o(1)$, which implies that the added penalty term has a negligible effect on the objective function when evaluated at the true values. This condition on ν is satisfied in the subsequent theorems and corollaries.

4.2 Rate of Convergence

Similar to other high-dimensional settings such as those in Negahban et al. (2012), we impose an invertibility condition involving another restricted set, denoted as $\mathcal{A}$, which we describe next and use to derive our main results. Before stating the next condition, we first introduce some notation.

Let $\Pi_0 = UDV'$ represent the singular value decomposition (SVD) of Π_0, where U and V are the matrices of singular vectors corresponding to all singular values. In particular, let $U_0 \in \mathbb{R}^{N \times r}$ and $V_0 \in \mathbb{R}^{T \times r}$ denote the columns of U and V associated with the non-zero singular values. For an $N \times T$ matrix Δ, we define the operators

$$\mathcal{P}(\Delta) \coloneqq U_0 U_0' \Delta V_0 V_0', \qquad \mathcal{M}(\Delta) \coloneqq \Delta - \mathcal{P}(\Delta) \tag{4.1}$$

Essentially, $\mathcal{P}(\cdot)$ can be thought of as a projection onto the subspace spanned by the columns of U_0 and V_0, constituting the " low-rank" space of Π_0. Similarly, $\mathcal{M}(\cdot)$ is the projection onto the space orthogonal to this low-rank space.

Next, we define the restricted set as

$$\mathcal{A} = \left\{ (\delta, \Delta) \in \mathbb{R}^p \times \mathbb{R}^{N \times T}, \text{s.t.} \|\Delta\|_* - 4\|\mathcal{P}(\Delta)\|_* - \sqrt{NT}\,\|\delta\| \leq 0 \right\} \tag{4.2}$$

It can be shown that $\left(\hat{\theta} - \theta_0, \hat{\Pi} - \Pi_0\right)$ lies in this "cone" under certain conditions. Namely, the set contains matrices Π that are close to Π_0, in the sense that the part that cannot be explained by λ_{0i} and f_{0t} is small in terms of nuclear norm.

Assumption 4 (Restricted Strong Convexity)**.** *For all $(\delta, \Delta) \in \mathcal{A}$, there exists a universal constant $c_{RSC} > 0$ such that*

$$\mathcal{E}(\theta_0 + \delta, \Pi_0 + \Delta) \geq c_{RSC}\,\|\delta\|^2 + \frac{c_{RSC}}{NT}\,\|\Delta\|_F^2$$

Assumption 4 is based on Restricted Strong Convexity (RSC), which relaxes the definition of strong convexity by only needing strong convexity in certain directions or over a subset of the ambient space (Negahban, Ravikumar, Wainwright, and Yu, 2012, Wainwright, 2019). This assumption represents a version of a widely used condition in the matrix completion literature, although verifying it typically requires imposing additional structure on the parameter space. See also the discussions in Moon and Weidner (2019) and Miao, Phillips, and Su (2022), following their Assumptions 1 and 2, respectively, for the linear case. A detailed discussion for the single-index model is provided in Appendix B.3.

Finally, to control the empirical risk associated with the estimation errors, we define the set

$$\mathcal{V} = \left\{ (\theta, \Pi) \in \mathcal{B} : \|\delta\|^2 + \frac{1}{NT}\,\|\Delta\|_F^2 \leq c_l \right\}.$$

We also introduce the norm $\rho(\cdot, \cdot)$, defined as

$$\rho(\delta, \Delta) = \left[\|\delta\|^2 + \frac{1}{NT}\,\|\Delta\|_*^2 \right]^{1/2}.$$

According to Lemma 1, it suffices to analyze the convergence rate of the empirical risk within the restricted set $\mathcal{V}$.

Proposition 1. *Under Assumptions 1 and 2, and with $\nu = c_\nu \frac{\psi_{NT} c_{\varepsilon,NT}}{\sqrt{NT}}$ for some constant $c_\nu \geq 2$, there exist positive sequences $\{\psi_{NT}\}$, and $\{c_{\varepsilon,NT}\}$ such that*

$$\lim_{(N,T)\to\infty} \mathrm{P}\left(\sup_{(\theta,\Pi)\in\mathcal{V}} \frac{\left|\tilde{\mathcal{L}}_{NT}(\theta, \Pi) - \tilde{\mathcal{L}}_{NT}(\theta_0, \Pi_0)\right|}{\rho(\theta - \theta_0, \Pi - \Pi_0) \vee c_{\varepsilon,NT}} \leq \psi_{NT} c_{\varepsilon,NT} \right) = 1$$

where $c_T = \lceil 2\mu^{-1} \log(NT) \rceil$, $d_T = \lfloor T/(2c_T) \rfloor$, $c_{\varepsilon,NT} = \frac{\sqrt{c_T}}{\sqrt{N \wedge d_T}}$ *and* $\psi_{NT}/(C_L \log(c_T + 1)) \to \infty$. *Moreover, Assumption UC holds.*

The proof of Proposition 1 is based on the empirical process theory and does not rely on the differentiability of the loss function. It derives sufficient conditions under which Assumption UC holds and provides a stochastic bound that will be used for the subsequent theorem. Specifically, we have $c_{\varepsilon,NT} = O\left(\sqrt{\log(NT)/(N \wedge T)}\right)$, since, under our exponential β-mixing setting, the dependence length satisfies $c_T = O(\log(NT))$. The scaling sequence ψ_{NT} is allowed to diverge slowly, at a rate exceeding $O(\log(\log(NT)))$, so that the probabilistic bound in Proposition 1 holds uniformly over the local parameter space $\mathcal{V}$.

We now present our first main result for estimating (θ_0, Π_0).

Theorem 1 (Convergence Rate)**.** *Under Assumptions 1-4, let* $\nu = c_\nu \frac{\psi_{NT} c_{\varepsilon,NT}}{\sqrt{NT}}$ *for some constant* $c_\nu \geq 2$. *With probability approaching one, we have*

$$\left\|\hat{\theta} - \theta_0\right\|^2 + \frac{1}{NT}\left\|\hat{\Pi} - \Pi_0\right\|_F^2 \lesssim \frac{\psi_{NT}^2 r c_T}{N \wedge d_T}$$

where $c_{\varepsilon,NT}$ *and* ψ_{NT} *are specified in Proposition 1.*

Theorem 1 establishes the convergence rates of the estimation errors for $\hat{\theta}$ and $\hat{\Pi}$ in the ℓ_2 norm. Here, θ is a low-dimensional parameter vector, while Π is a high-dimensional matrix with NT elements; hence, the Frobenius norm of Π is normalized by $1/\sqrt{NT}$ to ensure comparability. For simplicity, consider the case of i.i.d. panel data across i over t. When the regularization parameter is set to $\nu \asymp \frac{\sqrt{N} \vee \sqrt{T}}{NT}$, the convergence rate of our estimator under the Euclidean norm is of the order of $\sqrt{1/(N \wedge T)}$, with all other factors ignored. These results are consistent with previous work on penalized mean and quantile regression models for panel data (Athey, Bayati, Doudchenko, Imbens, and Khosravi, 2021, Belloni, Chen, Madrid Padilla, and Wang, 2023, Feng, 2023, Moon and Weidner, 2019), while the present paper develops a unified M-estimation framework that extends these results to a broader class of models. Finally, note that Theorem 1 does not require differentiability of the loss function; however, subsequent asymptotic normality results rely on smoothness conditions to establish limiting distributions.

Next, we impose an assumption on the common factors and factor loadings.

Assumption 5 (Strong factors)**.** *Let* $\frac{1}{N}\sum_{i=1}^N \lambda_{0i}\lambda_{0i}' \xrightarrow{p} \Sigma_\Lambda > 0$ *and* $\frac{1}{T}\sum_{t=1}^T f_{0t} f_{0t}' \xrightarrow{p} \Sigma_F >$ *for some positive definite matrices* $\Sigma_\Lambda, \Sigma_F \in \mathbb{R}^{r \times r}$. *Furthermore, the matrix* Π_0 *has* r *distinct nonzero singular values.*

Assumption 5 formalizes the strong factor condition commonly imposed in the literature; see, for example, Bai (2009) and Chen, Fernández-Val, and Weidner (2021). The requirement that the singular values $\sigma_1, \ldots, \sigma_r$ are distinct is analogous to Assumption G in Bai (2003), and closely related to Assumption 4.2 in Chernozhukov, Hansen, Liao, and Zhu (2023). In particular, Theorem 1 does not depend on Assumption 5.[5]

The following corollary establishes the consistency of $\hat{r}$ and the mean squared convergence rates of $\hat{\Lambda}$ and $\hat{F}$.

Corollary 1. *Suppose that the conditions in Theorem 1 and Assumption 5 hold. Without loss of generality, we impose the normalization that* $F'F/T = \mathbb{I}_r$ *and* $\Lambda'\Lambda/N$ *is diagonal with nonincreasing diagonal elements. Then,*
(i) $P(\hat{r} = r) \to 1$ *as* $N, T \to \infty$*;*
(ii) $\frac{1}{\sqrt{N}} \left\| \hat{\Lambda} - \Lambda_0 \hat{S} \right\|_F = O_p(\gamma_{NT})$*, and* $\frac{1}{\sqrt{T}} \left\| \hat{F} - F_0 \hat{S} \right\|_F = O_p(\gamma_{NT})$*, where* $\hat{S} = \operatorname{sgn}\left(\hat{F}' F_0\right)$ *and* $\gamma_{NT} = \psi_{NT} c_{\varepsilon,NT}$*, with* ψ_{NT} *and* $c_{\varepsilon,NT}$ *specified in Proposition 1.*

Nuclear-norm regularization does not require prior knowledge or specification of the number of factors r. Corollary 1 indicates that $\hat{r}$ is a consistent estimator of r. Thus, in what follows, we assume that the number of factors has been correctly selected. Without loss of generality, we further assume that $\hat{\mathrm{S}} = \mathbb{I}_r$ to simplify the notation.

4.3 Asymptotic Normality

Here, we use the shorthand notation $\ell_{it}(\theta, \pi) = \ell(W_{it}; \theta, \pi)$ for convenience, where $\theta \in \Theta$ and $\pi \in \Phi$. Additionally, we omit the function arguments when they are evaluated at the true parameter values $(\theta_0, \pi_{0,it})$, e.g., $\ell_{it} = \ell(W_{it}; \theta_0, \pi_{0,it})$.

To study the asymptotic behavior of the iterative estimator $\hat{\theta}^{(m+1)}$, it is necessary to introduce additional quantities that characterize how the parameter θ interacts with the incidental components π_{it}.

Let Ξ_{it} denote a p-dimensional vector defined by the following population weighted least squares projection for each component of $\mathbb{E}[\partial_{\theta_k \pi} \ell_{it}]$ onto the space spanned by the incidental parameters, under a metric given by $\mathbb{E}[\partial_{\pi^2} \ell_{it}]$. Specifically, $\Xi_{it,k} = \lambda_{i,k}^{*\prime} f_{0t} + \lambda_{0i}' f_{t,k}^*$, where $\left(\lambda_{i,k}^*, f_{t,k}^*\right)$ is defined as the solution to the following optimization problem,

$$\left(\lambda_{i,k}^*, f_{t,k}^*\right) \in \underset{\lambda_{i,k}, f_{t,k}}{\operatorname{argmin}} \sum_{i,t} \mathbb{E}[\partial_{\pi^2} \ell_{it}] \left(\frac{\mathbb{E}[\partial_{\theta_k \pi} \ell_{it}]}{\mathbb{E}[\partial_{\pi^2} \ell_{it}]} - \lambda_{i,k}' f_{0t} - \lambda_{0i}' f_{t,k} \right)^2$$

[5] Armstrong, Weidner, and Zeleneev (2023) use NNR and study robust inference for weak factors in the linear panel model.

In addition, we define the operator $D_{\theta\pi^q}\ell_{it} = \partial_{\theta\pi^q}\ell_{it} - \Xi_{it}\partial_{\pi^{q+1}}\ell_{it}$ for $q = 0, 1, 2$. Intuitively, these operators remove the component of $\partial_{\theta\pi^q}\ell_{it}$ that can be explained by the individual and time fixed effects. Furthermore, define a $p \times p$ matrix

$$\bar{W}_{NT} := \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{E}\left[\partial_{\theta\theta}\ell_{it} - \partial_{\pi^2}\ell_{it}\Xi_{it}\Xi_{it}'\right]$$

which will be used to characterize the information matrix for θ.

We now introduce the regularity conditions that are required for the asymptotic results.

Assumption 6. *Let $\mathcal{B}_\epsilon^0 \subset \mathbb{R}^{p+1}$ denote a bounded set that contains a ϵ-neighborhood of $(\theta_0, \pi_{0,it})$ for all $i \in [N]$ and $t \in [T]$, uniformly over N and T.*
(i) The function $(\theta, \pi) \mapsto \ell_{it}(\theta, \pi)$ is four times continuously differentiable over $\mathcal{B}_\epsilon^0$. The partial derivatives of $\ell_{it}(\theta, \pi)$ with respect to the elements of (θ, π) up to the fourth order are bounded in absolute value over $(\theta, \pi) \in \mathcal{B}_\epsilon^0$ by a function $M(W_{it}) > 0$ such that $\max_{i\in[N],t\in[T]}\mathbb{E}\left[|M(W_{it})|^{8+\iota}\right]$ is uniformly bounded for some $\iota > 0$ over all N and T;
(ii) There exist positive constants $b_{\min}$ and $b_{\max}$ such that for all $(\theta, \pi) \in \mathcal{B}_\epsilon^0$, $b_{\min} \leq \mathbb{E}\left[\partial_{\pi^2}\ell_{it}(\theta, \pi)\right] \leq b_{\max}$;
(iii) The matrix $\bar{W}_{NT}$ is uniformly positive definite, i.e., $\inf_{N,T}\sigma_{\min}\left(\bar{W}_{NT}\right) \geq c > 0$.

Assumption 6 strengthens the conditions used for Theorem 1.

Assumption 6(i) and (ii) require the loss function to exhibit sufficient smoothness and local strong convexity. Assumption 6(iii) is a generalized noncollinearity condition, similar to those commonly imposed in the factor literature, such as Assumption A in Bai (2009) and Assumption 1(vi) in Chen, Fernández-Val, and Weidner (2021)). In the linear case, Assumption 6(iii) simplifies to requiring $\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{E}\left[(X_{it} - \Xi_{it})(X_{it} - \Xi_{it})'\right]$ to be positive definite. Intuitively, this condition ensures that the covariates exhibit sufficient variation across individuals and over time, thus guaranteeing the identification of θ.

Theorem 2. *Suppose Assumptions 1-6 hold, then*
(i) For $m = 0, 1, 2, ...$, we have

$$\begin{aligned}\hat{\theta}^{(m+1)} - \theta_0 = C^{(0)}\left(\hat{\theta}^{(m)} - \theta_0\right) + C^{(1)} \\ + o_p\left(\left\|\hat{\theta}^{(m)} - \theta_0\right\|\right) + o_p\left((NT)^{-1/2}\right)\end{aligned}$$

where $C^{(0)} \xrightarrow{p} \bar{C}^{(0)}$ with $\|\bar{C}^{(0)}\| \in [0, 1)$, and $C^{(1)} = O_p\left((NT)^{-1/2}\right)$;

(ii) For any $m \geq -\left(\frac{1}{2}\log(NT) + \log(\gamma_{NT})\right) / \log\left(\|\bar{C}^{(0)}\|\right) - 1$ *with* $\bar{C}^{(0)} \neq 0$,

$$\hat{\theta}^{(m+1)} - \theta_0 = O_p\left((NT)^{-1/2}\right)$$

where $\gamma_{NT} = \psi_{NT} c_{\varepsilon,NT}$, *with* ψ_{NT} *and* $c_{\varepsilon,NT}$ *specified in Proposition 1.*

Theorem 2 examines the numerical convergence properties of $\hat{\theta}^{(m+1)}$. Specifically, Theorem 2(i) guarantees the convergence of the iterative procedure, while Theorem 2(ii) indicates that after $O(\log(NT))$ iterations, the iterative estimator $\hat{\theta}^{(m+1)}$ achieves the desired convergence rate for inference. It converges to the true value θ_0 much faster than γ_{NT}, which is the rate at which the initial estimator $\hat{\theta}$ converges in probability. The upper bound of our contraction parameter, specifically the spectral norm of $\bar{C}^{(0)}$, is crucial in the iterative procedure. The closer $\|\bar{C}^{(0)}\|$ is to 0, the faster is the contraction and the numerical convergence of the iterative procedure. Conversely, if $\|\bar{C}^{(0)}\|$ is close to 1, the numerical convergence of $\hat{\theta}^{(m+1)}$ is slow.

The following theorem establishes the asymptotic distribution of the second-step estimator $\hat{\theta}^{(m+1)}$.

Theorem 3. *Suppose Assumptions 1-6 hold, and the following limits exist,*

$$\bar{W}_\infty := \lim_{N,T\to\infty} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbb{E}\left[\partial_{\theta\theta}\ell_{it} - \partial_{\pi^2}\ell_{it}\Xi_{it}\Xi_{it}'\right]$$

$$\bar{B}_\infty := -\lim_{N,T\to\infty} \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \left\{ \left[\sum_{\tau=t}^{T} f_{0t}' H_{(\lambda\lambda)i}^{-1} f_{0\tau} \mathbb{E}\left[\partial_{\pi}\ell_{it} D_{\theta\pi}\ell_{i\tau}\right]\right] + \frac{1}{2} f_{0t}' H_{(\lambda\lambda)i}^{-1} f_{0t} \mathbb{E}\left[D_{\theta\pi^2}\ell_{it}\right] \right\}$$

$$\bar{D}_\infty := -\lim_{N,T\to\infty} \frac{1}{T} \sum_{i=1}^{N} \sum_{t=1}^{T} \lambda_{0i}' H_{(ff)t}^{-1} \lambda_{0i} \mathbb{E}\left[\partial_{\pi}\ell_{it} D_{\theta\pi}\ell_{it} + \frac{1}{2} D_{\theta\pi^2}\ell_{it}\right]$$

where $H_{(\lambda\lambda)i} = \sum_{t=1}^{T} \mathbb{E}\left[\partial_{\pi^2}\ell_{it}\right] f_{0t} f_{0t}'$, $H_{(ff)t} = \sum_{i=1}^{N} \mathbb{E}\left[\partial_{\pi^2}\ell_{it}\right] \lambda_{0i}\lambda_{0i}'$, *and* $\bar{W}_\infty > 0$. *Then, for any* $m \geq -\frac{1}{2}\log(NT)/\log\left(\|\bar{C}^{(0)}\|\right) - 1$, *it follows that*

$$\sqrt{NT}\left(\hat{\theta}^{(m+1)} - \theta_0 - \frac{1}{T}\bar{W}_\infty^{-1}\bar{B}_\infty - \frac{1}{N}\bar{W}_\infty^{-1}\bar{D}_\infty\right) \xrightarrow{d} \mathcal{N}\left(0, \bar{W}_\infty^{-1}\right) \tag{4.3}$$

Theorem 3 indicates that $\hat{\theta}^{(m+1)}$ contains two asymptotic bias terms associated with $\frac{1}{T}\bar{B}_\infty$ and $\frac{1}{N}\bar{D}_\infty$, respectively. For convex objective functions, the estimator is asymptotically equivalent to the one obtained from (3.3), as studied by Chen, Fernández-Val, and Weidner (2021).

4.4 Bias Correction

This section describes methods for removing the asymptotic bias of the second-step estimator. We first outline the analytical bias-correction procedure and then discuss alternative approaches, including the split-panel jackknife and the bootstrap.

The analytical correction is constructed using sample analogs of the expressions in Theorem 3, replacing the true values of (θ, π) with their second-step estimates. Both analytical bias correction and variance estimation require consistent estimators of the quantities $\bar{B}\infty$, $\bar{D}\infty$, and $\bar{W}_\infty$ defined in Theorem 3. Let $\hat{B}$, $\hat{D}$, and $\hat{W}$ denote the corresponding sample analogs, obtained by substituting sample averages for expectations and replacing the true parameters with their second-step estimates. For example,

$$\hat{W} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \partial_{\theta\theta}\hat{\ell}_{it} - \partial_{\pi^2}\hat{\ell}_{it}\hat{\Xi}_{it}\hat{\Xi}'_{it}$$

where $\partial_{\theta\theta}\hat{\ell}_{it} = \partial_{\theta\theta}\ell_{it}\left(W_{it}; \hat{\theta}, \hat{\lambda}_i{}'\hat{f}_t\right)$, $\partial_{\pi^2}\hat{\ell}_{it} = \partial_{\pi^2}\ell_{it}\left(W_{it}; \hat{\theta}, \hat{\lambda}_i{}'\hat{f}_t\right)$, and $\hat{\Xi}_{it}$ is a p-dimensional vector with elements $\hat{\Xi}_{it,k} = \lambda^{\#\prime}_{i,k}\hat{f}_t + \hat{\lambda}'_i f^{\#}_{t,k}$ where the pair $\left(\lambda^{\#}_{i,k}, f^{\#}_{t,k}\right)$ is defined by

$$\left(\lambda^{\#}_{i,k}, f^{\#}_{t,k}\right) \in \underset{\lambda_{i,k}, f_{t,k}}{\operatorname{argmin}} \sum_{i,t}\left(\partial_{\pi^2}\hat{\ell}_{it}\right)\left(\frac{\partial_{\theta_k\pi}\hat{\ell}_{it}}{\partial_{\pi^2}\hat{\ell}_{it}} - \lambda'_{i,k}\hat{f}_t - \hat{\lambda}'_i f_{t,k}\right)^2$$

Once those sample analogs are constructed, then the analytic bias correction of $\hat{\theta}$ reads

$$\hat{\theta}_{ABC} = \hat{\theta}^{(m+1)} - \frac{1}{T}\hat{W}^{-1}\hat{B} - \frac{1}{N}\hat{W}^{-1}\hat{D}$$

Theorem 4. *Suppose that the assumptions in Theorem 3 hold. Then* $\hat{W} \xrightarrow{p} \bar{W}_\infty$ *and*

$$\sqrt{NT}\left(\hat{\theta}_{ABC} - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(0, \bar{W}_\infty^{-1}\right)$$

Although the analytical correction offers a convenient closed-form adjustment, it requires estimating high-order derivatives and moment matrices that may be sensitive to model specification and sample size. In practice, alternative approaches—such as the split-panel jackknife and the bootstrap—can remove the leading bias terms without explicit analytical derivations.

Following Dhaene and Jochmans (2015) and Chen, Fernández-Val, and Weidner (2021), bias correction can also be implemented using the split-panel jackknife method. Let

$\hat{\theta}^1_{N/2,T}$ and $\hat{\theta}^2_{N/2,T}$ be the second-step estimators based on the subsamples $\{(i,t) : i = 1,\ldots,N/2; t = 1,\ldots,T\}$ and $\{(i,t) : i = N/2+1,\ldots,N; t = 1,\ldots,T\}$, respectively. Similarly, let $\hat{\theta}^1_{N,T/2}$ and $\hat{\theta}^2_{N,T/2}$ be the estimators obtained from the subsamples $\{(i,t) : i = 1,\ldots,N; t = 1,\ldots,T/2\}$ and $\{(i,t) : i = 1,\ldots,N; t = T/2+1,\ldots,T\}$, respectively. The jackknife bias-corrected estimator is

$$\hat{\theta}_{\text{JBC}} = 3\hat{\theta} - \frac{1}{2}\left(\hat{\theta}^1_{N/2,T} + \hat{\theta}^2_{N/2,T}\right) - \frac{1}{2}\left(\hat{\theta}^1_{N,T/2} + \hat{\theta}^2_{N,T/2}\right).$$

Under suitable homogeneity and stationarity conditions ensuring that the asymptotic biases of all estimators converge to the same limit,

$$\sqrt{NT}\left(\hat{\theta}_{\text{JBC}} - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(0, \bar{W}_\infty^{-1}\right)$$

Other bias-correction methods include the leave-one-out jackknife of Hughes (2022) and the bootstrap procedures of Higgins and Jochmans (2024) and Sun and Kim (2010). However, these approaches are not directly applicable to models with interactive fixed effects. Extending them, particularly adapting the bootstrap of Higgins and Jochmans (2024) to account for time-specific or interactive effects, remains an interesting avenue for future research.

5 Monte Carlo Simulations

In this section, we assess the finite sample performance of our proposed approach using Monte Carlo simulations. Section 5.1 outlines the main steps of the first-step estimation algorithm. Section 5.2 discusses the choice of the tuning parameter ν. Section 5.3 presents the simulation designs and results.

5.1 Implementation

This section outlines the computational procedures for the proposed estimation algorithms. We first present the single-index panel model and its implementation; the binary logit specification in Example 3 is treated as a special case. We then describe the random coefficient logit model in Example 4. Additional details and formal descriptions can be found in Section A.

Single-Index Panel Model Algorithm 1 summarizes the first-stage estimation procedure for the class of single-index panel models, which includes the binary logit specification

in Example 3 as a special case. We introduce a slack variable Z_Π to separate the low-rank interactive effects from the linear index component. This reformulation allows the optimization problem in (3.2) to be expressed in an equivalent and computationally convenient form:

$$\begin{aligned} \min_{\theta\in\mathbb{R}^p,\Pi\in\mathbb{R}^{N\times T}} & \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\ell\left(Y_{it},V_{it}\right)+\nu\|\Pi\|_* \\ \text{s.t. } & V=\sum_{j=1}^{p}X_j\theta_j+Z_\Pi, \quad Z_\Pi-\Pi=0 \end{aligned} \tag{5.1}$$

To solve the minimization problem (5.1), we use an Alternating Direction Method of Multipliers (ADMM) algorithm, which relies on the following augmented Lagrangian

$$\begin{aligned} \mathscr{L}\left(\theta,\Pi,V,Z_\Pi,U_p,U_v\right) = & \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\ell\left(Y_{it},V_{it}\right)+\nu\|\Pi\|_* \\ & +\frac{\eta}{2NT}\left\|V-\sum_{j=1}^{p}X_j\theta_j-Z_\Pi+U_p\right\|_F^2+\frac{\eta}{2NT}\left\|Z_\Pi-\Pi+U_v\right\|_F^2 \end{aligned}$$

Here, U_p and U_v are the scaled dual variables corresponding to the linear constraints $V=\sum_{j=1}^{p}X_j\theta_j+Z_\Pi$ and $Z_\Pi-\Pi=0$, and $\eta>0$ is the penalty parameter for constraint violations. Due to the separability of the parameters in $\mathscr{L}$, the ADMM algorithm proceeds by iteratively minimizing the augmented Lagrangian in blocks with respect to the original variables, in this case (V,θ,Π) and Z_Π, and then then updating the dual variables U_p and U_v.

As shown in Line 9 of Algorithm 1, Π is updated via singular value thresholding (SVT). Specifically, recall that the singular value decomposition (SVD) of a matrix A is

$$A=U_A\Sigma_AV_A'\in\mathbb{R}^{N\times T}, \quad \Sigma_A=\operatorname{diag}\left(\{\sigma_s\}_{s\in[N\wedge T]}\right)$$

For $\iota>0$, the soft-thresholding operator $S_\iota(\cdot)$ is defined as

$$S_\iota(A)\coloneqq U_AS_\iota\left(\Sigma_A\right)V_A', \quad S_\iota\left(\Sigma_A\right)=\operatorname{diag}\left(\{\sigma_s-\iota\}_+\right)$$

where t_+ is the positive part of t, namely, $t_+=\max(0,t)$. See Theorem 2.1 in Cai, Candes, and Shen (2008) for details.

Algorithm 1 ALM Algorithm: Single-Index Panel Model

Input: Observed data (Y_{it}, X_{it}), and initialization for $V^{(0)}$, $\theta^{(0)}$, $\Pi^{(0)}$, $Z_{\Pi}^{(0)}$, $U_p^{(0)}$, and $U_v^{(0)}$, parameters ν, η.
Initialize: $k = 0$
while not converged **do**
    Update parameters at iteration $k+1$**:**
    $V_{it}^{(k+1)} \leftarrow V_{it}^{(k)} + h(Y_{it}, V_{it}^{(k)}) - \eta(V_{it}^{(k)} - X_{it}'\theta^{(k)} - Z_{\Pi,it}^{(k)} + U_{p,it}^{(k)})$
    where $h(y, v) = \nabla_v \ell(y, v)$
    $\theta^{(k+1)} \leftarrow \left(\sum_{i=1}^N \sum_{t=1}^T X_{it}X_{it}'\right)^{-1} \left(\sum_{i=1}^N \sum_{t=1}^T X_{it}A_{it}^{(k)}\right)$
    where $A^{(k)} = V^{(k+1)} - Z^{(k)} + U_p^{(k)}$
    $\Pi^{(k+1)} \leftarrow S_{NT\frac{\nu}{\eta}}(Z^{(k)} + U_v^{(k)})$
    $Z_{\Pi}^{(k+1)} \leftarrow \frac{1}{2}(V^{(k+1)} - \sum_{j=1}^p X_j\theta_j^{(k+1)} + U_p^{(k)} + \Pi^{(k+1)} - U_v^{(k)})$
    $U_p^{(k+1)} \leftarrow V^{(k+1)} - \sum_{j=1}^p X_j\theta_j^{(k+1)} - Z_{\Pi}^{(k+1)} + U_p^{(k)}$
    $U_v^{(k+1)} \leftarrow Z_{\Pi}^{(k+1)} - \Pi^{(k+1)} + U_v^{(k)}$
    $k \leftarrow k+1$
end while
Output: Updated parameters $\theta^{(k+1)}$, $\Pi^{(k+1)}$.

Random Coefficient Logit Panel Algorithm 2 outlines the first-stage estimation algorithm for the binary random coefficient logit panel model in Example 4 using a majorization-maximization (MM) approach. For simplicity, we assume that all coefficients β are random, but incorporating fixed coefficients, such as $\beta_{j,it} = \beta_j$ for some j, is straightforward. Our algorithm extends the work of Train (2009) and James (2017) by introducing a novel surrogate function to handle the penalty term. Specifically, to update Π at step $k+1$, we use the surrogate function $Q\left(\Pi|\theta^{(k)}, \Pi^{(k)}\right)$, defined as

$$
\begin{aligned}
Q\left(\Pi \mid \theta^{(k)}, \Pi^{(k)}\right) =& \mathcal{L}_{NT}\left(\theta^{(k)}, \Pi^{(k)}\right) - \frac{1}{2NT}\sum_{i=1}^N\sum_{t=1}^T\left(\sum_{r=1}^R w_{itr}^{(k)} h\left(Y_{it}, X_{it}'\beta_{itr}^{(k)} + \pi_{it}^{(k)}\right)\right)^2 \\
&+ \frac{1}{2NT}\sum_{i=1}^N\sum_{t=1}^T\left(\pi_{it}^{(k)} - \sum_{r=1}^R w_{itr}^{(k)} h\left(Y_{it}, X_{it}'\beta_{itr}^{(k)} + \pi_{it}^{(k)}\right) - \pi_{it}\right)^2 + \nu\|\Pi\|_*
\end{aligned}
$$

where the closed-form solution is given by

$$
\Pi^{(k+1)} \leftarrow S_{NT\nu}\left(\Pi^{(k)} - \left(\sum_{r=1}^R w_{itr}^{(k)} h\left(Y_{it}, X_{it}'\beta_{itr} + \pi_{it}^{(k)}\right)\right)_{i,t}\right)
$$

with $h(y, v) = \Lambda(v) - y$ and $\Lambda(v) = \frac{1}{1+e^{-v}}$. This is detailed in Line 11 of Algorithm 2.

Algorithm 2 MM Algorithm: Binary Random Coefficient Logit Panel Model

Input: Observed data (Y_{it}, X_{it}), and initialization for $\beta^{(0)}$, $\Sigma^{(0)}$, $\Pi^{(0)}$, parameters ν.
Initialize: $k = 0$
while not converged **do**
    Step 1: Simulate and Calculate Weights
    Given $\left(\theta^{(k)}, \Pi^{(k)}\right)$, for each (i,t), draw R values of β from $N\left(\bar{\beta}^{(k)}, \Sigma^{(k)}\right)$, labeled β_{itr}
    For each β_{itr}, calculate $p_{it}(\beta_{itr}, \pi_{it}^{(k)})$ and the likelihood $L_{it}(\beta_{itr}, \pi_{it}^{(k)})$ using Equations (A.2) and (A.3).
    Using $L_{it}(\beta_{itr}\pi_{it}^{(k)})$, compute the weights $w_{itr}^{(k)}$ using Equation (A.4).
    Step 2: Update Parameters
    $\bar{\beta}^{(k+1)} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\sum_{r=1}^{R} w_{itr}^{(k)}\beta_{itr}\right)$
    $\Sigma^{(k+1)} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\sum_{r=1}^{R} w_{itr}^{(k)}\beta_{itr}\beta_{itr}'\right) - \bar{\beta}^{(k+1)}\bar{\beta}^{(k+1)'}$
    $\Pi^{(k+1)} \leftarrow S_{NT\nu}\left(\Pi^{(k)} - \left(\sum_{r=1}^{R} w_{itr}^{(k)} h\left(Y_{it}, X_{it}'\beta_{itr} + \pi_{it}^{(k)}\right)\right)_{i,t}\right)$
    $k \leftarrow k + 1$
end while
Output: Updated parameters $\beta^{(k+1)}$, $\Sigma^{(k+1)}$, and $\Pi^{(k+1)}$.

5.2 Selection of Tuning Parameters

The selection of the tuning parameter ν is crucial for the proper implementation of the estimator. The objective is to choose ν such that it dominates the "score" of our penalized estimator. This ensures a desirable rate of convergence and permits the consistent estimation of the rank of Π_0, even in scenarios where it is not known a priori. Nonetheless, the computation of the "score" is generally infeasible, as it depends on unknown true parameters (θ_0, Π_0). This challenge becomes more complex in the context of time series data, where the selection of ν must also account for the degree of temporal dependence.

Several general approaches are commonly used to choose the optimal tuning parameters. As is standard in the statistics and machine learning literature, the most popular data-driven method is cross-validation. Another common approach involves information criteria (IC). Methods like grid search or randomized search are also employed, where a range or a sample of tuning parameters is systematically explored to identify the best value.

In this paper, we select the tuning parameter ν using a modified information criterion (IC), motivated by prior work on low-rank and grouped panel estimators (e.g., Bai and Ng (2002); Belloni, Chen, Madrid Padilla, and Wang (2023); Su, Shi, and Phillips (2016)) and provides a practical way to balance model complexity and statistical fit.

Specifically, for each candidate value of ν, we compute the corresponding rank estimate

$\hat{r}(\nu)$, as defined in Equation (3.4),[6] and evaluate

$$\mathrm{IC}(\nu) = \mathcal{L}_{NT}\left(\hat{\theta}(\nu), \hat{\Pi}(\nu)\right) + \varrho_{NT} \cdot \hat{r}(\nu) \tag{5.2}$$

The first term on the right-hand side of (5.2) measures the overall fit to the data, and the second term introduces a penalty proportional to model complexity through the factor ϱ_{NT}. We experimented with several alternatives and found that $\varrho_{NT} = \frac{1}{2}\log(N \wedge T)\frac{N \vee T}{NT}$ works fairly well in Design 1 and so does $\varrho_{NT} = \frac{1}{2}\frac{\log\log(\sqrt{NT})}{\sqrt{NT}}$ in Design 2. Accordingly, we adopt these specifications for ϱ_{NT} in the analysis that follows.

5.3 Finite-Sample Performance

Data Generating Process To evaluate the finite-sample performance of the estimation procedure, we consider two data generating processes (DGPs) that cover models with convex and non-convex objective functions.

Design 1: The data are generated from the following model

$$y_{it} = \mathbb{I}\left\{X_{it}'\theta_0 + \sum_{s=1}^{2}\lambda_{is}f_{ts} - \varepsilon_{it} \geq 0\right\}$$

where $\mathbb{I}\{\cdot\}$ denotes the indicator function, and the error term ε_{it} follows a logistic distribution. We set $p = 3$ and $r = 2$ and take the coefficients θ_0 to satisfy $\theta_0 = (1,1,1)'$. The regressor vector $X_{it} = (x_{it,1}, x_{it,2}, x_{it,3})'$ consists of three variables, $\lambda_i = (\lambda_{i1}, \lambda_{i2})'$, and $f_t = (f_{t1}, f_{t2})'$. Specifically, we define $x_{it,k} = \tilde{x}_{it,k} + \rho(\lambda_{ik}^2 + f_{tk}^2)$ for $k = 1, 2$ and $x_{it,3} = \tilde{x}_{it,3}$. The parameter $\rho = 0.2$ governs the correlation between the covariates and the fixed effects. Here, λ_i, f_t, and $\tilde{x}_{it}$ are mutually independent across both i and t and $\lambda_i \sim \mathcal{N}\left(\binom{1}{1}, I_2\right)$ and $f_t \sim \mathcal{N}\left(\binom{0}{0}, I_2\right)$.

We consider two alternative specifications for the distribution of $\tilde{x}_{it}$:

- DGP 1: The covariates are independently distributed as $\tilde{x}_{it} \sim \mathcal{N}\left(\begin{pmatrix}0\\0\\0\end{pmatrix}, 4I_3\right)$ where I_k denotes k by k identity matrix.
- DGP 2: The components of $\tilde{x}_{it}$ follow a stationary AR(1) process. For $k = 1, 2, 3$, $\tilde{x}_{ik,t} = \varrho\tilde{x}_{ik,t-1} + \sigma u_{ik,t}$, where $u_{ik,t} \sim \mathcal{N}(0,1)$ with $\varrho = 0.2$, $\sigma = 2$, and initial values drawn from $\mathcal{N} \sim \left(0, \frac{\sigma^2}{1-\varrho^2}\right)$.

[6] The consistency of the rank estimator is established in Corollary 1.

Design 2: We extend Design 1 by allowing random coefficients β_{it}. With a slight abuse of notation,

$$y_{it} = \mathbb{I}\left\{ X_{it}'\beta_{0,it} + \sum_{s=1}^{2} \lambda_{is} f_{ts} - \varepsilon_{it} \geq 0 \right\}$$

The covariates X_{it}, factor loadings λ_{is}, and factors f_{ts} follow the same distributions as in Design 1, and we adopt the same specifications of $\tilde{x}_{it}$. The coefficients $\beta_{0,it} \sim \mathcal{N}\left(\bar{\beta}_0, \Sigma_0\right)$ with $\bar{\beta}_0 = (1,1,1)'$ and $\Sigma_0 = 0.3I_3$.

Performance For each model design, we consider different combinations of sample sizes $(N,T) \in \{(100,100), (150,100), (200,100), (150,150), (200,150), (200,200)\}$ to examine how the first-step and second-step estimators behave as the sample size increases. Each experiment is replicated 100 times. The ADMM algorithm is used to estimate the model in Design 1 and Design 2, while the MM algorithm is applied to estimate the model in Design 3. Detailed descriptions of the implementation are provided in Section 5.1.

To evaluate the performance of my estimator, we report the root mean squared error (RMSE) for both steps, defined as $\text{RMSE}(\hat{\theta}) = \sqrt{S^{-1}\sum_{s=1}^{S} |\hat{\theta}^{(s)} - \theta_0|^2}/\|\theta_0\|$, and the average estimated rank, $\mathbb{E}[\hat{r}] = S^{-1}\sum_{s=1}^{S} \hat{r}_s$. We set $S = 100$ Monte Carlo replications. These measures summarize the finite-sample accuracy of the estimators and the typical rank selected across simulations.

Table 1 Simulation Results: $\hat{\theta}$ in Design 1

(N,T)	DGP 1			DGP 2		
	First step		Second step	First step		Second step
	RMSE	$\mathbb{E}[\hat{r}]$	RMSE	RMSE	$\mathbb{E}[\hat{r}]$	RMSE
(100,100)	6.11	2.00	3.06	5.99	2.00	2.93
(150,100)	4.95	2.00	2.41	5.61	2.00	2.88
(200,100)	5.35	2.00	2.33	5.61	2.00	2.54
(150,150)	4.35	2.00	2.11	4.56	2.00	2.24
(200,150)	2.11	2.00	1.75	2.52	2.00	1.73
(200,200)	1.88	2.00	1.57	2.03	2.00	1.65

Note: This table reports the RMSE for both the first-step and second-step estimators, as well as the average estimated rank from the first-step procedure. Results are presented for $\hat{\theta}$ under Design 1. RMSE values are expressed as percentages relative to the true parameter values (%), and $\mathbb{E}[\hat{r}]$ denotes the mean estimated number of factors computed across 100 Monte Carlo replications.

Table 2 Simulation Results: $\hat{\tilde{\beta}}$ in Design 2

(N,T)	DGP 1			DGP 2		
	First step		Second step	First step		Second step
	RMSE	$\mathbb{E}[\hat{r}]$	RMSE	RMSE	$\mathbb{E}[\hat{r}]$	RMSE
(100,100)	8.38	2.05	7.38	8.19	2.04	7.51
(150,100)	7.68	2.00	6.50	7.68	2.00	6.71
(200,100)	6.97	2.00	5.70	6.90	2.00	5.62
(150,150)	6.08	2.00	4.53	6.33	2.00	4.90
(200,150)	5.20	2.00	3.63	5.35	2.00	3.90
(200,200)	3.02	2.00	2.76	2.74	2.00	2.13

Note: This table reports the RMSE for both the first-step and second-step estimators, as well as the average estimated rank from the first-step procedure. Results are presented for $\hat{\tilde{\beta}}$ under Design 2. RMSE values are expressed as percentages relative to the true parameter values (%), and $\mathbb{E}[\hat{r}]$ denotes the mean estimated number of factors computed across 100 Monte Carlo replications.

Table 1 summarizes the finite-sample performance of the proposed estimators. Several key observations can be made from the results. (i) At each sample size, the second-step estimators exhibit smaller RMSE compared to the first-step estimators; (ii) As the sample size increases, particularly when both N and T grow, the RMSE of the first-step estimators tends to decrease; (iii) Similarly, the RMSE of the second-step estimators generally decreases with increasing sample size; (iv) The rank estimator correctly estimates the rank as the sample size increases, with the average rank stabilizing at $\hat{r} = 2$ for larger samples. These observations confirm the consistency results for both the first-step and the second-step estimators. For example, the RMSE decreases from 3.06% when $N = T = 100$ to 1.57% at $N = T = 200$, consistent with the theoretical prediction that estimator performance improves with larger sample sizes.

Table 2 summarizes the results for Design 2, which incorporates random coefficients. The findings parallel those of Design 1: the second-step estimator consistently improves upon the initial estimator, and both bias and RMSE decline with sample size. In this case, we focus on the accuracy of the estimated mean of the random coefficients, which is reliably recovered in larger samples.

In general, the simulations confirm the main theoretical insights. The two-step procedure delivers substantial efficiency gains, the estimators are consistent as both dimensions of the panel grow, and rank selection is reliable in sufficiently large samples.

6 Empirical Application

In this section, we apply our methodology to revisit the joint financing behavior of US non-financial corporations, focusing on how firms adjust debt issuance and equity repurchase decisions in response to relative valuations across capital markets. The objective is to illustrate the empirical applicability of our approach by extending the baseline model of Ma (2019) and to compare its benchmark fixed-effects logit with our proposed specification.

Empirical studies suggest that firms often act as cross-market arbitrageurs, substituting between debt and equity to exploit differences in financing costs. When debt is relatively inexpensive, firms tend to issue debt and repurchase equity, whereas undervalued equity prompts them to issue equity and retire debt. In a recent study, Ma (2019) documents that a substantial share of financing flows originate from non-financial firms that simultaneously issue in one market while repurchasing in another. About 45 percent of quarterly net equity repurchases (by value) occur in periods when firms are concurrently net issuers of debt, and roughly 50 percent of seasoned equity issuance takes place among firms that are net retiring debt. These coordinated patterns indicate that firms actively rebalance their capital structures to exploit valuation differentials across markets, with this behavior most pronounced among large, profitable, and financially unconstrained firms.

To quantify this relationship, the study estimates the probability that a firm simultaneously issues equity and retires debt using a panel logit model with firm fixed effects:

$$\Pr\left(y_{it}=1 \mid X_{D,it-1}, X_{E,it-1}, Z_{it}\right)=\Lambda\left(\lambda_i+\beta_D' X_{D,it-1}+\beta_E' X_{E,it-1}+\gamma' Z_{it}\right) \tag{6.1}$$

where $\Lambda(\cdot)$ denotes the logistic cumulative distribution function. The binary outcome variable y_{it} equals one if the firm issues equity (negative net equity repurchases) and retires debt (negative net debt issuance) in the same quarter, and zero otherwise. Net equity repurchases are measured as purchases minus sales of common and preferred stock (PRSTKC−SSTK), and net debt issuance is defined as long-term debt issuance minus long-term debt reduction (DLTIS−DLTR). The vectors $X_{D,it-1}$ and $X_{E,it-1}$ capture debt and equity market valuations measured at the end of the quarter $t-1$, while Z_{it} includes additional firm-level controls that can affect financing behavior.

The model uses three firm-level indicators of market valuation. Two bonds-based measures, the credit spread and the term spread, capture relative valuations in the debt market ($X_{D,it-1}$). The firm-level credit spread is computed as the face-value-weighted average yield differential between the firm's bonds and the nearest-maturity Treasury, while the term spread is the yield difference between the nearest-maturity Treasury and the three-

month Treasury bill.

The valuation of the equity market ($X_{E,it-1}$) is proxied by the firm's specific value-price ratio (V / P), where V denotes the intrinsic value of the equity estimated from the residual income model and P represents the market price of equity (Dong, Hirshleifer, and Teoh, 2012). Control variables include net income, cash holdings, capital expenditures, deviations from target leverage, asset growth, and firm size.[7]

Before introducing our model, we first examine whether the data exhibit latent time-varying structure by estimating a two-way fixed-effects (TWFE) specification following Fernández-Val and Weidner (2016) and analyzing its residuals,

$$\widehat{u}_{TWFE,it} = y_{it} - \Lambda\left(\hat{\lambda}_i + \hat{f}_t + \hat{\beta}_D' X_{D,i,t-1} + \hat{\beta}_E' X_{E,i,t-1} + \hat{\gamma}' Z_{it}\right)$$

Figure A2 shows systematic temporal and cross-sectional patterns: firms display persistent deviations across quarters (horizontal streaks), and many firms experience synchronized movements within the same periods (vertical color bands). These residual patterns point to time-varying unobserved heterogeneity.

6.1 Data and Model

We begin with a brief overview of the data and the model. Following Ma (2019), this analysis integrates several primary data sources, combining firm-level information on bond prices, equity valuations, and accounting fundamentals with aggregate market indicators. Bond data is obtained primarily from the Trade Reporting and Compliance Engine (TRACE) and supplemented by Datastream and Mergent's Fixed Income Securities Database (FISD). Equity returns and valuation measures come from CRSP and the Institutional Brokers' Estimate System (IBES), while balance sheet and cash flow variables are drawn from Compustat.[8] The sample covers the period 2003-2024, producing a panel of $N = 212$ firms observed in $T = 88$ quarters.[9] Summary statistics for key variables appear in Table 3.

We extend this framework by incorporating interactive fixed effects and random coefficients, allowing for unobserved common shocks and heterogeneous firm responses to market

[7] See Ma (2019) for detailed definitions and data construction.

[8] Access to TRACE, Compustat, CRSP, and IBES is provided through Wharton Research Data Services (WRDS), while Datastream is accessed via the LESG portal.

[9] Additional details on data sources, variable definitions, and sample construction are provided in Appendix D.1.

valuations. Formally, the latent index is given by

$$\Pr\left(y_{it}=1 \mid X_{D,it-1}, X_{E,it-1}, Z_{it}\right)=\Lambda\left(\lambda_i' f_t+\beta_{D,it}' X_{D,it-1}+\beta_{E,it}' X_{E,it-1}+\gamma' Z_{it}\right) \tag{6.2}$$

where f_t denotes latent factors that capture aggregate market shocks, λ_i are the corresponding firm-specific factor loadings, and $\beta_{D,it}$ and $\beta_{E,it}$ represent random coefficients reflecting heterogeneity in sensitivities to debt and equity valuations. Specifically, we assume that $\begin{pmatrix}\beta_{D,it}\\ \beta_{E,it}\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}\bar{\beta}_D\\ \bar{\beta}_E\end{pmatrix}, \Sigma_\beta\right)$. We employ the same set of market valuation and control variables as in (6.1).

Table 3 Summary Statistics of Key Variables

Variable	**N**	**Mean**	**SD**	$\mathbf{P}_{10}$	**Median**	$\mathbf{P}_{90}$
Credit spread	14,667	2.03	1.76	0.73	1.48	3.90
Term spread	14,554	1.19	1.07	−0.28	1.23	2.54
V/P ratio	14,667	1.08	0.63	0.42	0.95	1.88
Cash holdings	14,557	7.68	7.99	0.83	5.15	17.39
Asset growth	14,622	6.34	14.45	−5.71	4.57	19.97
Size	14,667	9.38	1.20	7.88	9.29	10.90
Net income	14,639	1.52	2.06	0.05	1.47	3.52
Capital expenditures	14,573	1.30	1.21	0.28	0.94	2.66

6.2 Analysis

We report our estimation results in Table 4. Model A corresponds to the conventional panel logit specification with individual fixed effects, as defined in (6.1), and serves as the benchmark. It is estimated using the conditional maximum likelihood approach following Ma (2019). In contrast, Model B adopts the specification in (6.2), which extends the benchmark by allowing for random coefficients on lagged valuation measures and by incorporating interactive fixed effects. The rank of the interactive component is estimated using the procedure described in (3.4); the first-step estimation yields an estimated rank of $\hat{r}=2$. Accordingly, Table 4 reports the second-step estimates, including the means and standard deviations of the random coefficients.

In general, the results re-confirm the importance of market valuations in shaping joint financing decisions of firms and are consistent with economic intuition: firms are more likely to simultaneously issue equity and retire debt when the cost of debt is high (e.g.,

elevated bond spreads or expected excess returns) and the cost of equity is low (e.g., a low V/P ratio or low expected excess equity returns).

We find that the signs of the estimated effects are generally robust to the inclusion of latent factors, with variations in their magnitudes across specifications. For instance, the effect of lagged term spread is larger in our model than in the benchmark specification: a one-standard-deviation decrease in term spread multiplies the odds of joint equity issuance and debt retirement by about 1.33 on average across firms, compared with roughly 1.25 in Model A. We also uncover substantial heterogeneity across firms in their sensitivities to valuation measures. In particular, the estimated standard deviation of the random coefficient on the lagged credit spread is 0.363 and statistically significant, indicating meaningful variation in firms' responses to credit market conditions.

We regard the comparison between our results and those from the benchmark model as an illustration that highlights the empirical value of allowing richer forms of unobserved heterogeneity. Firms adjust their financing decisions in response to evolving macrofinancial conditions and to unobserved, time-varying shocks—such as shifts in credit availability, monetary policy, or market liquidity—that jointly influence financing behavior and asset valuations. At the same time, firms differ in the extent to which they adjust their financing in response to observed market valuations. By incorporating interactive fixed effects and random coefficients, the proposed approach flexibly captures both common dynamics and firm-specific heterogeneity, demonstrating its applicability to complex empirical settings.

Table 4 Empirical Illustration: Comparison of Results

	Model A		Model B			
	β		Mean ($\bar{\beta}$)		SD (σ_β)	
Variable	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
L.Credit spread	0.077	0.015	0.124	0.014	0.363	0.055
L.Term spread	0.212	0.027	0.267	0.037	0.335	0.159
L.V/P	−0.092	0.046	−0.118	0.056	0.316	0.341
Net income	−0.039	0.013	−0.044	0.013	–	
L.Cash holding	−0.006	0.005	0.023	0.003	–	
CAPX	−0.242	0.037	−0.034	0.020	–	
L.Size	−0.746	0.057	−0.268	0.008	–	
L.Asset growth	0.011	0.002	0.031	0.002	–	

7 Conclusion

This paper explores nonlinear panel data models with interactive fixed effects and tackles nonconvex objective functions in large N and T settings. We propose a two-step estimation method using nuclear norm regularization and an iterative process. First, we obtain preliminary estimates of the slope coefficient, factors, and factor loadings using NNR. The second step refines these estimates through an iterative process. We establish the consistency of the preliminary estimator and derive the asymptotic distribution for the second-step estimator. The performance of the estimator is demonstrated through Monte Carlo simulations and an empirical application to a random coefficient binary logit model.

Several extensions remain open for future research. First, it would be valuable to explore whether post-NNR inference can be achieved within a finite number of iterations, extending the work of Chernozhukov et al. (2023) and Choi, Kwon, and Liao (2024) to nonlinear panel settings. Second, while this paper accounts for certain heterogeneity in slope coefficients within the random coefficient logit panel model, allowing for even richer forms of coefficient heterogeneity, as discussed by Bonhomme and Denis (2024), presents an interesting avenue for further study. These topics are left for future exploration.

References

Agarwal, Alekh, Sahand N. Negahban, and Martin J. Wainwright. 2012. "Noisy Matrix Decomposition via Convex Relaxation: Optimal Rates in High Dimensions." *The Annals of Statistics* 40 (2).

Agarwal, Anish, Munther Dahleh, Devavrat Shah, and Dennis Shen. 2021. "Causal Matrix Completion." .

Ahn, Seung C. and Alex R. Horenstein. 2013. "Eigenvalue Ratio Test for the Number of Factors." *Econometrica* 81 (3):1203–1227.

Ando, Tomohiro, Jushan Bai, and Kunpeng Li. 2022. "Bayesian and Maximum Likelihood Analysis of Large-Scale Panel Choice Models with Unobserved Heterogeneity." *Journal of Econometrics* 230 (1):20–38.

Armstrong, Timothy B, Martin Weidner, and Andrei Zeleneev. 2023. "Robust Estimation and Inference in Panels with Interactive Fixed Effects." .

Athey, Susan, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. 2021. "Matrix Completion Methods for Causal Panel Data Models." *Journal of the American Statistical Association* 116 (536):1716–1730.

Bai, Jushan. 2003. "Inferential Theory for Factor Models of Large Dimensions." *Econometrica* 71 (1):135–171.

———. 2009. "Panel Data Models with Interactive Fixed Effects." *Econometrica* 77 (4):1229–1279.

Bai, Jushan and Serena Ng. 2002. "Determining the Number of Factors in Approximate Factor Models." *Econometrica* 70 (1):191–221.

Belloni, Alexandre, Mingli Chen, Oscar Hernan Madrid Padilla, and Zixuan (Kevin) Wang. 2023. "High-Dimensional Latent Panel Quantile Regression with an Application to Asset Pricing." *The Annals of Statistics* 51 (1).

Belloni, Alexandre and Victor Chernozhukov. 2011. "ℓ1-Penalized Quantile Regression in High-Dimensional Sparse Models." *The Annals of Statistics* 39 (1).

Belloni, Alexandre, Victor Chernozhukov, Denis Chetverikov, Christian Hansen, and Kengo Kato. 2018. "High-Dimensional Econometrics and Regularized GMM."

Berbee, Henry. 1987. "Convergence Rates in the Strong Law for Bounded Mixing Sequences." *Probability Theory and Related Fields* 74 (2):255–270.

Beyhum, Jad and Eric Gautier. 2019. "Square-Root Nuclear Norm Penalized Estimator for Panel Data Models with Approximately Low-Rank Unobserved Heterogeneity."

Beyhum, Jad and François Portier. 2024. "High-Dimensional Nonconvex LASSO-type M-estimators." *Journal of Multivariate Analysis* 202:105303.

Boneva, Lena and Oliver Linton. 2017. "A Discrete-Choice Model for Large Heterogeneous Panels with Interactive Fixed Effects with an Application to the Determinants of Corporate Bond Issuance." *Journal of Applied Econometrics* 32 (7):1226–1243.

Bonhomme, Stephane and Angela Denis. 2024. "Fixed Effects and Beyond: Bias Reduction, Groups, Shrinkage, and Factors in Panel Data." .

Boyd, Stephen, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. 2011.

Bryzgalova, Svetlana, Sven Lerner, Martin Lettau, and Markus Pelger. 2022. "Missing Financial Data." *SSRN Electronic Journal* .

Cahan, Ercument, Jushan Bai, and Serena Ng. 2023. "Factor-Based Imputation of Missing Values and Covariances in Panel Data of Large Dimensions." *Journal of Econometrics* 233 (1):113–131.

Cai, Jian-Feng, Emmanuel J. Candes, and Zuowei Shen. 2008. "A Singular Value Thresholding Algorithm for Matrix Completion." https://arxiv.org/abs/0810.3286v1.

Candes, Emmanuel J and Yaniv Plan. 2010. "Matrix Completion With Noise." *Proceedings of the IEEE* 98 (6):925–936.

Candès, Emmanuel J. and Benjamin Recht. 2009. "Exact Matrix Completion via Convex Optimization." *Foundations of Computational Mathematics* 9 (6):717–772.

Chatterjee, Sourav. 2015. "Matrix Estimation by Universal Singular Value Thresholding." *The Annals of Statistics* 43 (1):177–214.

Chen, Liang, Juan J. Dolado, and Jesús Gonzalo. 2021. "Quantile Factor Models." *Econometrica* 89 (2):875–910.

Chen, Liang and Minyuan Zhang. 2025. "Common Correlated Effects Estimation of Nonlinear Panel Data Models." *The Econometrics Journal* 28 (2):295–317.

Chen, Mingli, Iván Fernández-Val, and Martin Weidner. 2021. "Nonlinear Factor Models for Network and Panel Data." *Journal of Econometrics* 220 (2):296–324.

Chen, Mingli, Marc Rysman, Shuang Wang, and Krzysztof Wozniak. 2022. "An MM Algorithm for Fixed Effects in Multinomial Models with an Application to Payment Choice in Grocery Stores." .

Chen, Xiaohong and Timothy Christensen. 2015. "Optimal Uniform Convergence Rates and Asymptotic Normality for Series Estimators Under Weak Dependence and Weak Conditions." *Journal of Econometrics* 188 (2):447–465.

Chen, Yixiao, Ke Miao, and Liangjun Su. 2025. "High Dimensional Discrete Choice Models With Interactive Fixed Effects Applied to Causal Inference." *Journal of Applied Econometrics* :1–19.

Chernozhukov, Victor, Christian Hansen, Yuan Liao, and Yinchu Zhu. 2019. "Inference for Heterogeneous Effects Using Low-Rank Estimation of Factor Slopes." .

———. 2023. "Inference for Low-Rank Models." *The Annals of Statistics* 51 (3):1309–1330.

Chetverikov, Denis and Jesper Riis-Vestergaard Sørensen. 2025. "Selecting Penalty Parameters of High-Dimensional M-Estimators Using Bootstrapping after Cross Validation." *Journal of Political Economy* 133 (10):3208–3248.

Choi, Jungjun, Hyukjun Kwon, and Yuan Liao. 2024. "Inference for Low-Rank Completion without Sample Splitting with Application to Treatment Effect Estimation." *Journal of Econometrics* 240 (1):105682.

Dhaene, Geert and Koen Jochmans. 2015. "Split-Panel Jackknife Estimation of Fixed-effect Models." *The Review of Economic Studies* 82 (3):991–1030.

Dong, Ming, David Hirshleifer, and Siew Hong Teoh. 2012. "Overvalued Equity and Financing Decisions." *The Review of Financial Studies* 25 (12):3645–3683.

Fan, Jianqing, Wenyan Gong, and Ziwei Zhu. 2017. "Generalized High-Dimensional Trace Regression via Nuclear Norm Regularization." .

Feng, Junlong. 2023. "Nuclear Norm Regularized Quantile Regression With Interactive Fixed Effects." *Econometric Theory* :1–31.

Fernández-Val, Iván and Martin Weidner. 2016. "Individual and Time Effects in Nonlinear Panel Models with Large N , T." *Journal of Econometrics* 192 (1):291–312.

———. 2018. "Fixed Effects Estimation of Large-T Panel Data Models." *Annual Review of Economics* 10 (1):109–138.

Freyberger, Joachim, Bjoern Hoeppner, Andreas Neuhierl, and Michael Weber. 2025. "Missing Data in Asset Pricing Panels." *The Review of Financial Studies* 38 (3):760–802.

Gao, Jiti, Fei Liu, Bin Peng, and Yayi Yan. 2023. "Binary Response Models for Heterogeneous Panel Data with Interactive Fixed Effects." *Journal of Econometrics* 235 (2):1654–1679.

Greenwood, Robin and Samuel G. Hanson. 2013. "Issuer Quality and Corporate Bond Returns." *The Review of Financial Studies* 26 (6):1483–1525.

Hahn, Jinyong and Guido Kuersteiner. 2002. "Asymptotically Unbiased Inference for a Dynamic Panel Model with Fixed Effects When Both n and T Are Large." *Econometrica* 70 (4):1639–1657.

Hastie, Trevor, Robert Tibshirani, and Martin Wainwright. 2015. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC.

Higgins, Ayden and Koen Jochmans. 2024. "Bootstrap Inference for Fixed-Effect Models." *Econometrica* 92 (2):411–427.

Hong, Shengjie, Liangjun Su, and Tao Jiang. 2023. "Profile GMM Estimation of Panel Data Models with Interactive Fixed Effects." *Journal of Econometrics* 235 (2):927–948.

Huang, Shimeng and Henry Wolkowicz. 2018. "Low-Rank Matrix Completion Using Nuclear Norm Minimization and Facial Reduction." *Journal of Global Optimization* 72 (1):5–26.

Hughes, David William. 2022. "Estimating Nonlinear Network Data Models with Fixed Effects."

Hunter, David R and Kenneth Lange. 2004. "A Tutorial on MM Algorithms." *The American Statistician* 58 (1):30–37.

James, Jonathan. 2017. "MM Algorithm for General Mixed Multinomial Logit Models." *Journal of Applied Econometrics* 32 (4):841–857.

Lange, Kenneth. 2016. *MM Optimization Algorithms*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Lee, Lung-Fei. 1995. "Asymptotic Bias in Simulated Maximum Likelihood Estimation of Discrete Choice Models." *Econometric Theory* 11 (3):437–483.

Ma, Shujie, Liangjun Su, and Yichong Zhang. 2021. "Detecting Latent Communities in Network Formation Models."

Ma, Yueran. 2019. "Nonfinancial Firms as Cross-Market Arbitrageurs." *The Journal of Finance* 74 (6):3041–3087.

Mazumder, Rahul, Trevor Hastie, and Robert Tibshirani. 2010. "Spectral Regularization Algorithms for Learning Large Incomplete Matrices." *Journal of Machine Learning Research* 11 (80):2287–2322.

Miao, Ke, Peter C.B. Phillips, and Liangjun Su. 2022. "High-Dimensional VARs with Common Factors." *Journal of Econometrics* .

Moon, Hyungsik Roger and Martin Weidner. 2015. "Linear Regression for Panel With Unknown Number of Factors as Interactive Fixed Effects." *Econometrica* 83 (4):1543–1579.

———. 2017. "Dynamic Linear Panel Regression Models With Interactive Fixed Effects." *Econometric Theory* 33 (1):158–195.

———. 2019. "Nuclear Norm Regularized Estimation of Panel Regression Models."

Negahban, Sahand and Martin J. Wainwright. 2011. "Estimation of (near) Low-Rank Matrices with Noise and High-Dimensional Scaling." *The Annals of Statistics* 39 (2).

Negahban, Sahand N., Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. 2012. "A Unified Framework for High-Dimensional Analysis of M-Estimators with Decomposable Regularizers." *Statistical Science* 27 (4).

Pesaran, M. Hashem. 2006. "Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure." *Econometrica* 74 (4):967–1012.

Robertson, Donald and Vasilis Sarafidis. 2015. "IV Estimation of Panels with Factor Residuals." *Journal of Econometrics* 185 (2):526–541.

Semenova, Vira, Matt Goldman, Victor Chernozhukov, and Matt Taddy. 2023. "Inference on Heterogeneous Treatment Effects in High-Dimensional Dynamic Panels under Weak Dependence." *Quantitative Economics* 14 (2):471–510.

Städler, Nicolas, Peter Bühlmann, and Sara van de Geer. 2010. "L1-Penalization for Mixture Regression Models." *TEST* 19 (2):209–256.

Su, Liangjun, Zhentao Shi, and Peter C. B. Phillips. 2016. "Identifying Latent Structures in Panel Data." *Econometrica* 84 (6):2215–2264.

Sun, Yixiao and Min Seong Kim. 2010. "K-Step Bootstrap Bias Correction for Fixed Effects Estimators in Nonlinear Panel Models." *SSRN Electronic Journal* .

Train, Kenneth E. 2009. *Discrete Choice Methods with Simulation*. Cambridge: Cambridge University Press, 2 ed.

van der Vaart, Aad W. and Jon A. Wellner. 1996. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. New York, NY: Springer New York.

Vu, Van H. 2007. "Spectral Norm of Random Matrices." *Combinatorica* 27 (6):721–736.

Wainwright, Martin J. 2019. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press.

Wang, Fa. 2022. "Maximum Likelihood Estimation and Inference for High Dimensional Generalized Factor Models with Application to Factor-Augmented Regressions." *Journal of Econometrics* 229 (1):180–200.

Wang, Yu, Wotao Yin, and Jinshan Zeng. 2019. "Global Convergence of ADMM in Nonconvex Nonsmooth Optimization." *Journal of Scientific Computing* 78:29–63.

Yu, Bin. 1994. "Rates of Convergence for Empirical Processes of Stationary Mixing Sequences." *The Annals of Probability* 22 (1).

Yuan, Xiaoming and Junfeng Yang. 2013. "Sparse and low-rank matrix decomposition via alternating direction method." *Pacific Journal of Optimization* 9:167.

Zeleneev, Andrei and Weisheng Zhang. 2025. "Tractable Estimation of Nonlinear Panels with Interactive Fixed Effects." URL https://arxiv.org/abs/2511.15427.

Appendix

A Implementation of the First-Step Estimator

This section describes the implementation of the first-step estimator. Conventional matrix-completion methods are based on linear formulations and are therefore ineffective for data with nonlinear structures. To address this limitation, we employ two optimization algorithms specifically tailored to our models. For the convex single-index panel model, we adopt the Alternating Direction Method of Multipliers (ADMM), which decomposes the overall optimization problem into smaller, more manageable subproblems. This decomposition enables efficient parallelization and ensures convergence in convex settings (Boyd, Parikh, Chu, Peleato, and Eckstein, 2011). ADMM is particularly well suited to large-scale problems with separable objective functions.[A1]

On the other hand, we develop a new method based on the Majorization-Minimization (MM) algorithm for the random coefficient panel model. This algorithm, which can be seen as a generalization of the Expectation- Maximization (EM) algorithm, has been developed in the statistics literature (Hunter and Lange, 2004, Lange, 2016). We utilize the MM algorithm to separate random coefficients and interactive fixed effects. That is, we propose a new surrogate function to linearize the logit model so that we can find the closed-form solution to the IFEs. Sequential fixed effects estimation and majorization deliver estimates with significantly reduced computational and memory costs.

A.1 Single-Index Panel Model

We develop an Alternating Direction Method of Multipliers (ADMM) algorithm for estimating the first-stage parameters in the general single-index panel model. The binary logit model in Example 3 serves as a leading illustration of this broader framework.

To reformulate the original problem, we also introduce slack variables. As a result,

[A1] Although ADMM is primarily designed for convex problems, it has been successfully applied to some non-convex settings; see, for example, Wang, Yin, and Zeng (2019).

(3.2) is equivalent to

$$\min_{\theta\in\mathbb{R}^p,\Pi\in\mathbb{R}^{N\times T}} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\ell\left(Y_{it},V_{it}\right)+\nu\|\Pi\|_* \\ \text{s.t. } V=\sum_{j=1}^{p}X_j\theta_j+Z_\Pi,\quad Z_\Pi-\Pi=0 \tag{A.1}$$

To solve the minimization problem (A.1), the proposed algorithm uses the following augmented Lagrangian

$$\mathscr{L}\left(\theta,\Pi,V,Z_\Pi,U_p,U_v\right)=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\ell\left(Y_{it},V_{it}\right)+\nu\|\Pi\|_* \\ +\frac{\eta}{2NT}\left\|V-\sum_{j=1}^{p}X_j\theta_j-Z_\Pi+U_p\right\|_F^2+\frac{\eta}{2NT}\left\|Z_\Pi-\Pi+U_v\right\|_F^2$$

Here, U_p and U_v are the scaled dual variables corresponding to the linear constraints $V=\sum_{j=1}^{p}X_j\theta_j+Z_\Pi$ and $Z_\Pi-\Pi=0$, and $\eta>0$ is the penalty parameter for constraint violations. Due to the separability of the parameters in $\mathscr{L}$, the ADMM algorithm proceeds by iteratively minimizing the augmented Lagrangian in blocks with respect to the original variables, in this case (V,θ,Π) and Z_Π, and then then updating the dual variables U_p and U_v.

We solve the minimization problems iteratively:

$$V^{(k+1)}\leftarrow\arg\min_{V}\sum_{i=1}^{N}\sum_{t=1}^{T}\ell\left(Y_{it},V_{it}\right)+\frac{\eta}{2}\left\|V-\sum_{j=1}^{p}X_j\theta_j^{(k)}-Z_\Pi^{(k)}+U_p^{(k)}\right\|_F^2$$

$$\theta^{(k+1)}\leftarrow\arg\min_{\theta}\frac{\eta}{2}\left\|V^{(k+1)}-\sum_{j=1}^{p}X_j\theta_j-Z_\Pi^{(k)}+U_p^{(k)}\right\|_F^2$$

$$\Pi^{(k+1)}\leftarrow\arg\min_{\Pi}\frac{\eta}{2}\left\|Z_\Pi^{(k)}-\Pi+U_v^{(k)}\right\|_F^2+NT\nu\|\Pi\|_*$$

$$Z_\Pi^{(k+1)}\leftarrow\arg\min_{Z_\Pi}\frac{\eta}{2}\left\|V^{(k+1)}-\sum_{j=1}^{p}X_j\theta_j^{(k+1)}-Z_\Pi+U_p^{(k)}\right\|_F^2+\frac{\eta}{2}\left\|Z_\Pi-\Pi^{(k+1)}+U_v^{(k)}\right\|_F^2$$

$$U_p^{(k+1)}\leftarrow\arg\min_{U_p}\frac{\eta}{2}\left\|V^{(k+1)}-\sum_{j=1}^{p}X_j\theta_j^{(k+1)}-Z_\Pi^{(k+1)}+U_p\right\|_F^2$$

$$U_v^{(k+1)}\leftarrow\arg\min_{U_v}\frac{\eta}{2}\left\|Z_\Pi^{(k+1)}-\Pi^{(k+1)}+U_v\right\|_F^2$$

For single-index models, the loss function $\ell(y, v)$ is differentiable in v. Let $h(y, v) = \nabla_v \ell(y, v)$ denote its derivative with respect to the index. Then the first-step update for V can be written as

$$V^{(k+1)} \leftarrow V^{(k)} - \eta \left(V^{(k)} - X\theta^{(k)} - Z_{\Pi}^{(k)} + U_p^{(k)} \right) - h\left(Y, V^{(k)}\right),$$

which corresponds to a gradient-descent or Fisher-scoring step depending on the form of ℓ.

For instance, in the binary logit model considered in Example 3, the loss function is $\ell(y, v) = -yv + \log(1 + \exp(v))$, whose derivative is $h(y, v) = \Lambda(v) - y$ with $\Lambda(v) = (1 + e^{-v})^{-1}$. Hence,

$$h(Y_{it}, V_{it}^{(k)}) = \Lambda(V_{it}^{(k)}) - Y_{it}$$

recovering the update rule used in the binary logit specification of Simulation Design 1.

Additionally, following Cai, Candes, and Shen (2008) and Yuan and Yang (2013), we update Π using singular value thresholding (SVT). Let $A = U_A \Sigma_A V_A'$ represents the singular value decomposition (SVD) of matrix $A \in \mathbb{R}^{N \times T}$, where $\Sigma_A = \text{diag}\left(\{\sigma_s\}_{s \in [N \wedge T]}\right)$. For $\iota > 0$, the soft-thresholding operator $S_\iota(\cdot)$ is defined as

$$S_\iota(A) \coloneqq U_A S_\iota(\Sigma_A) V_A', \quad S_\iota(\Sigma_A) = \text{diag}\left(\{\sigma_s - \iota\}_+\right)$$

where t_+ is the positive part of t, namely, $t_+ = \max(0, t)$. Further details can be found in Theorem 2.1 of Cai, Candes, and Shen (2008).

Lastly, the remaining parameters, θ and slack variables, are solved using standard least-squares estimation, which admits closed-form solutions.

A.2 Random Coefficient Logit Panel

To expose the nature of the MM algorithm, we temporarily disregard the penalty term. Suppose our goal is to minimize $\mathcal{L}_{NT}(\Psi) : \mathbb{R}^{\dim(\Psi)} \mapsto \mathbb{R}$. Let $\Psi^{(k)}$ denote the current iterate in the minimization process. As suggested by the name, the majorization-minimization algorithm proceeds in two steps. First, we construct a surrogate function $Q_{NT}\left(\Psi|\Psi^{(k)}\right)$. Specifically, to update Ψ at step $k+1$, we use a surrogate function $Q_{NT}\left(\Psi|\Psi^{(k)}\right)$ that satisfies

1. $Q_{NT}(\Psi^{(k)}|\Psi^{(k)}) = \mathcal{L}_{NT}(\Psi^{(k)})$
2. $Q_{NT}(\Psi|\Psi^{(k)}) \geq \mathcal{L}_{NT}(\Psi) \quad$ for all Ψ

Here, the function $Q_{NT}\left(\Psi|\Psi^{(k)}\right)$ is said to majorize $\mathcal{L}_{NT}(\Psi)$ at $\Psi^{(k)}$. The current iterate

$\Psi^{(k)}$ is treated as constant. In the second step in the MM algorithm, we update $\Psi$by choosing $\Psi^{(k+1)}$ to minimize $Q_{NT}\left(\Psi|\Psi^{(k)}\right)$. However, constructing a surrogate function that both majorizes $\mathcal{L}_{NT}(\Psi)$ at $\Psi^{(k)}$ and is easy to minimize is often a challenge. This step is analogous to the expectation (E) step in a well-structured Expectation-Maximization (EM) algorithm.

For the random coefficient logit panel model, we define $\theta = \left(\bar{\beta}, \Sigma\right)$ and $\Psi = (\theta, \Pi)$. The MM algorithm updates the parameters $\theta = \left(\bar{\beta}, \Sigma\right)$ and Π separately by using two surrogate functions $Q_{NT}\left(\theta|\Psi^{(k)}\right)$ and $Q_{NT}\left(\Pi|\Psi^{(k)}\right)$. The surrogate function $Q_{NT}\left(\theta|\Psi^{(k)}\right)$ follows from Train (2009), and we extend the work of Train (2009), James (2017) and Chen, Rysman, Wang, and Wozniak (2022) by introducing a new surrogate function $Q_{NT}\left(\Pi|\Psi^{(k)}\right)$ which effectively handles the penalty term.

In Example 4, the probability of $y_{it} = 1$ for individual i and time t, given the random coefficient β and fixed effect π_{it}, is expressed as

$$p_{it}\left(\beta, \pi_{it}\right) = \frac{\exp\left(X_{it}'\beta + \pi_{it}\right)}{1 + \exp\left(X_{it}'\beta + \pi_{it}\right)} \tag{A.2}$$

The likelihood is given by

$$L_{it}(\beta, \pi_{it}) = p_{it}(\beta, \pi_{it})^{y_{it}}(1 - p_{it}(\beta, \pi_{it}))^{1-y_{it}} \tag{A.3}$$

Thus, the negative log-likelihood function is computed by integrating the individual-specific and time-specific probabilities over all observed data:

$$\mathcal{L}_{NT}\left(\Psi\right) = -\sum_{i=1}^{N}\sum_{t=1}^{T}\log\left[\int L_{it}(\beta, \pi_{it}) f\left(\beta\right) d\beta\right]$$

The integral in the negative log-likelihood does not have a known closed-form expression and is typically simulated numerically in estimation.

Rather than trying to directly optimize $\mathcal{L}_{NT}\left(\Psi\right)$, the strategy of the MM algorithm is to transfer optimization to simpler surrogate functions that guarantees descent of the negative log-likelihood. It refines the EM algorithm, as reviewed in Train (2009). In mixed logit models, the EM surrogate function takes the form,

$$Q_{NT}\left(\Psi|\Psi^{(k)}\right) = -\sum_{i=1}^{N}\sum_{t=1}^{T}\int \log\left[L_{it}(\beta, \pi_{it}) f\left(\beta|\theta\right) g_{it}\left(\beta|\Psi^{(k)}\right)\right] d\beta$$

In this expression, $g_{it}\left(\beta_{it}|\Psi^{(k)}\right)$ is an individual-and-time-specific probability density func-

tion of the unobserved data, β_{it}, conditional on conditioned on the observed data for individual at time t, and evaluated at $\Psi^{(k)}$. This density is derived using Bayes' rule,

$$g_{it}\left(\beta|\Psi^{(k)}\right) = \frac{L_{it}(\beta, \pi_{it}^{(k)}) f\left(\beta|\theta^{(k)}\right)}{\int_{\beta'} L_{it}(\beta', \pi_{it}^{(k)}) f\left(\beta'|\theta^{(k)}\right) d\beta'}$$

Evaluating this integral requires simulation. For each individual i at time t, we draw R samples from $N\left(\bar{\beta}^{(k)}, \Sigma^{(k)}\right)$, label each as β_{itr}, and compute the weights using the conditional likelihood function,

$$w_{itr}^{(k)} = \frac{L_{it}(\beta_{itr}, \pi_{it}^{(k)})}{\sum_{r'=1}^{R} L_{it}(\beta_{itr'}, \pi_{it}^{(k)})} \tag{A.4}$$

The superscript k on the weights indicates that they are a function of $\Psi^{(k)}$. As in maximum simulated likelihood, R must be large enough to eliminate the bis from simulating the denominator of the weights Lee (1995), Train (2009).

Using the approximation to $g_{it}\left(\beta|\Psi^{(k)}\right)$, the EM surrogate function becomes

$$\begin{aligned}
\tilde{Q}_{NT}\left(\Psi|\Psi^{(k)}\right) &= -\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{r=1}^{R} w_{itr}^{(k)} \log\left[L_{it}(\beta_{itr}, \pi_{it}) f\left(\beta_{itr}|\theta\right)\right] \\
&= \underbrace{-\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{r=1}^{R} w_{itr}^{(k)} \log\left[L_{it}(\beta_{itr}, \pi_{it})\right]}_{\tilde{Q}_{NT}\left(\Pi|\Psi^{(k)}\right)} \underbrace{-\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{r=1}^{R} w_{itr}^{(k)} \log\left[f\left(\beta_{itr}|\theta\right)\right]}_{Q_{NT}\left(\theta|\Psi^{(k)}\right)}
\end{aligned}$$

Since the EM surrogate function is additively separable in the parameters θ and Π under the log operator, it simplifies the optimization process by allowing the minimization to be performed over two independent functions rather than a complex joint maximization. Specifically, $Q_{NT}\left(\theta|\Psi^{(k)}\right)$ is the likelihood function for a multivariate normal distribution with weighted observations on β, where the solution for mean and covariance is obtained through sample analogs (Train, 2009). As a result, $\theta^{(k+1)} = \left(\bar{\beta}^{(k+1)}, \Sigma^{(k+1)}\right)$ admits a closed-form solution:

$$\begin{aligned}
\bar{\beta}^{(k+1)} &= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\sum_{r=1}^{R} w_{itr}^{(k)} \beta_{itr}\right) \\
\Sigma^{(k+1)} &= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\sum_{r=1}^{R} w_{itr}^{(k)} \beta_{itr}\beta_{itr}'\right) - \bar{\beta}^{(k+1)}\bar{\beta}^{(k+1)\prime}
\end{aligned}$$

Note that $\tilde{Q}_{NT}\left(\Psi|\Psi^{(k)}\right)$ serves as the surrogate function for $\mathcal{L}_{NT}\left(\Psi\right)$ without the penalty

term. Instead of optimizing over $\tilde{Q}_{NT}\left(\Pi|\Psi^{(k)}\right) + NT\nu\|\Pi\|_*$[A2], this paper derives an alternative surrogate function, providing a closed-form solution for Π. The main insight is to exploit a natural feature of discrete-choice models and recognize that $\tilde{Q}_{NT}\left(\Pi|\Psi^{(k)}\right)$, the weighted standard logit log-likelihood can be majorized with a quadratic upper bound function. We define $Q_{NT}\left(\Pi \mid \Psi^{(k)}\right)$ as the surrogate function for the penalized objective $\frac{1}{NT}\tilde{Q}_{NT}\left(\Pi \mid \Psi^{(k)}\right) + \nu\|\Pi\|_*$, which includes the scaled log-likelihood and nuclear-norm regularization. Specifically,

$$\begin{aligned} Q\left(\Pi \mid \theta^{(k)}, \Pi^{(k)}\right) =& \mathcal{L}_{NT}\left(\theta^{(k)}, \Pi^{(k)}\right) - \frac{1}{2NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\sum_{r=1}^{R} w_{itr}^{(k)} h\left(Y_{it}, X_{it}'\beta_{itr}^{(k)} + \pi_{it}^{(k)}\right)\right)^2 \\ &+ \frac{1}{2NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\pi_{it}^{(k)} - \sum_{r=1}^{R} w_{itr}^{(k)} h\left(Y_{it}, X_{it}'\beta_{itr}^{(k)} + \pi_{it}^{(k)}\right) - \pi_{it}\right)^2 + \nu\|\Pi\|_* \end{aligned}$$

where $h(y, v) = \Lambda(v) - y$ and $\Lambda(v) = \frac{1}{1+e^{-v}}$. Here, the closed-form solution for $\mathrm{Q}\left(\Pi|\Psi^{(k)}\right)$ is derived through SVT,

$$\Pi^{(k+1)} \leftarrow S_{NT\nu}\left(\Pi^{(k)} - \left(\sum_{r=1}^{R} w_{itr}^{(k)} h\left(Y_{it}, X_{it}'\beta_{itr} + \pi_{it}^{(k)}\right)\right)_{i,t}\right)$$

Here, the threshold $NT\nu$ follows directly from the scaling of the quadratic term.

B Proofs of Consistency and Convergence Rate Results

As in Berbee (1987), we employ the blocking argument to deal with β-mixing data. Similar arguments exist in different settings, e.g. Belloni, Chen, Madrid Padilla, and Wang (2023), Chen and Christensen (2015), Semenova, Goldman, Chernozhukov, and Taddy (2023), Yu (1994). Let block size be $c_T = \lceil 2\mu^{-1}\log(NT)\rceil$ with $1 \le c_T \le T/2$. For each $i \in [N]$, consider a partition of $\{W_{it} = (Y_{it}, X_{it})\}$ into $2d_T$ blocks of size c_T each, where $d_T = \lfloor T/2c_T \rfloor$. We denote random variables that correspond to block p as $\{W_{it}\}_p = \{W_{i,(p-1)c_T+1}, ...W_{i,pc_T}\}$ for $p \in [2d_T]$. Then we obtain the existence of sequence of random variables $\{W_{it}^*\}_{i\in[N],t\in[T]}$ such that

- $\{W_{it}^*\}_{i\in[N],t\in[T]}$ is independent of $\{W_{it}\}_{i\in[N],t\in[T]}$;

[A2] Note that the original problem is reweighted by NT.

- For a fixed t, $\{W_{it}^*\}_{i\in[N]}$ are independent;
- For a fixed i, $\{W_{it}^*\}_p$ and $\{W_{it}\}_p$ are identically distributed for each block p;
- For a fixed i, the odd-numbered blocks $\{\{W_{it}^*\}_1, \{W_{it}^*\}_3, ..., \{W_{it}^*\}_{2p-1}\}$ are independent, and the even-numbered blocks $\{\{W_{it}^*\}_2, \{W_{it}^*\}_4, ..., \{W_{it}^*\}_{2p}\}$ are independent.

Hence, denote H_l as the odd-numbered block, H_l' as the even-number block, and H_R as the remainder block of length $T - 2c_T d_T$. That is, for $l \in [d_T]$

$$\begin{aligned} H_l &= \{t : 1 + 2(l-1)c_T \leq t \leq (2l-1)c_T\} \\ H_l' &= \{t : 1 + (2l-1)c_T \leq t \leq 2lc_T\} \\ H_R &= \{t : 2c_T d_T + 1 \leq t \leq T\} \end{aligned}$$

Throughout the appendix, let $\underline{c}$ and $\bar{c}$ denote generic positive constants that may vary across their occurrences.

B.1 Proof of Results in Section 4.1

Lemma B.1. *We have $(\theta_0, \Pi_0) \in \mathcal{B}$ and if $\mathcal{L}_{NT}(\theta_0, \Pi_0) = \bar{\mathcal{L}}(\theta_0, \Pi_0) + o_p(1)$, then $\left(\hat{\theta}, \hat{\Pi}\right) \in \mathcal{B}$ with probability approaching one.*

Proof of Lemma B.1. Recall the definition of $\mathcal{B}$ is

$$\mathcal{B} = \left\{(\theta, \Pi) \in \Theta \times \Phi^{N\times T} : \|\Pi\|_* \leq \nu^{-1}\left(\bar{\mathcal{L}}(\theta_0, \Pi_0) + 1\right) + \|\Pi_0\|_*\right\}$$

Since $\bar{\mathcal{L}}(\theta_0, \Pi_0)$ is nonnegative, the first statement is clear, i.e., $(\theta_0, \Pi_0) \in \mathcal{B}$. For the second statement, since $0 \geq \mathcal{L}_{NT}\left(\hat{\theta}, \hat{\Pi}\right) - \mathcal{L}_{NT}(\theta_0, \Pi_0) + \nu\left(\|\hat{\Pi}\|_* - \|\Pi_0\|_*\right)$ and $\mathcal{L}_{NT}$ is nonnegative, we have $\nu\|\hat{\Pi}\|_* \leq \bar{\mathcal{L}}(\theta_0, \Pi_0) + o_p(1) + \nu\|\Pi_0\|_*$. Hence, $\left(\hat{\theta}, \hat{\Pi}\right) \in \mathcal{B}$ with probability approaching one. ■

Proof of Lemma 1. Let $\eta > 0$. By Assumption 3, there is $\varepsilon > 0$ such that for $(\theta, \Pi) \in \Theta \times \Phi^{N\times T}$ with $\|\theta - \theta_0\|^2 + \frac{1}{NT}\|\Pi - \Pi_0\|_F^2 \geq \eta^2$, we have $\bar{\mathcal{L}}(\theta, \Pi) - \bar{\mathcal{L}}(\theta_0, \Pi_0) \geq \varepsilon$. Thus, by the union bound,

$$\begin{aligned} &\mathrm{P}\left(\|\theta - \theta_0\|^2 + \frac{1}{NT}\|\Pi - \Pi_0\|_F^2 \geq \eta^2\right) \\ &\quad \leq \mathrm{P}\left(\left\{\bar{\mathcal{L}}(\theta, \Pi) - \bar{\mathcal{L}}(\theta_0, \Pi_0) \geq \varepsilon\right\} \cap \left\{\left(\hat{\theta}, \hat{\Pi}\right) \in \mathcal{B}\right\}\right) + \mathrm{P}\left(\left(\hat{\theta}, \hat{\Pi}\right) \notin \mathcal{B}\right) \end{aligned}$$

Since $(\theta_0, \Pi_0) \in \mathcal{B}$, we have $\mathcal{L}_{NT}(\theta_0, \Pi_0) = \bar{\mathcal{L}}(\theta_0, \Pi_0) + o_p(1)$ by Assumption UC, and $\mathrm{P}\left(\left(\hat{\theta}, \hat{\Pi}\right) \notin \mathcal{B}\right) = o(1)$ by Proposition B.1.

We have $\bar{\mathcal{L}}\left(\hat{\theta}, \hat{\Pi}\right) \geq \bar{\mathcal{L}}(\theta_0, \Pi_0)$ and $\mathcal{L}_{NT}\left(\hat{\theta}, \hat{\Pi}\right) - \mathcal{L}_{NT}(\theta_0, \Pi_0) \leq \nu\left(\|\Pi_0\|_* - \|\hat{\Pi}\|_*\right)$. It follows that, on the event $\left\{\left(\hat{\theta}, \hat{\Pi}\right) \in \mathcal{B}\right\}$,

$$\begin{aligned}
0 &\leq \bar{\mathcal{L}}\left(\hat{\theta}, \hat{\Pi}\right) - \bar{\mathcal{L}}(\theta_0, \Pi_0) \\
&\leq 2 \sup_{(\theta,\Pi)\in\mathcal{B}} \left|\mathcal{L}_{NT}(\theta, \Pi) - \bar{\mathcal{L}}(\theta, \Pi)\right| + \nu\left(\|\Pi_0\|_* - \|\hat{\Pi}\|_*\right) \\
&\leq 2 \sup_{(\theta,\Pi)\in\mathcal{B}} \left|\mathcal{L}_{NT}(\theta, \Pi) - \bar{\mathcal{L}}(\theta, \Pi)\right| + \nu\|\Pi_0\|_*
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathrm{P}\left(\left\{\bar{\mathcal{L}}(\theta, \Pi) - \bar{\mathcal{L}}(\theta_0, \Pi_0) \geq \varepsilon\right\} \cap \left\{\left(\hat{\theta}, \hat{\Pi}\right) \in \mathcal{B}\right\}\right) \\
&\leq \mathrm{P}\left(\left\{2 \sup_{(\theta,\Pi)\in\mathcal{B}} \left|\mathcal{L}_{NT}(\theta, \Pi) - \bar{\mathcal{L}}(\theta, \Pi)\right| + \nu\|\Pi_0\|_* \geq \varepsilon\right\} \cap \left\{\left(\hat{\theta}, \hat{\Pi}\right) \in \mathcal{B}\right\}\right) \\
&\leq \mathrm{P}\left(2 \sup_{(\theta,\Pi)\in\mathcal{B}} \left|\mathcal{L}_{NT}(\theta, \Pi) - \bar{\mathcal{L}}(\theta, \Pi)\right| + \nu\|\Pi_0\|_* \geq \varepsilon\right)
\end{aligned}$$

In spirit of Assumption UC and that $\nu\|\Pi_0\|_* = o(1)$, the above term goes to 0. Thus, the desired result follow. ■

B.2 Proof of Results in Section 4.2

Recall the definition of norm $\rho(\cdot, \cdot)$ as $\rho(\delta, \Delta) = \left[\|\delta\|^2 + \frac{1}{NT}\|\Delta\|_*^2\right]^{1/2}$. The empirical risk function over some bounded set $\overline{\mathcal{A}}$ is given by

$$\epsilon(\gamma) := \sup_{(\delta,\Delta)\in\overline{\mathcal{A}}} \left|\tilde{\mathcal{L}}_{NT}(\theta_0 + \delta, \Pi_0 + \Delta) - \tilde{\mathcal{L}}_{NT}(\theta_0, \Pi_0)\right|$$

where

$$\overline{\mathcal{A}} = \left\{(\delta, \Delta) \in \mathbb{R}^p \times \mathbb{R}^{N\times T} : (\theta_0 + \delta, \Pi_0 + \Delta) \in \Theta \times \Phi^{N\times T}\right\} \cap \left\{\rho^2(\delta, \Delta) \leq \gamma^2\right\}$$

The following auxiliary lemma is used to establish Proposition 1.

Lemma B.2. *Under Assumptions 1 and 2,* $\mathrm{P}(\epsilon(\gamma) \geq \kappa) = \zeta_{2,NT}$ *for any* $\kappa > 0$*, where*

$\zeta_{2,NT}$ is defined in (B.3). Furthermore, with probability at least $1-\zeta_{3,NT}$,

$$\epsilon(\gamma) \leq \frac{\psi_{NT}\sqrt{c_T}}{\sqrt{N \wedge d_T}}\gamma$$

where $\zeta_{3,NT}$ is defined in *(B.4) and is independent of γ. Moreover, $\zeta_{2,NT}$* and *$\zeta_{3,NT}$ approach zero when $(N,T) \to \infty$.*

Proof of Lemma B.2. Define the event $\Omega_1 = \left\{\max_{i\in[N],t\in[T]} \mathbb{E}\left[L\left(W_{it}\right)^4\right] \leq C_L^4\right\}$, where $C_L < \infty$ under Assumption 2. On Ω_1, $L(W_{it})$ has a uniform fourth-moment bound for all (i,t). We want to bound the empirical risk function $\epsilon(\gamma)$. To do so, we split the proof into five steps.

Step 1: To begin with, write $\tilde{\theta} = \theta_0 + \delta$ and $\tilde{\pi}_{it} = \pi_{0,it} + \Delta_{it}$. For any $\kappa > 0$, we have

$$\begin{aligned}
\mathrm{P}(\epsilon(\gamma)\sqrt{NT} \geq \kappa) \leq & 2\mathrm{P}\left(\sup_{(\delta,\Delta)\in\overline{\mathcal{A}}} \frac{1}{\sqrt{Nd_T}} \left|\sum_{i=1}^{N}\sum_{l=1}^{d_T}\sum_{t\in H_l} \frac{Z_{it}^*(\delta,\Delta_{it})}{\sqrt{c_T}}\right| \geq \frac{\kappa}{3}\right) \\
& + \mathrm{P}\left(\sup_{(\delta,\Delta)\in\overline{\mathcal{A}}} \frac{1}{\sqrt{Nd_T}} \left|\sum_{i=1}^{N}\sum_{t\in H_R} \frac{Z_{it}^*(\delta,\Delta_{it})}{\sqrt{c_T}}\right| \geq \frac{\kappa}{3}\right) + 2pNd_T\beta\left(c_T\right) \\
:= & \, 2A_1 + A_2 + 2pNd_T\beta\left(c_T\right)
\end{aligned} \tag{B.1}$$

where we define $Z_{it}^*(\delta,\Delta_{it})$ as

$$Z_{it}^*(\delta,\Delta_{it}) = \ell\left(W_{it}^*;\tilde{\theta},\tilde{\pi}_{it}\right) - \ell\left(W_{it}^*;\theta_0,\pi_{0,it}\right) - \mathbb{E}\left[\ell\left(W_{it}^*;\tilde{\theta},\tilde{\pi}_{it}\right) - \ell\left(W_{it}^*;\theta_0,\pi_{0,it}\right)\right]$$

Here, we use Berbee's coupling lemma (see Lemma 2.1 in Berbee (1987). Note that the lemma is also valid for non-stationary sequences. Next, we proceed to bound A_1 and A_2 in (B.1).

Before bounding A_1, note that $\|\Delta\|_F \leq \|\Delta\|_*$. Then, by Assumption 2, we have the following bound on the variance of the process,

$$\begin{aligned}
& \sup_{(\delta,\Delta)\in\overline{\mathcal{A}}} \mathbb{E}\left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{\ell\left(W_{it}^*;\tilde{\theta},\tilde{\pi}_{it}\right) - \ell\left(W_{it}^*;\theta_0,\pi_{0,it}\right)\right\}^2\right] \\
\leq & \sup_{(\delta,\Delta)\in\overline{\mathcal{A}}} \mathbb{E}\left[\frac{1}{NT}\sum_{i=1}^{}\sum_{t=1}^{}\left\{L\left(W_{it}\right)\left[\|\theta_1-\theta_2\| + |\pi_1-\pi_2|\right]\right\}^2\right] \leq \tilde{C}_L^2\gamma^2
\end{aligned}$$

where $\tilde{C}_L = \sqrt{2}C_L$, and $\max_{i\in[N],t\in[T]} \mathbb{E}\left[L(W_{it})^4\right] \leq C_L^4$.

For any fixed δ and Δ with $\|\delta\|^2 + \frac{1}{NT}\|\Delta\|_*^2 \leq \gamma^2$, we have the following bound on the

variance of the process,

$$\sup_{(\delta,\Delta)\in\overline{\mathcal{A}}} \text{Var}\left(\sum_{i=1}^{N}\sum_{l=1}^{d_T}\sum_{t\in H_l} Z_{it}^{*}\left(\delta,\Delta_{it}\right)\right) \leq NT\tilde{C}_L^2\gamma^2$$

Let $\{\varepsilon_{i,l}\}_{i\in[N],l\in[d_T]}$ be i.i.d Rademacher variables independent of the data. Given that $\mathbb{E}\left[Z_{it}^{*}\left(\delta,\Delta_{it}\right)\right]=0$ by construction, we apply the symmetrization lemma as in Lemma 2.3.7 in van der Vaart and Wellner (1996),

$$\begin{aligned}
&\text{P}\left(\sup_{(\delta,\Delta)\in\overline{\mathcal{A}}}\frac{1}{\sqrt{Nd_T}}\left|\sum_{i=1}^{N}\sum_{l=1}^{d_T}\sum_{t\in H_l}\frac{Z_{it}^{*}(\delta,\Delta_{it})}{\sqrt{c_T}}\right|\geq\kappa\right)\\
&\leq\frac{\text{P}\left(A^0(\gamma)\geq\frac{\kappa}{4}\mid\Omega_1\right)+P\left(\Omega_1^c\right)}{1-\frac{4}{NT\kappa^2}\sup_{(\delta,\Delta)\in\overline{\mathcal{A}}}\text{Var}\left(\sum_{i=1}^{N}\sum_{l=1}^{d_T}\sum_{t\in H_l}Z_{it}^{*}(\delta,\Delta_{it})\right)}\\
&\leq\frac{\text{P}\left(A^0(\gamma)\geq\frac{\kappa}{4}\mid\Omega_1\right)+P\left(\Omega_1^c\right)}{1-4\tilde{C}_L^2\gamma^2/\kappa^2},
\end{aligned}$$

where

$$A^0(\gamma):=\sup_{(\delta,\Delta)\in\overline{\mathcal{A}}}\left|\frac{1}{\sqrt{Nd_T}}\sum_{i=1}^{N}\sum_{l=1}^{d_T}\varepsilon_{i,l}\left(\sum_{t\in H_l}\frac{\ell\left(W_{it}^{*};\tilde{\theta},\tilde{\pi}_{it}\right)-\ell\left(W_{it}^{*};\theta_0,\pi_{0,it}\right)}{\sqrt{c_T}}\right)\right|$$

Next, we define $B_1^0(\gamma)$ and $B_2^0(\gamma)$ as:

$$B_1^0(\gamma)=\sup_{(\delta,\Delta)\in\overline{\mathcal{A}}}\left|\frac{1}{\sqrt{Nd_T}}\sum_{i=1}^{N}\sum_{l=1}^{d_T}\varepsilon_{i,l}\left(\sum_{t\in H_l}\frac{L\left(W_{it}^{*}\right)\|\delta\|}{\sqrt{c_T}}\right)\right|$$

and

$$B_2^0(\gamma)=\sup_{(\delta,\Delta)\in\overline{\mathcal{A}}}\left|\frac{1}{\sqrt{Nd_T}}\sum_{i=1}^{N}\sum_{l=1}^{d_T}\varepsilon_{i,l}\left(\sum_{t\in H_l}\frac{L\left(W_{it}^{*}\right)\Delta_{it}}{\sqrt{c_T}}\right)\right|$$

By Assumption 2 and union bound we obtain that

$$\text{P}\left(A^0(\gamma)\geq\kappa\mid\Omega_1\right)\leq\text{P}\left(B_1^0(\gamma)\geq\kappa\mid\Omega_1\right)+\text{P}\left(B_2^0(\gamma)\geq\kappa\mid\Omega_1\right) \tag{B.2}$$

It remains to bound two terms in the RHS of (B.2).

Step 2: We first consider the bound on $B_1^0(\gamma)$. The argument is similar to the proof

of Lemma D.3 in Belloni et al. (2018). Note that

$$B_1^0(\gamma) \leq c_T \max_{m\in[c_T]} \sup_{(\delta,\Delta)\in\overline{\mathcal{A}}} \left| \frac{1}{\sqrt{Nd_T}} \sum_{i=1}^{N} \sum_{l=0}^{d_T-1} \varepsilon_{i,l} \frac{L\left(W^*_{i,2lc_T+m}\right) \|\delta\|}{\sqrt{c_T}} \right|$$
$$= \max_{m\in[c_T]} \sup_{(\delta,\Delta)\in\overline{\mathcal{A}}} \left| \frac{\sqrt{c_T}}{\sqrt{Nd_T}} \sum_{i=1}^{N} \sum_{l=0}^{d_T-1} \varepsilon_{i,l} L\left(W^*_{i,2lc_T+m}\right) \|\delta\| \right|$$

Applying Markov's inequality and the union bound, we get

$$\begin{aligned}
&\mathrm{P}\left(B_1^0(\gamma) \geq \kappa\right) \\
&\leq c_T \max_{m\in[c_T]} \mathrm{P}\left(\sup_{(\delta,\Delta)\in\overline{\mathcal{A}}} \left| \frac{\sqrt{c_T}}{\sqrt{Nd_T}} \sum_{i=1}^{N} \sum_{l=0}^{d_T-1} \varepsilon_{i,l} L\left(W^*_{i,2lc_T+m}\right) \|\delta\| \right| \geq \kappa \right) \\
&\leq c_T \max_{m\in[c_T]} \inf_{\tau>0} \exp\left(-\tau\kappa\right) \mathbb{E}\left[\exp\left(\tau \sup_{(\delta,\Delta)\in\overline{\mathcal{A}}} \left| \frac{\sqrt{c_T}}{\sqrt{Nd_T}} \sum_{i=1}^{N} \sum_{l=0}^{d_T-1} \varepsilon_{i,l} L\left(W^*_{i,2lc_T+m}\right) \|\delta\| \right| \right) \right]
\end{aligned}$$

Now, fix $m \in [c_T]$. In this step, we will condition throughout on $\{W^*_{it}\}$ but omit explicit notation for this conditioning to keep the notation lighter. We have

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\tau \sup_{(\delta,\Delta)\in\overline{\mathcal{A}}} \left| \frac{\sqrt{c_T}}{\sqrt{Nd_T}} \sum_{i=1}^{N} \sum_{l=0}^{d_T-1} \varepsilon_{i,l} L\left(W^*_{i,2lc_T+m}\right) \|\delta\| \right| \right) \right] \\
&\leq 2\mathbb{E}\left[\exp\left(\frac{\tau\gamma\sqrt{c_T}}{\sqrt{Nd_T}} \sum_{i=1}^{N} \sum_{l=0}^{d_T-1} \varepsilon_{i,l} L\left(W^*_{i,2lc_T+m}\right) \right) \right] \\
&\leq 2\exp\left(\tau^2 c_T \gamma^2 C_L^2\right)
\end{aligned}$$

The first inequality follows from the definition of $\overline{\mathcal{A}}$ and the last inequality follows from the law of iterated expectations combined with both Hoeffding inequality and Chebyshev's inequality. Hence,

$$\begin{aligned}
&\mathrm{P}\left(B_1^0(\gamma) \geq \kappa\right) \\
&\leq 2c_T \inf_{\tau>0} \exp\left(-\tau\kappa\right) \exp\left(\tau^2 c_T \gamma^2 C_L^2\right) \\
&\leq 2c_T \exp\left(-\frac{\kappa^2}{4c_T\gamma^2 C_L^2} \right)
\end{aligned}$$

Step 3: Here, we bound $B_2^0(\gamma)$. First, by the definition of $\overline{\mathcal{A}}$, we have $\sup_{(\delta,\Delta)\in\overline{\mathcal{A}}} \|\Delta\|_* \leq \sqrt{NT}\gamma$ and we define $\zeta_{1,NT} = \sqrt{NT}$.

Notice that

$$B_2^0(\gamma) \leq c_T \max_{m\in[c_T]} \sup_{(\delta,\Delta)\in\overline{\mathcal{A}}} \left| \frac{1}{\sqrt{Nd_T}} \sum_{i=1}^{N} \sum_{l=0}^{d_T-1} \varepsilon_{i,l} \frac{L\left(W^*_{i,2lc_T+m}\right)\Delta_{i,2lc_T+m}}{\sqrt{c_T}} \right|$$
$$= \max_{m\in[c_T]} \sup_{(\delta,\Delta)\in\overline{\mathcal{A}}} \left| \frac{\sqrt{c_T}}{\sqrt{Nd_T}} \sum_{i=1}^{N} \sum_{l=0}^{d_T-1} \varepsilon_{i,l} L\left(W^*_{i,2lc_T+m}\right)\Delta_{i,2lc_T+m} \right|$$

We proceed to bound the moment generating function of $B_2^0(\gamma)$ and use that to obtain an upper bound on $B_2^0(\gamma)$. Now fix $m \in [c_T]$,

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\tau \sup_{(\delta,\Delta)\in\overline{\mathcal{A}}} \left| \frac{\sqrt{c_T}}{\sqrt{Nd_T}} \sum_{i=1}^{N} \sum_{l=0}^{d_T-1} \varepsilon_{i,l} L\left(W^*_{i,2lc_T+m}\right)\Delta_{i,2lc_T+m} \right|\right)\right] \\
&\leq \mathbb{E}\left[\sup_{(\delta,\Delta)\in\overline{\mathcal{A}}} \exp\left(\tau \frac{\sqrt{c_T}}{\sqrt{Nd_T}} \sum_{i=1}^{N} \sum_{l=0}^{d_T-1} C_L \left\| \left(\varepsilon_{il} L\left(W^*_{i,2lc_T+m}\right)\right)_{i,l} \right\| \|\Delta\|_*\right)\right] \\
&\leq \mathbb{E}\left[\sup_{(\delta,\Delta)\in\overline{\mathcal{A}}} \exp\left(\tau \frac{\sqrt{c_T}}{\sqrt{Nd_T}} C_L \mathbb{E}\left\| \left(\varepsilon_{il} L\left(W^*_{i,2lc_T+m}\right)\right)_{i,l} \right\| \|\Delta\|_*\right)\right. \\
&\left. \cdot \sup_{(\delta,\Delta)\in\overline{\mathcal{A}}} \exp\left(\tau \frac{\sqrt{c_T}}{\sqrt{Nd_T}} C_L \left(\left\| \left(\varepsilon_{il} L\left(W^*_{i,2lc_T+m}\right)\right)_{i,l} \right\| - \mathbb{E}\left\| \left(\varepsilon_{il} L\left(W^*_{i,2lc_T+m}\right)\right)_{i,l} \right\|\right) \|\Delta\|_*\right)\right] \\
&\leq \exp\left(\overline{c}\tau \frac{\sqrt{c_T}}{\sqrt{Nd_T}} C_L \zeta_{1,NT}\left(\sqrt{N}\vee\sqrt{d_T}\right)\gamma\right) \exp\left(\overline{c}\tau^2 \frac{c_T}{Nd_T} C_L^2 \zeta_{1,NT}^2 \gamma^2\right)
\end{aligned}$$

for a positive constant $\overline{c} > 0$. The first inequality holds due to the duality between the spectral and nuclear norms, while the second inequality follows from the triangle inequality. The two upper bounds in the final inequality are derived from Theorem 3.4 in Chatterjee (2015) and Theorem 1.2 in Vu (2007), respectively.

Therefore, by Markov inequality, we have

$$\begin{aligned}
&\mathrm{P}\left(B_2^0(\gamma) \geq \kappa \mid \Omega_1\right) \\
&\leq c_T \max_{m\in[c_T]} \mathrm{P}\left(\sup_{(\delta,\Delta)\in\overline{\mathcal{A}}} \left| \frac{\sqrt{c_T}}{\sqrt{Nd_T}} \sum_{i=1}^{N} \sum_{l=0}^{d_T-1} \varepsilon_{i,l} L\left(W^*_{i,2lc_T+m}\right)\Delta_{di,2lc_T+m} \right| \geq \kappa\right) \\
&\leq c_T \inf_{\tau>0} \exp\left(-\tau\kappa\right) \exp\left(\overline{c}\tau \frac{\sqrt{c_T}}{\sqrt{Nd_T}} C_L \zeta_{1,NT}\gamma\left(\sqrt{N}\vee\sqrt{d_T}\right)\right) \exp\left(\overline{c}\tau^2 \frac{c_T}{Nd_T} C_L^2 \zeta_{1,NT}^2 \gamma^2\right) \\
&\leq \overline{c} c_T \exp\left(-\frac{\sqrt{N\wedge d_T}\kappa}{\sqrt{c_T} C_L \zeta_{1,NT}\gamma}\right)
\end{aligned}$$

for a positive constant $\overline{c} > 0$. Similarly, we repeat the arguments above to bound A_2 in

(B.1).

Step 4: To prove the first part of Lemma B.2, note that for any positive γ and $\kappa > 0$, we have

$$\begin{aligned}
&\mathrm{P}(\epsilon(\gamma) \geq \kappa) \\
&\quad \leq 3\frac{\overline{c}\exp\left(-\frac{NT\kappa^2}{c_T C_L^2 \gamma^2} + \ln\left(pc_T\right)\right) + \overline{c}\exp\left(-\frac{\sqrt{N \wedge d_T}\kappa}{\sqrt{c_T}C_L\gamma} + \log\left(c_T\right)\right)}{1 - \overline{c}\tilde{C}_L^2\gamma^2/\left(NT\kappa^2\right)} \\
&\quad + 2pNd_T\beta\left(c_T\right) \coloneqq \zeta_{2,NT}
\end{aligned} \tag{B.3}$$

Step 5: To prove the second part of Lemma B.2, we set

$$\kappa = \frac{\psi_{NT}\sqrt{c_T}\zeta_{1,NT}}{\sqrt{N \wedge d_T}}\gamma$$

From this, we obtain that

$$\begin{aligned}
&\mathrm{P}(\epsilon(\gamma)\sqrt{NT} \geq \kappa) \\
&\quad \leq 3\frac{\overline{c}\exp\left(-\psi_{NT}^2 c_T\left(N \vee d_T\right) + \log\left(pc_T\right)\right) + \overline{c}\exp\left(-\psi_{NT} + \log\left(c_T\right)\right)}{1 - \overline{c}\left[\psi_{NT}^2 c_T\left(N \vee T\right)\right]^{-1}} \\
&\quad + 2pNd_T\beta\left(c_T\right) \coloneqq \zeta_{3,NT}
\end{aligned} \tag{B.4}$$

For fixed $\kappa > 0$, as $(N,T) \to \infty$, (B.4) converges to zero by the definition of ψ_{NT}. Therefore,

$$\epsilon(\gamma) \leq \kappa/\sqrt{NT} = \frac{\psi_{NT}\sqrt{c_T}}{\sqrt{N \wedge d_T}}\gamma$$

with probability at least $1 - \zeta_{3,NT}$, where $\zeta_{3,NT}$ is independent of γ and $\zeta_{3,NT} \to 0$ as $(N,T) \to \infty$. ■

Before we proceed to prove Proposition 1, we first bound the empirical risk over some set $\mathcal{V}$. Define $|\mathcal{V}|$ as

$$|\mathcal{V}| = \sup_{(\theta,\Pi)\in\mathcal{V}} \rho\left(\theta - \theta_0, \Pi - \Pi_0\right)$$

Proof of Proposition 1. Define $\mathcal{C} = \left\{(\theta,\Pi) \in \mathcal{V} : \rho\left(\theta - \theta_0, \Pi - \Pi_0\right) \leq c_{\varepsilon,NT}\right\}$ and the collection of rings $\mathcal{C}_k = \left\{(\theta,\Pi) \in \mathcal{V} : 2^k \leq \rho\left(\theta - \theta_0, \Pi - \Pi_0\right) \leq 2^{k+1}\right\}, k \in \mathbb{Z}$. We denote

$$A = \sup_{(\theta,\Pi)\in\mathcal{C}} \left|\tilde{\mathcal{L}}_{NT}(\theta,\Pi) - \tilde{\mathcal{L}}_{NT}(\theta_0,\Pi_0)\right|$$
$$A_k = \sup_{(\theta,\Pi)\in\mathcal{C}_k} \left|\tilde{\mathcal{L}}_{NT}(\theta,\Pi) - \tilde{\mathcal{L}}_{NT}(\theta_0,\Pi_0)\right|$$

Define $u_n = \left\lfloor \frac{\log(c_{\varepsilon,NT})}{\log(2)} \right\rfloor$ and $v_n = \left\lceil \frac{\log(|\mathcal{V}|)}{\log(2)} \right\rceil$, assuming $c_{\varepsilon,NT} \leq |\mathcal{V}|$ (otherwise, only A needs to be bounded). Since $\{\rho(\delta,\Delta) \leq c_{\varepsilon,NT}\} \cup \left\{\cup_{k=u_n}^{v_n} A_k\right\}$ covers the set $\mathcal{V}$ and because $\rho(\delta,\Delta) \geq 2^k$ on $\mathcal{C}_k$, it holds that

$$\frac{\left|\tilde{\mathcal{L}}_{NT}(\theta,\Pi) - \tilde{\mathcal{L}}_{NT}(\theta_0,\Pi_0)\right|}{\rho(\theta-\theta_0,\Pi-\Pi_0) \vee c_{\varepsilon,NT}} \leq \left(A\left(c_{\varepsilon,NT}\right)^{-1}\right) \vee \left(\max_{u_n\leq k\leq v_n} A_k 2^{-k}\right)$$

We now examine individually each term on the right-hand side. By using the fact that $\|\Delta\| \leq \|\Delta\|_*$ and setting $\zeta_{1,NT} = \sqrt{NT}$, we invoke Lemma B.2. Hence, we have

$$\mathrm{P}\left(A\left(c_{\varepsilon,NT}\right)^{-1} \geq \psi_{NT} c_{\varepsilon,NT}\right) \leq \zeta_{3,NT}$$

Similarly, it holds that $\mathrm{P}\left(A_k \geq \psi_{NT} c_{\varepsilon,NT} 2^{k+1}\right) \leq \zeta_{3,NT}$. The union bound then yields,

$$\mathrm{P}\left(\max_{u_n\leq k\leq v_n} A_k 2^{-k} \geq 2\psi_{NT} c_{\varepsilon,NT}\right) \leq (v_n - u_n + 1)\,\zeta_{3,NT}$$

The union bound and the two inequalities together gives that

$$\begin{aligned}
&\mathrm{P}\left(\sup_{(\theta,\Pi)\in\mathcal{V}} \frac{\left|\tilde{\mathcal{L}}_{NT}(\theta,\Pi) - \tilde{\mathcal{L}}_{NT}(\theta_0,\Pi_0)\right|}{\rho(\theta-\theta_0,\Pi-\Pi_0) \vee c_{\varepsilon,NT}} \geq \psi_{NT} c_{\varepsilon,NT}\right) \\
&\leq (v_n - u_n + 2)\,\zeta_{2,NT} \\
&\leq \left(4 + 2\log_2\left(\sqrt{N \wedge T}\right)\right)\zeta_{2,NT} \\
&\leq 3\frac{\bar{c}\exp\left(-\psi_{nt}^2 (N \vee d_T) + \bar{c}\log(pc_T)\right) + \bar{c}\exp\left(-\psi_{nt} + \bar{c}\log(c_T)\right)}{1 - \bar{c}\left[\psi_{nt}^2 c_T (N \vee T)\right]^{-1}} \\
&+ 2pNT\beta(c_T) \\
&\coloneqq \zeta_{4,NT}
\end{aligned}$$

for some constant $\bar{c} > 0$. To derive the penultimate inequality, we use the fact that $\lceil x \rceil \leq x + 1$ and that $|\mathcal{V}| \leq \left(\sqrt{p} + \sqrt{N \wedge T}\right) c_l'$. Furthermore, without loss of generosity, we assume that the fixed number of coefficients p is less than N and T. Hence, the last inequality stands. As a result, the desired outcome follows, with $\zeta_{4,NT}$ tending to zero as

$(N,T) \to \infty$. ■

Lemma B.3. *For any $N \times T$ matrix Π_0 and Δ, we have the following results*
(i) $\|\Pi_0 + \mathcal{M}(\Delta)\|_* = \|\Pi_0\|_* + \|\mathcal{M}(\Delta)\|_*$;
(ii) $\|\Delta\|_F^2 = \|\mathcal{M}(\Delta)\|_F^2 + \|\mathcal{P}(\Delta)\|_F^2$;
(iii) $\operatorname{rank}(\mathcal{P}(\Delta)) \leq 2\operatorname{rank}(\Pi_0)$;
(iv) $\|\Delta\|_*^2 \leq \|\Delta\|_F^2 \operatorname{rank}(\Delta)$;
(v) For any conformable matrices Δ_1 and Δ_2, $|\operatorname{tr}(\Delta_1 \Delta_2)| \leq \|\Delta_1\| \|\Delta_2\|_*$.

Proof of Lemma B.3. The proof follows from that of Lemma C.2 in Chernozhukov et al. (2019). ■

Proof of Theorem 1. Let

$$\begin{aligned}
\mathcal{A}_1 &= \left\{ \|\delta\|^2 + \frac{1}{NT}\|\Delta\|_F^2 > \gamma_{NT}^2 \vee c'_{\varepsilon,nt} \right\} \\
\mathcal{A}_2 &= \{(\hat{\theta}, \hat{\Pi}) \in \mathcal{V}\} \\
\mathcal{A}_3 &= \left\{ \left| \widetilde{\mathcal{L}}_{NT}(\hat{\theta}, \hat{\Pi}) - \widetilde{\mathcal{L}}_{NT}(\theta_0, \Pi_0) \right| \leq \psi_{NT} c_{\varepsilon,NT} \rho\left(\hat{\theta} - \theta_0, \hat{\Pi} - \Pi_0\right) \right\}
\end{aligned}$$

with $\gamma_{NT} = \frac{2(c_\nu+1)\sqrt{r}}{c_{RSC}} \psi_{NT} c_{\varepsilon,NT}$. We have

$$\begin{aligned}
\mathrm{P}(\mathcal{A}_1) &= P(\mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3) + P(\mathcal{A}_1 \cap (\mathcal{A}_2 \cap \mathcal{A}_3)^c) \\
&= P(\mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3) + P(\mathcal{A}_1 \cap (\mathcal{A}_2^c \cup \mathcal{A}_3^c)) \\
&\leq P(\mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3) + P(\mathcal{A}_1 \cap \mathcal{A}_2^c) + P(\mathcal{A}_1 \cap \mathcal{A}_3^c) \\
&:= P_1 + P_2 + P_3
\end{aligned}$$

where the second inequality is due to the union bound. It holds $P_2 \to 0$ by Lemma 1. Concerning P_3, when conditional on $\mathcal{A}_1 \cap \mathcal{A}_3^c$, it holds that $\rho\left(\hat{\theta} - \theta_0, \hat{\Pi} - \Pi_0\right) \vee c_{\varepsilon,NT} = \rho\left(\hat{\theta} - \theta_0, \hat{\Pi} - \Pi_0\right)$. Therefore, over $\mathcal{A}_1 \cap \mathcal{A}_3^c$, we find that

$$\left| \tilde{\mathcal{L}}_{NT}(\theta, \Pi) - \tilde{\mathcal{L}}_{NT}(\theta_0, \Pi_0) \right| \geq \psi_{NT} c_{\varepsilon,NT} \left(\rho(\theta - \theta_0, \Pi - \Pi_0) \vee c_{\varepsilon,NT} \right)$$

which holds with probability approaching zero by Lemma B.2. Hence, it suffices to show that $P_1 = 0$. Note that

$$\begin{aligned}
0 &\geq \mathcal{L}_{NT}\left(\hat{\theta}, \hat{\Pi}\right) - \mathcal{L}_{NT}(\theta_0, \Pi_0) + \nu\left(\|\hat{\Pi}\|_* - \|\Pi_0\|_*\right) \\
&\geq \mathcal{E}(\hat{\theta}, \hat{\Pi}) - \tilde{\mathcal{L}}_{NT}(\theta_0, \Pi_0) + \tilde{\mathcal{L}}_{NT}\left(\hat{\theta}, \hat{\Pi}\right) + \nu\left(\|\mathcal{P}(\Delta)\|_* - \|\mathcal{M}(\Delta)\|_*\right)
\end{aligned}$$

where the first inequality holds by the definition of $\mathcal{L}_{NT}\left(\hat{\theta}, \hat{\Pi}\right)$ and the second inequality holds by Lemma B.3. Specifically,

$$\begin{aligned}\|\hat{\Pi}\|_* - \|\Pi_0\|_* =& \|\Pi_0 + \mathcal{M}(\Delta) + \mathcal{P}(\Delta)\|_* - \|\Pi_0\|_* \\ \geq& \|\Pi_0 + \mathcal{M}(\Delta)\|_* - \|\mathcal{P}(\Delta)\|_* - \|\Pi_0\|_* \\ =& \|\mathcal{M}(\Delta)\|_* - \|\mathcal{P}(\Delta)\|_*\end{aligned} \tag{B.5}$$

Next, the definition of $\mathcal{C}$ implies that the empirical risk function is bounded above. Since $\mathcal{E}(\hat{\theta}, \hat{\Pi}) \geq 0$, and $\nu = c_\nu \frac{\psi_{NT} c_{\varepsilon,NT}}{\sqrt{NT}}$, we have

$$\begin{aligned}0 &\leq \psi_{NT} c_{\varepsilon,NT}\left(\|\delta\| + \frac{1}{\sqrt{NT}}\|\Delta\|_*\right) + c_\nu \frac{\psi_{NT} c_{\varepsilon,NT}}{\sqrt{NT}}\left(\|\mathcal{P}(\Delta)\|_* - \|\mathcal{M}(\Delta)\|_*\right) \\ &\leq \sqrt{NT}\|\delta\| + (c_\nu + 1)\|\mathcal{P}(\Delta)\|_* - (c_\nu - 1)\|\mathcal{M}(\Delta)\|_* \\ &\leq \sqrt{NT}\|\delta\| + 2c_\nu\|\mathcal{P}(\Delta)\|_* - (c_\nu - 1)\|\Delta\|_*\end{aligned}$$

Hence, within $\mathcal{B} \cap \mathcal{C}$, we have $(\delta, \Delta) \in \mathcal{A}$, where $\mathcal{A}$ is defined in (4.2). Therefore, by Assumptions 4, the excess risk function is bounded below by

$$\begin{aligned}&\inf_{\substack{(\theta,\Pi)\in\mathcal{A}_2\cap\mathcal{A}_3 \\ \|\delta\|^2 + \frac{1}{NT}\|\Delta\|_F^2 = \gamma_{NT}^2}} \mathcal{E}(\theta, \Pi) \\ &\geq \inf_{\substack{(\theta,\Pi)\in\mathcal{A}_2\cap\mathcal{A}_3 \\ \|\delta\|^2 + \frac{1}{NT}\|\Delta\|_F^2 = \gamma_{NT}^2}} c_l \mathbb{E}\left[\frac{1}{NT}\left\|\sum_{j=1}^{p} X_j \delta_j + \Delta\right\|_F^2\right] \\ &\geq \inf_{\substack{(\delta,\Delta)\in\mathcal{A}_2\cap\mathcal{A}_3 \\ \|\delta\|^2 + \frac{1}{NT}\|\Delta\|_F^2 = \gamma_{NT}^2}} c_l c_{RSC}\left[\|\delta\|^2 + \frac{1}{NT}\|\Delta\|_F^2\right] \\ &= s c_{RSC} \gamma_{NT}^2\end{aligned}$$

Summing these up, over the event $\mathcal{A}_2 \cap \mathcal{A}_3$, we have

$$\begin{aligned}
c_{RSC}\gamma_{NT}^2 &\leq \psi_{NT}c_{\varepsilon,NT}\left(\|\delta\| + \frac{1}{\sqrt{NT}}\|\Delta\|_*\right) + c_\nu\frac{\psi_{NT}c_{\varepsilon,NT}}{\sqrt{NT}}\left(\|\mathcal{P}(\Delta)\|_* - \|\mathcal{M}(\Delta)\|_*\right) \\
&\leq \psi_{NT}c_{\varepsilon,NT}\|\delta\| + \frac{\psi_{NT}c_{\varepsilon,NT}}{\sqrt{NT}}\left\{(c_\nu+1)\|\mathcal{P}(\Delta)\|_* - (c_\nu-1)\|\mathcal{M}(\Delta)\|_*\right\} \\
&\leq \psi_{NT}c_{\varepsilon,NT}\|\delta\| + \frac{\psi_{NT}c_{\varepsilon,NT}}{\sqrt{NT}}(c_\nu+1)\|\mathcal{P}(\Delta)\|_* \\
&\leq \psi_{NT}c_{\varepsilon,NT}\left(\|\delta\| + \frac{c_\nu+1}{\sqrt{NT}}\sqrt{2r}\|\Delta\|_F\right) \\
&\leq (c_\nu+1)\sqrt{2r}\psi_{NT}c_{\varepsilon,NT}\left(\|\delta\| + \frac{1}{\sqrt{NT}}\|\Delta\|_F\right) \\
&\leq (c_\nu+1)\sqrt{4r}\psi_{NT}c_{\varepsilon,NT}\gamma_{NT}
\end{aligned}$$

The first inequality follows from (B.5). The third inequality holds because $\|\mathcal{M}(\Delta)\|_*$ is positive and $c_\nu \geq 2$. The fourth inequality is derived from Lemma B.3(iii) and (iv). Consequently, $\gamma_{NT} \leq \frac{(c_\nu+1)\sqrt{4r}}{c_{RSC}}\psi_{NT}c_{\varepsilon,NT}$, leading to $P_1 = 0$ w.p.a.1. ■

Corollary 1 presents two key results. First, it establishes the consistency of our rank estimator $\hat{r}$. Second, it outlines the convergence rates of $\hat{\Lambda}$ and $\hat{F}$, which form the foundation for the iterative localized estimation in the second step. The proof for the second part is analogous to that of Lemma 2 in Chen, Dolado, and Gonzalo (2021), but we provide it here for the sake of completeness.

Proof of Corollary 1. Denote $\gamma_{NT} = \psi_{NT}\sqrt{\frac{rc_T}{N\wedge d_T}}$ and thus from Theorem 1, we have $\frac{1}{\sqrt{NT}}\left\|\widehat{\Pi} - \Pi_0\right\|_F = O_p(\gamma_{NT})$. Let $\sigma_1 > \cdots > \sigma_r > 0$ denote the nonzero singular values of Π_0. Note that Assumption 5 implies that here exist constant $c > 0$ such that, with probability approaching one, for all $s \leq r$: $c^{-1}\sqrt{NT} \leq \sigma_s \leq c\sqrt{NT}$, and $\sigma_{s-1}^2 - \sigma_s^2 \geq c\sqrt{NT}$.

Part (i): To show that our rank estimator $\hat{r}$ is consistent, we first use Weyl's inequality to obtain the error bound of singular values. That is, w.p.a.1,

$$\max_{s\in\{1,\ldots,N\wedge T\}}\left\{\left|\sigma_s\left(\hat{\Pi}\right) - \sigma_s\right|\right\} \leq \left\|\hat{\Pi} - \Pi_0\right\| \leq \left\|\hat{\Pi} - \Pi_0\right\|_F$$

Based on the results of Theorem 1 and the assumption that $\sigma_{r+1}(\Pi_0) = \cdots = \sigma_{N\wedge T}(\Pi_0) = 0$ (i.e., the true rank is r),

$$\max\left\{\max_{s\leq r}\left|\sigma_s\left(\hat{\Pi}\right) - \sigma_s(\Pi_0)\right|, \max_{r+1\leq s\leq N\wedge T}\left|\sigma_s\left(\hat{\Pi}\right)\right|\right\} \leq \sqrt{NT}\gamma_{NT}$$

By Assumption 5, we have $\sigma_r \asymp \sqrt{NT}c_r$. Thus,

$$\min_{s\leq r} \sigma_s\left(\hat{\Pi}\right) \geq (c_r - \gamma_{NT})\sqrt{NT} + \gamma_{NT}\sqrt{NT} \gtrsim NT\nu$$
$$\max_{r+1\leq s\leq N\wedge T} \sigma_s\left(\hat{\Pi}\right) \leq \gamma_{NT} = o_P(NT\nu)$$

This proves the consistency of $\hat{r}$.

Part (ii): To simplify notation, we omit the hat notation for the estimator and instead define $\frac{1}{\sqrt{NT}}\|\Pi - \Pi_0\|_F = \gamma_{NT}$, where $\Pi = \Lambda F'$, with Λ and F normalized such that $F'F/T = \mathbb{I}_r$ and $\Lambda'\Lambda/N$ is diagonal with non-increasing diagonal elements. Under this normalization, Λ and F are uniquely identified.

First, let $U \in \mathbb{R}^{r\times r}$ be a diagonal matrix whose diagonal elements are either 1 or -1. Since $F'F/T = F_0'F_0/T = \mathbb{I}_r$ and $\|\Lambda_0\|/\sqrt{N} \leq \bar{c}$ by Assumption refassu:factor, we have

$$\begin{aligned}\|\Lambda - \Lambda_0 U\|_F/\sqrt{N} &= \|(\Lambda - \Lambda_0 U)F'\|_F/\sqrt{NT} = \|\Lambda F' - \Lambda_0 F_0' + \Lambda_0 F_0' - \Lambda_0 U F'\|_F/\sqrt{NT} \\ &\leq \|\Lambda F' - \Lambda_0 F_0'\|_F/\sqrt{NT} + \|\Lambda_0\|_F/\sqrt{N}\cdot\|F - F_0 U\|_F/\sqrt{T} \\ &\leq \gamma_{NT} + \bar{c}\,\|F - F_0 U\|/\sqrt{T}.\end{aligned}$$

Thus,

$$\|\Lambda - \Lambda_0 U\|_F/\sqrt{N} + \|F - F_0 U\|_F/\sqrt{T} \leq \gamma_{NT} + (1+\bar{c})\|F - F_0 U\|_F/\sqrt{T}$$

Second,

$$\begin{aligned}\|F - F_0 U\|_F/\sqrt{T} &= \|F_0 U - F(F'F_0 U/T) + F(F'F_0 U/T) - F\|_F/\sqrt{T} \\ &\leq \|F_0 U - F(F'F_0 U/T)\|_F/\sqrt{T} + \|F(F'F_0 U/T) - F\|_F/\sqrt{T} \\ &= \|M_F F_0\|_F/\sqrt{T} + \|F'F_0/T - \mathrm{U}\|_F\end{aligned}$$

where $P_A = A(A'A)^{-1}A'$ and $M_A = \mathbb{I} - P_A$.

Third, we have

$$\begin{aligned}\frac{1}{\sqrt{NT}}\|(\Lambda F' - \Lambda_0 F_0')M_F\|_F &\leq \sqrt{\operatorname{rank}\left[(\Lambda F' - \Lambda_0 F_0')M_F\right]}\cdot\|M_F\|\cdot\|\Lambda F' - \Lambda_0 F_0'\|/\sqrt{NT} \\ &\lesssim \|\Lambda F' - \Lambda_0 F_0'\|_F/\sqrt{NT} = \gamma_{NT} \qquad \text{(B.6)}\end{aligned}$$

and since

$$\begin{aligned}\left\|\left(\Lambda F' - \Lambda_0 F_0'\right) M_F\right\|_F / \sqrt{NT} &= \left\|\Lambda_0 F_0' M_F\right\|_F / \sqrt{NT} \\ &= \sqrt{\mathrm{Tr}\left[\left(\Lambda_0'\Lambda_0/N\right) \cdot \left(F_0' M_F F_0/T\right)\right]} \\ &\geq \sqrt{\sigma_{Nr}} \sqrt{\mathrm{Tr}\left(F_0' M_F F_0/T\right)} \\ &= \sqrt{\sigma_{Nr}} \left\|M_F F_0\right\|_F / \sqrt{T} \end{aligned} \tag{B.7}$$

It follows from (B.6) and (B.7) that

$$\left\|M_F F_0\right\|_F / \sqrt{T} \lesssim \sqrt{\frac{1}{\sigma_{Nr}}} \gamma_{NT} \tag{B.8}$$

Similarly, it can be shown that

$$\left\|M_{F_0} F\right\|_F / \sqrt{T} \lesssim \sqrt{\frac{1}{\mu_{\min}\left(\Lambda'\Lambda/N\right)}} \gamma_{NT} \tag{B.9}$$

Fourth, we have

$$\frac{1}{\sqrt{NT}} \left\|\left(\Lambda F' - \Lambda_0 F_0'\right) P_F\right\|_F \leq \frac{1}{\sqrt{NT}} \left\|\Lambda F' - \Lambda_0 F_0'\right\|_F \cdot \left\|P_F\right\|_F = \sqrt{r}\gamma_{NT}$$

so

$$\begin{aligned}&\frac{1}{\sqrt{NT}} \left\|\left(\Lambda F' - \Lambda_0 F_0'\right) P_F\right\|_F \\ &= \frac{1}{\sqrt{NT}} \left\|\Lambda F' - \Lambda_0 \left(F_0' F/T\right) F'\right\|_F \\ &= \frac{1}{\sqrt{N}} \left\|\Lambda - \Lambda_0 \left(F_0' F/T\right)\right\|_F \leq \sqrt{r}\gamma_{NT}\end{aligned} \tag{B.10}$$

Likewise, it can be shown that

$$\frac{1}{\sqrt{N}} \left\|\Lambda_0 - \Lambda\left(F' F_0/T\right)\right\|_F \leq \sqrt{r}\gamma_{NT} \tag{B.11}$$

Fifth, define $R_T = F' F_0/T$. Note that $F R_T = F F' F_0/T = P_F F_0$, thus

$$\begin{aligned}\mathbb{I}_r = F_0' F_0/T &= R_T' \left(F' F/T\right) R_T + F_0' F_0/T - R_T' \left(F' F/T\right) R_T \\ &= R_T' R_T + F_0' F_0/T - F_0' F R_T/T + F_0' F R_T/T - R_T' \left(F' F/T\right) R_T \\ &= R_T' R_T + F_0' \left(F_0 - F R_T\right)/T = R_T' R_T + F_0' M_F F_0/T\end{aligned} \tag{B.12}$$

because

$$F_0'FR_T/T - R_T'\left(F'F/T\right)R_T = \left(F_0 - FR_T\right)'FR_T/T = F_0'M_FFR_T/T = 0$$

In addition,

$$\begin{aligned}
&\Lambda_0'\Lambda_0/N \\
&\quad = R_T'\left(\Lambda'\Lambda/N\right)R_T + \left(\Lambda_0'\Lambda_0/N - R_T'\left(\Lambda'\Lambda/N\right)R_T\right) \\
&\quad = R_T'\left(\Lambda'\Lambda/N\right)R_T + \Lambda_0'\left(\Lambda_0 - \Lambda R_T\right)/N + \left(\Lambda_0 - \Lambda R_T\right)'\Lambda R_T/N
\end{aligned} \tag{B.13}$$

Similar to the proof of (B.12), we have

$$\mathbb{I}_r = R_TR_T' + F'\left(F - F_0R_T'\right)/T = R_TR_T' + F'M_{F_0}F/T \tag{B.14}$$

From (B.13), it holds that

$$\begin{aligned}
\Lambda_0'\Lambda_0/N =& R_T'\left(\Lambda'\Lambda/N\right)\left(R_T'\right)^{-1}R_T'R_T + \Lambda_0'\left(\Lambda_0 - \Lambda R_T\right)/N + \left(\Lambda_0 - \Lambda R_T\right)'\Lambda R_T/N \\
=& R_T'\left(\Lambda'\Lambda/N\right)\left(R_T'\right)^{-1} + R_T'\left(\Lambda'\Lambda/N\right)\left(R_T'\right)^{-1}\left(R_T'R_T - \mathbb{I}_r\right) \\
&+ \Lambda_0'\left(\Lambda_0 - \Lambda R_T\right)/N + \left(\Lambda_0 - \Lambda R_T\right)'\Lambda R_T/N
\end{aligned}$$

and then it follows from the above equation and (B.12) that

$$\left(\Lambda_0'\Lambda_0/N + D_{NT}\right)R_T' = R_T'\left(\Lambda'\Lambda/N\right)$$

where

$$D_{NT} = R_T'\left(\Lambda'\Lambda/N\right)\left(R_T'\right)^{-1}F_0'M_FF_0/T - \Lambda_0'\left(\Lambda_0 - \Lambda R_T\right)/N - \left(\Lambda_0 - \Lambda R_T\right)'\Lambda R_T/N$$

From (B.8), (B.10) and (B.11), we have $\|D_{NT}\|_F \lesssim \gamma_{NT}$. Hence, by the Bauer-Fike theorem, it holds that

$$\left|\sigma_{\min}\left[\Lambda'\Lambda/N\right] - \sigma_{\min}\left[\Lambda_0'\Lambda_0/N\right]\right| \leq \|D_{NT}\| \leq \|D_{NT}\|_F \lesssim \gamma_{NT} \tag{B.15}$$

Moreover, by Assumption 5 and the perturbation theory for eigenvectors,

$$\|R_T'V_T\ \mathrm{S} - \mathbb{I}_r\|_F = \|R_T'V_T - \mathrm{S}\|_F \lesssim d\left(\theta, \theta_0\right)$$

where $V_T = \mathrm{diag}\left(\left(R_{T,1}R_{T,1}'\right)^{-1/2}, \ldots, \left(R_{T,r}R_{T,r}'\right)^{-1/2}\right)$, and $R_{T,j}'$ is the jth column of R_T'.

Furthermore, (B.9) and (B.15) imply that $\sigma_{\min}[\Lambda'\Lambda/N]$ is bounded below by a positive constant, and that $\|M_{F_0}F\|_F/\sqrt{T} \lesssim \gamma_{NT}$. From (B.14), we then have

$$\|V_T - \mathbb{I}_r\|_F \lesssim \|R_T R_T' - \mathbb{I}_r\|_F = \|M_{F_0}F\|_F^2/T \lesssim \gamma_{NT}^2$$

Note that the triangular inequality implies that

$$\|R_T' - \mathrm{S}\|_F \leq \|R_T'V_T - \mathrm{S}\|_F + \|R_T'V_T - R_T\|_F \leq \|R_T'V_T - \mathrm{S}\|_F + \|R_T\|_F \cdot \|V_T - \mathbb{I}_r\|_F$$

Thus,

$$\|F'F_0/T - \mathrm{S}\|_F = \|R_T - \mathrm{S}\|_F \lesssim \gamma_{NT}$$

Finally, setting $U = S$, we then have $\|F - F_0\mathrm{S}\|_F/\sqrt{T} \lesssim \gamma_{NT}$ and similarly $\|\Lambda - \Lambda_0\mathrm{S}\|/\sqrt{N} \lesssim \gamma_{NT}$. ■

B.3 Discussion of Assumptions in Single-Index Models

In this section, we specialize to the case of single-index models and introduce two auxiliary assumptions that together lead to the Restricted Strong Convexity condition in Assumption 4. We further provide primitive, low-level conditions that verify one of these assumptions.

We first recall that the single-index specification assumes that the conditional distribution of Y_{it} given X_{it} depends only on the scalar index $X_{it}'\theta + \pi_{it}$. The common parameter θ captures the marginal effect of observed covariates, while π_{it} accounts for latent, time-varying or individual-specific unobserved heterogeneity. Formally, the loss function can be written as

$$\ell(W_{it}; \theta, \pi_{it}) = \ell\big(Y_{it},\, X_{it}'\theta + \pi_{it}\big),$$

where $W_{it} = (Y_{it}, X_{it})$. In this setting, the population model (2.1) continues to hold, and estimation proceeds via the regularized sample criterion in (3.2).

Restricted Strong Convexity We now introduce two auxiliary assumptions that together imply the high-level Restricted Strong Convexity (RSC) condition in Assumption 4. These conditions are formulated to separate local strong convexity of the objective function from separability of Δ from the covariate X.

Assumption B.1. *There exist constants c_l and c_l' such that for all (N,T), the inequality*

$\|\delta\|^2 + \frac{1}{NT}\|\Delta\|_F^2 \leq c_l'$, *where* δ *is a* p*-dimensional vector and* Δ *is an* $N \times T$ *matrix, implies*

$$\mathcal{E}\left(\theta_0 + \delta, \Pi_0 + \Delta\right) \geq c_l \mathbb{E}\left[\frac{1}{NT}\left\|\sum_{j=1}^{p} X_j \delta_j + \Delta\right\|_F^2\right]$$

Assumption B.1 is a local strong convexity condition, similar to those commonly used in the literature on non-convex low-dimensional M-estimators. Specifically, the condition is imposed within an ℓ_2-ball of fixed radius around the true parameters. In the low-dimensional setting, i.e., when $\Delta = 0$, Assumption B.1 simplifies to the requirement that $\mathcal{E}\left(\theta_0 + \delta, 0\right) \geq \underline{c}\|\delta\|^2$ for some constant $\underline{c}$, under the assumption that $\min_{i,t} \mu_{\min}\left(\mathbb{E}\left[X_{it}X_{it}'\right]\right) \geq \bar{c}$. When fixed effects are present and the loss function is strongly convex and smooth (i.e., the second derivative exists and is bounded away from zero), Assumption B.1 is easily satisfied. For more general loss functions, sufficient conditions for Assumption B.1 are provided in Proposition B.2 below.

Assumption B.2. *For all* $(\delta, \Delta) \in \mathcal{A}$*, there exists a universal constant* $c_{RSC} > 0$ *such that*

$$\mathbb{E}\left[\left\|\sum_{j=1}^{p} X_j \delta_j + \Delta\right\|_F^2\right] \geq NT \cdot c_{RSC}'\|\delta\|^2 + c_{RSC}'\|\Delta\|_F^2$$

Intuitively, Assumption B.2 requires that Δ cannot be fully explained by X, and thus is separable. Its verification is challenging without imposing further conditions on parameter space. See also discussions in Moon and Weidner (2019) and Miao, Phillips, and Su (2022) following their Assumption 1 and Assumption 2, respectively.

Proposition B.1. *Suppose Assumptions B.1 and B.2 hold, then Assumption 4 holds.*

Proof. The result follows directly from the definitions of Assumptions B.1 and B.2. ■

Low-Level Sufficient Conditions. We first establish sufficient conditions for Assumption 3(ii), which guarantees identification, and then for Assumption B.1, which ensures local strong convexity of the expected criterion in single-index settings.

With a slight abuse of notation, we define the index function as $Z_{it} = X_{it}'\theta + \pi_{it} \in \mathcal{Z}$, where $\mathcal{Z} = \{z \in \mathbb{R} \mid z = x'\theta + \pi \text{ s.t. } (x, \theta, \pi) \in \mathcal{X} \times \Theta \times \Phi\}$. Define

$$m(z, x) = \mathbb{E}[\ell(Y_{it}, z) \mid X_{it} = x], \quad (z, x) \in \mathcal{Z} \times \mathcal{X},$$

and let $Z_{0,it} = X_{it}'\theta_0 + \pi_{0,it}$ denote the index function evaluated at the true parameter values.

We define the expected score function as $s(z,x) = \frac{\partial}{\partial z} m(z,x)$ and the expected Hessian as $H(z,x) = \frac{\partial^2}{\partial z \partial z'} m(z,x)$. We denote $s_{it} = s(Z_{it}, X_{it})$ and $s_{0,it} = s(Z_{0,it}, X_{it})$. Similarly, we denote $H_{it} = H(Z_{it}, X_{it})$ and $H_{0,it} = H(Z_{0,it}, X_{it})$.

Proposition B.2. *Suppose Assumption 3(ii) holds, along with the following conditions:*
(i) There exists a constant $\bar{c} > 0$ such that $\max_{i,t} \mu_{\max}\left(\mathbb{E}\left[X_{it}X_{it}'\right]\right) \leq \bar{c}$;
(ii) The partial derivatives of $m(z,x)$ with respect to z up to the third order are uniformly bounded in absolute value over $(z,x) \in \mathcal{Z} \times \mathcal{X}$;
(iii) It holds that $\sum_{i=1}^{N}\sum_{t=1}^{T} \mathbb{E}\left[s_{0,it}\left(Z_{it} - Z_{0,it}\right)\right] \geq 0$;
(iv) The expected Hessian $H_{0,it}$ is positive definite and, in fact,

$$\inf_{i\in[N],t\in[T]} \inf_{X_{it}\in\mathcal{X}} \mu_{min}\left(H_{0,it}\right) \geq c_{min}$$

for some constant $c_{min} > 0$.
(v) The following holds,

$$0 < q := \inf_{(\delta,\Delta)\in\mathcal{V}_1, \delta\neq 0} \frac{\left(\mathbb{E}\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(X_{it}'\delta + \Delta_{it}\right)^2\right)\right)^{3/2}}{\mathbb{E}\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left|X_{it}'\delta + \Delta_{it}\right|^3\right)}$$

where $\mathcal{V}_1 = \left\{(\delta,\Delta) \in \mathbb{R}^p \times \mathbb{R}^{N\times T} : \|\delta\|^2 + \frac{1}{NT}\|\Delta\|_F^2 \leq \bar{c}_l\right\}$.
Then Assumption B.1 holds.

Condition (i) requires that the eigenvalues of $\mathbb{E}[X_{it}X_{it}']$ are bounded from above, a standard assumption in high-dimensional models. Condition (iv) ensures the positive definiteness of the expected Hessian matrix at the true parameter values, and Condition (v) imposes a nonlinearity condition as in Belloni, Chen, Madrid Padilla, and Wang (2023) and Belloni and Chernozhukov (2011).

Proof of Proposition B.2. Using a third-order Taylor expansion of $m(Z_{it}, X_{it})$ around $Z_{0,it} = X_{it}'\theta_0 + \pi_{0,it}$, we obtain

$$\begin{aligned}\mathcal{E}(Z) &= \mathbb{E}\left[\frac{1}{NT}\sum_{i,t}\left(m(Z_{it},X_{it}) - m(Z_{0,it},X_{it})\right)\right] \\ &= \frac{1}{NT}\sum_{i,t}\mathbb{E}\Big[s(Z_{0,it},X_{it})(Z_{it}-Z_{0,it}) + \tfrac{1}{2}H(Z_{0,it},X_{it})(Z_{it}-Z_{0,it})^2 + r_{Z,it}\Big],\end{aligned}$$

where the remainder term satisfies

$$|r_{Z,it}| \leq \frac{1}{6} \sup_{z \in \mathcal{Z}} \left| \frac{\partial^3}{\partial z^3} m(z, X_{it}) \right| |Z_{it} - Z_{0,it}|^3 \leq \frac{c_M}{6} |Z_{it} - Z_{0,it}|^3$$

for some constant $c_M > 0$ by Condition (ii).

Since $\mathbb{E}[s(Z_{0,it}, X_{it})] = 0$ and the smallest eigenvalue of the expected Hessian satisfies $\mu_{\min}(H_{0,it}) \geq c_{\min} > 0$ by Condition (iv), it follows that

$$\mathcal{E}(Z) \geq \frac{1}{NT} \sum_{i,t} \mathbb{E}\left[\frac{c_{\min}}{2} (Z_{it} - Z_{0,it})^2 - \frac{c_M}{6} |Z_{it} - Z_{0,it}|^3 \right].$$

By the definition of $\mathcal{V}_1$ and the boundedness of X_{it}, we can choose a constant $c_l' \leq \bar{c}_l$ such that

$$\mathbb{E}\left[\frac{1}{NT} \sum_{i,t} (Z_{it} - Z_{0,it})^2 \right] \leq \left(\frac{3qc_{\min}}{c_M} \right)^2.$$

Substituting this into the previous inequality yields

$$\mathcal{E}(Z) \geq \frac{c_{\min}}{6} \mathbb{E}\left[\frac{1}{NT} \sum_{i,t} (Z_{it} - Z_{0,it})^2 \right],$$

which verifies Assumption B.1 with c_l' as defined above and $c_l = c_{\min}/6$. ■

C Proofs of Inference Results

C.1 Notation and Choice of Norms

In this section, I adopt the notations introduced by Fernández-Val and Weidner (2016) and Chen, Fernández-Val, and Weidner (2021). I extend their asymptotic results for "joint estimators" to the framework of my iterative procedure. We collect all these effects in the $r\,(N+T)$-vector $\phi_{NT} = (\lambda_1, \ldots, \lambda_N, f_1, \ldots, f_T)'$. The model parameter θ is the coefficient of interest, and the vector ϕ_{NT} is treated as a nuisance parameter. The true values of the parameters, denoted by θ_0 and $\phi_0 = (\lambda_{01}', \ldots, \lambda_{0N}', f_{01}', \ldots, f_{0T}')'$, are the solution to the population conditional maximum likelihood.

As discussed in Remark 1, the factors and factor loadings are not separately identified without suitable normalization. In the first-step estimation we adopt a standard normalization commonly used in linear and nonlinear factor models; however, for deriving the asymptotic distribution we employ an alternative normalization, following Chen, Fernández-Val,

and Weidner (2021). Specifically, we impose the restriction that $\sum_{i=1}^{N} \lambda_{0i}\lambda_{0i}' = \sum_{t=1}^{T} f_{0t}f_{0t}'$, and define the restricted parameter set as

$$\Phi := \left\{ \phi \in \mathbb{R}^{d_\phi} : \sum_{i=1}^{N} \lambda_{0i}\lambda_i' = \sum_{t=1}^{T} f_t f_{0t}' \right\}$$

Imposing $\hat{\phi} \in \Phi$ is not feasible in practice because the true parameters appear in the definition of Φ. Nonetheless, all asymptotic results concern the estimator $\hat{\theta}$, which is invariant to the choice of normalization for $\hat{\phi}$. Thus, treating $\hat{\phi} \in \Phi$ as if it were imposed is a technical device used solely for the proofs. With slight abuse of notation, we redefine $L_{NT}(\theta, \Lambda, F)$ for $\mathcal{L}_{NT}(\theta, \Lambda, F)$ in the main context, and introduce the following normalized criterion function

$$\mathcal{L}_{NT}(\theta, \Lambda, F) = L_{NT}(\theta, \Lambda, F) - \frac{b}{2NT} \|\Lambda_0'\Lambda - F'F_0\|_F^2$$

where $b > 0$ for some constant.

To simplify notations, we suppress the dependence on NT of all the sequences of functions and parameters, e.g. we use $\mathcal{L}$ for $\mathcal{L}_{NT}$ and ϕ for ϕ_{NT}. Partial derivatives are denoted with subscripts, e.g. $\partial_\theta \mathcal{L}(\theta, \phi)$ denotes $\partial \mathcal{L}(\theta, \phi)/\partial\theta$ and so on. Let

$$\mathcal{S}(\theta, \phi) = \partial_\phi \mathcal{L}(\theta, \phi), \quad \mathcal{H}(\theta, \phi) = \partial_{\phi\phi'} \mathcal{L}(\theta, \phi)$$

where $\partial_x f$ is the partial derivative of f with respect to x, and additional subscripts denote higher order partial derivatives. We refer to the $\dim \phi$-vector $\mathcal{S}(\theta, \phi)$ as the incidental parameter score, and to the $\dim \phi \times \dim \phi$ matrix $\mathcal{H}(\theta, \phi)$ as the incidental parameter Hessian. We drop arguments when evaluating at the true parameter values (θ_0, ϕ_0), e.g. $\mathcal{H} = \mathcal{H}(\theta_0, \phi_0)$. We use a bar for expectations given ϕ, such as $\partial_\theta \bar{\mathcal{L}} = \mathbb{E}[\partial_\theta \mathcal{L}]$, and a tilde for variables in deviations from expectations, such as $\partial_\theta \tilde{\mathcal{L}} = \partial_\theta \mathcal{L} - \partial_\theta \bar{\mathcal{L}}$.

We follow Fernández-Val and Weidner (2016) (FVM) in using the Euclidean norm $\|\cdot\|$ for vectors, and the norm induced by the Euclidean norm for matrices and tensors, i.e.

$$\|\partial_{\theta\theta}\mathcal{L}(\theta, \phi)\| = \max_{u,v \in \mathbb{R}^{\dim\theta} : \|u\|=1, \|v\|=1} \left\| \sum_{k,l=1}^{\dim\theta} u_k v_l \partial_{\theta\theta_k\theta_l} \mathcal{L}(\theta, \phi) \right\|$$

For matrices this induced norm is the spectral norm. Since the number of fixed effect parameters in the model grows with N and T, the choice of norm for vectors and matrices is important. Following FVW, we choose the ℓ_q-norm for $\dim \phi$ vectors and the corresponding

induced norms for matrices and tensors

$$\|\partial_{\phi\phi\phi}\mathcal{L}(\theta,\phi)\|_q = \max_{u,v\in\mathbb{R}^{\dim\phi}:\|u\|=1,\|v\|=1} \left\|\sum_{k,l=1}^{\dim\beta} u_k v_l \partial_{\phi\phi_k\phi_l}\mathcal{L}(\theta,\phi)\right\|_q$$

Note that for $w, x \in \mathbb{R}^{\dim\phi}$ and $q \geq 2$,

$$|w'x| \leq \|w\|_q \|x\|_{q/(q-1)} \leq (\dim\phi)^{(q-2)/q}\|w\|_q\|x\|_q$$

See FVW for more details on these norms. We also define the sets $\mathcal{B}(d,\theta_0) = \{\theta : \|\theta-\theta_0\| \leq d\}$ and $\mathcal{B}(d,\phi_0) = \{\phi : \|\phi-\phi_0\| \leq d\}$ for $d > 0$.

In the remainder of the section, we assume that all assumptions from Section 4 hold for all lemmas, unless explicitly stated otherwise. Additionally, we use the fact that the radius of our localized region $d_{NT} = o_p\left((NT)^{-1/4+\epsilon}\right)$, for $0 < \epsilon \leq \frac{1}{8} - \frac{1}{2q}$ with $q > 4$, when N and T go to infinity at the same rate.

C.2 Proof of Theorem 2 and 3

The first lemma is important for the stochastic expansion of $\hat{\theta}^{(m+1)}$, as it states that the expected incidental parameter Hessian matrix is asymptotically diagonal.

Lemma C.1. *We have*

$$\left\|\bar{\mathcal{H}}^{-1} - \bar{\mathcal{H}}_d^{-1}\right\| = O_p(1)$$

where $\bar{\mathcal{H}}_d = \operatorname{diag}\left(\bar{\mathcal{H}}^*_{(\lambda\lambda)}, \bar{\mathcal{H}}^*_{(ff)}\right)$. *Furthermore, for* $\theta \in \mathcal{B}(d_{NT},\theta_0)$ *and* $\phi \in \mathcal{B}\left(\sqrt{N\vee T}d_{NT},\phi_0\right)$, *there exists a constant* $c > 0$ *such that* $\sqrt{NT}\mathcal{H}(\theta,\phi) \geq c\mathbb{I}_{r(N+T)}$ *w.p.a.1.*

Proof of Lemma C.1. The result follows directly from Lemma 2 and Lemma 3 in Chen, Fernández-Val, and Weidner (2021); we therefore omit the proof. ■

Lemma C.2. *Let* $q = 8$, $\epsilon = 1/(16+2\iota)$, *and* $d_{NT} = o_p\left((NT)^{-1/4+\epsilon}\right)$. *Then, we have*
(i) For the q-norms defined before, we have

$$\|\mathcal{S}\|_q = O_p\left((NT)^{-3/4+1/(2q)}\right), \quad \|\partial_\theta\mathcal{L}\| = O_p\left((NT)^{-1/2}\right), \quad \|\tilde{\mathcal{H}}\|_q = o_p\left((NT)^{-1/2}\right),$$
$$\|\partial_{\theta\phi'}\mathcal{L}\|_q = O_p\left((NT)^{-1/2+1/(2q)}\right), \quad \|\partial_{\theta\theta'}\mathcal{L}\| = O_p(1), \quad \|\partial_{\theta\phi\phi}\mathcal{L}\|_q = O_p\left((NT)^{-1/2+\epsilon}\right),$$
$$\|\partial_{\phi\phi\phi}\mathcal{L}\|_q = O_p\left((NT)^{-1/2+\epsilon}\right),$$

and

$$\|\mathcal{S}\| = O_p\left((NT)^{-1/2}\right), \quad \left\|\mathcal{H}^{-1}\right\| = O_p\left(\sqrt{NT}\right), \quad \left\|\bar{\mathcal{H}}^{-1}\right\| = O_p\left(\sqrt{NT}\right)$$
$$\left\|\mathcal{H}^{-1} - \bar{\mathcal{H}}^{-1}\right\| = o_P\left((NT)^{3/8}\right), \quad \left\|\mathcal{H}^{-1} - \left(\bar{\mathcal{H}}^{-1} - \bar{\mathcal{H}}^{-1}\tilde{\mathcal{H}}\bar{\mathcal{H}}^{-1}\right)\right\| = o_P\left((NT)^{1/4}\right),$$
$$\|\partial_{\theta\phi'}\mathcal{L}\| = O_p\left((NT)^{-1/4}\right), \quad \|\partial_{\theta\phi\phi}\mathcal{L}\| = O_p\left((NT)^{-1/2+\epsilon}\right)$$
$$\left\|\sum_g \partial_{\phi\phi'\phi_g}\mathcal{L}\left[\mathcal{H}^{-1}\mathcal{S}\right]_g\right\| = O_p\left((NT)^{-3/4+1/(2q)+\epsilon}\right)$$
$$\left\|\sum_g \partial_{\phi\phi'\phi_g}\mathcal{L}\left[\bar{\mathcal{H}}^{-1}\mathcal{S}\right]_g\right\| = O_p\left((NT)^{-3/4+1/(2q)+\epsilon}\right)$$

(ii) Moreover,

$$\|\tilde{\mathcal{H}}\| = o_P\left((NT)^{-5/8}\right), \quad \left\|\partial_{\theta\theta'}\tilde{\mathcal{L}}\right\| = o_P(1), \quad \left\|\partial_{\theta\phi\phi}\tilde{\mathcal{L}}\right\| = o_P\left((NT)^{-5/8}\right),$$
$$\left\|\partial_{\theta\phi'}\tilde{\mathcal{L}}\right\| = O_p\left((NT)^{-1/2}\right), \quad \left\|\sum_{g,h=1}^{\dim\phi} \partial_{\phi\phi_g\phi_h}\tilde{\mathcal{L}}\left[\overline{\mathcal{H}}^{-1}\mathcal{S}\right]_g\left[\overline{\mathcal{H}}^{-1}\mathcal{S}\right]_h\right\| = o_P\left((NT)^{-3/4}\right).$$

(iii) For all $\theta \in \mathcal{B}\left(d_{NT}, \theta_0\right)$ *and* $\phi \in \mathcal{B}\left(\sqrt{N \vee T}d_{NT}, \phi_0\right)$*, we have*

$$\|\partial_{\theta\theta'}\mathcal{L}(\theta,\phi)\| = O_p(1), \quad \|\partial_{\theta\phi'}\mathcal{L}(\theta,\phi)\|_q = O_p\left((NT)^{-1/2+1/(2q)}\right), \quad \|\partial_{\theta\theta\theta}\mathcal{L}(\theta,\phi)\| = O_p(1)$$
$$\|\partial_{\theta\theta\phi}\mathcal{L}(\theta,\phi)\|_q = O_p\left((NT)^{-1/2+1/(2q)}\right), \quad \|\partial_{\theta\phi\phi}\mathcal{L}(\theta,\phi)\|_q = O_p\left((NT)^{-1/2+\epsilon}\right),$$
$$\|\partial_{\phi\phi\phi}\mathcal{L}(\theta,\phi)\|_q = O_p\left((NT)^{-1/2+\epsilon}\right), \quad \|\partial_{\theta\theta\phi\phi}\mathcal{L}(\theta,\phi)\|_q = O_p\left((NT)^{-1/2+\epsilon}\right),$$
$$\|\partial_{\theta\phi\phi\phi}\mathcal{L}(\theta,\phi)\|_q = O_p\left((NT)^{-1/2+\epsilon}\right), \quad \|\partial_{\phi\phi\phi\phi}\mathcal{L}(\theta,\phi)\|_q = O_p\left((NT)^{-1/2+\epsilon}\right)$$

Proof. The proofs follow those of Theorem 4 in Chen, Fernández-Val, and Weidner (2021) and Theorem C.1 and Lemma S.7 in Fernández-Val and Weidner (2016). ■

Lemma C.3. *The following stochastic expansion holds at the updated parameter* θ*:*

$$\partial_\theta\mathcal{L}\left(\hat{\theta}^{(m+1)}, \hat{\phi}^{(m)}\right) = \bar{W}\left(\theta^{(m+1)} - \theta_0\right) - U^{(0)}\left(\hat{\theta}^{(m)} - \theta_0\right) - U^{(1)} - U^{(2)} + R_\theta^{(m)}$$

where

$$\begin{aligned}
\bar{W} &= \partial_{\theta\theta'}\bar{\mathcal{L}} \\
U^{(0)} &= \left(\partial_{\theta\phi'}\bar{\mathcal{L}}\right)\bar{\mathcal{H}}^{-1}\left(\partial_{\phi\theta'}\bar{\mathcal{L}}\right) \\
U^{(1)} &= -\partial_{\theta}\mathcal{L} + \left(\partial_{\theta\phi'}\bar{\mathcal{L}}\right)\bar{\mathcal{H}}^{-1}\mathcal{S} \\
U^{(2)} &= \left(\left[\partial_{\theta\phi'}\tilde{\mathcal{L}}\right]\bar{\mathcal{H}}^{-1}\mathcal{S} - \left[\partial_{\theta\phi'}\bar{\mathcal{L}}\right]\bar{\mathcal{H}}^{-1}\tilde{\mathcal{H}}\bar{\mathcal{H}}^{-1}\mathcal{S}\right) \\
&\quad + \frac{1}{2}\sum_{g\in[\dim(\phi)]}\left(\partial_{\theta\phi'\phi_g}\bar{\mathcal{L}} + \left[\partial_{\theta\phi'}\bar{\mathcal{L}}\right]\bar{\mathcal{H}}^{-1}\left[\partial_{\phi\phi'\phi_g}\bar{\mathcal{L}}\right]\right)\bar{\mathcal{H}}^{-1}\mathcal{S}\left(\bar{\mathcal{H}}^{-1}\mathcal{S}\right)_g
\end{aligned}$$

and the remainder term satisfies $R_\theta^{(m)} = o_p\left(\left\|\hat{\theta}^{(m+1)} - \theta_0\right\|\right) + o_p\left(\left\|\hat{\theta}^{(m)} - \theta_0\right\|\right) + o_P\left((NT)^{-1/2}\right)$.

Proof of Lemma C.3. Based on Theorem B.1(i) in Fernández-Val and Weidner (2016), we have

$$\begin{aligned}
&\widehat{\phi}(\theta) - \phi_0 \\
&\quad = -\mathcal{H}^{-1}\left(\partial_{\phi\theta'}\mathcal{L}\right)(\theta - \theta_0) - \mathcal{H}^{-1}\mathcal{S} - \frac{1}{2}\mathcal{H}^{-1}\sum_g\left(\partial_{\phi\phi'\phi_g}\mathcal{L}\right)\mathcal{H}^{-1}\mathcal{S}\left(\mathcal{H}^{-1}\mathcal{S}\right)_g + R_{1\phi}(\theta)
\end{aligned}$$

and the remainder term of the expansion satisfies

$$\begin{aligned}
\left\|R_{1\phi}(\theta)\right\|_q &= \sup_{\|v\|_{q/(q-1)}=1} v' R_{1\phi}(\theta) \\
&= O_P\left[(NT)^{1/q+\epsilon}\|\theta-\theta_0\|^2 + (NT)^{-1/4+1/q+\epsilon}\|\theta-\theta_0\| + (NT)^{-3/4+3/(2q)+2\epsilon}\right] \\
&= o_P\left((NT)^{1/8+1/(2q)}\|\theta-\theta_0\|^2\right) + o_P\left((NT)^{-1/8+1/(2q)}\|\theta-\theta_0\|\right) \\
&\quad + o_P\left((NT)^{-1/2+1/(2q)}\right)
\end{aligned}$$

We can further decompose the deviation of the incidental parameter estimator as

$$\begin{aligned}
&\hat{\phi}(\theta) - \phi_0 \\
&\quad = -\bar{\mathcal{H}}^{-1}\left(\partial_{\phi\theta'}\bar{\mathcal{L}}\right)\left(\theta^{(m)} - \theta_0\right) - \bar{\mathcal{H}}^{-1}\mathcal{S} + \bar{\mathcal{H}}^{-1}\tilde{\mathcal{H}}\bar{\mathcal{H}}^{-1}\mathcal{S} \\
&\quad - \frac{1}{2}\bar{\mathcal{H}}^{-1}\sum_g\left(\partial_{\phi\phi'\phi_g}\bar{\mathcal{L}}\right)\bar{\mathcal{H}}^{-1}\mathcal{S}\left(\bar{\mathcal{H}}^{-1}\mathcal{S}\right)_g + R_{1\phi}(\theta) + R_{2\phi}(\theta) \qquad \text{(C.1)}
\end{aligned}$$

where

$$
\begin{aligned}
R_{2\phi}(\theta) \quad &= \left[\mathcal{H}^{-1}\left(\partial_{\phi\theta'}\mathcal{L}\right) - \bar{\mathcal{H}}^{-1}\left(\partial_{\phi\theta'}\bar{\mathcal{L}}\right)\right](\theta-\theta_0) + \left[\mathcal{H}^{-1} - \left(\bar{\mathcal{H}}^{-1} - \bar{\mathcal{H}}^{-1}\tilde{\mathcal{H}}\bar{\mathcal{H}}^{-1}\right)\right]\mathcal{S} \\
&+ \frac{1}{2}\left[\mathcal{H}^{-1}\sum_g \left(\partial_{\phi\phi'\phi_g}\mathcal{L}\right)\mathcal{H}^{-1}\mathcal{S}\left(\mathcal{H}^{-1}\mathcal{S}\right)_g - \bar{\mathcal{H}}^{-1}\sum_g \left(\partial_{\phi\phi'\phi_g}\bar{\mathcal{L}}\right)\bar{\mathcal{H}}^{-1}\mathcal{S}\left(\bar{\mathcal{H}}^{-1}\mathcal{S}\right)_g\right]
\end{aligned}
$$

Again, by Lemma C.2, we have

$$
\begin{aligned}
&\|R_\phi(\theta)\| \\
&\leq \left\|\mathcal{H}^{-1} - \bar{\mathcal{H}}^{-1}\right\| \|\partial_{\phi'\theta}\mathcal{L}\| \|\theta-\theta_0\| + \left\|\bar{\mathcal{H}}^{-1}\right\| \left\|\partial_{\phi'\theta}\tilde{\mathcal{L}}\right\| \|\theta-\theta_0\| \\
&+ \left\|\mathcal{H}^{-1} - \left(\bar{\mathcal{H}}^{-1} - \bar{\mathcal{H}}^{-1}\tilde{\mathcal{H}}\bar{\mathcal{H}}^{-1}\right)\right\| \|\mathcal{S}\| \\
&+ \frac{1}{2}\left(\left\|\mathcal{H}^{-1}\right\| + \left\|\bar{\mathcal{H}}^{-1}\right\|\right)\left\|\mathcal{H}^{-1} - \bar{\mathcal{H}}^{-1}\right\| \|\mathcal{S}\| \left\|\sum_g \partial_{\phi\phi'\phi_g}\mathcal{L}\left[\mathcal{H}^{-1}\mathcal{S}\right]_g\right\| \\
&+ \frac{1}{2}\left\|\mathcal{H}^{-1} - \bar{\mathcal{H}}^{-1}\right\|\left\|\bar{\mathcal{H}}^{-1}\right\| \|\mathcal{S}\| \left\|\sum_g \partial_{\phi\phi'\phi_g}\mathcal{L}\left[\bar{\mathcal{H}}^{-1}\mathcal{S}\right]_g\right\| \\
&+ \frac{1}{2}\left\|\bar{\mathcal{H}}^{-1}\right\| \|\mathcal{S}\| \left\|\sum_{g,h} \partial_{\phi\phi_g\phi_h}\tilde{\mathcal{L}}\left[\bar{\mathcal{H}}^{-1}\mathcal{S}\right]_g \tilde{\mathcal{L}}\left[\bar{\mathcal{H}}^{-1}\mathcal{S}\right]_h\right\| \\
&= o_p\left((NT)^{1/8}\|\theta-\theta_0\|\right) + o_p\left((NT)^{-1/4}\right)
\end{aligned}
$$

uniformly over $\theta \in \mathcal{B}(d_{NT}, \theta_0)$. Note that by Lemma C.2, we have

$$
\begin{aligned}
\left\|\widehat{\phi}(\theta) - \phi_0\right\|_q &\leq \left\|\mathcal{H}^{-1}\right\|_q \|\mathcal{S}\|_q + \left\|\mathcal{H}^{-1}\right\|_q \|\partial_{\phi\theta'}\mathcal{L}\|_q \|\theta-\theta_0\|_q \\
&+ \frac{1}{2}\left\|\mathcal{H}^{-1}\right\|_q^3 \|\partial_{\phi\phi\phi}\mathcal{L}\|_q \|\mathcal{S}\|_q^2 + \|R_{1\phi}(\theta)\|_q \\
&= O_P\left((NT)^{-1/4+1/(2q)}\right) + O_p\left((NT)^{1/(2q)}\|\theta-\theta_0\|\right) \\
&= O_p\left((NT)^{1/(2q)}\|\theta-\theta_0\|\right)
\end{aligned}
$$

Next, We expand $\partial_\theta\mathcal{L}\left(\hat{\theta}^{(m+1)}, \hat{\phi}^{(m)}\right)$ around (θ_0, ϕ_0) and get

$$
\begin{aligned}
\partial_\theta\mathcal{L}\left(\hat{\theta}^{(m+1)}, \hat{\phi}^{(m)}\right) =& \partial_\theta\mathcal{L} + \left(\partial_{\theta\theta'}\mathcal{L}\right)\left(\hat{\theta}^{(m+1)} - \theta_0\right) + \left(\partial_{\theta\phi'}\mathcal{L}\right)\left(\hat{\phi}^{(m)} - \phi_0\right) \\
&+ \frac{1}{2}\sum_{g\in[\dim(\phi)]}\left[\partial_{\theta\phi'\phi_g}\mathcal{L}\right]\mathcal{H}^{-1}\mathcal{S}\left(\mathcal{H}^{-1}\mathcal{S}\right)_g + R_{1\theta}^{(m)}
\end{aligned}
$$

and the remainder term satisfies

$$\begin{aligned}
\left\|R_{1\theta}^{(m)}\right\| &= \sup_{\|v\|=1} v' R_{1\theta}^{(m)} \\
&\leq \frac{1}{2}\left\|\sum_{g\in[\dim(\phi)]}\left[\partial_{\theta\phi'\phi_g}\mathcal{L}\right]\left(\hat{\phi}^{(m)}-\phi_0\right)\left(\hat{\phi}^{(m)}-\phi_0\right)_g - \sum_{g\in[\dim(\phi)]}\left[\partial_{\theta\phi'\phi_g}\mathcal{L}\right]\mathcal{H}^{-1}\mathcal{S}\left(\mathcal{H}^{-1}\mathcal{S}\right)_g\right\|_q \\
&+ \frac{1}{2}\left\|\partial_{\theta\theta\theta}\mathcal{L}\left(\theta^*,\phi_0\right)\right\|\left\|\hat{\theta}^{(m+1)}-\theta_0\right\|^2 + (NT)^{1/2-1/q}\left\|\partial_{\theta\theta\phi}\mathcal{L}\left(\theta_0,\phi^*\right)\right\|_q \\
&\cdot \left\|\hat{\phi}^{(m)}-\phi_0\right\|_q\left\|\hat{\theta}^{(m+1)}-\theta_0\right\| + \frac{1}{6}(NT)^{1/2-1/q}\left\|\partial_{\theta\phi\phi\phi}\mathcal{L}\left(\theta^*,\phi_0\right)\right\|_q\left\|\hat{\phi}^{(m)}-\phi_0\right\|_q^3 \\
&= o_p\left(\left\|\hat{\theta}^{(m+1)}-\theta_0\right\|\right) + o_p\left(\left\|\hat{\theta}^{(m)}-\theta_0\right\|\right) + O_p\left(\left\|\hat{\theta}^{(m+1)}-\theta_0\right\|^2\right) + o_P\left((NT)^{-1/2}\right) \\
&= o_p\left(\left\|\hat{\theta}^{(m+1)}-\theta_0\right\|\right) + o_p\left(\left\|\hat{\theta}^{(m)}-\theta_0\right\|\right) + o_P\left((NT)^{-1/2}\right)
\end{aligned}$$

where θ^* lies between $\hat{\theta}^{(m+1)}$ and θ_0 and ϕ^* lies between $\hat{\phi}^{(m)}$ and ϕ_0.

We can further decompose

$$\begin{aligned}
&\partial_\theta\mathcal{L}(\hat{\theta}^{(m+1)},\hat{\phi}^{(m)}) \\
&\quad= \partial_\theta\mathcal{L} + \left(\partial_{\theta\theta'}\bar{\mathcal{L}}\right)\left(\hat{\theta}^{(m+1)}-\theta_0\right) + \left(\partial_{\theta\phi'}\bar{\mathcal{L}}\right)\left(\hat{\phi}^{(m)}-\phi_0\right) \\
&\quad+ \frac{1}{2}\sum_g\left[\partial_{\theta\phi'\phi_g}\bar{\mathcal{L}}\right]\mathcal{H}^{-1}\mathcal{S}\left(\mathcal{H}^{-1}\mathcal{S}\right)_g + R_\theta^{(m)} \\
&\quad= \bar{W}\left(\hat{\theta}^{(m+1)}-\theta_0\right) - U^{(0)}\left(\hat{\theta}^{(m)}-\theta_0\right) - U^{(1)} - U^{(2)} + R_\theta^{(m)}
\end{aligned}$$

where the last equality uses (C.1), and $R_\theta^{(m)}$ takes the form

$$\begin{aligned}
R_\theta^{(m)} &= R_{1\theta}^{(m)} + \left(\partial_{\theta\theta'}\tilde{\mathcal{L}}\right)\left(\hat{\theta}^{(m+1)}-\theta_0\right) + \left(\partial_{\theta\phi'}\tilde{\mathcal{L}}\right)\bar{\mathcal{H}}^{-1}\left(\partial_{\phi\theta'}\bar{\mathcal{L}}\right)\left(\theta^{(m)}-\theta_0\right) - \left(\partial_{\theta\phi'}\tilde{\mathcal{L}}\right)\bar{\mathcal{H}}^{-1}\tilde{\mathcal{H}}\bar{\mathcal{H}}^{-1}\mathcal{S} \\
&+ \frac{1}{2}\left[\sum_g\left(\partial_{\theta\phi'\phi_g}\mathcal{L}\right)\mathcal{H}^{-1}\mathcal{S}\left(\mathcal{H}^{-1}\mathcal{S}\right)_g - \sum_g\left(\partial_{\theta\phi'\phi_g}\bar{\mathcal{L}}\right)\bar{\mathcal{H}}^{-1}\mathcal{S}\left(\bar{\mathcal{H}}^{-1}\mathcal{S}\right)_g\right] \\
&+ \frac{1}{2}\left(\partial_{\theta\phi'}\tilde{\mathcal{L}}\right)\bar{\mathcal{H}}^{-1}\sum_g\left(\partial_{\phi\phi'\phi_g}\bar{\mathcal{L}}\right)\bar{\mathcal{H}}^{-1}\mathcal{S}\left(\bar{\mathcal{H}}^{-1}\mathcal{S}\right)_g + \left(\partial_{\phi\theta'}\bar{\mathcal{L}}\right)\left(R_{1\phi}\left(\theta^{(m)}\right) + R_{2\phi}\left(\theta^{(m)}\right)\right)
\end{aligned}$$

Furthermore, from Lemma C.2, we have

$$\begin{aligned}
\left\|R_\theta^{(m)}\right\| \leq & \left\|R_{1\theta}^{(m)}\right\| + \left\|\partial_{\theta\theta'}\tilde{\mathcal{L}}\right\| \left\|\hat{\theta}^{(m+1)} - \theta_0\right\| + \left\|\partial_{\theta\phi'}\tilde{\mathcal{L}}\right\| \left\|\bar{\mathcal{H}}^{-1}\right\| \left\|\partial_{\phi\theta'}\bar{\mathcal{L}}\right\| \left\|\hat{\theta}^{(m)} - \theta_0\right\| \\
& + \left\|\bar{\mathcal{H}}^{-1}\right\|^2 \left\|\partial_{\theta\phi'}\tilde{\mathcal{L}}\right\| \left\|\tilde{\mathcal{H}}\right\| \|\mathcal{S}\| \\
& + \frac{1}{2} \left\|\partial_{\theta\phi\phi}\mathcal{L}\right\| \left(\left\|\mathcal{H}^{-1}\right\| + \left\|\bar{\mathcal{H}}^{-1}\right\|\right) \left\|\mathcal{H}^{-1} - \bar{\mathcal{H}}^{-1}\right\| \|\mathcal{S}\|^2 \\
& + \frac{1}{2} \left\|\bar{\mathcal{H}}^{-1}\right\|^2 \left\|\partial_{\theta\phi\phi}\tilde{\mathcal{L}}\right\| \|\mathcal{S}\|^2 + (NT)^{1/2-1/q} \left\|\partial_{\phi\theta'}\bar{\mathcal{L}}\right\|_q \left\|R_{1\phi}\left(\theta^{(m)}\right)\right\|_q + \left\|\partial_{\phi\theta'}\bar{\mathcal{L}}\right\| \left\|R_{1\theta}^{(m)}\right\| \\
= & \left\|R_{1\theta}^{(m)}\right\| + o_P\left(\left\|\hat{\theta}^{(m+1)} - \theta_0\right\|\right) + O_P\left((NT)^{-1/4} \left\|\hat{\theta}^{(m)} - \theta_0\right\|\right) \\
& + o_P\left((NT)^{-1/8+\epsilon} \left\|\hat{\theta}^{(m)} - \theta_0\right\|\right) + o_P\left((NT)^{-1/2}\right) \\
= & \, o_p\left(\left\|\hat{\theta}^{(m+1)} - \theta_0\right\|\right) + o_p\left(\left\|\hat{\theta}^{(m)} - \theta_0\right\|\right) + o_P\left((NT)^{-1/2}\right)
\end{aligned}$$

The desirable results then follow. ■

Before proceeding to the next lemma, we introduce several quantities. For any $N \times T$ matrix A we define the $N \times T$ matrix $\mathbb{P}A$ as follows

$$(\mathbb{P}A)_{it} = \lambda_i^{*\prime} f_{0t} + \lambda_{0i}' f_t^*, \quad (\lambda^*, f^*) \in \arg\min_{\lambda_i, f_t} \sum_{i,t} \mathbb{E}\left[\partial_{\pi^2}\ell_{it}\right] \left(A_{it} - \lambda_i' f_{0t} - \lambda_{0i}' f_t\right)^2$$

Note that $\mathbb{P}\mathbb{P} = \mathbb{P}$. We also define the linear operator $\tilde{\mathbb{P}}$ as

$$\tilde{\mathbb{P}}A = \mathbb{P}\tilde{A}, \quad \tilde{A}_{it} = \frac{A_{it}}{\mathbb{E}\left[\partial_{\pi^2}\ell_{it}\right]} \tag{C.2}$$

The next lemma gives a convenient algebraic result, which is used extensively in deriving the asymptotic expansion for the estimated common parameter. This is a straightforward extension of Lemma S.8 in FVW.

Lemma C.4. *Let A, B and C be $N \times T$ matrices, and let the expected incidental parameter Hessian $\bar{\mathcal{H}}$ be invertible. Define the $r(N+T)$ vectors $\mathcal{A}$ and $\mathcal{B}$ and the $r(N+T) \times r(N+T)$ matrix $\mathcal{C}$ as follows*[A3]

$$\mathcal{A} = \frac{1}{NT} \begin{pmatrix} \left(\sum_t A_{it} f_{0t}\right)_{i\in[N]} \\ \left(\sum_i A_{it} \lambda_{0i}\right)_{t\in[T]} \end{pmatrix}, \quad \mathcal{B} = \frac{1}{NT} \begin{pmatrix} \left(\sum_t B_{it} f_{0t}\right)_{i\in[N]} \\ \left(\sum_i B_{it} \lambda_{0i}\right)_{t\in[T]} \end{pmatrix}$$

[A3]Here, $\left(\sum_t A_{it} f_{0t}\right)_{i\in[N]}$ is a rN-dimensional vector with its i-th block to be a r-dimensional vector, $\sum_t A_{it} f_{0t}$.

and

$$\mathcal{C} = \frac{1}{NT}\begin{pmatrix} \text{diag}\left(\left(\sum_t C_{it} f_{0t} f_{0t}'\right)_{i\in[N]}\right) & \left(C_{it}\lambda_{0i} f_{0t}'\right)_{i\in[N],t\in[T]} \\ \left(C_{it}\lambda_{0i} f_{0t}'\right)'_{i\in[N],t\in[T]} & \text{diag}\left(\left(\sum_i C_{it}\lambda_{0i}\lambda_{0i}'\right)_{t\in[T]}\right) \end{pmatrix}$$

Then
(i) $\mathcal{A}'\bar{\mathcal{H}}^{-1}\mathcal{B} = \frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T \left(\tilde{\mathbb{P}}A\right)_{it} B_{it} = \frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T \left(\tilde{\mathbb{P}}B\right)_{it} A_{it}$;
(ii) $\mathcal{A}'\bar{\mathcal{H}}^{-1}\mathcal{B} = \frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T \mathbb{E}\left[\partial_{\pi^2}\ell_{it}\right]\left(\tilde{\mathbb{P}}A\right)_{it}\left(\tilde{\mathbb{P}}B\right)_{it}$;
(iii) $\mathcal{A}'\bar{\mathcal{H}}^{-1}\mathcal{C}^{-1}\bar{\mathcal{H}}^{-1}\mathcal{B} = \frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T \left(\tilde{\mathbb{P}}A\right)_{it} C_{it}\left(\tilde{\mathbb{P}}B\right)_{it}$.

Proof. To simplify the notations, we consider the case $r = 1$, but the proof can be easily generalized to the case $r > 1$. Let $\lambda_i^* f_{0t} + \lambda_{0i} f_t^* = \left(\mathbb{P}\tilde{A}\right)_{it} = \left(\tilde{\mathbb{P}}A\right)_{it}$, with $\tilde{A}_{it}$ defined in (C.2). The first order condition of the minimization prblem in the definition of $\left(\mathbb{P}\tilde{A}\right)_{it}$ can be written as $\bar{\mathcal{H}}\begin{pmatrix}\lambda^*\\ f^*\end{pmatrix} = \mathcal{A}$. One solution to this is $\begin{pmatrix}\lambda^*\\ f^*\end{pmatrix} = \bar{\mathcal{H}}^{-1}\mathcal{A}$. Therefore, $\mathcal{A}'\bar{\mathcal{H}}^{-1}\mathcal{B} = \begin{pmatrix}\lambda^*\\ f^*\end{pmatrix}\mathcal{B} = \frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T \left(\tilde{\mathbb{P}}A\right)_{it} B_{it}$. This is the first equality of the statement (i), and the second equality in statement (i) follows by symmetry. Statement (ii) is a special case of statement (iii) with $\mathcal{C} = \bar{\mathcal{H}}$. Let $\lambda_i f_{0t} + \lambda_{0i} f_t = \left(\mathbb{P}\tilde{B}\right)_{it} = \left(\tilde{\mathbb{P}}B\right)_{it}$ with $\tilde{B}_{it} = \frac{B_{it}}{\mathbb{E}\left[\partial_{\pi^2}\ell_{it}\right]}$. Analogous to the above, choose $\begin{pmatrix}\lambda\\ f\end{pmatrix} = \bar{\mathcal{H}}^{-1}\mathcal{B}$ as one solution to the minimization problem. Then $\mathcal{A}'\bar{\mathcal{H}}^{-1}\mathcal{C}^{-1}\bar{\mathcal{H}}^{-1}\mathcal{B} = \frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T \left(\tilde{\mathbb{P}}A\right)_{it} C_{it}\left(\tilde{\mathbb{P}}B\right)_{it}$. ■

Lemma C.5. *The approximate Hessian and the terms of the score defined in Lemma C.3 can be written as*

$$\begin{aligned}
\bar{W} &= \frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T \mathbb{E}\left[\partial_{\theta\theta}\ell_{it}\right] \\
U^{(0)} &= \frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T \mathbb{E}\left[\partial_{\pi^2}\ell_{it}\Xi_{it}\Xi_{it}'\right] \\
U^{(1)} &= \frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T -D_\theta\ell_{it} \\
U^{(2)} &= -\frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T \Lambda_{it}\left(D_{\theta\pi}\ell_{it} - \mathbb{E}\left[D_{\theta\pi}\ell_{it}\right]\right) + \frac{1}{2NT}\sum_{i=1}^N\sum_{t=1}^T \Lambda_{it}^2\mathbb{E}\left[D_{\theta\pi^2}\ell_{it}\right]
\end{aligned}$$

where $D_{\theta\pi^q}\ell_{it} = \partial_{\theta\pi^q}\ell_{it} - \Xi_{it}\partial_{\pi^{q+1}}\ell_{it}$ *with* $q = 0, 1, 2$.

Proof of Lemma C.5. First, it is by definition that

$$\bar{W} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{E}\left[\partial_{\theta\theta}\ell_{it}\right]$$

The terms $U^{(0)}$ is easy by Lemma C.4(ii). Since $\mathbb{E}\left[\partial_\theta \mathcal{L}\right] = 0$ and $\mathbb{E}\left[\mathcal{S}\right] = 0$, we have $\mathbb{E}\left[U^{(1)}\right] = 0$. Also, from Lemma C.3 and Lemma C.4(i), we have

$$U^{(1)} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(-\partial_\theta \ell_{it} + \Xi_{it}\partial_\pi \ell_{it}\right) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} -D_\theta \ell_{it}$$

It remains to show the decomposition of $U^{(2)}$. Define $\Lambda_{d,it}$ to be

$$\Lambda_{it} \coloneqq -\frac{1}{NT}\sum_{j=1}^{N}\sum_{\tau=1}^{T}\left(f_{0t}'\bar{\mathcal{H}}^{-1}_{(\lambda\lambda)ij}f_{0\tau} + f_{0t}'\bar{\mathcal{H}}^{-1}_{(\lambda f)i\tau}\lambda_{0j} + \lambda_{0i}'\bar{\mathcal{H}}^{-1}_{(f\lambda)tj}f_{0\tau} + \lambda_{0i}'\bar{\mathcal{H}}^{-1}_{(ff)t\tau}\lambda_{0j}\right)\partial_\pi \ell_{j\tau} \tag{C.3}$$

In addition, with Lemma C.4(iii),

$$\begin{aligned}
U^{(2)} &= \left[\partial_{\theta\phi'}\tilde{\mathcal{L}}\right]\bar{H}^{-1}\mathcal{S} - \left[\partial_{\theta\phi'}\bar{\mathcal{L}}\right]\bar{\mathcal{H}}^{-1}\tilde{\mathcal{H}}\bar{\mathcal{H}}^{-1}\mathcal{S} \\
&+ \frac{1}{2}\sum_{g=1}\left(\partial_{\theta\phi'\phi_g}\bar{\mathcal{L}} + \left[\partial_{\theta\phi'}\bar{\mathcal{L}}\right]\bar{\mathcal{H}}^{-1}\left[\partial_{\phi\phi'\phi_g}\bar{\mathcal{L}}\right]\right)\bar{\mathcal{H}}^{-1}\mathcal{S}\left(\bar{\mathcal{H}}^{-1}\mathcal{S}_g\right) \\
&= -\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Lambda_{it}\left(\partial_{\theta\pi}\tilde{\ell}_{it} + \Xi_{it}\partial_{\pi^2}\tilde{\ell}_{it}\right) \\
&+ \frac{1}{2NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Lambda_{it}^2\left(\mathbb{E}\left[\partial_{\theta\pi^2}\ell_{it}\right] + \left(\partial_{\theta\phi'}\bar{\mathcal{L}}\right)\bar{\mathcal{H}}^{-1}\mathbb{E}\left[\partial_\phi\partial_{\pi^2}\ell_{it}\right]\right) \\
&= \underbrace{-\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Lambda_{it}\left(D_{\theta\pi}\ell_{it} - \mathbb{E}\left[D_{\theta\pi}\ell_{it}\right]\right)}_{U^{(2a)}} \underbrace{+\frac{1}{2NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Lambda_{it}^2\mathbb{E}\left[D_{\theta\pi^2}\ell_{it}\right]}_{U^{(2b)}}
\end{aligned} \tag{C.4}$$

where the penultimate equality uses Lemma C.4(i). For each i and t, $\partial_\phi\partial_{\pi^2}\ell_{it}$ is a dim-ϕ vector, which can be written as $\partial_\phi\partial_{\pi^2}\ell_{it} = \begin{pmatrix} A1_{rT} \\ A'1_{rN} \end{pmatrix}$ for an $r\left(N\times T\right)$ matrix A with elements $A_{j\tau} = \partial_{\pi^3}\ell_{j\tau}$ if $j = i$ and $\tau = t$, and $A_{j\tau} = 0$ otherwise. Thus, Lemma C.4(i) gives $\left(\partial_{\theta\phi'}\bar{\mathcal{L}}\right)\bar{\mathcal{H}}^{-1}\partial_\phi\partial_{\pi^2}\ell_{it} = -\sum_{j,\tau}\Xi_{j\tau}\delta_{(i=j)}\delta_{(t=\tau)}\partial_{\pi^3}\ell_{it} = -\Xi_{it}\partial_{\pi^3}\ell_{it}$. ■

Lemma C.6. *The terms of the score defined in Lemma C.3 have the following limiting behaviors,* $U^{(0)} = O_p\left(1/\sqrt{NT}\right)$, $U^{(1)} = O_p\left(1/\sqrt{NT}\right)$, *and* $U^{(2)} \xrightarrow{p} \bar{B}_\infty/T + \bar{D}_\infty/N$,

where $\bar{B}_\infty$ and $\bar{D}_\infty$ are defined in Theorem 3.

Proof of Lemma C.6. It is clear that $U^{(0)} = O_p\left(\frac{1}{\sqrt{NT}}\right)$, and by CLT, $\sqrt{NT}U^{(1)} \xrightarrow{d} N\left(0, \bar{\Sigma}_\infty\right)$, where

$$\bar{\Sigma}_\infty = \lim_{N,T\to\infty} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbb{E}\left[\left(D_\theta \ell_{it}\right)\left(D_\theta \ell_{it}\right)'\right]$$

As for $U^{(2)}$, the main focus is to show that the bias formulas by taking account the specific structure of the incidental parameters Hessian. Decompose

$$\Lambda_{it} = \Lambda_{it}^{(1)} + \Lambda_{it}^{(2)} + \Lambda_{it}^{(3)} + \Lambda_{it}^{(4)}$$

with

$$\begin{aligned}
\Lambda_{it}^{(1)} &= -\frac{1}{NT} \sum_{j=1}^{N} f_{0t}' \bar{\mathcal{H}}_{(\lambda\lambda)ij}^{-1} \sum_{\tau=1}^{T} \partial_\pi \ell_{j\tau} f_{0\tau} \\
\Lambda_{it}^{(2)} &= -\frac{1}{NT} \sum_{\tau=1}^{T} f_{0t}' \bar{\mathcal{H}}_{(\lambda f)i\tau}^{-1} \sum_{j=1}^{N} \partial_\pi \ell_{j\tau} \lambda_{0j} \\
\Lambda_{it}^{(3)} &= -\frac{1}{NT} \sum_{j=1}^{N} \lambda_{0i}' \bar{\mathcal{H}}_{(f\lambda)tj}^{-1} \sum_{\tau=1}^{T} \partial_\pi \ell_{j\tau} f_{0\tau} \\
\Lambda_{it}^{(4)} &= -\frac{1}{NT} \sum_{\tau=1}^{T} \lambda_{0i}' \bar{\mathcal{H}}_{(ff)t\tau}^{-1} \sum_{j=1}^{N} \partial_\pi \ell_{j\tau} \lambda_{0j}
\end{aligned}$$

We look at two decomposition terms $U^{(2a)}$ and $U^{(2b)}$ as specified in (C.4). Let

$$U^{(2a)} = U^{(2a,1)} + U^{(2a,2)} + U^{(2a,3)} + U^{(2a,4)}$$

where

$$\begin{aligned}
U^{(2a,1)} &= \frac{1}{(NT)^2} \sum_{i,j,t} f_{0t}' \bar{\mathcal{H}}_{(\lambda\lambda)ij}^{-1} \left(\sum_{\tau} \partial_\pi \ell_{j\tau} f_{0\tau}\right) \left(D_{\theta\pi} \ell_{it} - \mathbb{E}\left[D_{\theta\pi} \ell_{it}\right]\right) \\
U^{(2a,2)} &= \frac{1}{(NT)^2} \sum_{i,\tau,t} f_{0t}' \bar{\mathcal{H}}_{(\lambda f)i\tau}^{-1} \left(\sum_{j} \partial_\pi \ell_{j\tau} \lambda_{0j}\right) \left(D_{\theta\pi} \ell_{it} - \mathbb{E}\left[D_{\theta\pi} \ell_{it}\right]\right) \\
U^{(2a,3)} &= \frac{1}{(NT)^2} \sum_{t,j,i} \lambda_{0i}' \bar{\mathcal{H}}_{(f\lambda)tj}^{-1} \left(\sum_{\tau} \partial_\pi \ell_{j\tau} f_{0\tau}\right) \left(D_{\theta\pi} \ell_{it} - \mathbb{E}\left[D_{\theta\pi} \ell_{it}\right]\right) \\
U^{(2a,4)} &= \frac{1}{(NT)^2} \sum_{t,\tau,i} \lambda_{0i}' \bar{\mathcal{H}}_{(ff)t\tau}^{-1} \left(\sum_{j} \partial_\pi \ell_{j\tau} \lambda_{0j}\right) \left(D_{\theta\pi} \ell_{it} - \mathbb{E}\left[D_{\theta\pi} \ell_{it}\right]\right)
\end{aligned}$$

Use $\left\|\bar{\mathcal{H}}^{-1} - \bar{\mathcal{H}}_d^{-1}\right\| = O_p(1)$ by Lemma C.1, and apply the Cauchy-Schwarz inequality to the sum over t in $U^{(2a,3)}$, and since both $\lambda_{0i}'\bar{\mathcal{H}}^{-1}_{(f\lambda)}\partial_\pi \ell_{j\tau} f_{0\tau}$ and $(D_{\theta\pi}\ell_{it} - \mathbb{E}[D_{\theta\pi}\ell_{it}])$ are mean zero, independent across i,

$$\begin{aligned}\left(U^{(2a,3)}\right)^2 &\leq \frac{1}{(NT)^4}\left[\sum_t\left(\sum_{j,\tau}\lambda_{0i}'\bar{\mathcal{H}}^{-1}_{(f\lambda)tj}\partial_\pi\ell_{j\tau}f_{0\tau}\right)^2\right]\left[\sum_t\left(\sum_i\left(D_{\theta\pi}\ell_{it} - \mathbb{E}[D_{\theta\pi}\ell_{it}]\right)\right)^2\right] \\ &= \frac{1}{(NT)^4}\left[\sum_t O_p(NT)\right]\left[\sum_t O_p(N)\right] = O_p\left(1/\left(N^2T\right)\right) = o_p\left(\frac{1}{NT}\right)\end{aligned}$$

Thus, $U^{(2a,3)} = o_p(1/\sqrt{NT})$. Analogously, $U^{(2a,2)} = o_p(1/\sqrt{NT})$.

According to Lemma C.1, $\bar{\mathcal{H}}^{-1}_{(\lambda\lambda)} = \text{diag}\left[\left(\frac{1}{NT}\sum_{t=1}^T \mathbb{E}\left[\partial_{\pi^2}\ell_{it}\right] f_{0t}f_{0t}'\right)^{-1}\right] + O_p(1)$. Analogously to the proof of $U^{(2a,3)}$, the $O_p(1)$ part of $\bar{\mathcal{H}}^{-1}_{(\lambda\lambda)}$ has an asymptotically negligible contribution to $U^{(2a,1)}$.

$$\begin{aligned}&U^{(2a,1)} \\ &= \frac{1}{(NT)^2}\sum_{i,j} f_{0t}'\bar{\mathcal{H}}^{-1}_{(\lambda\lambda)ij}\left(\sum_\tau \partial_\pi\ell_{j\tau}f_{0\tau}\right)\sum_t\left(D_{\theta\pi}\ell_{it} - \mathbb{E}[D_{\theta\pi}\ell_{it}]\right) \\ &= \frac{1}{NT}\sum_i\left\{f_{0t}'\left(\frac{1}{NT}\sum_{t=1}^T\mathbb{E}\left[\partial_{\pi^2}\ell_{it}\right]f_{0t}f_{0t}'\right)^{-1}\cdot\right. \\ &\quad \left.\left(\sum_\tau\partial_\pi\ell_{i\tau}f_{0\tau}\right)\sum_t\left(D_{\theta\pi}\ell_{it} - \mathbb{E}[D_{\theta\pi}\ell_{it}]\right)\right\} + o_p\left(\frac{1}{\sqrt{NT}}\right) \\ &= \underbrace{\frac{1}{T}\cdot\frac{1}{N}\sum_{i=1}^N\sum_{t=1}^T\sum_{\tau=t}^T f_{0t}'\left(\sum_{t=1}^T\mathbb{E}\left[\partial_{\pi^2}\ell_{it}\right]f_{0t}f_{0t}'\right)^{-1}f_{0\tau}\mathbb{E}\left[\partial_\pi\ell_{it}D_{\theta\pi}\ell_{i\tau}\right]}_{=:\bar{B}^{(1)}} + o_p\left(\frac{1}{\sqrt{NT}}\right)\end{aligned}$$

Note, previous assumptions guarantee that $\mathbb{E}\left[\left(U_i^{(2a,1)}\right)^2\right] = O_p\left(1/(NT)\right)$ uniformly over i. For the numerator, both $\partial_\pi\ell_{i\tau}f_{0\tau}$ and $(D_{\theta\pi}\ell_{it} - \mathbb{E}[D_{\theta\pi}\ell_{it}])$ are mean zero weakly correlated processes, hence the sum over which is of order $\sqrt{T}$ each. The denominator of of $U_i^{(1a,1)}$ is of order T as it sums over T. The last equality applies the WLLN over i, $\frac{1}{N}\sum_i U_i^{(2a,1)} = \frac{1}{N}\mathbb{E}U_i^{(2a,1)} + o_P(1)$, and by using $\mathbb{E}\left[\partial_\pi\ell_{it}D_{\theta\pi}\ell_{i\tau}\right] = 0$ for $t > \tau$. Analogously,

$$U^{(2a,4)} = \underbrace{\frac{1}{N}\cdot\frac{1}{T}\sum_{t=1}^T\sum_{i=1}^N\lambda_{0i}'\left(\sum_{i=1}^N\mathbb{E}\left[\partial_{\pi^2}\ell_{it}\right]\lambda_{0i}\lambda_{0i}'\right)^{-1}\lambda_{0i}\mathbb{E}\left[\partial_\pi\ell_{it}D_{\theta\pi}\ell_{it}\right]}_{=:\bar{D}^{(1)}} + o_p\left(\frac{1}{\sqrt{NT}}\right)$$

The definition of Λ_{it} induces the following decomposition of $U^{(2b)}$,

$$U^{(2b)} = \sum_{s,l=1}^{4} U^{(2b,s,l)}, \quad U^{(2b,s,l)} = \frac{1}{2NT} \sum_{i,t} \Lambda_{it}^{(s)} \Lambda_{it}^{(l)} \mathbb{E}\left[D_{\theta\pi^2} \ell_{it}\right]$$

Due to the symmetry $U^{(2b,s,l)} = U^{(2b,l,s)}$, this decomposition has 10 distinct terms. Starting with $U^{(2b,1,3)}$, we have

$$\begin{aligned}
U^{(2b,1,3)} &= \frac{1}{NT} \sum_{i=1}^{N} U_i^{(2b,1,3)} \\
U_i^{(2b,1,3)} &= \frac{1}{2T} \sum_{t=1}^{T} \mathbb{E}\left[D_{\theta\pi^2} \ell_{it}\right] f_{0t}' U_{it}^{(2b,1,3)} \lambda_{0i} \\
U_{it}^{(2b,1,3)} &= \frac{1}{N^2} \sum_{j_1,j_2=1}^{N} \left\{ \bar{\mathcal{H}}_{(\lambda\lambda)ij_1}^{-1} \left(\frac{1}{\sqrt{T}} \sum_{\tau=1}^{T} \partial_\pi \ell_{j_1\tau} f_{0\tau} \right) \left(\frac{1}{\sqrt{T}} \sum_{\tau=1}^{T} \partial_\pi \ell_{j_2\tau} f_{0\tau}' \right) \bar{\mathcal{H}}_{(f\lambda)tj_2}^{-1} \right\}
\end{aligned}$$

Given that $\mathbb{E}\left[\sum_t \partial_\pi \ell_{it} f_{0t}\right] = 0$ and $\mathbb{E}\left[\sum_t \partial_\pi \ell_{it} f_{0t} \sum_\tau \partial_\pi \ell_{j\tau} f_{0\tau}\right] = 0$ for $i \neq j$, along with the properties of the inverse expected Hessian from Lemma C.1, we have $\mathbb{E}\left[U_i^{(2b,1,3)}\right] = O_p(1/N)$ uniformly over i, $\mathbb{E}\left[\left(U_i^{(2b,1,3)}\right)^2\right] = O_p(1)$ uniformly over i, and $\mathbb{E}\left[U_i^{(2b,1,3)} U_j^{(2b,1,3)}\right] = O_p(1/N)$ uniformly over $i \neq j$. This implies that $\mathbb{E}\left[\sqrt{NT} U^{(2b,1,3)}\right] = O_p(1/N)$, and $\mathbb{E}\left[NT\left(U^{(2b,1,3)} - \mathbb{E}\left[U^{(2b,1,3)}\right]\right)^2\right] = O_p(1/\sqrt{N})$. Therefore $U^{(2b,1,3)} = o_p\left(1/\sqrt{NT}\right)$. By similar arguments, we have $U^{(2b,s,l)} = o_p\left(1/\sqrt{NT}\right)$ for all combinations of $s, l = 1, 2, 3, 4$, except for $s = l = 1$ and $s = l = 4$. For $s = l = 1$,

$$\begin{aligned}
U^{(2b,1,1)} &= \frac{1}{NT} \sum_{i=1}^{N} U_i^{(2b,1,1)} \\
U_i^{(2b,1,1)} &= \frac{1}{2T} \sum_{t=1}^{T} \mathbb{E}\left[D_{\theta\pi^2} \ell_{it}\right] f_{0t}' U_{it}^{(2b,1,1)} f_{0t} \\
U_{it}^{(2b,1,1)} &= \frac{1}{N^2} \sum_{j_1,j_2=1}^{N} \left\{ \bar{\mathcal{H}}_{(\lambda\lambda)ij_1}^{-1} \left(\frac{1}{\sqrt{T}} \sum_{\tau=1}^{T} \partial_\pi \ell_{j_1\tau} f_{0\tau} \right) \left(\frac{1}{\sqrt{T}} \sum_{\tau=1}^{T} \partial_\pi \ell_{j_2\tau} f_{0\tau}' \right) \bar{\mathcal{H}}_{(\lambda\lambda)ij_2}^{-1} \right\}
\end{aligned}$$

Analogous to the result for $U^{(2b,1,3)}$, we have $\mathbb{E}\left[NT\left(U^{(2b,1,1)} - \mathbb{E}U^{(2b,1,1)}\right)^2\right] = O_p(1/\sqrt{N})$.

Thus,

$$\begin{aligned}
U^{(2b,1,1)} &= \mathbb{E}\left[U^{(2b,1,1)}\right] + o_p\left(\frac{1}{\sqrt{NT}}\right) \\
&= -\frac{1}{2NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{E}\left[D_{\theta\pi^2}\ell_{it}\right] f'_{0t}\left(\sum_{t=1}^{T}\mathbb{E}\left[\partial_{\pi^2}\ell_{it}\right] f_{0t}f'_{0t}\right)^{-1}\mathbb{E}\left[\left(\partial_{\pi}\ell_{it}f_{0t}\right)\left(\partial_{\pi}\ell_{it}f_{0t}\right)'\right] \\
&\quad\cdot\left(\sum_{t=1}^{T}\mathbb{E}\left[\partial_{\pi^2}\ell_{it}\right] f_{0t}f'_{0t}\right)^{-1} f_{0t} + o_p\left(\frac{1}{\sqrt{NT}}\right) \\
&= \frac{1}{T}\cdot\underbrace{\frac{1}{2N}\sum_{i=1}^{N}\sum_{t=1}^{T} f'_{0t}\left(\sum_{t=1}^{T}\mathbb{E}\left[\partial_{\pi^2}\ell_{it}\right] f_{0t}f'_{0t}\right)^{-1} f_{0t}\mathbb{E}\left[D_{\theta\pi^2}\ell_{it}\right]}_{=:\bar{B}^{(2)}} + o_p\left(\frac{1}{\sqrt{NT}}\right)
\end{aligned}$$

Similarly,

$$\begin{aligned}
U^{(2b,4,4)} &= \mathbb{E}U^{(2b,4,4)} + o_p\left(\frac{1}{\sqrt{NT}}\right) \\
&= \frac{1}{N}\cdot\underbrace{\frac{1}{2T}\sum_{t=1}^{T}\sum_{i=1}^{N}\lambda'_{0i}\left(\sum_{i=1}^{N}\mathbb{E}\left[\partial_{\pi^2}\ell_{it}\right]\lambda_{0i}\lambda'_{0i}\right)^{-1}\lambda_{0i}\mathbb{E}\left[D_{\theta\pi^2}\ell_{it}\right]}_{=:\bar{D}^{(2)}} + o_p\left(\frac{1}{\sqrt{NT}}\right)
\end{aligned}$$

Summing up the above terms, we have $U^{(2a)} = \frac{1}{T}\bar{B}^{(1)} + \frac{1}{N}\bar{D}^{(1)} + o_p\left(\frac{1}{\sqrt{NT}}\right)$ and $U^{(2b)} = \frac{1}{T}\bar{B}^{(2)} + \frac{1}{N}\bar{D}^{(2)} + o_p\left(\frac{1}{\sqrt{NT}}\right)$. Since $\bar{B}_\infty = \lim_{N,T\to\infty}\left(\bar{B}^{(1)} + \bar{B}^{(2)}\right)$ and $\bar{D}_\infty = \lim_{N,T\to\infty}\left(\bar{D}^{(1)} + \bar{D}^{(2)}\right)$, then $U^{(2)} = \frac{1}{T}\bar{B}_\infty + \frac{1}{N}\bar{D}_\infty + o_p(1)$. We have shown that

$$U^{(2)} \xrightarrow{p} \frac{1}{T}\bar{B}_\infty + \frac{1}{N}\bar{D}_\infty \tag{C.5}$$

■

Now, we can proceed to prove our main theorems.

Proof of Theorem 2(i). Now we have

$$\begin{aligned}
&\theta^{(m+1)} - \theta_0 \\
&\quad= \bar{W}^{-1}U^{(0)}\left(\theta^{(m)} - \theta_0\right) + \bar{W}^{-1}U^{(1)} + \bar{W}^{-1}U^{(2)} \\
&\quad+ o_p\left(\left\|\theta^{(m+1)} - \theta_0\right\|\right) + o_p\left(\left\|\theta^{(m)} - \theta_0\right\|\right) + o_P\left((NT)^{-1/2}\right)
\end{aligned}$$

Let $\bar{C}^{(0)} = \bar{W}^{-1}\left(\partial_{\theta\phi'}\bar{\mathcal{L}}\right)\bar{\mathcal{H}}^{-1}\left(\partial_{\phi\theta'}\bar{\mathcal{L}}\right)$. We want to show that $\sigma_{\max}\left(\bar{C}^{(0)}\right) \in (0,1]$. Note that

by the Bartlett identities, $\mathbb{E}\left[\partial_\theta \mathcal{L} \partial_{\theta'} \mathcal{L}\right] = \partial_{\theta\theta'} \bar{\mathcal{L}}$, $\mathbb{E}\left[\partial_\theta \mathcal{L} \mathcal{S}'\right] = \partial_{\theta\phi'} \bar{\mathcal{L}}$, and $\mathbb{E}\left[\mathcal{S}\mathcal{S}'\right] = \bar{\mathcal{H}}$, then

$$\begin{aligned}
&\left(\partial_\theta \mathcal{L} - \left(\partial_{\theta\phi'} \bar{\mathcal{L}}\right) \bar{\mathcal{H}}^{-1} \mathcal{S}\right) \left(\partial_\theta \mathcal{L} - \left(\partial_{\theta\phi'} \bar{\mathcal{L}}\right) \bar{\mathcal{H}}^{-1} \mathcal{S}\right)' \\
&\quad = \partial_\theta \mathcal{L} \partial_{\theta'} \mathcal{L} - 2 \partial_\theta \mathcal{L} \mathcal{S} \bar{\mathcal{H}}^{-1} \left(\partial_{\theta\phi'} \bar{\mathcal{L}}\right)' + \left(\partial_{\theta\phi'} \bar{\mathcal{L}}\right) \bar{\mathcal{H}}^{-1} \mathcal{S}\mathcal{S}' \bar{\mathcal{H}}^{-1} \left(\partial_{\theta\phi'} \bar{\mathcal{L}}\right)' \\
&\quad \xrightarrow{p} \partial_{\theta\theta'} \bar{\mathcal{L}} - \left(\partial_{\theta\phi'} \bar{\mathcal{L}}\right) \bar{\mathcal{H}}^{-1} \left(\partial_{\phi\theta'} \bar{\mathcal{L}}\right)
\end{aligned}$$

Therefore, $\bar{W}\left(\mathbb{I}_p - \bar{C}^{(0)}\right) = \partial_{\theta\theta'} \bar{\mathcal{L}} - \left(\partial_{\theta\phi'} \bar{\mathcal{L}}\right) \bar{\mathcal{H}}^{-1} \left(\partial_{\phi\theta'} \bar{\mathcal{L}}\right) > 0$. By Fact 8.14.20 in Bernstein (2005, p.329),

$$\begin{aligned}
\sigma_{\min}\left(\mathbb{I}_p - \bar{C}^{(0)}\right) &\geq \sigma_{\max}\left(\bar{W}^{-1}\right) \sigma_{\min}\left(\bar{W}\left(\mathbb{I}_p - \bar{C}^{(0)}\right)\right) \\
&= \sigma_{\max}^{-1}\left(\bar{W}\right) \sigma_{\min}\left(\bar{W}\left(\mathbb{I}_p - \bar{C}^{(0)}\right)\right) := \bar{\rho} \in (0, 1]
\end{aligned}$$

where we use that $\bar{W} - \bar{W}\left(\mathbb{I}_p - \bar{C}^{(0)}\right) = \bar{W}\bar{C}^{(0)} \geq 0$. This implies that $\|\bar{C}^{(0)}\| = \sigma_{\max}\left(\bar{C}^{(0)}\right) = 1 - \sigma_{\min}\left(\mathbb{I}_p - \bar{C}^{(0)}\right) \leq 1 - \bar{\rho} \in [0, 1)$ w.p.a.1. ■

Proof of Theorem 2(ii). By Theorem 2(i) and the fact that $\bar{W}^{-1}U^{(1)} + \bar{W}^{-1}U^{(2)} = O_p\left((NT)^{-1/2}\right)$, we have $\left\|\hat{\theta}^{(m+1)} - \theta_0\right\| = O_p\left(\left\|\hat{\theta}^{(m)} - \theta_0\right\|\right)$. Continuous backward substitutions then give

$$\hat{\theta}^{(m+1)} - \theta_0 = \left[\bar{W}^{-1}U^{(0)} + o_p(1)\right]^{m+1} \left(\hat{\theta}^{(0)} - \theta_0\right) \tag{C.6}$$

$$+ \sum_{s=0}^{m} \left[\bar{W}^{-1}U^{(0)}\right]^s \left\{\bar{W}^{-1}U^{(1)} + \bar{W}^{-1}U^{(2)} + o_P\left((NT)^{-1/2}\right)\right\} \tag{C.7}$$

Thus, when $m + 1 \geq -\left(\frac{1}{2}\log(NT) + \log(\gamma_{NT})\right) / \log\left(\|\bar{C}^{(0)}\|\right)$, we can readily have

$$\left[\bar{W}^{-1}U^{(0)} + o_p(1)\right]^{m+1} \left(\hat{\theta}^{(0)} - \theta_0\right) = O_p\left((NT)^{-1/2}\right)$$

by Theorem 1. This, in conjunction with the fact that $\bar{W}^{-1}U^{(1)} + \bar{W}^{-1}U^{(2)} = O_p\left((NT)^{-1/2}\right)$, implies $\theta^{(m+1)} - \theta_0 = O_p\left((NT)^{-1/2}\right)$. ■

Proof of Theorem 3. For any $m \geq -\frac{1}{2}\log(NT)/\log\left(\|\bar{C}^{(0)}\|\right) - 1$, the cumulative effect (C.6) can be controlled as follows

$$\|(\text{C.6})\| \leq \sigma_{\max}^{m+1}\left(\bar{C}^{(0)}\right) O_p(\gamma_{NT}) = O_p\left((NT)^{-1/2}\gamma_{NT}\right)$$

Then

$$\begin{aligned}\sqrt{NT}\left(\hat{\theta}^{(m+1)}-\theta_0\right) &= \sum_{s=0}^{m}\left[\bar{W}^{-1}U^{(0)}\right]^s\left\{\bar{W}^{-1}U^{(1)}+\bar{W}^{-1}U^{(2)}\right\}+o_p(1)\\ &= \left(\mathbb{I}-\bar{W}^{-1}U^{(0)}\right)^{-1}\left(\mathbb{I}-\left[\bar{W}^{-1}U^{(0)}\right]^{m+1}\right)\bar{W}^{-1}\sqrt{NT}\left(U^{(1)}+U^{(2)}\right)+o_p(1)\\ &= \left(\mathbb{I}-\bar{W}^{-1}U^{(0)}\right)^{-1}\bar{W}^{-1}\sqrt{NT}\left(U^{(1)}+U^{(2)}\right)+o_p(1)\\ &= \left(\bar{W}-U^{(0)}\right)^{-1}\sqrt{NT}\left(U^{(1)}+U^{(2)}\right)+o_p(1)\end{aligned}$$

where the second equality follows from the summation formula for geometric sequences and the penultimate equality holds since for $m \geq -\frac{1}{2}\log(NT)/\log\left(\|\bar{C}^{(0)}\|\right)-1$,

$$\left[\bar{W}^{-1}U^{(0)}\right]^{m+1}\bar{W}^{-1}\sqrt{NT}\left(U^{(1)}+U^{(2)}\right)=O_p\left(\bar{W}^{-1}\left(U^{(1)}+U^{(2)}\right)\right)=O_p\left((NT)^{-1/2}\right)$$

Thus, by the definition of $\bar{W}$, $U^{(0)}$, and Ξ_{it}, we have

$$\bar{W}-U^{(0)}=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{E}\left[\partial_{\theta\theta}\ell_{it}-\partial_{\pi^2}\ell_{it}\Xi_{it}\Xi_{it}'\right]$$

By CLT and Bartlett identities,

$$\sqrt{NT}U^{(1)}=\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}-D_\theta\ell_{it}\xrightarrow{d}\mathcal{N}\left(0,\bar{W}_\infty\right)$$

We also have $U^{(2)}\xrightarrow{p}\frac{1}{T}\bar{B}_\infty+\frac{1}{N}\bar{D}_\infty$ from (C.5). The desired result then follows. ■

D Additional Details on Empirical Application

D.1 Data Sources and Sample Construction

Our empirical analysis is based on firm and market data following the framework of Ma (2019). The firm-level dataset combines information from Compustat, CRSP, IBES, TRACE, Datastream, and Mergent's Fixed Investment Securities Database (FISD). Flow variables such as net equity repurchases, net debt issuance, and capital expenditures are measured at the quarterly frequency and normalized by lagged total assets, while stock variables such as cash holdings and leverage are normalized by contemporaneous assets. Outliers are winsorized at the 1% level in all firm-level analyses. The sample excludes financial firms (SIC 6000–6999) and firms with missing bond or equity identifiers.

Firm-level bond variables are constructed from TRACE and FISD. The credit spread is the face-value-weighted difference between each bond's yield and the yield on the nearest-maturity Treasury, excluding convertible, asset-backed, and foreign currency bonds and those with less than one year to maturity. When yields are unavailable after November 2008, they are imputed from TRACE prices and FISD coupon information. The term spread is defined analogously using Treasury yields of different maturities. On the equity side, we construct the value-to-price (V/P) ratio following Dong, Hirshleifer, and Teoh (2012), combining Compustat book equity, CRSP prices, and IBES earnings forecasts.

Quarterly net equity repurchases are defined as share repurchases minus equity issuance (PRSTKC–SSTK), and net debt issuance as long-term debt issued minus retired (DLTIS–DLTR), each scaled by lagged assets. Additional controls include net income, cash holdings, capital expenditures, and deviations from target leverage, all from Compustat and defined as in Ma (2019). The aggregate bond yields and returns come from the Barclays Capital / Lehman Brothers indices compiled by Greenwood and Hanson (2013), updated with Bank of America Merrill Lynch data. Treasury yields are sourced from FRED, and stock market data is sourced from Robert Shiller's historical dataset.

To ensure comparability in the estimation of firm-specific effects, we focus on the period 2003-2024, when the coverage of TRACE becomes comprehensive. The earlier years used in Ma (2019), particularly the pre-2000 period, are highly unbalanced - especially for bond-level variables - and exhibit systematic patterns of missingness related to limited TRACE reporting and incomplete data integration across sources. We define an observation as missing if at least one of the key or control variables is unavailable for a given firm–quarter. Consequently, we restrict the sample to the period 2003Q1–2024Q4 and construct a high-coverage panel comprising 212 non-financial firms observed over 88 quarters. Firms are retained in the sample as long as they have at least 44 quarters of observed data.

Figure A1 visualizes the distribution of missing observations across firms and quarters in the final sample. Large blocks of missing observations near the sample edges primarily correspond to firms entering the panel mid-period or exiting early, while the few short gaps observed in the middle of the sample likely reflect random reporting noise rather than systematic non-coverage. Accordingly, we proceed under the assumption that the data are missing at random (MAR) in our empirical application.[A4]

[A4] Missing observations are common in empirical corporate-finance and asset-pricing panels. The objective of this paper is to achieve consistent estimation in high-coverage panels rather than to model systematic missingness. For broader treatments of structured or non-MAR missingness in financial panels, see Freyberger, Hoeppner, Neuhierl, and Weber (2025), Cahan, Bai, and Ng (2023), and Bryzgalova, Lerner, Lettau, and Pelger (2022). Extending the proposed framework to accommodate such forms of missingness is a promising direction for future research.

Figure A1 **Data availability across firms and quarters.** Each cell represents the availability of firm-level observations in the final estimation sample (212 firms, 2003Q1–2024Q4). White cells indicate available observations, and black cells denote missing quarters.

Our updated estimation algorithm explicitly handles these limited gaps under the assumption that missing observations are missing at random (MAR). The first-step estimator employs the Soft-Impute algorithm of Mazumder, Hastie, and Tibshirani (2010), while the second-step estimator inherently accommodates MAR data, following Bai (2009) and Chen, Fernández-Val, and Weidner (2021).

D.2 Additional Results

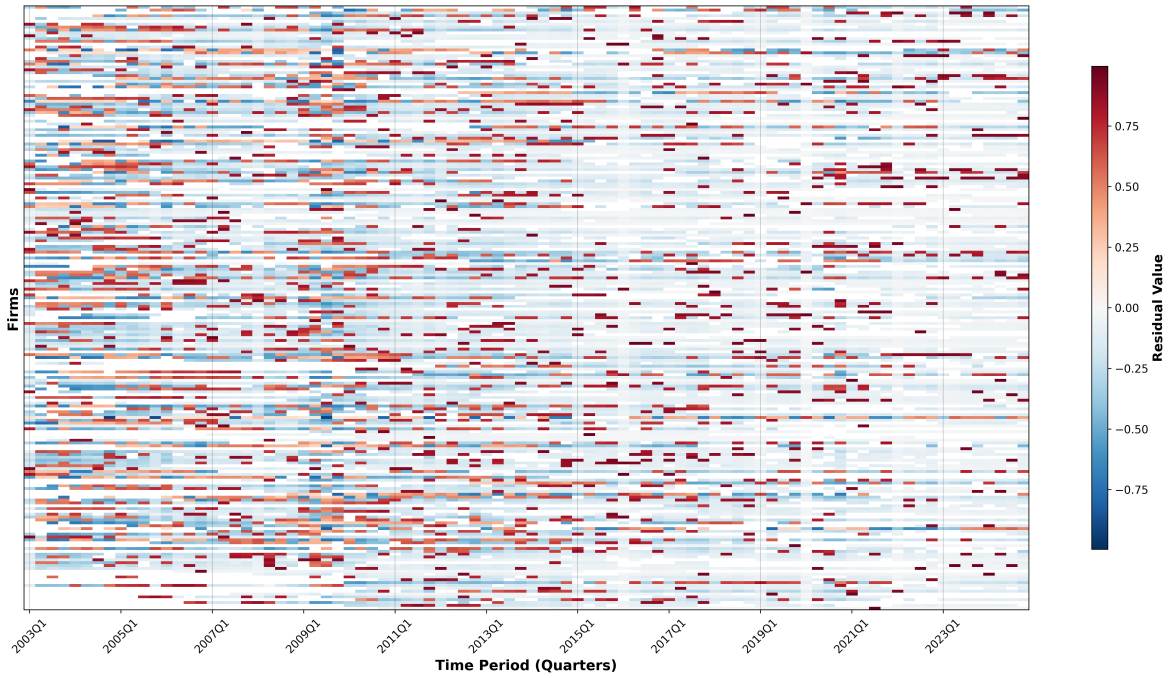

Figure A2 **Residual heatmap from TWFE estimation.** The figure plots firm–quarter residuals from the TWFE estimator following Fernández-Val and Weidner (2016).