5-2020

Essays on heterogeneous large panel data models

Ke MIAO
Singapore Management University, ke.miao.2015@phdecons.smu.edu.sg

Citation

MIAO, Ke. Essays on heterogeneous large panel data models. (2020). 1-199.
Available at: https://ink.library.smu.edu.sg/etd_coll/290

HETEROGENEOUS LARGE PANEL DATA MODELS

KE MIAO

SINGAPORE MANAGEMENT UNIVERSITY

2020

HETEROGENEOUS LARGE PANEL DATA MODELS

KE MIAO

A DISSERTATION

IN

ECONOMICS

Presented to the Singapore Management University in Partial Fulfilment

of the Requirements for the Degree of Doctor of Philosophy in Economics

2020

____

Supervisor of Dissertation

____

PhD in Economics, Programme Director

Heterogeneous Large Panel Data Models

by

Ke MIAO

Submitted to School of Economics in partial fulfilment of the requirements
for the Degree of Doctor of Philosophy in Economics

Dissertation Committee:

Liangjun Su (Supervisor/Chair)
Lee Kong Chian Professor of Economics
Singapore Management University

Peter Phillips
Sterling Professor of Economics & Professor of Statistics
Yale University
University of Auckland
University of Southampton
Singapore Management University

Sainan Jin
Professor of Economics
Singapore Management University

Qu Feng
Associate Professor of Economics
Nanyang Technological University

Singapore Management University
2020

Abstract

This dissertation consists of three papers which contribute to the estimation and inference theory of the heterogeneous large panel data models. The first chapter studies a panel threshold model with interactive fixed effects. The least-squares estimators in the shrinking-threshold-effect framework are explored. The inference theory on both slope coefficients and the threshold parameter is derived, and a test for the presence of the threshold effect is proposed. The second chapter considers the least-squares estimation of a panel structure threshold regression (PSTR) model, where parameters may exhibit latent group structures. Under some regularity conditions, the latent group structure can be correctly estimated with probability approaching one. A likelihood-ratio-based test on the group-specific threshold parameters is studied. Two specification tests are proposed: one tests whether the threshold parameters are homogeneous across groups, and the other tests whether the threshold effects are present. The third chapter studies high-dimensional vector autoregressions (VARs) augmented with common factors. An ℓ_1-nuclear-norm regularized estimator is considered. A singular value thresholding procedure is used to determine the correct number of factors with probability approaching one. Both a LASSO estimator and a conservative LASSO estimator are employed to improve estimation. The conservative LASSO estimates of the non-zero coefficients are shown to be asymptotically equivalent to the oracle least squares estimates. Monte Carlo studies are conducted to check the finite sample performance of the proposed test and estimators. Empirical applications are conducted in each chapter to illustrate the usefulness of the proposed methods.

Contents

Acknowledgement

Firstly, I would like to express my sincere gratitude to my supervisor Professor Liangjun Su, for providing guidance and feedback throughout my Ph.D. study. I appreciate all his contributions of time, ideas, and funding to make my Ph.D. experience productive and stimulating. His passion for both teaching and research was contagious and motivational for me. It has been an honor to be his student, and I could not have imagined having a better advisor.

My gratitude extends to Associate Professor Anthony Tay as the director of the Ph.D. program for his advice and support. I would also like to thank Qiu Ling Thor for providing excellent support in administrative issues.

I would like to thank my co-authors, Professor Kunpeng Li, Associate Professor Xun Lu, Professor Sainan Jin, Professor Peter C.B. Phillips, and Associate Professor Wendun Wang. They all have generously given me great comments and advice. I am lucky to learn from them during my early research career. I am particularly indebted to Professor Peter C.B. Phillips and Professor Sainan Jin. They kindly wrote a reference letter for me for my job seeking.

I will forever be thankful to family and friends. My special thanks to my parents and my life partner Liyao Li for their understanding, encouragement, companionship, and love. I consider myself nothing without them.

Chapter 1

Introduction

In the last two decades, there is a fast development of panel data econometrics, as panel data sets have become widely available to empirical researchers (see e.g. Hsiao 2014 and Pesaran 2015). A major advantage of using panel data is the ability to control for unobserved heterogeneity. With long panel data sets, we are able to identify and measure effects that are otherwise not detetable. Accompanied by the benefit, there is a cost. Technical difficulties arise when the models are built to capture various heterogeneity and relex exogeneity assumptions. Hence, it is an area of interest and importance.

In this dissertation, we extend the literature by analyzing three large hetergeneous panel data models. We have considered the unobserved heterogeneity due to three reasons: (1) threshold effect; (2) interactive fixed effects (IFEs), and (3) slope coefficients heterogeneity. All these three effects have received considerable attentions in recent empirical studies. The threshold effect of government debt on economic output has been well documented in the literature; see Reinhart and Rogoff (2010), Cecchetti et al. (2011), and Checherita-Westphal and Rother (2012), among others. Chudik et al. (2017) finds that the existing studies on the debt-growth nexus fail to incorporate cross-sectionally dependent errors that may exist across countries and consider modeling the cross-sectionally dependent errors via interactive fixed effects. In addition, Durlauf (2001), Su and Chen (2013), and Browning and Carro (2007) document that slope homogeneity among individuals is usually a vulnerable assumption.

In the second chapter, we propose to extend the panel threshold regression models by replacing the two-way fixed effects by the IFEs, which allow for the presence of a larger degree of unobserved heterogeneity and cross-sectional dependence. We have established the asymptotic theory for the least squares estimators and proposed a likelihood ratio test to make inference on the threshold parameter. The most challenging part of our analysis is to conduct a nonstandard analysis on the threshold parameter estimator jointly with large dimensional incidental parameters estimators. Due to the presence of large dimensional estimators, the analyses for consistency and convergence rate require some novel arguments that are quite different from the existing literature. We run Monte Carlo simulations to examine the finite sample performance of the LS estimators and the tests. To illustrate the usefulness of the proposed model, we consider an empirical application on the relationship between economic growth and financial development with the World Development Indicators (WDI) data.

In the third chapter, we propose a new panel threshold model that allows the slope and threshold coefficients to vary across individual units. We model individual heterogeneity via a grouped pattern, such that all the members within the same group share the same slope and threshold coefficients, whereas these coefficients can differ across groups in an arbitrary manner. We allow the group membership structure (i.e., which individuals belong to which group) to be unknown and estimated from the data. We refer to our model as a panel structure threshold regression (PSTR) model. To estimate the PSTR model, we consider a least-squares-type estimator that minimizes the sum of squared errors. Under some regularity conditions, we show that our estimators of the slope and threshold coefficients are asymptotically equivalent to the corresponding infeasible estimators of the group-specific parameters that are obtained by using individual group identity information. We evaluate the finite-sample performance of the proposed tests and estimation methods via extensive simulation studies. Our estimation method performs well in heterogeneous panels with threshold effects in finite samples. We illustrate the usefulness of our methods through two real-data examples. First, we revisit the relationship between capital market imperfections and firms' investment behavior. Next, we examine the impact of bank regulation, particularly branch deregulation, on income inequality in US, allowing observed and unobserved heterogeneity in their impact.

In the fourth chapter, we propose and study a high-dimensional vector autoregressions (VARs) model augmented with common factors (CFs) that allow for strong cross section dependence. To estimate the high dimensional VAR model with CFs, our approach uses a three-step procedure. The first step employs ℓ_1-nuclear norm regularized estimation that minimizes the sum of squared residuals with an ℓ_1-norm penalty on the transition matrices and a nuclear norm penalty on the low rank matrix representing the common component. In the second step, we include the estimated CFs as regressors and consider a generalized LASSO estimator to obtain an estimate of the transition matrices. We show that the estimation errors can be uniformly controlled, which facilitates the construction of weights for subsequent estimation by conservative LASSO in the third step. Under some regularity conditions, we show that this third step conservative LASSO estimator of the transition matrices achieves sign consistency (see Zhao and Yu 2006) asymptotically. Besides, the third step estimator of transition matrices, factors and factor loadings are asymptotically equivalent to the corresponding oracle least squares estimators that are obtained by using detailed information about the form of the true regression model. We illustrate the usefulness of this methodology through a real-data example. We revisit the financial connectedness measures proposed by Diebold and Yilmaz (2014) and document strong evidence of the existence of common factors in the volatilities of 23 sector exchange traded funds (ETFs).

Chapter five concludes and some technical results are provided in the appendix. Additional technical results can be found in the online supplement.

Chapter 2

Panel Threshold Models with Interactive Fixed Effects

2.1 Introduction

Both threshold effects and interactive fixed effects (IFEs) are of practical relevance and have received considerable attentions in recent empirical studies. In this paper, we propose a panel threshold model with IFEs, which includes both important effects in a model. The proposed model allows us to study threshold effects, IFEs, or both in a unified way. The proposed model has a wide range of applications. In a recent study, Chudik et al. (2017) investigate the debt-threshold effect on output. The threshold effect of government debt on economic output has been well documented in the literature; see Reinhart and Rogoff (2010), Cecchetti et al. (2011), and Checherita-Westphal and Rother (2012), among others. However, as argued by Chudik et al. (2017), the existing studies on the debt-growth nexus fail to incorporate cross-sectionally dependent errors that may exist across countries. This motivates them to consider a panel threshold model with heterogeneous coefficients and cross-sectionally dependent errors where the latter are modeled via the use of IFEs to deal with strong cross-sectional dependence. In this paper, we will consider another empirical example, the nexus of financial depth and economic growth where numerous studies have documented the presence of both threshold effects and unobserved heterogeneity where the latter is controlled via the use of one-way individual fixed effects or two-way additive fixed effects. In this paper, we propose to replace the two-way fixed effects by the IFEs, which allow for a larger degree of unobserved heterogeneity and cross-sectional dependence. It is interesting to know whether one can continue to find the evidence of threshold effects in the presence of IFEs. Using the World Development Indicators (WDI) data across 50 countries ranging from 1971 to 2015, we find strong evidence of threshold effects and IFEs, and the IFEs cannot be simplified into the two-way fixed effects. This confirms the necessity of incorporating the two effects into one model.

Apparently, the proposed model is related to two distinct branches of the econometrics literature, namely, the threshold models and the panel data models with IFEs. Threshold models can be traced back to Tong (1978), and have experienced substantial advancements over the last four decades. Early developments of threshold models focus much on fixed threshold effects. An undesirable consequence of this framework is that it is very difficult to conduct inference on the threshold value. As shown in Theorem 2 of Chan (1993), the limiting distribution of the least squares (LS) estimator of the threshold parameter is a functional of compound Poisson process, which involves many nuisance parameters such as the marginal distribution of the regressors and the regression coefficients. For this reason, most subsequent studies assume shrinking threshold effects to facilitate the inference in the threshold parameter. For example, Hansen (2000) develops a full statistical theory of the LS estimator for a linear cross-sectional regression model with threshold effects. Seo and Linton (2007) consider a smoothed LS estimation and establish the inferential theory for their semiparametric estimators in the framework of both shrinking and fixed threshold effects. As regard to panel threshold models, Hansen (1999) studies a static panel threshold model where the slope coefficient estimator is subject to the celebrated incidental parameter issue of Neyman and Scott

(1948). Dang et al. (2012) propose to apply the GMM technique to estimate a dynamic panel threshold model under the traditional large-N and short-T setup where N and T denote the number of cross-sectional units and the number of time periods, respectively. Ramírez-Rondán (2015) considers a similar model and advocates the use of maximum likelihood estimation. All the above studies assume that either the regressors or the threshold variable or both are exogenous. This assumption is restrictive in some empirical applications. To allow for endogenous regressors, in an empirical paper Kremer et al. (2013) estimate a dynamic panel threshold model by combining the forward orthogonal deviation transformation with the instrumental variable technique. In a recent paper, Seo and Shin (2016) propose a GMM method by extending the approaches of Hansen (1999, 2000) and Caner and Hansen (2004) to estimate a dynamic panel model with endogenous threshold variable and regressors. They show that if the threshold variable is endogenous, the estimator of the threshold parameter would lose super-consistency. It is worth mentioning that none of the above papers emphasize the issue of cross-sectional dependence within the data. If the cross-sectional dependence is not fully captured by model, the estimator would typically suffer inconsistency. There is a rapidly growing literature on panel data models with IFEs; see Bai (2009), Bai and Liao (2016), Moon and Weidner (2015, 2017), Li et al. (2016), Lu and Su (2016), among others. In a typical panel data model with IFEs, the unobserved errors are specified to have a factor structure, in which both factors and factor loadings can have arbitrary correlations with the regressors. This specification generalizes the traditional two-way additive scale form to a two-way multiplicative vector form. As a result, the IFEs model allows more richness of the unobserved heterogeneity that may vary across both time and individuals. Since the allowance and control of unobserved heterogeneity is one of most attractive features of panel data models, panel data model with IFEs becomes very appealing to empirical studies. Most of the studies on IFEs so far are limited to linear models; see, e.g., Bai (2009), Bai and Liao (2016), Moon and Weidner (2015, 2017). Other studies, such as Chen et al. (2014), consider the nonlinear models with IFEs but their analysis relies on the assumption of continuous differentiability of the objective function. In the proposed model, the nonlinear part is not differentiable at the change point, which greatly complicates the asymptotic analysis.

In this paper, we employ the least squares method to estimate the proposed model under the large-N and large-T setup. We consider the framework of diminishing threshold effects in the spirit of Hansen (2000). That is, the threshold effects shrink to zero as the sample size tends to infinity. We allow the regressors to be arbitrarily correlated with the IFEs provided that they have enough variations after projecting out the IFEs. We show that the threshold parameter can be estimated at a rate related to the magnitude of the threshold effect and the asymptotic distribution of the threshold parameter estimator is asymptotically pivotal up to a scale nuisance parameter. Under some regularity conditions, this rate, together with the shrinking rate of threshold effects, ensures the estimation of the threshold parameter to have no asymptotic effect on the estimation of slope coefficients. In other words, the slope coefficients can be estimated as if one knew the true threshold value. The most challenging part of our analysis is to conduct a nonstandard analysis on the threshold parameter estimator jointly with large dimensional incidental parameters estimators. Although a few previous studies, such as Hansen (1999) and Seo and Shin (2016), also have incidental parameters in their threshold models, the analyses in these models essentially only involve with low dimensional estimators because the incidental parameters can be concentrated out by the within-group transformation or first differencing, and the resultant objective function eventually depends on the finite dimensional regression coefficients and threshold parameter. This is in contrast with the current study in the high dimensional estimators are always present and have to be addressed throughout the whole analysis. Due to the presence of large dimensional estimators, the analyses for consistency and convergence rate require some novel arguments that are quite different from the existing literature. To see this point, we note that, in a standard threshold model or a panel data model with IFEs, the consistency can be established by only working with the objective function. But in the current model, the analysis is much complicated, and we take three steps to achieve this goal. In the first step, we work with the objective function to obtain the consistency of the estimators of all parameters but the threshold parameter, where the consistency of the large dimensional incidental parameters estimators is defined under some chosen norm invariant to rotational indeterminacy. In the second step, we work with the first order condition to derive some preliminary convergence rates for the slope coefficient estimators. In the third step, we work with the rescaled objective function to derive the consistency of the threshold-value estimator. Since the rescaled value is possibly large due to the fast shrinking threshold effects, the objective function in this step needs to be carefully chosen to offset the impact of the slow convergence rates of the incidental parameters estimators.[1] To the best of our knowledge, the methodology to prove consistency in this paper is new in the econometrics literature and can be useful to other discontinuous regression models in the presence of incidental parameters. For the convergence rate, a primary tool to deal with large dimensional parameters in panel data models with IFEs is the Cauchy-Schwarz inequality. Unfortunately, this tool does not provide useful bound when deriving the convergence rate of the threshold parameter estimator because of the special property of indicator function. Some new arguments are therefore developed in this paper to deal with this issue. In view of the fact that the unknown parameter in the limiting distri-

[1] In the M-estimation framework, if $\mathcal{L}(\theta)$ is the objective function of the unknown parameters θ that is to be maximized or minimized, then $\mathcal{L}(\theta) - c$ is also a valid one for any constant c. In classical analysis, c is suggested to be $\mathcal{L}(\theta^0)$ where θ^0 is the true values. However, due to the presence of large dimensional parameters, $\mathcal{L}(\theta) - \mathcal{L}(\theta^0)$ is not a good objective function in this paper.

bution of the threshold parameter estimator cannot be estimated accurately, we follow the lead of Hansen (2000) and propose a likelihood ratio (LR) test to facilitate inference on the threshold parameter. Again, since the estimators of the large dimensional incidental parameters would change under different threshold values, the analysis of the LR statistic is quite different from that in Hansen (2000). We find that the LR statistic is asymptotically pivotal in the case of conditional homoskedasticity. When conditional heteroskedasticity is present, the limiting distribution involves an unknown parameter that can be consistently estimated nonparametrically. We also consider the hypothesis testing on the presence of threshold effects, and propose a sup-Wald statistic. We also propose a procedure to obtain asymptotically correct critical values via simulations. We run Monte Carlo simulations to examine the finite sample performance of the LS estimators. For both dynamic and static models, the estimators are well-behaved in terms of asymptotic bias, standard deviation and coverage probability of the 95% confidence interval. Our slope coefficient estimators behave similarly to the infeasible estimators that are obtained when the threshold value is observed *a priori*. For the test of threshold effect, our simulations indicate that the rejection rate is close to the nominal level under the null hypothesis of the absence of threshold effects and the test has reasonable power under the alternative. In a nutshell, the simulations indicate that the LS estimators perform well in various data generating processes. To illustrate the usefulness of the proposed model, we consider an empirical application on the relationship between economic growth and financial development with the World Development Indicators (WDI) data. We find that both threshold effects and IFEs are present in the model, and that the financial development is beneficial to economic growth when it is below some threshold level and it harms growth otherwise. The former finding justifies the study of panel threshold model with IFEs and the latter is consistent with the conventional wisdom.

The outline of the paper is as follows. We introduce our model, discuss the estimation methods and list some basic assumptions in Section 2.2. We derive the asymptotic properties of these estimators in Section 2.3. We also study the likelihood ratio test on the threshold value and investigate several relevant issues associated with our model such as the threshold effect in the error variance, the determination of the number of factors, and the test of IFEs versus the two-way additive fixed effects in this section. We consider the hypothesis testing on the presence of threshold effect in Section 2.4. Section 2.5 reports the Monte Carlo simulation findings. Section 2.6 applies our method to study the relationship between economic growth and financial development. Section 2.7 concludes. The proofs of the main results in the paper are given in Appendix A. Additional materials can be found in the Online Supplemental of Miao et al. (2020a).

Notation. Let I_m denote an $m \times m$ identity matrix. For a real $m \times n$ matrix $A = (A_{ij})$, we use $\|A\|$ and $\|A\|_{\text{sp}}$ to denote its Frobenius norm and spectral norm, respectively. Let A' denote the transpose of A. When A has rank n, let $\mathbb{P}_A = A(A'A)^{-1}A'$ and $\mathbb{M}_A = I_m - \mathbb{P}_A$. When A is symmetric, we use $\mu_r(A)$ to denote its rth largest eigenvalue; $\mu_{\max}(A)$ and $\mu_{\min}(A)$ denote the largest and smallest eigenvalues of A, respectively. Let $\mathbf{1}\{\cdot\}$ be the indicator function. The symbol $\xrightarrow{p}$ denotes convergence in probability, $\xrightarrow{d}$ convergence in distribution, and *plim* probability limit. We use $(N, T) \to \infty$ to signify that N and T pass to infinity jointly.

2.2 Model, estimation method and assumptions

2.2.1 Model

Let N be the number of cross-sectional units and T the number of time periods. Consider the model

$$y_{it} = \beta^{0\prime} x_{it} + \delta^{0\prime} x_{it} d_{it}(\gamma^0) + \lambda_i^{0\prime} f_t^0 + e_{it}, \quad i = 1, ..., N, \; t = 1, ..., T, \tag{2.1}$$

where x_{it} is a $K \times 1$ vector of observable regressors, β^0 is a $K \times 1$ vector of slope coefficients, δ^0 is a $K \times 1$ vector of slope coefficients representing the threshold effect, γ^0 is a scalar threshold coefficient, $d_{it}(\gamma) \equiv \mathbf{1}\{q_{it} \leq \gamma\}$, q_{it} is a scalar threshold variable, λ_i^0 is an $R^0 \times 1$ vector of unobserved factor loadings, f_t^0 is an $R^0 \times 1$ vector of unobserved common factors, and e_{it} is the idiosyncratic error term. Throughout the paper, we use the superscript zero to signify the true parameter value. We use f_{tr}^0 and λ_{ir}^0 to denote the rth component of f_t^0 and λ_i^0, respectively, where $r = 1, \ldots, R^0$. We assume $\gamma^0 \in \Gamma \equiv [\underline{\gamma}, \overline{\gamma}]$, where $\underline{\gamma}$ and $\overline{\gamma}$ are two fixed constants. Following Hansen (2000), we consider the shrinking threshold effect framework by assuming that $\delta^0 \equiv \delta_{NT}^0 \to 0$ with the convergence rate specified below as $(N, T) \to \infty$.

Let $\Lambda^0 \equiv (\lambda_1^0, \ldots, \lambda_N^0)'$, $F^0 \equiv (f_1^0, \ldots, f_T^0)'$, $e_t = (e_{1t}, \ldots, e_{Nt})'$, $Y_t \equiv (y_{1t}, \ldots, y_{Nt})'$, $X_t \equiv (x_{1t}, \ldots, x_{Nt})'$ and $X_t(\gamma) \equiv (x_{1t} d_{1t}(\gamma), \ldots, x_{Nt} d_{Nt}(\gamma))'$. We can write the model in (2.1) in vector form

$$Y_t = X_t \beta^0 + X_t(\gamma^0)\delta^0 + \Lambda^0 f_t^0 + e_t = X_{t,\gamma^0} \theta^0 + \Lambda^0 f_t^0 + e_t, \tag{2.2}$$

where $t = 1, \ldots, T$, $\theta^0 \equiv (\beta^{0\prime}, \delta^{0\prime})'$ and $X_{t,\gamma} \equiv (X_t, X_t(\gamma))$.

For the moment we assume that the true number of factors R^0 is known and given by R. In Section 2.3.6 below, we propose a way to consistently estimate the number of factors. In factor analysis, it is well known that Λ and

F can only be identified up to a rotation. We follow Bai and Ng (2002) and Bai (2003) and consider the following set of identification restrictions:

(i) $\Lambda'\Lambda/N = I_R$, and (ii) $F'F$ is a diagonal matrix with diagonal elements ordered in descending order.

2.2.2 Estimation method

Given R, we can concentrate out the $T \times R$ matrix F and obtain the following Gaussian QML estimate of $(\theta, \Lambda, \gamma)$:

$$(\widehat{\theta}, \widehat{\Lambda}, \widehat{\gamma}) \equiv \operatorname{argmin}_{(\theta,\Lambda,\gamma)\in\mathbb{R}^{2K}\times\mathbb{L}\times\Gamma}\mathcal{L}(\theta, \Lambda, \gamma),$$

where

$$\mathcal{L}(\theta, \Lambda, \gamma) \equiv \sum_{t=1}^{T}(Y_t - X_{t,\gamma}\theta)'\mathbb{M}_\Lambda(Y_t - X_{t,\gamma}\theta), \tag{2.3}$$

$(\theta, \Lambda, \gamma) \in \mathbb{R}^{2K} \times \mathbb{L} \times \Gamma$ and $\mathbb{L} \equiv \{\Lambda \in \mathbb{R}^{N\times R} : \Lambda'\Lambda/N = I_R\}$.

The above minimization problem can be solved in two steps:

(i) In the first step, we keep γ fixed so that the objective function in (2.3) can be minimized as in Bai (2009), Moon and Weidner (2015, 2017) and Lu and Su (2016) to obtain the estimate $(\widehat{\theta}(\gamma), \widehat{\Lambda}(\gamma))$. Let $\mathcal{L}^*(\gamma) \equiv \mathcal{L}(\widehat{\theta}(\gamma), \widehat{\Lambda}(\gamma), \gamma)$.

(ii) In the second step, one can search over the interval $\Gamma \equiv [\underline{\gamma}, \overline{\gamma}]$ to minimize $\mathcal{L}^*(\gamma)$.

Because $\mathcal{L}^*(\gamma)$ is a step function that takes on less than NT distinct values, we can follow Hansen (2000) and search for γ over $\Gamma_n = \Gamma \cap \{q_{it}, 1 \le i \le N, 1 \le t \le T\}$. When NT is large, we can approximate Γ by grid of n points for some $n \le NT$. For example, let $q_{(j)}$ denote the $(\eta_0 + \frac{j-1}{n-1}(1 - 2\eta_0))$-th quantile of the sample $\{q_{it},\ 1 \le i \le N, 1 \le t \le T\}$ for $j = 1, \dots, n$ and $\bar{\Gamma}_n = \{q_{(1)}, \dots, q_{(n)}\}$. We then define $\widehat{\gamma}_n = \operatorname{argmin}_{\gamma\in\bar{\Gamma}_n}\mathcal{L}^*(\gamma)$, which will provide a good approximation to $\widehat{\gamma}$. Hansen (1999) recommends choosing $\eta_0 = 1\%$ or

5%. Given $\widehat{\gamma}$, the estimates of θ and Λ are calculated according to $\widehat{\theta} \equiv \widehat{\theta}(\widehat{\gamma})$ and $\widehat{\Lambda} \equiv \widehat{\Lambda}(\widehat{\gamma})$, respectively. Once the estimates of θ, γ, and Λ are obtained, the estimate of F can be constructed by the plug-in method (see Section 2.3).

Remark 2.1. This paper adopts the approach to concentrate out the factors first under the identification restrictions stated at the end of Section 2.1. One can alternatively consider concentrating out the factor loadings under the identification restrictions that $F'F = I_R$ and $\Lambda'\Lambda$ is a diagonal matrix with descending diagonal elements. The estimates of θ and γ under the two sets of identification restrictions would be the same, and so are their asymptotic distributions. In this paper, we consider the data scenario of $N/T \to \kappa$, as $(N,T) \to \infty$. Under this scenario, it does not make a big difference in computational time for the two concentration strategies. In a more general case where N and T diverge at different rates, it is desirable to concentrate out the larger dimension matrix.

Remark 2.2. As emphasized in the introduction, the objective function $\mathcal{L}(\theta, \Lambda, \gamma)$ is non-differentiable with respect to γ, which would have significant consequence on the asymptotic properties of the LS estimators. Alternatively, one can adopt the idea of Seo and Linton (2007) and use a smoothed objective function to estimate the model. Let $\mathscr{K}(\cdot)$ be a bounded function such that $\lim_{s\to-\infty} \mathscr{K}(s) = 0$ and $\lim_{s\to\infty} \mathscr{K}(s) = 1$. The smoothed objective function is defined as

$$\mathcal{L}^s(\theta, \Lambda, F, \gamma, h) = \sum_{i=1}^{N}\sum_{t=1}^{T}\left[y_{it} - x_{it}'\beta - x_{it}'\delta\mathscr{K}(\frac{\gamma - q_{it}}{h}) - \lambda_i' f_t\right]^2,$$

where $h = o(1)$ is a bandwidth parameter. The estimators of θ and γ can be obtained by minimizing the above function. This estimation method has the benefit that one can analyze the estimators in a unified way, irrespective of the fixed threshold effects or shrinking threshold effects; see Seo and Linton (2007) for details. But the computation burden becomes larger and one has to determine h.

Remark 2.3. In this paper, we do not consider the case in which the regressor x_{it} is correlated with the idiosyncratic error e_{it}. When the correlation is present, we need to apply the instrumental variable (IV) method to estimate the model. For simplicity, suppose that all the regressors are endogenous, but we have d_w-dimensional instrumental variables w_{it} with $d_w \geq K$. Let $W_{t,\gamma}$ be defined the same as $X_{t,\gamma}$. The estimation consists of two steps. In this first step, for each given θ and γ, we estimate $\widehat{\phi}(\theta, \gamma)$ and $\widehat{\Lambda}(\theta, \gamma)$ according to

$$\left(\widehat{\phi}(\theta,\gamma), \widehat{\Lambda}(\theta,\gamma)\right) = \underset{(\phi,\Lambda)\in\Phi\times\mathbb{L}}{\operatorname{argmax}} \sum_{t=1}^{T} \left(Y_t - X_{t,\gamma}\theta - W_{t,\gamma}\phi\right)' \mathbb{M}_{\Lambda} \left(Y_t - X_{t,\gamma}\theta - W_{t,\gamma}\phi\right),$$

where Φ is some compact parameter space for ϕ. In the second step, we obtain the estimator $\widehat{\theta}$ and $\widehat{\gamma}$ by

$$(\widehat{\theta}, \widehat{\gamma}) = \underset{(\theta,\gamma)\in\Theta\times\Gamma}{\operatorname{argmax}} \widehat{\phi}(\theta,\gamma)' \mathbb{W}_{NT}^{-1} \widehat{\phi}(\theta,\gamma).$$

where $\mathbb{W}_{NT}$ is a weighting matrix that can be chosen as a consistent estimate of $\operatorname{var}(\widehat{\phi}(\theta^0, \gamma^0))$ based on some preliminary consistent estimate $(\tilde{\theta}, \tilde{\gamma})$ of (θ, γ). Intuitively, if θ and γ are chosen at their true values, the estimator $\widehat{\phi}$ in the first step would be very close to 0. So by minimizing the weighted norm of $\widehat{\phi}$ in the second step, we would obtain the consistent estimators for θ^0 and γ^0. Such an estimation idea has been used by Chernozhukov and Hansen (2008) and Su and Hoshino (2016) in the IV quantile model, and by Lee et al. (2012) and Moon et al. (2018) in the IFEs framework.

We end this subsection with two equations, which serve as the bases for our asymptotic analyses. By definition, $(\widehat{\theta}, \widehat{\Lambda}) = \operatorname{argmin}_{(\theta,\Lambda)\in\mathbb{R}^{2K}\times\mathbb{L}} \mathcal{L}(\theta, \Lambda, \widehat{\gamma})$. This implies that given $\widehat{\gamma}$, the estimates $\widehat{\theta}$ and $\widehat{\Lambda}$ solve the following system of equations

$$\widehat{\theta} = \left(\sum_{t=1}^{T} X'_{t,\widehat{\gamma}} \mathbb{M}_{\widehat{\Lambda}} X_{t,\widehat{\gamma}}\right)^{-1} \sum_{t=1}^{T} X'_{t,\widehat{\gamma}} \mathbb{M}_{\widehat{\Lambda}} Y_t \tag{2.4}$$

and

$$\left[\frac{1}{NT} \sum_{t=1}^{T} (Y_t - X_{t,\widehat{\gamma}}\widehat{\theta})(Y_t - X_{t,\widehat{\gamma}}\widehat{\theta})'\right] \widehat{\Lambda} = \widehat{\Lambda} V_{NT}, \tag{2.5}$$

where V_{NT} is a diagonal matrix whose diagonal elements are the R largest eigenvalues of the matrix in the square brackets, arranged in decreasing order.

2.2.3 Assumptions

We first introduce some notations for ease of exposition. Let $x_{k,it}$ denote the kth element of x_{it}. Let $\mathbf{X}_k$ and $\mathbf{X}_k(\gamma)$ be two $N \times T$ matrices with (i,t)th entry being $x_{k,it}$ and $x_{k,it}d_{it}(\gamma)$, respectively. Define

$$\mathbf{X}_{k,\gamma} = \begin{cases} \mathbf{X}_k & \text{if } k \leq K, \\ \mathbf{X}_{k-K}(\gamma) & \text{if } K < k \leq 2K \end{cases}.$$

Define $\mathbf{X} \equiv (\mathbf{X}_1, \ldots, \mathbf{X}_K)$, $\mathbf{X}(\gamma) \equiv (\mathbf{X}_1(\gamma), \ldots, \mathbf{X}_K(\gamma))$ and $\mathbf{X}^*_\gamma \equiv (\mathbf{X}, \mathbf{X}(\gamma))$. Then we can rewrite the model (2.1) in matrix form:

$$\mathbf{Y} = \beta^0 \odot \mathbf{X} + \delta^0 \odot \mathbf{X}(\gamma^0) + \Lambda^0 F^{0\prime} + \mathbf{e},$$
$$\text{where } \beta^0 \odot \mathbf{X} \equiv \sum_{k=1}^{K} \beta^0_k \mathbf{X}_k, \quad \delta^0 \odot \mathbf{X}(\gamma^0) \equiv \sum_{k=1}^{K} \delta^0_k \mathbf{X}_k(\gamma^0),$$

and $\mathbf{Y}$ and $\mathbf{e}$ are $N \times T$ matrices with the (i,t)th entry y_{it} and e_{it}, respectively. It can be readily shown that $\widehat{\Lambda}$ is $\sqrt{N}$ times the first R left-singular vectors of the matrix $\mathbf{Y} - \widehat{\beta} \odot \mathbf{X} - \widehat{\delta} \odot \mathbf{X}(\widehat{\gamma})$. According the PC method, the estimator $\widehat{F}$ is given by $\widehat{F} = \big[\mathbf{Y} - \widehat{\beta} \odot \mathbf{X} - \widehat{\delta} \odot \mathbf{X}(\widehat{\gamma})\big]'\widehat{\Lambda}/N$.

With the above symbols, we further introduce the following notations that will be used in the subsequent assumptions. Define $NT \times 2K$ matrix $\mathcal{Z}(\Lambda, \gamma)$ as

$$\mathcal{Z}(\Lambda,\gamma) \equiv \begin{bmatrix} \mathcal{Z}_1(\Lambda,\gamma) \\ \mathcal{Z}_2(\Lambda,\gamma) \\ \vdots \\ \mathcal{Z}_T(\Lambda,\gamma) \end{bmatrix} = (\mathbb{M}_{F^0} \otimes \mathbb{M}_\Lambda) \begin{bmatrix} X_1 & X_1(\gamma) \\ X_2 & X_2(\gamma) \\ \vdots & \vdots \\ X_T & X_T(\gamma) \end{bmatrix} = (\mathbb{M}_{F^0} \otimes \mathbb{M}_\Lambda)\,(\mathrm{vec}(\mathbf{X}_{1,\gamma}), \ldots, \mathrm{vec}(\mathbf{X}_{2K,\gamma})),$$

where $\mathcal{Z}_k(\Lambda,\gamma)$, the k-th column of $\mathcal{Z}(\Lambda,\gamma)$, is equal to $\mathrm{vec}(\mathbb{M}_\Lambda \mathbf{X}_{k,\gamma} \mathbb{M}_{F^0})$ by the identity $\vec{(ABC)} = (C' \otimes A)\vec{(B)}$ and "$\otimes$" denotes the Kronecker product. From this result, it is obvious that $\mathcal{Z}_k(\Lambda,\gamma)$ has smaller variation compared to $\mathrm{vec}(\mathbf{X}_{k,\gamma})$. The matrix $\mathcal{Z}(\Lambda,\gamma)$ plays an important role in the identification of θ^0; see Assumption A.1 below. Let $X_t(\gamma_a,\gamma_b) \equiv X_t(\gamma_a) - X_t(\gamma_b)$ for $\gamma_a, \gamma_b \in \Gamma$.

Similar to $\mathcal{Z}(\Lambda,\gamma)$, we define an $NT \times K$ matrix:

$$\widetilde{\mathcal{Z}}(\Lambda,\gamma) \equiv (\mathbb{M}_{F^0} \otimes \mathbb{M}_{\Lambda}) \begin{bmatrix} X_1(\gamma,\gamma^0) \\ X_2(\gamma,\gamma^0) \\ \vdots \\ X_T(\gamma,\gamma^0) \end{bmatrix}.$$

The matrix $\widetilde{\mathcal{Z}}(\Lambda,\gamma)$ is related to the identification of γ^0.

Let $\mathcal{D} \equiv \sigma(F^0,\Lambda^0)$, the minimal sigma-field generated from F^0 and Λ^0, and $\Pr_{\mathcal{D}}(\cdot) \equiv \Pr(\cdot|\mathcal{D})$ and $E_{\mathcal{D}}(\cdot) \equiv E(\cdot|\mathcal{D})$. Let $\mathcal{F}_{NT,t} \equiv \sigma(\{(x_{it},q_{it},e_{i,t-1}),(x_{i,t-1},q_{i,t-1},e_{i,t-2}),\dots\}_{i=1}^{N})$. Let $f_{it,\mathcal{D}}(\gamma)$ denote the probability density function (PDF) of q_{it} conditional on $\mathcal{D}$, and $E_{\mathcal{D}}(\cdot|\gamma) \equiv E_{\mathcal{D}}(\cdot|q_{it}=\gamma)$. Let M denote a generic positive constant that may vary across places. We make the following assumptions for the asymptotic analysis.

Assumption A.1. There exists some constant $\tau > 0$ such that, as $(N,T) \to \infty$,

(i) $\Pr(\min_{(\Lambda,\gamma)\in\mathbb{L}\times\Gamma} \mu_{\min}[\mathcal{B}(\Lambda,\gamma)] \geq \tau) \to 1$, where $\mathcal{B}(\Lambda,\gamma) = \frac{1}{NT}\mathcal{Z}(\Lambda,\gamma)'\mathcal{Z}(\Lambda,\gamma)$;

(ii) $\Pr(\min_{\gamma\in\Gamma} \mu_{\min}[\mathcal{I}(\gamma)] \geq \tau \min\{1,|\gamma-\gamma^0|\}) \to 1$, where

$$\mathcal{I}(\gamma) = \frac{1}{NT}\widetilde{\mathcal{Z}}(\Lambda^0,\gamma)'\mathbb{M}_{\mathcal{Z}(\Lambda^0,\gamma)}\widetilde{\mathcal{Z}}(\Lambda^0,\gamma) \text{ with,}$$

$$\mathbb{M}_{\mathcal{Z}(\Lambda^0,\gamma)} = I_{NT} - \mathcal{Z}(\Lambda^0,\gamma)[\mathcal{Z}(\Lambda^0,\gamma)'\mathcal{Z}(\Lambda^0,\gamma)]^{-1}\mathcal{Z}(\Lambda,\gamma)'.$$

Assumption A.2. (i) $E_{\mathcal{D}}(e_{it}^{8+\epsilon})$ and $E_{\mathcal{D}}(\|x_{it}\|^{8+\epsilon})$ are uniformly bounded by a non-random constant for some constant $\epsilon > 0$.

(ii) $E\left\|f_t^0\right\|^8 \leq M$ and $\frac{1}{T}\sum_{t=1}^{T} f_t^0 f_t^{0\prime} \xrightarrow{p} \Sigma_f > 0$ for some $R \times R$ matrix Σ_f as $T \to \infty$;

(iii) $E\left\|\lambda_i^0\right\|^8 \leq M$ and $\frac{1}{N}\sum_{i=1}^{N} \lambda_i^0 \lambda_i^{0\prime} \xrightarrow{p} \Sigma_\lambda > 0$ for some $R \times R$ matrix Σ_λ as $N \to \infty$;

(iv) $\|\mathbf{e}\|_{\text{sp}} = O_p(\sqrt{N}+\sqrt{T})$.

Assumption A.3. The threshold effect δ^0 satisfies that $\delta^0 = (NT)^{-\alpha} C^0$ for some $\alpha \in (0, 1/2)$, $C^0 \in \mathbb{R}^K$ and $C^0 \neq 0$.

Assumption A.4. (i) For each $i = 1, \ldots, N$, $\{(x_{it}, q_{it}, e_{it}) : t = 1, 2, \ldots\}$ is conditional strong mixing given $\mathcal{D}$ with the mixing coefficients $\{\alpha^{\mathcal{D}}_{NT,i}(\cdot)\}$; $\alpha_m \equiv \sup_{1 \leq i \leq N} \alpha^{\mathcal{D}}_{NT,i}(m)$ satisfies that $\alpha_m = O(m^{-\zeta})$ for some $\zeta > \frac{12p}{4p-1}$ and $p > 4$;

(ii) (x_{it}, q_{it}, e_{it}), $i = 1, \ldots, N$, are mutually independent of each other conditional on $\mathcal{D}$;

(iii) For each $i = 1, \ldots, N$, $E(e_{it}|\mathcal{D} \vee \mathcal{F}_{NT,t}) = 0$ a.s.;

(iv) There exists a constant $c_f < \infty$ such that $\max_{i,t} \sup_{\gamma \in \Gamma} f_{it,\mathcal{D}}(\gamma) < c_f$;

(v) There exist $\mathcal{D}$-dependent variables $M_{it,\mathcal{D}}$ such that $\sup_{\gamma \in \Gamma} E_{\mathcal{D}}(\|x_{it}\|^4 | q_{it} = \gamma) \leq M_{it,\mathcal{D}}$, $\sup_{\gamma \in \Gamma} E_{\mathcal{D}}(\|x_{it} e_{it}\|^4 | q_{it} = \gamma) \leq M_{it,\mathcal{D}}$. We have

$$\Pr(\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} M^2_{it,\mathcal{D}} < M^2) \to 1$$

for some $M < \infty$ as $(N, T) \to \infty$.

Assumption A.1 is an identification condition for θ and γ. Assumption A.1(i) extends Assumption A in Bai (2009) to require the non-colinearity of the regressors uniformly over $\gamma \in \Gamma$. More specifically, it requires that the residuals from the linear projections of any linear combination of the form $\alpha^* \odot \mathbf{X}^*_\gamma$ on the column space of F^0 first and then on that of Λ^0 be asymptotically nondegenerate with $\alpha^* \in \mathbb{R}^{2K}$. Assumption A.1(ii) requires that $\widetilde{\mathcal{Z}}(\Lambda^0, \gamma)$ and $\mathcal{Z}(\Lambda^0, \gamma)$ be not colinear uniformly. To gain more intuitions of these two conditions, consider model (2.2), which is equivalent to

$$\mathbf{Y} = \sum_{k=1}^{K} \mathbf{X}_k \beta^0_k + \sum_{k=1}^{K} \mathbf{X}_k(\gamma^0) \delta^0 + \Lambda^0 F^{0\prime} + \mathbf{e}.$$

As pointed out in Bai (2009), the large dimensional nuisance parameters Λ^0 and F^0 are eliminated through projection matrices in the least squares estimation.

Following this intuition, we therefore premultiply $\mathbb{M}_{\Lambda^0}$ and postmultiply $\mathbb{M}_{F^0}$ on both sides to obtain

$$\mathbb{M}_{\Lambda^0}\mathbf{Y}\mathbb{M}_{F^0} = \sum_{k=1}^{K}\mathbb{M}_{\Lambda^0}\mathbf{X}_k\mathbb{M}_{F^0}\beta_k^0 + \sum_{k=1}^{K}\mathbb{M}_{\Lambda^0}\mathbf{X}_k(\gamma^0)\mathbb{M}_{F^0}\delta_k^0 + \mathbb{M}_{\Lambda^0}\mathbf{e}\mathbb{M}_{F^0}.$$

Taking vector operation on both sides and using the definition of $\mathcal{Z}(\Lambda, \gamma)$, we have

$$\begin{aligned}(\mathbb{M}_{F^0} \otimes \mathbb{M}_{\Lambda^0})\mathrm{vec}(\mathbf{Y}) &= (\mathbb{M}_{F^0} \otimes \mathbb{M}_{\Lambda^0})\left[\sum_{k=1}^{K}\mathrm{vec}(\mathbf{X}_k)\beta_k^0 + \sum_{k=1}^{K}\mathrm{vec}(\mathbf{X}_k(\gamma^0))\delta_k^0 + \mathrm{vec}(\mathbf{e})\right] \\ &= \mathcal{Z}(\Lambda^0, \gamma^0)\theta^0 + (\mathbb{M}_{F^0} \otimes \mathbb{M}_{\Lambda^0})\mathrm{vec}(\mathbf{e}). \end{aligned} \tag{2.6}$$

If we treat equation (2.6) as a linear regression model on $\theta^0 = (\beta^{0\prime}, \delta^{0\prime})'$, it is natural to impose the full column rank assumption on $\mathcal{Z}(\Lambda^0, \gamma^0)$ to identify the parameter θ^0, which is equivalent to assuming $\Pr(\mu_{\min}[\mathcal{B}(\Lambda^0, \gamma^0)] \geq \tau) \to 1$. In our model, Λ^0 and γ^0 are both unobserved and simultaneously estimated with θ^0. So one may expect that $\Pr(\mu_{\min}[\mathcal{B}(\widehat{\Lambda}, \widehat{\gamma})] \geq \tau) \to 1$ still holds. Since $\widehat{\Lambda}$ and $\widehat{\gamma}$ do not have the consistency property so far, and can be any values in the parameter space, this motivates us to impose Assumption A.1(i). To understand Assumption A.1(ii), we first consider Hansen (2000)'s model: $Y = X\beta^0 + X_{\gamma^0}\delta^0 + e$ where the definitions of notations are self-evident. We note that the identification condition for γ^0 in the Hansen's model is that the function

$$\mathcal{F}(\gamma) = \operatorname*{plim}_{N\to\infty}\mu_{\min}\Big(\frac{1}{N}X_{\gamma^0}'\mathbb{M}_{X_\gamma}X_{\gamma^0}\Big) = \operatorname*{plim}_{N\to\infty}\mu_{\min}\Big(\frac{1}{N}(X_\gamma - X_{\gamma^0})'\mathbb{M}_{X_\gamma}(X_\gamma - X_{\gamma^0})\Big)$$

has the minimum value 0 at γ^0. In the transformed model (2.6), the matrices $\widetilde{\mathcal{Z}}(\Lambda^0, \gamma)$ and $\mathcal{Z}(\Lambda^0, \gamma)$ play the same roles as $X_\gamma - X_{\gamma^0}$ and X_γ in Hansen's model. We therefore impose Assumption A.1(ii) to guarantee the identification of γ^0. Note that Assumption A.1(ii) is equivalent to the condition that the matrix $\mathcal{I}(\gamma)$ has minimum value 0 at γ^0. More specifically, if γ falls in some neighborhood of γ^0, the minimum eigenvalue behaves like the function $f(\gamma) = |\gamma - \gamma^0|$ which achieves 0 at γ^0, and if γ falls outside of this neighborhood, the minimum eigenvalue is always greater than some $\tau > 0$. In Hansen

(2000), he gives some more primitive conditions to identify γ^0. However, this is feasible because of the special property that $X'_\gamma X_{\gamma^0} = X'_{\gamma^0} X_{\gamma^0}$ for $\gamma \geq \gamma^0$ and $X'_\gamma X_{\gamma^0} = X'_\gamma X_\gamma$ for $\gamma < \gamma^0$ which only holds in his linear model. In model (2.6), we generally do not have $\mathscr{Z}^\dagger(\Lambda^0, \gamma)' \mathscr{Z}^\dagger(\Lambda^0, \gamma^0) = \mathscr{Z}^\dagger(\Lambda^0, \gamma^0)' \mathscr{Z}^\dagger(\Lambda^0, \gamma^0)$ for $\gamma \geq \gamma^0$ due to the presence of large dimensional incidental parameters, where $\mathscr{Z}^\dagger(\Lambda, \gamma) = (\mathbb{M}_{F^0} \otimes \mathbb{M}_\Lambda)\big[\vec{(}\mathbf{X}_1(\gamma)), \ldots, \vec{(}\mathbf{X}_k(\gamma))\big]$. So it seems infeasible to give more primitive conditions like Hansen (2000) in this paper.

Assumption A.2(i)-(iii) imposes some moment conditions on the regressors, error terms, factors and factor loadings. Note that we only consider strong factors here. Assumption A.2 (iv) is frequently assumed in the literature; see, e.g., Su and Chen (2013) and Moon and Weidner (2015). Assumption A.3 specifies a diminishing threshold effect. The same assumption is also imposed in other studies; see Hansen (1999), Caner and Hansen (2004), among others.

Assumption A.4 is similar to Assumption A.2 of Su and Chen (2013). We assume conditional strong mixing across t in Assumption A.4(i) and conditional independence across i in Assumption A.4(ii). As Su and Chen (2013) remark, the conditional strong mixing in Assumption A.4(i) can be replaced by the unconditional strong mixing if we assume that the factor loadings are nonrandom. Assumption A.4(ii) does not rule out the possibility of *unconditional* cross-sectional dependence among $\{x_{it}, q_{it}, e_{it}\}$ arising from the common factors. The martingale difference sequence (m.d.s.) condition in Assumption A.4(iii) simplifies the asymptotic analysis. It allows for conditional heteroskedasticity, skewness, or kurtosis of unknown form in e_{it} but rules out serial correlations. Note that Assumption A.4(iii) does allow for the presence of lagged dependent or independent variables. If serial correlations are suspected to exist among errors, we can add lagged dependent or independent variables into the model to remove them. So this assumption is not restrictive. Assumption A.4 (iv)-(v) imposes some conditions on the conditional PDF and moments of x_{it} given $\mathcal{D}$. Assumption A.4(iv) assumes the conditional PDF of

q_{it} is uniformly bounded; Assumption A.4(v) assumes that the fourth order conditional moments of x_{it} and $x_{it}e_{it}$ are well behaved. Note that $M_{it,\mathcal{D}}$ is not assumed to be bounded uniformly over (i,t) but its sample second moment is well behaved. Therefore Assumption A.4(v) is not restrictive.

2.3 Asymptotic Property

In this section, we study the asymptotic properties of the estimators. We first establish the consistency of $(\widehat{\theta}, \widehat{\Lambda}, \widehat{\gamma})$ and their convergence rates, derive the asymptotic distributions of $\widehat{\theta}$ and $\widehat{\gamma}$, and consider the statistical inference on γ based on a likelihood ratio test statistic. Then we investigate several relevant issues associated with our model such as the threshold effect in the error variance, the determination of the number of factors, and the test of IFEs versus the two-way additive fixed effects.

2.3.1 Consistency

This subsection establishes the consistency of the LS estimators defined in Section 2.2.2. We achieve this goal in three steps. We first show the consistency of $\widehat{\theta}$ and $\widehat{\Lambda}$ in Theorem 2.1. With consistency, we next give a preliminary convergence rate of $\widehat{\theta}$, which is given in Proposition A.2 in the appendix. Based on this convergence rate, we finally establish the consistency of $\widehat{\gamma}$ in Theorem 2.2.

Theorem 2.1. ***(Consistency of*** $\widehat{\theta}$ ***and*** $\widehat{\Lambda}$***)*** *Suppose that Assumptions A.1-A.4 hold. Then*

(i) $\widehat{\theta} - \theta^0 \xrightarrow{p} 0$;

(ii) The matrix $N^{-1}\Lambda^{0\prime}\widehat{\Lambda}$ *is invertible and* $\|\mathbb{P}_{\widehat{\Lambda}} - \mathbb{P}_{\Lambda^0}\| = o_p(1)$.

Theorem 2.1 establishes the consistency of the estimators $\widehat{\theta}$ and $\widehat{\Lambda}$, which is analogous to Proposition 1 in Bai (2009). It provides the basis for subsequent analyses. For example, with the consistency of $\widehat{\theta}$, the terms involved with $\|\widehat{\theta} -$

$\theta^0\|^2$ are of smaller order than $\|\widehat{\theta} - \theta^0\|$ and become asymptotically negligible.

The following theorem shows the consistency of $\widehat{\gamma}$. The proof of this theorem requires considerable amount of work. To appreciate the difficulty, we note that the estimators of the large dimensional incidental parameters (i.e., λ_i's or f_t's) have slow convergence rates $N^{-1/2}$ or $T^{-1/2}$. However, Assumption A.3 specifies an $(NT)^{-\alpha}$ threshold effect, so we have to multiply $(NT)^{2\alpha}$ on the objective function $\mathcal{L}(\theta, \Lambda, \gamma)$ to obtain a non-shrinking threshold effect. As α is close to 1/2 from the below, $(NT)^{2\alpha}$ is slightly smaller than NT. This gives rise to a challenging issue: the estimation errors coming from incidental parameters cause serious problems to our analyses because of their slow convergence rates. In contrast, under the fixed threshold effect framework, the proof of consistency for $\widehat{\gamma}$ is much easier as the normalization scale over there is equal to 1.

Theorem 2.2. ***(Consistency of*** $\widehat{\gamma}$***)*** *Under Assumptions A.1-A.4, with* $N/T \to \kappa > 0$ *as* $(N, T) \to \infty$, *we have* $\widehat{\gamma} - \gamma^0 = o_p(1)$.

The proof of Theorem 2.2 consists of two steps. Using the first order condition (2.4), together with the consistency established in the previous theorem, we first show that the LS estimator $\widehat{\theta}$ has a preliminary convergence rate $(NT)^{-\alpha}$. Next, we show that the rescaled objective function

$$\mathcal{L}^*(\gamma) = (NT)^{2\alpha}\Big[\mathcal{L}\Big(\widehat{\theta}(\gamma), \widehat{\Lambda}(\gamma), \gamma\Big) - \mathcal{L}\Big(\widehat{\theta}_{\gamma^0}, \widehat{\Lambda}_{\gamma^0}, \gamma^0\Big)\Big]$$

behaves like the one of a standard linear regression model (2.6), where $\widehat{\theta}_{\gamma^0}$ and $\widehat{\Lambda}_{\gamma^0}$ are the LS estimator when the threshold value γ^0 is observed a priori. Then invoking Assumption A.1(ii), whose implications are discussed in the linear regression model (2.6) in Section 2.2.3, we obtain the consistency of $\widehat{\gamma}$. Note that the magnitude of $\|\widehat{\theta}_{\gamma^0} - \theta^0\|$ is $O_p(\frac{1}{N} + \frac{1}{T})$, as documented in the previous studies such as Bai (2009) and Moon and Weidner (2017), implying $(NT)^{\alpha}\|\widehat{\theta}_{\gamma^0} - \theta^0\| = o_p(1)$ as $N/T \to \kappa$. So the estimation error in $\widehat{\theta}_{\gamma^0}$ is asymptotically negligible. We emphasize that the adjusting constant here should be $\mathcal{L}\big(\widehat{\theta}_{\gamma^0}, \widehat{\Lambda}_{\gamma^0}, \gamma^0\big)$, instead of $\mathcal{L}(\theta^0, \Lambda^0, \gamma^0)$ as the classical analysis suggests. The reason is that with $(NT)^{2\alpha}$ rescaled value, the estimation errors from Λ^0 have an asymptotically non-negligible effect on the objective function. But $\mathcal{L}\big(\widehat{\theta}_{\gamma^0}, \widehat{\Lambda}_{\gamma^0}, \gamma^0\big)$ contains the same estimation errors. So we use it as the adjusting constant to remove this effect.

2.3.2 Convergence rates

Given the consistency results in Theorems 2.1 and 2.2, we next establish the convergence rates for $\widehat{\theta}$ and $\widehat{\gamma}$. Because we do not have explicit expressions for these estimators, the derivations of these rates are tedious. We need the following assumption for the theoretical analysis.

Assumption A.5. Let $M_{\mathcal{D}}(\gamma) \equiv (NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} E_{\mathcal{D}}(x_{it}x_{it}'|\gamma) f_{it,\mathcal{D}}(\gamma)$, then the following statements hold:

(i) $M_{\mathcal{D}}(\gamma^0) > 0$ a.s. and $M_{\mathcal{D}}(\gamma)$ is continuous at $\gamma = \gamma^0$;

(ii) For all $\epsilon > 0$, there exist constants $\overline{N}$, $\overline{T}$, $B > 0$ and $\tau_1 > 0$, such that for all $N \geq \overline{N}$ and $T \geq \overline{T}$

$$\Pr\left(\inf_{|\gamma-\gamma^0|<B} \mu_{\min}(M_{\mathcal{D}}(\gamma)) > \tau_1\right) > 1 - \epsilon.$$

Assumption A.5 (i)-(ii) assumes that the square matrix $M_{\mathcal{D}}(\gamma)$ is well behaved in the neighborhood of γ^0.

The following theorem establishes the convergence rates of $\widehat{\gamma}$ and $\widehat{\theta}$.

Theorem 2.3. ***(Convergence rates of $\widehat{\gamma}$ and $\widehat{\theta}$)*** *Suppose that Assumptions A.1-A.5 hold and $N/T \to \kappa > 0$ as $(N,T) \to \infty$. Then*

(i) $(NT)^{1-2\alpha}(\widehat{\gamma} - \gamma^0) = O_p(1)$;

(ii)$\sqrt{NT}(\widehat{\theta} - \theta^0) = O_p(1)$.

Theorem 2.3(i) shows that $\widehat{\gamma} - \gamma^0 = O_p((NT)^{-1+2\alpha})$, which hinges on the magnitude of threshold effects. This result is quite intuitive. When α is close to zero, the threshold effects are large, we expect a more precise estimation of γ^0, leading to a faster convergence rate. On the other hand, when α is close to

1/2, the threshold effects are small, we expect a less precise estimation, which corresponds to a slower convergence rate. Also note that by equation (2.1),

$$\begin{aligned}\widehat{e}_{it} &= y_{it} - x_{it}'\widehat{\beta} - x_{it}'\widehat{\delta} d_{it}(\widehat{\gamma}) - \widehat{\lambda}_i'\widehat{f}_t \\ &= e_{it} + (\lambda_i^{0\prime} f_t^0 - \widehat{\lambda}_i'\widehat{f}_t) - x_{it}'(\widehat{\beta} - \beta^0) - x_{it}' d_{it}(\gamma^0)(\widehat{\delta} - \delta^0) - x_{it}'[d_{it}(\widehat{\gamma}) - d_{it}(\gamma^0)]\widehat{\delta}.\end{aligned}$$

Theorem 2.3(i) implies that for any $\epsilon > 0$, there is a C_ϵ such that $\Pr((NT)^{2\alpha-1}|\widehat{\gamma} - \gamma^0| > C_\epsilon) < \epsilon$. On the set $(NT)^{2\alpha-1}|\gamma - \gamma^0| \leq C_\epsilon$, we have $E\big[\|x_{it}(d_{it}(\gamma) - d_{it}(\gamma^0))\|\big] = O(|\gamma - \gamma^0|) = O((NT)^{2\alpha-1})$. This result, in conjunction with the fact $\widehat{\delta} = O_p((NT)^{-\alpha})$ and the inequality $\Pr(A) \leq \Pr(A|B) + \Pr(B^c)$, implies that the last term in the last displayed equation is $O_p((NT)^{\alpha-1}) = o_p(\frac{1}{\sqrt{NT}})$. Compared with the third and fourth terms, which are $O_p(\frac{1}{\sqrt{NT}})$ due to the result in Theorem 2.3(ii), we see that the last term is asymptotically negligible. However, when $\widehat{\gamma} = \gamma^0$, the last term is gone. So the estimation error associated with γ^0 has asymptotically negligible effects on the residual $\widehat{e}_{it}$. Since our least squares estimation aims to minimize $\sum_{i=1}^N \sum_{t=1}^T \widehat{e}_{it}^2$, we expect that the estimator of θ^0 with an unknown γ^0 is asymptotically equivalent to that with a known γ^0.

Theorem 2.3 is derived under the shrinking threshold effects assumption. When the threshold effects are fixed, we conjecture that $\widehat{\gamma} - \gamma^0 = O_p((NT)^{-1})$ and $\sqrt{NT}(\widehat{\theta} - \theta^0) = O_p(1)$. The first result is a natural extension of the result in Theorem 2.3(i) by letting $\alpha \to 0$. As regard to the second result, note that the estimation of θ^0 under the fixed threshold effects cannot be worse than the one under the shrinking case because of the stronger signal for γ^0 under the former, but cannot be better than the one when γ^0 is observed *a priori*. In both the shrinking threshold case and the observed γ^0 case, the estimators are $\sqrt{NT}$-consistent. This leads us to conjecture the second result. Furthermore, since the limiting distribution of the slope coefficient estimators under an unknown γ^0 in the shrinking case are the same as the one in the observed-γ^0 case, we conjecture that the limiting distribution of $\widehat{\theta} - \theta^0$ are the same as the result given in Theorem 2.4 below due to the same arguments.

In this paper, we assume that N and T pass to infinity at the same rate as in Moon and Weidner (2015, 2017). It is possible to allow N and T to diverge at different rates as in Bai (2009) and Lu and Su (2016). In this case, the estimators for slope coefficients would have three asymptotic bias terms entailed by weak exogeneity and heteroskedasticity, which are of order T^{-1} and N^{-1}. Then the result in Theorem 2.3(ii) should be changed to $\widehat{\theta} - \theta^0 = O_p(N^{-1} + T^{-1})$.

2.3.3 Asymptotic distributions of $\widehat{\theta}$ and $\widehat{\gamma}$

Given the convergence rates of $\widehat{\gamma}$ and $\widehat{\theta}$, we next present the limiting distributions. Our analysis indicates that $\widehat{\theta}$ has an asymptotically non-negligible bias. So the explicit expression of bias is also our target. For ease of exposition, we introduce the following notations.

Let $\mathbf{Z}_{k,\gamma} = \mathbb{M}_{\Lambda^0}\mathbf{X}_{k,\gamma}\mathbb{M}_{F^0}$, where $\mathbf{X}_{k,\gamma} = \mathbf{X}_k$ if $k \leq K$, and $\mathbf{X}_{k-K}(\gamma)$ if $K < k \leq 2K$. Let $z_{k,it,\gamma}$ be the (i,t)-th entry of the matrix $\mathbf{Z}_{k,\gamma}$ and $z_{it,\gamma} = (z_{1,it,\gamma}, \ldots, z_{2K,it,\gamma})'$ be the $2K$-dimensional vector composed of $z_{k,it,\gamma}$. Define

$$\begin{aligned}
\omega_{NT}(\gamma_1,\gamma_2) &= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} z_{it,\gamma_1} z'_{it,\gamma_2},\\
\Omega_{NT}(\gamma_1,\gamma_2) &= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} z_{it,\gamma_1} z'_{it,\gamma_2} e_{it}^2,\\
\mathbb{B}_{1,kNT}(\gamma) &= \frac{1}{N}\mathrm{tr}\Big[\mathbb{P}_{F^0} E_{\mathcal{D}}(\mathbf{e}'\mathbf{X}_{k,\gamma})\Big],\\
\mathbb{B}_{2,kNT}(\gamma) &= \frac{1}{T}\mathrm{tr}\Big[E_{\mathcal{D}}(\mathbf{e}\mathbf{e}')\mathbb{M}_{\Lambda^0}\mathbf{X}_{k,\gamma}F^0(F^{0\prime}F^0)^{-1}(\Lambda^{0\prime}\Lambda^0)^{-1}\Lambda^{0\prime}\Big],\\
\mathbb{B}_{3,kNT}(\gamma) &= \frac{1}{N}\mathrm{tr}\Big[E_{\mathcal{D}}(\mathbf{e}'\mathbf{e})\mathbb{M}_{F^0}\mathbf{X}'_{k,\gamma}\Lambda^0(\Lambda^{0\prime}\Lambda^0)^{-1}(F^{0\prime}F^0)^{-1}F^{0\prime}\Big].
\end{aligned}$$

We impose the following assumption on the above terms.

Assumption A.6. (i) The probability limits $\omega(\gamma_1,\gamma_2) \equiv \mathrm{plim}_{(N,T)\to\infty}\omega_{NT}(\gamma_1,\gamma_2)$ and $\Omega(\gamma_1,\gamma_2) \equiv \mathrm{plim}_{(N,T)\to\infty}\Omega_{NT}(\gamma_1,\gamma_2)$ are present and non-random, and are finite uniformly over $(\gamma_1,\gamma_2) \in \Gamma \times \Gamma$; (ii) The probability limits $\mathbb{B}_{\ell,k}(\gamma) \equiv$

$\text{plim}_{(N,T)\to\infty}\mathbb{B}_{\ell,kNT}(\gamma)$ for $\ell = 1,2,3$ and $k = 1,\ldots,2K$ are present and non-random, and are finite uniformly over $\gamma \in \Gamma$.

Assumption A.6 imposes some high level conditions on the terms appearing in the limiting distribution of $\widehat{\theta}$. The non-randomness of $\omega(\gamma_1,\gamma_2)$ and $\Omega(\gamma_1,\gamma_2)$ is crucial since it consists of the prerequisite conditions for the martingale central limit theorem. It is desirable to specify some primitive conditions to guarantee that the limits are fixed values. However, we emphasize that any effort on resorting to the primitive conditions would inevitably specify the internal structure of $E_{\mathcal{D}}(X_k)$, which, however, is generally unknown in practice, and has no guidance from the economic theories. Following the treatment of IFEs in the literature, we directly make high-level conditions. In our simulations, we generate x_{it} in a particular way. With this additional information on x_{it}, we can verify Assumption A.6 directly.

Let $\mathbb{B}_\ell(\gamma^0) = (\mathbb{B}_{\ell,1}(\gamma^0),\ldots,\mathbb{B}_{\ell,2K}(\gamma^0))'$ for $\ell = 1,2,3$. The following theorem reports the asymptotic distribution of $\widehat{\theta}$.

Theorem 2.4. ***(Asymptotic normality of*** $\widehat{\theta}$***)*** *Suppose that Assumptions A.1-A.6 hold and $N/T \to \kappa > 0$ as $(N,T)\to\infty$. Then*

$$\sqrt{NT}(\widehat{\theta}-\theta^0) \xrightarrow{d} N(\omega_0^{-1}\mathbb{B}, \omega_0^{-1}\Omega_0\omega_0^{-1}),$$

where $\omega_0 \equiv \omega(\gamma^0,\gamma^0)$, $\mathbb{B} \equiv -\kappa^{1/2}\mathbb{B}_1(\gamma^0) - \kappa^{-1/2}\mathbb{B}_2(\gamma^0) - \kappa^{1/2}\mathbb{B}_3(\gamma^0)$ and $\Omega_0 \equiv \Omega(\gamma^0,\gamma^0)$.

Theorem 2.4 indicates that $\widehat{\theta}$ has three asymptotically non-negligible bias terms associated with $\mathbb{B}_1(\gamma^0)$, $\mathbb{B}_2(\gamma^0)$ and $\mathbb{B}_3(\gamma^0)$. $\mathbb{B}_1(\cdot)$ is related to the possible appearance of the lagged dependent variables and it vanishes if the regressors are conditionally strictly exogenous in the sense that $E_{\mathcal{D}}(\mathbf{e}'\mathbf{X}_{k,\gamma}) = 0$ (i.e., $E_{\mathcal{D}}(e_{it}x_{is}) = 0$ and $E_{\mathcal{D}}[e_{it}x_{is}d_{is}(\gamma^0)] = 0$ for all t,s). In our setting both $E_{\mathcal{D}}(\mathbf{e}'\mathbf{e})$ and $E_{\mathcal{D}}(\mathbf{e}\mathbf{e}')$ are diagonal matrices. The second term, $\mathbb{B}_2(\cdot)$, vanishes in the absence of cross sectional heterogeneity in which case $E_{\mathcal{D}}(\mathbf{e}\mathbf{e}')$ is proportional to an identity matrix. Similarly, the third term, $\mathbb{B}_3(\cdot)$, vanishes in the absence of heterogeneity along the time dimension in which case $E_{\mathcal{D}}(\mathbf{e}'\mathbf{e})$ is proportional to an identity matrix. In the general case, we allow for weakly exogenous regressors and both cross-sectional and time series heteroskedasticity so that all three bias terms are present. In practice, one can follow Bai (2009) and Moon and Weidner (2017) to construct the bias-corrected estimates. The procedure is standard and omitted here for brevity.

To make inference on θ^0, we also need to estimate the asymptotic variance $\omega_0^{-1}\Omega_0\omega_0^{-1}$ consistently. Let $\widehat{\lambda}_i'$ be the ith row of $\widehat{\Lambda}$ and $\widehat{f}_t'$ be the tth row of $\widehat{F}$, where $\widehat{f}_t = (\widehat{\Lambda}'\widehat{\Lambda})^{-1}\widehat{\Lambda}'(Y_t - X_{t,\widehat{\gamma}}\widehat{\theta}) = \widehat{\Lambda}'(Y_t - X_{t,\widehat{\gamma}}\widehat{\theta})/N$. A consistent estimator, by the plug-in method, is given by $\widehat{\omega}^{-1}\widehat{\Omega}\widehat{\omega}^{-1}$, where

$$\widehat{\omega} \equiv \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\widehat{z}_{it,\widehat{\gamma}}\widehat{z}_{it,\widehat{\gamma}}', \qquad \widehat{\Omega} \equiv \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\widehat{z}_{it,\widehat{\gamma}}\widehat{z}_{it,\widehat{\gamma}}'\widehat{e}_{it}^2,$$

$\widehat{z}_{it,\gamma}$ is defined as a $2K \times 1$ vector with kth entry equal to (i,t)th entry of $\mathbb{M}_{\widehat{\Lambda}}\mathbf{X}_{k,\gamma}\mathbb{M}_{\widehat{F}}$ and $\widehat{e}_{it} = y_{it} - x_{it,\widehat{\gamma}}'\widehat{\theta} - \widehat{\lambda}_i'\widehat{f}_t$.

Next, we establish the asymptotic distribution of $\widehat{\gamma}$. Let $f_{it}(\cdot)$ denote the PDF of q_{it}. Let

$$D_{f,NT}(\gamma) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}E(x_{it}x_{it}'|q_{it}=\gamma)f_{it}(\gamma), \text{ and}$$
$$V_{f,NT}(\gamma) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}E(x_{it}x_{it}'e_{it}^2|q_{it}=\gamma)f_{it}(\gamma).$$

Then we add the following assumption.

Assumption A.7. (i) The limits $D_f(\gamma) \equiv \lim_{(N,T)\to\infty} D_{f,NT}(\gamma)$ and $V_f(\gamma) \equiv \lim_{(N,T)\to\infty} V_{f,NT}(\gamma)$ exist uniformly over $\gamma \in \Gamma$ and are continuous at $\gamma = \gamma_0$;

(ii) There is a constant M such that

$$\mathrm{Var}\Big(\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}E_{\mathcal{D}}(\|g_{it}(\gamma_1,\gamma_2)\|^2)\Big) \le M\max_{i,t}E\Big(\|g_{it}(\gamma_1,\gamma_2)\|^4\Big),$$

where $g_{it}(\gamma_1,\gamma_2) = x_{it}\,|d_{it}(\gamma_1) - d_{it}(\gamma_2)|$ and $x_{it}e_{it}\,|d_{it}(\gamma_1) - d_{it}(\gamma_2)|$.

Assumption A.7(i) requires the limits $D_f(\gamma)$ and $V_f(\gamma)$ to be well defined. It also assumes the continuity of $V_f(\gamma)$ at γ^0. The discontinuous case can be allowed and is discussed in Section 2.3.5 below. Assumption A.7(ii) requires the conditional expectations of $\|x_{it}\|^2|d_{it}(\gamma_1)-d_{it}(\gamma_2)|$ and $\|x_{it}e_{it}\|^2|d_{it}(\gamma_1)-d_{it}(\gamma_2)|$ to be well behaved.

Let $D_f^0 \equiv D_f(\gamma^0)$ and $V_f^0 \equiv V_f(\gamma^0)$. The following theorem gives the asymptotic distribution of $\widehat{\gamma}$.

Theorem 2.5. ***(Asymptotic distribution of*** $\widehat{\gamma}$*) Suppose that Assumptions A.1-A.5 and A.7 hold, and $N/T \to \kappa > 0$ as $(N,T) \to \infty$. Then $(NT)^{1-2\alpha}(\widehat{\gamma}-\gamma^0) \xrightarrow{d} \phi\xi$, where $\xi = \operatorname{argmax}_{-\infty<r<\infty}[-\frac{1}{2}|r| + W(r)]$, $\phi = C^{0\prime}V_f^0C^0/(C^{0\prime}D_f^0C^0)^2$, and $W(\cdot)$ is a two-sided standard Brownian motion on the real line.*

A two-sided Brownian motion $W(\cdot)$ on the real line is defined as

$$W(r) = W_1(-r)\mathbf{1}\{r \leq 0\} + W_2(r)\mathbf{1}\{r > 0\},$$

where $W_1(\cdot)$ and $W_2(\cdot)$ are two independent standard Brownian motions on $[0,\infty)$. Theorem 2.5 implies that the pseudo statistic $(NT)^{1-2\alpha}(\widehat{\gamma}-\gamma^0)/\phi$ has an asymptotically pivotal distribution, ξ. This result is similar to Theorem 1 in Hansen (2000).

The asymptotic result in Theorem 2.5 relies critically on the shrinking effect assumption. In the fixed threshold effect framework (i.e., $\alpha = 0$), it is possible to demonstrate $NT(\widehat{\gamma}-\gamma^0) = O_p(1)$. But deriving the asymptotic distribution is not an easy task. Based on Theorem 2 of Chan (1993), we conjecture that the limiting distribution is the maximizer of some compound Poisson process, which may involve the marginal distribution of x_{it} and the regression coefficients.

2.3.4 The likelihood ratio test

To make inference on γ^0, one may be tempted to apply the asymptotic distribution result in Theorem 2.5. But the limiting distribution of $\widehat{\gamma}$ depends on the scale parameter ϕ which is hard to estimate accurately. Inferences based on Theorem 2.5 tend to be poor in finite samples. Following the lead of Hansen (2000), we consider a likelihood ratio (LR) statistic in this subsection to test the null hypothesis $H_0 : \gamma = \gamma^0$. Let $(\widehat{\theta}(\gamma), \widehat{\Lambda}(\gamma))$ be the estimator when the threshold value γ is given and $\mathcal{L}^*(\gamma) = \mathcal{L}(\widehat{\theta}(\gamma), \widehat{\Lambda}(\gamma), \gamma)$. Define

$$LR_{NT}(\gamma) = NT \frac{\mathcal{L}^*(\gamma) - \mathcal{L}^*(\widehat{\gamma})}{\mathcal{L}^*(\widehat{\gamma})},$$

where $\mathcal{L}^*(\widehat{\gamma}) = \mathcal{L}(\widehat{\theta}(\widehat{\gamma}), \widehat{\Lambda}(\widehat{\gamma}), \widehat{\gamma}) = \mathcal{L}(\widehat{\theta}, \widehat{\Lambda}, \widehat{\gamma})$.

The following theorem reports the asymptotic distribution of $LR_{NT}(\gamma^0)$.

Theorem 2.6. ***(Likelihood ratio test)*** *Suppose that Assumptions A.1-A.7 hold and $N/T \to \kappa > 0$ as $(N,T) \to \infty$. Then under $H_0 : \gamma = \gamma^0$, we have*

$$LR_{NT}(\gamma^0) \xrightarrow{d} \eta^2 \Xi,$$

where $\eta^2 = C^{0\prime} V_f^0 C^0 / (\sigma^2 C^{0\prime} D_f^0 C^0)$, $\sigma^2 = \text{plim}_{(N,T)\to\infty} \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T e_{it}^2$, and $\Xi = \max_{-\infty<r<\infty} [-|r| + 2W(r)]$ has the distribution function characterized by $\Pr(\Xi \leq z) = (1 - e^{-z/2})^2$.

The result in Theorem 2.6 is essentially same to Theorem 2 in Hansen (2000). The major difference lies in the appearance of $\widehat{\Lambda}(\gamma)$ in our definition of $\mathcal{L}^*(\gamma)$, which is a large dimensional matrix estimator. Bounding the change of the likelihood value arising from the change of such a large dimensional matrix estimator is an indispensable step in our theoretical analysis and it requires the use of the celebrated Davis and Kahan's (1970) sin(Θ) theorem. Apparently, such a step is not needed in Hansen's analysis.

As Theorem 2.6 suggests, we still have a nuisance parameter η^2. In the special case that the errors are homoskedastic over the cross-section and time series dimensions, this parameter is equal to 1. But for the general case, one need to estimate η^2 consistently. Noting that

$$\eta^2 = \frac{\text{plim}_{(N,T)\to\infty} \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T E[(\delta^{0\prime} x_{it} e_{it})^2 | q_{it} = \gamma^0] f_{it}(\gamma^0)}{\sigma^2 \text{plim}_{(N,T)\to\infty} \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T E[(\delta^{0\prime} x_{it})^2 | q_{it} = \gamma^0] f_{it}(\gamma^0)},$$

we propose to estimate η^2 by

$$\widehat{\eta}^2 = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} K_h(\widehat{\gamma} - q_{it})(\widehat{\delta}' x_{it}\widehat{e}_{it})^2}{\widehat{\sigma}^2 \sum_{i=1}^{N}\sum_{t=1}^{T} K_h(\widehat{\gamma} - q_{it})(\widehat{\delta}' x_{it})^2},$$

where $\widehat{\sigma}^2 = \frac{1}{NT}\mathcal{L}^*(\widehat{\gamma})$, $K_h(u) = h^{-1}K(u/h)$, $h \to 0$ is bandwidth parameter and $K(\cdot)$ is a kernel function. We can readily show that $\widehat{\sigma}^2 = \sigma^2 + o_p(1)$ and $\widehat{\eta}^2 = \eta^2 + o_p(1)$ under some regularity conditions on h and $K(\cdot)$. Given the consistent estimate of η^2, we can consider the normalized LR statistic

$$NLR_{NT}(\gamma^0) = LR_{NT}(\gamma^0)/\widehat{\eta}^2.$$

We can easily tabulate the asymptotic critical value for $NLR_{NT}(\gamma^0)$. We can also invert this statistic to obtain the asymptotic $1-\alpha$ confidence interval for $\gamma:$

$$CI_{1-\alpha} = \{\gamma \in \Gamma : NLR_{NT}(\gamma) \leq \Xi_{1-\alpha}\}$$

where $\Xi_{1-\alpha}$ denotes the $1-\alpha$ quantile of Ξ. For example $\Xi_{1-\alpha}$ =5.94, 7.35, and 10.59 for α =0.10, 0.05, and 0.01, respectively.

2.3.5 Threshold effect in the error variance

So far, the asymptotic results in the previous subsections are derived under the assumption that $V_f(\gamma)$ is continuous at γ^0 (see Assumption A.7(i)). This entails a neat and symmetric asymptotic distribution of $\widehat{\gamma}$. However, there are cases where this assumption is violated. For example, the conditional distribution of the regressors given the threshold variable changes across the threshold value, or the distribution of the error term changes across the threshold value, or both. Among these cases, a particularly interesting case is that the variance of the error term has threshold effect; see, e.g., Seo and Linton (2007). In this subsection, we consider the extension to allow threshold effects in the error variance.

A direct consequence of the presence of threshold effects in the error variance is that $V_f(\gamma)$ has a jump at γ^0. We assume that the left and right limits exist with $V^0_{f,-} = \lim_{\gamma\uparrow\gamma^0} V_f(\gamma)$ and $V^0_{f,+} = \lim_{\gamma\downarrow\gamma^0} V_f(\gamma)$. The results in Theorems 2.5-2.6 now need to be modified to accommodate the discontinuity fact. However, we emphasize that the arguments and methods of deriving the results in Theorems 2.5-2.6 are only slightly affected in this more general case. In fact, we now need to derive the asymptotic results separately for the $\gamma \leq \gamma^0$ and $\gamma > \gamma^0$ cases. But even under Assumption 7(i), the proofs of Theorems 2.5-2.6 are conducted separately for the subcases that $\gamma \leq \gamma^0$ and $\gamma > \gamma^0$. So the only difference in the general case is how to unify the asymptotic results of the two subcases. Now, the asymptotic distribution of $(NT)^{1-2\alpha}(\widehat{\gamma} - \gamma^0)$ changes to

$$\operatorname*{argmax}_{-\infty<r<\infty} \big[\phi_L\big(-\frac{1}{2}|r| + W_1(-r)\big)\mathbf{1}\{r \leq 0\} + \phi_R\big(-\frac{1}{2}|r| + W_2(r)\big)\mathbf{1}\{r > 0\}\big],$$

with $\phi_L = C^{0\prime}V^0_{f,-}C^0/(C^{0\prime}D^0_f C^0)^2$ and $\phi_R = C^{0\prime}V^0_{f,+}C^0/(C^{0\prime}D^0_f C^0)^2$. Since ϕ_L and ϕ_R do not average out as they do in model (2.1) without a threshold effect in the error variance, the leading term of $(NT)^{1-2\alpha}(\widehat{\gamma} - \gamma^0)$ now is not asymptotically pivotal up to a scalar.

The LR test statistic $LR_{NT}(\gamma^0)$ has the asymptotic distribution

$$\max_{-\infty<r<\infty} \big[\eta^2_L\big(-|r| + 2W_1(-r)\big)\mathbf{1}\{r \leq 0\} + \eta^2_R\big(-|r| + 2W_2(r)\big)\mathbf{1}\{r > 0\}\big], \quad (2.7)$$

where $\eta^2_L = C^{0\prime}V^0_{f,-}C^0/(\sigma^2 C^{0\prime}D^0_f C^0)$ and $\eta^2_R = C^{0\prime}V^0_{f,+}C^0/(\sigma^2 C^{0\prime}D^0_f C^0)$. Again, the asymptotic distribution of $LR_{NT}(\gamma^0)$ is not pivotal up to a scalar any more. We can construct a nonparametric estimator to estimate η^2_L:

$$\widehat{\eta}^2_L = \frac{NT\sum_{i=1}^N\sum_{t=1}^T K_h(\widehat{\gamma} - q_{it})(\widehat{\delta}' x_{it}\widehat{e}_{it})^2\mathbf{1}\{q_{it} \leq \widehat{\gamma}\}}{\widehat{\sigma}^2[\sum_{i=1}^N\sum_{t=1}^T K_h(\widehat{\gamma} - q_{it})\mathbf{1}\{q_{it} \leq \widehat{\gamma}\}][\sum_{i=1}^N\sum_{t=1}^T K_h(\widehat{\gamma} - q_{it})(\widehat{\delta}' x_{it})^2]}.$$

The estimator for η^2_R can be constructed analogously. Note that the expression in (2.7) is equivalent to

$$\max\big[\max_{0<r<\infty} \eta^2_L\big(-r + 2W_1(r)\big), \max_{0<r<\infty} \eta^2_R\big(-r + 2W_2(r)\big)\big].$$

Given the fact that $\Pr(\max_{0<r<\infty}\{-\frac{1}{2}r + W(r)\} < z) = 1 - e^{-z}$ for a standard

Wiener process, the relationship of the asymptotic p-value and $LR_{NT}(\gamma^0)$ can be easily derived as

$$\text{Asy. p-value} = 1 - \left[1 - e^{-\frac{1}{2\hat{\eta}_L^2}LR_{NT}(\gamma^0)}\right]\left[1 - e^{-\frac{1}{2\hat{\eta}_R^2}LR_{NT}(\gamma^0)}\right].$$

Alternatively, we can calculate the critical value under nominal size α, C_α, by solving the following equation:

$$\left[1 - e^{-\frac{C_\alpha}{2\hat{\eta}_L^2}}\right]\left[1 - e^{-\frac{C_\alpha}{2\hat{\eta}_R^2}}\right] = 1 - \alpha.$$

The finite sample performance of the above discussed procedure is examined via Monte Carlo simulations in Section F of the online supplement of Miao et al. (2020a).

2.3.6 Determination of R^0

When implementing the least squares estimation on the proposed model, one has to determine the number of factors (R^0). This is an intrinsic issue related to factor analyses. For approximate factor models, there are various ways to determine R^0; see Bai and Ng (2002), Onatski (2010), and Ahn and Horenstein (2013), among others. However, it seems difficult to extend existing methods to the current model because of the nonlinearity arising from the threshold effects. Recently, Moon and Weidner (2019) and Chernozhukov et al. (2019) consider the nuclear norm regularized estimation of panel regression models and provide methods to determine the number of factors by singular value thresholding (SVT). This subsection extends these methods to determine the number of factors in our panel threshold model. Let $s_r(A)$ denote the rth largest singular value of an $N \times T$ matrix A. Let $\|A\|_* = \sum_{r=1}^{\min(N,T)} s_r(A)$ denote the nuclear norm of A. The procedure goes as follows:

1. Conduct the nuclear norm regularized estimation:

$$\begin{aligned}(\widehat{\theta}_\psi, \widehat{\gamma}_\psi, \widehat{\Theta}_\psi) &\equiv \text{argmin}_{(\theta,\gamma,\Theta)\in\mathbb{R}^{2K}\times\Gamma\times\mathbb{R}^{N\times T}}\mathcal{L}_{n,\psi}(\theta,\gamma,\Theta), \text{ where}\\ \mathcal{L}_{n,\psi}(\theta,\gamma,\Theta) &= \frac{1}{2NT}\Big\|\mathbf{Y}-\beta\odot\mathbf{X}-\delta\odot\mathbf{X}(\gamma)-\Theta\Big\|^2+\frac{\psi}{\sqrt{NT}}\|\Theta\|_*,\end{aligned}$$

   for some tuning parameter $\psi \asymp (N^{-1/2}+T^{-1/2})$.

2. Estimate R^0 by $\widehat{R} = \sum_{r=1}^{\min(N,T)} \mathbf{1}\{s_r(\widehat{\Theta}_\psi) \geq \chi_{NT}\}$ for some singular value threshold $\chi_{NT} = \log(\sqrt{N}+\sqrt{T})(\sqrt{NT}\psi\|\widehat{\Theta}_\psi\|_{\text{sp}})^{1/2}$.

The idea of the above proposed method is similar to that of Moon and Weidner (2019) and Chernozhukov et al. (2019). The $N\times T$ matrix estimator $\widehat{\Theta}_\psi$ estimates $\Lambda^0 F^{0\prime}$. Under some regularity conditions, we can show that $||\widehat{\Theta}_\psi - \Lambda^0 F^{0\prime}||_{\text{sp}} = O_p(\sqrt{N}+\sqrt{T})$. Suppose $R^0 > 0$, the first R^0 singular values of $\Lambda^0F^{0\prime}$ are of the order $O_p(\sqrt{NT})$ and the remaining ones are equal to zero, implying that $s_r(\widehat{\Theta}_\psi) \asymp \sqrt{NT}$ for $r \leq R^0$ and $s_r(\widehat{\Theta}_\psi) = O_p(\sqrt{N}+\sqrt{T})$ for $r > R^0$. Then one can readily show that $\Pr(\widehat{R} = R^0) \to 1$ as $(N,T)\to\infty$ if $\chi_{NT} \asymp \log(\sqrt{N}+\sqrt{T})(N^{1/2}T^{1/4}+N^{1/4}T^{1/2})$. Similar analysis can be applied to the special case of $R^0 = 0$. The factor $\log(\sqrt{N}+\sqrt{T})$ in χ_{NT} helps us to handle the case $R^0 = 0$. If $R^0 > 0$, this factor can be ignored by setting $\chi_{NT} = (\sqrt{NT}\psi\|\widehat{\Theta}_\psi\|_{\text{sp}})^{1/2}$.

In Section D of the online supplement of Miao et al. (2020a), we provide a rigorous proof for the consistency of $\widehat{R}$. There, we allow N and T to diverge to infinity at different rates. We also allow the threshold effects to be either fixed or shrink to zero. In Section F of the online supplement of Miao et al. (2020a), we provide some simulation results to demonstrate that the SVT method works fairly well in finite samples.

In practice, we need to choose the tuning parameter ψ. As discussed in section 2.5 of Chernozhukov et al. (2019), a desired tuning parameter ψ equals $c||\mathbf{e}||_{\text{sp}}$ for some constant $c \in (0,1)$. To quantify ψ, we can follow Chernozhukov et al. (2019) to compute an appropriate tuning parameter via simulation under the Gaussian assumption. In our framework, as e_{it}'s do not have cross-sectional or serial correlation, we can generate u_{it} that is independent across both (i, t) and $u_{it} \sim \mathcal{N}(0, \sigma^2)$. Then the tuning parameter is given by $c\|\mathbf{u}\|_{\mathrm{sp}}$, where $\mathbf{u}$ is a stack of u_{it}'s into an $N \times T$ matrix. In practice, we can replace σ^2 with an initial consistent estimator. As the least squares estimator in our model has similar robust properties as shown in Moon and Weidner (2015) that as long as the number of factors is not underestimated, we can first estimate the model with $R_{max} > R^0$ and obtain an estimate of σ^2.

2.3.7 Two-way additive effects v.s. interactive effects

Our model is attractive in its flexibility to model the unobserved heterogeneity and cross-sectional dependence in real data via the IFEs. In traditional panel data models, the unobserved heterogeneity is usually addressed via the two-way additive fixed effects. This naturally gives rise to the question on which model should be used in empirical applications. Let α_i and ν_t be the individual and time fixed effects. Then

$$\alpha_i + \nu_t = \lambda_i' f_t$$

with $\lambda_i = (\alpha_i, 1)'$ and $f_t = (1, \nu_t)'$. This implies that the two-way additive fixed effects is a very special case of IFEs with the number of factors equal to 2 and a factor and a factor loading set to 1. As a result, if the unobserved heterogeneity in the data is of two-way additive fixed effects form but one uses our panel threshold model with IFEs to estimate it, the resultant estimators are still consistent but inefficient. The inefficiency is due to the fact that the useful information of partial factors and factor loadings being observed are not properly accounted in the IFEs estimation. On the other hand, if the heterogeneity is of the IFEs form and cannot be simplified to the two-way additive form, but one uses the within-group method to estimate the model, the resultant estimators would be generally inconsistent because the endogeneity arising from the unobserved factors and factor loadings are not fully controlled.

Based on the above discussion, a plausible procedure is that we first invoke the LS estimation method to estimate the panel threshold model with IFEs, and then we test whether the estimated heterogeneity can be further reduced to the two-way additive form. If the test is passed, we turn to the within-group estimator to achieve efficiency; otherwise we can continue to employ the LS estimator studied in this paper.

Motivated by the above discussions, we propose a formal statistic to test the null of two-way additive fixed-effects against the alternative of more general IFEs in Section E of the online supplement of Miao et al. (2020a). We focus on the case of $R^0 = 2$ there and propose a test statistic that is asymptotically standard normal under the null. Alternatively, we can extend the sup-type test of Castagnetti et al. (2015) to our framework. In addition, we conduct some simulations in Section F of the online supplement of Miao et al. (2020a) to evaluate the finite sample performance of the test.

2.4 Testing the existence of threshold effect

Testing the existence of threshold effect is non-standard because the threshold level γ^0 is not identified when $\delta^0 = 0$. This issue has been well documented in the econometrics literature; see Andrews (1993), Hansen (1996) and references therein. In the current paper, we are interested in testing the null hypothesis $\mathbb{H}_0 : \delta^0 = 0$ versus the alternative $\mathbb{H}_1 : \delta^0 \neq 0$. To study the local power of our test, we consider the sequence of Pitman local alternatives $\mathbb{H}_{1n} : \delta^0 = \frac{c}{\sqrt{NT}}$. Note that δ^0 shrinks to zero at a faster rate than the one specified in Assumption A.3. So under the Pitman local alternatives the parameters γ^0 is not identified. Apparently, the case of $c = 0$ corresponds to the null of no threshold effect.

For each $\gamma \in \Gamma = [\underline{\gamma}, \overline{\gamma}]$, we can obtain the estimator $\widehat{\theta}(\gamma)$, $\widehat{\Lambda}(\gamma)$, $\widehat{F}(\gamma)$ and bias-corrected estimator $\widetilde{\theta}(\gamma)$. Then we construct the asymptotic variance estimators

$$\widehat{\omega}_{NT}(\gamma,\gamma) \equiv \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\check{z}_{it,\gamma}\check{z}_{it,\gamma}', \quad \text{and} \quad \widehat{\Omega}_{NT}(\gamma,\gamma) \equiv \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\check{z}_{it,\gamma}\check{z}_{it,\gamma}'\widehat{e}_{it}(\gamma)^2,$$

where $\check{z}_{it,\gamma}$ is a $2K \times 1$ vector with kth entry equal to (i,t)th entry of $\mathbb{M}_{\widehat{\Lambda}(\gamma)}\mathbf{X}_{k,\gamma}\mathbb{M}_{\widehat{F}(\gamma)}$ and $\widehat{e}_{it}(\gamma) = y_{it} - x_{it,\gamma}'\widehat{\theta}(\gamma) - \widehat{\lambda}_i(\gamma)'\widehat{f}_t(\gamma)$. The sup-Wald statistic is defined as

$$\text{sup}W_{NT} \equiv \sup_{\gamma\in\Gamma} W_{NT}(\gamma),$$

where $W_{NT}(\gamma) = NT \cdot \widetilde{\theta}(\gamma)'L\widehat{\mathbb{K}}_{NT}^{-1}(\gamma)L'\widetilde{\theta}(\gamma)$ with $L = [0_{K\times K}, I_K]'$ a section matrix, and $\widehat{\mathbb{K}}_{NT}(\gamma) = L'\widehat{\omega}_{NT}(\gamma,\gamma)^{-1}\widehat{\Omega}_{NT}(\gamma,\gamma)\widehat{\omega}_{NT}(\gamma,\gamma)^{-1}L$, the estimated covariance matrix for $\sqrt{NT}L'\widetilde{\theta}(\gamma)$.

The asymptotic property of the sup$W_{NT}(\gamma)$ statistic is presented in the following theorem.

Theorem 2.7. ***(Wald test)*** *Suppose that Assumptions A.1-A.2, A.4 and A.6 hold and $N/T \to \kappa > 0$ as $(N,T) \to \infty$. Then under $\mathbb{H}_{1n} : \delta^0 = c/\sqrt{NT}$, we have*

$$\text{sup}W_{NT} \xrightarrow{d} \sup_{\gamma\in\Gamma} W^c(\gamma)$$

where

$$W^c(\gamma) = \left[\overline{S}(\gamma) + \overline{Q}(\gamma)c\right]'\overline{K}(\gamma,\gamma)^{-1}\left[\overline{S}(\gamma) + \overline{Q}(\gamma)c\right], \quad \overline{Q}(\gamma) = L'\omega(\gamma,\gamma)^{-1}\omega(\gamma,\gamma^0)L,$$

and $\overline{S}(\gamma) = L'\omega(\gamma,\gamma)^{-1}S(\gamma)$ is a mean-zero Gaussian process with covariance kernel $\overline{K}(\gamma_1,\gamma_2) = L'\omega(\gamma_1,\gamma_1)^{-1}\ \Omega(\gamma_1,\gamma_2)\omega(\gamma_2,\gamma_2)^{-1}L$.

Under $\mathbb{H}_0$, $c = 0$ and $\sup_{\gamma\in\Gamma} W^0(\gamma) = \sup_{\gamma\in\Gamma}\overline{S}(\gamma)'\overline{K}(\gamma,\gamma)^{-1}\overline{S}(\gamma)$. Clearly, the limiting null distribution of supW_{NT} depends on the Gaussian process $\overline{S}(\gamma)$ and is not pivotal. We cannot tabulate the asymptotic critical values for the sup-Wald statistic. Nevertheless, given the simple structure of $\overline{S}(\gamma)$, we can follow the literature (e.g., Hansen (1996)) and simulate the critical values via the following procedure:

1. Generate $\{v_{it}, i = 1, \ldots, N, t = 1, \ldots, T\}$ independently from the standard normal distribution;
2. Calculate $\widehat{S}_{NT}(\gamma) = \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=1}^{T} \check{z}_{it,\gamma} \widehat{e}_{it}(\gamma) v_{it}$;
3. Computesup$W_{NT}^{*} \equiv \sup_{\gamma \in \Gamma} \widehat{S}_{NT}(\gamma)' \widehat{\omega}_{NT}(\gamma, \gamma)^{-1} L \widehat{\mathbb{K}}_{NT}^{-1}(\gamma) L' \widehat{\omega}_{NT}(\gamma, \gamma)^{-1} \widehat{S}_{NT}(\gamma)$;
4. Repeat Steps 1-3 B times and denote the resulting supW_{NT}^{*} test statistics as sup$W_{NT,j}^{*}$ for $j = 1, \ldots, B$.
5. Calculate the simulated/bootstrap p-value for the supW_{NT} test as $p_W^* = \frac{1}{B} \sum_{j=1}^{B} \mathbf{1}\{\text{sup} W_{NT,j}^{*} \geq \text{sup} W_{NT}\}$ and reject the null when p_W^* is smaller than some prescribed nominal level of significance.

The next theorem justifies the asymptotic validity of the above procedure.

Theorem 2.8. ***(Bootstrap validity)*** *Suppose that Assumptions A.1-A.2, A.4-A.7 hold and $N/T \to \kappa > 0$ as $(N, T) \to \infty$. Then* $\text{supW}_{NT}^{*} \xrightarrow{d} \sup_{\gamma \in \Gamma} W^0(\gamma)$.

Theorem 2.8 indicates that the bootstrap statistic can mimic the asymptotic null distribution of the statistic supW_{NT}. When B is sufficiently large, the asymptotic critical value of the level α test based on supW_{NT} is approximately given by the empirical upper α-quantile of $\{\text{sup}W_{NT,j}^{*}, j = 1, \ldots, B\}$. Therefore, we can reject the null hypothesis $\mathbb{H}_0 : \delta^0 = 0$ if the simulated p-value p_W^* is smaller than α.

2.5 Monte Carlo simulations

In this section, we conduct Monte Carlo simulations to evaluate the finite sample performance of our estimators and test statistics.

2.5.1 Data generating processes

We consider four data generating processes:

DGP 1: $y_{it} = \beta^0 y_{i,t-1} + \delta^0 y_{i,t-1} \mathbf{1}\{y_{i,t-1} \leq \gamma^0\} + \lambda_i^{0\prime} f_t^0 + e_{it}$;

DGP 2: $y_{it} = \beta^0 y_{i,t-1} + \delta^0 y_{i,t-1}\mathbf{1}\{q_{it} \leq \gamma^0\} + \lambda_i^{0\prime} f_t^0 + e_{it}$, where the threshold variable q_{it} is independently and identically distributed (i.i.d.) from $N(2,1)$;
DGP 3: $y_{it} = \beta^0 x_{it} + \delta^0 x_{it}\mathbf{1}\{x_{it} \leq \gamma^0\} + \lambda_i^{0\prime} f_t^0 + e_{it}$, where $x_{it} = 2 + v_{it} + (\lambda_i^0 + \lambda_i^*)'(f_t^0 + 0.5 f_{t-1}^0)$, v_{it} is i.i.d. $N(0,1)$ and λ_{ir}^* is i.i.d. $N(1,1)$ for $r = 1, 2$;
DGP 4: $y_{it} = \beta^0 x_{it} + \delta^0 x_{it}\mathbf{1}\{q_{it} \leq \gamma^0\} + \lambda_i^{0\prime} f_t^0 + e_{it}$, where x_{it} is generated in the same way as in DGP 3 and q_{it} is generated in the same way as in DGP 2.

We set $\beta^0 = 0.3$, $\delta^0 = (NT)^{-0.2}$ and $\gamma^0 = 2$ for all four DGPs. In all the above four DGPs, both λ_i^0 and f_t^0 are 2×1 vectors with each entry of λ_i^0 being i.i.d. $N(1,1)$ and each entry of f_t^0 being i.i.d. $0.7 \times N(1,1)$. The factors and factor loadings are mutually independent of each other. We also generate the idiosyncratic error terms e_{it} independently from the student t-distribution with nine degrees of freedom. For the dynamic DGPs 1 and 2, we throw away the first 1000 time periods of observations to get rid of the start-up effect. For the static DGPs 3 and 4, we generate correlations between the regressors x_{it} and the factors and factor loadings.

DGP 1 is a dynamic panel where the lagged dependent variable $y_{i,t-1}$ also serves as the threshold variable. DGP 2 is a dynamic panel with a strictly exogenous variable q_{it} being the threshold variable. DGP 3 is a static panel with the exogenous regressor x_{it} serving as the threshold variable and DGP 4 is a static panel where the threshold variable q_{it} is different from the regressor x_{it}. Note that the high level assumption, Assumption A.6, is satisfied in our DGPs. Take DGP1 as an example. It is easy to see that y_{it} is independent with y_{jt} condition on $F^0 = (f_1^0, \ldots, f_T^0)'$ and y_{it} is weakly correlated y_{is} with exponential-decay correlations conditional on λ_i^0. Then we prove the law of large numbers simply by directly showing that the variance converges to zero. With the partition arguments as used in the proof of Theorem 2.4.1 in Vaart et al. (1996), the point convergence can be readily extended to the uniform convergence.

We are interested in the performance of our estimators and test statistics in all the above four scenarios. In addition, we also consider generating conditional heteroskedastic errors as in Su and Chen (2013). The simulation performance is similar to that reported here.

2.5.2 Implementation and estimation results

For each DGP, we consider the feasible and infeasible bias-corrected estimators, where the infeasible estimators are obtained by using the information of true threshold parameter γ^0. For all bias-corrected estimators, we correct all three bias terms based on the formula in Section 2.3 by ignoring the fact that there is no need to correct some bias terms in some DGPs. For example, the slope coefficient estimator has only the bias term $\mathbb{B}_1(\gamma^0)$ in DGPs 1 and 2, and has no bias term in DGPs 3 and 4. For the slope coefficients (β^0, δ^0) and the threshold parameter γ^0, we report the empirical bias (Bias), standard deviation (Std), and coverage probability (CP) of the 95% confidence interval for the corresponding true value. The number of repetitions for each case is set to be 500.

Table 2.1 reports the estimation and inference results with the unknown R^0 being estimated by the SVT method in Section 2.3.6. The tuning parameter ψ is chosen following the description at the end of Section 2.3.6. Specifically, we chose $c = 0.3$ in ψ. The number of factors is estimated quite accurately. When both N and T are not less than 50 for all DGPs, the correct estimation rate is above 99%. For each combination of N and T in each DGP, the table reports the estimation and inference results for the feasible estimates of β^0 and δ^0 on a row, followed by those of the infeasible estimates in the next row. We summarize some important findings from Table 2.1. First, as expected, the infeasible estimates of β^0 and δ^0 tend to outperform the feasible estimates slightly in terms of bias and standard deviation, which is especially true for the estimate of threshold effect δ^0. In addition, we find that the two estimates behave similarly, which supports the theoretical claim that they are asymptot-

Table 2.1: Performance of the least square estimators

				β			δ			γ	
	N	T	Bias	Std	CP	Bias	Std	CP	Bias	Std	CP
1	25	50	-0.0049	0.0304	93.6%	0.0055	0.0455	91.2%	0.0006	0.2525	90.2%
			-0.0039	0.0298	94.2%	-0.0016	0.0427	92.0%			
	50	25	-0.0153	0.0345	88.2%	0.0099	0.0452	92.6%	-0.0194	0.2264	91.2%
			-0.0141	0.0336	89.4%	0.0011	0.0408	93.6%			
	50	50	-0.0036	0.0213	93.8%	0.0032	0.0301	92.8%	-0.0015	0.1277	91.4%
			-0.0028	0.0214	93.4%	-0.0011	0.0292	93.6%			
	50	100	-0.0024	0.0140	94.8%	0.0022	0.0196	95.4%	0.0001	0.0738	94.0%
			-0.0019	0.0140	94.6%	0.0000	0.0196	95.2%			
	100	50	-0.0037	0.0152	91.0%	0.0021	0.0205	92.2%	0.0025	0.0759	93.4%
			-0.0033	0.0151	91.4%	0.0001	0.0197	94.0%			
	100	100	-0.0016	0.0099	95.0%	0.0014	0.0144	94.0%	0.0008	0.0445	94.0%
			-0.0013	0.0099	95.2%	0.0002	0.0145	93.0%			
2	25	50	-0.0059	0.0306	89.4%	0.0021	0.0209	93.2%	0.0019	0.0279	90.4%
			-0.0054	0.0305	90.2%	0.0010	0.0210	92.2%			
	50	25	-0.0140	0.0289	89.4%	0.0018	0.0196	93.6%	-0.0003	0.0266	92.6%
			-0.0140	0.0285	90.4%	0.0010	0.0192	93.8%			
	50	50	-0.0045	0.0201	92.4%	0.0000	0.0139	93.4%	-0.0001	0.0150	94.4%
			-0.0044	0.0201	92.4%	-0.0005	0.0138	93.6%			
	50	100	-0.0021	0.0138	94.6%	0.0000	0.0103	92.8%	0.0006	0.0098	95.8%
			-0.0019	0.0138	94.8%	-0.0002	0.0102	93.4%			
	100	50	-0.0043	0.0148	93.2%	0.0005	0.0097	93.0%	0.0006	0.0128	93.2%
			-0.0040	0.0146	93.2%	0.0001	0.0097	93.2%			
	100	100	-0.0011	0.0102	93.2%	0.0003	0.0074	92.8%	0.0002	0.0068	94.4%
			-0.0010	0.0101	94.0%	0.0001	0.0074	93.2%			
3	25	50	0.0018	0.0232	92.2%	0.0064	0.0409	92.6%	-0.0178	0.2124	91.4%
			0.0019	0.0219	93.2%	-0.0010	0.0399	94.8%			
	50	25	0.0036	0.0247	89.2%	0.0089	0.0461	88.2%	-0.0242	0.2751	91.2%
			0.0034	0.0229	91.4%	0.0002	0.0428	91.4%			
	50	50	-0.0012	0.0157	92.2%	0.0047	0.0282	92.2%	-0.0099	0.1384	91.2%
			-0.0007	0.0157	92.4%	0.0002	0.0272	93.4%			
	50	100	-0.0003	0.0111	93.0%	0.0013	0.0202	91.6%	-0.0089	0.1328	92.2%
			0.0000	0.0110	93.2%	-0.0012	0.0196	92.4%			
	100	50	0.0001	0.0111	93.4%	0.0028	0.0207	92.0%	-0.0094	0.0899	94.4%
			0.0003	0.0110	93.6%	0.0004	0.0211	92.2%			
	100	100	0.0000	0.0075	93.4%	0.0016	0.0132	94.0%	0.0004	0.0418	94.4%
			0.0002	0.0075	93.6%	0.0005	0.0133	93.8%			
4	25	50	0.0023	0.0299	88.6%	0.0042	0.0355	91.4%	-0.0103	0.2241	91.0%
			0.0016	0.0265	90.0%	0.0005	0.0357	90.8%			
	50	25	0.0021	0.0289	85.8%	0.0022	0.0353	91.6%	-0.0063	0.1931	92.4%
			0.0021	0.0268	86.8%	-0.0013	0.0357	90.6%			
	50	50	0.0010	0.0173	92.2%	0.0003	0.0231	93.8%	-0.0015	0.0748	94.6%
			0.0014	0.0167	92.4%	-0.0015	0.0228	94.4%			
	50	100	0.0006	0.0114	94.6%	0.0011	0.0153	95.6%	-0.0029	0.0517	93.4%
			0.0009	0.0113	95.2%	-0.0002	0.0154	94.4%			
	100	50	0.0004	0.0123	92.0%	0.0010	0.0160	94.4%	0.0022	0.0436	95.2%
			0.0009	0.0122	92.2%	-0.0001	0.0161	94.4%			
	100	100	-0.0001	0.0083	93.4%	0.0004	0.0109	95.2%	0.0016	0.0292	96.4%
			0.0002	0.0083	93.8%	-0.0002	0.0109	95.0%			

Note. We report the bias, std and CP for the feasible estimates followed those for the infeasible estimates with true γ^0, where CP refers to the coverage probability for the 95% confidence intervals.

ically equivalent. Second, for both sets of estimates, the biases and standard deviations decrease to zero as either N or T increases. In terms of inference, the 95% confidence intervals for the slope coefficients in these DGPs tend to be under-covered, but their performance generally improves as either N or T increases. Third, the biases are of asymptotically smaller order compared to the standard deviations, indicating that the bias-corrected estimator performs as our theory predicts. As the slope coefficients estimator $\widehat{\theta}$ has bias term in DGPs 1-2, we observe slightly larger biases of the bias-corrected estimator compared to that in DGPs 3-4. Fourth, for the estimate of γ^0, the biases and standard deviations tend to decrease as N or T becomes large. The 95% confidence intervals tend to under-cover slightly when N or T is small for DGPs 1, 3, and 4, but the coverage probability approaches the nominal level (95%) quickly as both N and T increase.

2.5.3 Test for the threshold effect

We next consider the test for the presence of threshold effects. We also consider both dynamic and static cases and all four DGPs. The main difference is that now we set $\delta^0 = \delta^0_{NT} = 0,\ 2(NT)^{-1/2}$ and $10(NT)^{-1/2}$ in order to evaluate both the size and local power performance of our test statistic. We consider three frequently-used nominal levels in empirical studies, i.e., 1%, 5% and 10%. As before, we employ the bias-corrected estimates with three potential bias terms corrected in all cases. The number of repetitions is 500.

Table 2.2 reports the test results. First, under $\mathbb{H}_0 : \delta^0 = 0$, the rejection rates for all DGPs are close to the the nominal levels with moderate deviations when N or T is not large enough. But as N and T becomes large, the size of our test improves quickly. Second, the rejection rates increase fast as δ^0 deviates from 0 further and further. Note that our asymptotic estimation theory considers the case where δ^0 is of order $(NT)^{-\alpha}$ with $\alpha < 1/2$. In order to see the local power to approach 1, we need to have a large constant c for

Table 2.2: Rejecting frequency for testing the threshold effect at 1%, 5% and 10% nominal levels

DGP	N	T	$\delta^0_{NT} = 0$			$\delta^0_{NT}= 2/\sqrt{NT}$			$\delta^0_{NT}= 10/\sqrt{NT}$		
			1%	5%	10%	1%	5%	10%	1%	5%	10%
1	50	50	0.018	0.074	0.106	0.100	0.290	0.406	1.000	1.000	1.000
	50	100	0.018	0.068	0.142	0.100	0.268	0.358	1.000	1.000	1.000
	100	50	0.018	0.076	0.110	0.112	0.256	0.356	1.000	1.000	1.000
	100	100	0.006	0.064	0.122	0.094	0.236	0.372	1.000	1.000	1.000
2	50	50	0.024	0.086	0.156	0.460	0.698	0.794	1.000	1.000	1.000
	50	100	0.018	0.084	0.144	0.458	0.674	0.776	1.000	1.000	1.000
	100	50	0.016	0.068	0.118	0.476	0.666	0.750	1.000	1.000	1.000
	100	100	0.012	0.064	0.132	0.400	0.616	0.748	1.000	1.000	1.000
3	50	50	0.024	0.084	0.144	0.150	0.352	0.448	1.000	1.000	1.000
	50	100	0.014	0.052	0.104	0.138	0.332	0.424	1.000	1.000	1.000
	100	50	0.016	0.080	0.162	0.150	0.362	0.470	1.000	1.000	1.000
	100	100	0.010	0.066	0.120	0.140	0.324	0.458	1.000	1.000	1.000
4	50	50	0.026	0.076	0.140	0.136	0.334	0.436	1.000	1.000	1.000
	50	100	0.006	0.072	0.136	0.120	0.332	0.454	1.000	1.000	1.000
	100	50	0.026	0.098	0.158	0.146	0.276	0.384	1.000	1.000	1.000
	100	100	0.024	0.058	0.116	0.122	0.280	0.400	1.000	1.000	1.000

the threshold effect ($c/\sqrt{NT}$). Third, the local power performance of our test for the static panel is similar to that for the dynamic panel.

2.6 Empirical Application

In this section, we apply our method to study the relationship between financial depth and economic growth.

2.6.1 Literature

An important research topic in the economic growth literature is about the relationship between financial depth and economic growth. Recent empirical studies frequently suggest that there is a turning point in the effect of financial development on economic growth; see Levine (2003, 2005), Law and Singh (2014) and Arcand et al. (2015), among others. Levine (2005) provides an extensive survey of the theoretical literature that emphasizes how the services provided by the financial sector would contribute to economic growth. Law and Singh (2014) construct a sample of 87 countries for the period 1980-2010 from several datasets including World Development Indicators (WDI), Penn World

Table 6.3, International Country Risk Guide (ICRG), and the Barro and Lee's dataset. They consider the short panel framework by averaging the time series observations for each country over five-year periods. They find the presence of threshold effect in the finance–growth relationship. In particular, the level of financial development is beneficial to growth only up to a certain level of its value; beyond the level further development of finance tends to adversely affect growth. Similarly, Arcand et al. (2015) find that financial depth starts to have a negative effect on output growth for high-income countries when credit to the private sector reaches 100% of GDP.

Although the effect of financial sector on economic growth has been studied in a wide range, a potential common drawback of the previous studies, such as Law and Singh (2014) and Arcand et al. (2015), is that the cross-sectional dependence among the data that arises from the unobserved common factors is largely ignored. In this section we revisit the relationship between financial depth and economic growth by explicitly modeling the cross-sectional dependence with a factor structure.

2.6.2 Model

We extend the panel threshold model of Law and Singh (2014) to allow for the presence of IFEs. The model is given by

$$\begin{aligned} GROWTH_{it} = {} & a \cdot \mathbf{1}\{FIN_{it} \leq \gamma\} + \beta_1 FIN_{it} \cdot \mathbf{1}\{FIN_{it} \leq \gamma\} \\ & + \beta_2 FIN_{it} \cdot \mathbf{1}\{FIN_{it} > \gamma\} \quad + \varphi' x_{it} + \lambda_i' f_t + e_{it}, \end{aligned}$$

where $GROWTH_{it}$ denotes the economic growth for country i in year t, FIN_{it} denotes the level of financial development for country i in year t, x_{it} is a vector of control variables, and the remaining symbols are the same as in the theory part of the paper.

The above model is slightly different from that given by equation (2.1). First, we do not consider the threshold effect in the coefficient of control vari-

Table 2.3: Descriptive statistics

	Unit of measurement	mean	std dev	median	min	max
Growth	%	3.604	4.520	3.892	-36.700	39.487
Private Sector Credit	log(% of GDP)	3.304	0.900	3.266	0.329	5.570
Liquid Liability	log(% of GDP)	3.620	0.645	3.595	1.495	5.445
Domestic Credit	log(% of GDP)	3.407	0.912	3.363	0.432	5.743
Lag GDP Per Capita	log(US$) 2010 constant price	8.358	1.552	8.227	5.390	11.425
Population Growth	%	1.781	1.084	1.871	-1.475	6.366
Trade openness	log(% of GDP)	4.054	0.600	4.050	1.844	6.090
Government expenditure	log(% of GDP)	2.629	0.369	2.617	1.169	3.772

ables x_{it}'s. Second, as we do not allow for the intercept in the regressors, we only put $\mathbf{1}\{FIN_{it} \leq \gamma\}$ as a regressor. Third, we directly write β_1 and β_2 as the regime one and the regime two coefficients of FIN_{it}.

We follow Law and Singh (2014) and collect annual data from the WDI database between 1971 and 2015. The dataset is a balanced panel with $N = 50$ and $T = 45$. We consider three measures of financial development, namely, *private sector credit* (PSC), *liquid liabilities* (LL), and *domestic credit* (DC). All these three banking sector development indicators are expressed as ratios to GDP. The control variables include: *initial per capita GDP*, *trade openness*, *government expenditure,* and *population growth.* Table 2.3 reports the descriptive statistics of the variables used in our regression. It is seen that the economic growth exhibits a large variation among the 50 countries over the 45 years period under our investigation. In contrast, the three measures of financial development and control variables have relatively small variations.

2.6.3 Test for the presence of threshold effects

To conduct the hypothesis testing for the presence of threshold effects, we first need to specify the number of factors. Note that under the null hypothesis, the model reduces to a standard panel data model with IFEs. So we have a large room of choices on the methods of determining the number of factors, such as Bai and Ng (2002), Onatski (2010) and Ahn and Horenstein (2013). Nevertheless, we find that different methods produce different estimates of R^0

Table 2.4: Test for the presence of threshold effect

R	Private Sector Credit		Liquid Liability		Domestic Credit	
	supW_{NT}	p-value	supW_{NT}	p-value	supW_{NT}	p-value
1	44.661	0.000	39.849	0.000	41.839	0.000
2	74.643	0.000	63.684	0.000	66.610	0.000
3	109.856	0.000	69.551	0.000	84.437	0.000
4	36.041	0.000	22.737	0.000	41.788	0.000
5	38.052	0.000	23.387	0.000	49.027	0.000

for our dataset. For this reason, we are conservative to consider the tests under the various numbers of factors. Specifically, we consider $R = 1, \ldots, 5$ and report the corresponding test statistics and p-values for each case.

Table 2.4 reports the supW_{NT} statistics and the associated bootstrap p-values based on 500 bootstrap replications when all the three measures of financial development are considered. Apparently, all these supW_{NT} statistics suggest that we reject the null hypothesis of no threshold effect at the 1% level.

2.6.4 Number of factors and the test against the additive fixed effects

To estimate the model, we can also consider different choices of R. We apply the SVT method in Section 2.3.6 to determine the number of factors. We also try to modify the existing methods to determine the number of factors in our framework. First, we estimate the model with $R_{\max} = 5$ and obtain the residuals. Then we conduct the eigenvalue distribution (ED) of Onatski (2010), the growth rate (GR) and eigenvalue ratio (ER) of Ahn and Horenstein (2013) and PC and IC of Bai and Ng (2002) to estimate the number of factors in the residuals.

The results are summarized in Table 2.5. ED, GR and ER choose $R = 1$ for all three specifications. PC and IC of Bai and Ng (2002) choose R to be 3 and 2, respectively. Our SVT method chooses $R = 2$ for all three specifications. Moon and Weidner (2015) find that in the linear panel data models with IFEs, we can still estimate the slope coefficients consistently when the number of factors is overspecified. To be conservative, we focus on the case where $R = 3$.

Table 2.5: Number of factors determined by various methods

	Bai and Ng		Onatski-ReStat	AH		SVT
FIN	PC_{p1}	IC_{p1}	ED	ER	GR	
PSC	3	2	1	1	1	2
LL	3	2	1	1	1	2
DC	3	2	1	1	1	2

Note: Bai and Ng refers to Bai and Ng (2002), Onatski-ReStat refers to Onatski (2010), and AH refers to Ahn and Horenstein (2013).

As our SVT method chooses $R = 2$, there is some possibility that the IFEs can be captured by the two-way additive fixed effects (AFEs). In Section E of the online supplement of Miao et al. (2020a), we propose a method to test AFEs against IFEs. We conduct the test for all three specifications here. The test statistics are 18.11, 17.73, and 18.52 with PSC, LL, and DC serving as FIN_{it} respectively. Under the null of AFEs, the test statistic is $N(0,1)$ asymptotically. Therefore, we find a strong evidence to support the IFEs. An implication of this result is that the existing studies may have endogeneity issue because the unobserved heterogeneity are not fully controlled by the traditional two-way additive fixed effects.

2.6.5 Estimation results

Table 2.6 reports the regression results with $R = 3$. The estimates of threshold coefficient γ are 4.2322, 4.304 and 4.559, respectively when *private sector credit*, *liquid liabilities*, and *domestic credit* are used as a measure for the financial development. In terms of the original percentage scale, these numbers correspond to 68.87%, 71.30% and 95.13%, respectively, where, e.g., the first percentage suggests the turning point for the model with *private sector credit* occurs when the private sector credit over the GDP ratio reaches 68.87%, a number that is substantially smaller than 100%, and a number suggested by Arcand et al. (2015). For these three estimates of γ, we find there are about 84.3%, 85.1% and 87.5% of observations in the data that are smaller than the corresponding estimate of γ. In some sense, the estimated threshold

Table 2.6: Estimates of the slope and threshold coefficients

	Model 1		Model 2		Model 3	
Fin Development	PSC		LL		DC	
Threshold Estimate:						
Threshold level ($\widehat{\gamma}$)	4.232 (log68.87)		4.304 (log74.00)		4.559 (log95.45)	
95% CI	[4.106, 4.268]		[4.225, 4.360]		[4.542, 4.567]	
Sample Quantile	84.3%		85.1%		87.5%	
Impact of finance:						
Regime one ($\widehat{\beta}_1$)	0.274	(0.282)	0.298	(0.393)	0.078	(0.267)
Regime two ($\widehat{\beta}_2$)	-3.101	(0.698)	-3.820	(0.967)	-3.014	(0.965)
Impact of Covariates:						
Lag GDP Per Capita	-2.595	(0.490)	-2.561	(0.507)	-2.595	(0.497)
Population Growth	1.353	(0.235)	1.188	(0.229)	1.344	(0.237)
Trade Openness	3.098	(0.350)	3.143	(0.351)	3.204	(0.356)
Gov Expenditure	-2.827	(0.423)	-2.867	(0.436)	-2.785	(0.430)
Intercept	-13.618	(3.406)	-17.031	(4.275)	-12.784	(4.556)

Note: The values without parentheses (the left column) are the least square estimates and the values in parentheses (the right column) are the corresponding standard errors.

coefficient is far apart from the tail of the distribution of the threshold variable. In Table 2.6, we also report the 95% confidence intervals that are based on the likelihood-ratio test. The three confidence intervals are quite narrow due to the fact that threshold effects are not so small.

The estimates of β_1 and β_2 in model (2.8) suggest that the financial development is a positive but not statistically significant determinant of economic growth if it is less than a certain threshold level, and its effect becomes negative and statistically significant when it is higher than the threshold level. This empirical finding is roughly in line with that in Law and Singh (2014) and supports the conventional wisdom that more finance is definitely not always better and it tends to harm economic growth after a turning point. However, we emphasize that although the final results are changed not so much in comparison with Law and Singh (2014), our model should be used in this type of empirical studies since we find strong evidence that both threshold effects and IFEs are present in the data.

2.7 Conclusion

In this paper, we consider the least squares estimation of the panel threshold models with IFEs. We study the asymptotic properties of the least squares estimators in the shrinking threshold effect framework and propose a likelihood ratio test for the threshold parameter in the model. We also propose a test for the presence of threshold effects. Our simulations suggest that our estimators and test statistics perform well in finite samples. We apply our method to study the effect of financial development on economic growth and find strong evidence to support the proposed model.

There are several interesting topics for further research. First, it is interesting to consider panel threshold regressions with both IFEs and endogeneity. Endogeneity has become a serious concern in recent cross-sectional threshold models; see, e.g., Yu and Phillips (2018). The extension to our framework will be complicated by the presence of IFEs. Second, we only consider a panel threshold model where the regression parameters exhibit homogeneity over both cross-sectional and time dimensions. In the large N and large T setup, there is a possibility for unobserved parameter heterogeneity or latent group structure over the cross-sectional dimensions (e.g., Ando and Bai (2016), Su et al. (2016) and Su and Ju (2018)) and structural changes along the time dimension (e.g., Qian and Su (2016), Li et al. (2016), and Okui and Wang (2020)). We leave these topics for future research.

Chapter 3
Panel Threshold Regressions with Latent Group Structures

3.1 Introduction

Threshold models have a wide variety of applications in economics; see Durlauf and Johnson (1995), Potter (1995), Kremer et al. (2013), and Arcand et al. (2015), among others. In both the cross sectional and time series framework, asymptotic theory for estimation and inference in threshold models has been well developed. See, e.g., Chan (1993) and Hansen (2000) on asymptotic distribution theory for the threshold estimator in the fixed-threshold-effect and shrinking-threshold-effect frameworks, respectively, and Hansen (2011) for a review on the development and applications of threshold regression models in economics. Both Chan (1993) and Hansen (2000) require the exogeneity of the regressors. Endogeneity has been considered in some existing papers; see, e.g., Caner and Hansen (2004), Kourtellos et al. (2016), and Yu and Phillips (2018). In the panel setup, Hansen (1999) studies static panel threshold models with exogenous regressors and threshold variables; Seo and Shin (2016) propose a GMM method to estimate dynamic panel threshold models with additive fixed effects, where either the regressors or the threshold variables can be endogenous; and Miao et al. (2020a) study estimation and inference in dynamic panel threshold regression with interactive fixed effects.

All existing studies in panel threshold models assume that the slope coefficients and threshold parameters are common across all individual units.

However, such an assumption of homogeneity is vulnerable in practice given that individual heterogeneity has been widely documented in empirical studies using panel data. See, e.g., Durlauf (2001) and Su and Chen (2013) for cross-country evidence and Browning and Carro (2007) for ample microeconomic evidences. In panel threshold regressions, heterogeneity can exist in not only the slopes but also the threshold coefficients. Neglecting latent heterogeneity in any aspect can lead to inconsistent estimation and misleading inferences. In particular, pooling individuals with different threshold values would bias the threshold and the slope coefficient estimation, and it can even lead to a failure in detecting any threshold effect in finite samples since heterogeneous threshold effects may offset each other. Even if all units share the same threshold coefficient, ignoring heterogeneity in the slopes would also lead to inconsistent estimates.

In this paper, we propose a new panel threshold model that allows the slope and threshold coefficients to vary across individual units. We model individual heterogeneity via a grouped pattern, such that all the members within the same group share the same slope and threshold coefficients, whereas these coefficients can differ across groups in an arbitrary manner. Hence, the latent group structure may result from two sources of heterogeneity: that in the slope coefficients and that in the threshold level coefficients. We allow the group membership structure (i.e., which individuals belong to which group) to be unknown and estimated from the data. We refer to our model as a *panel structure threshold regression* (PSTR) model.

Using a panel structure model that imposes a group pattern is a convenient way to model unobserved heterogeneity, and they have recently received much attention; see Lin and Ng (2012), Bonhomme and Manresa (2015), Ando and Bai (2016, 2017), Su et al. (2016), Lu and Su (2017), Liu et al. (2020), Su and Ju (2018), Su et al. (2019), and Okui and Wang (2020), among others. An important advantage of the panel structure model is that it allows flexible forms of unobserved heterogeneity while remaining parsimonious at the same time. As group structure is latent in such a model, the determination of an individual's membership is the key question. Several approaches have been proposed to address this issue. Sun (2005), Kasahara and Shimotsu (2009) , and Browning and Carro (2007) consider finite mixture models. Su et al. (2016) propose a variant of the Lasso procedure (C-Lasso) to achieve a classification in this regard, and this method has been extended to allow for two-way component errors, interactive fixed effects, nonstationary regressor, and semiparametric specification, respectively, in Lu and Su (2017), Su and Ju (2018), Huang et al. (2020) , and Su et al. (2019). Lin and Ng (2012), Bonhomme and Manresa (2015), Sarafidis and Weber (2015), and Liu et al. (2020) extend the K-means algorithms to the panel regression framework. Wang et al. (2018) and Wang and Su (2020) propose to identify the latent group structure based on the Lasso or spectral clustering techniques in the statistics literature. In the nonparametric literature, Vogt and Linton (2017, 2020) consider procedures to estimate the unknown group structures for nonparametric regression curves.

To estimate the PSTR model, we consider a least-squares-type estimator that minimizes the sum of squared errors. We choose the least-squares approach for classification because the group, slope, and threshold parameters can be estimated in the same framework, which facilitates the theory. The disadvantage is that we cannot allow for endogeneity in the regressors and threshold variables. Cases with endogenous regressors or threshold variables require different and more complicated analysis and will be left for future research. Due to the presence of the latent group structure and threshold parameters, we do not have an analytically closed-form solution to the problem. We propose to employ an EM-type iterative algorithm to find the solution with multiple starting values. Under some regularity conditions, we show that our estimators of the slope and threshold coefficients are asymptotically equivalent to the corresponding infeasible estimators of the group-specific parameters that are obtained by using individual group identity information.

To study the asymptotic properties of the estimators of the threshold coefficients, we follow the lead of Hansen (2000) and consider the shrinking-threshold-effect framework, where the threshold effect is diminishing as the sample size approaches infinity. In this framework, we can make inferences regarding each threshold parameter by constructing a likelihood ratio (LR) statistic. We show that the LR statistics are asymptotically pivotal in the case of conditional homoskedasticity and that they depend on a scale nuisance parameter otherwise. Such a scale parameter can be consistently estimated nonparametrically when conditional heteroskedasticity is suspected.

We also consider two specification test statistics. The first one is designed to test the homogeneity of the threshold parameters across each group via the LR principle. The corresponding LR test statistic is non-standard and involves a linear combination of two-sided Brownian motions. We show how one can obtain the simulated p-value with estimated parameters in our discussion. This test is useful since pooling units, if their threshold coefficients pass the homogeneity test, improves the efficiency of threshold estimation, especially in small samples. The second is designed to test the absence of the threshold effect under the null by adopting the method proposed by Hansen (1996). In our latent group structure framework, one may suspect the presence of a subset of threshold effects among all groups, and we also need to take into account the uncertainty caused by the unknown group structure when studying the asymptotic behavior of the test.

We evaluate the finite-sample performance of the proposed tests and estimation methods via extensive simulation studies. First, the proposed information criterion can determine the correct number of groups with a large probability, regardless of whether any threshold effect is present. Given the number of groups, the next task is to test the existence of threshold effects. Our proposed test has an appropriate size and non-trivial power in detecting the threshold effect. The power is an increasing function of the strength of both the threshold effect and sample size. A nice feature of the test is that it performs well regardless of whether the threshold is heterogeneous across units. If the threshold effect is present, one can further test whether the threshold parameters differ across groups. We demonstrate that our test for the homogeneity of the threshold is also well-behaved in terms of size and its power improves as the degree of threshold heterogeneity and sample sizes increase. Finally, after the model and the number of groups are specified, we can proceed with parameter estimation. Our estimation method performs well in heterogeneous panels with threshold effects in finite samples. With this method, we can precisely estimate group membership, and the clustering accuracy improves as the number of time periods increases. Both the threshold parameters and slope coefficients can be precisely estimated. Moreover, we find that when the threshold parameters are homogeneous across groups, pooling observations with a common threshold does improve the efficiency of threshold estimation, which in turn highlights the importance of testing the homogeneity of the threshold parameters.

We illustrate the usefulness of our methods through two real-data examples. First, we revisit the relationship between capital market imperfections and firms' investment behavior. We document a large degree of heterogeneity in firms' investment behavior, which is bound by various types of financial constraints, such as cash flow, Tobin's Q, and leverage. Such heterogenous threshold effects cannot be captured by the conventional panel threshold regressions. Next, we examine the impact of bank regulation, particularly branch deregulation, on income inequality in US, allowing observed and unobserved heterogeneity in their impact. We find a group pattern of heterogeneity in the impact of deregulation across states even after controlling for the threshold effect. The group structure coincides with geographic locations to some extent but not perfectly, and the threshold effects appear to be salient in each group.

This application again demonstrates the usefulness of the PSTR since it allows us to capture both observed heterogeneity through thresholds and unobserved heterogeneity through the latent group structure.

The remainder of the paper is organized as follows. In Section 3.2, we introduce our model and estimation method. In Section 3.3, we introduce the assumptions and examine the asymptotic properties of the estimators of the latent group structure and the slope and threshold coefficients. In Section 3.4, we introduce the inference procedure on the threshold parameters and propose a specification test for the homogeneity of the threshold parameters across groups. In Section 3.5, we consider the specification test for the presence of threshold effects. In Section 3.6, we propose a BIC-type information criterion to determine the number of groups. We conduct Monte Carlo experiments to evaluate the finite sample performance of our estimators and tests in Section 3.7. We apply our model to study the relationship between investment and financing constraints and the relationship between bank regulation and income distribution in Section 3.8. Section 3.9 concludes. The proofs of the main results in the paper are relegated to the Appendix. Further technical details can be found in the online supplemental materials.

To proceed, we adopt the following notation. The indicator function is denoted as $\mathbf{1}(\cdot)$. $\mathbf{0}_{a\times b}$ denotes an $a \times b$ matrix of zeros. For two constants a and b, we denote $\max(a, b)$ as $a \vee b$ and $\min(a, b)$ as $a \wedge b$. For an $m \times n$ real matrix A, we denote its transpose as A' and its Frobenius norm as $\|A\|$ $(\equiv [tr(AA')]^{1/2})$ where $\equiv$ means "is defined as". For a real symmetric matrix A, we denote its minimum eigenvalue as $\lambda_{\min}(A)$. The operators $\xrightarrow{p}$ and $\xrightarrow{d}$ denote convergence in probability and distribution, respectively. We use $(N, T) \to \infty$ to denote the joint convergence of N and T when N and T pass to infinity simultaneously. Alternatively, as the co-editor suggests, one can consider the pathwise asymptotics as in Phillips and Moon (1999) and Vogt and Linton (2020).

3.2 The Model and Estimates

In this section we first present the panel threshold model with latent group structures and then introduce the estimators of all the parameters in the model.

3.2.1 The Model

Let N denote the number of cross-sectional units and T the number of time periods. We consider the model

$$y_{it} = x'_{it}\beta^0_{g^0_i} + x'_{it}\delta^0_{g^0_i} \cdot d_{it}(\gamma^0_{g^0_i}) + \mu_i + \varepsilon_{it}, \quad i = 1, ..., N,\ t = 1, ..., T, \tag{3.1}$$

where x_{it} is a $K \times 1$ vector of observable regressors, $d_{it}(\gamma) \equiv \mathbf{1}(q_{it} \leq \gamma)$, q_{it} is a scalar threshold variable, μ_i is the individual fixed effects and ε_{it} is the idiosyncratic error term. Note that we allow both the slope and threshold coefficients to be group specific: γ^0_g is a scalar threshold coefficient, β^0_g is a $K \times 1$ vector of regression coefficients that lies in a compact parameter space $\mathcal{B}$, and δ^0_g is a $K \times 1$ vector of threshold-effect coefficients for $g \in \mathcal{G} \equiv \{1, ..., G\}$, where G is a fixed integer known as the number of groups. The group-membership variable $g^0_i \in \mathcal{G}$ indicates to which group individual unit i belongs. This group-membership variable is unknown and has to be estimated from the data. All members in group g have the same coefficients $(\beta^{0\prime}_g, \delta^{0\prime}_g, \gamma^0_g)'$. We assume $\gamma^0_g \in \mathbf{\Gamma} = [\underline{\gamma}, \overline{\gamma}]$ for all $g \in \mathcal{G}$, where $\underline{\gamma}$ and $\overline{\gamma}$ are two fixed constants. Following the lead of Hansen (2000), we will work in the shrinking-threshold-effect framework by assuming that $\delta^0_g \equiv \delta^0_{g,NT} \to 0$ as $(N, T) \to \infty$ for each $g \in \mathcal{G}$ unless specified otherwise.

Let $\mathbf{D} \equiv (\gamma_1, ..., \gamma_G)' \in \Gamma^G$, $\mathbf{G} \equiv (g_1, ..., g_N)' \in \mathcal{G}^N$ and $\mathbf{\Theta} \equiv (\theta'_1, ..., \theta'_G)' \in \mathcal{B}^G$, where $\theta_g \equiv (\beta'_g, \delta'_g)' \in \mathcal{B} \subset \mathbb{R}^{2K}$. For any given group structure $\mathbf{G}$, we let $\mathbf{G}_g = \{i |\ g_i = g, 1 \leq i \leq N\}$ be the index set of the members in group $g \in \mathcal{G}$. We denote the true parameters as $(\mathbf{\Theta}^0, \mathbf{D}^0, \mathbf{G}^0)$, where $\mathbf{\Theta}^0 \equiv (\theta^{0\prime}_1, ..., \theta^{0\prime}_G)'$, $\mathbf{D}^0 \equiv (\gamma^0_1, ..., \gamma^0_G)'$ and $\mathbf{G}^0 \equiv (g^0_1, ..., g^0_N)'$. Analogously, we denote the true members in group $g \in \mathcal{G}$ by $\mathbf{G}^0_g = \{i |\ g^0_i = g, 1 \leq i \leq N\}$.

For the moment, we assume that the true number of groups G^0 is known and given by G. In Section 3.6, we will discuss how to determine G^0 in practice.

3.2.2 Estimation

To remove the individual-specific fixed effects μ_i, we employ the usual within-transformation which leads to

$$\tilde{y}_{it} = \tilde{x}_{it}'\beta^0_{g_i^0} + \tilde{x}_{it}(\gamma^0_{g_i^0})'\delta^0_{g_i^0} + \tilde{\varepsilon}_{it}, \quad i = 1, ..., N, \; t = 1, ..., T, \tag{3.2}$$

where $\tilde{x}_{it}(\gamma) \equiv x_{it}d_{it}(\gamma) - \frac{1}{T}\sum_{s=1}^{T} x_{is}d_{is}(\gamma)$, and $\tilde{x}_{it}$, $\tilde{y}_{it}$ and $\tilde{\varepsilon}_{it}$ are defined analogously. Let $z_{it}(\gamma) \equiv (x_{it}', x_{it}'d_{it}(\gamma))'$ and $\tilde{z}_{it}(\gamma) \equiv z_{it}(\gamma) - \frac{1}{T}\sum_{s=1}^{T} z_{is}(\gamma)$. Then the model in (3.2) can be rewritten as

$$\tilde{y}_{it} = \tilde{z}_{it}(\gamma^0_{g_i^0})'\theta^0_{g_i^0} + \tilde{\varepsilon}_{it}, \quad i = 1, ..., N, \; t = 1, ..., T. \tag{3.3}$$

Given G, we can obtain the following least squares estimator of $(\Theta, \mathbf{D}, \mathbf{G})$:

$$(\hat{\Theta}, \hat{\mathbf{D}}, \hat{\mathbf{G}}) = \underset{(\Theta, \mathbf{D}, \mathbf{G}) \in \mathcal{B}^G \times \Gamma^G \times \mathcal{G}^N}{\arg\min} \mathcal{Q}(\Theta, \mathbf{D}, \mathbf{G}),$$

where

$$\mathcal{Q}(\Theta, \mathbf{D}, \mathbf{G}) = \sum_{i=1}^{N}\sum_{t=1}^{T} \left[\tilde{y}_{it} - \tilde{z}_{it}(\gamma_{g_i})'\theta_{g_i}\right]^2 . \tag{3.4}$$

For any given threshold $\mathbf{D}$ and group structure $\mathbf{G}$, the slope coefficients θ_g, $g = 1, ..., G$, can be estimated by

$$\hat{\theta}_g(\mathbf{D}, \mathbf{G}) = \left(\sum_{i=1}^{N}\sum_{t=1}^{T} \mathbf{1}(g_i = g)\tilde{z}_{it}(\gamma_g)\tilde{z}_{it}(\gamma_g)'\right)^{-1} \sum_{i=1}^{N}\sum_{t=1}^{T} \mathbf{1}(g_i = g)\tilde{z}_{it}(\gamma_g)\tilde{y}_{it}.$$

Concentrating out Θ, we can estimate the threshold $\mathbf{D}$ and group structure $\mathbf{G}$ by

$$(\hat{\mathbf{D}}, \hat{\mathbf{G}}) = \underset{(\mathbf{D}, \mathbf{G}) \in \Gamma^G \times \mathcal{G}^N}{\arg\min} \dot{\mathcal{Q}}(\mathbf{D}, \mathbf{G}), \tag{3.5}$$

where $\dot{\mathcal{Q}}(\mathbf{D}, \mathbf{G}) \equiv \mathcal{Q}(\hat{\Theta}(\mathbf{D}, \mathbf{G}), \mathbf{D}, \mathbf{G})$ and $\hat{\Theta}(\mathbf{D}, \mathbf{G}) = (\hat{\theta}_1(\mathbf{D}, \mathbf{G})', ..., \hat{\theta}_G(\mathbf{D}, \mathbf{G})')'$.

To find the solution to the above optimization problem, we need to search over the space of $(\mathbf{D}, \mathbf{G})$ to minimize the objective function in (3.5). We propose to employ the following EM-type iterative algorithm to conduct the searching process:

1. Set $\mathbf{G}^{(0)}$ as a random initialization of the group structure $\mathbf{G}$ and let $s = 0$.

2. Given a s conduct:

   (a) For given $\mathbf{G}^{(s)}$, compute

   $$\mathbf{D}^{(s)} = \arg\min{}_{\mathbf{D}\in\Gamma^G} \dot{\mathcal{Q}}(\mathbf{D}, \mathbf{G}^{(s)}).$$

   (b) For given $\mathbf{D}^{(s)} = \{\gamma_g^{(s)},\, g = 1, ..., G\}$ and $\mathbf{G}^{(s)} = \{g_i^{(s)},\, i = 1, ..., N\}$, compute the slope coefficients for each group $g \in \mathcal{G}$

   $$\hat{\theta}_g^{(s)} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T} \mathbf{1}(g_i^{(s)} = g)\tilde{z}_{it}(\gamma_g^{(s)})\tilde{z}_{it}(\gamma_g^{(s)})'\right)^{-1} \sum_{i=1}^{N}\sum_{t=1}^{T} \mathbf{1}(g_i^{(s)} = g)\tilde{z}_{it}(\gamma_g^{(s)})\tilde{y}_{it}.$$

   (c) Compute for all $i \in \{1, \ldots, N\}$,

   $$g_i^{(s+1)} = \arg\min_{g\in\mathcal{G}} \sum_{t=1}^{T} [\tilde{y}_{it} - \tilde{z}_{it}(\gamma_g^{(s)})'\hat{\theta}_g^{(s)}]^2.$$

   (d) Set $s = s + 1$. Repeat Steps (a)-(c) until numerical convergence.

The above algorithm is similar to Algorithm 1 in Bonhomme and Manresa (2015, BM hereafter) and it alternates among three steps. Steps (a) and (b) are the "update" steps where one updates the estimates of the threshold parameter and those of the slope coefficients in turn. Step (c) is an "assignment" step where each individual i is re-assigned to the group $g_i^{(s+1)}$. The objective function is non-increasing in the number of iterations and we find through simulations that numerical convergence is typically very fast. Nevertheless, it is hard to ensure that the obtained solution is globally optimal because it depends on the chosen starting values. In practice, one can start with multiple random starting values and select the solution that yields the lowest objective value.

3.3 Asymptotic Theory

In this section, we study the asymptotic properties of the estimators of the group structure, slope and threshold parameters. We first show the consistency of the group structure estimator and then establish the asymptotic properties of the estimators of the slope and threshold coefficients.

3.3.1 The estimator of the group structure

We establish the consistency of the group structure estimator in this subsection. Let $\mathcal{F}_{NT,t} \equiv \sigma(\{(x_{it}, q_{it}, \varepsilon_{i,t-1}), (x_{i,t-1}, q_{i,t-1}, \varepsilon_{i,t-2}), ...\}_{i=1}^{N})$ where $\sigma(A)$ denotes the minimal sigma-field generated from A. Let $X_i = (x_{i1}, ..., x_{iT})'$, $\varepsilon_i = (\varepsilon_{i1}, ..., \varepsilon_{iT})'$ and $q_i = (q_{i1}, ..., q_{iT})'$. We use N_g to denote the number of individuals belonging to group g : $N_g = \left|\mathbf{G}_g^0\right|$. That is, $\left|\mathbf{G}_g^0\right|$ denotes the cardinality of $\mathbf{G}_g^0$. For any group structure $\mathbf{G}$, let

$$M_{NT}(g, \tilde{g}, \mathbf{D}, \mathbf{G}) \equiv \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{1}(g_i^0 = g)\mathbf{1}(g_i = \tilde{g})\tilde{z}_{it}(\gamma_{\tilde{g}})\tilde{z}_{it}(\gamma_{\tilde{g}})'.$$

Let $0 < C < \infty$ denote a generic constant that may vary across places. Let $\max_i = \max_{1\leq i\leq N}$, $\max_t = \max_{1\leq t\leq T}$ and $\max_{i,t} = \max_{1\leq i\leq N}\max_{1\leq t\leq T}$. We first make the following assumptions.

Assumption A.1: (i.1) For each $i = 1, ..., N$, $t = 1, ..., T$, $E(\varepsilon_{it}|\mathcal{F}_{NT,t-1}) = 0$ a.s., or (i.2) for each $i = 1, ..., N$, $t = 1, ..., T$, $E(\varepsilon_{it}|X_i, q_i) = 0$ a.s.;

(ii) $\{(x_{it}, q_{it}, \varepsilon_{it}) : t = 1, 2, \ldots\}$ are mutually independent of each other across i;

(iii) The process $\{(x_{it}, q_{it}, \varepsilon_{it}), \ t \geq 1\}$ is a strong mixing process with mixing coefficients $\alpha_i[t]$ satisfying $\max_{1\leq i\leq N}\alpha_i[t] \leq c_\alpha\rho^t$ for some constants $c_\alpha > 0$ and $\rho \in (0, 1)$.

(iv) The parameter space $\mathcal{B}$ and Γ are compact so that $\sup_{\theta\in\mathcal{B}}\|\theta\| \leq C$ and $\Gamma= [\underline{\gamma}, \overline{\gamma}]$;

(v) $\max_{i,t} E\|x_{it}\|^{8+\epsilon_0} \leq C$ and $\max_{i,t} E(\|\varepsilon_{it}\|^{8+\epsilon_0}) \leq C$ for some $\epsilon_0 > 0$;

(vi) The threshold effect satisfies $\delta_g^0 = (NT)^{-\alpha}C_g^0$ for some constants $\alpha \in (0, 1/2)$ and $C_g^0 \neq 0$ for all $g \in \mathcal{G}$.

Assumption A.2: There exists a constant $\underline{c}_\lambda > 0$ such that for all $g \in \mathcal{G}$,

$$\Pr\left(\inf_{(\mathbf{G},\mathbf{D})\in\mathcal{G}^N\times\Gamma^G}\max_{\tilde{g}\in\mathcal{G}}\{\lambda_{\min}[M_{NT}(g, \tilde{g}, \mathbf{D}, \mathbf{G})]\} > \underline{c}_\lambda\right) \to 1 \text{ as } (N, T) \to 1.$$

Assumption A.3: (i) For all $g, \tilde{g} \in \mathcal{G}$ with $g \neq \tilde{g}$, we have $\left\|\beta_g^0 - \beta_{\tilde{g}}^0\right\| > \underline{c}_\beta$ for some constant $\underline{c}_\beta > 0$;

(ii) For any $g \neq \tilde{g}$ and $1 \leq i \leq N$, we have $E[\tilde{x}_{it}'(\beta_{\tilde{g}}^0 - \beta_g^0)]^2 \equiv \underline{c}_{g\tilde{g},i} \geq \underline{c}_{g\tilde{g}}$ for some constant $\underline{c}_{g\tilde{g}} > 0$;

(iii) For all $g \in \mathcal{G} : \lim_{N\to\infty} N_g/N = \pi_g > 0$.

(iv) $N = O(T^2)$ and $T = O(N^2)$ as $(N, T) \to \infty$.

Assumption A.1(i)–(iii) is similar to Assumption A.2(a)–(c) in Su and Chen (2013) . The major differences lie in four aspects. First, Su and Chen (2013) consider linear panel data models with interactive fixed effects and the sigma-field $\mathcal{F}_{NT,t}$ there also incorporates the factors and factor loadings, whereas we consider the panel threshold regression models with a latent group structure and the additive fixed effects. Second, Su and Chen (2013) only consider Assumption A.1(i.1) and allow for lagged dependent variables to appear in the regressor vector. Here we consider both scenarios in Assumption A.1(i): the martingale difference sequence (m.d.s.) condition in A.1(i.1) and the strict exogeneity condition in A.1(i.2), where we allow for dynamic panels in the first scenario and assume strict exogeneity in the second scenario. In the second scenario, we allow for serial correlation of an unknown form in the error term. When A.1(i.1) holds, we have asymptotic biases for the estimators of the slope coefficients. When A.1(i.2) holds and serial correlation is likely to appear, we have to use the HAC estimator for the asymptotic variance of the slope estimators. Third, due to the potential appearance of the lagged dependent variables in the regression model, Su and Chen (2013) use the notion of conditional strong mixing for the process while we focus on the case of unconditional strong mixing in our model in Assumption A.1(iii). In other words, we follow Hahn and Kuersteiner (2011) and treat the fixed effects $\{\mu_i\}$ to be nonrandom in our setting in the dynamic case. If $\{\mu_i\}$ are random, we can modify the unconditional strong mixing conditions to the conditional strong mixing conditions as in Su and Chen (2013). Fourth, Su and Chen

(2013) assume conditional cross-sectional independence whereas we assume cross-sectional independence in Assumption A.1(ii).

A.1(iv) is imposed to facilitate the proof as we do not have closed form solutions to our optimization problem. Assumption A.1(v) imposes some moment conditions on the regressors and error terms, which are weaker than the exponential tail assumption in BM (2015). Assumption A.1(vi) assumes shrinking threshold effect as in Hansen (2000). In this framework, the asymptotic distribution of the estimator of γ_g is pivotal up to a scale effect, which facilities the inference procedure. In part E of the online supplement we study the asymptotic properties of our estimators in the fixed threshold effect framework. In the latter case, the inference becomes difficult in practice and one can consider extending the smoothed least squares estimation of Seo and Linton (2007) to our PSTR model.

Assumption A.2 is similar to Assumption 1(g) in BM (2015). Given any conjectured group structure $\mathbf{G}$ and for each $g \in \mathcal{G}$, we cannot assume $\lambda_{\min}[M_{NT}(g, \tilde{g}, \mathbf{D}, \mathbf{G})] > \underline{c}_{\lambda}$ for any $\tilde{g} \in \mathcal{G}$ due to the possibility of very few individuals assigned to be in group $\tilde{g}$. However, there exists some group $\tilde{g} \in \mathcal{G},$ in which a positive proportion of N members are assigned. As BM (2015) remark, such an assumption is reminiscent of the full rank condition in standard regression models.

Assumption A.3(i) and (iii) parallels Assumption A1(vi)–(vii) in Su, Shi, and Phillips (2016, SSP hereafter). A.3(i) requires that the group-specific slope coefficients be separated from each other, and it can be relaxed to allow the differences between the group-specific slope coefficients to shrink to zero at some slow rates at the cost of more lengthy arguments. It is worth emphasizing that the latent group structure is identified through the separation of group-specific slope coefficients and we find that the potential separation of the threshold parameters is not necessary; see the remarks after Theorem 3.1 for futher discussions. A.3(iii) implies that each group has an asymptotically non-negligible proportion of individuals as $N \to \infty$. Noting that

$E[\tilde{x}_{it}'(\beta_{\tilde{g}}^0 - \beta_g^0)]^2 = (\beta_{\tilde{g}}^0 - \beta_g^0)' E(\tilde{x}_{it}\tilde{x}_{it}')(\beta_{\tilde{g}}^0 - \beta_g^0)$, A.3(ii) is automatically satisfied under A.3(i) provided that the minimum eigenvalue of $E(\tilde{x}_{it}\tilde{x}_{it}')$ is bounded away from zero. Apparently, x_{it} cannot contain time-invariant regressors under Assumption A.3(ii). Assumption A.3(iv) puts some restrictions on the relative magnitudes of N and T, which can be easily met in many macro and financial applications. If we follow BM (2015) and assume exponentially-decaying tails, we can relax the conditions on (N, T) to $N/T^v \to 0$ as $(N, T) \to \infty$ for some $v > 0$. If we follow Vogt and Linton (2019) and consider the pathwise asymptotics by setting $N = g(T)$ for some divergent function $g(\cdot)$ and passing $T \to \infty$. Then Assumption A.3(iv) can be satisfied when $g(T)/T^2 + T/g(T)^2$ converges to some positive finite constant as $T \to \infty$.

The following theorem reports the consistency of the estimators of the group membership for all individuals.

Theorem 3.1. *Suppose that Assumptions A.1–A.3 hold. Then*

$$\Pr\left(\sup_{1\le i\le N} \mathbf{1}(\hat{g}_i \neq g_i^0) = 1\right) \to 0 \text{ as } (N, T) \to \infty.$$

Theorem 3.1 is similar to Theorem 2 of BM (2015). This theorem states that as $(N, T) \to \infty$, we can correctly estimate the group structure with probability approaching one (w.p.a.1). From the proof of the above theorem, we can see that the identification of the true group structure highly hinges on Assumption A.3(i). In particular, since we permit $\delta_g^0 = \delta_{g,NT}^0 \to 0$ as $(N, T) \to \infty$ under the shrinking-threshold-effect framework, the proof of Theorem 3.1 mainly relies on the differences of β_g^0's across groups. In this case, as long as the slope coefficients in one regime are separate from each other across the G groups, they are also separate from each other asymptotically in the other regime and whether the threshold parameters in different groups differ from each other does not matter. In other words, the threshold parameters do not need to separate from each other. In the online Supplementary Material, we give a proof of Theorem 3.1 under the fixed-threshold-effect framework. We show that in that case, either the separation among θ_g^0's or that among γ_g^0's is sufficient for identifying the latent group structure under some regularity conditions. To stay focused, we will work in the shrinking-threshold-effect framework below.

3.3.2 The estimators of the slope and threshold coefficients

Given the fact that the latent group structure can be recovered from the data at a sufficiently fast rate (see Lemma A.3 in the appendix), we will show that the estimators of the slope and threshold coefficients are asymptotically equivalent to the infeasible estimators that are obtained as if the true group structure were known. Then we derive the asymptotic distributions of the coefficient estimators.

To establish the asymptotic equivalence, we add some notation. Let $\tilde{x}_{it}(\gamma,\gamma^*) = \tilde{x}_{it}(\gamma) - \tilde{x}_{it}(\gamma^*)$. Let $f_{it}(\cdot)$ denote the probability density function (PDF) of q_{it}. For all $g \in \mathcal{G}$, define

$$\begin{aligned}
w_g(\gamma) &= \frac{1}{N_g T}\sum_{i\in\mathbf{G}_g^0}\sum_{t=1}^{T}\tilde{z}_{it}(\gamma)\tilde{z}_{it}(\gamma)', \\
\tilde{w}_g(\gamma) &= \frac{1}{N_g T}\sum_{i\in\mathbf{G}_g^0}\sum_{t=1}^{T}\tilde{x}_{it}(\gamma,\gamma_g^0)\tilde{x}_{it}(\gamma,\gamma_g^0)' \\
&\quad -\frac{1}{N_g T}\sum_{i\in\mathbf{G}_g^0}\sum_{t=1}^{T}\tilde{x}_{it}(\gamma,\gamma_g^0)\tilde{z}_{it}(\gamma)'\left[w_g(\gamma)\right]^{-1}\frac{1}{N_g T}\sum_{i\in\mathbf{G}_g^0}\sum_{t=1}^{T}\tilde{z}_{it}(\gamma)\tilde{x}_{it}(\gamma,\gamma_g^0)'. \\
M_{g,NT}(\gamma) &= \frac{1}{N_g T}\sum_{i\in\mathbf{G}_g^0}\sum_{t=1}^{T}E[x_{it}x_{it}'d_{it}(\gamma)], \\
D_{g,NT}(\gamma) &= \frac{1}{N_g T}\sum_{i\in\mathbf{G}_g^0}\sum_{t=1}^{T}E(x_{it}x_{it}'|q_{it}=\gamma)f_{it}(\gamma), \text{ and} \\
V_{g,NT}(\gamma) &= \frac{1}{N_g T}\sum_{i\in\mathbf{G}_g^0}\sum_{t=1}^{T}E(x_{it}x_{it}'\varepsilon_{it}^2|q_{it}=\gamma)f_{it}(\gamma).
\end{aligned}$$

Let $M_g(\gamma) = \lim_{(N,T)\to\infty} M_{g,NT}(\gamma)$, $D_g(\gamma) = \lim_{(N,T)\to\infty} D_{g,NT}(\gamma)$, $V_g(\gamma) = \lim_{(N,T)\to\infty} V_{g,NT}(\gamma)$, $D_g^0 = D_g(\gamma_g^0)$, and $V_g^0 = V_g(\gamma_g^0)$. We add the following two assumptions.

Assumption A.4: (i) There exists a constant $\tau > 0$ such that

$$\Pr\left(\min_{\gamma\in\Gamma}\lambda_{\min}\left[w_g(\gamma)\right] \geq \tau\right) \to 1$$

as $(N,T) \to \infty$ for all $g \in \mathcal{G}$;

(ii) There exists a constant $\tau > 0$ such that

$$\min_{\gamma\in\Gamma}\left\{\Pr(\lambda_{\min}[\tilde{w}_g(\gamma)] \geq \tau \min[1, \left|\gamma-\gamma_g^0\right|])\right\} \to 1$$

as $(N,T)\to\infty$ for each $g\in\mathcal{G}$.

Assumption A.5: (i) $\max_{\gamma\in\Gamma}\max_{i,t} E(\|\xi_{it}\|^4 \, |q_{it}=\gamma) \leq C$ for $\xi_{it}=x_{it}$ and $x_{it}\varepsilon_{it}$;

(ii) $f_{it}(\gamma)$ is continuous over Γ and $\max_{i,t}\sup_{\gamma\in\Gamma} f_{it}(\gamma) \leq c_f < \infty$.

(iii) For $g\in\mathcal{G}$, $D_g(\gamma)$ and $V_g(\gamma)$ are continuous at $\gamma=\gamma^0$;

(iv) There exists a constant $c>0$ such that $\inf_{\gamma\in\Gamma}\lambda_{\min}[M_g(\gamma)] \geq c$ for all $g\in\mathcal{G}$.

Assumption A.4(i) is a non-colinearity assumption for the regressors and A.4(ii) holds because $E\left\|x_{it}(\gamma)-x_{it}(\gamma^*)\right\| \asymp |\gamma-\gamma^*|$ under some regularity conditions on $\{x_{it}, q_{it}\}$, where $a\asymp b$ means and both a/b and b/a are bounded away from zero. It's natural to expect that the first term in the definition of $\tilde{w}_g(\gamma)$ is of the same probability order as $\left|\gamma-\gamma_g^0\right|$. A.4(ii) requires that after projecting $\tilde{x}_{it}(\gamma,\gamma_g^0)$ onto $\tilde{z}_{it}(\gamma)$, the associated residual exhibits the same probability order of variations groupwise. Assumption A.5 imposes some conditions on the conditional PDF and moments of x_{it} and $x_{it}\varepsilon_{it}$. A.5(i) requires that the fourth order conditional moment of $x_{it}\varepsilon_{it}$ and x_{it} be well behaved; A.5(ii) requires that the PDF of q_{it} be uniformly bounded; A.5(iii)–(iv) requires the probability limits of some quantities associated with the asymptotic variance be well behaved.

To state the next theorem, we define the infeasible estimators of the slope and threshold coefficients that are obtained with known group structures:

$$(\check{\Theta},\check{\mathbf{D}}) \equiv \underset{(\Theta,\mathbf{D})\in\mathcal{B}^G\times\Gamma^G}{\arg\min}\ \check{\mathcal{Q}}(\Theta,\mathbf{D}), \tag{3.6}$$

where $\check{\mathcal{Q}}(\Theta,\mathbf{D}) \equiv \mathcal{Q}(\Theta,\mathbf{D},\mathbf{G}^0)$. With the knowledge of the true group structure $\mathbf{G}^0$, we can split the N individuals into G groups perfectly and estimate the group-specific parameters for each group. Let $\check{\mathcal{Q}}_g(\theta,\gamma)=\sum_{i\in\mathbf{G}_g^0}\sum_{t=1}^T[\tilde{y}_{it}-$

$\tilde{z}_{it}(\gamma)'\theta]^2$. Then we have

$$\check{\mathcal{Q}}(\Theta, \mathbf{D}) = \sum_{g=1}^{G} \check{\mathcal{Q}}_g(\theta, \gamma) \text{ and } (\check{\theta}_g, \check{\gamma}_g) = \underset{(\theta,\gamma)\in\mathcal{B}\times\Gamma}{\arg\min}\, \check{\mathcal{Q}}_g(\theta, \gamma) \text{ for each } g \in \mathcal{G}.$$

The following theorem establishes the asymptotic equivalence between the feasible estimator $(\hat{\Theta}, \hat{\mathbf{D}})$ and the infeasible estimator $(\check{\Theta}, \check{\mathbf{D}})$.

Theorem 3.2. *Suppose that Assumptions A.1–A.5 hold with* $\alpha \in (0, 1/3)$ *in Assumption A.1(vi). Let* $\alpha_{NT} = (NT)^{1-2\alpha}$. *Then we have* $(NT)^{1/2} \left\| \hat{\Theta} - \check{\Theta} \right\| \xrightarrow{p} 0$ *and* $\alpha_{NT}(\hat{\mathbf{D}} - \check{\mathbf{D}}) \xrightarrow{p} 0$.

Theorem 3.2 shows that $\hat{\Theta} - \check{\Theta} = o_p((NT)^{-1/2})$ and $\hat{\mathbf{D}} - \check{\mathbf{D}} = o_p(\alpha_{NT}^{-1})$ by restricting $\alpha \in (0, 1/3)$ in Assumption A.1(vi). Under Assumptions A.1–A.5, we can show that $\check{\Theta} - \Theta^0 = O_p((NT)^{-1/2} + T^{-1})$ and $\check{\mathbf{D}}$ has α_{NT}-rate of convergence. Therefore, the estimator $(\hat{\Theta}, \hat{\mathbf{D}})$ has the same asymptotic distribution as that of $(\check{\Theta}, \check{\mathbf{D}})$. Then we can establish the asymptotic distribution of our least squares estimator.

To report the asymptotic distributions of $\hat{\theta}_g$ and $\hat{\gamma}_g$, we add some notation:

$$\begin{aligned}
\omega_{g,NT}(\gamma, \gamma^*) &\equiv \frac{1}{N_g T} \sum_{i\in\mathbf{G}_g^0} \sum_{t=1}^{T} \tilde{z}_{it}(\gamma)\tilde{z}_{it}(\gamma^*)', \\
\Omega_{g,NT1}(\gamma, \gamma^*) &\equiv \frac{1}{N_g T} \sum_{i\in\mathbf{G}_g^0} \sum_{t=1}^{T} \tilde{z}_{it}(\gamma)\tilde{z}_{it}(\gamma^*)'\varepsilon_{it}^2, \\
\Omega_{g,NT2}(\gamma, \gamma^*) &\equiv \frac{1}{N_g T} \sum_{i\in\mathbf{G}_g^0} \sum_{t=1}^{T} \sum_{s=1}^{T} \tilde{z}_{it}(\gamma)\tilde{z}_{is}(\gamma^*)'\varepsilon_{is}\varepsilon_{it}, \text{ and} \\
\mathbb{B}_{g,NT}(\gamma) &\equiv \frac{1}{N_g T} \sum_{i\in\mathbf{G}_g^0} \sum_{t=2}^{T} \sum_{s<t} E\left[z_{it}(\gamma)\varepsilon_{is}\right].
\end{aligned}$$

Assumption A.6: (i) For each $g \in \mathcal{G}$, the following probability limits exist and are finite: $\omega_g(\gamma, \gamma^*) = \operatorname{p\,lim}_{(N,T)\to\infty} \omega_{g,NT}(\gamma, \gamma^*)$, $\Omega_{g,\ell}(\gamma, \gamma^*) = \operatorname{p\,lim}_{(N,T)\to\infty} \Omega_{g,NT\ell}(\gamma, \gamma^*)$ for $\ell = 1, 2$, and $\mathbb{B}_g(\gamma) = \lim_{(N,T)\to\infty} \mathbb{B}_{g,NT}(\gamma)$.

(ii) $\omega_{g,NT}(\gamma, \gamma^*) \xrightarrow{p} \omega_g(\gamma, \gamma^*)$ and $\Omega_{g,NT\ell}(\gamma, \gamma^*) \xrightarrow{p} \Omega_{g,\ell}(\gamma, \gamma^*)$ for $\ell = 1, 2$ uniformly in $\gamma, \gamma^* \in \Gamma$.

Assumption A.6 imposes some conditions on the probability limits of random quantities that are associated with the asymptotic variance and bias of $\hat{\Theta}$.

Here, we follow Hansen (2000) and assume directly that $\omega_{g,NT}$ and $\Omega_{g,NT\ell}$ for $\ell = 1, 2$ converge uniformly to some limits. The uniformity greatly facilitates the proofs of Theorem 3.3 below.

We establish the asymptotic distribution of our estimators in the following theorem.

Theorem 3.3. *Suppose that Assumptions A.1–A.6 hold with $\alpha \in (0, 1/3)$ in Assumption A.1(vi). Let $\alpha_{N_gT} = (N_gT)^{1-2\alpha}$, $\omega_g^0 = \omega_g(\gamma_g^0, \gamma_g^0)$, $\mathbb{B}_g^0 = \mathbb{B}_g(\gamma_g^0)$, and $\Omega_{g,\ell}^0 = \Omega_{g,\ell}(\gamma_g^0, \gamma_g^0)$ for $\ell = 1, 2$. Then for each $g \in \mathcal{G}$,*

(i)$\sqrt{N_gT}(\hat{\theta}_g - \theta_g^0) - (\omega_g^0)^{-1}\sqrt{\frac{N_g}{T}}\mathbb{B}_g^0 \xrightarrow{d} \mathcal{N}(0, (\omega_g^0)^{-1}\Omega_{g,1}^0(\omega_g^0)^{-1})$ under Assumption A.1(i.1) and $\sqrt{N_gT}(\hat{\theta}_g - \theta_g^0) \xrightarrow{d} \mathcal{N}(0, (\omega_g^0)^{-1}\Omega_{g,2}^0(\omega_g^0)^{-1})$ under Assumption A.1(i.2);

(ii) $\alpha_{N_gT}(\hat{\gamma}_g - \gamma_g^0) \xrightarrow{d} \varpi_g \mathcal{T}_g$, where $\varpi_g = \frac{C_g^{0\prime}V_g^0C_g^0}{(C_g^{0\prime}D_g^0C_g^0)^2}$, $\mathcal{T}_g = \arg\max_{r\in\mathbb{R}}\left[-\frac{1}{2}|r| + W_g(r)\right]$, and $W_g(\cdot)$, $g \in \mathcal{G}$, are mutually independent two-sided Brownian motions.

Theorem 3.3 establishes the asymptotic distributions of the estimators of the slope and threshold coefficients. Note that we strengthen Assumption A.1(vi) slightly to require $\alpha \in (0, 1/3)$. From the proof of Lemma B.7 that is used in the proof of the above theorem, we can easily find that such an extra condition is not needed if we only consider the case where $N/T \to \kappa$ for some $\kappa \in (0, \infty)$.

When we allow for dynamics in Assumption A.1(i.1), the estimator $\hat{\theta}_g$ of the group-specific slope coefficient θ_g^0 exhibits a bias term to be corrected as in standard dynamic panels. One can conduct the bias correction by estimating ω_g^0 and $\mathbb{B}_{g,0}$ consistently by

$$\hat{\omega}_g \equiv \frac{1}{\hat{N}_gT}\sum_{i\in\hat{\mathbf{G}}_g}\sum_{t=1}^{T}\tilde{z}_{it}(\hat{\gamma}_g)\tilde{z}_{it}(\hat{\gamma}_g)' \text{ and } \hat{\mathbb{B}}_g = \frac{1}{\hat{N}_gT}\sum_{i\in\hat{\mathbf{G}}_g}\sum_{t=2}^{T}\sum_{s<t} z_{it}(\hat{\gamma}_g)\hat{\varepsilon}_{is},$$

where $\hat{N}_g = \left|\hat{\mathbf{G}}_g\right|$ denotes the cardinality of $\hat{\mathbf{G}}_g$, $\hat{\mathbf{G}}_g \equiv \{i : \hat{g}_i = g\}$ for $g \in \mathcal{G}$, and $\hat{\varepsilon}_{it} = \tilde{y}_{it} - \tilde{z}_{it}(\hat{\gamma}_g)'\hat{\theta}_g$. Similarly, it is easy to show that a consistent estimator of the asymptotic variance of $\hat{\theta}_g$ in this case is given by $\hat{\omega}_g^{-1}\hat{\Omega}_{g,1}\hat{\omega}_g^{-1}$, where $\hat{\Omega}_{g,1} = \frac{1}{\hat{N}_gT}\sum_{i\in\hat{\mathbf{G}}_g}\sum_{t=1}^{T}\tilde{z}_{it}(\hat{\gamma}_g)\tilde{z}_{it}(\hat{\gamma}_g)'\hat{\varepsilon}_{it}^2$. When (X_i, q_i) is strictly exogenous in Assumption A.1(i.2), we allow for serial correlation in the error terms. In this case, we propose to estimate the asymptotic variance of $\hat{\theta}_g$ by $\hat{\omega}_g^{-1}\hat{\Omega}_{g,2}\hat{\omega}_g^{-1}$, where $\hat{\Omega}_{g,2}$ is a panel heteroskedasticity and autocorrelation consistent (HAC) estimator:

$$\hat{\Omega}_{g,2} = \frac{1}{\hat{N}_g} \sum_{i \in \hat{\mathbf{G}}_g} \left[\hat{\Lambda}_{i,0} + \sum_{s=1}^{J_T} w_{Ts} (\hat{\Lambda}_{is} + \hat{\Lambda}'_{is}) \right],$$

where $w_{Ts} = 1 - |s| / J_T$, J_T satisfies $1/J_T + J_T^3/T \to 0$ as $T \to \infty$, and $\hat{\Lambda}_{is} = \frac{1}{T}\sum_{t=s+1}^{T} \tilde{z}_{it}(\hat{\gamma}_g)\tilde{z}_{i,t-s}(\hat{\gamma}_g)' \times \hat{\varepsilon}_{it}\hat{\varepsilon}_{i,t-s}$. Following Su and Jin (2012) and the results in Theorems 3.2–3.3, we can show that $\hat{\Omega}_{g,2}$ and $\hat{\omega}_g^{-1}\hat{\Omega}_{g,2}\hat{\omega}_g^{-1}$ are consistent estimators of $\Omega_{g,2}^0$ and $(\omega_g^0)^{-1}\Omega_{g,2}^0(\omega_g^0)^{-1}$, respectively.

Theorem 3.3(ii) indicates that the asymptotic distribution of $\hat{\gamma}_g$ is pivotal up to a scale parameter ϖ_g, which is similar to that given by Theorem 1 in Hansen (2000). It is well known that this result highly relies on the assumption that the threshold effect converges to zero as $(N, T) \to \infty$. Under the fixed-threshold-effect framework ($\alpha = 0$), it is possible to demonstrate $NT(\hat{\gamma}_g - \gamma_g^0) = O_p(1)$ but the asymptotic distribution of $\hat{\gamma}_g$ will not be asymptotically pivotal even after appropriate normalization. In addition, it is well known that the above scale parameter ϖ_g cannot be consistently estimated. To make inference on the threshold parameters, we propose to apply the likelihood ratio test in the next section.

3.4 Inference on the Threshold Parameter

In this section, we consider inference on the threshold parameter $\mathbf{D} = (\gamma_1, ..., \gamma_G)'$. We consider three cases. The first case is to test the null hypothesis on the threshold parameter γ_g for a single group $g \in \mathcal{G}$:

$$H_{01} : \gamma_g = \gamma_g^0 \text{ for some } \gamma_g^0 \in \Gamma.$$

Next, we consider testing the homogeneity of the threshold parameters:

$$H_{02} : \gamma_1^0 = ... = \gamma_G^0 = \gamma^0 \text{ for some } \gamma^0 \in \Gamma.$$

If one fails to reject the hypothesis of common threshold parameter for all groups, one can estimate the model with a common threshold parameter, γ, say. Then we can study the inference on the common threshold parameter

$$H_{03} : \gamma = \gamma^0 \text{ for some } \gamma^0 \in \Gamma.$$

3.4.1 Likelihood ratio test for a single γ_g

To test the null hypothesis $H_{01} : \gamma_g = \gamma_g^0$, a standard approach is to use the likelihood ratio (LR) test. If we know the true group structure, the likelihood ratio test statistic can be constructed as in Hansen (2000). In our framework, we need to construct the test statistic based on the estimated group structure $\{\hat{\mathbf{G}}_g, g \in \mathcal{G}\}$. Let $\bar{\theta}_g(\gamma) \equiv \arg\min_{\theta \in \mathcal{B}} \bar{\mathcal{Q}}_g(\theta, \gamma)$, where $\bar{\mathcal{Q}}_g(\theta, \gamma) \equiv \sum_{i \in \hat{\mathbf{G}}_g} \sum_{t=1}^{T} [\tilde{y}_{it} - \tilde{z}_{it}(\gamma)'\theta]^2$. We follow the lead of Hansen (2000) and propose to employ the following LR test statistic for γ_g :

$$\mathcal{L}_{g,NT}(\gamma) \equiv \hat{N}_g T \frac{\bar{\mathcal{Q}}_g(\bar{\theta}_g(\gamma), \gamma) - \bar{\mathcal{Q}}_g(\hat{\theta}_g, \hat{\gamma}_g)}{\bar{\mathcal{Q}}_g(\hat{\theta}_g, \hat{\gamma}_g)}.$$

The major difference is that we consider the minimization of $\bar{\mathcal{Q}}_g(\theta, \gamma)$ instead of the infeasible version $\check{\mathcal{Q}}_g(\theta, \gamma)$. In the proof of Theorem 3.4 below, we show that $\bar{\mathcal{Q}}_g(\theta, \gamma)$ and $\check{\mathcal{Q}}_g(\theta, \gamma)$ are asymptotically equivalent so that we can study the asymptotic distribution of the LR test statistic based on the minimization of the infeasible objective function.

For each $g \in \mathcal{G}$, let $\sigma_g^2 = \lim_{(N,T)\to\infty} \frac{1}{N_g T} \sum_{i \in \mathbf{G}_g^0} \sum_{t=1}^{T} E(\varepsilon_{it}^2)$, $w_{g,V} = C_g^{0\prime} V_g^0 C_g^0$ and $w_{g,D} = C_g^{0\prime} D_g^0 C_g^0$. Let $\sigma^2 = \lim_{(N,T)\to\infty} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} E(\varepsilon_{it}^2)$. The following theorem establishes the asymptotic null distribution of the above LR test statistic.

Theorem 3.4. *Suppose that Assumptions A.1–A.6 hold with $\alpha \in (0, 1/3)$ in Assumption A.1(vi). Then under $H_{01} : \gamma_g = \gamma_g^0$, we have*

$$\mathcal{L}_{g,NT}(\gamma_g^0) \xrightarrow{d} \eta_g^2 \xi_g \text{ for each } g \in \mathcal{G},$$

where $\eta_g^2 = \frac{w_{g,V}}{w_{g,D}\sigma_g^2}$ and $\xi_g = \max_{s \in \mathbb{R}}[2W_g(s) - |s|]$ has the distribution function characterized by $\Pr(\xi_g \leq x) = (1 - e^{-x/2})^2$.

Theorem 3.4 indicates that the asymptotic distribution of the LR test statistic constructed from the estimated group structure is asymptotically equivalent to that of the infeasible test statistic obtained from the true group structure. Now, we still have a nuisance parameter η_g^2. In the special case where we have conditional homoskedasticity along both the cross-section and time dimensions, $\eta_g^2 = 1$ and the LR statistic is asymptotically free of any nuisance parameter. If heteroskedasticity is suspected, then we need to estimate η_g^2 consistently. Noting that

$$\eta_g^2 = \frac{\text{plim}_{(N,T)\to\infty} \frac{1}{N_g T} \sum_{i\in\mathbf{G}_g^0} \sum_{t=1}^T E[(\delta_g^{0\prime} x_{it}\varepsilon_{it})^2 | q_{it} = \gamma_g^0] f_{it}\left(\gamma_g^0\right)}{\sigma_g^2 \text{plim}_{(N,T)\to\infty} \frac{1}{N_g T} \sum_{i\in\mathbf{G}_g^0} \sum_{t=1}^T E[(\delta_g^{0\prime} x_{it})^2 | q_{it} = \gamma_g^0] f_{it}\left(\gamma_g^0\right)},$$

we propose to estimate η_g^2 by

$$\hat{\eta}_g^2 = \frac{\sum_{i\in\hat{\mathbf{G}}_g} \sum_{t=1}^T K_h(\hat{\gamma}_g - q_{it})(\hat{\delta}_g' x_{it}\hat{\varepsilon}_{it})^2}{\hat{\sigma}_g^2 \sum_{i\in\hat{\mathbf{G}}_g} \sum_{t=1}^T K_h(\hat{\gamma}_g - q_{it})(\hat{\delta}_g' x_{it})^2},$$

where $\hat{\sigma}_g^2 = \tilde{\mathcal{Q}}_g(\hat{\theta}_g, \hat{\gamma}_g)/(\hat{N}_g T)$, $K_h(u) = h^{-1}K(u/h)$, $h \to 0$ is the bandwidth parameter and $K(\cdot)$ is a kernel function. We can readily show that $\hat{\sigma}_g^2 = \sigma_g^2 + o_p(1)$ and $\hat{\eta}_g^2 = \eta_g^2 + o_p(1)$ under some standard weak conditions on h and $K(\cdot)$. Given the consistent estimate of η_g^2, we can consider the normalized LR statistic

$$\mathcal{NL}_{g,NT}(\gamma_g^0) = \mathcal{L}_{g,NT}(\gamma_g^0)/\hat{\eta}_g^2.$$

We can easily tabulate the asymptotic critical value for $\mathcal{NL}_{g,NT}(\gamma_g^0)$. We can also invert this statistic to obtain the asymptotic $1-a$ confidence interval for γ :

$$CI_{1-a} = \left\{\gamma \in \Gamma : \mathcal{NL}_{g,NT} \leq \xi_{1-a}\right\},$$

where ξ_{1-a} denotes the $1-a$ percentile of ξ. For example, $\xi_{1-\alpha} = 5.94$, 7.35, and 10.59 for $a = 0.10$, 0.05, and 0.01, respectively.

3.4.2 Test for common threshold parameters

In applications, it is likely that all individuals share a common threshold parameter, although their slope coefficients may still vary across groups. In this case, estimating the model with the common-threshold restriction imposed improves the asymptotic efficiency of the threshold estimator. Thus motivated, one may wish to test the homogeneity of the threshold parameter prior to estimation. In this section, we consider testing the null hypothesis

$$H_{02} : \gamma_1^0 = ... = \gamma_G^0 = \gamma^0 \text{ for some } \gamma^0 \in \Gamma.$$

Let $\mathcal{D}_r = \{\mathbf{D} = (\gamma, ..., \gamma)', \gamma \in \Gamma\} \subseteq \Gamma^G$ be the restricted parameter space and $\mathbf{D}_{r,\gamma} \equiv (\gamma, ..., \gamma)' \in \mathcal{D}_r$. Then the null hypothesis can be equivalently rewritten as $H_{02} : \mathbf{D}^0 \in \mathcal{D}_r$. We can estimate the model by restricting $\mathbf{D} \in \mathcal{D}_r$ under H_{02} :

$$(\hat{\Theta}_r, \hat{\mathbf{D}}_r, \hat{\mathbf{G}}_r) = \underset{(\Theta,\mathbf{D},\mathbf{G})\in\mathcal{B}^G\times\mathcal{D}_r\times\mathcal{G}^N}{\arg\min} \mathcal{Q}(\Theta, \mathbf{D}, \mathbf{G}).$$

Then we can construct the LR test statistic by

$$\mathcal{L}_{NT} = NT\frac{\mathcal{Q}(\hat{\Theta}_r, \hat{\mathbf{D}}_r, \hat{\mathbf{G}}_r) - \mathcal{Q}(\hat{\Theta}, \hat{\mathbf{D}}, \hat{\mathbf{G}})}{\mathcal{Q}(\hat{\Theta}, \hat{\mathbf{D}}, \hat{\mathbf{G}})}.$$

The following theorem studies the asymptotic distribution of $\mathcal{L}_{NT}$ under H_{02}.

Theorem 3.5. *Suppose that Assumptions A.1–A.6 hold with* $\alpha \in (0, 1/3)$ *in Assumption A.1(vi). Under the null hypothesis* $H_{02} : \mathbf{D}^0 \in \mathcal{D}_r$*, we have*

$$\mathcal{L}_{NT} \xrightarrow{d} \sum_{g=1}^{G} \tilde{\eta}_g^2 \max_{s_g\in\mathbb{R}}[2W_g(s_g) - |s_g|] - \max_{s\in\mathbb{R}} \left[\sum_{g=1}^{G} \tilde{\eta}_g^2 \left(2W_g(\rho_g s) - |\rho_g s|\right)\right] \equiv \Xi,$$

where $\rho_g = \frac{w_{g,D}}{w_{1,D}}\pi_g/\tilde{\eta}_g^2$*, and* $\tilde{\eta}_g^2 = w_{g,V}/(w_{g,D}\sigma^2)$*.*

Theorem 3.5 indicates that the limiting distribution Ξ of $\mathcal{L}_{NT}$ involves two sets of nuisance parameters, viz, $\tilde{\eta}_g^2$ and ρ_g for $g \in \mathcal{G}$. Under conditional homoskedasticity, we have $\tilde{\eta}_g^2 = 1$ for each g. If heteroskedasticity is suspected, then we need to estimate $\tilde{\eta}_g^2$ consistently. If ρ_g is homogeneous across g, we do not need to estimate it. However, ρ_g is generally not homogeneous across g and we need to estimate it via estimating $\tilde{\eta}_g^2$, $\frac{w_{g,D}}{w_{1,D}}$, and π_g. Using Theorem 3.1, it is easy to show that a consistent estimator of π_g is given by $\hat{\pi}_g = \hat{N}_g/N$. Noting that $\tilde{\eta}_g^2 = \frac{\sigma_g^2}{\sigma^2}\eta_g^2$ and

$$\frac{w_{g,D}}{w_{1,D}} = \frac{\text{plim}_{(N,T)\to\infty}\frac{1}{N_gT}\sum_{i\in\mathbf{G}_g^0}\sum_{t=1}^{T} E[(\delta_g^{0\prime}x_{it})^2|q_{it} = \gamma^0]f_{it}(\gamma^0)}{\text{plim}_{(N,T)\to\infty}\frac{1}{N_1T}\sum_{i\in\mathbf{G}_1^0}\sum_{t=1}^{T} E[(\delta_1^{0\prime}x_{it})^2|q_{it} = \gamma^0]f_{it}(\gamma^0)},$$

we propose to estimate $\tilde{\eta}_g^2$ and $\frac{w_{g,D}}{w_{1,D}}$ respectively by

$$\widehat{\tilde{\eta}}_g^2 = \hat{\eta}_g^2 \frac{\hat{\sigma}_g^2}{\hat{\sigma}^2} \quad \text{and} \quad \frac{\hat{w}_{g,D}}{\hat{w}_{1,D}} = \frac{\frac{1}{\hat{N}_g T} \sum_{i \in \hat{\mathbf{G}}_g} \sum_{t=1}^{T} K_h(\hat{\gamma}_g - q_{it})(\hat{\delta}_g' x_{it})^2}{\frac{1}{\hat{N}_g T} \sum_{i \in \hat{\mathbf{G}}_1} \sum_{t=1}^{T} K_h(\hat{\gamma}_1 - q_{it})(\hat{\delta}_1' x_{it})^2},$$

where $\hat{\sigma}^2 = \frac{1}{NT} \sum_{g=1}^{G} \sum_{i \in \hat{\mathbf{G}}_g} \sum_{t=1}^{T} [\tilde{y}_{it} - \tilde{z}_{it}(\hat{\gamma}_g)' \hat{\theta}_g]^2$. It is easy to show that the above estimators are consistent under standard conditions and a consistent estimator of ρ_g is given by $\hat{\rho}_g = \frac{\hat{w}_{g,D}}{\hat{w}_{1,D}} \hat{\pi}_g / \widehat{\tilde{\eta}}_g^2$. To find out the p-value, we can simulate the asymptotic distribution with these estimates. Basically, we can estimate G independent two-sided Brownian motions $W_g(\cdot)$ and construct the corresponding statistic where the nuisance parameters are replaced with their consistent estimates. Simulating a large number of times, we can mimic the asymptotic distribution sufficiently well. Then, we can reject the null hypothesis at the prescribed a level, if the test statistic is larger than $1 - \alpha$ quantile of the simulated distribution.

3.4.3 Likelihood ratio test for common threshold parameter

Suppose we have common threshold parameters, we can use the restricted estimator $(\hat{\Theta}_r, \hat{\mathbf{D}}_r, \hat{\mathbf{G}}_r)$ defined in the last subsection. Even in this case, the estimators of the group-specific slope coefficients share the same asymptotic distribution as the unrestricted estimators studied in the last section due to the asymptotic independence between the estimators of the slope coefficients and that of the threshold parameter.

To make inference on the common threshold parameter γ, we also consider an LR test for $H_{03} : \gamma = \gamma^0$. The LR test statistic is now defined by

$$\mathcal{L}_{NT}^{c}(\gamma) = NT \frac{\mathcal{Q}(\hat{\Theta}(\mathbf{D}_{r,\gamma}, \hat{\mathbf{G}}_r), \mathbf{D}_{r,\gamma}, \hat{\mathbf{G}}_r) - \mathcal{Q}(\hat{\Theta}_r, \hat{\mathbf{D}}_r, \hat{\mathbf{G}}_r)}{\mathcal{Q}(\hat{\Theta}_r, \hat{\mathbf{D}}_r, \hat{\mathbf{G}}_r)},$$

where $\hat{\Theta}(\mathbf{D}_{r,\gamma}, \hat{\mathbf{G}}_r)$ is defined as in Section 3.2.1 and the superscript c is an abbreviation for "common". Note that $H_{03} : \gamma = \gamma^0$ can be equivalently rewritten as $H_{03} : \mathbf{D}^0 = \mathbf{D}_{r,\gamma^0}$.

The next theorem establishes the asymptotic distribution of $\mathcal{L}^c_{NT}(\gamma)$ under H_{03}.

Theorem 3.6. *Suppose that Assumptions A.1–A.6 hold with $\alpha \in (0, 1/3)$ in Assumption A.1(vi). Under the null $H_{03} : \mathbf{D}^0 = \mathbf{D}_{r,\gamma^0}$, we have*

$$\mathcal{L}^c_{NT}(\gamma^0) \xrightarrow{d} \eta^2 \max_{s\in\mathbb{R}} \left[W(s) - |s|\right],$$

where $\eta^2 = \left(\sum_{g=1}^G \pi_g w_{g,V}\right) / \left(\sigma^2 \sum_{g=1}^G \pi_g w_{g,D}\right)$.

Like before, we can estimate the nuisance parameter η^2 consistently by the nonparametric estimator:

$$\hat{\eta}^2 = \frac{\sum_{i=1}^N \sum_{t=1}^T K_h(\hat{\gamma} - q_{it})(\hat{\delta}'_{\hat{g}_i} x_{it}\hat{\varepsilon}_{it})^2}{\hat{\sigma}^2 \sum_{i=1}^N \sum_{t=1}^T K_h(\hat{\gamma} - q_{it})(\hat{\delta}'_{\hat{g}_i} x_{it})^2},$$

where $\hat{\gamma}$ is the estimator of the common threshold parameter γ under H_{02}, $K_h(u) = h^{-1}K(u/h)$, $h \to 0$ is the bandwidth parameter and $K(\cdot)$ is a kernel function.

3.5 Test for the Presence of Threshold Effect

In application, one may suspect that a set of groups do not have the threshold effect. In this case, we can verify the existence of threshold effects for $P \leq G$ groups by testing the null hypothesis

$$\mathbb{H}_0 : \delta^0_{g_1} = ... = \delta^0_{g_P} = 0$$

versus the alternative hypothesis $\mathbb{H}_1 : \delta^0_{g_l} \neq 0$ for some $g_l \in \mathcal{G}_s$, where $\mathcal{G}_s \equiv \{g_l, l = 1, ..., P\} \subset \mathcal{G}$. To study the local power of our test, we consider the following sequence of Pitman local alternatives

$$\mathbb{H}_{1NT} : \delta^0_{g_l} = c_l/\sqrt{NT} \text{ for } g_l \in \mathcal{G}_s.$$

Let $\mathbf{c} \equiv (c'_1, ..., c'_P)'$ and $\mathbb{L} \equiv (e_{g_1}, ..., e_{g_P})' \otimes L$, where $\otimes$ denotes Kronecker product, $L \equiv [\mathbf{0}_{K\times K}, I_K]$ and e_{g_l} is a $G \times 1$ vector with g_lth entry being 1 and other entries equal to zero. Then we can rewrite the null and local alternative hypotheses respectively as

$$\mathbb{H}_0 : \mathbb{L}\Theta^0 = \mathbf{0}_{KP\times 1} \text{ and } \mathbb{H}_{1NT} : \mathbb{L}\Theta^0 = \mathbf{c}/\sqrt{NT}.$$

Note that $\mathbf{c} = \mathbf{0}_{KP\times 1}$ corresponds to the null hypothesis of no threshold effects and we allow $\delta^0_{g_l}$ for $g_l \in \mathcal{G}_s$ to shrink to zero at the $(NT)^{-1/2}$-parametric rate under the local alternative. Under $\mathbb{H}_{1NT}$, the early estimators of Θ^0 and $\mathbf{G}^0$ continue to be consistent with any $\mathbf{D} \in \Gamma^G$ despite the fact that we cannot identify $\mathbf{D}^0$.

As we do not know the true group structure, we need to rely on the estimated group structure $\hat{\mathbf{G}}$. For any fixed $\mathbf{D}$ and a preliminary estimate of group structure $\hat{\mathbf{G}}$, we can obtain the bias-corrected estimator $\bar{\Theta}^{\mathrm{bc}}(\mathbf{D}, \hat{\mathbf{G}}) = (\bar{\theta}^{\mathrm{bc}}_1(\gamma_1)', ..., \bar{\theta}^{\mathrm{bc}}_G(\gamma_G)')'$. Let

$$\hat{\Pi} = \operatorname{diag}(\hat{\pi}_1, ..., \hat{\pi}_G) \otimes I_{2K} \text{ and } \hat{\mathbb{K}}_{NT}(\mathbf{D}) = \mathbb{L}\hat{\omega}(\mathbf{D})^{-1}\hat{\mathbf{\Omega}}(\mathbf{D})\hat{\omega}(\mathbf{D})^{-1}\mathbb{L}',$$

where

$$\hat{\omega}(\mathbf{D}) = \begin{bmatrix} \hat{\omega}_1(\gamma_1,\gamma_1) & & \\ & \ddots & \\ & & \hat{\omega}_G(\gamma_G,\gamma_G) \end{bmatrix} \text{ and } \hat{\mathbf{\Omega}}(\mathbf{D}) = \begin{bmatrix} \hat{\Omega}_{1,1}(\gamma_1,\gamma_1) & & \\ & \ddots & \\ & & \hat{\Omega}_{G,1}(\gamma_G,\gamma_G) \end{bmatrix}.$$

We can construct the sup-Wald test statistic $\mathcal{W}_{NT} = \sup_{\mathbf{D}\in\Gamma^G} W_{NT}(\mathbf{D})$, where

$$W_{NT}(\mathbf{D}) = NT \cdot \bar{\Theta}^{\mathrm{bc}}(\mathbf{D}, \hat{\mathbf{G}})'\hat{\Pi}^{1/2}\mathbb{L}' \left(\hat{\mathbb{K}}_{NT}(\mathbf{D})\right)^{-1} \mathbb{L}\hat{\Pi}^{1/2}\bar{\Theta}^{\mathrm{bc}}(\mathbf{D}, \hat{\mathbf{G}}).$$

Let $S_{g,NT}(\gamma) = \frac{1}{\sqrt{N_g T}} \sum_{i\in\mathbf{G}^0_g} \sum_{t=1}^T \{z_{it}(\gamma) - \frac{1}{T}\sum_{s=1}^T E[z_{is}(\gamma)]\}\varepsilon_{it}$. Let $S_g(\gamma)$ be a zero mean Gaussian process with covariance kernel $\Omega_g(\gamma, \gamma^*)$. Let $\mathbb{K}(\mathbf{D}) = \mathbb{L}\omega(\mathbf{D})^{-1}\mathbf{\Omega}(\mathbf{D})\omega(\mathbf{D})^{-1}\mathbb{L}'$, $\overline{\mathbf{S}}(\mathbf{D}) = \mathbb{L}\omega(\mathbf{D})^{-1}\mathbf{S}(\mathbf{D})$, $\mathbf{S}(\mathbf{D}) = (S_1(\gamma_1)', ..., S_G(\gamma_G)')'$, and $\overline{\mathbf{Q}}(\mathbf{D}) = \mathbb{L}\omega(\mathbf{D})^{-1}\mathbf{Q}(\mathbf{D})\Pi^{1/2}\mathbb{L}'$, where Π =diag(π_1, ..., π_G) $\otimes I_{2K}$,

$$\begin{aligned} \mathbf{Q}(\mathbf{D}) &= \begin{bmatrix} \omega_1(\gamma_1,\gamma_1^0) & & \\ & \ddots & \\ & & \omega_G(\gamma_G,\gamma_G^0) \end{bmatrix}, \mathbf{\Omega}(\mathbf{D}) = \begin{bmatrix} \Omega_{1,1}(\gamma_1,\gamma_1) & & \\ & \ddots & \\ & & \Omega_{G,1}(\gamma_G,\gamma_G) \end{bmatrix}, \text{ and} \\ \omega(\mathbf{D}) &= \begin{bmatrix} \omega_1(\gamma_1,\gamma_1) & & \\ & \ddots & \\ & & \omega_G(\gamma_G,\gamma_G) \end{bmatrix}. \end{aligned}$$

To state the next theorem, we add one assumption.

Assumption A.7: For each $g \in \mathcal{G}$, $S_{g,NT}(\gamma) \Rightarrow S_g(\gamma)$ on the compact set Γ, where $\Rightarrow$ denotes the usual weak convergence.

The following theorem establishes the asymptotic distribution of our sup-Wald test statistic under $\mathbb{H}_{1NT}$.

Theorem 3.7. *Suppose that Assumptions A.1(i.1) and (ii)–(v), and A.2–A.7 hold. Then under* $\mathbb{H}_{1NT} : \mathbb{L}\Theta^0 = \mathbf{c}/\sqrt{NT}$, *we have*

$$\mathcal{W}_{NT} \overset{d}{\to} \sup_{\mathbf{D}\in\Gamma^G} W^{\mathbf{c}}(\mathbf{D}),$$

where $W^{\mathbf{c}}(\mathbf{D}) = \left[\overline{\mathbf{S}}(\mathbf{D}) + \overline{\mathbf{Q}}(\mathbf{D})\mathbf{c}\right]' [\mathbb{K}(\mathbf{D})]^{-1} \left[\overline{\mathbf{S}}(\mathbf{D}) + \overline{\mathbf{Q}}(\mathbf{D})\mathbf{c}\right]$.

Under $\mathbb{H}_0$, $\mathbf{c} = 0$ and $\overline{w}^0 \equiv \sup_{\mathbf{D}\in\Gamma^G} W^0(\mathbf{D}) = \sup_{\mathbf{D}\in\Gamma^G} \overline{\mathbf{S}}(\mathbf{D})' [\mathbb{K}(\mathbf{D})]^{-1} \overline{\mathbf{S}}(\mathbf{D})$. Clearly, the limiting null distribution of $\mathcal{W}_{NT}$ depends on the Gaussian process $\overline{\mathbf{S}}(\mathbf{D})$ and is not pivotal. We cannot tabulate the asymptotic critical values for the above sup-Wald statistic. Nevertheless, given the simple structure of $\overline{\mathbf{S}}(\mathbf{D})$, we can follow the literature (e.g., Hansen 1996) and simulate the critical values via the following procedure:

1. Generate $\{v_{it}, i = 1, ..., N, t = 1, ..., T\}$ independently from the standard normal distribution;

2. Calculate $\hat{\mathbf{S}}_{g,NT}(\mathbf{D}) = \frac{1}{\sqrt{\hat{N}_g T}} \sum_{i\in\hat{\mathbf{G}}_g} \sum_{t=1}^{T} \tilde{z}_{it}(\gamma_g)\hat{\varepsilon}_{it}(\gamma_g)v_{it}$;

3. Compute $\mathcal{W}^*_{NT} \equiv \sup_{\mathbf{D}\in\Gamma^G} \hat{\mathbf{S}}(\mathbf{D})'\omega(\mathbf{D})^{-1}\mathbb{L}'[\hat{\mathbb{K}}_{NT}(\mathbf{D})]^{-1}\mathbb{L}\hat{\omega}(\mathbf{D})^{-1}\hat{\mathbf{S}}(\mathbf{D})$;

4. Repeat Steps 1–3 B times and denote the resulting $\mathcal{W}^*_{NT}$ test statistics as $\mathcal{W}^*_{NT,j}$ for $j = 1, ..., B$.

5. Calculate the simulated/bootstrap p-value for the $\mathcal{W}_{NT}$ test as $p^*_W = \frac{1}{B}\sum_{j=1}^{B} \mathbf{1}\{\mathcal{W}^*_{NT,j} \geq \mathcal{W}_{NT}\}$ and reject the null when p^*_W is smaller than some prescribed level of significance.

The above discussion was based on the m.d.s. condition in Assumption A.1(i.1). If we consider the case of static panels such that Assumption A.1(i.2) holds, then the covariance kernel is given by
$\Omega_g(\gamma, \gamma^*) = \lim_{(N,T)\to\infty} \frac{1}{N_g T} \sum_{i\in\mathbf{G}^0_g} \sum_{s=1}^{T} \sum_{t=1}^{T} E[\tilde{z}_{is}(\gamma)\tilde{z}_{it}(\gamma^*)'\varepsilon_{it}\varepsilon_{is}]$ for $g \in \mathcal{G}$. Now, the above simulation procedure needs to be modified because $\hat{\mathbf{S}}_{g,NT}(\mathbf{D})$

constructed in Step 2 will not mimic the Gaussian process $\overline{\mathbf{S}}(\mathbf{D})$ in this case. Instead of generating the independently and identically distributed (i.i.d.) standard normal random variables $\{v_{it}\}$ in Step 1, we can generate $v_i = (v_{i1}, ..., v_{iT})'$ independently from a zero mean multivariate normal distribution with the variance-covariance matrix $\Sigma = \{\sigma_{ts}\}$ given by $\sigma_{ts} = [1 - (|t-s|/p_T)]\,\mathbf{1}(|t-s| \leq p_T)$ for some p_T such that $1/p_T + p_T^3/T \to 0$. Then

$$E_w\left[\hat{\mathbf{S}}_{g,NT}(\mathbf{D})\hat{\mathbf{S}}_{g,NT}(\mathbf{D})'\right] = \frac{1}{\hat{N}_g T}\sum_{i\in\hat{G}_g}\sum_{t=1}^{T}\sum_{s=1}^{T} k(\frac{t-s}{p_T})\tilde{z}_{it}(\gamma_g)\tilde{z}_{is}(\gamma_g)\hat{\varepsilon}_{it}(\gamma_g)\hat{\varepsilon}_{is}(\gamma_g),$$

where $E_w(\cdot)$ denotes the expectation conditional on the sample $w \equiv \{x_{it}, q_{it}, \varepsilon_{it}, i = 1, ..., N, t = 1, ..., T\}$ and $k(s) = [1 - |s|/p_T]\,\mathbf{1}(|s| \leq p_T)$. Apparently, $E_w[\hat{\mathbf{S}}_{g,NT}(\mathbf{D})\hat{\mathbf{S}}_{g,NT}(\mathbf{D})']$ converges in probability to $\Omega_g(\gamma_g, \gamma_g)$ and the modified simulation procedure will generate statistics that follow the same asymptotic distribution as that of $\mathcal{W}_{NT}$.

In practice, we frequently consider testing the presence of threshold effects in all G groups, that is, testing $\mathbb{H}_0 : \delta_1^0 = ... = \delta_G^0 = 0$. In this case, $\mathbb{L} = I_G \otimes L$ and we can readily rewrite our Wald statistic $\mathcal{W}_{NT}$ as

$$\mathcal{W}_{NT} = \sup_{(\gamma_1,\dots,\gamma_G)\in\Gamma^G}\sum_{g=1}^{G} W_{gNT}(\gamma_g) \equiv \mathcal{W}_{NT}^{\text{sum}},$$

where $W_{gNT}(\gamma_g) = \hat{N}_g T \cdot \bar{\delta}_g^{\text{bc}}(\gamma_g)' [\hat{\mathbf{K}}_{gNT}(\gamma_g)]^{-1} \bar{\delta}_g^{\text{bc}}(\gamma_g)$, $\hat{\mathbf{K}}_{gNT}(\gamma_g) = L\hat{\omega}_g(\gamma_g,\gamma_g)^{-1}\hat{\Omega}_{g,1}(\gamma_g,\gamma_g) \times\hat{\omega}_g(\gamma_g,\gamma_g)^{-1}L'$, and $\bar{\delta}_g^{\text{bc}}(\gamma_g) = L\bar{\theta}_g^{\text{bc}}(\gamma_g)$. Here, $W_{gNT}(\gamma_g)$ is the Wald statistic used for testing whether $\delta_g^0 = 0$ for the gth group. For this reason, we can also refer to $\mathcal{W}_{NT}$ as a sup-sum-type of Wald statistic ($\mathcal{W}_{NT}^{\text{sum}}$). Alternatively, we can also consider a sup-sup-type of Wald statistic:

$$\mathcal{W}_{NT}^{\text{sup}} = \sup_{1\leq g\leq G}\sup_{\gamma_g\in\Gamma} W_{gNT}(\gamma_g).$$

Following the proof of Theorem 3.7, we can readily find the limiting null distribution of $\mathcal{W}_{NT}^{\text{sup}}$. As before, when we allow for serial correlation in the error terms, we should use $\hat{\Omega}_{g,2}$ in place of $\hat{\Omega}_{g,1}$ and modify the simulation procedure correspondingly to obtain the simulated p-values. We will compare the performance of $\mathcal{W}_{NT}^{\text{sum}}$ with that of $\mathcal{W}_{NT}^{\text{sup}}$ via simulations in Section 3.7.

3.6 Determining the number of groups

In practice, the true number of groups G^0 is typically unknown. In this case, we can consider a BIC-type information criterion (IC) to determine the number of groups. Following BM (2015) and SSP (2016), we consider the following IC:

$$IC(G) = \ln(\hat{\sigma}^2(G)) + \lambda_{NT} GK, \tag{3.7}$$

where $\hat{\sigma}^2(G) = (NT)^{-1}\, \mathcal{Q}(\hat{\Theta}^{(G)}, \hat{\mathbf{D}}^{(G)}, \hat{\mathbf{G}}^{(G)})$, where we make the dependence of $\hat{\Theta}, \hat{\mathbf{D}}, \hat{\mathbf{G}}$ on the group number G explicit, and λ_{NT} is a tuning parameter that plays the role of $\ln(NT)/(NT)$ in the standard BIC for linear panel data models. The estimated number of groups is given by

$$\hat{G} = \underset{G \in \{1,\ldots,G_{\max}\}}{\arg\min}\; IC(G),$$

where $G_{\max}$ is an upper bound for G^0 that does not grow with (N, T). Following the arguments in SSP (2016), we can readily show that $\Pr(\hat{G} < G^0) \to 0$ provided $\lambda_{NT} = o(1)$ under the standard condition that $\hat{\sigma}^2(G) \xrightarrow{p} \sigma^2(G) > \sigma^2$ whenever $G < G^0$. This implies that $\hat{G} \geq G^0$ w.p.a.1. As in BM (2015), it is difficult to further show that $\Pr(\hat{G} = G^0) \to 1$ as $(N, T) \to \infty$ without further restrictions given the use of the K-means-type iterative algorithm in our estimation procedure.

On the other hand, if we require each estimated group should contain a minimum proportion ν of individuals (e.g., $\nu = 0.05$),[1] then we can show that when $G > G^0$, the threshold parameters and slope coefficients can also be estimated consistently and it is possible to show that $\hat{\sigma}^2(G) - \hat{\sigma}^2(G^0) = O_p(T^{-1})$ under some conditions stated in the online supplement. In this case, a choice of λ_{NT} such that $T \cdot \lambda_{NT} \to \infty$ as $(N, T) \to \infty$ would help to eliminate the over-selected model. Then we can prove the following theorem.

Theorem 3.8. *Suppose that Assumptions A.1–A.5 hold. Suppose that As-*

[1] If a group contains less than $\lfloor \nu N \rfloor$ members, the members in this group can be merged into other groups.

sumptions D.1-D.2 in the online supplement holds. Then $\Pr(\hat{G} = G^0) \to 1$ *as* $(N,T) \to \infty$.

Theorem 3.8 shows that the use of the IC helps to determine the correct number of groups w.p.a.1. SSP and Liu et al. (2020) propose a similar IC to ours. SSP also require that $\lambda_{NT} \to 0$ and $\lambda_{NT}T \to \infty$ as $(N,T) \to \infty$ for general nonlinear models but remark this condition can be relaxed substantially for linear panel data models. In contrast, Liu et al. (2020) require that $\lambda_{NT} \to 0$ and $\lambda_{NT}T^{\frac{1}{2(1+\epsilon)}} \to \infty$ for some $\epsilon > 0$, which is much stronger than our requirement on λ_{NT}. The main reason is that they consider general nonlinear regression models and do not explore the properties of their objective function. They suggests using the tuning parameter $\lambda_{NT} \asymp T^{-1/4}$, which satisfies our theoretical requirement but tends to be too large to be useful in practice. In the simulations in the next section, we find that by setting $\lambda_{NT} = 0.1 \ln(NT)/T$, the above IC works fairly well in determining the true number of groups.

3.7 Monte Carlo Simulations

In this section we evaluate the finite sample performance of our tests and estimates via a set of Monte Carlo experiments.

3.7.1 Data generation processes

We consider three main cases. The first two cases concern static panels with different error structures, and the third case examines the dynamic panel. In each case, we consider two subcases that differ regarding whether the threshold value is group specific or common across individual units. Thus, we have six data generating processes (DGPs) in total.

DGP 1: We generate the data from the following static panel structure model:

$$y_{it} = \mu_i + \beta_{1,g_i} x_{it} \mathbf{1}(q_{it} \leq \gamma_{g_i}) + \beta_{2,g_i} x_{it} \mathbf{1}(q_{it} > \gamma_{g_i}) + \varepsilon_{it}, \qquad (3.8)$$

where $\mu_i = T^{-1}\sum_{t=1}^{T} x_{it}$, and we generate x_{it} from an i.i.d. standard normal distribution. The slope coefficient vector $\beta_{g_i} = (\beta'_{1,g_i}, \beta'_{2,g_i})'$ has a group pattern of heterogeneity with the number of groups $G = 3$, and it is specified as

$$(\beta_{1,1}, \beta_{1,2}, \beta_{1,3}) = (1, 1.75, 2.5), \quad \text{and} \quad (\beta_{2,1}, \beta_{2,2}, \beta_{2,3}) = (1, 1.75, 2.5) + c_1(NT)^{-0.1},$$

where c_1 controls the size of the threshold effect and we set $c_1 = 1$ if not especially mentioned. Let π_g be the proportion of units in group g for $g = 1, 2, 3$, and we fix the ratio of units among groups such that $\pi_1 : \pi_2 : \pi_3 = 0.3 : 0.3 : 0.4$. The threshold variable q_{it} follows i.i.d. $N(1, 1)$. The error term ε_{it} is heteroskedastic, generated as $\varepsilon_{it} = \sigma_{it} e_{it}$, where $\sigma_{it} = (s + 0.1x_{it}^2)^{1/2}$, with s controlling for the signal-to-noise ratio, and $e_{it} \sim$ i.i.d. $N(0, 1)$. We set $s = 0.5$, leading to R^2 of about 0.85. Let $\mathbf{D} = (\gamma_1, \gamma_2, \gamma_3)'$. We consider two subcases: group-specific and homogeneous threshold value, i.e.

$$\text{DGP 1.1}: \ \mathbf{D} = (0.5, 1, 1.5)', \qquad \text{DGP 1.2}: \ \mathbf{D} = (1, 1, 1)'.$$

DGP 2: This is the same as DGP 1 except that the error term is generated from an autoregressive process,

$$\varepsilon_{it} = 0.4\varepsilon_{it-1} + e_{it}, \quad e_{it} \sim \text{ i.i.d. } N(0, 1).$$

As above, we consider two subcases, with group-specific and homogeneous threshold values, and we label these two subcases DGP 2.1 and DGP 2.2, respectively.

DGP 3: In this case, we consider dynamic panel data models,

$$y_{it} = \mu_i + (\alpha_{1,g_i}, \beta_{1,g_i})X_{it}\mathbf{1}(q_{it} \le \gamma_{g_i}) + (\alpha_{2,g_i}, \beta_{2,g_i})X_{it}\mathbf{1}(q_{it} > \gamma_{g_i}) + \varepsilon_{it}, \quad (3.9)$$

where $X_{it} = (y_{i,t-1}, x_{it})'$ and $\mu_i = T^{-1}\sum_{t=1}^{T} x_{it}$. The slope coefficient of $y_{i,t-1}$ is set as

$$(\alpha_{1,1}, \alpha_{1,2}, \alpha_{1,3}) = (0.2, 0.4, 0.6), \quad \text{and} \quad (\alpha_{2,1}, \alpha_{2,2}, \alpha_{2,3}) = (0.2, 0.4, 0.6) + c_2(NT)^{-0.1},$$

with $c_2 = 1/4$ if not especially mentioned. The slope coefficient β_{g_i}, the threshold variable q_{it}, and the error term ε_{it} are all generated in the same manner as that in DGP 1. We likewise consider two subcases with different types of threshold values, and we label them DGP 3.1 and DGP 3.2.

For each DGP, we consider two cross-sectional sample sizes, $N = (50, 100)$, and two time series periods, $T = (30, 60)$, leading to four combinations of cross-sectional and time series dimensions. The number of replications is set to 1000 for the estimation and 500 for the hypothesis testing.

3.7.2 Determining the number of groups

As both of our testing and estimation procedures require specifications of the number of groups, we first examine the accuracy of the IC in determining the number of groups, measured by the empirical probability of selecting a particular number. The proposed IC is calculated by assuming the presence of the threshold effect. Nevertheless, researchers typically do not have prior knowledge of the existence of the threshold effect, and tests for the threshold effect in turn require input of the number of groups. Therefore, we examine the performance of IC for the PSTR model in both scenarios with and without the threshold effect ($c_1 = 1$ and $c_2 = 1/4$ in the former case and $c_1 = c_2 = 0$ in the latter). In practice, we need to choose an appropriate λ_{NT} for the information criterion. We experiment with many alternatives and find that $\lambda_{NT} = 0.1 \ln(NT)/T$ works fairly well.

TABLE 3.1 around here.

Table 3.1 displays the empirical probability of selecting a particular number of groups in the three DGPs, and the highest probability in each case is highlighted in bold. The left panel displays the selection frequency when there is no threshold effect but only group-specific slope coefficients, and the right panel considers the cases in the presence of the threshold effect. In both cases, our IC can select the correct number of groups with a large probability, more than 96% in all cases, and this probability increases as either N or T increases. This result suggests that the proposed IC can correctly determine the number of groups regardless whether the there is a threshold effect, and this further allows us to implement tests and estimation given the true number of groups.

3.7.3 Test for the existence of threshold effect

Next, we investigate the performance of the two Wald statistics ($\mathcal{W}_{NT}^{\text{sum}}$ and $\mathcal{W}_{NT}^{\text{sup}}$) to test the existence of a panel structure threshold effect at three conventional significance levels, namely, 1%, 5%, and 10%. These tests are evaluated given the correct number of groups, say $G^0 = 3$. Prior to the test, one is typically ignorant whether the threshold is heterogeneous across groups. Hence, we implement our tests assuming that the threshold is group specific. To facilitate computation and avoid ill behavior for the test statistic, we truncate the top and bottom 10% of the threshold values and use the grid $\{11\%, 12\%, \ldots, 89\%\}$. The critical values for the two test statistics are simulated based on $B = 600$ replications.

TABLEs 3.2 and 3.3 around here.

Table 3.2 presents the rejection frequency of the two tests when the threshold is group specific. The left panel presents the size of the test, i.e. the rejection frequency under the null hypothesis with $c_1 = 0$ in DGP 1 and 2 and $c_1 = c_2 = 0$ in DGP 3. Since the classification is based on the discrepancy of slope coefficients, heterogeneity in the threshold does not contribute to group separation. Hence, the size of both tests is generally well controlled. We find that both tests tend to be oversized when $N = 50$ and $T = 30$, but the sizes improve when either N or T increases. The middle panel shows the power of the tests in the presence of a weak threshold effect ($c_1 = 1/5$, $c_2 = 1/15$). Both tests demonstrate non-trivial power in detecting the threshold effect, and for the fixed DGP and nominal level, the power function monotonically increases as either dimension of the sample size grows. Finally, the right panel considers a stronger threshold effect with $c_1 = 1/2$ and $c_2 = 1/10$. We find that the rejection frequency of both tests increases as the threshold effect increases, and it reaches 1 with large samples.

Table 3.3 considers the case in which the threshold is homogeneous across groups. Again, both tests demonstrate reasonably good size and power properties. We find that both tests tend to over reject the null hypothesis when there are indeed no threshold effects, especially when $T = 30$. As T increases, the rejection frequency approaches the nominal level under the null. Under the alternative hypothesis, the rejection frequency in the presence of homogeneous thresholds seems to be higher than that in case of heterogeneous thresholds. This arises potentially because we estimate the threshold for each group, ignoring the feature of homogeneity. The inefficiency of threshold estimates may inflate the rejection frequency.

3.7.4 Test for homogeneity of threshold parameters across groups

If there exists a threshold effect, the next issue is whether the threshold is common for individuals. We test the homogeneity of the threshold using the LR-based statistic discussed in Section 3.4.2. As above, we use the grid $\{11\%, 12\%, \ldots, 89\%\}$ to facilitate the computation. To estimate η_g^2, we employ the nonparametric method detailed in Section 3.4.2 and follow Hansen (2000) in using the Epanechnikov kernel and the bandwidth selected according to a minimum mean square error criterion. The rejection frequency is displayed in Table 3.4.

TABLE 3.4 around here.

The left panel of Table 3.4 presents the rejection frequency under the null hypothesis of homogeneous thresholds with $\mathbf{D} = (1, 1, 1)'$. The size of the test statistic is generally close to the nominal levels in all DGPs, except that it is undersized for the 10% level test in DGP 2 and 3. The middle panel reports the rejection frequency under the alternative hypothesis of weakly heterogeneous threshold values, i.e., $\mathbf{D} = (0.85, 1, 1.15)'$; the right panel considers the case in which the threshold is strongly heterogeneous, i.e., $\mathbf{D} = (0.5, 1, 1.5)'$. As the degree of heterogeneity increases, we observe a stable increase in the power function. The power is also increasing as either N or T increases for the fixed degree of heterogeneity and nominal level. This indicates that our test has reasonably good power in detecting the heterogeneity of threshold values.

3.7.5 Estimation results

Finally, we consider the estimation of the PSTR model in the case of both homogeneous and group-specific thresholds. When the thresholds are expected to be common across groups, we impose an equality restriction for threshold estimation, but we still allow group-specific slope coefficients. We evaluate the performance of the proposed method with respect to three aspects: clustering, slope coefficient estimates, and threshold estimates. The accuracy of classification is measured by the average of the misclassification frequency (MF) across replications, defined as

$$\mathrm{MF} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}(\hat{g}_i \neq g_i^0).$$

For slope coefficient estimates, we focus on the bias, root mean squared error (RMSE), and coverage probability (CP) of the two-sided nominal 95% confidence interval, while the threshold parameter estimates are evaluated based on the bias, 95% coverage probability, and average confidence interval length. In the dynamic panels (DGP 3), the evaluation is based on the bias-corrected slope coefficient estimates.

TABLE 3.6 around here.

Table 3.5 presents the average misclassification rate across replications. In general, the method can correctly estimate the group membership, and the misclassification rate decreases quickly as T increases. In the static panel with heteroskedastic error (DGP 1), PSTR can correctly classify at least 96% of individuals when $T = 30$ and roughly 99.7% when $T = 60$. When the errors are serially correlated (DGP 2), PSTR can correctly estimate the group membership for more than 90% of individuals in the worst case. Allowing for dynamics does not deteriorate the good performance of classification, and the misclassification rate remains low in all cases. Interestingly, we find that the misclassification rate is lower in the case of homogeneous threshold parameters than in the case of group-specific thresholds. This is consistent with our theoretical prediction that group identification requires the separation of group-specific slope coefficients instead of heterogeneity among the threshold parameters.

TABLEs 3.6–3.8 around here.

Next, we examine the estimates of the slope coefficients and threshold parameters, and the results are presented in Tables 3.6–3.8. In each DGP, the slope coefficients can be accurately estimated with a small bias, and the coverage probability is generally close to the 95% nominal level. Again, allowing for group-specific thresholds leads to poorer slope and threshold estimates. We find that when the threshold is group specific in DGP 2.1, the RMSE of the slope estimates sometimes decreases disproportionally faster than the speed of the increase in T. This occurs because the relatively large misclassification rate in DGP 2.1 is remarkably reduced by increasing T, and precise classification contributes to better slope estimates.

The threshold parameter is also estimated accurately in all cases, and the average length of the confidence interval shrinks as both N and T increase. We find that the average length of the confidence interval is generally much smaller in the case of a homogeneous threshold than the group specific threshold. This suggests that pooling does improve the efficiency of the threshold estimation for common threshold groups.

3.8 Empirical Applications

We illustrate our procedure through two empirical applications. Our first application examines the investment decision of firms in the presence of financing constraints using the popular data of Hansen (1996). As a second application, we examine the impact of bank deregulation on the distribution of income using the historical data of US states.

3.8.1 Investment and financing constraints

We first apply the proposed PSTR estimator to revisit the question whether capital market imperfections affect firms' investment behavior. An influential and seminal study by Fazzari et al. (1987) suggest that firms' investment is associated with its cash flow only when the firm is constrained by external financing. To investigate the threshold effect of financial constraints, Hansen (1999) examine three investment determinants, i.e., Tobin's Q, cash flow, and leverage, allowing the impact of cash flow to vary depending on whether a firm is financially constrained. This study assume that firms are all homogeneous, such that they face the same threshold parameters and share a common effect of determinants. A number of evidence, however, has shown that firms behave heterogeneously in their financial activities, including investment decisions (see, for example, Spearot (2012), Bernard et al. (2007), and Fazzari et al. (1987)). Heterogeneity may occur not only in the effect of financial variables on investment (even after differentiating constrained and unconstrained firms), but also in threshold parameters. Firms with diversified characteristics may be subjected to distinct threshold levels.

Thus motivated, we revisit the determinants of investment and consider the following model

$$Inv_{it} = \alpha_i + \beta_{1,g_i} x_{i,t-1} \mathbf{1}(q_{i,t-1} \leq \gamma_{g_i}) + \beta_{2,g_i} x_{i,t-1} \mathbf{1}(q_{i,t-1} > \gamma_{g_i}) + \varepsilon_{it}, \quad (3.10)$$

where Inv_{it} is the ratio of investment to capital and α_i denotes the firm fixed effects. We follow Lang et al. (1996) and Hansen (1999) to consider the potential determinants $x_{it} = (Q_{it}, CF_{it}, L_{it})$, where Q_{it} is Tobin's Q, CF_{it} is the ratio of cash flows to capital, and L_{it} denotes leverage. q_{it} is the threshold variable, which we specify as Tobin's Q, cash flow, or leverage, all of which proxy for a certain degree of financial constraints. The lagged values of Q, CF, and L are used as regressors and threshold variables to avoid possible endogeneity (see also Hansen (1999) and Gonzalez et al. (2017)). This model allows a *time-invariant* group pattern of heterogeneity in both slope coefficients and the threshold parameter as well as *time-varying* heterogeneity depending on the realization of the threshold variable. We use the same data set as Hansen (1999) that contains 565 firms over 15 years.

Figure 3.1 around here.

To estimate (3.10), we first determine the number of groups chosen based on the IC. Figure 3.1 displays the value of the IC when we choose the number of groups ranging from 1 to 8 under the three specifications of the threshold variable. For each given number of groups, we estimate the parameters in (3.10) based on 1000 initializations. The IC selects four groups when we use cash flow and Tobin's Q as the threshold variable, while it suggests five groups when leverage is used. We next test the existence of threshold effects using $\mathcal{W}_{NT}^{\text{sum}}$ and $\mathcal{W}_{NT}^{\text{sup}}$ defined in Section 3.5. Both tests (based on 600 bootstrap replications) suggest the presence of threshold effects for the three specifications of the threshold variable, and the common-threshold test tends to reject the null hypothesis of homogeneity in all cases.

TABLE 3.9 around here.

Table 3.9 summarizes the estimation results of (3.10) with three specifications of the threshold variable. When we specify the threshold variable as Tobin's Q, the estimates of the threshold are 10.721, 2.800, 0.854, and 0.282 for the four groups, such that 93%, 87%, 56%, and 15% of the sample fall below the threshold in each group, respectively. In most groups, both Tobin's Q and cash flow are positively associated with investment, as expected. Leverage generally has a negative impact on investment, and this impact is stronger for constrained firms than for unconstrained firms. This result supports the over-investment hypothesis that leverage serves as a disciplining device that prevents firms from over-investing (see, e.g., Jensen (1986) and Seo and Shin (2016)). Group 1 is characterized by relatively low average investment but high average Tobin's Q, while firms in Group 2 are mostly undervalued but still invest aggressively. Group 3 contains very "unsuccessful" firms with highest average leverage as well as lowest average cash flow and Tobin's Q. By contrast, Group 4 is featured by the highest average cash flow and Tobin's Q but lowest average leverage, indicating that firms in this group can be well operated and active in the market. The estimated thresholds for both Groups 1 and 2 occur at the upper quantiles, whereas the effects of cash flow and leverage differ remarkably across the two groups. The effect of cash flow is strongly and positively significant for overvalued firms in Group 2 but less clear for the same type of firms in Group 1. When Tobin's Q is below the threshold, the leverage effect is stronger for firms in Group 2 than for firms in Group 1. For the very "unsuccessful" firms in Group 3, investment is more sensitive to Tobin's Q and cash flow compared with Groups 1 and 2. This is in line with the expectation that the marginal benefit from extra cash and a high asset value is especially high for firms that lack financial resources. Most firms in Group 4 are “successful”, with average Tobin’s Q greater than 1. For a few firms in this group that are severely undervalued and thus financially constrained, both the positive impact of Tobin’s Q and negative impact of leverage are pronounced.

Next, we examine the case in which we use cash flow as the threshold variable. Again, we find a large degree of heterogeneity in the estimates of threshold parameters and slope coefficients. Group 1 contains the burgeoning firms with the largest average cash flow and Tobin’s Q. Most firms in this group fall below the lower threshold regime, with significantly positive effects of Tobin’s Q and cash flow and a negative effect of leverage. The threshold effect in Group 2 is particularly prominent, since the impact of Tobin’s Q and cash flow on investment is much stronger for cash-constrained firms than for unconstrained firms. We find that the effects of Tobin’s Q and cash flow are both negative and sizable for extremely cash-constrained firms in Group 3. Further examination reveals that such firms may borrow money to expand, such that they still invest aggressively when they face a shortage of cash flow. This also explains a large positive effect of leverage when they are cash constrained.

Finally, we use leverage as a threshold variable. In this case, the IC suggests five groups. The first three groups share the same threshold at zero, but the slope coefficient estimates differ. Firms in Groups 1 and 2 generally have a low investment level, but firms in Group 1 are mostly overvalued, while those in Group 2 are often undervalued. When these firms have non-zero debt, their investment is positively affected by their cash flow and Tobin’s Q. The investment behavior of Group 3 is more sensitive to cash flow than that of Groups 1 and 2. Group 4 contains a number of overvalued firms with large cash flow, and the negative effect of leverage on investment in this group is particularly strong in comparison with that of other groups. Group 5, as an extra group, emerges in this case because of seven firms with especially high investment. Such firms also have an abundance of cash and well-valued assets.

These are possibly the aggressive firms, for which we find a strong and positive impacts of cash flow and leverage on investment.

In general, we find a large degree of heterogeneity across firms, which is potentially driven by unobserved firm characteristics, such as their market performance, investment strategy, and managerial risk-taking behavior. Such heterogeneity cannot be captured by conventional threshold regressions. The group pattern varies to some extent for different specifications of the threshold variable. This suggests that the three candidate threshold variables capture distinct aspects of financial constraints.

3.8.2 Bank regulation and income distribution

Our second application concerns the relationship between bank regulation and the distribution of income. Bank regulation plays a crucial role in governing the financial market. It subjects banks to certain restrictions and guidelines regarding, for example, bank mergers, acquisitions, and branching, in the hope of creating a transparent environment for banking institutions, individuals, and corporations. Bank regulations generally consist of two components: (1) licensing that sets requirements for starting a new bank and (2) governmental supervision of the bank's activities. Hence, with stiffer regulations, there could be fewer banks in operation in the market, and banking activities can be more restricted. In shaping regulation policies, income inequality is always one of the central concerns. There exists a theoretical debate on the impact of bank regulation on the distribution of income. On the one hand, imposing stiffer regulatory restrictions on bank mergers and branching is likely to create and protect local banking monopolies, which further leads to higher fixed fees that hurt the poor. Thus, the main motivation for deregulation is to intensify bank competition and improve bank performance. On the other hand, objection on deregulation is also raised due to the fears that centralized banking power would discriminatively curtail the financial opportunities of the poor (Kroszner and Strahan (1999)) and thus amplify inequality.

We revisit the relationship between bank regulation, particularly branch deregulation, and the distribution of income by applying the PSTR estimator. This analysis was first undertaken by Beck et al. (2010) using US state-level data in a standard (fixed effects) panel framework. We employ the same data set that covers 49 US states for 31 years from 1976 to 2006.[2] The impact of branch deregulation may vary remarkably across states depending on their financial market situations, economic performance, demographic features, and so forth. For example, Beck et al. (2010) suggested that the impact of bank deregulation is more prominent if bank performance prior to deregulations is more severely hurt by intrastate branching restrictions. Moreover, deregulation may disproportionately affect different income groups that are characterized by heterogeneous demographic features, and its impact on the distribution of income could also differ across states depending on their economic and financial market performance.

To model the heterogeneous impact of bank deregulation on the distribution of income, we consider the panel structure threshold model as follows:

$$Inc_{it} = \alpha_i + (\beta_{1,g_i} d_{it} + \beta_{1,g_i} x_{it})\mathbf{1}(q_{it} \leq \gamma_{g_i}) + (\beta_{2,g_i} d_{it} + \beta_{2,g_i} x_{it})\mathbf{1}(q_{it} > \gamma_{g_i}) + \varepsilon_{it}, \tag{3.11}$$

where Inc_{it} represents the distribution of income, which is measured by the logistic transformation of the Gini coefficient following Beck et al. (2010) and α_i denotes the state fixed effects.[3] d_{it} is a dummy variable that equals one if a state has implemented deregulation and zero otherwise, and the date of deregulation refers to that on which a state permitted branching via mergers and acquisitions. The control variables in x_{it} include two salient and robust demographic determinants of income inequality based on the cornerstone

[2] The dataset contains 50 US states and the District of Columbia but excludes Delaware and South Dakota.

[3] We also consider alternative measures of the distribution of income, such as the logarithm of the Gini coefficients and Theil index, and the results are qualitatively unchanged.

study of Beck et al. (2010), namely, the percentage of high school dropouts (Dropout) and the unemployment rate (Unemp). We consider four specifications of the threshold variable q_{it}: the two demographic variables in the covariates (Dropout and Unemp), the initial share of small banks, and the initial share of small firms. Obviously, these two demographic variables allow us to examine the potentially heterogeneous impact of deregulation, which depends on the demographic features of the state. The initial share of small banks reflects the degree of bank competition at the date of deregulation, which may disproportionately determine the impact of deregulation. The initial share of small firms also plays a role in influencing the impact of deregulation because the barriers to obtaining credit from distant banks is greater for small firms than for larger firms, leading to a heterogeneous impact across states with different initial shares of small firms. To analyze the effect of the two share variables, we have to use a subsample of the data with 37 states if we wish to have a balanced panel. Detailed information on the dataset and its source can be found in Beck et al. (2010).

The moderate effect of the two initial share variables was first proposed and analyzed by Beck et al. (2010) in a difference-in-difference (DiD) framework. The advantages of (3.11) compared to the conventional DiD approach are as follows: (1) DiD can only report a positive or negative (linear) effect of the moderating variables, (e.g., the same value for all levels of the initial share of small firms), while PSTR provides information on how such an effect varies (possibly non-linearly) across different levels of these variables; (2) DiD captures only *observed* heterogeneity that is driven by the moderating variables, while PSTR allows us to model the *unobserved* heterogeneity as the group pattern is fully unrestricted.

FIGURE 3.2 around here.

We first examine the optimal number of groups chosen by the IC. Figure

3.2 displays the value of IC when we choose the number of groups ranging from 1 to 8 under four specifications of threshold variables. The IC robustly chooses two groups as the optimal specification in all cases. The p-values of $\mathcal{W}_{NT}^{\text{sup}}$ and $\mathcal{W}_{NT}^{\text{sum}}$ suggest that the impact of explanatory variables does exhibit threshold effects for all four specifications of the threshold variable, although to different extents.

TABLE 3.10 around here.

FIGURE 3.3 around here.

Table 3.10 presents the estimated threshold and effects of the explanatory variables. In general, we find a large degree of heterogeneity both across groups and across different levels of the threshold variables. We first examine the impact of deregulation if we specify the threshold variable as the rate of high school dropouts. In this case, the test for the common threshold rejects the null of homogeneity with p-value 0.03; thus, we allow the threshold coefficient to vary across groups in our estimation. The estimation is based on 10000 initial values, and the same number of initializations is used for the estimation with other threshold variables below. Our method assigns 26 states into Group 1 and 23 states into Group 2. Interestingly, the classification coincides with the geographic location to some extent (see Figure 3.3). Group 1 contains mainly coastal states, such as Washington, Oregon, California, New York, New Jersey, Massachusetts, Vermont, Virginia, and Florida. These states are generally characterized by good economic performance and active financial markets. Group 2 includes states with less active financial markets, including mostly inland and Southeastern states, such as Montana, North Dakota, Minnesota, Nebraska, Iowa, North and South Carolina, and Georgia. The two groups are distinguished by the effects of covariates and the threshold. The estimated threshold of Group 1 is 0.295, such that 73% of observations fall below the threshold. The effect of deregulation on income inequality is significantly negative (-0.0291) when the dropout rate is below the threshold, and it is of a similar size as reported by Beck et al. (2010) (see column (1) of Table II of Beck et al. (2010)). Nevertheless, this effect becomes insignificant when the dropout rate is particularly high. For Group 2, the estimated threshold is much smaller with 1.5% of the sample in the lower threshold regime, and a majority of the sample in this group reports a significantly negative impact of deregulation on inequality. Compared with Group 1, the inequality reduction induced by deregulation is much less sizeable in Group 2. This is possibly because bank competition is disproportionately intensified by deregulation in coastal states than in inland/south-eastern states, leading to better bank performance and further to a larger reduction in income inequality.

Next, we examine the deregulation effect when we specify the threshold variable as the unemployment rate. The p-value of the common-threshold test is 0.01, strongly favoring the hypothesis of the heterogeneous threshold coefficients. The group pattern estimated in this case is closely in line with the specification above, with only two states (Ohio and Wyoming) switching their group memberships. We again find a large degree of heterogeneity across the two groups. The estimated thresholds are 9.8 for Group 1 and 2.6 for Group 2, which leads to about 95% and 10% of the sample below the threshold, respectively. The impact of deregulation on inequality is significantly negative for the majority of the sample in both groups but insignificant for the minority. These results suggest that branching deregulation can reduce income inequality in most states, but the magnitude of reduction is bigger in Group 1. However, for the states with an extreme unemployment and dropout rate, deregulation does not significantly help reduce inequality and even enlarges inequality.

To explicitly examine how the degree of bank competition influences the impact of deregulation, we consider the threshold variable as the initial share of small bank. Owing to the unavailability of the initial share in some states, we employ a subsample of the data with 37 states. In this case, the test for the common threshold strongly suggests homogeneity; thus, we proceed with the estimation imposing the homogeneity restriction. The states are again classified into coastal and inland/south-eastern groups with only four states (Kentucky, New Hampshire, North Dakota, and West Virginia) switching their group memberships compared with the case of the dropout rate being the threshold variable. This confirms the heterogeneity of geographic locations and demonstrates the robustness of the estimated group pattern. The estimated threshold is 0.1723 for both groups (due to the common-threshold restriction), such that most observations are in the lower threshold regime. The impact of deregulation is negative in all groups and all regimes, but the magnitude of inequality reduction is larger when the share is beyond the threshold in both groups. This result is in line with the expectation that states with a comparatively high ratio of small banks benefit more from eliminating branching restrictions, as such restrictions that protect small banks from competition have been particularly harmful to bank operations. Since most states are in the lower threshold regime in both groups, we see that the magnitude of inequality reduction induced by deregulation is larger for the majority in Group 1 than the majority in Group 2 as in the previous states.

Finally, we consider the potential threshold effect induced by the initial share of small firms. Again, the test for the common threshold fails to reject the null of homogeneity; thus, we estimate the model restricting the two groups to share the same threshold. The estimated group pattern remains highly similar to the above case using the initial share of small banks as the threshold variable, with only one state changing its group membership. The estimated threshold in both groups is in the 0.783 quantile of the initial share of small firms. Interestingly, when we specify the threshold variable as the two initial-share variables, the estimated slope coefficients in Group 1 are close or even identical.

This is, of course, due to the robustness of the classification; moreover, it implies that the two share variables result in similar sample thresholding for Group 1. However, sample thresholding by the two share variables differs in Group 2, and the impact of deregulation is not significant in Group 2 when we use the initial shares of small firms as the threshold variable. In both groups, the inequality reduction is more sizable when the initial share of small firms is beyond the threshold. This confirms the theoretical argument that the impact of deregulation is more pronounced in states with a large ratio of small firms before deregulation, since the existence of branching restrictions impedes the growth of small firms that typically face greater barriers to obtaining credit from distant banks and thus enlarges inequality (Beck et al., 2010).

To summarize, the PSTR estimates provide at least two new important insights that are not provided by standard panel data models with interaction terms. First, we find a large degree of heterogeneity between the two groups even after controlling for the threshold effect, and the impact of deregulation is more sizeable in the group containing most coastal states. This result is robust regardless of the way in which we specify the threshold variable. The group structure coincides with the geographic locations to some extent but not precisely, and this latent group pattern is difficult, if not impossible, to recover using standard panel data approaches. Second, we find a clear threshold effect in each of the two groups. The degree of inequality reduction induced by deregulation depends on the demographic features and the composition of financial markets. Such a group pattern heterogeneity and nonlinear feature of threshold effects can be simultaneously captured by our PSTR model but not by the conventional DiD approach.

3.9 Conclusion

In this paper, we consider the least squares estimation of a panel structure threshold regression (PSTR) model, where both the slope coefficients and threshold parameters may exhibit latent group structures. We summarize the practical procedure of using this model as follows. The procedure starts with selecting the right number of groups using the IC. With the number of groups given, we first test the presence of threshold effects using the two proposed Wald-type statistics. If there are threshold effects, we then need to test whether the threshold coefficients also vary across groups. Next, we can proceed with the estimation with or without the homogeneity of thresholds imposed, depending on the results of the common-threshold test. We show that we can consistently estimate the latent group structure and estimators of the slope and the threshold coefficients are asymptotically equivalent to the infeasible estimators that are obtained as if the true group structures were known. Moreover, the standard inference based on LR test statistic can provide a correct coverage for the group-specific threshold parameters.

There are several interesting topics for further research. First, we only allow individual fixed effects in our PSTR model. It is possible to also allow for fixed time effects in the model, but this will complicate the analysis to a great deal. Second, it is very interesting but challenging to study the PSTR model with interactive fixed effects, which can incorporate strong cross-sectional dependence in many macro or financial data. Third, we do not allow the latent group structures to change over time. It is interesting and extremely challenging to study PSTR models with a time-varying latent group structure. Fourth, as mentioned in the introduction, we can also consider a PSTR model with endogenous regressors and threshold variables and latent group structures, which would require the use of GMM-type estimation. Fifth, one can also consider a PSTR model with multiple thresholds or multiple threshold variables by extending the works of Li and Ling (2012) and Seo and Linton (2007) to the panel setup with or without latent group structures.

Tables and Figures

Table 3.1: Group number selection frequency using IC when $G^0 = 3$

			No threshold effect					With threshold effect				
	N	T	1	2	3	4	5	1	2	3	4	5
DGP 1.1	50	30	0.000	0.000	**0.967**	0.033	0.000	0.000	0.000	**0.976**	0.024	0.000
	50	60	0.000	0.000	**0.972**	0.026	0.002	0.000	0.000	**0.997**	0.003	0.000
	100	30	0.000	0.000	**1.000**	0.000	0.000	0.000	0.000	**0.998**	0.002	0.000
	100	60	0.000	0.000	**1.000**	0.000	0.000	0.000	0.000	**1.000**	0.000	0.000
DGP 1.2	50	30	0.000	0.000	**0.974**	0.026	0.000	0.000	0.000	**0.976**	0.024	0.000
	50	60	0.000	0.000	**0.998**	0.002	0.000	0.000	0.000	**0.998**	0.002	0.000
	100	30	0.000	0.000	**0.996**	0.004	0.000	0.000	0.000	**1.000**	0.000	0.000
	100	60	0.000	0.000	**1.000**	0.000	0.000	0.000	0.000	**1.000**	0.000	0.000
DGP 2.1	50	30	0.000	0.000	**0.982**	0.018	0.000	0.000	0.000	**0.982**	0.016	0.002
	50	60	0.000	0.000	**0.996**	0.004	0.000	0.000	0.000	**0.992**	0.008	0.000
	100	30	0.000	0.000	**0.998**	0.002	0.000	0.000	0.000	**0.998**	0.002	0.000
	100	60	0.000	0.000	**1.000**	0.000	0.000	0.000	0.000	**1.000**	0.000	0.000
DGP 2.2	50	30	0.000	0.000	**0.994**	0.006	0.000	0.000	0.000	**0.946**	0.032	0.022
	50	60	0.000	0.000	**1.000**	0.000	0.000	0.000	0.000	**0.998**	0.002	0.000
	100	30	0.000	0.000	**0.996**	0.004	0.000	0.000	0.000	**0.997**	0.003	0.000
	100	60	0.000	0.000	**1.000**	0.000	0.000	0.000	0.000	**1.000**	0.000	0.000
DGP 3.1	50	30	0.000	0.000	**0.992**	0.008	0.000	0.000	0.000	**0.998**	0.002	0.000
	50	60	0.000	0.000	**1.000**	0.000	0.000	0.000	0.000	**1.000**	0.000	0.000
	100	30	0.000	0.000	**1.000**	0.000	0.000	0.000	0.000	**0.998**	0.000	0.000
	100	60	0.000	0.000	**1.000**	0.000	0.000	0.000	0.000	**1.000**	0.000	0.000
DGP 3.2	50	30	0.000	0.000	**0.998**	0.002	0.000	0.000	0.000	**0.994**	0.006	0.000
	50	60	0.000	0.000	**1.000**	0.000	0.000	0.000	0.000	**1.000**	0.000	0.000
	100	30	0.000	0.000	**1.000**	0.000	0.000	0.000	0.000	**0.998**	0.002	0.000
	100	60	0.000	0.000	**1.000**	0.000	0.000	0.000	0.000	**1.000**	0.000	0.000

Table 3.2: Rejection frequency of test for existence of threshold effect: Heterogeneous thresholds

			No threshold effect			Weak threshold effect			Strong threshold effect		
			($c_1 = 0$, $c_2 = 0$)			($c_1 = 1/5$, $c_2 = 1/15$)			($c_1 = 1/2$, $c_2 = 1/10$)		
	N	T	1%	5%	10%	1%	5%	10%	1%	5%	10%
							$\mathcal{W}_{NT}^{\text{sup}}$				
DGP 1.1	50	30	0.026	0.072	0.122	0.096	0.228	0.332	0.728	0.833	0.923
	50	60	0.006	0.044	0.088	0.160	0.304	0.496	0.918	0.985	1.000
	100	30	0.016	0.050	0.084	0.160	0.308	0.436	0.923	0.980	0.993
	100	60	0.010	0.044	0.080	0.276	0.512	0.606	1.000	1.000	1.000
DGP 2.1	50	30	0.036	0.094	0.138	0.108	0.202	0.308	0.533	0.755	0.878
	50	60	0.008	0.058	0.088	0.096	0.240	0.332	0.760	0.923	0.943
	100	30	0.024	0.074	0.120	0.126	0.294	0.332	0.788	0.930	0.968
	100	60	0.010	0.044	0.080	0.140	0.342	0.442	0.968	0.993	0.998
DGP 3.1	50	30	0.024	0.070	0.150	0.160	0.306	0.444	0.826	0.942	0.970
	50	60	0.012	0.050	0.106	0.260	0.526	0.642	0.992	1.000	1.000
	100	30	0.018	0.062	0.118	0.212	0.492	0.610	0.984	0.998	1.000
	100	60	0.006	0.058	0.086	0.520	0.770	0.868	1.000	1.000	1.000
							$\mathcal{W}_{NT}^{\text{sum}}$				
DGP 1.1	50	30	0.030	0.076	0.148	0.152	0.276	0.376	0.853	0.915	0.968
	50	60	0.012	0.042	0.086	0.224	0.358	0.544	0.980	1.000	1.000
	100	30	0.020	0.060	0.102	0.244	0.398	0.554	0.985	0.995	1.000
	100	60	0.016	0.044	0.080	0.378	0.622	0.686	1.000	1.000	1.000
DGP 2.1	50	30	0.042	0.106	0.154	0.148	0.260	0.342	0.673	0.855	0.928
	50	60	0.016	0.056	0.090	0.122	0.274	0.382	0.880	0.963	0.980
	100	30	0.032	0.112	0.186	0.216	0.418	0.450	0.925	0.973	0.980
	100	60	0.012	0.060	0.086	0.244	0.436	0.530	0.995	1.000	1.000
DGP 3.1	50	30	0.012	0.064	0.098	0.178	0.312	0.436	0.888	0.962	0.986
	50	60	0.004	0.030	0.080	0.302	0.574	0.668	0.996	1.000	1.000
	100	30	0.014	0.054	0.094	0.272	0.528	0.654	1.000	0.998	1.000
	100	60	0.004	0.036	0.068	0.596	0.798	0.886	1.000	1.000	1.000

Table 3.3: Rejection frequency of test for existence of threshold effect: Homogeneous thresholds

			No threshold effect			Weak threshold effect			Strong threshold effect		
			($c_1 = 0$, $c_2 = 0$)			($c_1 = 1/5$, $c_2 = 1/15$)			($c_1 = 1/2$, $c_2 = 1/10$)		
	N	T	1%	5%	10%	1%	5%	10%	1%	5%	10%
							$\mathcal{W}_{NT}^{\text{sup}}$				
DGP 1.2	50	30	0.024	0.072	0.118	0.126	0.356	0.434	0.818	0.964	0.990
	50	60	0.006	0.044	0.088	0.164	0.408	0.526	0.984	0.996	1.000
	100	30	0.016	0.050	0.095	0.208	0.400	0.512	0.978	0.996	1.000
	100	60	0.010	0.044	0.085	0.412	0.635	0.734	1.000	1.000	1.000
DGP 2.2	50	30	0.032	0.076	0.138	0.090	0.220	0.360	0.692	0.926	0.948
	50	60	0.016	0.066	0.118	0.140	0.282	0.404	0.906	0.986	0.994
	100	30	0.020	0.068	0.116	0.122	0.330	0.440	0.908	0.982	0.996
	100	60	0.012	0.052	0.096	0.264	0.474	0.620	0.998	0.998	0.998
DGP 3.2	50	30	0.024	0.094	0.174	0.256	0.474	0.626	0.940	0.990	1.000
	50	60	0.008	0.066	0.118	0.454	0.700	0.804	1.000	1.000	1.000
	100	30	0.012	0.086	0.134	0.398	0.670	0.730	1.000	1.000	1.000
	100	60	0.007	0.056	0.104	0.740	0.906	0.966	1.000	1.000	1.000
							$\mathcal{W}_{NT}^{\text{sum}}$				
DGP 1.2	50	30	0.029	0.076	0.140	0.198	0.400	0.454	0.962	0.992	0.996
	50	60	0.012	0.042	0.086	0.300	0.508	0.672	0.998	1.000	1.000
	100	30	0.018	0.060	0.114	0.362	0.540	0.652	0.998	1.000	1.000
	100	60	0.015	0.044	0.086	0.620	0.780	0.881	1.000	1.000	1.000
DGP 2.2	50	30	0.034	0.076	0.154	0.146	0.322	0.408	0.912	0.970	0.980
	50	60	0.008	0.070	0.124	0.190	0.400	0.548	0.986	1.000	1.000
	100	30	0.041	0.099	0.156	0.298	0.442	0.566	0.990	0.996	1.000
	100	60	0.014	0.056	0.096	0.394	0.628	0.734	1.000	1.000	1.000
DGP 3.2	50	30	0.012	0.068	0.138	0.324	0.520	0.626	0.990	1.000	1.000
	50	60	0.006	0.036	0.070	0.560	0.760	0.816	1.000	1.000	1.000
	100	30	0.010	0.064	0.090	0.480	0.734	0.816	1.000	1.000	1.000
	100	60	0.008	0.044	0.088	0.860	0.982	0.992	1.000	1.000	1.000

Table 3.4: Rejection frequency for the test of homogeneous thresholds

Threshold			Homogeneous			Weakly heterogeneous			Strongly heterogeneous		
			$\gamma = [1, 1, 1]$			$\gamma = [0.85, 1, 1.15]$			$\gamma = [0.5, 1, 1.5]$		
	N	T	1%	5%	10%	1%	5%	10%	1%	5%	10%
DGP 1	50	30	0.013	0.076	0.110	0.810	0.904	0.960	0.968	0.980	0.994
	50	60	0.018	0.061	0.114	0.994	0.986	0.996	1.000	1.000	1.000
	100	30	0.014	0.064	0.096	0.990	0.998	1.000	1.000	1.000	1.000
	100	60	0.012	0.046	0.112	1.000	1.000	1.000	1.000	1.000	1.000
DGP 2	50	30	0.014	0.034	0.052	0.344	0.592	0.690	0.116	0.312	0.408
	50	60	0.010	0.038	0.056	0.862	0.950	0.948	0.498	0.710	0.808
	100	30	0.010	0.052	0.058	0.844	0.932	0.956	0.498	0.714	0.794
	100	60	0.008	0.042	0.050	0.994	0.998	1.000	0.920	0.970	0.994
DGP 3	50	30	0.006	0.040	0.064	0.936	0.972	0.900	0.692	0.856	0.900
	50	60	0.010	0.046	0.048	1.000	1.000	1.000	0.968	0.994	0.998
	100	30	0.006	0.042	0.066	0.998	1.000	1.000	0.972	0.992	0.998
	100	60	0.010	0.036	0.040	1.000	1.000	1.000	1.000	1.000	1.000

Table 3.5: Average misclassification rate

	$N = 50$		$N = 100$	
	$T = 30$	$T = 60$	$T = 30$	$T = 60$
DGP 1.1	0.0365	0.0032	0.0316	0.0026
DGP 1.2	0.0203	0.0011	0.0179	0.0013
DGP 2.1	0.0963	0.0141	0.0697	0.0124
DGP 2.2	0.0509	0.0076	0.0470	0.0075
DGP 3.1	0.0041	0.0001	0.0028	0.0000
DGP 3.2	0.0011	0.0000	0.0015	0.0002

Table 3.6: Estimates of coefficients and threshold values: Heteroskedastic error (DGPs 1.1-1.2)

		β_1			β_2			γ		
					DGP 1.1: D^0= $(0.5, 1, 1.5)'$					
		Bias	RMSE	CP	Bias	RMSE	CP	Bias	CP	Length
$N = 50$	Group 1	-0.001	0.078	0.908	-0.002	0.056	0.915	0.009	0.958	0.549
$T = 30$	Group 2	0.003	0.097	0.895	0.015	0.107	0.893	0.018	0.923	0.373
	Group 3	0.002	0.078	0.920	0.004	0.103	0.890	0.001	0.960	0.545
$N = 50$	Group 1	-0.004	0.052	0.925	0.000	0.035	0.940	0.002	0.963	0.214
$T = 60$	Group 2	-0.001	0.042	0.925	-0.001	0.042	0.928	0.002	0.965	0.202
	Group 3	-0.003	0.037	0.948	-0.001	0.055	0.913	0.000	0.973	0.246
$N = 100$	Group 1	0.001	0.055	0.922	-0.002	0.038	0.898	-0.003	0.966	0.245
$T = 30$	Group 2	0.004	0.045	0.920	0.000	0.048	0.904	-0.003	0.948	0.207
	Group 3	0.007	0.035	0.928	-0.003	0.057	0.922	0.001	0.968	0.240
$N = 100$	Group 1	0.003	0.037	0.944	-0.002	0.024	0.938	0.000	0.972	0.125
$T = 60$	Group 2	0.003	0.030	0.938	-0.001	0.029	0.942	-0.002	0.970	0.108
	Group 3	0.000	0.025	0.920	-0.004	0.036	0.946	-0.004	0.962	0.119
					DGP 1.2: D^0= $(1, 1, 1)'$					
		Bias	RMSE	CP	Bias	RMSE	CP	Bias	CP	Length
$N = 50$	Group 1	0.001	0.057	0.938	-0.010	0.060	0.928	-0.002	0.928	0.073
$T = 30$	Group 2	-0.002	0.064	0.903	-0.010	0.060	0.923			
	Group 3	0.006	0.062	0.923	-0.011	0.061	0.913			
$N = 50$	Group 1	0.004	0.038	0.960	-0.003	0.040	0.943	0.001	0.933	0.049
$T = 60$	Group 2	0.001	0.043	0.927	-0.003	0.043	0.917			
	Group 3	0.004	0.042	0.940	-0.003	0.038	0.957			
$N = 100$	Group 1	0.001	0.042	0.930	-0.009	0.041	0.947	0.001	0.940	0.051
$T = 30$	Group 2	0.006	0.045	0.913	-0.006	0.041	0.933			
	Group 3	0.006	0.040	0.930	-0.011	0.039	0.963			
$N = 100$	Group 1	0.003	0.026	0.963	-0.002	0.028	0.950	-0.001	0.947	0.027
$T = 60$	Group 2	0.005	0.029	0.933	-0.002	0.029	0.940			
	Group 3	0.004	0.027	0.953	-0.004	0.029	0.943			

Table 3.7: Estimates of coefficients and threshold values: Autoregressive error (DGPs 2.1-2.2)

			β_1			β_2			γ	
					DGP 2.1: D^0= $(0.5, 1, 1.5)'$					
		Bias	RMSE	CP	Bias	RMSE	CP	Bias	CP	Length
$N = 50$	Group 1	-0.014	0.153	0.834	0.015	0.163	0.874	0.048	0.932	0.797
$T = 30$	Group 2	-0.008	0.198	0.812	0.032	0.225	0.802	-0.010	0.848	0.605
	Group 3	-0.024	0.140	0.858	0.001	0.203	0.856	-0.034	0.936	0.924
$N = 50$	Group 1	-0.008	0.092	0.914	-0.001	0.043	0.930	-0.006	0.966	0.374
$T = 60$	Group 2	-0.003	0.051	0.924	0.002	0.050	0.942	0.004	0.964	0.291
	Group 3	-0.005	0.050	0.922	0.005	0.073	0.892	-0.014	0.958	0.433
$N = 100$	Group 1	-0.021	0.080	0.894	-0.009	0.050	0.882	-0.015	0.960	0.380
$T = 30$	Group 2	-0.002	0.076	0.840	0.000	0.073	0.856	0.003	0.918	0.302
	Group 3	0.006	0.057	0.880	0.013	0.075	0.910	-0.003	0.946	0.331
$N = 100$	Group 1	0.002	0.045	0.944	0.002	0.031	0.932	0.001	0.980	0.195
$T = 60$	Group 2	-0.003	0.037	0.930	0.001	0.037	0.934	0.002	0.950	0.158
	Group 3	-0.002	0.031	0.942	0.000	0.046	0.940	0.000	0.972	0.181
					DGP 2.2: D^0= $(1, 1, 1)'$					
		Bias	RMSE	CP	Bias	RMSE	CP	Bias	CP	Length
$N = 50$	Group 1	-0.002	0.067	0.937	-0.008	0.074	0.920	0.001	0.960	0.181
$T = 30$	Group 2	0.012	0.108	0.877	0.000	0.091	0.923			
	Group 3	0.009	0.091	0.917	-0.008	0.097	0.927			
$N = 50$	Group 1	0.005	0.048	0.965	0.000	0.050	0.938	-0.001	0.985	0.079
$T = 60$	Group 2	0.001	0.053	0.918	-0.002	0.048	0.945			
	Group 3	0.004	0.051	0.930	-0.004	0.049	0.955			
$N = 100$	Group 1	-0.004	0.053	0.928	-0.017	0.061	0.851	-0.001	0.950	0.099
$T = 30$	Group 2	0.001	0.056	0.914	-0.002	0.057	0.910			
	Group 3	0.019	0.053	0.914	-0.002	0.051	0.932			
$N = 100$	Group 1	0.001	0.033	0.950	-0.005	0.036	0.930	0.000	0.980	0.051
$T = 60$	Group 2	-0.001	0.031	0.965	-0.001	0.033	0.965			
	Group 3	0.004	0.033	0.965	-0.001	0.036	0.920			

Table 3.8: Estimates of coefficients and threshold values: Dynamic panel (DGPs 3.1-3.2)

		β_1			β_2			γ		
					DGP 3.1: D^0= $(0.5, 1, 1.5)'$					
		Bias	RMSE	CP	Bias	RMSE	CP	Bias	CP	Length
$N = 50$	Group 1	-0.007	0.035	0.923	-0.010	0.025	0.920	-0.006	0.940	0.184
$T = 30$	Group 2	-0.003	0.017	0.963	-0.007	0.020	0.907	0.003	0.970	0.161
	Group 3	-0.002	0.012	0.923	-0.007	0.019	0.877	-0.008	0.947	0.147
$N = 50$	Group 1	-0.003	0.025	0.930	-0.005	0.017	0.950	0.001	0.973	0.095
$T = 60$	Group 2	-0.002	0.013	0.953	-0.003	0.012	0.943	-0.002	0.940	0.073
	Group 3	-0.001	0.007	0.960	-0.001	0.011	0.950	0.000	0.947	0.073
$N = 100$	Group 1	-0.007	0.027	0.927	-0.009	0.019	0.917	0.000	0.973	0.110
$T = 30$	Group 2	-0.003	0.014	0.937	-0.007	0.015	0.933	0.000	0.943	0.082
	Group 3	-0.002	0.008	0.940	-0.005	0.013	0.907	-0.002	0.947	0.074
$N = 100$	Group 1	-0.005	0.019	0.923	-0.004	0.013	0.927	-0.001	0.947	0.059
$T = 60$	Group 2	-0.002	0.009	0.930	-0.002	0.009	0.953	0.000	0.960	0.044
	Group 3	-0.001	0.005	0.950	-0.003	0.009	0.920	0.000	0.967	0.038
					DGP 3.2: D^0= $(1, 1, 1)'$					
		Bias	RMSE	CP	Bias	RMSE	CP	Bias	CP	Length
$N = 50$	Group 1	-0.008	0.029	0.957	-0.014	0.032	0.910	0.001	0.977	0.050
$T = 30$	Group 2	-0.004	0.019	0.930	-0.005	0.019	0.910			
	Group 3	0.000	0.012	0.937	-0.005	0.013	0.927			
$N = 50$	Group 1	-0.003	0.018	0.957	-0.005	0.021	0.937	0.000	0.950	0.024
$T = 60$	Group 2	-0.002	0.013	0.940	-0.002	0.012	0.940			
	Group 3	-0.001	0.008	0.953	-0.002	0.009	0.923			
$N = 100$	Group 1	-0.008	0.020	0.957	-0.010	0.025	0.897	0.001	0.983	0.029
$T = 30$	Group 2	-0.002	0.013	0.950	-0.006	0.014	0.933			
	Group 3	0.000	0.009	0.953	-0.005	0.010	0.893			
$N = 100$	Group 1	-0.006	0.017	0.948	-0.004	0.018	0.938	0.000	0.964	0.201
$T = 60$	Group 2	-0.002	0.011	0.942	-0.001	0.012	0.944			
	Group 3	-0.002	0.007	0.958	-0.003	0.008	0.922			

Table 3.9: Investment and financial constraint: Estimated threshold and slope coefficients

Threshold variable		Tobin's Q				
		Group 1	Group 2	Group 3	Group 4	
γ (Lower regime %)		10.721 (93%)	2.800 (87%)	0.854 (56%)	0.282 (15%)	
β_1	Q	0.0081***	0.0716***	0.1537***	1.3450***	
		(0.0008)	(0.0029)	(0.0146)	(0.0631)	
	CF	0.0918***	0.0977***	0.3278***	−4.6433***	
		(0.0051)	(0.0121)	(0.0366)	(0.1563)	
	L	−0.0158***	−0.0671***	0.0206	−0.8063***	
		(0.0039)	(0.0068)	(0.0204)	(0.1025)	
β_2	Q	0.0086***	0.0134***	0.0553***	−0.0004	
		(0.0010)	(0.0052)	(0.0084)	(0.0003)	
	CF	−0.0194*	0.3007***	−0.4886***	−0.0161***	
		(0.0116)	(0.0579)	(0.0617)	(0.0080)	
	L	0.0668	0.0798	0.1251***	−0.0143***	
		(0.0803)	(0.0830)	(0.0270)	(0.0061)	
Threshold variable		Cash flow				
		Group 1	Group 2	Group 3	Group 4	
γ (Lower regime %)		0.853 (98%)	0.279 (66%)	−0.084 (1.6%)	−0.343(0.2%)	
β_1	Q	0.0013***	0.1447***	−0.4135***	−0.0034***	
		(0.0004)	(0.0075)	(0.0295)	(0.0009)	
	CF	0.0684***	0.1545***	−2.0022***	−0.1496***	
		(0.0052)	(0.0216)	(0.0697)	(0.0411)	
	L	−0.0096*	0.0203*	52.6850*	−0.2208***	
		(0.0041)	(0.0106)	(0.1284)	(0.0435)	
β_2	Q	−0.0010***	0.0068***	0.0468***	0.0117***	
		(0.0005)	(0.0013)	(0.0028)	(0.0013)	
	CF	0.0806***	0.0054	−0.0835***	0.2958***	
		(0.0081)	(0.0138)	(0.0128)	(0.0110)	
	L	−0.0996***	0.1644***	−0.0399***	−0.0730***	
		(0.0193)	(0.0256)	(0.0060)	(0.0083)	
Threshold variable		Leverage				
		Group 1	Group 2	Group 3	Group 4	Group 5
γ (Lower regime %)		0 (8.5%)	0 (8.5%)	0 (8.5%)	0.002 (8.9%)	0.806 (98%)
β_1	Q	−0.0003	0.0957***	0.0014	0.0107***	0.0538***
		(0.0003)	(0.0109)	(0.0027)	(0.0012)	(0.0131)
	CF	0.0584***	−0.0047	0.2276***	−0.0519***	−0.8202***
		(0.0097)	(0.0509)	(0.0247)	(0.0132)	(0.1507)
	L	−0.0083	−0.0297	0.0816	−0.8464***	0.1648***
		(0.0165)	(0.0566)	(0.0549)	(0.0637)	(0.0337)
β_2	Q	0.0039***	0.0804***	0.0117***	0.0003	1.2055***
		(0.0008)	(0.0033)	(0.0021)	(0.0005)	(0.1284)
	CF	0.0304***	0.0423***	0.3854***	0.1164***	−4.3237***
		(0.0054)	(0.0133)	(0.0163)	(0.0086)	(0.2463)
	L	0.0024	−0.0535***	0.0258***	−0.1168***	0.2734***
		(0.0042)	(0.0072)	(0.0078)	(0.0099)	(0.0383)

Table 3.10: Impact of bank deregulation: Estimated threshold and slope coefficients

Threshold variable		Dropout rate		Unemployment rate	
		Group 1	Group 2	Group 1	Group 2
γ (Lower regime %)		0.295 (73%)	0.041 (1.5%)	9.80 (95%)	2.60 (10%)
β_1	Dereg	−0.0291***	0.2444***	−0.0316***	−0.0228
		(0.0082)	(0.0576)	(0.0080)	(0.0427)
	Dropout	−0.6749***	3.3793***	−0.6959***	−5.2629
		(0.0778)	(0.7635)	(0.0805)	(3.0658)
	Unemp	0.0032*	0.0390*	0.0007	0.1566***
		(0.0020)	(0.0198)	(0.0022)	(0.0489)
β_2	Dereg	−0.1672*	−0.0199***	0.0339	−0.0197***
		(0.0779)	(0.0086)	(0.0415)	(0.0088)
	Dropout	−1.1961***	−0.2286***	−0.4149	−0.2125***
		(0.2666)	(0.0614)	(0.6825)	(0.0629)
	Unemp	0.0626***	0.0263***	0.0212***	0.0263***
		(0.0118)	(0.0021)	(0.0051)	(0.0022)
Threshold variable		Ratio of small banks		Ratio of small firms	
		Group 1	Group 2	Group 1	Group 2
γ (Lower regime %)		0.1723 (94.5%)		0.8943 (78.3%)	
β_1	Dereg	−0.0291***	−0.0067	−0.0354***	0.0003
		(0.0092)	(0.0091)	(0.0091)	(0.0117)
	Dropout	−0.7805***	−0.2432***	−0.8015***	−0.3306***
		(0.0933)	(0.0791)	(0.0924)	(0.0968)
	Unemp	0.0038	0.0253***	0.0030	0.0244***
		(0.0026)	(0.0022)	(0.0025)	(0.0026)
β_2	Dereg	−0.0655	−0.1555***	−0.0655	−0.0089
		(0.0455)	(0.0479)	(0.0455)	(0.0141)
	Dropout	0.5417***	−1.7011***	0.5417***	−0.0295
		(0.2723)	(0.4793)	(0.2723)	(0.1294)
	Unemp	0.0573***	−0.0008	0.0573***	0.0303***
		(0.0179)	(0.0092)	(0.0179)	(0.0042)

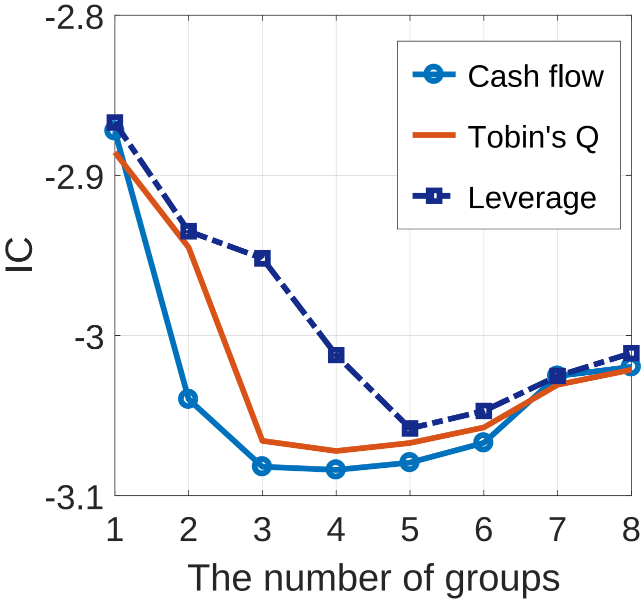

Figure 3.1: The information criterion for determining the number of groups in the investment and financial constraint application

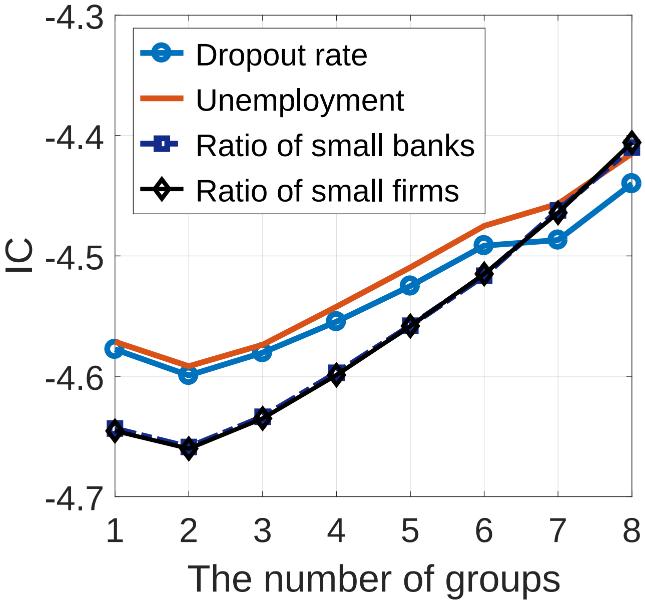

Figure 3.2: The information criterion for determining the number of groups in the bank deregulation application

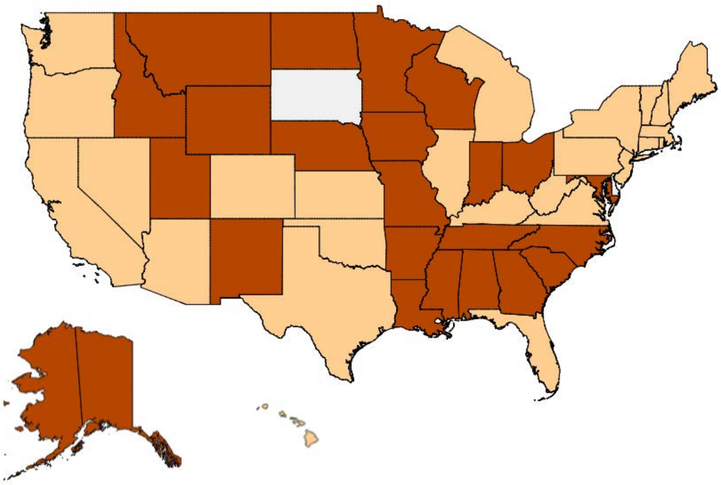

Figure 3.3: Estimates of the group memebership of US states ($G = 2$)

Chapter 4

High-dimensional VAR with Common Factors

4.1 Introduction

Vector autoregressions (VARs) were introduced and a limit theory for estimation and inference developed in a pathbreaking study by Mann and Wald (1943) that also considered structural VAR formulations.[1] The VAR approach was further developed and promoted for empirical macroeconomic research in an influential paper by Sims (1980). Since then the methodology has become one of the most heavily used tools in the applied finance and macroeconomic literatures, giving a simple and useful method of capturing rich dynamics and interconnectedness in multiple time series. Unrestricted VARs may be efficiently estimated by least squares regression, which makes them particularly attractive in applied research. But low dimensional VARs often suffer from omitted variable bias, which makes the approach vulnerable to misleading inference on both coefficients and impulse responses. In a series of articles Sims and coauthors have explored whether to include more variables in VAR formulations to raise their forecasting performance (see Sims (1992, 1993); Leeper et al. (1996)).

In the absence of restrictions the number of VAR coefficients increases

[1] The Mann and Wald extension to the structural VAR (SVAR) case was developed in the final section of their paper but seems largely to have been forgotten in the vast literature on that topic that has emerged in the last few decades. For further discussion, readers are referred to Hurn et al. (2019)

quadratically, making VAR estimation inevitably a high dimensional problem as the number of variables increases. The dynamic factor model (DFM), introduced by Geweke (1977), provides a tool to summarize information from a large number of time series while avoiding some of the problems of high dimensionality. Since then, a large literature has emerged on DFMs . Examples of theoretical work include Forni et al. (2000), Bai and Ng (2002), Bai (2003), and Hallin and Liška (2007); and in applied finance and macroeconomics, various studies document the useful capacity of DFMs in capturing comovements among macroeconomic or financial time series (e.g., Fama and French (1993); Stock and Watson (1999) and 2002; Giannone et al. (2004) ; Ludvigson and Ng (2007); and Cheng and Hansen (2015)). In other work, Bernanke et al. (2005) propose a factor-augmented VAR (FAVAR) model to assist in making structural inferences while avoiding the problem of information sparsity that occurs in low dimensional VAR systems. Although the presence of common factors helps in capturing additional variation and co-variation in the data, there is still evidence to suggest that misspecification continues to play a role in applied work with DFMs, particularly in forecasting. Stock and Watson (2005, 2002), for instance, test the ability of cross variation in forecasting, namely whether observations on another variable such as x_{jt} help in predicting x_{it} given lagged values of x_{it} and common factors using 132 U.S. macroeconomic time series. Their results suggest that exclusion of other variables like x_{jt} from the regression equation for x_{it} involves misspecification that can impair forecasting performance. A systematic approach to dealing with potential misspecification of this type is to emply modern machine learning methods that rely on regularized estimation. The present paper seeks to do this in the context of large dimensional FAVAR systems.

Regularized estimation has received intense recent attention in both econometrics and statistics. In the cross-sectional framework, among the most influential works are Tibshirani (1996), Zhao and Yu (2006), Zhao and Yu (2006), Candes and Tao (2007) and Huang et al. (2008). Inspired by the methods developed in these papers a growing body of literature on high dimensional autoregressive models has emerged. Haufe et al. (2010) proposed a method based on Group LASSO to discover causal effects in multivariate time series. Basu and Michailidis (2015) studied deviation bounds for Gaussian processes and investigated ℓ_1 regularized estimation of transition matrices in sparse VAR models. Kock and Callot (2015) establish oracle inequalities for high dimensional VAR models. Han et al. (2015) proposed a generalized Dantzig selector in high dimensional VARs. Guo et al. (2016) studied a class of VAR models with banded coefficient matrices. These methods have opened up new avenues for handling high dimensional VAR models in practical work. In particular, regularized estimation has now been employed in various empirical applications in economic and financial analysis, among which we mention the following recent studies: Smeekes and Wijler (2018) studied forecasting capabilities of penalized regression in cases where the generating process is a factor model; Medeiros et al. (2019) considered inflation forecasting with machine learning methods; Uematsu and Tanaka (2019) examined high-dimensional forecasting and variable selection via folded-concave penalized regressions; and high dimensional VARs were adopted to estimate networks and construct measures of financial sector connectedness (see Barigozzi and Brownlees (2019); Barigozzi and Hallin (2017); Demirer et al. (2018)).

All these studies assume that the model's idiosyncratic errors have at most weak cross-sectional dependence (c.f., Chudik et al. 2011). However, the vast literature on the DFM indicates that this assumption is fragile in applications. In response to this limitation, the present paper proposes a new high dimensional VAR model in which some common factors (CFs) figure in the determination of each time series besides the idiosyncratic errors and lagged values of the time series themselves. In earlier work, Chudik and Pesaran (2011) considered a factor-augmented infinite dimensional VAR model. For simplicity, they construct a model for which the strong cross section dependence that is due to the factors is explicitly separated from other sources of cross section dependence. They mention the possibility of using high dimensional VAR models with CFs but do not explicitly analyze the model. The FAVAR system in the present paper additionally allows for serial correlation among the CFs, which in turn leads to correlation between the CFs and the lagged time series. To properly control for the presence of CFs in this FAVAR system it is necessary to estimate factors, factor loadings, and transition matrices simultaneously. Practical implementation also requires determination of the number of factors and lag length.

To estimate the high dimensional VAR model with CFs, our approach uses a three-step procedure. The first step employs ℓ_1-nuclear norm regularized estimation that minimizes the sum of squared residuals with an ℓ_1-norm penalty on the transition matrices and a nuclear norm penalty on the low rank matrix Θ representing the common component. Imposing the ℓ_1-norm penalty helps to estimate sparse transition matrices. The nuclear norm penalty helps to estimate the low rank matrix arising from the CFs and the factor loadings. Nuclear norm regularized estimation, which has appealing computational efficiency and good theoretical properties in estimating low rank matrices, has been recently studied by Chernozhukov et al. (2019) and Moon and Weidner (2019). Under some regularity conditions, we establish nonasymptotic bounds for the estimation error of the transition matrices and the low rank matrix Θ. Applying a singular value thresholding procedure on the singular values of the estimate of the matrix Θ, we obtain an estimate of the number of factors. We also show that the true number of factors can be estimated correctly with probability approaching one (w.p.a.1). Then, given the estimated factor number, preliminary estimates of the common factors can be obtained.

In the second step, we include the estimated CFs as regressors and consider a generalized LASSO to obtain an estimate of the transition matrices. We show that the estimation errors can be uniformly controlled, which facilitates the construction of weights for subsequent estimation by adaptive (or conservative) LASSO in the third step. Under some regularity conditions, we show that this third step conservative LASSO estimator of the transition matrices achieves sign consistency (see Zhao and Yu 2006) asymptotically. Besides, the third step estimator of transition matrices, factors and factor loadings are asymptotically equivalent to the corresponding oracle least squares estimators that are obtained by using detailed information about the form of the true regression model. We also study the asymptotic properties of these oracle least squares estimators and find that they perform as well as if the true common factors were known.

We illustrate the usefulness of this methodology through a real-data example. We revisit the financial connectedness measures proposed by Diebold and Yilmaz (2014) and document strong evidence of the existence of common factors in the volatilities of 23 sector exchange traded funds (ETFs). The findings show that common factors account for a large proportion of the variation in these volatilities; and, conditional on the common factors, a high level of connectedness remains present among the idiosyncratic components. This empirical application demonstrates the particular usefulness of our high dimensional VAR with CFs model in its ability to allow for time series with strong cross section dependence while distinguishing variations that originates from different sources.

The remainder of the paper is organized as follows. In Section 4.2, we introduce our model and conduct a stationarity analysis. Section 4.3 introduces the estimation methods and examines their theoretical properties. In Section 4.4, we conduct Monte Carlo experiments to evaluate finite sample performance of the methodology. We apply the model and methods to study financial connectedness in Section 4.5. Section 4.6 concludes. Proofs of the main results in the paper are given in the Appendix C. Further technical details are provided in the online Supplementary Materials.

4.1.1 Notation

To proceed, we introduce some notation. Let $A = (a_{ij}) \in \mathbb{R}^{M\times N}$ and $v = (v_1, ..., v_N)' \in \mathbb{R}^N$ be a matrix and a vector, respectively. We denote v_I as the subvector of v whose entries are indexed by a set $I \subset [N] \equiv \{1, ..., N\}$. We denote $A_{I,J}$ as the submatrix of A whose rows are indexed by I and columns are indexed by J. Let $A_{*,J} \equiv A_{[N],J}$ be the submatrix of A whose columns are indexed by J, $A_{I,*} \equiv A_{I,[M]}$ be the submatrix of A whose rows are indexed by I. For notational simplicity, we also write the individual columns and rows of A respectively as $A_{*,j} = A_{*,\{j\}}$ for $j = 1, ..., N$ and $A_{i,*} = A_{\{i\},*}$ for $i = 1, ..., M$.

For $0 < q < \infty$, we define the ℓ_0, ℓ_q, and ℓ_∞ norms of a vector v to be

$$|v|_0 \equiv \sum_{i=1}^N \mathbf{1}(v_i \neq 0),\ |v|_q \equiv \left(\sum_{i=1}^N |v_i|^q\right)^{1/q}, \text{ and } |v|_\infty \equiv \max_{1\leq i\leq N} |v_i|,$$

where $\mathbf{1}(\cdot)$ is the indicator function. In the special case $q = 2$, $|\cdot|_2$ is the Euclidean norm of v, and we write $|v| \equiv |v|_2$ for notational simplicity.

For $0 < q < \infty$, we define the ℓ_q, $\ell_{\max}$, Frobenius (F), and nuclear ($_*$) norms of the matrix A to be:

$$||A||_q \equiv \max_{||v||_q=1} ||Av||_q,\ ||A||_{\max} \equiv \max_{i,j} |a_{ij}|,$$

$$||A||_{\mathrm{F}} \equiv \left(\sum_{i,j} |a_{ij}|^2\right)^{1/2} \text{ and } ||A||_* = \sum_{k=1}^{\min(N,M)} \psi_k(A),$$

where $\psi_k(\cdot)$ is the kth largest singular value of A for $k = 1, ..., \min(N, M)$. We also denote the largest and smallest singular value of A as $\psi_{\max}(A)$ and $\psi_{\min}(A)$. In the special case $q = 2$, the ℓ_2 matrix norm is given by $||A||_2 = ||A||_{\mathrm{op}} \equiv \psi_1(A)$. For a full rank $N \times R$ matrix F with $N > R$, we denote the corresponding orthogonal projection matrices as $\mathbb{P}_F = F(F'F)^{-1}F'$ and $\mathbb{M}_F = I_N - \mathbb{P}_F$, where I_N denotes the $N\times N$ identity matrix. Let $\mathrm{vec}(\cdot)$ denote the (columnwise) vectorization operator, and $\otimes$ be the (right hand) Kronecker operator. For a random variable or vector x, we denote its expectation and

ℓ_p-norm as $E(x)$ and $|||x|||_p \equiv [E(|x|_p^p)]^{1/p}$. The operators $\vee$ and $\wedge$ denote max and min, viz., $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$.

4.2 Model

For a N-dimensional vector-valued time series $\{Y_t\} = \{(y_{1t}, ..., y_{Nt})'\}$, the high-dimensional vector autoregression model of lag p with CFs is given by:

$$Y_t = \sum_{j=1}^{p} A_j^0 Y_{t-j} + \Lambda^0 f_t^0 + u_t, \quad t = 1, ..., T, \tag{4.1}$$

where $A_1^0, ..., A_p^0$ are $N \times N$ transition matrices, $\Lambda^0 = (\lambda_1^0, ..., \lambda_N^0)'$ is the $N \times R^0$ factor loading matrix, f_t^0 is an R^0-dimensional vector of common factors, and u_t is an N-dimensional vector of unobserved idiosyncratic errors. Throughout this paper we use the superscript 0 to denote true values. The coefficients of interest are the A_j^0's, Λ^0, and $F^0 \equiv (f_1^0, ..., f_T^0)'$. In practice, we need to determine the number of factors and lag length. We propose a method to consistently determine p in Section 4.3. Given p, the number of factors can be determined in the first step of our estimation procedure introduced in Section 4.3. We consider the framework that both the number of cross-sectional units N and the time periods T go to infinity. The estimation is a natural high-dimensional problem with the number of parameters $(N^2p + R^0N + R^0T)$ growing linearly with T and quadratically with N.

It is convenient to reformulate model (4.1) as a multivariate regression problem in the form

$$\underbrace{\begin{bmatrix} Y_1' \\ \vdots \\ Y_T' \end{bmatrix}}_{\mathbf{Y}} = \underbrace{\begin{bmatrix} Y_0' & \cdots & Y_{1-p}' \\ \vdots & \ddots & \vdots \\ Y_{T-1}' & \cdots & Y_{T-p}' \end{bmatrix}}_{\mathbf{X}} \underbrace{\begin{bmatrix} A_1^{0\prime} \\ \vdots \\ A_p^{0\prime} \end{bmatrix}}_{B^0} + \underbrace{\begin{bmatrix} f_1^{0\prime} \\ \vdots \\ f_T^{0\prime} \end{bmatrix}}_{F^0} \underbrace{\begin{bmatrix} \lambda_1^{0\prime} \\ \vdots \\ \lambda_N^{0\prime} \end{bmatrix}'}_{\Lambda^{0\prime}} + \underbrace{\begin{bmatrix} u_1' \\ \vdots \\ u_T' \end{bmatrix}}_{\mathbf{U}}, \tag{4.2}$$

where $\mathbf{Y} \in \mathbb{R}^{T \times N}$, $\mathbf{X} \in \mathbb{R}^{T \times Np}$, $B^0 \in \mathbb{R}^{Np \times N}$, and $\mathbf{U} \in \mathbb{R}^{T \times N}$. A key observation here is that $\Theta^0 \equiv F^0 \Lambda^{0\prime}$ is a low rank matrix. However, due to the correlation between $\mathbf{X}B^0$ and Θ^0, use of principal component analysis (PCA) on $\mathbf{Y}$ cannot deliver a consistent estimate of the common factors. Note that both $||\mathbf{X}B^0||_{\text{op}}$ and $||\Theta^0||_{\text{op}}$ are $O_P(\sqrt{NT})$ under some regularity conditions, and $||\mathbf{U}||_{\text{op}} = O_P(\sqrt{N} + \sqrt{T})$. We cannot separate the low rank matrix Θ^0 from $\mathbf{Y}$ without information about B^0. Besides, when the common factors are themselves serially correlated, pure VAR(p) estimation generally suffers from endogeneity bias issues.

4.2.1 Stationarity analysis

Let $X_t \equiv \mathbf{X}'_{t,*}$. The N-dimensional VAR(p) process $\{Y_t\}$ can be rewritten in companion form as an Np-dimensional VAR(1) process with CFs, viz.,

$$\underbrace{\begin{bmatrix} Y_t \\ Y_{t-1} \\ \vdots \\ Y_{t-p+1} \end{bmatrix}}_{X_{t+1}} = \underbrace{\begin{bmatrix} A_1^0 & A_2^0 & \cdots & A_{p-1}^0 & A_p^0 \\ I_N & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I_N & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & I_N & \mathbf{0} \end{bmatrix}}_{\Phi} \underbrace{\begin{bmatrix} Y_{t-1} \\ Y_{t-2} \\ \vdots \\ Y_{t-p} \end{bmatrix}}_{X_t} + \underbrace{\begin{bmatrix} \Lambda^0 f_t^0 \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}}_{\mathcal{F}_t} + \underbrace{\begin{bmatrix} u_t \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}}_{\mathcal{U}_t}. \tag{4.3}$$

If one treats $\mathcal{F}_t + \mathcal{U}_t$ as an impulse at period t, the process $\{X_{t+1}\}$ in (4.3) can be regarded as a high-dimensional VAR(1) process. We can write the reverse characteristic polynomial (Lütkepohl 2005) of Y_t as

$$\mathcal{A}(z) \equiv I_N - \sum_{j=1}^{p} A_j^0 z^p.$$

In the low-dimensional framework, the process is stationary if $\mathcal{A}(z)$ has no roots in and on the complex unit circle, or equivalently the largest modulus of eigenvalues of Φ is in unit circle. To achieve identification, we need to study the Gram or signal matrix $S_X \equiv \mathbf{X}'\mathbf{X}/T$ and $\Sigma_X = E(X_t X_t')$ in the later analysis. Basu and Michailidis (2015; hereafter BM) study the deviation bounds for the Gram matrix, using Gaussianity assumptions and boundedness of the spectral density function. Following their lead, we impose some conditions that will ensure S_X is well behaved.

To proceed, we write X_{t+1} as a moving average process of infinite order (MA(∞)):

$$\begin{aligned} X_{t+1} &= \sum_{j=0}^{\infty} \Phi^j (\mathcal{F}_{t-j} + \mathcal{U}_{t-j}) \equiv X_{t+1}^{(f)} + X_{t+1}^{(u)}, \text{ where} \qquad (4.4) \\ X_{t+1}^{(f)} &\equiv \sum_{j=0}^{\infty} \Phi^j \mathcal{F}_{t-j}, \text{ and } X_{t+1}^{(u)} \equiv \sum_{j=0}^{\infty} \Phi^j \mathcal{U}_{t-j}. \end{aligned}$$

Then we can study the stationarity of Y_t by studying $X_{t+1}^{(f)}$ and $X_{t+1}^{(u)}$, respectively. First, we consider $X_{t+1}^{(f)}$, which is the component due to the common factors. Note that the covariance matrix of $\mathcal{F}_t$ is a high-dimensional matrix with rank R^0 and explosive nonzero eigenvalues. Even if the largest modulus of eigenvalues of Φ is smaller than 1, the variances of entries of $X_{t+1}^{(f)}$ are not ensured to be uniformly bounded. Specifically, we consider $y_{it}^{(f)}$, which is the ith entry of $X_{t+1}^{(f)}$. Let $e_{j,M}$ be the jth unit M-dimensional vector. Noting that $y_{it}^{(f)} = (e_{1,p} \otimes e_{i,N})' X_{t+1}^{(f)}$, we can write y_{it} as an MA(∞) process

$$y_{it}^{(f)} = \sum_{j=0}^{\infty} (e_{1,p} \otimes e_{i,N})' \Phi^j (e_{1,p} \otimes \Lambda^0) f_{t-j}^0 \equiv \sum_{j=0}^{\infty} \alpha_{iN}^{(f)}(j) f_{t-j}^0,$$

where f_t^0 can be serially correlated. To ensure $y_{it}^{(f)} = O_P(1)$, we need to require the coefficients $\alpha_{iN}^{(f)}(j)$ to be $O(1)$ and summable. Note that we generally do not have $||\Phi||_{\text{op}} \leq 1$, as explained in the supplement of BM (2015). In assumption A.1, we impose sufficient conditions that ensure the $\alpha_{iN}^{(f)}(j)$ are well-behaved. The online supplementary material provides a discussion of these conditions.

For the process $X_{t+1}^{(u)}$, stationarity is assured if we assume the covariance matrix of u_t is well-behaved and u_t is serially uncorrelated as in BM (2015) and KC (2015). Similar to $y_{it}^{(f)}$, we define $y_{it}^{(u)}$ such that

$$y_{it} \equiv y_{it}^{(f)} + y_{it}^{(u)}, \qquad (4.5)$$

where

$$y_{it}^{(u)} = \sum_{j=0}^{\infty} \alpha_{iN}^{(u)}(j) u_{t-j} \text{ and } \alpha_{iN}^{(u)}(j) \equiv (e_{1,p} \otimes e_{i,N})' \Phi^j (e_{1,p} \otimes I_N).$$

Again, imposing zero serial correlation and weak cross-sectional correlation across u_{it}'s is not enough to ensure $y_{it}^{(u)} = O_P(1)$ uniformly.

Let $\underline{c}$ and $\bar{c}$ denote generic constants that may vary across occurrences. Throughout the paper, we will treat Λ^0 as nonrandom. To ensure the stationarity of $\{Y_t\}$, we impose the following assumption.

Assumption A.1. (i) $u_t = C^{(u)} \epsilon_t^{(u)}$, *where* $\epsilon_t^{(u)} = (\epsilon_{1,t}^{(u)}, ..., \epsilon_{m,t}^{(u)})'$, $\epsilon_{i,t}^{(u)}$*'s are i.i.d. random variables across* (i, t) *with mean zero and variance 1, and* $C^{(u)}$ *is an* $N \times m$ *matrix such that* $C^{(u)} C^{(u)\prime} = \Sigma_u$ *and* $\underline{c} \le \psi_{\min}(\Sigma_u) \le \psi_{\max}(\Sigma_u) \le \bar{c}$*;*

(ii) $\{f_t^0\}$ *follows a strictly stationary linear process:*

$$f_t^0 - \mu_f = \sum_{j=0}^{\infty} C_j^{(f)} \epsilon_{t-j}^{(f)},$$

where $\epsilon_t^{(f)} \equiv (\epsilon_{1,t}^{(f)}, ..., \epsilon_{R^0,t}^{(f)})'$ *are i.i.d.* $(0, I_{R^0})$ *across* t, *sup*$_{m\ge1}(m+1)^{\alpha} \sum_{j=m}^{\infty} ||C_j^{(f)}||_{\max} \le \bar{c} < \infty$ *for some constant* $\alpha > 1$*;*

(iii) $\max_{1\le r\le R^0} |||\epsilon_{r,t}^{(f)}|||_q < \bar{c}$ *and* $\max_{1\le i\le m} |||\epsilon_{i,t}^{(u)}|||_q < \bar{c}$ *for some* $q > 4$;

(iv) $\{\epsilon_t^{(u)}\}$ *is independent of* $\{\epsilon_t^{(f)}\}$;

(v) *the largest modulus of the eigenvalues of* Φ *is bounded by some constant* $0 < \rho < 1$;

(vi) $||(\Phi^j)_{[N],[N]}||_{\text{op}} \le \bar{c}\rho^j$ *and* $|\alpha_{iN}^{(f)}(j)| < \bar{c}\rho^j$;

(vii) $\max_{|z|=1} \Psi_{\max}(\mathcal{A}^*(z)\mathcal{A}(z)) \le \bar{c}$, *where* $|z|$ *denotes the modulus of* z *in the complex plane, and* $\mathcal{A}^*(z)$ *denotes the conjugate transpose of* $\mathcal{A}(z)$.

Assumption A.1(i) is frequently made in high dimensional time series analysis; see, e.g., Bai and Saranadasa (1996), Chen and Qin (2010) and Ma et al. (2020). It requires that u_t be independent over t and weakly dependent across i. At the cost of more complicated notations, one can allow $\psi_{\min}(\Sigma_u)$ to converge to zero and $\psi_{\max}(\Sigma_u)$ to diverge to infinity, both at a slow rate. Assumption A.1(ii) assumes the common factors to be stationary and allows for weak serial correlation. The factors can have nonzero mean so that y_{it}'s can also have nonzero mean. Assumption A.1(iii) requires that $\epsilon_{i,t}^{(u)}$ and $\epsilon_{i,t}^{(f)}$ have finite qth order moments, which is a weak assumption compared to the Gaussianity distribution assumption of BM (2015) and KC (2015). Assumption A.1(iv) requires independence between $\{\epsilon_t^{(u)}\}$ and $\{\epsilon_t^{(f)}\}$, which facilitates separate study of $y_{it}^{(f)}$ and $y_{it}^{(u)}$. Assumption A.1(v) is a standard assumption to ensure stationarity. Assumption A.1(vi) is a high level condition to ensure that $E(y_{it}^2)$ is uniformly bounded. Assumption A.1 (vii) helps to bound the minimum eigenvalue of Σ_X. By the inequalities

$$\max_{|z|=1} \Lambda_{\max}(\mathcal{A}^*(z)\mathcal{A}(z)) \leq (\max_{|z|=1} ||\mathcal{A}(z)||_{\text{op}})^2 \leq 1 + \sum_{k=1}^{p} ||A_j^0||_{\text{op}},$$

we can see that requiring all the A_j^0's to have finite operator norms is a sufficient condition.

The online Supplementary Material provides further discussion of the Assumption A.1(vi)-(vii). The following proposition ensures stationarity of the y_{it} and establishes a lower bound for $\psi_{\min}(\Sigma_X)$.

Proposition 4.1. *Suppose that Assumption A.1 holds. (i) Then Y_t is a stationary process, $\sup_i E(y_{it}^2) < \infty$, and*

$$\psi_{\min}(\Sigma_X) \geq \frac{\psi_{\min}(\Sigma_u)}{\max_{|z|=1} \psi_{\max}(\mathcal{A}^*(z)\mathcal{A}(z))}.$$

(ii) Let $\Sigma_{XF} \equiv E(X_t f_t^{0\prime})$, and $\Sigma \equiv \Sigma_X - \Sigma_{XF}\Sigma_F^{-1}\Sigma_{XF}'$. We also have $\psi_{\min}(\Sigma) \geq \frac{\psi_{\min}(\Sigma_u)}{\max_{|z|=1} \psi_{\max}(\mathcal{A}^*(z)\mathcal{A}(z))}$.

4.3 Estimation method and theoretical results

This section develops an estimation procedure for the model and establishes some its properties, both asymptotic and non-asymptotic. The procedure assumes at this point that the lag length p is known and that R^0 is unknown. Lag length is actually data-determined in the manner explained later in Section 4.3.4. The number of factors can be determined consistently in the first estimation step.

4.3.1 First-step estimator

In the first step, we propose an ℓ_1-nuclear norm regularized estimation procedure to estimate the coefficient B^0 and the low rank matrix Θ^0 simultaneously. Moon and Weidner (2018; hereafter MW) and Chernozhukov et al. (2019) show that nuclear norm regularized estimation can achieve consistent estimation of the low rank matrix. In our model, we impose a sparsity condition on B^0 and use ℓ_1-norm regularization to achieve regressor selection. For Θ^0, the nuclear norm regularization helps to achieve consistent estimation. The first step estimator is given by the following procedure.

The first-step estimator: *Let* $\gamma_1 = \gamma_1(N,T) \equiv c_1 T^{-1/2} logN$ *and* $\gamma_2 = \gamma_2(N,T) \equiv c_2(N^{-1/2} + T^{-1/2})$ *for some constants* c_1 *and* c_2.

1. *Estimate the coefficient* B *and the low rank matrix* Θ *by running the following* ℓ_1*-nuclear norm regularized regression:*

$$\begin{aligned}(\tilde{B},\tilde{\Theta}) &= \arg\min{}_{(B,\Theta)}\mathcal{L}(B,\Theta), \text{ where}\\ \mathcal{L}(B,\Theta) &\equiv \frac{1}{2NT}||\mathbf{Y}-\mathbf{X}B-\Theta||_{\mathrm{F}}^2 + \frac{\gamma_1}{N}|\mathrm{vec}(B)|_1 + \frac{\gamma_2}{\sqrt{NT}}||\Theta||_*. \quad (4.6)\end{aligned}$$

2. *Estimate the number of factors* R^0 *by singular value thresholding (SVT) as:*

$$\hat{R} = \sum_{i=1}^{N\wedge T} \mathbf{1}\{\psi_i(\tilde{\Theta}) \geq (\gamma_2\sqrt{NT}||\tilde{\Theta}||_{\mathrm{op}})^{1/2}\}.$$

3. *Obtain a preliminary estimate of* F^0. *Let the singular value decomposition (SVD) of* $\tilde{\Theta}$ *be* $\tilde{\Theta} = \tilde{U}\tilde{D}\tilde{V}'$, *where* $\tilde{D}$ *=diag*$(\psi_1(\tilde{\Theta}),...,\psi_{N\wedge T}(\tilde{\Theta}))$. *Let* $\tilde{F} = \sqrt{T}\tilde{U}_{*,[\hat{R}]}$.

Remark 4.1 The objective function $\mathcal{L}(B,\Theta)$ minimizes the sum of squared residuals with both the nuclear norm regularization on Θ and ℓ_1-regularization on B. To obtain the numerical solution, we can apply an EM type algorithm. In the E-step, we fix B and update the estimate of Θ. The solution can be obtained following the result of Lemma 1 of MW (2018).[2] In the M-step, we fix Θ and update B. The optimization problem can be decomposed to N LASSO-type linear regression problems.

Non-asymptotic results for the first-step estimator

In this subsubsection we establish non-asymptotic properties of the first step estimator. In particular, for $\tilde{B}$ and $\tilde{\Theta}$, we establish a non-asymptotic inequality for the estimation error. For $\hat{R}$, we show that $\hat{R} = R^0$ w.p.a.1.

To proceed, we introduce some notation and assumptions. We first introduce a key invertibility condition for the operator $(\Delta^{(1)}, \Delta^{(2)}) :\rightarrow \mathbf{X}\Delta^{(1)} + \Delta^{(2)}$ when $(\Delta^{(1)}, \Delta^{(2)})$ is restricted to lie in a 'cone'. A similar condition is imposed in MW (2018) and Chernozhukov et al. (2018). Following their lead, we refer to the condition as '*restricted strong convexity*'. To define the 'cone', let $J_i \subset [Np]$ be an index set such that $j \in J_i$ if and only if $B^0_{ji} \neq 0$. Let $J_i^c = [Np] \backslash J_i$. For a $T \times N$ matrix $\Delta^{(2)}$, define the operators

$$\mathcal{P}(\Delta^{(2)}) \equiv U_{*,[R^0]} U_{*,[R^0]}{}' \Delta^{(2)} V_{*,[R^0]} V'_{*,[R^0]} \text{ and } \mathcal{M}(\Delta^{(2)}) \equiv \Delta^{(2)} - \mathcal{P}(\Delta^{(2)}).$$

Hence, the operator $\mathcal{P}(\cdot)$ projects a matrix onto a 'low-rank' space which contains Θ^0. For some $c > 0$, the 'cone' $\mathcal{C}_{NT}(c) \subset \mathbb{R}^{Np \times N} \times \mathbb{R}^{T \times N}$ is a set of $(\Delta^{(1)}, \Delta^{(2)})$ satisfying the restriction:

$$\frac{\gamma_1 \sum_{i=1}^N |\Delta^{(1)}_{J_i^c,i}|_1}{N} + \frac{\gamma_2 \left\|\mathcal{M}(\Delta^{(2)})\right\|_*}{\sqrt{NT}} \leq c\frac{\gamma_1 \sum_{i=1}^N |\Delta^{(1)}_{J_i,i}|_1}{N} + c\frac{\gamma_2 \left\|\mathcal{P}(\Delta^{(2)})\right\|_*}{\sqrt{NT}}.$$

We impose the following condition.

Assumption A.2 (Restricted strong convexity) *If* $(\Delta^{(1)}, \Delta^{(2)}) \in \mathcal{C}_{NT}(c)$ *for some* $c > 0$*, then there exist constants* κ_c *and* κ'_c *such that*

$$\left\|\mathbf{X}\Delta^{(1)} + \Delta^{(2)}\right\|_{\mathrm{F}}^2 \geq T \cdot \kappa'_c \left\|\Delta^{(1)}\right\|_{\mathrm{F}}^2 + \kappa_c \left\|\Delta^{(2)}\right\|_{\mathrm{F}}^2.$$

The next assumption involves a regularity condition on the errors and a sparsity

[2] Let the SVD of A be $A = USV'$, where $S = diag(s_1, ..., s_q)$, with $q = rank(A)$. Then $\arg\min_\Theta \left(\frac{1}{2}||A - \Theta||_{\mathrm{F}}^2 + \gamma||\Theta||_*\right)$ is given by $U \cdot diag((s_1 - \gamma)_+, ..., (s_q - \gamma)_+) \cdot V'$, where $(s)_+ = \max(0, s)$.

condition on the transition matrix:

Assumption A.3 (i) $\|\mathbf{U}\|_{\text{op}}/\sqrt{NT} \leq \gamma_2/2$, *where* γ_2 *is the tuning parameter for nuclear norm regularization;*

(ii) *each column of* B^0 *contains at most* K_J *nonzero entries.*

Assumption A.3 (i) requires the idiosyncratic error matrix to have operator norm of order $O_P(\sqrt{N}+\sqrt{T})$. This condition is also assumed in MW (2018) and Chernozhukov et al. (2018). It holds w.p.a.1 if the $\epsilon_{it}^{(u)}$'s are i.i.d. sub-Gaussian (see, e.g., Vershynin 2018). Assumption A.3(ii) is a sparsity assumption. We allow K_J goes to infinity at a slow rate. This sparsity condition can be relaxed to the approximate sparsity condition as in Belloni et al. (2012) but that extension is not pursued here.

Theorem 4.1. *Suppose that Assumptions A.1-A.3 hold. Then we have* $N^{-1/2}\left\|\tilde{B}-B^0\right\|_F \leq \bar{c}(\gamma_1\sqrt{K_J}\vee\gamma_2)$ *and* $(NT)^{-1/2}\left\|\tilde{\Theta}-\Theta^0\right\|_F \leq \bar{c}(\gamma_1\sqrt{K_J}\vee\gamma_2)$, *with probability at least* $1-\bar{c}'(N^2T^{1-q/4}(logN)^{-q/2}+N^{2-\underline{c}logN})$ *for some finite positive constants* $\underline{c}$, $\bar{c}$, *and* $\bar{c}'$.

Theorem 4.1 establishes a non-asymptotic inequality for the estimation errors of $\tilde{B}$ and $\tilde{\Theta}$. The inequality is valid when both $N^2T^{1-q/4}(\log N)^{-q/2}$ and $N^{2-\underline{c}\log N}$ are small. In general, the first term dominates the second one for finite q and divergent N and T. If the error terms are sub-exponential, we can allow q to diverge to infinity in which case the second term could dominate the first one. To prove the above theorem, we need to establish a bound for $T^{-1}||\mathbf{U}'\mathbf{X}||_{\max}$. Specifically, we need to find a sharp probability bound for a partial sum like $T^{-1}\sum_{t=1}^{N} y_{i,t-k}u_{jt}$. We resort to a Nagaev-type inequality, as introduced by Wu (2005) and Wu and Wu (2016), allowing for both dependence among summands and non-Gaussianity. The summand $y_{i,t-k}u_{jt}$ has a nonlinear Wold presentation $y_{i,t-k}u_{jt} = g_{ijk}(\ldots,\epsilon_{t-1},\epsilon_t)$, where $\epsilon_t \equiv (\epsilon_t^{(u)\prime},\epsilon_t^{(f)\prime})'$ is i.i.d. random variables under Assumption A.1. Then one can verify that the *dependence adjusted norm* (see Wu and Wu, 2016) of $y_{i,t-k}u_{jt}$ is well bounded so that one can obtain a sharp probability bound using the Nagaev-type inequality for nonlinear processes.

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

Assumption A.4 (i) *There exists an $\bar{N}$ such that for all $N > \bar{N}$, $||\Lambda^{0\prime}\Lambda^0/N - \Sigma_\Lambda||_{\max} \leq \bar{c}N^{-1/2}$ for an $R^0 \times R^0$ matrix Σ_Λ and $||\Lambda^0||_{\max} \leq \bar{c}$;*

(ii) *Let $\Sigma_F = E(f_t^0 f_t^{0\prime})$, there are constants $s_1 > \cdots > s_{R^0} > 0$ so that s_j equals the jth largest eigenvalue of $\Sigma_F^{1/2}\Sigma_\Lambda\Sigma_F^{1/2}$.*

Assumption A.4 requires that the factors and the factor loadings are strong/ pervasive with well-behaved sample second moments. Assumption (ii) requires distinct eigenvalues of $\Sigma_F^{1/2}\Sigma_\Lambda\Sigma_F^{1/2}$ in order to identify the corresponding eigenvectors.

The next theorem establishes the consistency of $\hat{R}$ and the mean-square convergence rate of $\tilde{F}$:

Theorem 4.2. *Suppose Assumptions A.1-A.4 hold. There exist positive constants $\underline{c}$, $\bar{c}$ and $\bar{c}'$, and a random matrix $\tilde{H}$ depending on (F^0, Λ^0) such that (i) $\hat{R} = R^0$ and (ii) $||\tilde{F} - F^0\tilde{H}||_F/\sqrt{T} \leq \bar{c}(\gamma_1\sqrt{K_J} \vee \gamma_2)$, both with probability larger than $1 - \bar{c}'(N^2T^{1-q/4}(logN)^{-q/2} + N^{2-\underline{c}logN})$.*

Theorem 4.2 establishes the consistency of $\hat{R}$ and the mean-square convergence rate of $\tilde{F}$. Intuitively, since $\tilde{\Theta}$ is a consistent estimator for $\Theta^0 \equiv F^0\Lambda^{0\prime}$ with well-controlled estimation errors, we expect the first R^0 singular values of $\tilde{\Theta}$ to be $O_P(\sqrt{NT})$ and the other singular values to be $O_P[\sqrt{NT}(\gamma_1 \vee \gamma_2)]$. Then the hard SVT procedure can distinguish the $\sqrt{NT}$-order singular values from those of smaller order. Alternatively, given the consistency of $\tilde{B}$ established in Theorem 4.1, we can regard the 'residual' $\mathbf{Y} - \mathbf{X}\tilde{B}$ as a approximation of $F^0\Lambda^{0\prime} + \mathbf{U}$. It is reasonable to conjecture that one can also apply the methods of Bai and Ng (2002), Onatski (2010) and Ahn and Horenstein (2013) to determine the number of factors. Theorem 4.2 (ii) establishes the convergence rate of $\tilde{F}$. The $R \times R$ transformation matrix $\tilde{H}$ is similar to the matrix H in

Bai (2003).

4.3.2 Second-step estimator

In this subsection, we introduce the second-step estimator. The second-step estimator is a generalization of LASSO estimator, which includes the estimated factor matrix $\tilde{F}$ as regressors. Our goal is to obatain an estimator which uniformly converges to the true parameter. Then the second-step estimator can be utilized to construct adaptive- or conservative- LASSO weights.

The second-step estimator: *Let* $\gamma_3 = c_3(\gamma_1\sqrt{K_J} \vee \gamma_2)$ *for some constant* c_3*. For each* $i = 1, ..., N$*, solve the minimization problem:*

$$(\dot{B}'_{*,i}, \dot{\lambda}'_j)' = \text{argmin}_{(v',\lambda')' \in \mathbb{R}^{Np+R^0}} \frac{1}{2T}||\mathbf{Y}_{*,i} - \mathbf{X}v - \tilde{F}\lambda||_F^2 + \gamma_3|v|_1, \qquad (4.7)$$

where the LASSO penalty is only imposed on coefficients of regressors X*. Then the second-step estimators of* B^0 *and* Λ^0 *are given by* $\dot{B} = (\dot{B}_{*,1}, ..., \dot{B}_{*,N})$ *and* $\dot{\Lambda} = (\dot{\lambda}_1, ..., \dot{\lambda}_N)'$*, respectively.*

Remark 4.2 In the proof of Theorem 4.3, we show that $\dot{B}_{*,i}$ solves the LASSO problem with dependent variable $\mathbb{M}_{\tilde{F}}\mathbf{Y}_{*,i}$ and regressors $\mathbb{M}_{\tilde{F}}\mathbf{X}$.

Below, we bound the convergence rate of the entries of $\dot{B}$ uniformly. Then the estimate can be used to construct the weights for the adaptive LASSO estimator in the third step.

Non-asymptotic results for the second step estimator

Recall $\Sigma \equiv \Sigma_X - \Sigma_{XF}\Sigma_F^{-1}\Sigma'_{XF}$ and let $\tilde{\Sigma} = \mathbf{X}'\mathbb{M}_{\tilde{F}}\mathbf{X}/T$. By Proposition 4.1, $\psi_{\min}(\Sigma)$ is bounded below by some constant. Hence, it is straightforward to see that

$$\min_{|v|\neq 0} \frac{v'\Sigma v}{|v|^2} \geq \psi_{\min}(\Sigma) > 0.$$

However, the matrix $\tilde{\Sigma}$ cannot be ensured to be positive definite. If $Np > T$, $\tilde{\Sigma}$ is singular, which leads to $\min_{|v|\neq 0} \frac{v'\tilde{\Sigma}v}{|v|^2} = 0$. In this case, we follow Bickel et al. (2009) and Kock and Callot (2015) to establish the **restricted eigenvalue**

condition. Specifically, we replace the above minimum by another minimum over a smaller set. Let $J \subset [Np]$ be an index set and $J^c = [Np] \backslash J$. We say the restricted eigenvalue condition is satisfied for some $1 \leq K \leq Np$ if

$$\min_{|J| \leq K} \min_{\substack{|v| \neq 0 \\ |v_{J^c}|_1 \leq 3|v_J|_1}} \frac{v' \tilde{\Sigma} v}{|v_J|^2} \equiv \kappa_{\tilde{\Sigma}}(K) > 0, \tag{4.8}$$

where $|J|$ denotes the cardinality of J. In (4.8), the minimum is restricted to those vectors that $||v_{J^c}||_1 \leq 3||v_J||_1$, where J has cardinality below K. In this restricted space, we can show that (4.8) is satistied with a high probability for $K = K_J$.

The following theorem establishes the $\ell_{\max}$-norm bound for the estimation error of $\dot{B}$.

Theorem 4.3. *Suppose that Assumptions A.1-A.4 hold. Then we have*

$$||\dot{B} - B^0||_{max} \leq \max_{1 \leq i \leq N} |\dot{B}_{*,i} - B^0_{*,i}|_1 \leq \frac{48}{[\psi_{\min}(\Sigma_X)]^2} K_J \gamma_3,$$

with probability larger than $1 - \bar{c}(N^2 T^{1-q/4}(logN)^{-q/2} + N^{2-\underline{c}logN})$, *for some finite positive constants* $\underline{c}$, *amd* $\bar{c}$.

4.3.3 Third-step estimator

In the first and second step, we impose penalty on every parameter, which introduces the asymptotic bias into the estimators of transition matrices. Zou (2006) proposed the adaptive LASSO technique in a linear regression framework, which penalizes the true zero parameters more than the non-zero ones. Then he shows that the adaptive LASSO estimator is asymptotically equivalent to the oracle least-squares estimator, which is obtained with information of relevant regressors. Kock and Callot (2015) also explore the adaptive LASSO method in the high-dimensional VAR framework.

In practice, the regressors with zero estimates in the preliminary stage, which are usually plain LASSO estimates, are excluded in adaptive LASSO. Hence, any incorrect regressor exclusion by the preliminary stage estimates directly leads to wrong regressor selection of adaptive LASSO. To solve this problem. the conservative LASSO, which gives regressors that are excluded by the initial estimator a second chance, is introduced (Caner and Kock 2018). In this subsection, we extend the conservative LASSO estimator to our high dimensional VAR model with CFs framework.

The third-step estimator (Conservative LASSO): *Conduct the following procedure:*

1. *Let* $\gamma_4 = \gamma_4(N,T)$ and $\hat{F}^{(0)} = \tilde{F}$. *Let* W *be a* $Np \times N$ *matrix with entries*
$$w_{ki} = \begin{cases} 1 & \text{if } |\dot{B}_{ki}| < \alpha\gamma_4, \\ 0 & \text{if } |\dot{B}_{ki}| \geq \alpha\gamma_4, \end{cases} \tag{4.9}$$
where $i = 1, ..., Np$, $i = 1, ..., N$, and $\alpha > 0$.

2. *For integer* $\ell \geq 1$, *update the estimate of* B and Λ:
$$(\hat{B}^{(\ell)\prime}_{*,i}, \hat{\lambda}^{(\ell)\prime}_i)' = \text{argmin}_{(v,\lambda)' \in \mathbb{R}^{NP+\hat{R}}} \frac{1}{2T} \left\| \mathbf{Y}_{*,i} - \mathbf{X}v - \hat{F}^{(\ell-1)}\lambda \right\|_F^2 + \gamma_4 \sum_{k=1}^{pN} w_{ki} |v_k|,$$
where v_k *is the kth entry of* v, $i = 1, ..., N$. *Let* $\hat{B}^{(\ell)} \equiv (\hat{B}^{(\ell)}_{*,1}, ..., \hat{B}^{(\ell)}_{*,N})$.

3. *Obtain the SVD of* $\mathbf{Y} - \mathbf{X}\hat{B}^{(\ell)}$ *as* $\mathbf{Y} - \mathbf{X}\hat{B}^{(\ell)} = \hat{U}^{(\ell)}\hat{D}^{(\ell)}\hat{V}^{(\ell)\prime}$. *Obtain an updated estimate of* F^0 *as* $\hat{F}^{(\ell)} = \sqrt{T}\hat{U}^{(\ell)}_{*,[\hat{R}]}$.

4. *Iterate steps 2-3 until numerical convergence. Denote the final estimators as* $\hat{B}$, $\hat{F}$ and $\hat{\Lambda}$.

Remark 4.3 The weights w_{ki}'s can take various forms. For example, Caner and Kock (2018) also consider $w_{ki} \equiv \frac{\gamma_{\text{prec}}}{|\dot{B}_{ki}| \vee \gamma_{\text{prec}}}$, where $\gamma_{\text{prec}} = O(\gamma_4)$.

Asymptotic properties of the third-step estimator

We establish two results: (i) the conservative LASSO estimator $\hat{B}^{(\ell)}$ has the variable-selection consistency w.p.a.1; (ii) $\hat{B}$ is asymptotically equivalent to the oracle least squares estimator.

First, we introduce some notations. Following Zhao and Yu (2006) and Huang et al. (2008) , we say that $\hat{B}^{(\ell)} =_s B^0$, or $\hat{B}^{(\ell)}$ is sign-consistent for B^0, if and only if

$$\text{sgn}(\hat{B}^{(\ell)}_{*,i}) = \text{sgn}(B^0_{*,i}), \text{ for all } i \in [N], \text{ where } \text{sgn}(B_{*,i}) \equiv [\text{sgn}(B_{1,i}), ..., \text{sgn}(B_{Np,i})]',$$

and

$$\text{sgn}(B_{ki}) \equiv \begin{cases} 1, & \text{if } B_{ki} > 0; \\ 0, & \text{if } B_{ki} = 0; \\ -1, & \text{if } B_{ki} < 0. \end{cases}$$

Assumption A.5 (i) *The magnitude of nonzero coefficients has appropriate order:* $\gamma_4 = o(\min_{i\in[N]} \min_{k\in J_i} |B^0_{ki}|)$;

(ii) $\sum_{i=1}^N |J_i|/N \leq C$ for some constant $0 < C < \infty$ as $N \to \infty$.

Assumption A.5 (i) assumes the nonzero entries of B^0 are uniformly bounded away from zero. This is a standard assumption in the adaptive LASSO literature. The lower bound $\min_{i\in[N]} \min_{k\in J_i} |B^0_{ki}|$ is allowed to tend to zero at a slow rate. As N increases, the 'average magnitude' of nonzero coefficients often decreases to ensure stationarity. By Theorem 4.3, and Assumption A.5 (i), there are constants $\underline{c}$ and $\bar{c}$ such that

$$\max_{k\in J_i} w_{ki} = 0 \text{ and } \min_{k\in J_i^c} w_{ki} = 1$$

w.p.a.1. In this case, we only put penalty on zero entries. Assumption A.5(ii) assumes that the number of nonzero coefficients is proportional to N. This assumption ensures that $||\mathbf{X}(\hat{B}^{(\ell)} - B^0)||_\text{F}$ has desird convergence rate.

The following theorem establishes the variable selection consistency of $\hat{B}^{(\ell)}$ and a preliminary convergence rate of $\hat{B}^{(\ell)}$ and $\hat{F}^{(\ell)}$.

Theorem 4.4. *Suppose that Assumptions A.1-A.5 hold,* $(K_J^{3/2}T^{-1/2}logN + K_J^{1/2}N^{-1/2}) = o(\gamma_4)$, *and* $N^2T^{1-q/4}(logN)^{-q/2} \to 0$, *as* $(N,T) \to \infty$. *Then*

(i) $P(\hat{B}^{(\ell)} =_s B^0) \to 1$, *as* $(N,T) \to \infty$;

(ii) $||\mathbf{X}(\hat{B}^{(\ell)} - B^0)||_F/\sqrt{NT} = O_P(\gamma_1\sqrt{K_J} + \gamma_2)$;

(iii) $||\hat{F}^{(\ell)} - F^0\tilde{H}||_F/\sqrt{T} = O_P(\gamma_1\sqrt{K_J} + \gamma_2)$.

Theorem 4.4 shows that $\hat{B}^{(\ell)}$ has the oracle property in that it selects the correct variables w.p.a.1. Due to the presence of common factors, we can only obtain a preliminary rate $O_P(\gamma_1\sqrt{K_J}+\gamma_2)$. To improve the rate of convergence, we study the final estimators $\hat{B}$, $\hat{F}$ and $\hat{\Lambda}$. Now, $\hat{F}$ corresponds to the first $\hat{R}$ eigenvectors of $(\mathbf{Y}-\mathbf{X}\hat{B})(\mathbf{Y}-\mathbf{X}\hat{B})'$, scaled by $\sqrt{T}$, and one can expand $\hat{F}-F^0\tilde{H}$ following the lead of Bai and Ng (2002) and Bai (2009). By looking at the product of $\hat{F}-F^0\tilde{H}$ and other terms, we can bound the smaller order terms with sharper bounds. Hence, we can finally improve the probability order of each element in $\hat{B}_{J_i,i}-B^0_{J_i,i}$ to $1/\sqrt{T}$.

The following theorem gives the asymptotic distribution of $\hat{B}_{J_i,i}$.

Theorem 4.5. *Suppose Assumptions A.1-A.5 hold, $N^2T^{1-q/4}(logN)^{-q/2}\to 0$ and $N/T^2\to 0$ as $(N,T)\to\infty$. Let S_i denote an $L\times|J_i|$ selection matrix such that $\|S_i\|_F$ is finite and L is an fixed integer. Conditional on the event $\{\hat{B}=_s B^0\}$, for $i=1,...,N$, we have*

$$\sqrt{T}S_i(\hat{B}_{J_i,i}-B^0_{J_i,i})\xrightarrow{d}N(\mathbf{0},\sigma_i^2S_i(\Sigma_{J_i,J_i})^{-1}S_i').$$

Note that, we specify a selection matrix S_i in Theorem 4.5 that is not needed if $|J_i|$ is fixed. Intuitively, we allow $|J_i|$ to diverge to infinity as $(N,T)\to\infty$ and we cannot derive the asymptotic normality of $\hat{B}_{J_i,i}$ directly when $|J_i|\to\infty$. Instead, we follow standard practice on estimation and inference with a diverging number of parameters (see, e.g., Fan and Peng 2004; Lam and Fan 2008; Qian and Su 2016) and prove asymptotic normality for arbitrary linear combinations of the elements of $\hat{B}_{J_i,i}$. In the special case where $|J_i|$ is fixed, we can take $S_i=I_{|J_i|}$ and obtain the usual joint asymptotic normal distribution of the $\hat{B}_{J_i,i}$'s.

4.3.4 Tuning parameter selection

In practice, we need to select the tuning parameters γ_ℓ, for $\ell=1,...,4$. For γ_2, which is the tuning parameter for the nuclear norm penalty, we adopt a simple plug-in approach similar to that introduced in Chernozhukov et al. (2018). An ideal tuning parameter for γ_2 is one such that

$$||\mathbf{U}||_{\mathrm{op}}/\sqrt{NT}\leq(1-c)\gamma_2$$

for some $c > 0$ with high probability. Suppose $\mathbf{U}$ is a random matrix with i.i.d. sub-Gaussian entries that have mean zero and variance σ_u^2, its operator norm is bounded by $C\sigma_u(\sqrt{N} + \sqrt{T})$, for some $C > 0$ with high probability (see Vershynin, 2018). One can first use $\gamma_2 = \frac{\hat{\sigma}_y}{C}(\sqrt{N} + \sqrt{T})$ for some $C > 1$ and $\hat{\sigma}_y$ is the sample standard deviation of Y. After obtaining an estimate $\hat{\sigma}_u$ of σ_u, we can calculate a suitable γ_2 via simulation. Specifically, we can simulate the random matrices $\mathbf{U}$ with i.i.d. $N(0, \hat{\sigma}_u^2)$. Then we let $\gamma_2 = Q(||\mathbf{U}||_{\text{op}}, 0.95)$, where $Q(x, \alpha)$ denote the α^{th} quantile of x.

For γ_1, γ_3, and γ_4, we propose to use the 5-fold cross validation (CV) process. For the first-step estimation, the procedure goes as follows:

1. Partition the data into 5 separate sets along the time dimension $(\mathbf{T}_1, ..., \mathbf{T}_5 \subset [T])$;

2. For $k = 1, ..., K$, fit the model to the training set by excluding the kth fold data. Denote the estimators by $\tilde{B}^{(\gamma,k)}$ and $\tilde{\Lambda}^{(\gamma,k)}$, where $\tilde{\Lambda}^{(\gamma,k)}$ is a $N \times R$ matrix containing the first R right singular vectors of $\tilde{\Theta}$. Calculate the sum of squared prediction errors
$$cv(\gamma, k) = \text{tr}[(\mathbf{Y}_{\mathbf{T}_k} - \mathbf{X}_{\mathbf{T}_k}\tilde{B}^{(\gamma,k)})\mathbb{M}_{\tilde{\Lambda}^{(\gamma,k)}}(\mathbf{Y}_{\mathbf{T}_k} - \mathbf{X}_{\mathbf{T}_k}\tilde{B}^{(\gamma,k)})'];$$

3. Compute the CV error for a fixed tuning parameter by $CV(\gamma) = \sum_{k=1}^{5} cv(\gamma, k)$.

4. Select $\gamma^* = \arg\min_\gamma CV(\gamma)$.

Remark 4.4 Once the sample $\mathbf{T}_k$ is excluded, we cannot obtain an estimate of $F_{\mathbf{T}_k,*}$. Hence we cannot obtain the residuals by deducting the estimate of $F_{\mathbf{T}_k,*}\Lambda'$. For this reason, we multiply $\mathbf{Y}_{\mathbf{T}_k} - \mathbf{X}_{\mathbf{T}_k}\tilde{B}^{(\gamma,k)}$ by $\mathbb{M}_{\tilde{\Lambda}^{(\gamma,k)}}$ to project out $F_{\mathbf{T}_k,*}\Lambda'$ in the above procedure.

For the second-step estimator, the CV procedure is standard. For the third step, we fix the tuning parameter before the iterations begin.

4.3.5 Lag length selection

In the estimation procedure, we have so far assumed that the lag length p is known. In practice, the lag length p is usually unknown and requires estimation. In this subsection, we propose a procedure to determine the lag length p. Suppose we estimate the model with some $p_{\max} \geq p^0$, where we use the superscript '0' to denote the true parameter. The model with $p_{\max}$ continues to be a correctly specified model except that $A_k^0 = \mathbf{0}$ for $k > p^0$. Due to LASSO regularization, the estimator $\hat{A}_p$ for $p > p^0$ should shrink to zero. Noting this point, we propose to determine the lag length by the following procedure:

1. Obtain the estimates $\hat{A}_k$ with lag length $p_{\max}$;

2. Calculate $a_k = ||\hat{A}_k||_{\mathrm{F}}^2 \vee c$ for some constant c and $k = 1, ..., p_{\max}$;

3. The criterion function we consider is given by the ratio
$$GR(p) = \frac{\sum_{k=p}^{p_{\max}} a_k}{\sum_{k=p+1}^{p_{\max}} a_k}, \ p = 1, ..., p_{\max} - 1.$$
The term GR refers to the growth ratio of $\sum_{k=p}^{p_{\max}} a_k$.

4. The estimator of p^0 is the maximizer of $GR(p) : \hat{p} = \arg\max_{1 \leq k < p_{\max}} GR\,(k)$.

Remark 4.5 (i) One can also simply run an ℓ_1-nuclear penalized regression with $p_{\max}$, which is the first step of the estimation procedure given in Section 3.1. We only require that $||\hat{A}_k - A_k^0||_{\mathrm{F}}$ converge to zero at a certain rate.

(ii) In practice, one may obtain a very small or even zero $||\hat{A}_k||_{\mathrm{F}}^2$ for large $k > p^0$. In this case, if we directly use $a_k = ||\hat{A}_k||_{\mathrm{F}}^2$, the growth ratio may possibly choose a larger p than p^0. To solve this problem, we bound a_k below by some constant $c > 0$.

(iii) The $GR(p)$ criterion function is constructed to allow for zeros A_k^0 for $k < p^0$. If we believe all A_k^0 are nonzero, one can also consider the criterion function $FR(p) = a_p/a_{p+1}$, where the term FR refers to Frobenius norm ratio.

4.4 Monte Carlo Simulations

In this section we evaluate the finite sample performance of our estimation procedure by means of a set of Monte Carlo experiments.

4.4.1 Data generating processes

We consider three main cases with $p = 1$. In each case, we consider both strict sparsity and approximate sparsity subcases. Thus, we have six data generating processes (DGPs) in total. For each DGP, we generate the data from the following high dimensional VAR(1) process with CFs:

$$Y_t = A_1^0 Y_{t-1} + \Lambda^0 f_t^0 + u_t, \tag{4.10}$$

where A_1^0 varies across different DGPs, $\Lambda^0 = (\lambda_1^0, ..., \lambda_N^0)'$. The factor loading λ_{ri}^0, for $r = 1, ..., R^0$, is independently and identically distributed (i.i.d.) standard normal random variables. The factors f_{tr}^0, for $r = 1, ..., R^0$, follow an autoregressive process:

$$f_{tr}^0 = \rho_f \cdot f_{t-1,r}^0 + \epsilon_{tr}^{(f)},$$

where $\rho_f = 0.6$ and $\epsilon_{tr}^{(f)}$ are i.i.d. $N(0, 1)$. The idiosyncratic error terms are generated as $u_{it} = s \cdot \epsilon_{it}^{(u)}$, where s controls the signal-to-noise ratio, and $\epsilon_{it}^{(u)}$ are i.i.d. $N(0, 1)$.

DGP 1 (Tridiagonal transition matrix): $(A_1^0)_{ij} = 0.3 \cdot \mathbf{1}(|i - j| \leq 1)$.

DGP 2 (Block-diagonal transition matrix): We generate a block-diagonal matrix A_1^0 =blkdiag$(S_1, ..., S_K)$, where the S_k's are 5×5 random matrices. The diagonal entries of S_k are fixed with $(S_k)_{i,i} = 0.3$. In each column of S_k, we randomly choose 2 out of 4 off-diagonal entries to be -0.3.

DGP 3 (Random matrix): We fix diagonal entries of A_1^0 to be 0.3 (i.e. $(A_1^0)_{ii} = 0.3$). In each row of A_1^0, we randomly choose 3 out of $N - 1$ entries to be -0.3.

FIGURE 4.1 around here

Figure 4.1 illustrates the structure of the random transition matrices used in our simulation. For each DGP, we consider $N = 30, 60$, and $T = 100, 200, 400$, leading to six combinations of cross-sectional and time series dimensions. The number of replications is set to 500.

4.4.2 Implementation and estimation results

For each DGP, we consider the feasible estimator proposed in this work and the oracle least squares estimator. The oracle estimators are obtained by using the information of the number of factors and the true regressors.

Table 4.1 reports the model selection accuracy. For each combination of N and T in each DGP, the fourth and fifth columns report the under- and over-estimation rate of $\hat{R}$, respectively. The TPR (true positive rate) columns report the average shares of relevant variables included. The FPR (false positive rate) columns report the average shares of irrelevant variables included. We summarize some important findings from Table 4.1. First, the proposed hard singular value thresholding procedure correctly determined the number of factors for each case. Second, with N fixed, the TPR increases with T in all cases as expected. All three step estimators can include almost all the true regressors when $T = 400$. Third, among the three estimators, the conservative LASSO (3rd step) estimator includes the least regressors with zero coefficients in almost all settings. In addition, only conservative LASSO estimator tends to exclude more irrelavent regressors as T increases, while the FPRs of the first and second step estimators increase as T grows.

TABLE 4.1 around here

Table 4.2 reports the estimation error of both the feasible estimators and the oracle least squares estimator. We report the root mean squared errors (RMSEs) for all entries and nonzero entries respectively. We summarize some important findings from Table 4.2. First, as expected, the oracle least squares estimator uniformly outperforms the feasible estimators. This is mainly due to the fact that the FPRs of feasible estimators were never zero. Second, the RMSE of the oracle estimator for nonzero entries decreases with T at a $\sqrt{T}$ rate and changes with N slightly. This is consistent with our theoretical prediction that the oracle least squares estimator converges to the true values at the $\sqrt{T}$ rate. Third, the conservative LASSO outperforms the other two feasible estimators in terms of RMSEs in all cases.

TABLE 4.2 around here

For all DGPs, we also consider estimation of a misspecified VAR(1) model, $Y_t = A_1^0 Y_{t-1} + u_t$, where the common factors are ignored. We first estimate the model with LASSO as in KC (2015). Then we construct the weights as in 4.9 and use conservative LASSO to estimate the misspecified model. Table 4.3 reports the performance of these two estimators. We summarize some important findings from Table 4.3. First, the FPRs for both estimators are quite high. This indicates that the misspecification may lead to non-sparse estimates of the transition matrices, in the presence of latent factors. Second, the estimators for the misspecified model also have higher RMSEs. Third, in many cases, the conservative LASSO estimator performs even worse than the LASSO estimator in terms of RMSEs.

TABLE 4.3 around here

4.5 Empirical application

4.5.1 Evaluating a network of financial assets volatilities

In recent years, financial asset connectedness has been an active topic in financial econometrics. Examples of contributions to this literature include Barigozzi and Brownlees (2019; hereafter BB), Barigozzi and Hallin (2017), Billio et al. (2012), Diebold and Yilmaz (2014; hereafter DY) and Diebold and Yilmaz (2015) , and Hautsch et al. (2014). Some of these authors directly model the large panel of time series as a vector autoregressive process without considering the existence of potential common factors. A LASSO type method is employed to estimate the transition matrices. However, Barigozzi and Hallin (2017) and BB (2019) documented evidence for the existence of a factor structure in volatility. Barigozzi and Hallin (2017) considered controlling for the presence of common factors by means of a dynamic factor model. BB (2019) use regression residuals of individual volatilities on observed factors (e.g., market volatility or sector-specific volatility) to represent the idiosyncratic components of the volatilities. Neither of these papers provides theoretical justification for the approach.

In this empirical application, we extend the measure of connectedness of DY and study the connectedness of financial assets. Specifically, we study the connectedness in a panel of volatility measures. As remarked by DY, the volatilities of financial assets can be interpreted as a form of 'investor fear'. Then volatility connectedness represents 'fear connectedness' across assets. In this scenario, it is natural to take into account common factors, which reflect confidence in the market. Spillover effects across assets is another reason for connectedness. We use the econometric methodology derived in the present work to analyze a panel of return volatilities of 23 sector ETF funds. The findings show that common factors account for 58% of the overall variability. Conditioning on these factors, the interdependence across individuals still captures a relatively high proportion of the variation.

Data description and empirical framework

We collect the weekly 'open price', 'close price', 'high price' and 'low price' of a series of sector ETF funds from Yahoo finance. A list of the fund names and tickers is given in Table 4.4.

Table 4.4 around here

They fall into several categories. The 'Energy', 'Financial' and 'Consumer cyclical' are three large categories which contain three to four funds. The other categories contain at most two funds. The sample spans July 2007 to August 2019, which corresponds to 688 weeks. As volatility is unobserved, we use the observed price data to estimate it. Specifically, we follow Garman and Klass (1980) and Alizadeh et al. (2002) to estimate asset volatility by the measure

$$\begin{aligned}\tilde{\sigma}_{it}^2 &= 0.511(H_{it}-L_{it})^2 - 0.019[(C_{it}-O_{it})(H_{it}+L_{it}-2O_{it}) - 2(H_{it}-O_{it})(L_{it}-O_{it})] \\ &\quad -0.383(C_{it}-O_{it})^2,\end{aligned}$$

where O_{it}, C_{it}, H_{it}, and L_{it} are natural logarithms of weekly 'open price', 'close price', 'high price' and 'low price', respectively. We present descriptive statistics for volatilities in Table 4.5. The kurtosis of each time series is quite large. We follow DY (2014) to normalize the data by taking natural logarithms and then center each time series, that is our y_{it} is given by $\log(\tilde{\sigma}_{it}^2) - \overline{\log(\tilde{\sigma}_{i\cdot}^2)}$.

Table 4.5 around here

Given the panel of volatilities, we fit the data in our VAR with common factors model in (4.1). By the decomposition (4.5), $y_{it} = y_{it}^{(f)} + y_{it}^{(u)}$, where $y_{it}^{(f)}$ is due to the common factors and $y_{it}^{(u)}$ is due to the idiosyncratic errors. In addition var(y_{it}) =var$(y_{it}^{(f)})$+var$(y_{it}^{(u)})$. Then $\nu_i \equiv$var$(y_{it}^{(f)})$/var(y_{it}) measures the proportion of variance in y_{it} that is due to common factors and $\bar{\nu} \equiv \sum_{i=1}^N \text{var}(y_{it}^{(f)}) / \sum_{i=1}^N \text{var}(y_{it})$ measures the proportion of variation in all time series.

For the idiosyncratic component $y_{it}^{(u)}$, we can calculate the measure of connectedness proposed by DY (2014). As discussed in the Section 4.2, we have $y_{it}^{(u)} = \sum_{j=0}^{\infty} \alpha_{iN}^{(u)}(j) C^{(u)} \epsilon_{t-j}^{(u)}$, where $\alpha_{iN}^{(u)}(j) = (e_{1,p} \otimes e_{i,N})' \Phi^j (e_{1,p} \otimes I_N)$ and $\epsilon_t^{(u)} \sim (0, I_N)$. One can treat the $\epsilon_{it}^{(u)}$'s as the idiosyncratic shocks to individual i. The variance of H-step ahead prediction error due to $\{\epsilon_{j,t+h}^{(u)}\}_{h=1}^H$ is $s_{ij}^H = \sum_{h=0}^{H-1} ([\alpha_{iN}^{(u)}(h) C^{(u)}]_j)^2$. If we can identify both Φ and $C^{(u)}$, we can easily estimate the variance decomposition matrix $\check{D}^H$ with (i,j)th entry $s_{ij}^H / \sum_{k=1}^N s_{ik}^H$. However, $C^{(u)}$ is not identified without further assumption. Although we cannot identify $C^{(u)}$, the matrix $\Sigma_u = C^{(u)} C^{(u)\prime}$ is identified. DY (2014) propose to calculate the H-step generalized variance decomposition matrix $D^H = [d_{ij}^H]_{N\times N}$, where

$$d_{ij}^H = \frac{\sigma_{jj}^{-1}\sum_{h=0}^{H-1}(\alpha_{iN}^{(u)}(h)\Sigma_u e_{j,N})^2}{\sum_{h=0}^{H-1}\alpha_{iN}^{(u)}(h)\Sigma_u\alpha_{iN}^{(u)}(h)'},$$ and $e_{j,N}$ is jth unit N dimensional vector.

Unlike $\check{D}^H$, the row sums of D^H are not necessarily unity. We normalize D^H to $\tilde{D}^H$ with (i,j)th entry $\tilde{d}_{ij}^H = d_{ij}^H / \sum_{k=1}^N d_{ik}^H$ so that $\sum_{j=1}^N \tilde{d}_{ij}^H = 1$ and $\sum_{i,j=1}^N \tilde{d}_{ij}^H = N$. Hence, overall connectedness in the $y_{it}^{(u)}$'s can be measured as $\tilde{d}^H = \sum_{i\neq j}\tilde{d}_{ij}^H/N$. In addition, we let $\tilde{d}_{i\leftarrow}^H \equiv \sum_{j\neq i}\tilde{d}_{ij}^H$. Following DY (2014), we call $\tilde{d}_{i\leftarrow}^H$ the 'FROM' index, as it measures the proportion of generalized variance decomposition that is due to other individuals. Similarly, we let $\tilde{d}_{\leftarrow j}^H \equiv \sum_{i\neq j}\tilde{d}_{ij}^H$ and call this the 'TO' index.

Estimation results

We use the procedure proposed in section 4.3.3 to determine the lag length with $p_{\max} = 8$. The result gives $\hat{p} = 4$. When we run the regression with $p = 4$, the number of factors is determined to be one ($\hat{R} = 1$).

Figure 4.2 around here

Figure 4.2 reports the heat map which represents the estimates of the $\hat{A}_k$'s. The element value is represented by scaled color. First, most of the nonzero entries are estimated to be positive. The positive coefficients represent the propagation of investor fear across assets. Second, the diagonal elements of $\hat{A}_k$'s are mostly nonzero. The magnitude of the diagonal elements is larger than that of the off diagonal elements on average. Third, the number of nonzero coefficients in $\hat{A}_k$ decreases as k increases and the average magnitude of the entries also decreases. More recent investor fear causes greater present investor fear. In all, 330 out of 2116 entries are nonzero.

Table 4.6 around here

Next we calculate the statistics introduced in the last subsection. Table 4.6 provides the estimates of ν_i, $\tilde{d}_{i\leftarrow}^H$, and $\tilde{d}_{\leftarrow j}^H$. Almost all the γ_i's are above 50%, and the overall variation due to the common factors is $\bar{\nu} = 58.4\%$. The market level investor fear is playing a dominant roll in investor trading behavior. After conditioning on the factors, we consider the idiosyncratic part by looking at $\tilde{d}^H_{i\leftarrow}$, $\tilde{d}^H_{\leftarrow j}$ and the H-step generalized variance decomposition matrix $\tilde{D}^H$. The 'FROM' index ranges between 28.7% and 72.7%. Interestingly, the 'energy' and 'finance' funds have higher 'FROM' index compared to other funds. A similar observation applies for the 'TO' index. Specifically, the 'TO' index of XLE and IYE are close to 100% and both are 'energy' funds. The energy industry therefore transmits considerable investor fear to the entire market. This finding is intuitive as the oil price has been extremely volatile in recent years and the energy price affects all industries. The fund GDX (VanEck Vectors Gold Miners ETF) has the least connectedness. It receives only 28.7% connectedness from other assets and transmits only 23.6% connectedness to others. The overall connectedness measure is 49.8%. Conditioning on the factors, there is still substantive transmission of investor fear across individuals. Figure 4.3 reports the heat map of the H-step generalized variance decomposition matrix $\tilde{D}^H$ at $H = 12$. We observe that the interconnections within the same category is high, whereas connectedness across categories is relatively low.

Figure 4.3 around here

The lower panel of Table 4.6 provides the measure of connectedness with the pure VAR model estimation as in Demirer et al. Without controlling for the common factors, the 'FROM; and 'TO' index of each fund becomes much larger. However, we observe little heterogeneity across categories. In this case, all the connectedness due to common factors is interpreted as the individual level connectedness, which potentially leads to wrong inference.

In sum, our framework extends traditional VAR(p) analysis of financial asset connectedness to control for the presence common factors in the determination of volatility. We have found that common factors account for more than half of the variation in the data. In addition to the connectedness that is due to common factors there is still a remarkable degree of connectedness that arises from spillover channels that operate among the assets themselves.

4.6 Conclusion

The methodology developed in this paper provides for regularized estimation of high dimensional VARs with unobserved common factors that allow for strong cross section dependence as well as serial dependence among the time series. Incorporating such dependence is particularly important in high dimensional disaggregated data where connectedness between the variables may arise through many different channels. This dependence and connectedness are seen to be especially relevant in studying the transmission of investor fear across financial assets in our study of asset price volatility.

The practical elements involved in implementing our procedure can be summarized as follows. Given VAR lag length, which is later chosen by means of a growth ratio criterion, preliminary estimates of the model are obtained using ℓ_1-nuclear norm regularized estimation. The number of factors and a preliminary estimate of the common factors are obtained and the correct number of factors can be estimated with probability approaching one. Next, we estimate the model using the generalized LASSO using the preliminary estimate of the common factors as regressors. Conservative LASSO is then used to obtain the final estimates, which are asymptotically equivalent to the oracle least squares estimates obtained as if the true regression model were known.

The methods and results open up multiple avenues for further research. First, following Barigozzi and Brownlees (2019) it may be useful in practice to impose some sparsity assumptions on the large dimensional error variance matrix and develop estimation methods to achieve this. Second, frequency domain methods can be used to estimate the common factor components. Third, the model studied here does not allow for structural change in the transition matrices or the factor loadings. It will also be interesting and challenging to study high dimensional VAR models with common factors that may involve time-varying transition matrices and factor loadings, which can help to capture empirically evolution in institutional and regulatory frameworks.

Tables and Figures

Table 4.1: Model Selection Accuracy

			Number of factors		Step 1		Step 2		Step 3	
DGP	N	T	UER	OER	TPR	FPR	TPR	FPR	TPR	FPR
1	30	100	0.0%	0.0%	97.4%	19.3%	98.8%	18.5%	93.7%	8.0%
	30	200	0.0%	0.0%	99.6%	19.1%	99.9%	18.1%	99.4%	5.8%
	30	400	0.0%	0.0%	99.9%	21.8%	100.0%	19.5%	99.9%	4.9%
	60	100	0.0%	0.0%	96.8%	12.7%	98.2%	12.2%	90.5%	5.1%
	60	200	0.0%	0.0%	99.9%	12.2%	100.0%	11.7%	99.1%	2.6%
	60	400	0.0%	0.0%	100.0%	11.9%	100.0%	11.1%	99.9%	1.7%
2	30	100	0.0%	0.0%	86.2%	21.8%	83.9%	18.9%	94.0%	15.7%
	30	200	0.0%	0.0%	95.3%	28.0%	93.7%	24.8%	99.4%	12.8%
	30	400	0.0%	0.0%	99.2%	37.0%	98.7%	33.3%	99.9%	8.2%
	60	100	0.0%	0.0%	76.7%	10.3%	76.5%	9.4%	90.6%	10.7%
	60	200	0.0%	0.0%	88.9%	12.5%	89.7%	12.0%	99.2%	8.9%
	60	400	0.0%	0.0%	96.4%	17.7%	95.8%	16.7%	100.0%	5.5%
3	30	100	0.0%	0.0%	93.2%	24.9%	92.3%	22.0%	96.5%	17.4%
	30	200	0.0%	0.0%	98.1%	31.4%	97.6%	27.6%	99.6%	11.7%
	30	400	0.0%	0.0%	99.5%	38.4%	99.3%	34.4%	99.7%	7.3%
	60	100	0.0%	0.0%	88.1%	12.8%	88.4%	11.8%	95.9%	11.8%
	60	200	0.0%	0.0%	96.1%	15.6%	95.5%	13.9%	99.8%	9.4%
	60	400	0.0%	0.0%	98.9%	19.5%	98.6%	17.9%	100.0%	4.5%

Note: We report the under/over-estimation rate (UER and OER) of the number of factors in the UER and OER column, respectively. The true positive rate (TPR) columns report the average shares of relevant variables included. The FPR (false positive rate) columns report the average shares of irrelevant variables included.

Table 4.2: Root mean squared errors of the feasible and oracle transition matrix estimators

			All entries				Nonzero entries			
DGP	N	T	Oracle	Step 1	Step 2	Step 3	Oracle	Step 1	Step 2	Step 3
1	30	100	0.019	0.063	0.059	0.050	0.062	0.145	0.132	0.117
	30	200	0.014	0.055	0.051	0.033	0.044	0.118	0.106	0.066
	30	400	0.010	0.052	0.049	0.029	0.033	0.100	0.092	0.047
	60	100	0.013	0.044	0.041	0.038	0.061	0.150	0.138	0.131
	60	200	0.010	0.035	0.032	0.021	0.043	0.108	0.098	0.066
	60	400	0.007	0.033	0.031	0.016	0.032	0.089	0.080	0.041
2	30	100	0.018	0.065	0.065	0.057	0.056	0.177	0.184	0.154
	30	200	0.012	0.055	0.055	0.038	0.039	0.142	0.150	0.103
	30	400	0.009	0.047	0.047	0.027	0.028	0.110	0.119	0.070
	60	100	0.012	0.050	0.049	0.044	0.054	0.204	0.205	0.179
	60	200	0.008	0.042	0.041	0.028	0.038	0.170	0.168	0.114
	60	400	0.006	0.035	0.035	0.019	0.027	0.138	0.143	0.081
3	30	100	0.019	0.065	0.064	0.055	0.051	0.150	0.155	0.127
	30	200	0.013	0.053	0.053	0.035	0.035	0.117	0.123	0.082
	30	400	0.009	0.047	0.047	0.027	0.025	0.095	0.100	0.058
	60	100	0.013	0.050	0.049	0.042	0.049	0.173	0.173	0.146
	60	200	0.009	0.039	0.040	0.024	0.034	0.135	0.140	0.085
	60	400	0.006	0.033	0.033	0.015	0.024	0.109	0.113	0.056

Note: We report the root mean squared errors (RMSEs) of the feasible and oracle transition matrix estimators. 4th-7th columns report the RMSEs of all entries. 8th-11th columns report the RMSEs of non-zero entries.

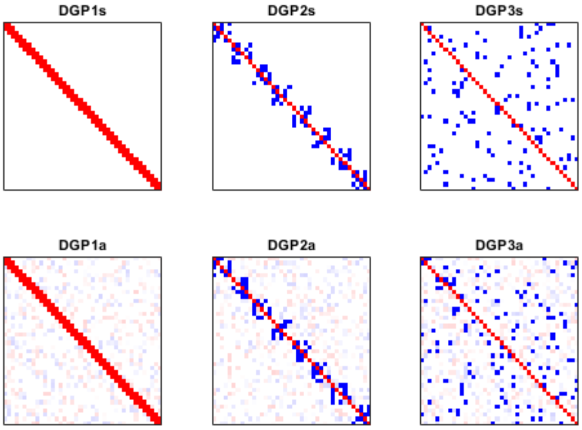

Figure 4.1: Structure of transition matrices

Table 4.3: Results of misspecified estiamtes

			LASSO				Conservative LASSO			
DGP	N	T	TPR	FPR	RMSE$_a$	RMSE$_b$	TPR	FPR	RMSE$_a$	RMSE$_b$
1	30	100	78.7%	34.9%	0.115	0.208	78.4%	45.2%	0.178	0.227
	30	200	88.9%	37.7%	0.094	0.178	88.1%	43.3%	0.129	0.173
	30	400	95.3%	45.0%	0.083	0.150	94.5%	43.0%	0.103	0.134
	60	100	71.0%	22.6%	0.086	0.216	72.8%	39.5%	0.161	0.240
	60	200	86.7%	25.7%	0.070	0.179	87.0%	38.9%	0.114	0.175
	60	400	94.9%	30.2%	0.058	0.148	95.3%	37.9%	0.083	0.128
2	30	100	86.2%	59.6%	0.150	0.202	81.9%	54.8%	0.211	0.233
	30	200	95.0%	61.5%	0.107	0.152	91.7%	51.4%	0.139	0.159
	30	400	98.9%	66.3%	0.080	0.113	97.7%	50.5%	0.098	0.110
	60	100	77.0%	46.6%	0.135	0.218	74.1%	48.9%	0.222	0.263
	60	200	91.6%	51.9%	0.100	0.165	86.8%	44.6%	0.143	0.175
	60	400	98.3%	56.1%	0.072	0.120	96.7%	44.4%	0.097	0.116
3	30	100	89.2%	59.2%	0.139	0.186	85.7%	55.9%	0.196	0.215
	30	200	96.2%	61.4%	0.102	0.141	94.0%	54.3%	0.133	0.148
	30	400	99.1%	67.1%	0.079	0.107	98.3%	53.2%	0.096	0.106
	60	100	82.0%	46.1%	0.126	0.203	79.8%	50.6%	0.208	0.247
	60	200	94.0%	51.7%	0.093	0.151	90.5%	46.6%	0.135	0.164
	60	400	98.8%	55.5%	0.068	0.110	97.6%	45.0%	0.091	0.109

Note: We report the true positive rate (TPR), false positive rate (FPR), root mean squared errors of all entries (RMSE$_a$) and nonzero entries (RMSE$_b$) of misspecified estimates. We consider the LASSO estimator as in Kock and Callot (2015) and a conservative LASSO estimator. The LASSO estimator was used to construct weights for conservative LASSO.

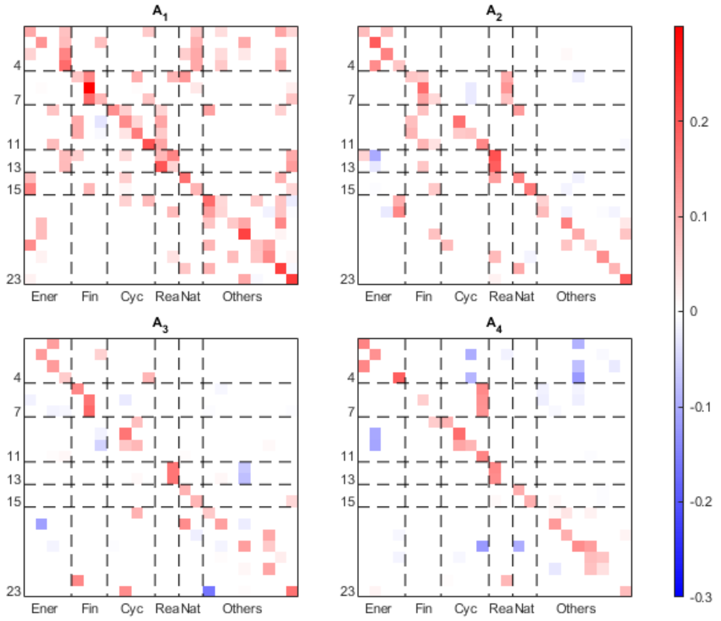

Figure 4.2: Heat map of the transition matrices

Table 4.4: Funds information

category	ticker	fund name	category	ticker	fund name
Energy	XLE	Energy Select Sector SPDR Fund	Natu	XLB	Materials Select Sector SPDR Fund
	XOP	Spdr S&P Oil & Gas Explo & Prod Etf		XME	SPDR S&P Metals & Mining ETF
	IYE	iShares U.S. Energy ETF	Tech	XLK	Technology Select Sector SPDR Fund
	OIH	VanEck Vectors Oil Services ETF		SMH	VanEck Vectors Semiconductor ETF
Financial	XLF	Financial Select Sector SPDR Fund	Heal	XLV	Health Care Select Sector SPDR Fund
	KBE	SPDR S&P Bank ETF		IBB	iShares Nasdaq Biotechnology ETF
	KRE	SPDR S&P Regional Banking ETF	Def	XLP	Consumer Staples Select Sector SPDR Fund
Cyc	XLY	Cons. Disc. Select Sector SPDR Fund	Util	XLU	Utilities Select Sector SPDR Fund
	XHB	Spdr S&P Homebuilders Etf	Indu	XLI	Industrial Select Sector SPDR Fund
	ITB	iShares U.S. Home Construction ETF	EPM	GDX	VanEck Vectors Gold Miners ETF
	XRT	Spdr S&P Retail Etf			
Rea	IYR	iShares U.S. Real Estate ETF			
	VNQ	Vanguard Real Estate Index Fund ETF			

Note. Cyc, Rea, Natu, Tech, Heal, Def, Util, Indu and EMP stand for consumer cyclical, real estate, natural resource, technology, health care, consumer defensive, utilities, industrials and equity precious metals, respectively.

Table 4.5: Descriptive statistics

TICKER	XLE	XOP	IYE	OIH	XLF	KBE	KRE	XLY
mean	0.00136	0.00246	0.00141	0.00220	0.00157	0.00194	0.00184	0.00082
median	0.00063	0.00130	0.00059	0.00128	0.00041	0.00059	0.00066	0.00029
max	0.06034	0.06290	0.11527	0.05856	0.05743	0.04793	0.09748	0.03063
min	0.00004	0.00005	0.00004	0.00008	0.00001	0.00002	0.00002	0.00001
std	0.00369	0.00472	0.00549	0.00418	0.00463	0.00484	0.00539	0.00214
skewness	10.954	7.604	15.469	8.159	7.645	5.823	11.530	8.869
kurtosis	151.595	77.386	291.137	88.226	77.152	44.720	175.439	102.667
TICKER	XHB	ITB	XRT	IYR	VNQ	XLB	XME	XLK
mean	0.00218	0.00251	0.00115	0.00137	0.00146	0.00098	0.00264	0.00071
median	0.00079	0.00102	0.00056	0.00039	0.00041	0.00047	0.00133	0.00031
max	0.05071	0.04660	0.03094	0.04847	0.04831	0.02948	0.05631	0.03112
min	0.00007	0.00001	0.00001	0.00003	0.00004	0.00004	0.00014	0.00002
std	0.00431	0.00473	0.00231	0.00377	0.00403	0.00205	0.00510	0.00187
skewness	5.305	4.936	7.783	6.789	6.958	8.059	6.912	9.814
kurtosis	41.414	33.799	83.839	61.695	64.487	90.224	62.231	128.784
TICKER	SMH	XLV	IBB	XLP	XLU	XLI	GDX	
mean	0.00111	0.00054	0.00105	0.00036	0.00062	0.00075	0.00263	
median	0.00069	0.00025	0.00058	0.00016	0.00030	0.00036	0.00154	
max	0.02010	0.02865	0.03488	0.02197	0.03903	0.02108	0.07009	
min	0.00004	0.00002	0.00003	0.00001	0.00003	0.00001	0.00010	
std	0.00153	0.00162	0.00207	0.00111	0.00193	0.00156	0.00439	
skewness	5.713	11.898	9.968	13.670	14.053	7.405	8.300	
kurtosis	52.259	176.016	135.878	237.109	250.309	76.935	102.080	

Table 4.6: Connectedness measures across funds

Connectedness measures by estimates of VAR with CFs model								
TICKER	XLE	XOP	IYE	OIH	XLF	KBE	KRE	XLY
ν_i	64.9%	59.1%	65.4%	58.0%	65.2%	56.8%	56.6%	72.0%
FROM	71.4%	65.4%	71.7%	64.3%	61.7%	61.3%	62.3%	51.8%
TO_i	106.8%	86.0%	103.9%	71.5%	57.8%	72.6%	51.4%	37.3%
TICKER	XHB	ITB	XRT	IYR	VNQ	XLB	XME	XLK
ν_i	53.6%	49.5%	60.1%	50.7%	49.7%	67.2%	56.9%	70.5%
FROM_i	60.5%	58.3%	36.5%	57.9%	58.6%	37.5%	44.1%	39.0%
TO_i	56.3%	41.7%	19.0%	79.7%	74.4%	26.3%	37.2%	37.3%
TICKER	SMH	XLV	IBB	XLP	XLU	XLI	GDX	average
ν_i	54.8%	64.3%	50.7%	61.3%	50.6%	67.7%	31.0%	$\bar{\nu}$ =56.1%
FROM_i	31.9%	38.3%	28.8%	30.7%	29.7%	40.9%	27.7%	$\bar{d}^{12}$ =49.8%
TO_i	23.3%	34.1%	33.0%	21.2%	19.6%	20.7%	19.1%	
Connectedness measures by estimates of pure VAR model								
TICKER	XLE	XOP	IYE	OIH	XLF	KBE	KRE	XLY
FROM_i	89.3%	87.1%	89.4%	87.0%	89.6%	86.8%	87.6%	90.9%
TO_i	105.0%	79.5%	103.0%	77.7%	112.9%	97.0%	89.1%	110.5%
TICKER	XHB	ITB	XRT	IYR	VNQ	XLB	XME	XLK
FROM_i	87.3%	86.3%	88.8%	85.7%	86.2%	90.1%	88.8%	89.8%
TO_i	95.8%	80.8%	79.1%	94.0%	89.6%	105.6%	80.1%	103.8%
TICKER	SMH	XLV	IBB	XLP	XLU	XLI	GDX	average
FROM_i	87.6%	88.1%	83.8%	88.4%	85.7%	89.8%	76.5%	$\bar{d}^{12}$ =87.40%
TO_i	74.8%	81.2%	60.8%	80.0%	60.0%	104.3%	45.8%	

Note. Cyc, Rea, Natu, Tech, Heal, Def, Util, Indu and EMP stand for consumer cyclical, real estate, natural resource, technology, health care, consumer defensive, utilities, industrials and equity precious metals, respectively.

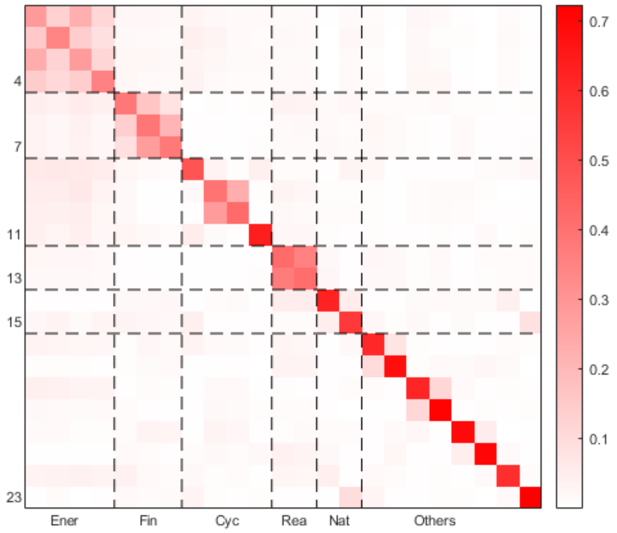

Figure 4.3: Heat map of $\tilde{D}^{12}$

Chapter 5
Conclusion

This study contribute to the estimation and inference theory of the heterogeneous large panel models. We have considered several types of heterogeneity in our three models: (1) slope heterogeneity due to the threshold effect; (2) slope and threshold parameters heterogeneity due to latent group structure; (3) time varying heterogeneity due to interactive fixed effects or common factors. The asymptotic properties of the proposed estimators of regression coefficients are established. For each model, we have proposed various tests to correctly specify the models, for instance, determining the number of factors or groups. Extensive Monte Carlo experiments are done to show the good performance of the proposed estimators and tests. In empirical applications, the methods are employed to study several problems in macro-economics and finance. We document significant level of heterogeneity in various real datasets.

In the future research, it is interesting to extend the models in several directions. First, one can consider heterogeneous panel threshold regressions with endogeneity. Second, one can extend the models by allowing for parameters to allow for structural change. Third, since we are in large T framework, it is interesting to extend the models to consider nonstationary time series data.

Appendix A

Technical Results for Chapter 2

Proof of the main results

In this appendix we prove the main results in the paper. The proof relies on some technical lemmas whose proofs are given in the online supplement.

Let $\pi_{NT} = \min(\sqrt{N}, \sqrt{T})$. As we require $N/T \to \kappa$ as $(N, T) \to \infty$, we may use the property $O(T) = O(N)$ in various places.

Proof of Theorem 2.1

To prove Theorem 2.1, we need the following two lemmas.

Lemma A.1. *Suppose Assumptions A.2 and A.4 hold. Then*

(i) $\frac{1}{NT}\|\mathbf{e}\mathbf{e}'\| = O_p(\pi_{NT}^{-1}), \quad \frac{1}{NT}\|\mathbf{e}'\mathbf{e}\| = O_p(\pi_{NT}^{-1}), \quad \frac{1}{NT}\|\mathbf{e}'\mathbf{e}F^0\| = O_p(\pi_{NT}^{-1}),$
$\frac{1}{NT}\|\Lambda^{0\prime}\mathbf{e}\mathbf{e}'\| = O_p(\pi_{NT}^{-1});$

(ii) $\frac{1}{\sqrt{NT}}\|\mathbf{e}F^0\| = O_p(1), \quad \frac{1}{\sqrt{NT}}\|\Lambda^{0\prime}\mathbf{e}\| = O_p(1), \quad \frac{1}{\sqrt{NT}}\|\Lambda^{0\prime}\mathbf{e}F^0\| = O_p(1);$

(iii) $\sup_{\gamma\in\Gamma}\frac{1}{NT}\|\mathbf{e}\mathbf{X}_{k,\gamma}'\| = O_p(\pi_{NT}^{-1})$ *for* $k = 1, \ldots, 2K;$

(iv) $\sup_{\gamma\in\Gamma}\frac{1}{\sqrt{NT}}|\mathrm{tr}(\mathbf{e}\mathbf{X}_{k,\gamma}')| = O_p(1), \quad \sup_{\gamma\in\Gamma}\frac{1}{\sqrt{NT}}|\mathrm{tr}(\mathbf{e}\mathbf{X}_{k,\gamma}'\mathbb{M}_{\Lambda^0})| = O_p(1)$ *for* $k = 1, \ldots, 2K.$

Lemma A.2. *Under Assumptions A.2 and A.4,*

(i) $\sup_{\Lambda\in\mathbb{L}}\left\|\frac{1}{NT}\sum_{t=1}^{T}X_t'\mathbb{M}_\Lambda e_t\right\| = O_p(\pi_{NT}^{-1}),$

(ii) $\sup_{(\Lambda,\gamma)\in\mathbb{L}\times\Gamma}\left\|\frac{1}{NT}\sum_{t=1}^{T}X_t(\gamma)'\mathbb{M}_\Lambda e_t\right\| = O_p(\pi_{NT}^{-1}),$

(iii) $\sup_{\Lambda\in\mathbb{L}}\left\|\frac{1}{NT}\sum_{t=1}^{T}f_t^{0\prime}\Lambda^{0\prime}\mathbb{M}_\Lambda e_t\right\| = O_p(\pi_{NT}^{-1}),$

(iv) $\sup_{\Lambda\in\mathbb{L}}\left|\frac{1}{NT}\sum_{t=1}^{T}e_t'\mathbb{P}_\Lambda e_t\right| = O_p(\pi_{NT}^{-2}),$

where $\mathbb{L} = \left\{\Lambda \in \mathbb{R}^{N\times R} : N^{-1}\Lambda'\Lambda = I_R\right\}.$

Proof of Theorem 2.1. (i) Let $X_t(\gamma_1,\gamma_2) = X_t(\gamma_1) - X_t(\gamma_2)$. Note that

$$Y_t - X_t\beta - X_t(\gamma)\delta \equiv \Lambda^0 f_t^0 + e_t + \Psi_t(\theta,\gamma),$$

where $\Psi_t(\theta,\gamma) \equiv -X_t(\beta-\beta^0) - X_t(\gamma)(\delta-\delta^0) - X_t(\gamma,\gamma^0)\delta^0 = -X_{t,\gamma}(\theta-\theta^0) - X_t(\gamma,\gamma^0)\delta^0$. The objective function can be rewritten as

$$\begin{aligned}\mathcal{L}(\theta,\Lambda,\gamma) &= \sum_{t=1}^{T}\left[\Lambda^0 f_t^0 + e_t + \Psi_t(\theta,\gamma)\right]'\mathbb{M}_\Lambda\left[\Lambda^0 f_t^0 + e_t + \Psi_t(\theta,\gamma)\right]\\ &= \mathcal{L}_1(\theta,\Lambda,\gamma) + \mathcal{L}_2(\theta,\Lambda,\gamma) + \sum_{t=1}^{T}e_t'\mathbb{M}_\Lambda e_t, \qquad (A.1)\end{aligned}$$

where

$$\mathcal{L}_1(\theta,\Lambda,\gamma) = \sum_{t=1}^{T}\Psi_t(\theta,\gamma)'\mathbb{M}_\Lambda\Psi_t(\theta,\gamma) + \sum_{t=1}^{T}f_t^{0\prime}\Lambda^{0\prime}\mathbb{M}_\Lambda\Lambda^0 f_t^0 + 2\sum_{t=1}^{T}\Psi_t(\theta,\gamma)'\mathbb{M}_\Lambda\Lambda^0 f_t^0, \text{ and}$$

$$\mathcal{L}_2(\theta,\Lambda,\gamma) = 2\sum_{t=1}^{T}f_t^{0\prime}\Lambda^{0\prime}\mathbb{M}_\Lambda e_t + 2\sum_{t=1}^{T}\Psi_t(\theta,\gamma)'\mathbb{M}_\Lambda e_t.$$

It is easy to verify that $\mathcal{L}(\theta^0,\Lambda^0,\gamma^0) = \sum_{t=1}^{T}e_t'\mathbb{M}_{\Lambda^0}e_t$. Then we have

$$\mathcal{L}(\theta,\Lambda,\gamma) - \mathcal{L}(\theta^0,\Lambda^0,\gamma^0) = \mathcal{L}_1(\theta,\Lambda,\gamma) + \mathcal{L}_2(\theta,\Lambda,\gamma) - \sum_{t=1}^{T}e_t'(\mathbb{P}_\Lambda - \mathbb{P}_{\Lambda^0})e_t. \qquad (A.2)$$

By Lemma A.2(i)-(iv), we have

$$\sup_{\|\theta\|\le M}\ \sup_{(\Lambda,\gamma)\in\mathbb{L}\times\Gamma}\left|\frac{1}{NT}\mathcal{L}_2(\theta,\Lambda,\gamma) - \frac{1}{NT}\sum_{t=1}^{T}e_t'(\mathbb{P}_\Lambda - \mathbb{P}_{\Lambda^0})e_t\right| = o_p(1).$$

This result, together with the fact that $\mathcal{L}(\widehat{\theta}, \widehat{\Lambda}, \widehat{\gamma}) - \mathcal{L}(\theta^0, \Lambda^0, \gamma^0) \leq 0$, implies $\frac{1}{NT}\mathcal{L}_1(\widehat{\theta}, \widehat{\Lambda}, \widehat{\gamma}) + o_p(1) \leq 0$.

Let $\Psi(\theta, \gamma) = (\Psi_1(\theta, \gamma), \ldots, \Psi_T(\theta, \gamma))$, which is an $N \times T$ matrix. Noting that

$$\begin{aligned}
\sum_{t=1}^{T} \Psi_t(\theta, \gamma)' \mathbb{M}_\Lambda \Psi_t(\theta, \gamma) &= \operatorname{tr}[\Psi(\theta, \gamma)' \mathbb{M}_\Lambda \Psi(\theta, \gamma)] \\
&= \operatorname{tr}[(\mathbb{M}_{F^0} + \mathbb{P}_{F^0}) \Psi(\theta, \gamma)' \mathbb{M}_\Lambda \Psi(\theta, \gamma)] \\
&= \operatorname{tr}[\mathbb{M}_{F^0} \Psi(\theta, \gamma)' \mathbb{M}_\Lambda \Psi(\theta, \gamma)] \\
&\quad + \operatorname{tr}\big[(F^{0\prime} F^0)^{-1/2} F^{0\prime} \Psi(\theta, \gamma)' \mathbb{M}_\Lambda \Psi(\theta, \gamma) F^0 (F^{0\prime} F^0)^{-1/2}\big], \\
\sum_{t=1}^{T} f_t^{0\prime} \Lambda^{0\prime} \mathbb{M}_\Lambda \Lambda^0 f_t^0 &= \operatorname{tr}[(F^{0\prime} F^0)^{1/2} \Lambda^{0\prime} \mathbb{M}_\Lambda \Lambda^0 (F^{0\prime} F^0)^{1/2}], \text{ and} \\
\sum_{t=1}^{T} \Psi_t(\theta, \gamma)' \mathbb{M}_\Lambda \Lambda^0 f_t^0 &= \operatorname{tr}[(F^{0\prime} F^0)^{1/2} \Lambda^{0\prime} \mathbb{M}_\Lambda \Psi(\theta, \gamma) F^0 (F^{0\prime} F^0)^{-1/2}],
\end{aligned}$$

we have

$$\mathcal{L}_1(\theta, \Lambda, \gamma) = \operatorname{tr}[\mathbb{M}_{F^0} \Psi(\theta, \gamma)' \mathbb{M}_\Lambda \Psi(\theta, \gamma)] + \operatorname{tr}[\Xi(\theta, \gamma)' \mathbb{M}_\Lambda \Xi(\theta, \gamma)],$$

where $\Xi(\theta, \gamma) \equiv \Lambda^0 (F^{0\prime} F^0)^{1/2} + \Psi(\theta, \gamma) F^0 (F^{0\prime} F^0)^{-1/2}$. Let $\mathcal{B}(\Lambda, \gamma)$, $\mathcal{Z}(\Lambda, \gamma)$ and $\widetilde{\mathcal{Z}}(\Lambda, \gamma)$ be the notations introduced in Section 2.3. Let

$$\widetilde{\mathcal{B}}(\Lambda, \gamma) \equiv \frac{1}{NT} \mathcal{Z}(\Lambda, \gamma)' \widetilde{\mathcal{Z}}(\Lambda, \gamma) \text{ and } \breve{\mathcal{B}}(\Lambda, \gamma) \equiv \frac{1}{NT} \widetilde{\mathcal{Z}}(\Lambda, \gamma)' \widetilde{\mathcal{Z}}(\Lambda, \gamma).$$

Using the properties that $\operatorname{tr}(B_1 B_2 B_3) = \text{vec}(B_1)' (B_2 \otimes I) \text{vec}(B_3')$ and

$$\operatorname{tr}(B_1 B_2 B_3 B_4) = vec\,(B_1)' \left(B_2 \otimes B_4'\right) vec\left(B_3'\right)$$

for any conformable matrices B_1, B_2, B_3, B_4 and an identity matrix I (see, e.g., Bernstein (2005, p.253)), we have

$$\operatorname{tr}[\mathbb{M}_{F^0} \Psi(\theta, \gamma)' \mathbb{M}_\Lambda \Psi(\theta, \gamma)] = \text{vec}\big(\Psi(\theta, \gamma)\big)' (\mathbb{M}_{F^0} \otimes \mathbb{M}_\Lambda) \, \text{vec}\big(\Psi(\theta, \gamma)\big).$$

Noting that $(\mathbb{M}_{F^0} \otimes \mathbb{M}_\Lambda) \vec{(}\Psi(\theta, \gamma)) = -\mathcal{Z}(\Lambda, \gamma)(\theta - \theta^0) - \widetilde{\mathcal{Z}}(\Lambda, \gamma)\delta^0$ and $\mathbb{M}_{F^0} \otimes \mathbb{M}_\Lambda$ is a projection matrix, we have $\frac{1}{NT} \operatorname{tr}[\mathbb{M}_{F^0} \Psi(\theta, \gamma)' \mathbb{M}_\Lambda \Psi(\theta, \gamma)]$

$$\begin{aligned}
&= \big[\theta - \theta^0 + \mathcal{B}(\Lambda, \gamma)^{-1} \widetilde{\mathcal{B}}(\Lambda, \gamma) \delta^0\big]' \mathcal{B}(\Lambda, \gamma) \big[\theta - \theta^0 + \mathcal{B}(\Lambda, \gamma)^{-1} \widetilde{\mathcal{B}}(\Lambda, \gamma) \delta^0\big] \\
&\quad + \delta^{0\prime} \big[\breve{\mathcal{B}}(\Lambda, \gamma) - \widetilde{\mathcal{B}}(\Lambda, \gamma)' \mathcal{B}(\Lambda, \gamma)^{-1} \widetilde{\mathcal{B}}(\Lambda, \gamma)\big] \delta^0 \\
&\equiv \frac{1}{NT} \mathcal{L}_{1,1}(\theta, \Lambda, \gamma) + \frac{1}{NT} \mathcal{L}_{1,2}(\Lambda, \gamma), \text{ say,} \qquad \text{(A.3)}
\end{aligned}$$

where the invertibility of $\mathcal{B}(\Lambda, \gamma)$ is ensured by Assumption A.1(i). By quadratic form, we have $(NT)^{-1}\mathcal{L}_{1,1}(\theta,\Lambda,\gamma)\geq 0$. Noting that $(NT)^{-1}\mathcal{L}_{1,2}(\Lambda,\gamma)$ can be written as $\delta^{0\prime}\widetilde{\mathcal{Z}}(\Lambda,\gamma)'\mathbb{M}_{\mathcal{Z}(\Lambda,\gamma)}\widetilde{\mathcal{Z}}(\Lambda,\gamma)\delta^0$, we also have $(NT)^{-1}\mathcal{L}_{1,2}(\Lambda,\gamma)\geq 0$. It is easy to verify that $\mathcal{B}(\Lambda,\gamma)$, $\widetilde{\mathcal{B}}(\Lambda,\gamma)$ and $\breve{\mathcal{B}}(\Lambda,\gamma)$ are $O_p(1)$ uniformly in (Λ,γ). Noting that $\delta^0=o(1)$, $(NT)^{-1}\mathcal{L}_{1,2}(\Lambda,\gamma)\geq 0$ and $(NT)^{-1}\text{tr}[\Xi(\theta,\gamma)'\mathbb{M}_\Lambda\Xi(\theta,\gamma)]\geq 0$ for all (γ,Λ), the fact that $\frac{1}{NT}\mathcal{L}_1(\widehat{\theta},\widehat{\Lambda},\widehat{\gamma})+o_p(1)\leq 0$ implies

$$(\widehat{\theta}-\theta^0)'\mathcal{B}(\widehat{\Lambda},\widehat{\gamma})(\widehat{\theta}-\theta^0)+o_p(1)<0.$$

This implies that $\|\widehat{\theta}-\theta^0\|=o_p(1)$ by Assumption A.1(i).

(ii) Given $\|\widehat{\theta}-\theta^0\|=o_p(1)$, we can readily show that

$$\frac{1}{NT}\text{tr}[\Psi(\widehat{\theta},\widehat{\gamma})'\mathbb{M}_{\widehat{\Lambda}}\Psi(\widehat{\theta},\widehat{\gamma})]=o_p(1)\text{ and }\frac{1}{NT}\text{tr}[\Psi(\widehat{\theta},\widehat{\gamma})'\mathbb{M}_{\widehat{\Lambda}}\Lambda^0F^{0\prime}]=o_p(1).$$

These results, in conjunction with the fact that $\frac{1}{NT}\mathcal{L}_1(\widehat{\theta},\widehat{\Lambda},\widehat{\gamma})=o_p(1)$, implies that $\frac{1}{NT}\text{tr}[\widehat{\Xi}'\mathbb{M}_{\widehat{\Lambda}}\widehat{\Xi}]=o_p(1)$ where $\widehat{\Xi}\equiv\Xi(\widehat{\theta},\widehat{\gamma})$. It follows that

$$\frac{1}{NT}\text{tr}\left(F^0\Lambda^{0\prime}\mathbb{M}_{\widehat{\Lambda}}\Lambda^0F^{0\prime}\right)=\text{tr}\left(\frac{\Lambda^{0\prime}\mathbb{M}_{\widehat{\Lambda}}\Lambda^0}{N}\frac{F^{0\prime}F^0}{T}\right)=o_p(1).$$

Because $T^{-1}F^{0\prime}F^0>0$, the above equation implies that

$$\frac{\Lambda^{0\prime}\mathbb{M}_{\widehat{\Lambda}}\Lambda^0}{N}=\frac{\Lambda^{0\prime}\Lambda^0}{N}-\frac{\Lambda^{0\prime}\widehat{\Lambda}}{N}\frac{\widehat{\Lambda}'\Lambda^0}{N}=o_p(1).$$

Hence, $\widehat{\Lambda}'\Lambda^0/N$ is invertible and it follows that $I_R-\widehat{\Lambda}'\mathbb{P}_{\Lambda^0}\widehat{\Lambda}=o_p(1)$. Consequently, we have

$$\left\|\mathbb{P}_{\widehat{\Lambda}}-\mathbb{P}_{\Lambda^0}\right\|^2=\text{tr}[\left(\mathbb{P}_{\widehat{\Lambda}}-\mathbb{P}_{\Lambda^0}\right)^2]=2\text{tr}(I_R-\widehat{\Lambda}'\mathbb{P}_{\Lambda^0}\widehat{\Lambda}/N)=o_p\left(1\right).$$

This completes the proof of (ii). ■

Proof of Theorem 2.2

To prove Theorem 2.2, we need two propositions. In proposition A.1 and A.2, we establish preliminary convergence rates of $\widehat{\Lambda}$ and $\widehat{\theta}$ respectively. Given the rates, we go back to equation (A.2) and show the consistency of $\widehat{\gamma}$.

Proposition A.1. *Let* $H\equiv(F^{0\prime}F^0/T)(\Lambda^{0\prime}\widehat{\Lambda}/N)V_{NT}^{-1}$. *Under Assumptions A.1-A.4,*

(i) $V_{NT}\xrightarrow{p}V$, *where* V *is a diagonal matrix consisting of eigenvalues of* $\Sigma_{\Lambda^0}\Sigma_{F^0}$*;*

(ii) H *is invertible and* $\frac{1}{N}\|\widehat{\Lambda}-\Lambda^0H\|^2=O_p(\|\widehat{\theta}-\theta^0\|^2+\|\delta^0\|^2+\pi_{NT}^{-2})$.

Proof of Proposition A.1. Let $\tilde{e}_t \equiv Y_t - X_t\widehat{\beta} - X_t(\widehat{\gamma})\widehat{\delta} = e_t + \Lambda^0 f_t^0 - X_{t,\gamma^0}(\widehat{\theta} - \theta^0) + X_t(\gamma^0, \widehat{\gamma})\widehat{\delta}$. By the eigenvalue equation $\big(\frac{1}{NT}\sum_{t=1}^T \tilde{e}_t\tilde{e}_t'\big)\widehat{\Lambda} = \widehat{\Lambda}V_{NT}$, we can obtain the following decomposition

$$\begin{aligned}
\widehat{\Lambda} - \Lambda^0 H = \frac{1}{NT}\Bigg\{ & \mathbf{e}\mathbf{e}' + \mathbf{e}F^0\Lambda^{0\prime} + \Lambda^0F^{0\prime}\mathbf{e}' - \sum_{k=1}^{2K}(\widehat{\theta}_k - \theta_k^0)\mathbf{e}\mathbf{X}_{k,\gamma^0}' \\
& - \sum_{k=1}^{2K}(\widehat{\theta}_k - \theta_k^0)\mathbf{X}_{k,\gamma^0}\mathbf{e}' + \sum_{k=1}^{K}\widehat{\delta}_k\mathbf{e}\mathbf{X}_k(\gamma^0,\widehat{\gamma})' + \sum_{k=1}^{K}\widehat{\delta}_k\mathbf{X}_k(\gamma^0,\widehat{\gamma})\mathbf{e}' \\
& - \sum_{k=1}^{2K}(\widehat{\theta}_k - \theta_k^0)\Lambda^0F^{0\prime}\mathbf{X}_{k,\gamma^0}' - \sum_{k=1}^{2K}(\widehat{\theta}_k - \theta_k^0)\mathbf{X}_{k,\gamma^0}F^0\Lambda^{0\prime} \\
& + \sum_{k=1}^{K}\widehat{\delta}_k\Lambda^0F^{0\prime}\mathbf{X}_k(\gamma^0,\widehat{\gamma})' + \sum_{k=1}^{K}\widehat{\delta}_k\mathbf{X}_k(\gamma^0,\widehat{\gamma})F^0\Lambda^{0\prime} \\
& + \sum_{k=1}^{2K}\sum_{k'=1}^{2K}(\widehat{\theta}_k - \theta_k^0)(\widehat{\theta}_{k'} - \theta_{k'}^0)\mathbf{X}_{k,\gamma^0}\mathbf{X}_{k',\gamma^0}' + \sum_{k=1}^{K}\sum_{k'=1}^{K}\widehat{\delta}_k\widehat{\delta}_{k'}\mathbf{X}_k(\gamma^0,\widehat{\gamma})\mathbf{X}_{k'}(\gamma^0,\widehat{\gamma})' \\
& - \sum_{k=1}^{K}\sum_{k'=1}^{2K}\widehat{\delta}_k(\widehat{\theta}_{k'} - \theta_{k'}^0)\mathbf{X}_k(\gamma^0,\widehat{\gamma})\mathbf{X}_{k',\gamma^0}' - \sum_{k=1}^{2K}\sum_{k'=1}^{K}\widehat{\delta}_{k'}(\widehat{\theta}_k - \theta_{k'}^0)\mathbf{X}_{k,\gamma^0}\mathbf{X}_{k'}(\gamma^0,\widehat{\gamma})'\Bigg\}\widehat{\Lambda} \\
\equiv {} & (I_1 + \cdots + I_{15})\widehat{\Lambda}V_{NT}^{-1}, \qquad \text{(A.4)}
\end{aligned}$$

where $\mathbf{X}_k(\gamma_1,\gamma_2) \equiv \mathbf{X}_k(\gamma_1) - \mathbf{X}_k(\gamma_2)$ for $k = 1,\ldots,K$.

For I_1, we have $\|I_1\| = \frac{1}{NT}\|\mathbf{e}\mathbf{e}'\| = O_p(\pi_{NT}^{-1})$ by Lemma A.1(ii). For I_2 and I_3, we have

$$\|I_2\| = \|I_3\| = \frac{1}{NT}\|\mathbf{e}F^0\Lambda^{0\prime}\| \leq \frac{1}{\sqrt{T}}\frac{\|\mathbf{e}F^0\|}{\sqrt{NT}}\frac{\|\Lambda^0\|}{\sqrt{N}} = O_p(\frac{1}{\sqrt{T}}).$$

For I_4 and I_5, noting that $\|\mathbf{e}\mathbf{X}_{k,\gamma^0}'\| = O_p(N\sqrt{T})$ and $|\widehat{\theta}_k - \theta_k^0| \leq \|\widehat{\theta} - \theta^0\|$, we have $\|I_4\| = \|I_5\| = O_p(T^{-1/2}\|\widehat{\theta} - \theta^0\|)$. For I_5 and I_6, noting that $\|\mathbf{e}\mathbf{X}_k(\gamma^0,\widehat{\gamma})'\| \leq 2\sup_{\gamma\in\gamma}\|\mathbf{e}\mathbf{X}_k(\gamma)'\| = O_p(N\sqrt{T})$ by Lemma A.1(i), we have $\|I_6\| = \|I_7\| = O_p(T^{-1/2}\|\widehat{\delta}\|)$. For I_8 and I_9, we cam show that

$$\frac{1}{NT}\|\Lambda^0F^{0\prime}\mathbf{X}_{k,\gamma^0}'\| \leq \frac{\|\Lambda^0\|}{\sqrt{N}}\frac{\|F^0\|}{\sqrt{T}}\frac{\|\mathbf{X}_{k,\gamma^0}\|}{\sqrt{NT}} = O_p(1),$$

which implies that $\|I_8\| = \|I_9\| = O_p(\|\widehat{\theta} - \theta^0\|)$. Similarly, $\|I_{10}\| = \|I_{11}\| = O_p(\|\widehat{\delta}\|)$. For $I_{12},\ldots,I_{15}$, we can bound their Frobenius norm by $O_p(\|\widehat{\theta} - \theta^0\| + \|\widehat{\delta}\|)$. By the triangle inequality, $\|\widehat{\delta}\| \leq \|\delta^0\| + \|\widehat{\delta} - \delta^0\| \leq \|\delta^0\| + \|\widehat{\theta} - \theta^0\|$. It follows that $\|I_1 + \cdots + I_{15}\| = O_p(\|\widehat{\theta} - \theta^0\| + \|\delta^0\| + \pi_{NT}^{-1})$.

Premultiplying $\Lambda^{0\prime}/N$ on both sides of (A.4), we can obtain that

$$\frac{\Lambda^{0\prime}\widehat{\Lambda}}{N}V_{NT} = \frac{\Lambda^{0\prime}\Lambda^0}{N}\frac{F^{0\prime}F^0}{T}\frac{\Lambda^{0\prime}\widehat{\Lambda}}{N} + o_p(1).$$

Noting that both $\frac{\Lambda^{0\prime}\Lambda^0}{N}\frac{F^{0\prime}F^0}{T}$ and $\frac{\Lambda^{0\prime}\widehat{\Lambda}}{N}$ are asymptotically nonsingular matrices, the above equality shows that the columns of $\frac{\Lambda^{0\prime}\widehat{\Lambda}}{N}$ are the (non-normalized) eigenvectors of the matrix $\frac{\Lambda^{0\prime}\Lambda^0}{N}\frac{F^{0\prime}F^0}{T}$, and V_{NT} consists of the eigenvalues of the same matrix (in the limit). Thus, $V_{NT} \xrightarrow{p} V$ where V is a diagonal matrix consisting of eigenvalues of $\Sigma_{\Lambda^0}\Sigma_{F^0}$.

(ii) Noting that V_{NT} is invertible, we have

$$N^{-1/2}\|\widehat{\Lambda} - \Lambda^0 H\| = N^{-1/2}\|(I_1 + \cdots + I_{15})\widehat{\Lambda}V_{NT}^{-1}\|$$

Checking the terms one by one, we can readily show that $N^{-1/2}\|\widehat{\Lambda}-\Lambda^0 H\| = O_p(\|\widehat{\theta}-\theta^0\| + \|\delta^0\| + \pi_{NT}^{-1})$, or equivalently, $\frac{1}{N}\|\widehat{\Lambda} - \Lambda^0 H\|^2 = O_p(\|\widehat{\theta} - \theta^0\|^2 + \|\delta^0\|^2 + \pi_{NT}^{-2})$. ■

Lemma A.3. *Under Assumptions A.1-A.4,*

(i) $\frac{1}{N}\Lambda^{0\prime}(\widehat{\Lambda} - \Lambda^0 H) = O_p(\|\widehat{\theta} - \theta^0\| + \|\delta^0\| + \pi_{NT}^{-2})$;

(ii) $\frac{1}{N}\widehat{\Lambda}'(\widehat{\Lambda} - \Lambda^0 H) = O_p(\|\widehat{\theta} - \theta^0\| + \|\delta^0\| + \pi_{NT}^{-2})$;

(iii) $HH' = (\frac{1}{N}\Lambda^{0\prime}\Lambda^0)^{-1} + O_p(\|\widehat{\theta} - \theta^0\| + \|\delta^0\| + \pi_{NT}^{-2})$;

(iv) $\|\mathbb{P}_{\widehat{\Lambda}} - \mathbb{P}_{\Lambda^0}\|^2 = O_p(\|\widehat{\theta} - \theta^0\| + \|\delta^0\| + \pi_{NT}^{-2})$.

Lemma A.4. *Under Assumptions A.1-A.4, as $N, T \to \infty$ and $N/T \to \kappa > 0$,*

(i) $\frac{1}{\sqrt{NT}}\|\mathbf{e}'(\widehat{\Lambda} - \Lambda^0 H)\| = O_p(\|\widehat{\theta} - \theta^0\| + \|\delta^0\| + \pi_{NT}^{-1})$;

(ii) $\frac{1}{(NT)^{1-2\alpha}}\Big|\sum_{t=1}^{T}\widetilde{\Psi}_t(\widehat{\theta},\widehat{\gamma})'\mathbb{M}_{\widehat{\Lambda}}e_t\Big| = o_p(1)$,

(iii) $\frac{1}{(NT)^{1-2\alpha}}\Big|\sum_{t=1}^{T}\widetilde{\Psi}_t(\widehat{\theta},\widehat{\gamma})'\mathbb{M}_{\widehat{\Lambda}}\Lambda^0 f_t^0 - \mathcal{R}\Big| = o_p(1)$,

(iv) $\frac{1}{(NT)^{1-2\alpha}}\Big|\sum_{t=1}^{T} f_t^{0\prime}\Lambda^{0\prime}\mathbb{M}_{\widehat{\Lambda}}e_t + \mathrm{tr}[\mathbf{e}'\mathbb{M}_{\Lambda^0}\mathbf{e}\mathbb{P}_{F^0}]\Big| = o_p(1)$,

(v) $\frac{1}{(NT)^{1-2\alpha}}\Big|\sum_{t=1}^{T} e_t'(\mathbb{P}_{\widehat{\Lambda}} - \mathbb{P}_{\Lambda^0})e_t\Big| = o_p(1)$,

(vi) $\frac{1}{(NT)^{1-2\alpha}}\Big|\sum_{t=1}^{T} f_t^{0\prime}\Lambda^{0\prime}\mathbb{M}_{\widehat{\Lambda}}\Lambda^0 f_t^0 - \mathrm{tr}(\mathbf{e}'\mathbb{M}_{\Lambda^0}\mathbf{e}\mathbb{P}_{F^0}) - \mathcal{R}\Big| = o_p(1)$, *where* $\widetilde{\Psi}(\theta,\gamma) = X_{t,\gamma}(\theta-\theta^0) + X_t(\gamma,\gamma^0)\delta^0$ *and*

$$\begin{aligned}\mathcal{R} = &\sum_{k=1}^{2K}\sum_{k'=1}^{2K}(\widehat{\theta}_k-\theta_k^0)(\widehat{\theta}_{k'}-\theta_{k'}^0)\frac{1}{NT}\mathrm{tr}\big[\mathbf{X}_{k,\widehat{\gamma}}\mathbb{M}_{\widehat{\Lambda}}\mathbf{X}_{k',\widehat{\gamma}}\mathbb{P}_{F^0}\big]\\ &+\sum_{k=1}^{K}\sum_{k'=1}^{K}\delta_k^0\delta_{k'}^0\frac{1}{NT}\mathrm{tr}\big[\mathbf{X}_k(\widehat{\gamma},\gamma^0)\mathbb{M}_{\widehat{\Lambda}}\mathbf{X}_{k'}(\widehat{\gamma},\gamma^0)\mathbb{P}_{F^0}\big]\\ &+2\sum_{k=1}^{2K}\sum_{k'=1}^{K}(\widehat{\theta}_k-\theta_k^0)\delta_{k'}^0\frac{1}{NT}\mathrm{tr}\big[\mathbf{X}_{k,\widehat{\gamma}}\mathbb{M}_{\widehat{\Lambda}}\mathbf{X}_{k'}(\widehat{\gamma},\gamma^0)\mathbb{P}_{F^0}\big].\end{aligned}$$

Proposition A.2. *Suppose that Assumptions A.1-A.4 hold and* $N/T\to\kappa>0$. *Then* $\|\widehat{\theta}-\theta^0\| = O_p((NT)^{-\alpha})$.

Proof of Proposition A.2. By definition,

$$\widehat{\theta} = \Big(\sum_{t=1}^{T}X'_{t,\widehat{\gamma}}\mathbb{M}_{\widehat{\Lambda}}X_{t,\widehat{\gamma}}\Big)^{-1}\Big(\sum_{t=1}^{T}X'_{t,\widehat{\gamma}}\mathbb{M}_{\widehat{\Lambda}}Y_t\Big).$$

Substituting $Y_t = X_{t,\widehat{\gamma}}\theta^0 + X_t(\gamma^0,\widehat{\gamma})\delta^0 + \Lambda^0 f_t^0 + e_t$ into the above equation yields

$$(\frac{1}{NT}\sum_{t=1}^{T}X'_{t,\widehat{\gamma}}\mathbb{M}_{\widehat{\Lambda}}X_{t,\widehat{\gamma}})(\widehat{\theta}-\theta^0) = \frac{1}{NT}\sum_{t=1}^{T}X'_{t,\widehat{\gamma}}\mathbb{M}_{\widehat{\Lambda}}e_t + \frac{1}{NT}\sum_{t=1}^{T}X'_{t,\widehat{\gamma}}\mathbb{M}_{\widehat{\Lambda}}\Lambda^0 f_t^0 + \frac{1}{NT}\sum_{t=1}^{T}X'_{t,\widehat{\gamma}}\mathbb{M}_{\widehat{\Lambda}}X_t(\gamma^0,\widehat{\gamma})\delta^0. \tag{A.5}$$

Consider the first term on the right hand side (RHS) of (A.5). By

$$\begin{aligned}\mathbb{P}_{\widehat{\Lambda}} - \mathbb{P}_{\Lambda^0} = &\frac{1}{N}(\widehat{\Lambda}-\Lambda^0H)(\widehat{\Lambda}-\Lambda^0H)' + \frac{1}{N}(\widehat{\Lambda}-\Lambda^0H)H'^{0\prime}\\ &+\frac{1}{N}\Lambda^0H(\widehat{\Lambda}-\Lambda^0H)' + \frac{1}{N}\Lambda^0\Big[HH' - (\frac{\Lambda'\Lambda}{N})^{-1}\Big]\Lambda^{0\prime},\end{aligned} \tag{A.6}$$

the first term on the RHS of (A.5) is equal to

$$\begin{aligned}&\frac{1}{NT}\sum_{t=1}^{T}X'_{t,\widehat{\gamma}}\mathbb{M}_{\Lambda^0}e_t - \frac{1}{NT}\sum_{t=1}^{T}X'_{t,\widehat{\gamma}}\Big[\frac{1}{N}(\widehat{\Lambda}-\Lambda^0H)(\widehat{\Lambda}-\Lambda^0H)'\Big]e_t - \frac{1}{NT}\sum_{t=1}^{T}X'_{t,\widehat{\gamma}}\Big[\frac{1}{N}(\widehat{\Lambda}-\Lambda^0H)H'^{0\prime}\Big]e_t\\ &-\frac{1}{NT}\sum_{t=1}^{T}X'_{t,\widehat{\gamma}}\Big[\frac{1}{N}\Lambda^0H(\widehat{\Lambda}-\Lambda^0H)'\Big]e_t - \frac{1}{NT}\sum_{t=1}^{T}X'_{t,\widehat{\gamma}}\Big[\frac{1}{N}\Lambda^0\Big(HH'-(\frac{\Lambda'\Lambda}{N})^{-1}\Big)\Lambda^{0\prime}\Big]e_t = \mathit{II}_1 - \cdots - \mathit{II}_5, \text{say}\end{aligned}$$

Let $\mathit{II}_{l,k}$ be the kth element of II_l for $l = 1,\ldots,5$ and $k = 1,\ldots,2K$. Because K is a finite value, it suffices to consider $\mathit{II}_{l,k}$ for each k. Term $\mathit{II}_{1,k}$ is $O_p(\frac{1}{\sqrt{NT}})$ since $\sup_{\gamma\in\Gamma}|\frac{1}{\sqrt{NT}}\sum_{t=1}^{T}X'_{t,\gamma}\mathbb{M}_{\Lambda^0}e_t| = O_p(1)$. For $\mathit{II}_{2,k}$,

$$|\mathit{II}_{2,k}| \le \frac{\|\widehat{\Lambda}-\Lambda^0H\|^2}{N}\frac{\|\mathbf{e}\mathbf{X}'_{k,\widehat{\gamma}}\|}{NT} = \frac{1}{\sqrt{T}}O_p(\|\widehat{\theta}-\theta^0\|^2 + \|\delta^0\|^2 + \pi_{NT}^{-2})$$

since $\sup_{\gamma\in\Gamma}\frac{1}{NT}\|\mathbf{e}\mathbf{X}'_{k,\gamma}\| = O_p(\frac{1}{\sqrt{T}})$. For $\mathit{II}_{3,k}$,

$$|\mathit{II}_{3,k}| \le \frac{1}{\sqrt{N}}\frac{\|\widehat{\Lambda}-\Lambda^0H\|}{\sqrt{N}}\|H'\|\frac{\|\Lambda^{0\prime}\mathbf{e}\mathbf{X}'_{k,\widehat{\gamma}}\|}{NT} = O_p(\frac{1}{\sqrt{NT}})O_p(\|\widehat{\theta}-\theta^0\| + \|\delta^0\| + \pi_{NT}^{-1})$$

since $\sup_{\gamma\in\Gamma}\frac{1}{NT}\|\Lambda^{0\prime}\mathbf{e}\mathbf{X}'_{k,\gamma}\| = O_p(\frac{1}{\sqrt{T}})$. For $II_{4,k}$,

$$|II_{4,k}| \le \|H\|\frac{\|\mathbf{e}\mathbf{X}'_{k,\widehat{\gamma}}\|}{NT}\frac{\|\Lambda^0\|}{\sqrt{N}}\frac{\|\widehat{\Lambda}-\Lambda^0H\|}{\sqrt{N}} = O_p(\frac{1}{\sqrt{T}})O_p(\|\widehat{\theta}-\theta^0\| + \|\delta^0\| + \pi_{NT}^{-1}).$$

For $II_{5,k}$,

$$|II_{5,k}| \le \frac{1}{\sqrt{N}}\frac{\|\Lambda^0\|}{\sqrt{N}}\frac{\|\Lambda^{0\prime}\mathbf{e}\mathbf{X}'_{k,\widehat{\gamma}}\|}{NT}\|HH'-(\frac{\Lambda'\Lambda}{N})^{-1}\| = O_p(\frac{1}{\sqrt{NT}})O_p(\|\widehat{\theta}-\theta^0\|+\|\delta^0\|+\pi_{NT}^{-2}).$$

Summarizing all the above results, we have

$$\frac{1}{NT}\sum_{t=1}^{T} X'_{t,\widehat{\gamma}}\mathbb{M}_{\widehat{\Lambda}}e_t = o_p(\|\widehat{\theta}-\theta^0\| + \|\delta^0\|) + O_p(\pi_{NT}^{-2}). \tag{A.7}$$

Next, we consider the second term on the right hand side of (A.5). By the fact that $\mathbb{M}_{\widehat{\Lambda}}\Lambda^0 = \mathbb{M}_{\widehat{\Lambda}}(\Lambda^0 - \widehat{\Lambda}H^{-1})$ and equation (A.4), we can write the kth entry of $\frac{1}{NT}\sum_{t=1}^{T} X'_{t,\widehat{\gamma}}\mathbb{M}_{\widehat{\Lambda}}\Lambda^0 f_t^0$ as

$$\begin{aligned}\frac{1}{NT}\mathrm{tr}\Big[(\Lambda^0-\widehat{\Lambda}H^{-1})'\mathbb{M}_{\widehat{\Lambda}}\mathbf{X}_{k,\widehat{\gamma}}F^0\Big] &= -\frac{1}{NT}\mathrm{tr}\Big[(I_1+\cdots+I_{15})'\mathbb{M}_{\widehat{\Lambda}}\mathbf{X}_{k,\widehat{\gamma}}F^0G'\widehat{\Lambda}'\Big]\\ &= -(J_{1,k}+\cdots+J_{15,k}), \qquad \text{say.}\end{aligned} \tag{A.8}$$

where $G = (N^{-1}\Lambda^{0\prime}\widehat{\Lambda})^{-1}(T^{-1}F^{0\prime}F^0)^{-1}$ and $J_{1,k},\ldots,J_{15,k}$ are implicitly defined in the above expression. For $J_{1,k}$, we use Lemma A.1(ii) and the fact that $\widehat{\Lambda} = \Lambda^0H + (\widehat{\Lambda}-\Lambda^0H)$ to obtain

$$\begin{aligned}|J_{1,k}| &= \frac{1}{N^2T^2}\Big|\mathrm{tr}(\widehat{\Lambda}'\mathbf{e}\mathbf{e}'\mathbb{M}_{\widehat{\Lambda}}\mathbf{X}_{k,\widehat{\gamma}}F^0G')\Big|\\ &\le \Big\{\frac{1}{\sqrt{N}}\|H\|\frac{\|\Lambda^{0\prime}\mathbf{e}\mathbf{e}'\|}{NT} + \frac{\|\widehat{\Lambda}-\Lambda^0H\|}{\sqrt{N}}\frac{\|\mathbf{e}\mathbf{e}'\|}{NT}\Big\}\frac{\|\mathbb{M}_{\widehat{\Lambda}}\mathbf{X}_{k,\widehat{\gamma}}\|}{\sqrt{NT}}\frac{\|F^0\|}{\sqrt{T}}\|G\|\\ &= O_p(\frac{1}{\sqrt{N}\pi_{NT}}) + O_p(\pi_{NT}^{-1}\|\widehat{\theta}-\theta^0\| + \pi_{NT}^{-1}\|\delta^0\| + \pi_{NT}^{-2}).\end{aligned}$$

For $J_{2,k}$, we have

$$J_{2,k} = \frac{1}{N^2T^2}\mathrm{tr}\Big(\mathbf{e}'\mathbb{M}_{\widehat{\Lambda}}\mathbf{X}_{k,\widehat{\gamma}}F^0G'\widehat{\Lambda}'\Lambda^0F^{0\prime}\Big) = \frac{1}{NT}\mathrm{tr}(\mathbf{e}'\mathbb{M}_{\widehat{\Lambda}}\mathbf{X}_{k,\widehat{\gamma}}\mathbb{P}_{F^0}).$$

Then we have

$$|J_{2,k}| \le \frac{1}{NT}\Big|\mathrm{tr}(\mathbf{e}'\mathbb{M}_{\Lambda^0}\mathbf{X}_{k,\widehat{\gamma}}\mathbb{P}_{F^0})\Big| + \frac{1}{NT}\Big|\mathrm{tr}[\mathbf{e}'(\mathbb{P}_{\Lambda^0}-\mathbb{P}_{\widehat{\Lambda}})\mathbf{X}_{k,\widehat{\gamma}}\mathbb{P}_{F^0}]\Big|.$$

The first term is $O_p(\frac{1}{\sqrt{NT}})$ since $\sup_{\gamma\in\Gamma}|\frac{1}{\sqrt{NT}}\mathrm{tr}(\mathbf{e}'\mathbb{M}_{\Lambda^0}\mathbf{X}_{k,\gamma}\mathbb{P}_{F^0})| = O_p(1)$. For the second term, by the expression of $\mathbb{P}_{\widehat{\Lambda}} - \mathbb{P}_{\Lambda^0}$, it is equal to

$$
\begin{aligned}
&-\frac{1}{NT}\mathrm{tr}\Big[\frac{1}{N}(\widehat{\Lambda}-\Lambda^0 H)(\widehat{\Lambda}-\Lambda^0 H)'\mathbf{X}_{k,\widehat{\gamma}}\mathbb{P}_{F^0}\mathbf{e}'\Big] - \frac{1}{NT}\mathrm{tr}\Big[\frac{1}{N}\Lambda^0 H(\widehat{\Lambda}-\Lambda^0 H)'\mathbf{X}_{k,\widehat{\gamma}}\mathbb{P}_{F^0}\mathbf{e}'\Big] \\
&-\frac{1}{NT}\mathrm{tr}\Big[\frac{1}{N}(\widehat{\Lambda}-\Lambda^0 H)H'^{0\prime}\mathbf{X}_{k,\widehat{\gamma}}\mathbb{P}_{F^0}\mathbf{e}'\Big] - \frac{1}{NT}\mathrm{tr}\Big[\frac{1}{N}\Lambda^0(HH' - \frac{1}{N}(\Lambda^{0\prime}\Lambda^0)^{-1})\Lambda^{0\prime}\mathbf{X}_{k,\widehat{\gamma}}\mathbb{P}_{F^0}\mathbf{e}'\Big]
\end{aligned}
$$

We use $II_{6,k},\ldots,II_{9,k}$ to denote the above four terms. For $II_{6,k}$,

$$
|II_{6,k}| \le \frac{\|\widehat{\Lambda}-\Lambda^0 H\|^2}{N}\frac{\|\mathbf{X}_{k,\widehat{\gamma}}\mathbb{P}_{F^0}\mathbf{e}'\|}{NT} = \frac{1}{\sqrt{T}}O_p(\|\widehat{\theta}-\theta^0\|^2 + \|\delta^0\|^2 + \pi_{NT}^{-2})
$$

since $\sup_{\gamma\in\Gamma}\frac{1}{NT}\|\mathbf{X}_{k,\gamma}\mathbb{P}_{F^0}\mathbf{e}'\| = O_p(\frac{1}{\sqrt{T}})$. For $II_{7,k}$,

$$
|II_{7,k}| \le \frac{1}{\sqrt{NT}}\|H\|\frac{\|\mathbf{X}_{k,\hat{\gamma}}F^0\|}{\sqrt{NT}}T\|(F^{0\prime}F^0)^{-1}\|\frac{\|F^{0\prime}e'^0\|}{\sqrt{NT}}\frac{\|\widehat{\Lambda}-\Lambda^0 H\|}{\sqrt{N}} = \frac{1}{\sqrt{NT}}O_p(\|\widehat{\theta}-\theta^0\| + \|\delta^0\| + \pi_{NT}^{-1})
$$

by $\sup_{\gamma\in\Gamma}\frac{1}{\sqrt{NT}}\|\mathbf{X}_{k,\gamma}F^0\| \le \|\mathbf{X}_k F^0\| = O_p(1)$. For $II_{8,k}$,

$$
|II_{8,k}| \le \frac{1}{\sqrt{T}}\|H\|\frac{\|\Lambda^0\|}{\sqrt{N}}\frac{\|\mathbf{X}_{k,\hat{\gamma}}F^0\|}{\sqrt{NT}}T\|(F^{0\prime}F^0)^{-1}\|\frac{\|F^{0\prime}e'\|}{\sqrt{NT}}\frac{\|\widehat{\Lambda}-\Lambda^0 H\|}{\sqrt{N}} = \frac{1}{\sqrt{T}}O_p(\|\widehat{\theta}-\theta^0\| + \|\delta^0\| + \pi_{NT}^{-1}).
$$

For $II_{9,k}$,

$$
|II_{9,k}| \le \frac{1}{\sqrt{T}}\frac{\|\Lambda^0\|}{\sqrt{N}}\frac{\|\mathbf{X}_{k,\hat{\gamma}}F^0\|}{\sqrt{NT}}T\|(F^{0\prime}F^0)^{-1}\|\frac{\|F^{0\prime}e'^0\|}{\sqrt{NT}}\Big\|HH' - \frac{1}{N}(\Lambda^{0\prime}\Lambda^0)^{-1}\Big\| = \frac{1}{\sqrt{T}}O_p(\|\widehat{\theta}-\theta^0\|+\|\delta^0\|+\pi_{NT}^{-2}).
$$

Summarizing the above results, we have $|J_{2,k}| = o_p(\|\widehat{\theta}-\theta^0\| + \|\delta^0\|) + O_p(\pi_{NT}^{-2})$. For $J_{3,k}$, we have

$$
\begin{aligned}
|J_{3,k}| &= \left|\frac{1}{N^2T^2}\mathrm{tr}\Big(\widehat{\Lambda}'\mathbf{e}F^0(\Lambda^0 - \widehat{\Lambda}H^{-1})'\mathbb{M}_{\widehat{\Lambda}}\mathbf{X}_{k,\widehat{\gamma}}F^0G'\Big)\right| \\
&\le \frac{1}{\sqrt{NT}}\frac{\|\widehat{\Lambda}'\mathbf{e}F^0\|}{\sqrt{NT}}\frac{\|\widehat{\Lambda}-\Lambda^0 H\|}{\sqrt{N}}\|H^{-1}\|\frac{\|\mathbb{M}_{\widehat{\Lambda}}\mathbf{X}_{k,\widehat{\gamma}}\|}{\sqrt{NT}}\frac{\|F^0G'\|}{\sqrt{T}} \\
&= \frac{1}{\sqrt{NT}}O_p(\|\widehat{\theta}-\theta^0\| + \|\delta^0\| + \pi_{NT}^{-1}) + \frac{1}{\sqrt{T}}O_p(\|\widehat{\theta}-\theta^0\|^2 + \|\delta^0\|^2 + \pi_{NT}^{-2}),
\end{aligned}
$$

where we use the fact that

$$
\frac{\|\widehat{\Lambda}'\mathbf{e}F^0\|}{\sqrt{NT}} \le \frac{\|H\|\|\Lambda^{0\prime}\mathbf{e}F^0\|}{\sqrt{NT}} + \sqrt{N}\frac{\|\widehat{\Lambda}-\Lambda^0 H\|}{\sqrt{N}}\frac{\|\mathbf{e}F^0\|}{\sqrt{NT}} = O_p(1) + \sqrt{N}O_p(\|\widehat{\theta}-\theta^0\|+\|\delta^0\|+\pi_{NT}^{-1}).
$$

Next, we consider $J_{4,k}$. Note that

$$
J_{4,k} = \sum_{k'=1}^{2K}(\widehat{\theta}_{k'} - \theta^0_{k'})\cdot\frac{1}{N^2T^2}\mathrm{tr}\Big(\mathbf{X}_{k',\gamma^0}\mathbf{e}'\mathbb{M}_{\widehat{\Lambda}}\mathbf{X}_{k,\widehat{\gamma}}F^0G'\widehat{\Lambda}'\Big) \equiv \sum_{k'=1}^{2K}(\widehat{\theta}_{k'} - \theta^0_{k'})\cdot J_{4,k}(k').
$$

We can show that $|J_{4,k}(k')|$ is bounded by

$$\frac{1}{\sqrt{T}}\frac{\|\mathbf{X}_{k',\gamma^0}\mathbf{e}'\|}{N\sqrt{T}}\frac{\|\mathbf{X}_{k,\widehat{\gamma}}\|}{\sqrt{NT}}\frac{\|F^0G'\widehat{\Lambda}'\|}{\sqrt{NT}}=O_p(\frac{1}{\sqrt{T}}).$$

It follows that $J_{4,k}=o_p(\|\widehat{\theta}-\theta^0\|)$. By the same token, we can show that $J_{5,k}=o_p(\|\widehat{\theta}-\theta^0\|)$, $J_{6,k}=o_p(\|\widehat{\delta}\|)=o_p(\|\delta^0\|+\|\widehat{\theta}-\theta^0\|)$ and $J_{7,k}=o_p(\|\delta^0\|+\|\widehat{\theta}-\theta^0\|)$. Next, we consider $J_{8,k}$. Note that

$$J_{8,k}=\sum_{k'=1}^{2K}(\widehat{\theta}_{k'}-\theta^0_{k'})\cdot\frac{1}{N^2T^2}\mathrm{tr}\left[\mathbf{X}_{k',\gamma^0}F^0(\Lambda^0-\widehat{\Lambda}H^{-1})'\mathbb{M}_{\widehat{\Lambda}}\mathbf{X}_{k,\widehat{\gamma}}F^0G'\widehat{\Lambda}'\right]\equiv\sum_{k'=1}^{2K}(\widehat{\theta}_{k'}-\theta^0_{k'})\cdot J_{8,k}(k'),$$

where

$$\begin{aligned}|J_{8,k}(k')|&\le\frac{\|\mathbf{X}_{k',\gamma^0}\|}{\sqrt{NT}}\frac{\|F^0\|}{\sqrt{T}}\frac{\|\widehat{\Lambda}-\Lambda^0H\|\|H^{-1}\|}{\sqrt{N}}\frac{\|\mathbf{X}_{k,\widehat{\gamma}}\|}{\sqrt{NT}}\frac{\|F^0G'\widehat{\Lambda}'\|}{\sqrt{NT}}\\&=O_p(\|\widehat{\theta}-\theta^0\|+\|\delta^0\|+\pi^{-1}_{NT}).\end{aligned}$$

Then $J_{8,k}=o_p(\|\widehat{\theta}-\theta^0\|)$. By the same token, we can show that $J_{10,k}=o_p(\|\widehat{\theta}-\theta^0\|+\|\delta^0\|)$. For $J_{9,k}$, we have that

$$J_{9,k}=\frac{1}{N^2T^2}\sum_{k'=1}^{2K}(\widehat{\theta}_k-\theta^0_k)\mathrm{tr}[\Lambda^0F^0\mathbf{X}'_{k',\gamma^0}\mathbb{M}_{\widehat{\Lambda}}\mathbf{X}_{k,\widehat{\gamma}}F^0G'\widehat{\Lambda}']=\frac{1}{NT}\sum_{k'=1}^{2K}(\widehat{\theta}_{k'}-\theta^0_{k'})\mathrm{tr}\Big[\mathbf{X}'_{k',\gamma^0}\mathbb{M}_{\widehat{\Lambda}}\mathbf{X}_{k,\widehat{\gamma}}\mathbb{P}_{F^0}\Big].$$

It follows that $J_{9,k}=O_p(\|\widehat{\theta}-\theta^0\|)$. For $J_{11,k}$, we have that

$$J_{11,k}=\frac{1}{NT}\sum_{k'=1}^{K}\widehat{\delta}_{k'}\mathrm{tr}\big[\mathbf{X}_{k'}(\gamma^0,\widehat{\gamma})\mathbb{M}_{\widehat{\Lambda}}\mathbf{X}_{k,\widehat{\gamma}}\mathbb{P}_{F^0}\big].$$

It follows that $J_{11,k}=O_p(\|\widehat{\theta}-\theta^0\|+\|\delta^0\|)$. For the terms $J_{12,k},\ldots,J_{15,k}$, we can readily show that they are all $o_p(\|\widehat{\theta}-\theta^0\|+\|\delta^0\|)$. Given the above results, with some manipulations on the terms $J_{9,k}$ and $J_{11,k}$, we have

$$\begin{aligned}\frac{1}{NT}\sum_{t=1}^{T}X'_{t,\widehat{\gamma}}\mathbb{M}_{\widehat{\Lambda}}\Lambda^0f^0_t&=\Big(\frac{1}{NT}\sum_{t=1}^{T}X'_{t,\widehat{\gamma}}\mathbb{M}_{\widehat{\Lambda}}\sum_{s=1}^{T}X_{s,\widehat{\gamma}}a_{st}\Big)(\widehat{\theta}-\theta)+\Big(\frac{1}{NT}\sum_{t=1}^{T}X'_{t,\widehat{\gamma}}\mathbb{M}_{\widehat{\Lambda}}\sum_{s=1}^{T}X_s(\widehat{\gamma},\gamma^0)a_{st}\Big)\delta^0\\&\quad+O_p(\pi^{-2}_{NT})+o_p(\|\widehat{\theta}-\theta^0\|+\|\delta^0\|).\end{aligned}\tag{A.9}$$

Substituting (A.7) and (A.9) into (A.5), we have

$$[\mathcal{B}(\widehat{\Lambda},\widehat{\gamma})+o_p(1)](\widehat{\theta}-\theta^0)=-[\widetilde{\mathcal{B}}(\widehat{\gamma},\widehat{\Lambda})+o_p(1)]\delta^0+O_p(\pi^{-2}_{NT}).\tag{A.10}$$

Multiplying $(NT)^{\alpha}$ on both sides, by Assumptions A.1(i) and Assumption A.3, we have $(NT)^{\alpha}\widehat{\theta}=O_p(1)$. ■

Proof of Theorem 2.2. We use Lemma A.4 (ii)-(vi) to prove this theorem. Consider the objective function $\mathcal{L}(\widehat{\theta}, \widehat{\Lambda}, \widehat{\gamma})$. By definition, we have

$$\begin{aligned}
\frac{1}{(NT)^{1-2\alpha}}\mathcal{L}(\widehat{\theta}, \widehat{\Lambda}, \widehat{\gamma}) &= \frac{1}{(NT)^{1-2\alpha}}\sum_{t=1}^{T}\Big[\Lambda^0 f_t^0 + e_t - \widetilde{\Psi}_t(\widehat{\theta}, \widehat{\gamma})\Big]'\mathbb{M}_{\widehat{\Lambda}}\Big[\Lambda^0 f_t^0 + e_t - \widetilde{\Psi}_t(\widehat{\theta}, \widehat{\gamma})\Big] \\
&= \frac{1}{(NT)^{1-2\alpha}}\sum_{t=1}^{T}\widetilde{\Psi}_t(\widehat{\theta}, \widehat{\gamma})'\mathbb{M}_{\widehat{\Lambda}}\widetilde{\Psi}_t(\widehat{\theta}, \widehat{\gamma}) + \frac{1}{(NT)^{1-2\alpha}}\sum_{t=1}^{T}e_t'\mathbb{M}_{\widehat{\Lambda}}e_t \\
&\quad + \frac{1}{(NT)^{1-2\alpha}}\sum_{t=1}^{T}f_t^{0\prime}\Lambda^{0\prime}\mathbb{M}_{\widehat{\Lambda}}\Lambda^0 f_t^0 - 2\frac{1}{(NT)^{1-2\alpha}}\sum_{t=1}^{T}\widetilde{\Psi}_t(\widehat{\theta}, \widehat{\gamma})'\mathbb{M}_{\widehat{\Lambda}}\Lambda^0 f_t^0 \\
&\quad - 2\frac{1}{(NT)^{1-2\alpha}}\sum_{t=1}^{T}\widetilde{\Psi}_t(\widehat{\theta}, \widehat{\gamma})'\mathbb{M}_{\widehat{\Lambda}}e_t + 2\frac{1}{(NT)^{1-2\alpha}}\sum_{t=1}^{T}f_t^{0\prime}\Lambda^{0\prime}\mathbb{M}_{\widehat{\Lambda}}e_t,
\end{aligned}$$

where $\widetilde{\Psi}_t(\theta, \gamma)$ is defined in Lemma A.4. Using results (a)-(e), together with the fact that

$$\begin{aligned}
\frac{1}{(NT)^{1-2\alpha}}\sum_{t=1}^{T}\widetilde{\Psi}_t(\widehat{\theta}, \widehat{\gamma})'\mathbb{M}_{\widehat{\Lambda}}\widetilde{\Psi}_t(\widehat{\theta}, \widehat{\gamma}) &= \sum_{k=1}^{2K}\sum_{k'=1}^{2K}(\widehat{\theta}_k - \theta_k^0)(\widehat{\theta}_{k'} - \theta_{k'}^0)\frac{1}{NT}\mathrm{tr}\Big[\mathbf{X}_{k,\widehat{\gamma}}\mathbb{M}_{\widehat{\Lambda}}\mathbf{X}_{k',\widehat{\gamma}}\Big] \\
&\quad + \sum_{k=1}^{K}\sum_{k'=1}^{K}\delta_k^0\delta_{k'}^0\frac{1}{NT}\mathrm{tr}\Big[\mathbf{X}_k(\widehat{\gamma}, \gamma^0)\mathbb{M}_{\widehat{\Lambda}}\mathbf{X}_{k'}(\widehat{\gamma}, \gamma^0)\Big] \\
&\quad + 2\sum_{k=1}^{2K}\sum_{k'=1}^{K}(\widehat{\theta}_k - \theta_k^0)\delta_{k'}^0\frac{1}{NT}\mathrm{tr}\Big[\mathbf{X}_{k,\widehat{\gamma}}\mathbb{M}_{\widehat{\Lambda}}\mathbf{X}_{k'}(\widehat{\gamma}, \gamma^0)\Big]
\end{aligned}$$

we have

$$\begin{aligned}
\frac{1}{(NT)^{1-2\alpha}}\mathcal{L}(\widehat{\theta}, \widehat{\Lambda}, \widehat{\gamma}) &= \sum_{k=1}^{2K}\sum_{k'=1}^{2K}(\widehat{\theta}_k - \theta_k^0)(\widehat{\theta}_{k'} - \theta_{k'}^0)\frac{1}{NT}\mathrm{tr}\Big[\mathbf{X}_{k,\widehat{\gamma}}\mathbb{M}_{\widehat{\Lambda}}\mathbf{X}_{k',\widehat{\gamma}}\mathbb{M}_{F^0}\Big] \\
&\quad + \sum_{k=1}^{K}\sum_{k'=1}^{K}\delta_k^0\delta_{k'}^0\frac{1}{NT}\mathrm{tr}\Big[\mathbf{X}_k(\widehat{\gamma}, \gamma^0)\mathbb{M}_{\widehat{\Lambda}}\mathbf{X}_{k'}(\widehat{\gamma}, \gamma^0)\mathbb{M}_{F^0}\Big] \\
&\quad + 2\sum_{k=1}^{2K}\sum_{k'=1}^{K}(\widehat{\theta}_k - \theta_k^0)\delta_{k'}^0\frac{1}{NT}\mathrm{tr}\Big[\mathbf{X}_{k,\widehat{\gamma}}\mathbb{M}_{\widehat{\Lambda}}\mathbf{X}_{k'}(\widehat{\gamma}, \gamma^0)\mathbb{M}_{F^0}\Big] \\
&\quad + \frac{1}{(NT)^{1-2\alpha}}\mathrm{tr}\Big[\mathbf{e}'\mathbb{M}_{\Lambda^0}\mathbf{e}\mathbb{M}_{F^0}\Big] + o_p(1).
\end{aligned}$$

Let $\widehat{\theta}_{\gamma^0}$ and $\widehat{\Lambda}_{\gamma^0}$ be the estimator defined by

$$(\widehat{\theta}_{\gamma^0}, \widehat{\Lambda}_{\gamma^0}) = \underset{(\theta,\Lambda)\in\Theta\times\mathbb{L}}{\operatorname{argmax}}\ \mathcal{L}(\theta, \Lambda, \gamma^0).$$

Note that $\widehat{\theta}_{\gamma^0}$ is the least squares estimator for a standard IFEs model with known γ^0. According to Bai (2009) and Moon and Weidner (2017), $\widehat{\theta}_{\gamma^0} - \theta^0 = O_p(\pi_{NT}^{-2})$, or equivalently $(NT)^{\alpha}(\widehat{\theta}_{\gamma^0} - \theta^0) = o_p(1)$ under $N/T \to \kappa$. Given this, with the same arguments in deriving the above expression, we have

$$\frac{1}{(NT)^{1-2\alpha}}\mathcal{L}(\widehat{\theta}_{\gamma^0}, \widehat{\Lambda}_{\gamma^0}, \gamma^0) = \frac{1}{(NT)^{1-2\alpha}}\text{tr}\Big[\mathbf{e}'\mathbb{M}_{\Lambda^0}\mathbf{e}\mathbb{M}_{F^0}\Big] + o_p(1).$$

However, we also have $\mathcal{L}(\widehat{\theta}, \widehat{\Lambda}, \widehat{\gamma}) \le \mathcal{L}(\widehat{\theta}_{\gamma^0}, \widehat{\Lambda}_{\gamma^0}, \gamma^0)$ due to the definition of $(\widehat{\theta}, \widehat{\Lambda}, \widehat{\gamma})$. Given this result, with some algebra manipulation on the first three terms of $\mathcal{L}(\widehat{\theta}, \widehat{\Lambda}, \widehat{\gamma})$, we have

$$\begin{aligned}\Big[(NT)^{\alpha}(\widehat{\theta} - \theta^0) + \mathcal{B}(\widehat{\Lambda}, \widehat{\gamma})^{-1}\widetilde{\mathcal{B}}(\widehat{\Lambda}, \widehat{\gamma})C^0\Big]'\mathcal{B}(\widehat{\Lambda}, \widehat{\gamma})\Big[(NT)^{\alpha}(\widehat{\theta} - \theta^0) + \mathcal{B}(\widehat{\Lambda}, \widehat{\gamma})^{-1}\widetilde{\mathcal{B}}(\widehat{\Lambda}, \widehat{\gamma})C^0\Big]& \\ + C^{0\prime}\Big[\breve{\mathcal{B}}(\widehat{\Lambda}, \widehat{\gamma}) - \widetilde{\mathcal{B}}(\widehat{\Lambda}, \widehat{\gamma})'\mathcal{B}(\widehat{\Lambda}, \widehat{\gamma})^{-1}\widetilde{\mathcal{B}}(\widehat{\Lambda}, \widehat{\gamma})\Big]C^0 \le o_p(1).&\end{aligned}$$

Noting that (l, k)th entry of $\mathcal{B}(\gamma, \widehat{\Lambda})$ is $\frac{1}{NT}\text{tr}[\mathbf{X}_{l,\gamma}'\mathbb{M}_{\widehat{\Lambda}}\mathbf{X}_{k,\gamma}\mathbb{M}_{F^0}]$, we have

$$\begin{aligned}\sup_{\gamma\in\Gamma}\Big\|\mathcal{B}(\gamma, \widehat{\Lambda}) - \mathcal{B}(\gamma, \Lambda^0)\Big\| &\le \Big[\sum_{l,k=1}^{2K}\Big(\frac{1}{NT}\text{tr}[\mathbb{M}_{F^0}\mathbf{X}_{l,\gamma}'(\mathbb{M}_{\widehat{\Lambda}} - \mathbb{M}_{\Lambda^0})\mathbf{X}_{k,\gamma}]\Big)^2\Big]^{1/2} \\ &\le \Big[\sum_{l,k=1}^{2K}\Big(\frac{1}{NT}\|\mathbb{M}_{F^0}\mathbf{X}_{l,\gamma}'\|\|\mathbf{X}_{k,\gamma}\|\|\mathbb{P}_{\widehat{\Lambda}} - \mathbb{P}_{\Lambda^0}\|\Big)^2\Big]^{1/2} \\ &= O_p(1)\cdot\|\mathbb{P}_{\widehat{\Lambda}} - \mathbb{P}_{\Lambda^0}\| = o_p(1),\end{aligned}$$

since $\sup_{\gamma\in\Gamma}\|\mathbf{X}_{k,\gamma}\|/\sqrt{NT} \le \|\mathbf{X}_k\|/\sqrt{NT} = O_p(1)$. Similarly we have that

$$\sup_{\gamma\in\Gamma}\Big\|\widetilde{\mathcal{B}}(\gamma, \widehat{\Lambda}) - \widetilde{\mathcal{B}}(\gamma, \Lambda^0)\Big\| = o_p(1), \text{ and } \Big\|\breve{\mathcal{B}}(\gamma, \widehat{\Lambda}) - \breve{\mathcal{B}}(\gamma, \Lambda^0)\Big\| = o_p(1).$$

Hence we can get

$$C^{0\prime}\Big[\breve{\mathcal{B}}(\widehat{\gamma}, \Lambda^0) - \widetilde{\mathcal{B}}(\widehat{\gamma}, \Lambda^0)'\mathcal{B}(\widehat{\gamma}, \Lambda^0)^{-1}\widetilde{\mathcal{B}}(\widehat{\gamma}, \Lambda^0)\Big]C^0 = o_p(1).$$

By Assumption A.1(ii), we have

$$o_p(1) = C^{0\prime}\Big[\breve{\mathcal{B}}(\widehat{\gamma}, \Lambda^0) - \widetilde{\mathcal{B}}(\widehat{\gamma}, \Lambda^0)'\mathcal{B}(\widehat{\gamma}, \Lambda^0)^{-1}\widetilde{\mathcal{B}}(\widehat{\gamma}, \Lambda^0)\Big]C^0 = C^{0\prime}\mathcal{I}(\widehat{\gamma})C^0 \ge \|C^0\|^2\tau\min[1, |\widehat{\gamma} - \gamma^0|],$$

for some constant $\tau > 0$ with probability approaching 1. Hence, we must have $|\widehat{\gamma} - \gamma^0| = o_p(1)$. ■

Proposition A.3. *Suppose that Assumptions A.1-A.4 hold and $N/T \to \kappa$. Then we have*

(i) $\|\widehat{\theta} - \theta^0\| = o_p((NT)^{-\alpha})$;

(ii) For Proposition A.1 and Lemmas A.3-A.4, we can strengthen $O_p(\|\widehat{\theta} - \theta^0\| + \|\delta^0\|)$ *to* $o_p((NT)^{-\alpha})$.

Proof of Proposition A.3. (i) Revisit (A.10). We have shown above that $\sup_{\gamma\in\Gamma}\|\mathcal{B}(\widehat{\Lambda}, \gamma) - \mathcal{B}(\Lambda^0, \gamma)\| = o_p(1)$ and $\sup_{\gamma\in\Gamma}\|\widetilde{\mathcal{B}}(\widehat{\Lambda}, \gamma) - \widetilde{\mathcal{B}}(\Lambda^0, \gamma)\| = o_p(1)$. However, we also have $\mathcal{B}(\Lambda^0, \widehat{\gamma}) \xrightarrow{p} \mathcal{B}(\Lambda^0, \gamma^0)$ and $\widetilde{\mathcal{B}}(\Lambda^0, \widehat{\gamma}) \xrightarrow{p} \widetilde{\mathcal{B}}(\Lambda^0, \gamma^0) = 0$ due to the continuous mapping theorem and the definition of $\widetilde{\mathcal{B}}$. Given this, we have that $\mu_{\min}(\mathcal{B}(\widehat{\Lambda}, \gamma)) \geq \tau + o_p(1)$ and $\widetilde{\mathcal{B}}(\widehat{\Lambda}, \widehat{\gamma}) = o_p(1)$. These two results, together with (A.10), give (i).

(ii) Note that in the previous analysis, all the terms involving δ^0 or $\widehat{\delta}$ must include the symbol $\mathbf{X}_k(\gamma^0, \widehat{\gamma})$. Given the consistency of $\widehat{\gamma}$, the "O_p" terms now change to "o_p" terms. This result, together with $(NT)^{\alpha}(\widehat{\delta} - \delta^0) = o_p(1)$, leads to (ii). ■

Proof of Theorem 2.3

Let $h_{it}(\gamma_1, \gamma_2) \equiv \|x_{it}e_{it}\||d_{it}(\gamma_1) - d_{it}(\gamma_2)|$, $k_{it}(\gamma_1, \gamma_2) \equiv \|x_{it}\||d_{it}(\gamma_1) - d_{it}(\gamma_2)|$, and $\alpha_{NT} \equiv (NT)^{1-2\alpha}$. Define

$$\begin{aligned}
J_{NT}(\gamma) &= \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=1}^{T} x_{it} e_{it} d_{it}(\gamma), \\
J^*_{NT}(\gamma) &= \frac{1}{\sqrt{NT}} \sum_{t=1}^{T} X_t(\gamma)' \mathbb{P}_{\Lambda^0} e_t, \\
K_{NT}(\gamma) &= \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} k_{it}(\gamma, \gamma^0)^2, \\
K^*_{NT}(\gamma) &= \frac{1}{N^2 T} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{t=1}^{T} \|x_{it}\| \|x_{jt}\| \left\|\lambda_j^0\right\| \left\|\lambda_i^0\right\| \left|d_{it}(\gamma^0) - d_{it}(\gamma)\right|, \\
G_{NT}(\gamma) &= \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (C^{0\prime} x_{it})^2 \left|d_{it}(\gamma) - d_{it}(\gamma^0)\right|, \text{ and} \\
G^*_{NT}(\gamma) &= \frac{1}{T} \sum_{t=1}^{T} \left\| \frac{1}{N} \sum_{i=1}^{N} x_{it} \lambda_i^{0\prime} d_{it}(\gamma, \gamma^0) \right\|^2.
\end{aligned}$$

To prove Theorem 2.3, we add the following proposition and three lemmas.

Lemma A.5. *Suppose that Assumptions A.4-A.5 hold. For all $\eta > 0$ and $\varepsilon > 0$, there exists some $\overline{v} < \infty$ such that for any $B < \infty$,*

$$\textit{(i)} \Pr\left(\sup_{\frac{\overline{v}}{\alpha_{NT}} \leq |\gamma - \gamma^0| \leq B} \frac{\left\|J_{NT}(\gamma) - J_{NT}(\gamma^0)\right\|}{\sqrt{\alpha_{NT}} \left|\gamma - \gamma^0\right|} > \eta \right) \leq \varepsilon;$$

$$\textit{(ii)} \Pr\left(\sup_{\frac{\overline{v}}{\alpha_{NT}} \leq |\gamma - \gamma^0| \leq B} \frac{\left\|J^*_{NT}(\gamma) - J^*_{NT}(\gamma^0)\right\|}{\sqrt{\alpha_{NT}} \left|\gamma - \gamma^0\right|} > \eta \right) \leq \varepsilon.$$

Lemma A.6. *Suppose Assumptions A.4-A.5 hold. There exist constants $B > 0$ and $0 < d$, $k < \infty$, such that for all $1 > \eta > 0$, $\varepsilon > 0$, and $c_{NT} \to 0$, $c_{NT}\alpha_{NT} \to \infty$,* *there exists a* $\overline{v} < \infty$ *such that for large enough* (N, T),

$$\Pr\left(\inf_{\frac{\overline{v}}{\alpha_{NT}} \le |\gamma-\gamma^0| \le B} \frac{G_{NT}(\gamma)}{|\gamma-\gamma^0|} < (1-\eta)d\right) \le \varepsilon, \qquad \Pr\left(\sup_{\frac{\overline{v}}{\alpha_{NT}} \le |\gamma-\gamma^0| \le c_{NT}} \frac{G^*_{NT}(\gamma)}{|\gamma-\gamma^0|} > \eta\right) \le \varepsilon,$$

$$\Pr\left(\sup_{\frac{\overline{v}}{\alpha_{NT}} \le |\gamma-\gamma^0| \le B} \frac{K_{NT}(\gamma)}{|\gamma-\gamma^0|} > (1+\eta)k\right) \le \varepsilon, \text{ and } \Pr\left(\sup_{\frac{\overline{v}}{\alpha_{NT}} \le |\gamma-\gamma^0| \le B} \frac{K^*_{NT}(\gamma)}{|\gamma-\gamma^0|} > (1+\eta)k\right) \le \varepsilon.$$

Lemma A.7. *Let*

$$H_{1,NT}(\gamma) \equiv \frac{1}{NT}\sum_{t=1}^{T} X_t(\gamma,\gamma^0)'(\mathbb{P}_{\widehat{\Lambda}} - \mathbb{P}_{\Lambda^0})X_{t,\gamma^0}, \qquad H_{2,NT}(\gamma) \equiv \frac{1}{(NT)^{1-\alpha}}\sum_{t=1}^{T} X_t(\gamma,\gamma^0)'(\mathbb{P}_{\widehat{\Lambda}} - \mathbb{P}_{\Lambda^0})\Lambda^0 f_t,$$

$$H_{3,NT}(\gamma) \equiv \frac{1}{(NT)^{1-\alpha}}\sum_{t=1}^{T} X_t(\gamma,\gamma^0)'(\mathbb{P}_{\widehat{\Lambda}} - \mathbb{P}_{\Lambda^0})e_t, \text{ and } \quad H_{4,NT}(\gamma) \equiv \frac{1}{NT}\sum_{t=1}^{T} X_t(\gamma,\gamma^0)'(\mathbb{P}_{\widehat{\Lambda}} - \mathbb{P}_{\Lambda^0})X_t(\gamma,\gamma^0).$$

Suppose Assumptions A.1-A.5 hold, and $N/T \to \kappa > 0$ *as* $(N, T) \to \infty$*. Then for arbitrary* $\epsilon > 0$ *and* $\eta > 0$*, there are constants* $B > 0$ *and* $\overline{v} > 0$ *such that*

$$\Pr\left(\sup_{\frac{\overline{v}}{\alpha_{NT}} \le |\gamma-\gamma^0| \le B} \frac{\|H_{l,NT}(\gamma)\|}{|\gamma-\gamma^0|} > \eta\right) \le \varepsilon, \qquad \textit{for } l = 1,2,3,4.$$

Proof of Theorem 2.3. (i) Now we have that $\widehat{\gamma} - \gamma^0 = o_p(1)$, $\widehat{\theta} - \theta^0 = o_p((NT)^{-\alpha})$ and $\frac{1}{N}\Lambda^{0\prime}\Lambda^0 \xrightarrow{p} \Sigma_\lambda > 0$. Let the constants B, d and k be as defined in Lemmas A.5-A.7, and $m \equiv 2||\Sigma_\lambda^{-1}||$. Let $M \equiv \max(d, k, m, ||C^0||, 1)$ and choose η and ν small enough such that $\max(\eta, \nu) < M$ and $d - M^3(18\nu + 22\eta + 20\nu\eta) > 0$. Let the event $\mathcal{E}_{NT}$ be the joint event that

(1) $\left|\widehat{\gamma} - \gamma^0\right| \le B$,

(2) $\|(N^{-1}\Lambda^{0\prime}\Lambda^0)^{-1}\| \le m$,

(3) $(NT)^\alpha\|\widehat{\theta} - \theta^0\| \le \nu$,

(4) $\inf_{\frac{\overline{v}}{\alpha_{NT}} \le |\gamma-\gamma^0| \le B} \frac{G_{NT}(\gamma)}{|\gamma-\gamma^0|} > (1-\eta)d$,

(5) $\sup_{\frac{\overline{v}}{\alpha_{NT}} \le |\gamma-\gamma^0| \le B} \frac{K_{NT}(\gamma)}{|\gamma-\gamma^0|} < (1+\eta)k$,

(6) $\sup_{\frac{\overline{v}}{\alpha_{NT}} \le |\gamma-\gamma^0| \le B} \frac{K^*_{NT}(\gamma)}{|\gamma-\gamma^0|} < (1+\eta)k$,

(7) $\sup_{\frac{\overline{v}}{\alpha_{NT}} \le |\gamma-\gamma^0| \le B} \frac{\left\|J^*_{NT}(\gamma) - J^*_{NT}(\gamma^0)\right\|}{\sqrt{\alpha_{NT}}|\gamma-\gamma^0|} < \eta$,

(8) $\sup_{\frac{\overline{v}}{\alpha_{NT}} \le |\gamma-\gamma^0| \le B} \frac{\left\|J_{NT}(\gamma) - J_{NT}(\gamma^0)\right\|}{\sqrt{\alpha_{NT}}|\gamma-\gamma^0|} < \eta$,

(9) $\sup_{\frac{\overline{v}}{\alpha_{NT}} \le |\gamma-\gamma^0| \le B} \frac{\left\|H_{l,NT}(\gamma)\right\|}{|\gamma-\gamma^0|} < \eta$, for $l = 1, 2, 3, 4$, and

(10) $\sup_{\frac{\overline{v}}{\alpha_{NT}} \le |\gamma-\gamma^0| \le B} \frac{G^*_{NT}(\gamma)}{|\gamma-\gamma^0|} < \eta$,

Fix $\epsilon > 0$, one can choose $\overline{v}$ for large enough (N, T) such that $\Pr(\mathcal{E}_{NT}) \geq 1 - \epsilon$, by the Assumption A.2(ii) and Lemmas A.5-A.7. Let $\widehat{\delta} = (NT)^{-\alpha}\widehat{C}$, we have $\|\widehat{C} - C^0\| \leq (NT)^{\alpha}\|\widehat{\theta} - \theta^0\| \leq \nu$.

It suffices to show that $\mathcal{E}_{NT}$ implies $\left|\widehat{\gamma} - \gamma^0\right| \leq \frac{\overline{v}}{\alpha_{NT}}$. Conditional on $\mathcal{E}_{NT}$, we consider $\frac{\overline{v}}{\alpha_{NT}} \leq \left|\gamma - \gamma^0\right| \leq c_{NT} \leq B$ and calculate

$$\begin{aligned}\frac{(NT)^{2\alpha-1}}{|\gamma-\gamma^0|}\left(\mathcal{L}(\widehat{\theta},\widehat{\Lambda},\gamma) - \mathcal{L}(\widehat{\theta},\widehat{\Lambda},\gamma^0)\right) &= \frac{(NT)^{2\alpha}}{|\gamma-\gamma^0|}\frac{1}{NT}\sum_{t=1}^{T}\widehat{\delta}'X_t(\gamma,\gamma^0)'\mathbb{M}_{\widehat{\Lambda}}X_t(\gamma,\gamma^0)\widehat{\delta} \\ &\quad - 2\frac{(NT)^{2\alpha}}{|\gamma-\gamma^0|}\frac{1}{NT}\sum_{t=1}^{T}\widehat{\delta}'X_t(\gamma,\gamma^0)'\mathbb{M}_{\widehat{\Lambda}}X_{t,\gamma^0}(\widehat{\theta}-\theta^0) \\ &\quad - 2\frac{(NT)^{2\alpha}}{|\gamma-\gamma^0|}\frac{1}{NT}\sum_{t=1}^{T}\widehat{\delta}'X_t(\gamma,\gamma^0)'\mathbb{M}_{\widehat{\Lambda}}\Lambda^0 f_t^0 \\ &\quad - 2\frac{(NT)^{2\alpha}}{|\gamma-\gamma^0|}\frac{1}{NT}\sum_{t=1}^{T}\widehat{\delta}'X_t(\gamma,\gamma^0)'\mathbb{M}_{\widehat{\Lambda}}e_t \\ &\equiv \tilde{\mathcal{L}}_1 + \cdots + \tilde{\mathcal{L}}_4. \end{aligned} \tag{A.11}$$

For $\tilde{\mathcal{L}}_1$, we have

$$\begin{aligned}\tilde{\mathcal{L}}_1 &= \frac{(NT)^{2\alpha}}{|\gamma-\gamma^0|}\frac{1}{NT}\sum_{t=1}^{T}\delta^{0\prime}X_t(\gamma,\gamma^0)'\mathbb{M}_{\Lambda^0}X_t(\gamma,\gamma^0)\delta^0 \\ &\quad + \frac{(NT)^{2\alpha}}{|\gamma-\gamma^0|}\frac{1}{NT}\sum_{t=1}^{T}(\widehat{\delta}+\delta^0)'X_t(\gamma,\gamma^0)'\mathbb{M}_{\Lambda^0}X_t(\gamma,\gamma^0)(\widehat{\delta}-\delta^0) \\ &\quad - \frac{(NT)^{2\alpha}}{|\gamma-\gamma^0|}\frac{1}{NT}\sum_{t=1}^{T}\widehat{\delta}'X_t(\gamma,\gamma^0)'(\mathbb{P}_{\widehat{\Lambda}} - \mathbb{P}_{\Lambda^0})X_t(\gamma,\gamma^0)\widehat{\delta} \\ &\equiv \tilde{\mathcal{L}}_{11} + \tilde{\mathcal{L}}_{12} + \tilde{\mathcal{L}}_{13}.\end{aligned}$$

By events (10) and (4) we have

$$\begin{aligned}\tilde{\mathcal{L}}_{11} &\geq \frac{(NT)^{2\alpha}}{|\gamma-\gamma^0|}\frac{1}{NT}\sum_{t=1}^{T}\delta^{0\prime}X_t(\gamma,\gamma^0)'X_t(\gamma,\gamma^0)\delta^0 \\ &\quad - \frac{(NT)^{2\alpha}}{|\gamma-\gamma^0|}\frac{1}{N^2T}\sum_{t=1}^{T}\|\delta^{0\prime}X_t(\gamma,\gamma^0)'\Lambda^0\|^2\left\|(\frac{1}{N}\Lambda^{0\prime}\Lambda^0)^{-1}\right\| \\ &> \frac{G_{NT}(\gamma)}{|\gamma-\gamma^0|} - \|C^0\|^2\frac{G^*_{NT}(\gamma)}{\|\gamma-\gamma^0\|}\left\|(\frac{1}{N}\Lambda^{0\prime}\Lambda^0)^{-1}\right\| \\ &> (1-\eta)d - \|C^0\|^2 m\eta > d - (M + M^3)\eta.\end{aligned}$$

For $\tilde{\mathcal{L}}_{12}$, we have

$$\begin{aligned}|\tilde{\mathcal{L}}_{12}| &\leq \|\widehat{C}-C^0\|\|\widehat{C}+C^0\|\frac{1}{|\gamma-\gamma^0|\,NT}\Big\|\sum_{t=1}^{T}X_t(\gamma,\gamma^0)'\mathbb{M}_{\Lambda^0}X_t(\gamma,\gamma^0)\Big\| \\ &\leq \|\widehat{C}-C^0\|\|\widehat{C}+C^0\|\frac{K_{NT}(\gamma)}{|\gamma-\gamma^0|} \\ &\leq \nu(2\|C^0\|+\nu)(1+\eta)k \leq 2M^2(1+\eta)\nu,\end{aligned}$$

by events (3) and (5). For $\tilde{\mathcal{L}}_{13}$, we have

$$|\tilde{\mathcal{L}}_{13}| \leq (\|C^0\|+\nu)^2\frac{\|H_{4,NT}(\gamma)\|}{|\gamma-\gamma^0|} \leq 4M^2\eta,$$

by events (3) and (9). For $\tilde{\mathcal{L}}_{2}$, we have

$$\begin{aligned}|\tilde{\mathcal{L}}_{2}| &\leq 2\frac{(NT)^{2\alpha}}{|\gamma-\gamma^0|}\Big\|\frac{1}{NT}\sum_{t=1}^{T}\widehat{\delta}'X_t(\gamma,\gamma^0)'X_{t,\gamma^0}(\widehat{\theta}-\theta^0)\Big\| \\ &\quad+2\frac{(NT)^{2\alpha}}{|\gamma-\gamma^0|}\Big\|\frac{1}{NT}\sum_{t=1}^{T}\widehat{\delta}'X_t(\gamma,\gamma^0)'\mathbb{P}_{\Lambda^0}X_{t,\gamma^0}(\widehat{\theta}-\theta^0)\Big\| \\ &\quad+2\frac{(NT)^{2\alpha}}{|\gamma-\gamma^0|}\Big\|\frac{1}{NT}\sum_{t=1}^{T}\widehat{\delta}'X_t(\gamma,\gamma^0)'(\mathbb{P}_{\widehat{\Lambda}}-\mathbb{P}_{\Lambda^0})X_{t,\gamma^0}(\widehat{\theta}-\theta^0)\Big\| \\ &\equiv \tilde{\mathcal{L}}_{21}+\tilde{\mathcal{L}}_{22}+\tilde{\mathcal{L}}_{23}.\end{aligned}$$

For $\tilde{\mathcal{L}}_{21}$, noting that

$$\Big\|\frac{1}{NT}\sum_{t=1}^{T}X_t(\gamma,\gamma^0)'X_{t,\gamma^0}\Big\| \leq \Big\|\frac{1}{NT}\sum_{t=1}^{T}X_t(\gamma,\gamma^0)'X_t\Big\|+\Big\|\frac{1}{NT}\sum_{t=1}^{T}X_t(\gamma,\gamma^0)'X_t(\gamma^0)\Big\| \leq 2K_{NT}(\gamma),$$

we have

$$\tilde{\mathcal{L}}_{21} \leq 4\|\widehat{C}\|\Big[(NT)^{\alpha}\|\widehat{\theta}-\theta^0\|\Big]\frac{K_{NT}(\gamma)}{|\gamma-\gamma^0|} \leq 8M^2(1+\eta)\nu,$$

by events (3) and (5). Similarly, we have

$$\begin{aligned}&\tilde{\mathcal{L}}_{22} \leq 4\|\widehat{C}\|\Big((NT)^{\alpha}\|\widehat{\theta}-\theta^0\|\Big)\Big\|(\frac{\Lambda^{0\prime}\Lambda^0}{N})^{-1}\Big\|\frac{K^*_{NT}(\gamma)}{|\gamma-\gamma^0|} \leq 8M^3(1+\eta)\nu, \text{ and} \\ &\tilde{\mathcal{L}}_{23} \leq 2\|\widehat{C}\|\Big((NT)^{\alpha}\|\widehat{\theta}-\theta^0\|\Big)\frac{H_{1,NT}(\gamma)}{|\gamma-\gamma^0|} \leq 2M\eta\nu,\end{aligned}$$

by events (3), (6) and (9). Next, we consider $\tilde{\mathcal{L}}_3$. We have

$$\begin{aligned}
|\tilde{\mathcal{L}}_3| &\leq 2\frac{(NT)^{2\alpha}}{|\gamma-\gamma^0|}\Big\|\frac{1}{NT}\sum_{t=1}^{T}\widehat{\delta}'X_t(\gamma,\gamma^0)'(\mathbb{P}_{\widehat{\Lambda}}-\mathbb{P}_{\Lambda^0})\Lambda^0 f_t^0\Big\| \\
&\leq 2\|\widehat{C}\|\Big\|\frac{1}{(NT)^{1-\alpha}}\sum_{t=1}^{T}X_t(\gamma,\gamma^0)'(\mathbb{P}_{\widehat{\Lambda}}-\mathbb{P}_{\Lambda^0})\Lambda^0 f_t\Big\|\frac{1}{|\gamma-\gamma^0|} \\
&= 2\|\widehat{C}\|\frac{H_{2,NT}(\gamma)}{|\gamma-\gamma^0|} \leq 4M\eta,
\end{aligned}$$

by events (3) and (9). For $\tilde{\mathcal{L}}_4$, we have

$$\begin{aligned}
|\tilde{\mathcal{L}}_4| &= 2\frac{(NT)^{2\alpha}}{|\gamma-\gamma^0|}\Big\|\frac{1}{NT}\sum_{t=1}^{T}\widehat{\delta}'X_t(\gamma,\gamma^0)'\mathbb{M}_{\widehat{\Lambda}}e_t\Big\| \\
&\leq 2\frac{(NT)^{2\alpha}}{|\gamma-\gamma^0|}\Big\|\frac{1}{NT}\sum_{t=1}^{T}\widehat{\delta}'X_t(\gamma,\gamma^0)'e_t\Big\| + 2\frac{(NT)^{2\alpha-1/2}}{|\gamma-\gamma^0|}\Big\|\frac{1}{\sqrt{NT}}\sum_{t=1}^{T}\widehat{\delta}'X_t(\gamma,\gamma^0)'\mathbb{P}_{\Lambda^0}e_t\Big\| \\
&\quad + 2\frac{(NT)^{2\alpha}}{|\gamma-\gamma^0|}\Big\|\frac{1}{NT}\sum_{t=1}^{T}\widehat{\delta}'X_t(\gamma,\gamma^0)'(\mathbb{P}_{\widehat{\Lambda}}-\mathbb{P}_{\Lambda^0})e_t\Big\| \\
&\equiv \tilde{\mathcal{L}}_{41}+\tilde{\mathcal{L}}_{42}+\tilde{\mathcal{L}}_{43}.
\end{aligned}$$

By events (3) and (7)-(9), we have

$$\begin{aligned}
\tilde{\mathcal{L}}_{41} &\leq 2\|\widehat{C}\|\frac{\|J_{NT}(\gamma)-J_{NT}(\gamma^0)\|}{\sqrt{\alpha_{NT}}\,|\gamma-\gamma^0|} \leq 4M\eta, \\
\tilde{\mathcal{L}}_{42} &\leq 2\|\widehat{C}\|\frac{\|J^*_{NT}(\gamma)-J^*_{NT}(\gamma^0)\|}{\sqrt{\alpha_{NT}}\,|\gamma-\gamma^0|} \leq 4M\eta, \\
\tilde{\mathcal{L}}_{43} &\leq 2\|\widehat{C}\|\frac{\|H_{3,NT}(\gamma)\|}{|\gamma-\gamma^0|} \leq 4M\eta.
\end{aligned}$$

Therefore, we conclude that

$$\begin{aligned}
\frac{(NT)^{2\alpha}}{|\gamma-\gamma^0|}(\mathcal{L}(\widehat{\theta},\widehat{\Lambda},\gamma)-\mathcal{L}(\widehat{\theta},\widehat{\Lambda},\gamma^0)) &\geq \tilde{\mathcal{L}}_1-|\tilde{\mathcal{L}}_2|-|\tilde{\mathcal{L}}_3|-|\tilde{\mathcal{L}}_4| \\
&\geq d - M^3(18\nu+22\eta+20\nu\eta) > 0
\end{aligned}$$

for $\overline{v}/\alpha_{NT} \leq |\gamma-\gamma^0| \leq c_{NT}$. We can conclude that when $\mathcal{E}_{NT}$ occurs, we have $|\widehat{\gamma}-\gamma^0| < \overline{v}/\alpha_{NT}$. Note that $\mathcal{E}_{NT}$ happens with probability larger than $1-\epsilon$. Therefore, for any $\epsilon > 0$, there is a constant $\overline{v}$ such that for (N,T) sufficiently large, we have

$$\Pr\left(|\widehat{\gamma}-\gamma^0| \geq \frac{\overline{v}}{\alpha_{NT}}\right) < \epsilon.$$

This shows that $|\widehat{\gamma}-\gamma^0| = O_p(\alpha_{NT}^{-1})$.

(ii) As mentioned in the proof of the second result of Proposition A.3, all the terms involving δ^0 or $\widehat{\delta}$ have $\mathbf{X}_k(\widehat{\gamma},\gamma^0)$. Given the obtained convergence rate of $\widehat{\gamma}$, these terms now are $O_p((NT)^{\alpha-1}) = o_p(\frac{1}{\sqrt{NT}})$. Now revisit the proof of Proposition A.2. Neglecting all the terms involving δ^0, we immediately obtain $\sqrt{NT}\|\widehat{\theta} - \theta^0\| = O_p(1)$ when $N/T \to \kappa$. In addition, all the results in Lemma A.3 can be sharpened to $O_p(\pi_{NT}^{-2})$, and the second result of Proposition A.1 can also be sharpened to $O_p(\pi_{NT}^{-2})$. ■

Proof of Theorem 2.4

To prove Theorem 2.4, we need the following two lemmas. Lemma A.8 finds the asymptotic bias terms and Lemma A.9 establishes a central limit theorem (CLT). The main difficulty is to find the asymptotic distribution of $\mathcal{C}_{NT}(\gamma) = (NT)^{-1}\sum_{i,t} z_{it,\gamma}e_{it}$. Because $z_{it,\gamma}$ is a variable depends on all entries of $\{\mathbf{X}_{k,\gamma}\}_{k=1}^{2K}$, it is difficult to show CLT directly. To establish the CLT, we follow the same strategy as used by Moon and Weidner (2017).

We define some notations. For each $k = 1, \ldots, 2K$, we define $N \times T$ matrices $\overline{\mathbf{X}}_{k,\gamma}$, $\widetilde{\mathbf{X}}_{k,\gamma}$ and $\widetilde{\mathbf{Z}}_{k,\gamma}$ as follows:

$$\overline{\mathbf{X}}_{k,\gamma} = E_{\mathcal{D}}(\mathbf{X}_{k,\gamma}), \quad \widetilde{\mathbf{X}}_{k,\gamma} = \mathbf{X}_{k,\gamma} - \overline{\mathbf{X}}_{k,\gamma}, \quad \text{and } \widetilde{\mathbf{Z}}_{k,\gamma} = \mathbb{M}_{\Lambda^0}\overline{\mathbf{X}}_{k,\gamma}\mathbb{M}_{F^0} + \widetilde{\mathbf{X}}_{k,\gamma}.$$

Because $\widetilde{\mathbf{X}}_{k,\gamma}$ is centered around zero conditional on $\mathcal{D}$, one can verify that $\|\widetilde{\mathbf{X}}_{k,\gamma}\mathbb{P}_{F^0}\|$ and $\|\mathbb{P}_{\Lambda^0}\widetilde{\mathbf{X}}_{k,\gamma}\|$ are $o_p(\sqrt{NT})$. In the proof of Lemma A.9, we can see that $e_{it}\widetilde{z}_{it,\gamma}$ is an m.d.s., where $\widetilde{z}_{it,\gamma}$ is defined analogously to $z_{it,\gamma}$.

Lemma A.8. *Suppose that Assumptions A.1-A.6 hold and $N/T \to \kappa > 0$. Then for $k = 1, \ldots, 2K$, we have*

(i) $(NT)^{-1/2}\text{tr}(\mathbb{M}_{\Lambda^0}\mathbf{X}_{k,\gamma}\mathbb{M}_{F^0}\mathbf{e}') = (NT)^{-1/2}\text{tr}(\mathbf{e}'\widetilde{\mathbf{Z}}_{k,\gamma}) - \sqrt{\kappa}\mathbb{B}_{1,kNT}(\gamma) + o_p(1)$, *uniformly on γ*

(ii) $\mathbb{B}_{2,NT}(\widehat{\gamma}) = \mathbb{B}_{2,NT}(\gamma^0) + o_p(1)$;

(iii) $\mathbb{B}_{3,NT}(\widehat{\gamma}) = \mathbb{B}_{3,NT}(\gamma^0) + o_p(1)$.

Lemma A.9. *Suppose that Assumptions A.2, A.4 and A.6 hold. Then*

$$\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T} e_{it}\widetilde{z}_{it,\gamma} \xrightarrow{d} \mathcal{G}(\gamma) \quad \text{in} \quad \ell^\infty(\Gamma),$$

where $\mathcal{G}(\gamma)$ is some Gaussian process with $E(\mathcal{G}(\gamma_1)\mathcal{G}(\gamma_2)') = \Omega(\gamma_1, \gamma_2)$ and $\ell^\infty(\Gamma)$ denotes the space of bounded functions over the compact set Γ endowed with the uniform metric.

Proof of Theorem 2.4. Revisit equation (A.5) in the proof of Proposition A.2. The first term on the RHS is $\frac{1}{NT}\sum_{t=1}^{T} X'_{t,\widehat{\gamma}}\mathbb{M}_{\widehat{\Lambda}}e_t$, whose kth element is studied in details. According to the analysis there, only $II_{1,k}$ and $II_{4,k}$ matter since the remaining terms are either $o_p(\|\widehat{\theta}-\theta^0\|)$, or $o_p(\pi_{NT}^{-3})$, or involve δ^0. For those terms involving δ^0, they are $O_p((NT)^{\alpha-1})$ due to the arguments in the proof of result (ii) of Theorem 2.3. For $II_{4,k}$, substituting (A.4) into it and some straightforward computation shows that

$$II_{4,k} = \frac{1}{N}\mathrm{tr}\Big[E_{\mathcal{D}}(\mathbf{e}'\mathbf{e})\mathbf{X}'^{0}_{k,\widehat{\gamma}}(\Lambda^{0\prime}\Lambda^0)^{-1}(F^{0\prime}F^0)^{-1}F^{0\prime}\Big]+O_p(\pi_{NT}^{-3})+o_p(\|\widehat{\theta}-\theta^0\|)+O_p((NT)^{\alpha-1}).$$

Given this, we have

$$\frac{1}{NT}\sum_{t=1}^{T}X'_{tk,\widehat{\gamma}}\mathbb{M}_{\widehat{\Lambda}}e_t = \frac{1}{NT}\sum_{t=1}^{T}X'_{tk,\widehat{\gamma}}\mathbb{M}_{\Lambda^0}e_t - \frac{1}{N}\mathrm{tr}\Big[E_{\mathcal{D}}(\mathbf{e}'\mathbf{e})\mathbf{X}'^{0}_{k,\widehat{\gamma}}(\Lambda^{0\prime}\Lambda^0)^{-1}(F^{0\prime}F^0)^{-1}F^{0\prime}\Big] + o_p(\frac{1}{\sqrt{NT}}),$$

where $X_{tk,\widehat{\gamma}}$ is the kth column of $X_{t,\widehat{\gamma}}$. Next, we consider the second term $\frac{1}{NT}\sum_{t=1}^{T} X'_{t,\widehat{\gamma}}\mathbb{M}_{\widehat{\Lambda}}\Lambda^0 f_t^0$, whose kth term is analyzed according to (A.8). By the same arguments, i.e., neglecting all the terms that are $o_p(\|\widehat{\theta}-\theta^0\|)$, or $O_p(\pi_{NT}^{-3})$, or involve δ^0, we only keep $J_{1,k}$, $J_{2,k}$ and $J_{9,k}$. This gives

$$\begin{aligned}\frac{1}{NT}\sum_{t=1}^{T}X'_{tk,\widehat{\gamma}}\mathbb{M}_{\widehat{\Lambda}}\Lambda^0 f_t^0 = &\frac{1}{NT^2}\sum_{t=1}^{T}\sum_{s=1}^{T}a_{st}X'_{tk,\widehat{\gamma}}\mathbb{M}_{\widehat{\Lambda}}X_{s,\gamma^0}(\widehat{\theta}-\theta^0)\\ &-\frac{1}{NT^2}\sum_{t=1}^{T}\sum_{s=1}^{T}a_{st}X'_{tk,\widehat{\gamma}}\mathbb{M}_{\widehat{\Lambda}}e_s - J_{1,k} + o_p(\frac{1}{\sqrt{NT}}).\end{aligned}$$

For the second term on the RHS of the above equation, it can be decomposed into five terms, which are given in the analysis of $J_{2,k}$ in the proof of Proposition A.2. The analysis there indicates that only the first term $\frac{1}{NT}\mathrm{tr}(\mathbf{e}'\mathbb{M}_{\Lambda^0}\mathbf{X}_{k,\widehat{\gamma}}\mathbb{P}_{F^0})$ and $II_{8,k}$ matter. Substituting (A.4) into $II_{8,k}$, we have

$$II_{8,k} = \frac{1}{N}\mathrm{tr}\Big[E_{\mathcal{D}}(\mathbf{e}'\mathbf{e})\mathbb{P}_{F^0}\mathbf{X}'^{0}_{k,\widehat{\gamma}}(\Lambda^{0\prime}\Lambda^0)^{-1}(F^{0\prime}F^0)^{-1}F^{0\prime}\Big]+O_p(\pi_{NT}^{-3})+o_p(\|\widehat{\theta}-\theta^0\|)+O_p((NT)^{\alpha-1}).$$

In addition, with the results in Lemma A.3, we can readily show that $J_{1,k} = \mathbb{B}_{2,k,NT}(\widehat{\gamma}) + o_p(\frac{1}{\sqrt{NT}})$. With the above results, we have

$$\begin{aligned}\frac{1}{NT}\sum_{t=1}^{T}X'_{tk,\widehat{\gamma}}\mathbb{M}_{\widehat{\Lambda}}\Lambda^0 f_t^0 = &\frac{1}{NT^2}\sum_{t=1}^{T}\sum_{s=1}^{T}a_{st}X'_{tk,\widehat{\gamma}}\mathbb{M}_{\widehat{\Lambda}}X_{s,\gamma^0}(\widehat{\theta}-\theta^0) - \frac{1}{NT^2}\sum_{t=1}^{T}\sum_{s=1}^{T}a_{st}X'_{tk,\widehat{\gamma}}\mathbb{M}_{\Lambda^0}e_s - \mathbb{B}_{2,k,NT}(\widehat{\gamma})\\ &+\frac{1}{N}\mathrm{tr}\Big[E_{\mathcal{D}}(\mathbf{e}'\mathbf{e})\mathbb{P}_{F^0}\mathbf{X}'^{0}_{k,\widehat{\gamma}}(\Lambda^{0\prime}\Lambda^0)^{-1}(F^{0\prime}F^0)^{-1}F^{0\prime}\Big] + o_p(\frac{1}{\sqrt{NT}}).\end{aligned}$$

Finally, we consider the last term:

$$\frac{1}{NT}\sum_{t=1}^{T}X_{tk,\widehat{\gamma}}\mathbb{M}_{\widehat{\Lambda}}X(\gamma^0,\widehat{\gamma})\delta^0 = O_p(|\widehat{\gamma}-\gamma^0|)O_p(\|\delta^0\|) = O_p((NT)^{\alpha-1}) = o_p(\frac{1}{\sqrt{NT}}).$$

Given the above analysis, we have

$$[\mathcal{B}(\Lambda^0,\widehat{\gamma})+o_p(1)]\sqrt{NT}(\widehat{\theta}-\theta^0)=\sqrt{NT}\mathcal{C}_{NT}(\widehat{\gamma})-\sqrt{\frac{T}{N}}\mathbb{B}_{2,NT}(\widehat{\gamma})-\sqrt{\frac{N}{T}}\mathbb{B}_{3,NT}(\widehat{\gamma})+o_p(1).$$

Given the convergence rate of $\widehat{\gamma}$, one can verify that

$$\begin{aligned}
\mathcal{B}(\Lambda^0,\widehat{\gamma}) &= \omega_{NT}(\gamma^0,\gamma^0)+o_p(1),\\
\sqrt{NT}\mathcal{C}_{NT}(\widehat{\gamma}) &= \frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}e_{it}\widetilde{z}_{it,\widehat{\gamma}}-\sqrt{\frac{N}{T}}\mathbb{B}_{1,NT}(\widehat{\gamma}),\\
\sqrt{NT}\mathbb{B}_{\ell,NT}(\widehat{\gamma}) &= \sqrt{NT}\mathbb{B}_{\ell,NT}(\gamma^0)+o_p(1), \quad \text{for } \ell=1,2,3,\\
\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}e_{it}\widetilde{z}_{it,\widehat{\gamma}} &= \frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}e_{it}\widetilde{z}_{it,\gamma^0}+o_p(1),
\end{aligned}$$

where the last result is due to Lemma A.9. It follows that

$$\begin{aligned}
\Big[\omega_{NT}(\gamma^0,\gamma^0)\Big]^{-1}\sqrt{NT}(\widehat{\theta}-\theta^0) &= \frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}e_{it}\widetilde{z}_{it,\gamma^0}-\sqrt{\frac{N}{T}}\mathbb{B}_{1,NT}(\gamma^0)\\
&\quad -\sqrt{\frac{T}{N}}\mathbb{B}_{2,NT}(\gamma^0)-\sqrt{\frac{N}{T}}\mathbb{B}_{3,NT}(\gamma^0)+o_p(1).
\end{aligned}$$

Then by Lemmas A.8 and A.9 as well as Assumption A.6, we have that as $N/T\to\kappa$,

$$\omega_0\sqrt{NT}(\widehat{\theta}-\theta^0)-\mathbb{B}\xrightarrow{d}N(0,\Omega_0),$$

where $\mathbb{B}=-\kappa^{1/2}\mathbb{B}_1(\gamma^0)-\kappa^{-1/2}\mathbb{B}_2(\gamma^0)-\kappa^{1/2}\mathbb{B}_3(\gamma^0)$ and $\omega_0=\omega(\gamma^0,\gamma^0)$. ■

Proof of Theorem 2.5

Let $\widetilde{g}\equiv C^{0\prime}D_f^0C^0$ and $\widetilde{h}\equiv C^{0\prime}V_f^0C^0$. Let

$$\begin{aligned}
&R_{NT}(v)\equiv\sqrt{\alpha_{NT}}\Big[J_{NT}(\gamma^0+\frac{v}{\alpha_{NT}})-J_{NT}(\gamma^0)\Big],\\
&\widetilde{G}_{NT}(v)\equiv\alpha_{NT}G_{NT}(\gamma^0+\frac{v}{\alpha_{NT}}), \text{ and } \widetilde{K}_{NT}(v)\equiv\alpha_{NT}K_{NT}(\gamma^0+\frac{v}{\alpha_{NT}}).
\end{aligned}$$

Define

$$\tilde{\mathcal{L}}_{NT}(v)=\mathcal{L}(\widehat{\theta},\widehat{\Lambda},\gamma^0)-\mathcal{L}(\widehat{\theta},\widehat{\Lambda},\gamma^0+\frac{v}{\alpha_{NT}}).$$

To prove Theorem 2.5, we need the following three lemmas.

Lemma A.10. *Suppose that Assumptions A.2 and A.4–A.5 and $N/T\to\kappa>0$. Then $R_{NT}(v)\Rightarrow B(v)$, where $B(v)$ is a Brownian motion with covariance matrix $E\left[B(u)B(v)'\right]=V_f^0\min(u,v)$.*

Lemma A.11. *Suppose that Assumptions A.2 and A.4–A.5 hold and $N/T\to\kappa>0$.*

Then $\widetilde{G}_{NT}(v) \xrightarrow{p} \widetilde{g}\,|v|$ and $\widetilde{K}_{NT}(v) \xrightarrow{p} \|D_f\|\,|v|$ uniformly in $v \in \Upsilon$, where $\widetilde{g} = C^{0\prime}D_f^0C^0$ and Υ is a compact set on the real line that includes 0 as its interior point.

Lemma A.12. *Suppose that Assumptions A.1-A.7 hold and $N/T \to \kappa > 0$. Then*

$$\tilde{\mathcal{L}}_{NT}(v) \Rightarrow \tilde{\mathcal{L}}(v) = -\widetilde{g}\,|v| + 2\sqrt{\widetilde{h}}W(v),$$

where $\widetilde{h} = C^{0\prime}V_f^0C^0$.

Proof of Theorem 2.5. Noting that $\widehat{\gamma} = \text{argmin}_\gamma\mathcal{L}(\widehat{\theta}(\gamma), \widehat{\Lambda}(\gamma), \gamma) = \text{argmin}_\gamma\mathcal{L}(\widehat{\theta}, \widehat{\Lambda}, \gamma)$ where $\widehat{\theta} = \widehat{\theta}(\widehat{\gamma})$ and $\widehat{\Lambda} = \widehat{\Lambda}(\widehat{\gamma})$, we have $\alpha_{NT}(\widehat{\gamma}-\gamma^0) = \text{argmax}_v\tilde{\mathcal{L}}_{NT}(v)$. By Theorem 2.3, $\alpha_{NT}(\widehat{\gamma} - \gamma^0) = O_p(1)$. So it suffices to confine the analysis on some compact set K. By Lemma A.12, $\tilde{\mathcal{L}}_{NT}(v) \Rightarrow \tilde{\mathcal{L}}(v)$ in $\ell^\infty(K)$, the space of all the bounded functions over some compact set K endowed with the uniform metric. The limit functional $\tilde{\mathcal{L}}(v)$ is continuous, has a unique maximum, and $\lim_{|v|\to\infty} \tilde{\mathcal{L}}(v) = -\infty$ almost surely. It therefore satisfies the conditions in Theorem 2.7 of Kim and Pollard (1990). By the argmax continuous mapping theorem (CMT), we have

$$\alpha_{NT}(\widehat{\gamma} - \gamma^0) \xrightarrow{d} \text{argmax}_{v\in\mathbb{R}}\tilde{\mathcal{L}}(v).$$

Following the proof of Theorem 1 in Hansen (2000), we have

$$\begin{aligned}
\text{argmax}_{v\in\mathbb{R}}\tilde{\mathcal{L}}(v) &= \phi\text{argmax}_{r\in\mathbb{R}}\left[-\widetilde{g}\phi\,|r| + 2\sqrt{\widetilde{h}}\sqrt{\phi}W(r)\right] \\
&= \phi\text{argmax}_{r\in\mathbb{R}}[-\widetilde{g}\phi\,|r| + 2\widetilde{g}\phi W(r)] \\
&= \phi\text{argmax}_{r\in\mathbb{R}}[-\frac{|r|}{2} + W(r)],
\end{aligned}$$

where we apply the change of variables with $v = \phi r$, $\phi = \widetilde{h}/\widetilde{g}^2$ and the distributional equality $W(a^2r) = aW(r)$. This completes the proof of the theorem. ■

Proof of Theorem 2.6

To prove theorem 2.6, we need the following lemma.

Lemma A.13. *Suppose Assumptions A.1-A.5 hold and $N/T \to \kappa > 0$. Then $\mathcal{L}(\widehat{\theta}_{\gamma^0}, \widehat{\Lambda}_{\gamma^0}, \gamma^0) - \mathcal{L}(\widehat{\theta}, \widehat{\Lambda}, \gamma^0) = o_p(1)$.*

Proof of Theorem 2.6. Recall that $\widehat{\sigma}^2 = \frac{1}{NT}\mathcal{L}(\widehat{\theta}, \widehat{\Lambda}, \widehat{\gamma})$. It is easy to show that $\widehat{\sigma}^2 \xrightarrow{p} \sigma^2$. By the definitions of $LR_{NT}(\cdot)$ and $\tilde{\mathcal{L}}_{NT}(\cdot)$, we have

$$\begin{aligned}
\widehat{\sigma}^2LR_{NT}(\gamma^0) - \tilde{\mathcal{L}}_{NT}(\widehat{v}) &= \left[\mathcal{L}(\widehat{\theta}_{\gamma^0}, \widehat{\Lambda}_{\gamma^0}, \gamma^0) - \mathcal{L}(\widehat{\theta}, \widehat{\Lambda}, \gamma^0 + \frac{\widehat{v}}{\alpha_{NT}})\right] - \left[\mathcal{L}(\widehat{\theta}, \widehat{\Lambda}, \gamma^0) - \mathcal{L}(\widehat{\theta}, \widehat{\Lambda}, \gamma^0 + \frac{\widehat{v}}{\alpha_{NT}})\right] \\
&= \mathcal{L}(\widehat{\theta}_{\gamma^0}, \widehat{\Lambda}_{\gamma^0}, \gamma^0) - \mathcal{L}(\widehat{\theta}, \widehat{\Lambda}, \gamma^0) = o_p(1),
\end{aligned}$$

where the last result is due to Lemma A.13. By Lemma A.12 and the CMT,

$$LR_{NT}(\gamma^0) = \frac{\tilde{\mathcal{L}}_{NT}(\widehat{v})}{\widehat{\sigma}^2} + o_p(1) = \frac{\sup_v \tilde{\mathcal{L}}_{NT}(v)}{\widehat{\sigma}^2} + o_p(1) \xrightarrow{d} \frac{\sup_v \tilde{\mathcal{L}}(v)}{\sigma^2}.$$

The limit distribution is

$$\begin{aligned}\frac{1}{\sigma^2}\sup_v \Big[-\widetilde{g}|v| + 2\sqrt{\widetilde{h}}W(v)\Big] &= \frac{1}{\sigma^2}\sup_u \Big[-\widetilde{g}\Big|\frac{\widetilde{h}}{\widetilde{g}^2}u\Big| + 2\sqrt{\widetilde{h}}W(\frac{\widetilde{h}}{\widetilde{g}^2}u)\Big] \\ &= \frac{\widetilde{h}}{\widetilde{g}\sigma^2}\sup_u \Big[-|u| + 2W(u)\Big] = \eta^2\Xi,\end{aligned}$$

where the first equality holds by the change of variables ($v = \frac{\widetilde{h}}{\widetilde{g}^2}u$) and the second equality holds by the fact that $W(a^2r) = aW(r)$.

Note that we can write $\Xi = 2\max(\Xi_1, \Xi_2)$, where $\Xi_1 = \sup_{r\le 0}[-\frac{1}{2}|r| + W(r)]$ and $\Xi_2 = \sup_{r\ge 0}[-\frac{1}{2}|r| + W(r)]$. Ξ_1 and Ξ_2 are independent exponential random variables with distribution function $\Pr(\Xi_1 \le x) = 1 - e^{-x}$. It follows that

$$\Pr(\Xi \le x) = \Pr(2\max(\Xi_1, \Xi_2) \le x) = \Pr(\Xi_1 \le x/2)\Pr(\Xi_2 \le x/2) = (1 - e^{-x/2})^{2.}.$$

This completes the proof of the theorem. ■

Proof of Theorem 2.7

Proof of Theorem 2.7. Let $\widetilde{\theta}(\gamma)$ be the bias corrected estimator that can be obtained as in Bai (2009) or Moon and Weidner (2017) by treating γ as known. Note that the model can be written as

$$y_{it} = x_{it}'\theta^0 + \lambda_i^{0\prime}f_t^0 + \Big(e_{it} + \frac{1}{\sqrt{NT}}x_{it}'cd_{it}(\gamma^0)\Big).$$

Treating the expression in the brackets as a new error term, by Bai (2009) or Moon and Weidner (2017),

$$\begin{aligned}\sqrt{NT}(\widetilde{\theta}(\gamma) - \theta^0) = {}& \omega_{NT}(\gamma,\gamma)^{-1}\frac{1}{\sqrt{NT}}\sum_{i=1}^N\sum_{t=1}^T \widetilde{z}_{it,\gamma}e_{it} \\ &+ \omega_{NT}(\gamma,\gamma)^{-1}\frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T \widetilde{z}_{it,\gamma}(\widetilde{z}_{it,\gamma} - \widetilde{z}_{it,\gamma^0})'Lc + o_{p\gamma}(1).\end{aligned}$$

where $o_{p\gamma}(1)$ denotes the terms which are $o_p(1)$ uniformly in γ. Recall that $S_{NT}(\gamma) \equiv \frac{1}{\sqrt{NT}}\sum_{i=1}^N\sum_{t=1}^T \widetilde{z}_{it,\gamma}e_{it}$ and let $\widetilde{\omega}_{NT}(\gamma_1,\gamma_2) \equiv \frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T \widetilde{z}_{it,\gamma_1}\widetilde{z}_{it,\gamma_2}'$. By $L'\theta^0 =$

$\delta^0, \sqrt{NT}\delta^0 = c$ and $c = L'Lc$, we have

$$\begin{aligned}\sqrt{NT}L'\widetilde{\theta}(\gamma) &= L'Lc + L'\widetilde{\omega}_{NT}(\gamma,\gamma)^{-1}S_{NT}(\gamma) + L'\widetilde{\omega}_{NT}(\gamma,\gamma)^{-1}\left[\widetilde{\omega}_{NT}(\gamma,\gamma^0) - \widetilde{\omega}_{NT}(\gamma,\gamma)\right]Lc + o_{p\gamma}(1)\\ &= L'\widetilde{\omega}_{NT}(\gamma,\gamma)^{-1}S_{NT}(\gamma) + L'\widetilde{\omega}_{NT}(\gamma,\gamma)^{-1}\widetilde{\omega}_{NT}(\gamma,\gamma^0)Lc + o_{p\gamma}(1)\\ &\Rightarrow L'\omega(\gamma,\gamma)^{-1}S(\gamma) + L'\omega(\gamma,\gamma)^{-1}\omega(\gamma,\gamma^0)Lc \equiv \overline{S}(\gamma) + \overline{Q}(\gamma)c,\end{aligned}$$

where $\overline{S}(\gamma) \equiv L'\omega(\gamma,\gamma)^{-1}S(\gamma)$ and $\overline{Q}(\gamma) \equiv L'\omega(\gamma,\gamma)^{-1}\omega(\gamma,\gamma^0)L$.

Next, it is standard to show that $L'\widehat{V}_{NT}(\gamma)L \xrightarrow{p} L'\omega(\gamma,\gamma)^{-1}\Omega(\gamma,\gamma)\omega(\gamma,\gamma)^{-1}L = \overline{K}(\gamma,\gamma)$ uniformly in γ. Then by the CMT, we have $W_{NT}(\gamma) \Rightarrow W^c(\gamma)$. ■

Proof of Theorem 2.8

Proof of Theorem 2.8. Let $\mathcal{W}_{NT} = \{(y_{it}, x_{it}, q_{it}), 1 \le i \le N, 1 \le t \le T\}$ and $S^*_{NT}(\gamma) = \frac{1}{\sqrt{NT}}\sum_{i=1}^N\sum_{t=1}^T z_{it,\gamma}e_{it}v_{it}$. Let $P^*(\cdot)$, $E^*(\cdot)$ and $\mathrm{Var}^*(\cdot)$ denote the probability, expectation and variance conditional on the random sample $\mathcal{W}_{NT}$. We say that $A_{NT} = o_{p*}(1)$ is $P_w(\|A_{NT}\| \ge \epsilon) = o_p(1)$ for any $\epsilon > 0$. Note that $A_{NT} = o_p(1)$ implies that $A_{NT} = o_{p*}(1)$. It suffices to prove the theorem by showing that (i) $S^*_{NT}(\gamma) \Rightarrow S(\gamma)$ conditional on w, (ii) $\widehat{\omega}_{NT}(\gamma,\gamma) = \omega(\gamma,\gamma) + o_{p^*}(1)$ uniformly in $\gamma \in \Gamma$, and (iii) $\widehat{S}_{NT}(\gamma) = S^*_{NT}(\gamma) + o_{p*}(1)$ uniformly in $\gamma \in \Gamma$, (iv) $\widehat{\Omega}_{NT}(\gamma,\gamma) = \Omega(\gamma,\gamma) + o_{p^*}(1)$ uniformly in $\gamma \in \Gamma$.

We first show (i). Note that by Assumption A.7(i)

$$E^*\left[S^*_{NT}(\gamma_1)S^*_{NT}(\gamma_2)'\right] = \frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T z_{it,\gamma_1}z'_{it,\gamma_2}e^2_{it} = \Omega_{NT}(\gamma_1,\gamma_2) = \Omega(\gamma_1,\gamma_2) + o_{p^*}(1).$$

So $S^*_{NT}(\gamma)$ is a zero-mean Gaussian process with covariance function kernel $\Omega(\gamma_1,\gamma_2)$ asymptotically. In addition, it is standard to show that the finite dimensional distribution of S^*_{NT} converges to that of S_{NT} as $(N,T) \to \infty$. The stochastic equicontinuity also holds by standard arguments. Then we have $S^*_{NT} \Rightarrow S$ conditional on $\mathcal{W}_{NT}$.

To show (ii), it suffices to show $\frac{1}{NT}\sum_{i,t}\|z_{it,\gamma} - \check{z}_{it,\gamma}\|^2 = o_{p^*}(1)$. We have

$$\frac{1}{NT}\sum_{i,t}\|z_{it,\gamma} - \check{z}_{it,\gamma}\|^2 = \frac{1}{NT}\sum_{k=1}^{2K}\|\mathbb{M}_{\Lambda^0}\mathbf{X}_{k,\gamma}\mathbb{M}_{F^0} - \mathbb{M}_{\widehat{\Lambda}(\gamma)}\mathbf{X}_{k,\gamma}\mathbb{M}_{\widehat{F}(\gamma)}\|^2.$$

We can readily show $\|\mathbb{P}_{\Lambda^0} - \mathbb{P}_{\widehat{\Lambda}(\gamma)}\| = O_p(\pi_{NT}^{-1})$ and $\|\mathbb{P}_{F^0} - \mathbb{P}_{\widehat{F}(\gamma)}\| = O_p(\pi_{NT}^{-1})$ uniformly. Hence, the result follows.

To show (iii), notice that

$$\begin{aligned}S^*_{NT}(\gamma) - \widehat{S}_{NT}(\gamma) &= \frac{1}{\sqrt{NT}}\sum_{i,t}z_{it,\gamma}\left(e_{it} - \widehat{e}_{it}(\gamma)\right)v_{it} + \frac{1}{\sqrt{NT}}\sum_{i,t}\left(z_{it,\gamma} - \check{z}_{it,\gamma}\right)\widehat{e}_{it}(\gamma)v_{it}\\ &\equiv A_1(\gamma) + A_2(\gamma). \text{ say.}\end{aligned}$$

For $A_1(\gamma)$, we have $\widehat{e}_{it}(\gamma) = y_{it} - x_{it,\gamma}{}'\widehat{\theta}(\gamma) - \widehat{\lambda}_i'\widehat{f}_t = e_{it} - x_{it,\gamma}'(\widehat{\theta}(\gamma) - \theta^0) - (\widehat{\lambda}_i(\gamma)'\widehat{f}_t(\gamma) - \lambda_i^{0\prime}f_t^0)$, and that

$$\begin{aligned} A_1(\gamma) &= -\frac{1}{\sqrt{NT}}\sum_{i,t} z_{it,\gamma}x_{it,\gamma}'(\widehat{\theta}(\gamma)-\theta^0)v_{it} - \frac{1}{\sqrt{NT}}\sum_{i,t} z_{it,\gamma}(\widehat{\lambda}_i(\gamma)'\widehat{f}_t(\gamma) - \lambda_i^{0\prime}f_t^0)v_{it} \\ &\equiv -A_{11}(\gamma) - A_{12}(\gamma), \text{ say.} \end{aligned}$$

The first term is bounded by $\|\frac{1}{\sqrt{NT}}\sum_{i,t} z_{it,\gamma}x_{it,\gamma}'v_{it}\|\|\widehat{\theta}(\gamma)-\theta^0\| = O_p(\|\widehat{\theta}-\theta^0\|(\ln N)^3) = o_{p*}(1)$ uniformly in γ because one can show that $\sup_{\gamma\in\Gamma}\|\frac{1}{\sqrt{NT}}\sum_{i,t} z_{it,\gamma}x_{it,\gamma}'v_{it}\| = O_{p^*}((\ln N)^3)$ by using Bernstein-type inequality for independent observations (see, e.g., the online supplement of Miao et al. (2020a) of Su et al. (2016)). Note that

$$\begin{aligned} \|\mathrm{Var}^*(A_{12})\| &= \|\frac{1}{NT}\sum_{i,t} z_{it,\gamma}z_{it,\gamma}'(\widehat{\lambda}_i(\gamma)'\widehat{f}_t(\gamma) - \lambda_i^{0\prime}f_t^0)^2\| \\ &\leq \Big(\frac{1}{NT}\sum_{i,t}\|z_{it,\gamma}\|^4\Big)^{1/2}\Big(\frac{1}{NT}\sum_{i,t}(\widehat{\lambda}_i(\gamma)'\widehat{f}_t(\gamma) - \lambda_i^{0\prime}f_t^0)^4\Big)^{1/2} = o_{p^*}(1) \end{aligned}$$

as we can readily show that $\frac{1}{NT}\sum_{i,t}(\widehat{\lambda}_i(\gamma)'\widehat{f}_t(\gamma) - \lambda_i^{0\prime}f_t^0)^4 = o_{p^*}(1)$ uniformly in γ.

Consider $A_2(\gamma)$, we have

$$\begin{aligned} A_2(\gamma) &= \frac{1}{\sqrt{NT}}\sum_{i,t}(z_{it,\gamma} - \check{z}_{it,\gamma})\,e_{it}v_{it} - \frac{1}{\sqrt{NT}}\sum_{i,t}(z_{it,\gamma} - \check{z}_{it,\gamma})\,x_{it}(\gamma)'(\widehat{\beta}(\gamma) - \beta)v_{it} \\ &\quad -\frac{1}{\sqrt{NT}}\sum_{i,t}(z_{it,\gamma} - \check{z}_{it,\gamma})\,(\widehat{\lambda}_i(\gamma)'\widehat{f}_t(\gamma) - \lambda_i^{0\prime}f_t^0)v_{it}. \\ &\equiv A_{21}(\gamma) - A_{22}(\gamma) - A_{23}(\gamma), \text{ say.} \end{aligned}$$

For $A_{21}(\gamma)$, we can calculate the variance of A_{21} conditional on $\mathcal{W}_{NT}$:

$$\begin{aligned} \|\mathrm{Var}^*(A_{21})\| &= \|\frac{1}{NT}\sum_{i,t}(z_{it,\gamma} - \check{z}_{it,\gamma})(z_{it,\gamma} - \check{z}_{it,\gamma})'e_{it}^2\| \\ &\leq \sup_{i,t} e_{it}^2 \cdot \frac{1}{NT}\sum_{i,t}\|z_{it,\gamma} - \check{z}_{it,\gamma}\|^2 \\ &= O_p((NT)^{2/(8+\epsilon)})\cdot O_p(\pi_{NT}^{-1}) = o_{p^*}(1). \end{aligned}$$

For $A_{22}(\gamma)$ and $A_{23}(\gamma)$, we can follow the arguments as used in the analysis of $A_{11}(\gamma)$ and $A_{12}(\gamma)$ and show they are $o_{p^*}(1)$ uniformly in γ. Then $\widehat{S}_{NT}(\gamma) - S_{NT}^*(\gamma) = o_{p^*}(1)$ uniformly in γ. Thus we have $\widehat{S}_{NT}(\gamma) \Rightarrow S(\gamma)$.

To show (iv), we can follow similar arguments to (iii). The analysis is tedious and omitted.

The final result follows from the CMT. ■

Appendix B
Technical Results for Chapter 3

In this appendix we prove the main results in the paper. The proofs rely on some technical lemmas whose proofs can be found in Appendix B of the online supplement of Miao et al. (2020b). They also call on some other technical lemmas in Appendix C of the online supplement of Miao et al. (2020b).

Proof of the main results

To prove Theorem 3.1, we first need three technical lemmas, viz, Lemmas B.1–B.2 below. To state these lemmas, we define some notation. First, we introduce the following auxiliary objective function:

$$\tilde{\mathcal{Q}}(\Theta, \mathbf{D}, \mathbf{G}) = \sum_{i=1}^{N}\sum_{t=1}^{T}\left[\tilde{x}_{it}'(\beta_{g_i^0}^0 - \beta_{g_i}) + \tilde{x}_{it}(\gamma_{g_i^0}^0)'\delta_{g_i^0}^0 - \tilde{x}_{it}(\gamma_{g_i})\delta_{g_i}\right]^2 + \sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{\varepsilon}_{it}^2. \qquad \text{(B.1)}$$

Lemma B.1 shows that the distance between $\tilde{\mathcal{Q}}(\Theta, \mathbf{D}, \mathbf{G})$ and $\mathcal{Q}(\Theta, \mathbf{D}, \mathbf{G})$ is $o_p(1)$ uniformly in $(\Theta, \mathbf{D}, \mathbf{G})$ so that we can study the asymptotic properties of $\hat{\Theta}$ through $\tilde{\mathcal{Q}}(\Theta, \mathbf{D}, \mathbf{G})$ in Lemma B.2. Now, define the Hausdorff distance $d_H : \mathcal{B}^G \times \mathcal{B}^G \to R$ as follows

$$d_H(a,b) \equiv \max\left\{\max_{g\in\mathcal{G}}\left(\min_{\tilde{g}\in\mathcal{G}}\|a_g - b_{\tilde{g}}\|\right),\ \max_{\tilde{g}\in\mathcal{G}}\left(\min_{g\in\mathcal{G}}\|a_g - b_{\tilde{g}}\|\right)\right\}.$$

Lemma B.1. *Suppose that Assumption A.1 holds. Then*

$$\sup_{(\Theta,\mathbf{D},\mathbf{G})\in\mathcal{B}^G\times\Gamma^G\times\mathcal{G}^N}\frac{1}{NT}|\mathcal{Q}(\Theta, \mathbf{D}, \mathbf{G}) - \tilde{\mathcal{Q}}(\Theta, \mathbf{D}, \mathbf{G})| = o_p(1).$$

Lemma B.2. *Suppose that Assumptions A.1–A.3 hold. Then* $d_H(\hat{\Theta}, \Theta^0) \xrightarrow{p} 0$ *as* $(N,T) \to \infty$.

Remark. The proof of Lemma B.2 shows that there exists a permutation $\sigma_{\hat{\Theta}}$ such that $\left\|\hat{\theta}_g - \theta^0_{\sigma_{\hat{\Theta}}(g)}\right\| = o_p(1)$. We can take $\sigma_{\hat{\Theta}}(g) = g$ by relabeling. In the following analysis, we shall write $\hat{\theta}_g - \theta_g^0 = o_p(1)$ without referring to the relabeling any more.

Lemma B.3. *Let* $\hat{g}_i(\Theta, \mathbf{D}) = \arg\min_{g\in\mathcal{G}}\sum_{t=1}^{T}\left[\tilde{y}_{it} - \tilde{z}_{it}(\gamma_g)'\theta_g\right]^2$. *Suppose Assumptions A.1–A.3 hold. For some* $\eta > 0$ *small enough and* (N,T) *large enough such that* $\max_{g\in\mathcal{G}}\left\|\delta_g^0\right\| \leq$

$\sqrt{\eta}$, *we have*

$$\Pr\left(\sup_{(\Theta,\mathbf{D})\in\mathcal{N}_\eta\times\Gamma^G}\left[\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}(\hat{g}_i(\Theta,\mathbf{D})\neq g_i^0)\right]\right)=o(T^{-4}),$$

where $\mathcal{N}_\eta=\left\{\Theta\in\mathcal{B}^G:\left\|\theta_g-\theta_g^0\right\|^2<\eta,\ g\in\mathcal{G}\right\}$.

Proof of Theorem 3.1: By Lemma B.2, we have $(\hat{\Theta},\hat{\mathbf{D}})\in\mathcal{N}_\eta\times\Gamma^G$. Therefore, we can conclude that $\frac{1}{N}\sum_{i=1}^{N}\Pr(\hat{g}_i\neq g_i^0)=o(T^{-4})$ by Lemma B.3 Hence, we have

$$\Pr\left(\sup_i\mathbf{1}\left(\hat{g}_i\neq g_i^0\right)=1\right)\leq\sum_{i=1}^{N}\Pr(\hat{g}_i\neq g_i^0)=N\cdot o(T^{-4})=o\left(NT^{-4}\right).\ \blacksquare.$$

To prove Theorem 3.2, we need Lemmas B.4–A.7.

Lemma B.4. *Suppose w_{it} is any random variable with $\frac{1}{NT}\sum_{i,t}E\left\|w_{it}\right\|^{3+\epsilon}\leq C$ for some constant $\epsilon>0$ and $C>0$. Suppose Assumptions A.1–A.5 hold. Then*

$$\left\|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{1}(\hat{g}_i\neq g_i^0)w_{it}\right\|=o_p((NT)^{-1}).$$

To state the next lemma, we define an auxiliary estimator $\check{\Theta}(\mathbf{D})\equiv(\check{\theta}_1(\gamma_1)',...,\check{\theta}_G(\gamma_G)')'$, which is the least squares estimator of Θ with fixed $\mathbf{D}$ and true group specification $\mathbf{G}^0$, that is,

$$\check{\theta}_g(\gamma)=\left(\sum_{i\in\mathbf{G}_g^0}\sum_{t=1}^{T}\tilde{z}_{it}(\gamma)\tilde{z}_{it}(\gamma)'\right)^{-1}\left(\sum_{i\in\mathbf{G}_g^0}\sum_{t=1}^{T}\tilde{z}_{it}(\gamma)\tilde{y}_{it}\right)\text{ for }g\in\mathcal{G}.$$

Then the infeasible estimator is given by $\check{\Theta}=\check{\Theta}(\check{\mathbf{D}})$ with $\check{\mathbf{D}}=\arg\min_{\mathbf{D}\in\Gamma^G}\check{\mathcal{Q}}(\check{\Theta}(\mathbf{D}),\mathbf{D})$. See also (3.6) in Section 3.3.1. In the online supplemental material we derive the asymptotic properties of $\check{\Theta}$. The next lemma establishes the asymptotic equivalence by exploiting the properties of infeasible estimators.

Lemma B.5. *Suppose that Assumptions A.1–A.5 hold. Then $(N,T)\to\infty$ we have $\hat{\theta}_g=\check{\theta}_g(\hat{\gamma}_g)+o_p((NT)^{-1})$ for all $g\in\mathcal{G}$.*

Lemma B.6. *Suppose that Assumptions A.1–A.5 hold and $\alpha\in(0,1/3)$. Then $\alpha_{NT}(\hat{\gamma}_g-\gamma_g^0)=O_p(1)$ for all $g\in\mathcal{G}$.*

Lemma B.7. *Suppose that Assumptions A.1–A.5 hold. For any $\gamma=\gamma_g^0+O_p(1/\alpha_{NT})$ and $g\in\mathcal{G}$, the following statement holds:*

$$\check{\theta}_g(\gamma)-\check{\theta}_g(\gamma_g^0)=o_p((NT)^{-1/2})\text{ and }\check{Q}_g(\check{\theta}_g(\gamma),\gamma)-\check{Q}_g(\check{\theta}_g,\gamma)=o_p(1).$$

Proof of Theorem 3.2: For the first result, we can show $\sqrt{NT}[\check{\theta}_g(\hat{\gamma}_g)-\check{\theta}_g]\to 0$ by Lemmas B.5–B.7. It suffices to show the second result. Given Lemma B.6, we can denote

$\hat{\gamma}_g \equiv \gamma_g^0 + \hat{v}_g/\alpha_{N_gT}$ and $\check{\gamma}_g \equiv \gamma_g^0 + \check{v}_g/\alpha_{N_gT}$. Let

$$\begin{aligned} Q_{g,NT}^{**}(v_g) &\equiv \check{Q}_g(\check{\theta}_g(\hat{\gamma}_g), \gamma_g^0) - \check{Q}_g(\check{\theta}_g(\hat{\gamma}_g), \gamma_g^0 + v_g/\alpha_{N_gT}) \text{ and} && \text{(B.2)} \\ Q_{g,NT}^{*}(v_g) &\equiv \check{Q}_g(\check{\theta}_g, \gamma_g^0) - \check{Q}_g(\check{\theta}_g, \gamma_g^0 + v_g/\alpha_{N_gT}). && \text{(B.3)} \end{aligned}$$

First we show that $Q_{g,NT}^{**}(v) - Q_{g,NT}^{*}(v) \xrightarrow{p} 0$ uniformly on any compact set Ψ. It is straightforward to calculate that

$$Q_{g,NT}^{**}(v) - Q_{g,NT}^{*}(v) = L_{g,NT}^{*}(v) - L_{g,NT}(v),$$

where $L_{g,NT}(v)$ is a remainder term that is defined in Lemma C.14 in the online Supplementary Material and $L_{g,NT}^{*}(v)$ can be defined analogously. We show in the proof of Lemma C.14 that $L_{g,NT}^{*}(v) \xrightarrow{p} 0$ uniformly on any compact set Ψ. Similar arguments can be used to show that $L_{g,NT}^{*}(v) \xrightarrow{p} 0$ uniformly on any compact set Ψ. Therefore, we have $Q_{g,NT}^{**}(v) - Q_{g,NT}^{*}(v) \xrightarrow{p} 0$ uniformly on any compact set Ψ.

Next, we have

$$\begin{aligned} Q_{g,NT}^{**}(\hat{v}_g) &= \check{Q}_g(\check{\theta}_g(\hat{\gamma}_g), \gamma_g^0) - \check{Q}_g(\check{\theta}_g(\hat{\gamma}_g), \gamma_g^0 + \hat{v}_g/\alpha_{N_gT}) \\ &= \check{Q}_g(\check{\theta}_g, \gamma_g^0) - \left[\check{Q}_g(\check{\theta}_g(\check{\gamma}_g), \gamma_g^0 + \hat{v}_g/\alpha_{N_gT}) + o_p(1)\right] \\ &= \check{Q}_g(\check{\theta}_g, \gamma_g^0) - \check{Q}_g(\check{\theta}_g, \gamma_g^0 + \check{v}_g/\alpha_{N_gT}) + o_p(1) \\ &= Q_{g,NT}^{*}(\check{v}_g) + o_p(1) \\ &= \max_{v\in\mathbb{R}} Q_{g,NT}^{*}(v) + o_p(1), \end{aligned}$$

where the first and second equalities hold by (B.2) and Lemma B.7, respectively, the fourth equality holds by (B.3) and the fact that $\check{\theta}_g = \check{\theta}_g(\check{\gamma}_g)$, and the last equality follows from the definition of $\check{\gamma}_g$. On the other hand side, $Q_{g,NT}^{**}(\hat{v}_g) = Q_{g,NT}^{*}(\hat{v}_g) + o_p(1)$ by the uniform convergence of $Q_{g,NT}^{**}(v) - Q_{g,NT}^{*}(v)$ in probability to zero. It follows that

$$Q_{g,NT}^{*}(\hat{v}_g) = \max_{v\in\mathbb{R}} Q_{g,NT}^{*}(v) + o_p(1).$$

Noting that $Q_{g,NT}^{*}(\cdot)$ converges weakly to a continuous stochastic process that has a unique maximum and $\check{v}_g = \arg\max_{v\in\mathbb{R}} Q_{g,NT}^{*}(v)$, we must have

$$\hat{v}_g = \arg\max_{v\in\mathbb{R}} Q_{g,NT}^{*}(v) + o_p(1) = \check{v}_g + o_p(1),$$

which implies $\alpha_{N_gT}(\hat{\gamma}_g - \check{\gamma}_g) = o_p(1)$. ■

Lemma B.8. *Suppose Assumptions A.1(ii)–(vi) and A.3–A.6 hold. Let $\mathbb{M}_0 = I_T - \frac{1}{T}\iota_T\iota_T'$ with ι_T being a $T \times 1$ vector of ones.*
(i) Under Assumption A.1(i.1) we have $\frac{1}{\sqrt{N_gT}}\sum_{i\in\mathbf{G}_g^0} Z_i(\gamma_g^0)'\mathbb{M}_0\varepsilon_i + \sqrt{\frac{N_g}{T}}\mathbb{B}_{g,NT}(\gamma_g^0) \xrightarrow{d} N(0, \Omega_{g,1}^0)$, where $\mathbb{B}_{g,NT}(\gamma_g^0) = \frac{1}{N_gT}\sum_{i\in\mathbf{G}_g^0}\sum_{t=1}^{T}\sum_{s<t} E\left[z_{it}(\gamma_g^0)\varepsilon_{is}\right]$ for each $g \in \mathcal{G}$;
(ii) Under Assumption A.1(i.2) we have: $\frac{1}{\sqrt{N_gT}}\sum_{i\in\mathbf{G}_g^0} Z_i(\gamma_g^0)'\mathbb{M}_0\varepsilon_i \xrightarrow{d} N(0, \Omega_{g,2}^0)$, where $\Omega_{g,2}(\gamma_g^0, \gamma_g^0)$ is as defined in Assumption A.6.

Proof of Theorem 3.3. (i) By Theorem 3.2, we only need to consider the infeasible estimator $\check{\Theta}$. By Lemma B.7, we have that

$$\begin{aligned}\sqrt{N_g T}(\check{\theta}_g - \theta_g^0) &= \sqrt{N_g T}(\check{\theta}_g(\gamma_g^0) - \theta_g^0) + o_p(1) \\ &= \left(\frac{1}{N_g T}\sum_{i\in\mathbf{G}_g^0} Z_i(\gamma_g^0)'\mathbb{M}_0 Z_i(\gamma_g^0)\right)^{-1} \frac{1}{\sqrt{N_g T}}\sum_{i\in\mathbf{G}_g^0} Z_i(\gamma_g^0)'\mathbb{M}_0\varepsilon_i + o_p(1).\end{aligned}$$

Then the result follows from Lemma B.8 and Assumption A.6.

(ii) The result follows from Theorem 3.2 and Lemma C.14 in the online supplement. ■

Proof of Theorem 3.4. First, using Lemma B.3, we can readily show that $\hat{N}_g/N_g \xrightarrow{p} 1$ and $\check{Q}_g(\check{\theta}_g, \check{\gamma}_g)/(N_g T) \xrightarrow{p} \sigma_g^2$. Let $\bar{\theta}_g(\gamma)$ be the minimizer of $\bar{Q}_g(\theta,\gamma)$ that is defined in Section 3.4.1. Following the proof of Lemma B.5, we can also show that $\bar{\theta}_g(\gamma_g^0) = \check{\theta}_g\left(\gamma_g^0\right) + o_p((NT)^{-1})$. With this and using Lemma B.4, we can readily show that

$$\bar{Q}_g(\bar{\theta}_g(\gamma_g^0), \gamma_g^0) = \bar{Q}_g(\check{\theta}_g\left(\gamma_g^0\right), \gamma_g^0) + o_p(1) = \check{Q}_g(\check{\theta}_g(\gamma_g^0), \gamma_g^0) + o_p(1). \tag{B.4}$$

On the one hand, by the definitions of $(\check{\theta}_g, \check{\gamma}_g)$ and $\{(\hat{\theta}_g, \hat{\gamma}_g), g = 1, ..., G\}$, we have

$$\check{Q}_g(\check{\theta}_g, \check{\gamma}_g) \le \check{Q}_g(\hat{\theta}_g, \hat{\gamma}_g) \text{ and } \sum_{g=1}^{G} \bar{Q}_g(\hat{\theta}_g, \hat{\gamma}_g) \le \sum_{g=1}^{G} \bar{Q}_g(\check{\theta}_g, \check{\gamma}_g).$$

On the other hand, we can apply Lemma B.4 to show that

$$\bar{Q}_g(\theta,\gamma) = \check{Q}_g(\theta,\gamma) + o_p(1). \tag{B.5}$$

This, in conjunction with the first inequality in the above displayed equation implies that $\check{Q}_g(\check{\theta}_g, \check{\gamma}_g) \le \bar{Q}_g(\hat{\theta}_g, \hat{\gamma}_g) + o_p(1)$ and hence $\sum_{g=1}^{G} \check{Q}_g(\check{\theta}_g, \check{\gamma}_g) \le \sum_{g=1}^{G} \bar{Q}_g(\hat{\theta}_g, \hat{\gamma}_g) + o_p(1)$. Combining this last inequality with the second inequality in the above displayed equation yields

$$\sum_{g=1}^{G} \check{Q}_g(\check{\theta}_g, \check{\gamma}_g) = \sum_{g=1}^{G} \bar{Q}_g(\hat{\theta}_g, \hat{\gamma}_g) + o_p(1),$$

which, in conjunction with $\check{Q}_g(\check{\theta}_g, \check{\gamma}_g) \le \bar{Q}_g(\hat{\theta}_g, \hat{\gamma}_g) + o_p(1)$ for each $g \in \mathcal{G}$, implies that

$$\bar{Q}_g(\hat{\theta}_g, \hat{\gamma}_g) = \check{Q}_g(\check{\theta}_g, \check{\gamma}_g) + o_p(1). \tag{B.6}$$

Noting that $\check{\theta}_g(\gamma_g^0) - \theta_g^0 = [\sum_{i\in\mathbf{G}_g^0} Z_i(\gamma_g^0)'\mathbb{M}_0 Z_i(\gamma_g^0)]^{-1} \sum_{i\in\mathbf{G}_g^0} Z_i(\gamma_g^0)'\mathbb{M}_0\varepsilon_i$ and using the analysis of $\check{\theta}_g - \theta_g^0$ in the proof of Theorem 3.3, we can readily show that $\check{\theta}_g(\gamma_g^0) - \check{\theta}_g = o_p(1/\sqrt{NT})$. With this, we can also show that

$$\check{Q}_g(\check{\theta}_g(\gamma_g^0), \gamma_g^0) - \check{Q}_g(\check{\theta}_g, \gamma_g^0) = o_p(1). \tag{B.7}$$

Then we have

$$
\begin{aligned}
\bar{Q}_g(\bar{\theta}_g(\gamma_g^0),\gamma_g^0)-\bar{Q}_g(\hat{\theta}_g,\hat{\gamma}_g) &= \check{Q}_g(\bar{\theta}_g(\gamma_g^0),\gamma_g^0)-\check{Q}_g(\check{\theta}_g,\check{\gamma}_g)+o_p(1)\\
&= \check{Q}_g(\check{\theta}_g(\gamma_g^0),\gamma_g^0)-\check{Q}_g(\check{\theta}_g,\check{\gamma}_g)+o_p(1)\\
&= [\check{Q}_g(\check{\theta}_g,\gamma_g^0)-\check{Q}_g(\check{\theta}_g,\check{\gamma}_g)]+[\check{Q}_g(\check{\theta}_g(\gamma_g^0),\gamma_g^0)-\check{Q}_g(\check{\theta}_g,\gamma_g^0)]+o_p(1)\\
&= \check{Q}_g(\check{\theta}_g,\gamma_g^0)-\check{Q}_g(\check{\theta}_g,\check{\gamma}_g)+o_p(1), \qquad \text{(B.8)}
\end{aligned}
$$

where the first equality follows from (B.5) and (B.6), the second and last equalities hold by (B.4) and (B.7), respectively. By Lemma C.14 in the online Supplementary Material, we have

$$
\check{Q}_g(\check{\theta}_g,\gamma_g^0)-\check{Q}_g(\check{\theta}_g,v/\alpha_{N_gT}+\gamma_g^0)\Rightarrow -\pi_g^{2\alpha}w_{g,D}\left|v\right|+2\sqrt{w_{g,V}\pi_g^{2\alpha}}W_g(v),
$$

where $w_{g,D}\equiv C_g^{0\prime}D_g^0C_g^0$ and $w_{g,V}\equiv C_g^{0\prime}V_g^0C_g^0$. Then by the continuous mapping theorem (CMT),

$$
\begin{aligned}
\check{Q}_g(\check{\theta}_g,\gamma_g^0)-\check{Q}_g(\check{\theta}_g,\check{v}_g/\alpha_{N_gT}+\gamma_g^0) &\Rightarrow \max_{v\in\mathbb{R}}\left[-\pi_g^{2\alpha}w_{g,D}\left|v\right|+2\sqrt{w_{g,V}\pi_g^{2\alpha}}W_g(v)\right]\\
&= \frac{w_{g,V}}{w_{g,D}}\max_{v\in\mathbb{R}}\left[-\frac{w_{g,D}^2}{w_{g,V}}\left|\pi_g^{2\alpha}v\right|+2\sqrt{\frac{w_{g,D}^2\pi_g^{2\alpha}}{w_{g,V}}}W_g(v)\right]\\
&= \frac{w_{g,V}}{w_{g,D}}\max_{v\in\mathbb{R}}\left[-\left|\frac{w_{g,D}^2}{w_{g,V}}\pi_g^{2\alpha}v\right|+2W_g(\frac{w_{g,D}^2}{w_{g,V}}\pi_g^{2\alpha}v)\right]\\
&= \frac{w_{g,V}}{w_{g,D}}\max_{r\in\mathbb{R}}\left[-\left|r\right|+2W_g(r)\right], \qquad \text{(B.9)}
\end{aligned}
$$

where the second equality holds by the distributional equality $aW_g(v)=W_g(a^2v)$ and the last equality follows from the change of variable (by setting $r\equiv\frac{w_{g,D}^2}{w_{g,V}}\pi_g^{2\alpha}v$). Lastly, we have

$$
\begin{aligned}
\mathcal{L}_{g,NT}\left(\gamma_g^0\right) &= \frac{\bar{Q}_g(\bar{\theta}_g(\gamma_g^0),\gamma_g^0)-\bar{Q}_g(\hat{\theta}_g,\hat{\gamma}_g)}{\bar{Q}_g(\hat{\theta}_g,\hat{\gamma}_g)/(N_gT)}=\frac{\check{Q}_g(\check{\theta}_g,\gamma_g^0)-\check{Q}_g(\check{\theta}_g,\check{v}_g/\alpha_{N_gT}+\gamma_g^0)}{\check{Q}_g(\check{\theta}_g,\check{\gamma}_g)/(N_gT)}+o_P\left(1\right)\\
&\overset{d}{\to}\frac{w_{g,V}}{\sigma_g^2w_{g,D}}\max_{r\in\mathbb{R}}\left[-\left|r\right|+2W_g(r)\right],
\end{aligned}
$$

where the first equality holds by (B.8) and (B.6), and the convergence follows from (B.9) and the fact that $\check{Q}_g(\check{\theta}_g,\check{\gamma}_g)/(N_gT)=\sigma_g^2+o_p\left(1\right).$ ■

Proof of Theorem 3.5. Under the null hypothesis, one can study the asymptotic property of $(\hat{\Theta}_r,\hat{\mathbf{D}}_r,\hat{\mathbf{G}}_r)$ similar to that of $(\hat{\Theta},\hat{\mathbf{D}},\hat{\mathbf{G}})$. Following the arguments as used in the proof of Lemma B.5, we can show that

$$
Q(\hat{\Theta}_r,\hat{\mathbf{D}}_r,\hat{\mathbf{G}}_r)=\check{Q}(\check{\Theta}(\check{\mathbf{D}}_r),\check{\mathbf{D}}_r)+o_p(1),
$$

where $\check{\mathbf{D}}_r=\arg\min_{\mathbf{D}\in\mathcal{D}_r}\check{Q}(\check{\Theta}(\mathbf{D}),\mathbf{D})$. This, in conjunction with the fact that $Q(\hat{\Theta},\hat{\mathbf{D}},\hat{\mathbf{G}})=$

$\check{Q}(\check{\Theta}, \check{\mathbf{D}}) + o_p(1)$, implies that

$$\begin{aligned} Q(\hat{\Theta}_r, \hat{\mathbf{D}}_r, \hat{\mathbf{G}}_r) - Q(\hat{\Theta}, \hat{\mathbf{D}}, \hat{\mathbf{G}}) &= \check{Q}(\check{\Theta}(\check{\mathbf{D}}_r), \check{\mathbf{D}}_r) - \check{Q}(\check{\Theta}, \check{\mathbf{D}}) + o_p(1) \\ &= \left[\check{Q}(\check{\Theta}(\check{\mathbf{D}}_r), \check{\mathbf{D}}_r) - \check{Q}(\check{\Theta}(\check{\mathbf{D}}_r), \mathbf{D}^0)\right] + \left[\check{Q}(\check{\Theta}, \mathbf{D}^0) - \check{Q}(\check{\Theta}, \check{\mathbf{D}})\right] \\ &\quad + \left[\check{Q}(\check{\Theta}(\check{\mathbf{D}}_r), \mathbf{D}^0) - \check{Q}(\check{\Theta}, \mathbf{D}^0)\right] + o_p(1) \\ &= \left[\check{Q}(\check{\Theta}, \mathbf{D}^0) - \check{Q}(\check{\Theta}, \check{\mathbf{D}})\right] - \left[\check{Q}(\check{\Theta}(\check{\mathbf{D}}_r), \mathbf{D}^0) - \check{Q}(\check{\Theta}(\check{\mathbf{D}}_r), \check{\mathbf{D}}_r)\right] + o_p(1), \end{aligned}$$

where we use the fact that $\check{Q}(\check{\Theta}(\check{\mathbf{D}}_r), \mathbf{D}^0) - \check{Q}(\check{\Theta}, \mathbf{D}^0) = o_p(1)$ that can be proved by following the same arguments as used to derive (B.7).

For $\check{Q}(\check{\Theta}, \mathbf{D}^0) - \check{Q}(\check{\Theta}, \check{\mathbf{D}})$, we have that under $H_{02} : \mathbf{D}^0 \in \mathcal{D}_r$ (i.e., $\gamma_1^0 = ... = \gamma_G^0 = \gamma^0$),

$$\begin{aligned} \check{Q}(\check{\Theta}, \mathbf{D}^0) - \check{Q}(\check{\Theta}, \check{\mathbf{D}}) &= \sum_{g=1}^{G} \left[\check{Q}_g(\check{\theta}_g, \gamma^0) - \check{Q}(\check{\theta}_g, \check{v}_g/\alpha_{N_gT} + \gamma^0)\right] = \sum_{g=1}^{G} Q_{g,NT}^*(\check{v}_g) \\ &\Rightarrow \sum_{g=1}^{G} \frac{w_{g,V}}{w_{g,D}} \max_{v_g \in \mathbb{R}} \left[- \left| \frac{w_{g,D}^2}{w_{g,V}} \pi_g^{2\alpha} v_g \right| + 2W_g(\frac{w_{g,D}^2}{w_{g,V}} \pi_g^{2\alpha} v_g) \right] \\ &= \sum_{g=1}^{G} \frac{w_{g,V}}{w_{g,D}} \max_{v_g \in \mathbb{R}} \left[- |v_g| + 2W_g(v_g) \right] \end{aligned}$$

by Lemma C.13 in the online supplement. Writing $\check{\mathbf{D}}_r = (\gamma^0 + \check{v}_r/\alpha_{NT}, ..., \gamma^0 + \check{v}_r/\alpha_{NT})'$, we have that under $H_{02} : \mathbf{D}^0 \in \mathcal{D}_r$,

$$\begin{aligned} \check{Q}(\check{\Theta}(\check{\mathbf{D}}_r), \mathbf{D}^0) - \check{Q}(\check{\Theta}(\check{\mathbf{D}}_r), \check{\mathbf{D}}_r) &= \sum_{g=1}^{G} \left[\check{Q}_g(\check{\theta}_g(\check{\mathbf{D}}_r), \gamma^0) - \check{Q}(\check{\theta}_g(\check{\mathbf{D}}_r), \pi_g^{1-2\alpha} \check{v}_r/\alpha_{N_gT} + \gamma^0)\right] \\ &= \sum_{g=1}^{G} Q_{g,NT}^*(\pi_g^{1-2\alpha} \check{v}_r) \\ &\Rightarrow \max_{v \in \mathbb{R}} \left(\sum_{g=1}^{G} \frac{w_{g,V}}{w_{g,D}} \left[- \left| \frac{w_{g,D}^2}{w_{g,V}} \pi_g v \right| + 2W_g(\frac{w_{g,D}^2}{w_{g,V}} \pi_g v) \right] \right) \\ &= \max_{u \in \mathbb{R}} \left(\sum_{g=1}^{G} \frac{w_{g,V}}{w_{g,D}} \left[- \left| \frac{w_{g,D}}{w_{1,D}} \frac{\sigma^2 w_{g,D}}{w_{g,V}} \pi_g u \right| + 2W_g(\frac{w_{g,D}}{w_{1,D}} \frac{\sigma^2 w_{g,D}}{w_{g,V}} \pi_g u) \right] \right), \end{aligned}$$

where the last equality is obtained by changing variable $u = v \cdot \sigma^2/w_{1,D}$. This completes our proof. ■

Proof of Theorem 3.6. This proof is analogous to the first half of that of Theorem 3.5 and thus omitted. ■

Proof of Theorem 3.7. Following the arguments as used in the proof of Theorem 3.2, the Wald test statistic is asymptotically equivalent to the infeasible Wald test statistic uniformly for $\mathbf{D}$. Therefore, we can focus on the study of the asymptotic property of the infeasible Wald test statistic. To avoid introducing new notations, we just assume $\hat{\mathbf{G}} = \mathbf{G}^0$, which occurs w.p.a.1. Then $\bar{\theta}_g^{\mathrm{bc}}(\gamma_g) = \check{\theta}_g^{\mathrm{bc}}(\gamma_g)$ w.p.a.1., where $\check{\theta}_g^{\mathrm{bc}}(\gamma_g)$ is the bias-corrected version of $\check{\theta}_g(\gamma_g)$ when necessary (e.g., in the dynamic case) and $\check{\theta}_g(\gamma_g)$ is defined before Theorem 3.2. Similarly, let $\check{\Theta}^{\mathrm{bc}}(\mathbf{D})$ be the bias corrected version of $\check{\Theta}(\mathbf{D})$ when necessary

For $g \in \mathcal{G}$, we can readily establish

$$\begin{aligned}\sqrt{N_g T}\left[\check{\theta}_g^{\mathrm{bc}}(\gamma_g) - \theta_g^0\right] &= \omega_{g,NT}(\gamma_g,\gamma_g)^{-1}\frac{1}{\sqrt{N_g T}}\sum_{i\in\mathbf{G}_g^0}\sum_{t=1}^{T}\tilde{z}_{it}(\gamma_g)[\tilde{x}_{it}(\gamma_g^0) - \tilde{x}_{it}(\gamma_g)]'\delta_g^0 \\ &\quad + \omega_{g,NT}(\gamma_g,\gamma_g)^{-1}S_{g,NT}(\gamma) + o_p(1).\end{aligned}$$

Note that $\omega_{g,NT}(\gamma_g,\gamma_g) \xrightarrow{p} \omega_g(\gamma_g,\gamma_g)$ uniformly in γ_g by Assumption A.6 and $S_{g,NT}(\gamma) \Rightarrow S_g(\gamma)$ on Γ by Assumption A.7. In addition, by Assumption A.6,

$$\begin{aligned}\frac{1}{\sqrt{N_g T}}\sum_{i\in\mathbf{G}_g^0}\sum_{t=1}^{T}\tilde{z}_{it}(\gamma_g)[\tilde{x}_{it}(\gamma_g^0) - \tilde{x}_{it}(\gamma_g)]'\delta_g^0 &= \sqrt{\pi_g}\frac{1}{N_g T}\sum_{i\in\mathbf{G}_g^0}\sum_{t=1}^{T}\tilde{z}_{it}(\gamma_g)[\tilde{z}_{it}(\gamma_g^0) - \tilde{z}_{it}(\gamma_g)]'L'c_g \\ &= \sqrt{\pi_g}\frac{1}{N_g T}\sum_{i\in\mathbf{G}_g^0}\sum_{t=1}^{T}\left[\tilde{z}_{it}(\gamma_g)\tilde{z}_{it}(\gamma_g^0)' - \tilde{z}_{it}(\gamma_g)\tilde{z}_{it}(\gamma_g)'\right]L'c_g \\ &\xrightarrow{p} \sqrt{\pi_g}\left[\omega_g(\gamma_g,\gamma_g^0) - \omega_g(\gamma_g,\gamma_g)\right]L'c_g,\end{aligned}$$

where the convergence follows by Assumption A.6. Then under $\mathbb{H}_{1NT} : \mathbb{L}\Theta^0 = \mathbf{c}/\sqrt{NT}$,

$$\begin{aligned}\sqrt{N_g T}L\check{\theta}_g^{\mathrm{bc}}(\gamma_g) &\Rightarrow \sqrt{N_g T}L\theta_g^0 + L\omega_g(\gamma_g,\gamma_g)^{-1}\left\{S_g(\gamma_g) + \sqrt{\pi_g}\left[\omega_g(\gamma_g,\gamma_g^0) - \omega_g(\gamma_g,\gamma_g)\right]L'c_g\right\} \\ &= \sqrt{\pi_g}c_g + L\omega_g(\gamma_g,\gamma_g)^{-1}[S_g(\gamma_g) + \sqrt{\pi_g}\omega_g(\gamma_g,\gamma_g^0)L'c_g] - \sqrt{\pi_g}LL'c_g \\ &= L\omega_g(\gamma_g,\gamma_g)^{-1}[S_g(\gamma_g) + \sqrt{\pi_g}\omega_g(\gamma_g,\gamma_g^0)L'c_g].\end{aligned}$$

Then by the CMT, we can conclude that

$$\sqrt{NT}\mathbb{L}\hat{\Pi}^{1/2}\check{\Theta}^{\mathrm{bc}}(\mathbf{D}) = \begin{pmatrix}\sqrt{N_{g_1}T}L\check{\theta}_{g_1}^{\mathrm{bc}}(\gamma_{g_1}) \\ \vdots \\ \sqrt{N_{g_P}T}L\check{\theta}_{g_P}^{\mathrm{bc}}(\gamma_{g_P})\end{pmatrix} \Rightarrow \mathbb{L}\omega(\mathbf{D})^{-1}\left[\mathbf{S}(\mathbf{D}) + \mathbf{Q}(\mathbf{D})\Pi^{1/2}\mathbb{L}'\mathbf{c}\right].$$

It is standard to show that $\hat{\mathbb{K}}_{NT}(\mathbf{D}) \xrightarrow{p} \mathbb{L}\omega(\mathbf{D})^{-1}\mathbf{\Omega}(\mathbf{D})\omega(\mathbf{D})^{-1}\mathbb{L}'$ uniformly in $\mathbf{D}$. Then we have $W_{NT}(\gamma) \Rightarrow W^c(\gamma)$ by the CMT. ■

Proof of Theorem 3.8. Using Theorem 3.2 and the analysis of the infeasible estimators in Section C of the online supplement, we can readily show that $\hat{\sigma}^2\left(G^0\right) \xrightarrow{p} \sigma^2$ as $(N,T) \to \infty$. Then $IC(G^0) = \ln\left(\hat{\sigma}^2\left(G^0\right)\right) + \lambda_{NT}G^0K \to \sigma^2$ by Assumption D.2(ii) in the online supplement, where σ^2 $=\lim_{(N,T)\to\infty}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}E(\varepsilon_{it}^2)$. When $1 \le G < G^0$, by Assumption D.2(ii) we have that w.p.a.1. $IC(G) = \ln\left(\hat{\sigma}^2\left(G\right)\right) + \lambda_{NT}GK \ge \ln(\bar{\sigma}^2) > \ln(\sigma^2)$ as $(N,T) \to \infty$. So we have

$$\Pr(\hat{G} < G^0) = \Pr(\exists 1 \le G < G^0,\ IC(G) < IC(G^0)) \to 0 \text{ as } (N,T) \to \infty. \qquad \text{(B.10)}$$

Next, we consider the case where $G^0 < G \le G_{\max}$. When $G > G^0$, we have by Proposition D.1 in the online supplement that $\max_{G^0<G\le G_{\max}}[\hat{\sigma}^2\left(G\right) - \hat{\sigma}^2(G^0)] = O_p(T^{-1})$.

It follows that

$$\begin{aligned}\Pr(\hat{G} > G^0) &= \Pr(\exists G^0 < G \leq G_{\max},\ IC(G) < IC(G^0)) \\ &= \Pr(\exists G^0 < G \leq G_{\max}, T[\ln(\hat{\sigma}^2\,(G)) - \ln(\hat{\sigma}^2(G^0))] > (G - G^0)T\lambda_{NT}) \\ &\to 0 \text{ as } (N,T) \to \infty, \qquad \text{(B.11)}\end{aligned}$$

where the last line follows from the fact that $T[\ln(\hat{\sigma}^2\,(G)) - \ln(\hat{\sigma}^2(G^0))] = T\ln(1 + \frac{\hat{\sigma}^2(G)-\hat{\sigma}^2(G^0)}{\hat{\sigma}^2(G^0)}) = O(T(\hat{\sigma}^2\,(G) - \hat{\sigma}^2(G^0)) = O_p\,(1)$ and $T\lambda_{NT} \to \infty$ as $(N,T) \to \infty$ by Assumption D.2(ii). Combining (B.10) and (B.11), we have $\Pr(\hat{G} = G^0) \to 1$ as $(N,T) \to \infty$. ■

Appendix C

Technical Results for Chapter 4

Proof of the main results

Proof of Proposition 4.1: (i) Under Assumption A.1 (vi), the MA(∞) representation in equation (4.4) is valid. However, X_{t+1} is a high dimensional vector and the properties of y_{it} have to be examined more carefully. By Assumption A.1 (iv), $y_{it}^{(u)}$'s and $y_{it}^{(f)}$'s are mutually independent. It suffices to study them separately. By Assumption A.1 (i), we can write $y_{it}^{(u)}$ as a linear process:

$$y_{it}^{(u)} = \sum_{j=0}^{\infty} \alpha_{iN}^{(u)}(j) u_{t-j} = \sum_{j=0}^{\infty} \alpha_{iN}^{(u)}(j) C^{(u)} \epsilon_{t-j}^{(u)} \equiv \sum_{j=0}^{\infty} C_j^{(i,u)} \epsilon_{t-j}^{(u)},$$

where $C_j^{(i,u)} \equiv \alpha_{iN}^{(u)}(j) C^{(u)}$. Under Assumption A.1 (vi), one can bound $|(e_{1,p} \otimes e_{i,N})' \Phi^j|$ by $\Psi_{\max}([\Phi^j]_{[N],[N]}) \leq \bar{c}\rho^j$. It follows that $|\alpha_{iN}^{(u)}(j)| \leq \bar{c}\rho^j$. Then the MA($\infty$) representation of $y_{it}^{(u)}$ is valid with $E(y_{it}^{(u)}) = 0$ and $\text{Var}(y_{it}^{(u)}) = \sum_{j=0}^{\infty} \alpha_{iN}^{(u)}(j) \Sigma_u \alpha_{iN}^{(u)}(j)' < \infty$.

Under Assumption A.1(vi), we can also show that $E(|y_{it}^{(f)}|) = \sum_{j=0}^{\infty} |\alpha_{iN}^{(f)}(j)\mu_f| < \infty$. The MA($\infty$) representation of $y_{it}^{(f)}$ is

$$y_{it}^{(f)} = E(y_{it}^{(f)}) + \sum_{j=0}^{\infty} \alpha_{iN}^{(f)}(j)(f_{t-j}^{0} - \mu_f) = E(y_{it}) + \sum_{j=0}^{\infty} C_j^{(i,f)} \epsilon_{t-j}^{(f)},$$

where $C_j^{(i,f)} \equiv \sum_{k=0}^{j} \alpha_{iN}^{(f)}(k) C_{j-k}^{(f)}$. Under Assumption A.1(vi), $|C_j^{(i,f)}| \leq \sum_{k=0}^{j} |\alpha_{iN}^{(f)}(k)| \cdot ||C_{j-k}^{(f)}||_{\text{op}}$. In addition, by Assumption A.1(ii),

$$\sum_{j=0}^{\infty} \sum_{k=0}^{j} \rho^k ||C_{j-k}^{(f)}||_{\max} = \sum_{k=0}^{\infty} \rho^k \sum_{j=k}^{\infty} ||C_{j-k}^{(f)}||_{\max} \leq \bar{c} \sum_{k=0}^{\infty} \rho^k (k+1)^{-\alpha},$$

for some constant $\bar{c} < \infty$. Hence $C_j^{(i,f)}$ is absolutely summable, $\text{Var}(y_{it}^{(f)}) = \sum_{j=0}^{\infty} C_j^{(i,f)} C_j^{(i,f)\prime} < \infty$, and the MA($\infty$) representation of $y_{it}^{(f)}$ is valid.

Similar to the decomposition (4.5), we can write $X_t = X_t^{(u)} + X_t^{(f)}$. For Σ_X, due to the independence between $X_t^{(u)}$ and $X_t^{(f)}$, we can also write it as $\Sigma_X = \Sigma_X^{(f)} + \Sigma_X^{(u)}$, where $\Sigma_X^{(u)} \equiv E(X_t^{(u)} X_t^{(u)\prime})$ and $\Sigma_X^{(f)} \equiv E(X_t^{(f)} X_t^{(f)\prime})$. By the fact that $\Sigma_X^{(f)}$ is positive semi definite, we have $\psi_{\min}(\Sigma_X) \geq \psi_{\min}(\Sigma_X^{(u)})$. It suffices to show $\Psi_{\min}(\Sigma_X^{(u)})$ is bounded below.

By Proposition 2.3 of BM (2015), we have

$$\psi_{\min}(\Sigma_X^{(u)}) \geq \frac{\psi_{\min}(\Sigma_u)}{\max_{|z|=1\psi\max}(\mathcal{A}^*(z)\mathcal{A}(z))}.$$

Given Assumption A.1 (vii), we have that $\psi_{\min}(\Sigma_X^{(u)})$ is bounded below by some constant.

(ii) By the independence between $X_t^{(u)}$ and $X_t^{(f)}$, one can also show that $\psi_{\min}(\Sigma) \geq \psi_{\min}(\Sigma_X^{(u)})$. ■

Theoretical analysis of the first-step estimators

Lemma C.1. *For the $T \times N$ matrices Θ^0 and Δ, we have*

(i) $\left\|\Theta^0 + \mathcal{M}(\Delta)\right\|_* = \left\|\Theta^0\right\|_* + \|\mathcal{M}(\Delta)\|_*$*;*

(ii) $\|\Delta\|_F^2 = \|\mathcal{M}(\Delta)\|_F^2 + \|\mathcal{P}(\Delta)\|_F^2$*;*

(iii) $rank(\mathcal{P}(\Delta)) \leq 2R^0$*;*

(iv) $\|\Delta\|_F^2 = \sum_j \psi_j(\Delta)^2$ *and* $\|\Delta\|_*^2 \leq \|\Delta\|_F^2 rank(\Delta)$;

For any conformable matrices M_1 and M_2, the following statement holds:

(v) $|tr(M_1M_2)| \leq \|M_1\|_{max} |vec(M_2)|_1$ *and* $|tr(M_1M_2)]| \leq \|M_1\|_{op} \|M_2\|_*$.

Lemma C.2. *Suppose that Assumption A.1 holds. There exists absolute constants c, $\underline{c}$, $\bar{c} \in (0,\infty)$ such that*

(i) $\|\mathbf{U}'\mathbf{X}\|_{max}/T \leq \gamma_1/2$ *with probability greater than* $1 - \bar{c}(N^2T^{1-q/4}(logN)^{-q/2} + N^{2-\underline{c}logN})$*;*

(ii) $\|\mathbf{U}'\mathbb{P}_{F^0}\mathbf{X}\|_{max}/T \leq c \cdot \gamma_1$ *with probability greater than* $1 - \bar{c}(NT^{1-q/4}(logN)^{-q/2} + N^{1-\underline{c}logN})$.

Proof of Theorem 4.1. Let $\tilde{\Delta}^{(1)} = \tilde{B} - B^0$, $\tilde{\Delta}^{(2)} = \tilde{\Theta} - \Theta^0$ and the event $\mathcal{E}_{NT}^{(1)} = \{\|\mathbf{U}'\mathbf{X}\|_{\max}/T \leq \gamma_1/2, \|\mathbf{U}\|_{\text{op}}/\sqrt{NT} \leq \gamma_2/2\}$. By Lemma C.2, and Assumption A.3(i), $\mathcal{E}_{NT}^{(1)}$ holds with probability at least $1 - \bar{c}[N^2T^{1-q/2}(\text{log}N)^{-q/2} + N^{2-\underline{c}\text{log}N}]$. By the definition of $(\tilde{B}, \tilde{\Theta})$, we have that

$$\begin{aligned} 0 &\geq \mathcal{L}(\tilde{B},\tilde{\Theta}) - \mathcal{L}(B^0,\Theta^0) \\ &= \frac{1}{2NT}(||\mathbf{Y} - \mathbf{X}\tilde{B} - \tilde{\Theta}||_{\text{F}}^2 - ||\mathbf{U}||_{\text{F}}^2) + \frac{\gamma_1}{N}(|\text{vec}(\tilde{B})|_1 - |\text{vec}(B^0)|_1) + \frac{\gamma_2}{\sqrt{NT}}(||\tilde{\Theta}||_* - ||\Theta^0||_*) \\ &= d_1 + d_2 + d_3. \end{aligned} \tag{C.1}$$

To establish the asymptotic property of $\tilde{B}$ and $\tilde{\Theta}$, we have to study the three terms d_1, d_2 and d_3 in order.

First, we consider d_1. By the identity $\mathbf{Y} = \mathbf{X}B^0 + \Theta^0 + \mathbf{U}$, we have

$$\left\|\mathbf{Y} - \mathbf{X}\tilde{B} - \tilde{\Theta}\right\|_{\text{F}}^2 - \|\mathbf{U}\|_{\text{F}}^2 = \left\|\mathbf{X}\tilde{\Delta}^{(1)} + \tilde{\Delta}^{(2)}\right\|_{\text{F}}^2 - 2\text{tr}[\mathbf{U}'(\mathbf{X}\tilde{\Delta}^{(1)} + \tilde{\Delta}^{(2)})].$$

For $\text{tr}[\mathbf{U}'(\mathbf{X}\tilde{\Delta}^{(1)} + \tilde{\Delta}^{(2)})]$, conditional on the event $\mathcal{E}_{NT}^{(1)}$, we have that

$$\begin{aligned} \frac{1}{NT}|\text{tr}[\mathbf{U}'(\mathbf{X}\tilde{\Delta}^{(1)} + \tilde{\Delta}^{(2)})]| &\leq \frac{1}{NT}\|\mathbf{U}'\mathbf{X}\|_{\max}|\text{vec}(\tilde{\Delta}^{(1)})|_1 + \frac{1}{NT}\|\mathbf{U}\|_{\text{op}}\left\|\tilde{\Delta}^{(2)}\right\|_* \\ &\leq \frac{\gamma_1}{2N}|\text{vec}(\tilde{\Delta}^{(1)})|_1 + \frac{\gamma_2}{2\sqrt{NT}}\left\|\tilde{\Delta}^{(2)}\right\|_*, \end{aligned}$$

where the first inequality holds by the triangle inequality and Lemma C.1(v). It follows that

$$\begin{aligned}
d_1 &\geq \frac{1}{2NT}\left\|\mathbf{X}\tilde{\Delta}^{(1)}+\tilde{\Delta}^{(2)}\right\|_{\mathrm{F}}^2-\frac{\gamma_1}{2N}|\mathrm{vec}(\tilde{\Delta}^{(1)})|_1-\frac{\gamma_2}{2\sqrt{NT}}\left\|\tilde{\Delta}^{(2)}\right\|_*\\
&\geq \frac{1}{2NT}\left\|\mathbf{X}\tilde{\Delta}^{(1)}+\tilde{\Delta}^{(2)}\right\|_{\mathrm{F}}^2-\frac{\gamma_1}{2N}\sum_{i=1}^{N}\left(|\tilde{\Delta}^{(1)}_{J_i,i}|_1+|\tilde{\Delta}^{(1)}_{J_i^c,i}|_1\right)\\
&\quad -\frac{\gamma_2}{2\sqrt{NT}}\left(\left\|\mathcal{P}(\tilde{\Delta}^{(2)})\right\|_*+\left\|\mathcal{M}(\tilde{\Delta}^{(2)})\right\|_*\right). \qquad (C.2)
\end{aligned}$$

Next, we consider d_2. By the identities $|\tilde{B}_{*,i}|_1=|\tilde{B}_{J_i,i}|_1+|\tilde{B}_{J_i^c,i}|_1$ and $|B^0_{*,i}|_1=|B^0_{J_i,i}|_1$, we have

$$\begin{aligned}
d_2 &= \frac{\gamma_1}{N}\sum_{i=1}^{N}(|\tilde{B}_{J_i,i}|_1+|\tilde{B}_{J_i^c,i}|_1-|B^0_{J_i,i}|_1)\\
&\geq \frac{\gamma_1}{N}\sum_{i=1}^{N}(|\tilde{\Delta}^{(1)}_{J_i^c,i}|_1-|\tilde{\Delta}^{(1)}_{J_i,i}|_1). \qquad (C.3)
\end{aligned}$$

where we use the fact that $|\tilde{B}_{J_i,i}|_1+|\tilde{\Delta}^{(1)}_{J_i,i}|_1\geq|B^0_{J_i,i}|_1$ by the triangle inequality and that $B^0_{J_i^c,i}=0$.

Now, we consider d_3. By the triangle inequality and Lemma C.1(i), we have

$$\begin{aligned}
\left\|\tilde{\Theta}\right\|_* &= \left\|\tilde{\Delta}^{(2)}+\Theta^0\right\|_*=\left\|\Theta^0+\mathcal{P}(\tilde{\Delta}^{(2)})+\mathcal{M}(\tilde{\Delta}^{(2)})\right\|_*\\
&\geq \left\|\Theta^0+\mathcal{M}(\tilde{\Delta}^{(2)})\right\|_*-\left\|\mathcal{P}(\tilde{\Delta}^{(2)})\right\|_*\\
&= \|\Theta^0\|_*+\left\|\mathcal{M}(\tilde{\Delta}^{(2)})\right\|_*-\left\|\mathcal{P}(\tilde{\Delta}^{(2)})\right\|_*.
\end{aligned}$$

It follows that

$$d_3\geq\frac{\gamma_2}{\sqrt{NT}}(\left\|\mathcal{M}(\tilde{\Delta}^{(2)})\right\|_*-\left\|\mathcal{P}(\tilde{\Delta}^{(2)})\right\|_*). \qquad (C.4)$$

Combining the results in (C.1)-(C.4), we have

$$\begin{aligned}
&\frac{1}{2NT}\left\|\mathbf{X}\tilde{\Delta}^{(1)}+\tilde{\Delta}^{(2)}\right\|_{\mathrm{F}}^2+\frac{\gamma_1}{2N}\sum_{i=1}^{N}||\tilde{\Delta}^{(1)}_{J_i^c,i}||_1+\frac{\gamma_2}{2\sqrt{NT}}\left\|\mathcal{M}(\tilde{\Delta}^{(2)})\right\|_*\\
\leq\ &\frac{3\gamma_1}{2N}\sum_{i=1}^{N}||\tilde{\Delta}^{(1)}_{J_i,i}||_1+\frac{3\gamma_2}{2\sqrt{NT}}\left\|\mathcal{P}(\tilde{\Delta}^{(2)})\right\|_*. \qquad (C.5)
\end{aligned}$$

The above inequality indicates that $(\tilde{\Delta}^{(1)},\tilde{\Delta}^{(2)})\in\mathcal{C}_{NT}(3)$. By Assumption A.2, we obtain that

$$\frac{1}{N}\left\|\tilde{\Delta}^{(1)}\right\|_{\mathrm{F}}^2+\frac{1}{NT}\left\|\tilde{\Delta}^{(2)}\right\|_{\mathrm{F}}^2\leq\kappa_3\frac{1}{NT}\left\|\mathbf{X}\tilde{\Delta}^{(1)}+\tilde{\Delta}^{(2)}\right\|_{\mathrm{F}}^2. \qquad (C.6)$$

By the inequality (C.5), we have

$$\begin{aligned}
\frac{1}{NT}\left\|\mathbf{X}\tilde{\Delta}^{(1)}+\tilde{\Delta}^{(2)}\right\|_{\mathrm{F}}^2 &\leq \frac{3\gamma_1}{N}\sum_{i=1}^{N}|\tilde{\Delta}^{(1)}_{J_i,i}|_1+\frac{3\gamma_2}{\sqrt{NT}}\left\|\mathcal{P}(\tilde{\Delta}^{(2)})\right\|_*\\
&\leq 3\gamma_1\sqrt{K_J}\frac{\left\|\tilde{\Delta}^{(1)}\right\|_{\mathrm{F}}}{\sqrt{N}}+3\sqrt{2R^0}\gamma_2\frac{\left\|\tilde{\Delta}^{(2)}\right\|_{\mathrm{F}}}{\sqrt{NT}}\\
&\leq 3(\gamma_1\sqrt{K_J}\vee(\sqrt{2R^0}\gamma_2))\sqrt{\frac{1}{N}\left\|\tilde{\Delta}^{(1)}\right\|_{\mathrm{F}}^2+\frac{1}{NT}\left\|\tilde{\Delta}^{(2)}\right\|_{\mathrm{F}}^2},\qquad (C.7)
\end{aligned}$$

where the second inequality is by Lemma C.1(ii)-(iv) and the inequality $\sum_{i=1}^{N}|\tilde{\Delta}_{J_i,i}^{(1)}|_1 \leq \sqrt{K_J N}(\sum_{i=1}^{N}|\tilde{\Delta}_{J_i,i}^{(1)}|^2)^{1/2} \leq \sqrt{K_J N}\left\|\tilde{\Delta}^{(1)}\right\|_{\mathrm{F}}$. Combining (C.6)-(C.7) yields

$$\frac{1}{N}\left\|\tilde{\Delta}^{(1)}\right\|_{\mathrm{F}}^2+\frac{1}{NT}\left\|\tilde{\Delta}^{(2)}\right\|_{\mathrm{F}}^2 \leq 6\kappa_6[(\gamma_1\sqrt{K_J})\vee(\sqrt{2R^0}\gamma_2)]\sqrt{\frac{1}{N}\left\|\tilde{\Delta}^{(1)}\right\|_{\mathrm{F}}^2+\frac{1}{NT}\left\|\tilde{\Delta}^{(2)}\right\|_{\mathrm{F}}^2},$$

which implies that $\frac{1}{\sqrt{N}}\left\|\tilde{\Delta}^{(1)}\right\|_{\mathrm{F}} \leq \bar{c}(\gamma_1\sqrt{K_J}\vee\gamma_2)$ and $\frac{1}{\sqrt{NT}}\left\|\tilde{\Delta}^{(2)}\right\|_{\mathrm{F}} \leq \bar{c}(\gamma_1\sqrt{K_J}\vee\gamma_2)$ with $\bar{c}=3\kappa_6(1\vee\sqrt{2R^0})<\infty$. This completes the proof. ■

Lemma C.3. *Suppose that Assumptions A.1 and A.3 holds. Let $S_F \equiv F^{0\prime}F^0/T$. Then for any $x>0$,*

$$P(T^{1/2}||S_F-\Sigma_F||_{max}>x) \leq C_1 x^{-q/2}T^{1-q/4}+C_2\exp\left(-C_3x^2\right)$$

for some absolute constants C_ℓ, $\ell=1,2,3$.

Proof of Theorem 4.2. We operate conditional on the event that $\mathcal{E}_{NT}^{(2)}=\{\|\mathbf{U}'\mathbf{X}\|_{\max}/T \leq \gamma_1/2,\ \|\mathbf{U}\|_{\mathrm{op}}/\sqrt{NT}\leq\gamma_2/2$ and $||S_F-\Sigma_F||_{\mathrm{op}}\leq c\}$. One can verify that

$$P(\mathcal{E}_{NT}^{(2)}) \geq 1-\bar{c}'(N^2T^{1-q/4}(\log N)^{-q/2}+N^{2-\underline{c}\log N}),$$

by Lemmas A.2-A.3. With Theorem 4.1, we have with probability at least $1-\bar{c}'(N^2T^{1-q/4}(\log N)^{-q/2}+N^{2-\underline{c}\log N})$,

$$(NT)^{-1/2}\left\|\tilde{\Theta}-\Theta^0\right\|_{\mathrm{op}} \leq (NT)^{-1/2}\left\|\tilde{\Theta}-\Theta^0\right\|_{\mathrm{F}} \leq \bar{c}(\gamma_1\sqrt{K_J}\vee\gamma_2).$$

Next, we show that $\mathcal{E}_{NT}^{(2)}$ implies the desired results.

Step 1: Bound the eigenvalues.

Let $S_\Lambda=\Lambda^{0\prime}\Lambda^0/N$ and $S_F=F^{0\prime}F^0/T$. Let $\sigma_1^2\geq\cdots\geq\sigma_{R^0}^2$ be the R^0 nonzero eigenvalues of $\frac{1}{NT}\Theta^0\Theta^{0\prime}=\frac{1}{T}F^{0\prime}S_\Lambda F^0$. Note that $\sigma_1^2,...,\sigma_{R^0}^2$ are the same as the eigenvalues of $S_F^{1/2}S_\Lambda S_F^{1/2}$. Conditional on the event $\mathcal{E}_{NT}^{(2)}$ and by Assumption A.4 (i)-(iii), we have

$$|\sigma_j^2-s_j|\leq\bar{c}(\sqrt{\log N}T^{-1/2}+N^{-1/2}) \text{ for some } \bar{c}<\infty \text{ and } j=1,...,R^0.$$

This also implies that $||\Theta^0||_{\mathrm{op}}=\sqrt{(s_1+o_P(1))NT}$. Let $\tilde{\sigma}_1^2\geq\cdots\geq\tilde{\sigma}_{N\wedge T}^2$ be the eigenvalues of $\frac{1}{NT}\tilde{\Theta}\tilde{\Theta}'$. Again by the Weyl's theorem, we have

$$\begin{aligned}\left|\tilde{\sigma}_j^2-s_j\right| &\leq \left|\tilde{\sigma}_j^2-\sigma_j^2\right|+\left|\sigma_j^2-s_j\right|\\ &\leq \frac{1}{NT}\left\|\tilde{\Theta}\tilde{\Theta}'-\Theta^0\Theta^{0\prime}\right\|_{\mathrm{op}}+\left|\sigma_j^2-s_j\right|\\ &\leq \frac{2}{NT}\left\|\Theta^0\right\|_{\mathrm{op}}\left\|\tilde{\Theta}-\Theta^0\right\|_{\mathrm{op}}+\frac{1}{NT}\left\|\tilde{\Theta}-\Theta^0\right\|_{\mathrm{op}}^2+\left|\sigma_j^2-s_j\right|,\end{aligned}$$

implying $\left|\tilde{\sigma}_j^2-s_j\right|\leq\bar{c}(\gamma_1\sqrt{K_J}\vee\gamma_2)$. Then for $r\leq R^0$, w.p.a.1,

$$\begin{aligned}\left|\sigma_{j-1}^2-\tilde{\sigma}_j^2\right| &\geq \left|\sigma_{j-1}^2-\sigma_j^2\right|-\left|\tilde{\sigma}_j^2-\sigma_j^2\right|\geq(s_{j-1}-s_j)/2\\ \left|\tilde{\sigma}_j^2-\sigma_{j+1}^2\right| &\geq \left|\sigma_j^2-\sigma_{j+1}^2\right|-\left|\tilde{\sigma}_j^2-\sigma_j^2\right|\geq(s_j-s_{j+1})/2 \qquad \text{(C.8)}\end{aligned}$$

with $\sigma_{j+1}^2 = s_{j+1} = 0$.

Step 2: Prove the consistency of $\hat{R}$.

Note that $\psi_r(\tilde{\Theta}) = \tilde{\sigma}_r\sqrt{NT}$. By step 1, we have that $\psi_r(\tilde{\Theta}) \geq \sqrt{[s_{R^0} + o(1)]NT}$, for all $r \leq R^0$, and $\psi_{R^0+1}(\tilde{\Theta}) = (s_1 + o(1))\sqrt{NT}(\gamma_1\sqrt{K_J} + \gamma_2)$. Noting that $\sqrt{NT}\left\|\tilde{\Theta}\right\|_{\mathrm{op}} = O(NT)$, implying that

$$\min_{i\leq R^0}\psi_i(\tilde{\Theta}) \geq \gamma_2\sqrt{NT}\left\|\tilde{\Theta}\right\|_{\mathrm{op}})^{1/2} \text{ and } \psi_{R^0+1}(\tilde{\Theta}) < (\gamma_2\sqrt{NT}\left\|\tilde{\Theta}\right\|_{\mathrm{op}})^{1/2},$$

when N and T are larger than some $\bar{N}$ and $\bar{T}$, respectively. In this case, we have $P(\hat{R} = R^0) \to 1$ as $(N,T) \to \infty$.

Step 3: Characterize the eigenvectors.

Next we show that there is an $R^0 \times R^0$ matrix $\tilde{H}$, so that the columns of $\frac{1}{\sqrt{T}}F^0\tilde{H}$ are the first R^0 eigenvectors of $\Theta^0\Theta^{0\prime}$. Let v be the $R^0 \times R^0$ matrix whose columns are the eigenvectors of $S_F^{1/2}S_\Lambda S_F^{1/2}$. Then $D = v'S_F^{1/2}S_\Lambda S_F^{1/2}v$ is a diagonal matrix of the eigenvalues of $S_F^{1/2}S_\Lambda S_F^{1/2}$ that are distinct according to Assumption A.3 and Weyl's theorem. Let $\tilde{H} = S_F^{-1/2}v$, then

$$\begin{aligned}
\frac{1}{NT}\Theta^0\Theta^{0\prime}F^0\tilde{H} &= \frac{1}{T}F^0S_\Lambda F^{0\prime}F^0\tilde{H} = F^0S_\Lambda S_F\tilde{H} = F^0S_\Lambda S_F^{1/2}v \\
&= F^0S_F^{1/2}S_F^{-1/2}S_\Lambda S_F^{1/2}v = F^0S_F^{1/2}vv'S_F^{-1/2}S_\Lambda S_F^{1/2}v \\
&= F^0\tilde{H}D.
\end{aligned}$$

In addition, we have $(F^0\tilde{H})'F^0\tilde{H}/T = v'S_F^{-1/2}\frac{F^{0\prime}F^0}{T}S_F^{-1/2}v = v'v = I_{R^0}$. So the columns of $\frac{1}{\sqrt{T}}F^0\tilde{H}$ are the eigenvectors of $\Theta^0\Theta^{0\prime}$, corresponding to the eigenvalues in D.

Step 4: Prove the convergence.

We bound $\left\|\tilde{F} - F^0\tilde{H}\right\|_{\mathrm{F}}$ conditional on the event $\hat{R} = R^0$. By the Davis-Kahan $\sin(\Theta)$ theorem (see, e.g., Yu et al. 2015) and (C.8),

$$\begin{aligned}
\frac{1}{\sqrt{T}}\left\|\tilde{F} - F^0\tilde{H}\right\|_{\mathrm{F}} &\leq \frac{\frac{1}{NT}\left\|\tilde{\Theta}\tilde{\Theta}' - \Theta^0\Theta^{0\prime}\right\|_{\mathrm{op}}}{\min_{j\leq R^0}\min\{\left|\sigma_{j-1}^2 - \tilde{\sigma}_j^2\right|, \left|\tilde{\sigma}_j^2 - \sigma_{j+1}^2\right|\}} \\
&\leq \bar{c}\frac{1}{NT}\left\|\tilde{\Theta}\tilde{\Theta}' - \Theta^0\Theta^{0\prime}\right\|_{\mathrm{op}} \leq \bar{c}(\gamma_1\sqrt{K_J} \vee \gamma_2).
\end{aligned}$$

Next we have

$$\begin{aligned}
\left\|\mathbb{P}_{\tilde{F}} - \mathbb{P}_{F^0}\right\|_{\mathrm{F}} &= \left\|\frac{\tilde{F}\tilde{F}'}{T} - \mathbb{P}_{F^0}\right\|_{\mathrm{F}} \leq 2\bar{c}\left\|\frac{1}{\sqrt{T}}\tilde{F} - \frac{1}{\sqrt{T}}F^0\tilde{H}\right\|_{\mathrm{F}} + \left\|\frac{F^0\tilde{H}\tilde{H}'F^{0\prime}}{T} - \mathbb{P}_{F^0}\right\|_{\mathrm{F}} \\
&\leq \bar{c}(\gamma_1\sqrt{K_J} \vee \gamma_2),
\end{aligned}$$

where the second equality is by the fact $\tilde{H}\tilde{H}' = S_F^{-1/2}vv'S_F^{-1/2} = S_F^{-1}$. This proves the result in (ii). ■

Theoretical analysis of the second-step estimators

Recall that v is the $R^0 \times R^0$ matrix whose columns are the eigenvectors of $S_F^{1/2}S_\Lambda S_F^{1/2}$. Define v^0 to as the $R^0 \times R^0$ matrix whose columns are the eigenvectors of $\Sigma_F^{1/2}\Sigma_\Lambda\Sigma_F^{1/2}$. Let

$H^0 \equiv \Sigma_F^{-1/2} v^0$. One can easily verify that $||H^0||_{\max} \leq \bar{c}$, for some absolute constant $\bar{c} < \infty$.

To prove Theorem **4.3**, we impose the next Lemma.

Lemma C.4. *Suppose that Assumptions A.1-A.3 hold. Let* $\tilde{\Sigma} \equiv T^{-1}\mathbf{X}'\mathbf{X} - T^{-2}\mathbf{X}'\tilde{F}\tilde{F}'\mathbf{X}$. *Then there exist some constants* $\underline{c}$, $\bar{c}$ *and* $\bar{c}'$ *such that with probability larger than* $1 - \bar{c}'(N^2T^{1-q/4}(logN)^{-q/2} + N^{2-\underline{c}logN})$ *we have*

(i) $||\tilde{H}||_{max} \leq ||\tilde{H}||_\infty \leq \bar{c}$,

(ii) $max_{1\leq j\leq pN}|\mathbf{X}_{*,j}|/\sqrt{T} < \bar{c}$, *and* $max_{1\leq j\leq N}|\mathbf{U}_{*,j}|/\sqrt{T} < \bar{c}$;

(iii) $||F^{0\prime}\mathbf{U}||_{max}/T \leq T^{-1/2}logN/(8\bar{c}^2)$ *and* $\left\|T^{-1}\mathbf{X}'F^0 - \Sigma_{XF}\right\|_{max} \leq \bar{c}T^{-1/2}logN$;

(iv) $||\tilde{\Sigma} - \Sigma||_{max} \leq \gamma_3$;

(v) suppose $16K_J\gamma_3 \leq \psi_{\min}(\Sigma)/2$, $\tilde{\Sigma}$ *satisfies the restricted eigenvalue condition for* K_J, *and* $\kappa_{\tilde{\Sigma}}(K_J) \geq \psi_{\min}(\Sigma)/2$.

Proof of Theorem 4.3. In this proof, we choose a large enough $\gamma_3 > (\gamma_1\sqrt{K_J} \vee \gamma_2)$ and fix $\bar{c}$ as in Lemma C.4. Let the $\mathcal{E}_{NT}^{(3)}$ be the joint event of

$$\begin{array}{ll} (1)\ T^{-1}\left\|\mathbf{U}'\mathbf{X}\right\|_{\max} \leq \gamma_3/4; & (2)\ \max_{1\leq j\leq pN}|\mathbf{X}_{*,j}|/\sqrt{T} \leq \bar{c}; \\ (3)\ \max_{1\leq j\leq N}|\mathbf{U}_{*,j}|/\sqrt{T} \leq \bar{c}; & (4)\ ||\tilde{F} - F^0\tilde{H}||_{\mathrm{F}}/\sqrt{T} \leq \gamma_3/(16\bar{c}^2); \\ (5)\ ||F^{0\prime}\mathbf{U}||_{\max}/T \leq \gamma_3/(16\bar{c}^2); & (6)\ ||\tilde{H}||_\infty \leq \bar{c}; \\ (7)\ \hat{R} = R^0; & \end{array}$$

and (8) $\tilde{\Sigma}$ satisfies the restricted eigenvalue condition for K_J with $\kappa_{\tilde{\Sigma}}(K_J) \geq \psi_{\min}(\Sigma)/2$. Under the Assumptions A.1-A.3, by Lemmas C.2 and Lemma C.4, $\mathcal{E}_{NT}^{(3)}$ holds with probability larger than $1-$ $\bar{c}'(N^2T^{1-q/4}(\mathrm{log}N)^{-q/2}+$ $N^{2-\underline{c}\mathrm{log}N})$. Conditional on the event $\mathcal{E}_{NT}^{(3)}$, we also have that

$$\begin{aligned} (9)\ T^{-1}||\tilde{F}'\mathbf{U}||_{\max} &\leq T^{-1}||(\tilde{F} - F^0\tilde{H})'\mathbf{U}||_{\max} + T^{-1}||\tilde{H}'F^{0\prime}\mathbf{U}||_{\max} \\ &\leq T^{-1}||\tilde{F} - F^0\tilde{H}||_{\mathrm{F}} \cdot \max_{1\leq j,N}||\mathbf{U}_{*,j}|| + ||\tilde{H}'||_\infty T^{-1}||F^{0\prime}\mathbf{U}||_{\max} \\ &\leq \gamma_3/(8\bar{c}), \end{aligned}$$

and

$$\begin{aligned} (10)\ \max_{1\leq i\leq N} T^{-1/2}|\lambda_i^{0\prime}F^{0\prime}\mathbb{M}_{\tilde{F}}| &\leq \max_{1\leq i\leq N} T^{-1/2}|\lambda_i^0| \cdot ||F^{0\prime}\mathbb{M}_{\tilde{F}}||_{\mathrm{F}} \\ &\leq \bar{c}T^{-1/2}||\tilde{F} - F^0\tilde{H}||_{\mathrm{F}} \leq \gamma_3/(8\bar{c}). \end{aligned}$$

Conditional on the event $\mathcal{E}_{NT}^{(3)}$, we establish the bound of $|\dot{\Delta}_{*,i}|_1 \equiv |\dot{B}_{*,i} - B_{*,i}^0|_1$, for $j = 1, ..., N$.

Step 1. Concentrating out λ.

The objective function (4.7) is a least squares objective function with respect to λ. Given $\dot{B}_{*,i}$, we have that

$$\dot{\lambda}_j = (\tilde{F}'\tilde{F})^{-1}\tilde{F}'(\mathbf{Y}_{*,i} - \mathbf{X}\dot{B}_{*,i}) = T^{-1}\tilde{F}'(\mathbf{Y}_{*,i} - \mathbf{X}\dot{B}_{*,i}),$$

where the second equality is by the identity $\tilde{F}'\tilde{F}/T = I_T$. After concentrating out λ, the optimization problem becomes

$$\dot{B}_{*,i} = \text{argmin}_{(v',\lambda')'\in\mathbb{R}^{Np+R^0}} \frac{1}{2T}||\mathbb{M}_{\tilde{F}}(\mathbf{Y}_{*,i} - \mathbf{X}v)||_F^2 + \gamma_3|v|_1, \qquad \text{(C.9)}$$

where $\mathbb{M}_{\tilde{F}} = I_T - \tilde{F}\tilde{F}/T$.

Step 2. Compare objective functions at $\dot{B}_{*,i}$ and $B^0_{*,i}$.

By the identity $\mathbf{Y}_{*,i} = \mathbf{X}B^0_{*,i} + F^0\lambda_i^0 + \mathbf{U}_{*,i}$ and the definition of $\dot{B}_{*,i}$, we have

$$\begin{aligned} 0 &\geq \frac{1}{2T}[||\mathbb{M}_{\tilde{F}}(\mathbf{Y}_{*,i} - \mathbf{X}\dot{B}_{*,i})||_{\text{F}}^2 - ||\mathbb{M}_{\tilde{F}}(F^0\lambda_i^0 + \mathbf{U}_{*,i})||_{\text{F}}^2] + \gamma_3(|\dot{B}_{*,i}|_1 - |B^0_{*,i}|_1) \\ &= \frac{1}{2T}||\mathbb{M}_{\tilde{F}}\mathbf{X}\dot{\Delta}_{*,i}||_{\text{F}}^2 - \frac{1}{T}\text{tr}[(F^0\lambda_i^0 + \mathbf{U}_{*,i})'\mathbb{M}_{\tilde{F}}\mathbf{X}\dot{\Delta}_{*,i}] + \gamma_3(|\dot{B}_{*,i}|_1 - |B^0_{*,i}|_1), \end{aligned}$$

where $\dot{\Delta} \equiv \dot{B} - B^0$ and $\dot{\Delta}_{*,i}$ denotes the ith column of $\dot{\Delta}$. By Lemma C.1 (v), the above inequality becomes

$$\begin{aligned} \frac{1}{T}||(F^0\lambda_i^0 + \mathbf{U}_{*,i})'\mathbb{M}_{\tilde{F}}\mathbf{X}||_{\max}|\dot{\Delta}_{*,i}|_1 &\geq \frac{1}{2T}||\mathbb{M}_{\tilde{F}}\mathbf{X}\dot{\Delta}_{*,i}||_F^2 + \gamma_3(|\dot{B}_{*,i}|_1 - |B^0_{*,i}|_1) \\ &\geq \frac{1}{2T}||\mathbb{M}_{\tilde{F}}\mathbf{X}\dot{\Delta}_{*,i}||_F^2 + \gamma_3|\dot{\Delta}_{J_i^c,i}|_1 - \gamma_3|\dot{\Delta}_{J_i,i}|_1. \end{aligned}$$

Step 3. Bound $T^{-1}\max_i[||(F^0\lambda_i^0 + U_{*,i})'M_{\tilde{F}}X||_{\max}]$, conditional on the event $\mathcal{E}_{NT}^{(3)}$.

By triangle inequality and Cauchy Schwarz inequality, we have

$$\begin{aligned} ||(F^0\lambda_i^0 + \mathbf{U}_{*,i})'\mathbb{M}_{\tilde{F}}\mathbf{X}||_{\max} &\leq ||\lambda_i^{0\prime}F^{0\prime}\mathbb{M}_{\tilde{F}}\mathbf{X}||_{\max} + ||\mathbf{U}'_{*,i}\mathbb{M}_{\tilde{F}}\mathbf{X}||_{\max} \\ &\leq \max_{1\leq j\leq Np}|\mathbf{X}_{*,j}|\cdot|\lambda_i^{0\prime}F^{0\prime}\mathbb{M}_{\tilde{F}}| + \max_{1\leq j\leq Np}|\mathbf{U}'_{*,i}\mathbf{X}_{*,j}| + T^{-1}||\mathbf{U}'_{*,i}\tilde{F}\tilde{F}'\mathbf{X}||_{\max} \\ &\leq \max_{1\leq j\leq Np}|\mathbf{U}'_{*,i}\mathbf{X}_{*,j}| + (|\mathbf{U}'_{*,i}\tilde{F}/\sqrt{T}| + |\lambda_i^{0\prime}F^{0\prime}\mathbb{M}_{\tilde{F}}|)\max_{1\leq j\leq Np}|\mathbf{X}_{*,i}|. \end{aligned}$$

Combining event (1) (9) and (10) of $\mathcal{E}_{NT}^{(3)}$, the right hand side of the above inequality is bounded by $\gamma_3/2$.

Step 4. Obtain the final bound for $||\dot{B}_{*,i} - B^0_{*,i}||_1$.

Combining the results of Steps 2-3, we obtain that

$$3\gamma_3|\dot{\Delta}_{J_i,i}|_1 \geq \frac{1}{T}||\mathbb{M}_{\tilde{F}}\mathbf{X}\dot{\Delta}_{*,i}||_{\text{F}}^2 + \gamma_3|\dot{\Delta}_{J_i^c,i}|_1.$$

It follows that $|\dot{\Delta}_{J_i^c,i}|_1 \leq 3|\dot{\Delta}_{J_i,i}|_1$ and

$$\begin{aligned} \dot{\Delta}_{*,i}\tilde{\Sigma}\dot{\Delta}_{*,i} &\leq 3\gamma_3|\dot{\Delta}_{J_i,i}|_1 \leq 3\sqrt{K_J}\gamma_3|\dot{\Delta}_{J_i,i}| \\ &\leq \frac{6\sqrt{K_J}}{\psi_{\min}(\Sigma)}\gamma_3\sqrt{\dot{\Delta}_{*,i}\tilde{\Sigma}\dot{\Delta}_{*,i}}, \end{aligned}$$

where the last inequality is by $\mathcal{E}_{NT}^{(3)}(8)$. It follows that $\sqrt{\dot{\Delta}'_{*,i}\tilde{\Sigma}\dot{\Delta}_{*,i},} \leq \frac{6\sqrt{K_J}}{\psi_{\min}(\Sigma)}\gamma_3$ and $|\dot{\Delta}_{J_i,i}|_1 \leq \frac{2\sqrt{K_J}}{\psi_{\min}(\Sigma)}\sqrt{\dot{\Delta}_{*,i}\tilde{\Sigma}\dot{\Delta}_{*,i},}$. Hence, we have established

$$|\dot{\Delta}_{*,i}|_1 \leq 4|\dot{\Delta}_{J_i,i}|_1 \leq \frac{48}{(\psi_{\min}(\Sigma))^2}K_J\gamma_3.$$

■

Theoretical analysis of the third-step estimators

Lemma C.5. *Suppose that Assumptions A.1-A.4 hold and* $N^2T^{1-q/4}(logN)^{-q/2}+N^{2-\underline{c}logN} \to 0$. *Then*

(i) For $i=1,...,N$, $\psi_{\min}(\tilde{\Sigma}_{J_i,J_i}) \geq \underline{c}$ *w.p.a.1 for some finite constant* $\underline{c}$*;*

(ii) $||\tilde{\Sigma}_{J_i^c,J_i}||_{max} \leq \bar{c}$ *w.p.a.1 for some finite constant* $\bar{c}$*.*

The same results also apply on $\hat{F}^{(\ell)}$*, once we have established that* $||\hat{F}^{(\ell)}-F^0\tilde{H}||_F/\sqrt{T} = O_P(\gamma_1\sqrt{K_J}+\gamma_2)$.

Proof of Theorem 4.4: For any n-dimensional vector $v=(v_1,...,v_n)'$, denote

$$\text{abs}(v) = (|v_1|,...,|v_n|)',$$

and say that $v < v'$ if and only if $v_i < v_i'$ for all $i=1,...,n$. Let $W^{(i)}$ =diag$(w_{1i},...,w_{Np,i})$, $W^{(1,i)} = W^{(i)}_{J_i,J_i}$ and $W^{(0,i)} = W^{(i)}_{J_i^c,J_i^c}$.

The following proof is by induction. Based on error bounds on $\hat{F}^{(\ell)}$'s, we show that results (i)-(iii) holds for $(\ell+1)$th-step estimators. Then the results follows as we already have $||\hat{F}^{(0)} - F^0\tilde{H}||_{\text{F}}/\sqrt{T} = O_P(\gamma_1\sqrt{K_J}+\gamma_2)$.

For notational simplicity, let $\tilde{\Sigma}$ denote $T^{-1}\mathbf{X}'\mathbb{M}_{\hat{F}^{(\ell)}}\mathbf{X}$ for $\ell = 0,1,2,\ldots$.

(i) For all (k,i)'s such that $B^0_{ki}=0$, $\sup_{(k,i):B^0_{ki}=0}|\dot{B}_{ki}| \leq ||\dot{B}-B^0||_{\max} \leq O_P(K_J\gamma_3) = o_P(\gamma_4)$. It follows that $W^{(0,i)} = I_{|J_i^c|}$ with probability approaching one. For all (k,i)'s such that $B^0_{ki} \neq 0$,

$$\begin{aligned} \min_{k,i:B^0_{ki}\neq 0}|\dot{B}_{ki}| &> \min_{i\in[N]}\min_{k\in J_i}|B^0_{ki}| - ||\dot{B}-B^0||_{\max} \\ &= \min_{i\in[N]}\min_{k\in J_i}|B^0_{ki}| - o_P(\gamma_4) \geq \alpha\gamma_4, \end{aligned}$$

with probability approaching one, by Assumption A.5. It follows that $W^{(1,i)} = \mathbf{0}$ with probability approaching one. For each $i\in[N]$, the estimator $\hat{B}^{(\ell)}_{*,i}$ can be written as

$$\hat{B}^{(\ell)}_{*,i} = \text{argmin}_{v\in\mathbb{R}^{NP}}\mathcal{L}^{(i)}(v,\hat{F}^{(\ell-1)}),$$

where

$$\mathcal{L}^{(i)}(v,F) \equiv \frac{1}{2T}(\mathbf{Y}_{*,i}-\mathbf{X}v)'\mathbb{M}_{\hat{F}^{(\ell-1)}}(\mathbf{Y}_{*,i}-\mathbf{X}v) + \gamma_4\sum_{k=1}^{pN}w_{ki}\,|v_k| \text{ for } i=1,...,N.$$

Following the proof of Proposition 1 of Zhao an Yu (2006), sgn$(\hat{B}^{(l)}_{*,i})$ =sgn$(B^0_{*,i})$ is implied by event $\mathcal{E}_{i,1}\cap\mathcal{E}_{i,2}$, where

$$\mathcal{E}_{i,1} \equiv \left\{\text{abs}[T^{-1/2}\tilde{\Sigma}^{-1}_{J_i,J_i}\mathbf{X}'_{*,J_i}\mathbb{M}_{\hat{F}^{(\ell-1)}}(\mathbf{U}_{*,i}+F^0\lambda^0_i)] < T^{1/2}\text{abs}(B^0_{J_i,i}) - T^{1/2}\gamma_4\text{abs}[\tilde{\Sigma}^{-1}_{J_i,J_i}W^{(1,i)}\text{sgn}(B^0_{J_i,i})]\right\};$$

and

$$\begin{aligned} \mathcal{E}_{i,2} &\equiv \{\text{abs}[T^{-1/2}(-\tilde{\Sigma}_{J_i^c,J_i}\tilde{\Sigma}^{-1}_{J_i,J_i}\cdot\mathbf{X}'_{*,J_i}+\mathbf{X}'_{*,J_i^c})\mathbb{M}_{\hat{F}^{(\ell-1)}}(\mathbf{U}_{*,i}+F^0\lambda^0_i)] \\ &< T^{1/2}\gamma_4W^{(0,i)}\cdot\iota_{|J_i^c|} - T^{1/2}\gamma_4\text{abs}[\tilde{\Sigma}_{J_i^c,J_i}\tilde{\Sigma}^{-1}_{J_i,J_i}W^{(1,i)}\text{sgn}(B^0_{J_i,i})]\}. \end{aligned}$$

We prove (i) by showing that $\mathcal{E}_{i,1}$ and $\mathcal{E}_{i,2}$ hold w.p.a.1.

First, we consider $\mathcal{E}_{i,1}$. It suffices to show that each entry of $T^{-1/2}\text{abs}[\tilde{\Sigma}^{-1}_{J_i,J_i}\mathbf{X}'_{*,J_i}\mathbb{M}_{\hat{F}^{(\ell-1)}}(\mathbf{U}_{*,i}+F^0\lambda^0_i)]$ is $o_P(\sqrt{T}\min_i\min_{k\in J_i}|B^0_{ki}|)$. Applying the triangle inequality, one has

$$\begin{aligned}&T^{-1/2}\text{abs}[\tilde{\Sigma}^{-1}_{J_i,J_i}\mathbf{X}'_{*,J_i}\mathbb{M}_{\hat{F}^{(\ell-1)}}(\mathbf{U}_{*,i}+F^0\lambda^0_i)] \qquad\qquad (C.10)\\ \leq &T^{-1/2}\text{abs}(\tilde{\Sigma}^{-1}_{J_i,J_i}\mathbf{X}'_{*,J_i}\mathbb{M}_{\hat{F}^{(\ell-1)}}\mathbf{U}_{*,i})+T^{-1/2}\text{abs}(\tilde{\Sigma}^{-1}_{J_i,J_i}\mathbf{X}'_{*,J_i}\mathbb{M}_{\hat{F}^{(\ell-1)}}F^0\lambda^0_i)\\ \leq &T^{-1/2}\text{abs}(\tilde{\Sigma}^{-1}_{J_i,J_i}\mathbf{X}'_{*,J_i}\mathbb{M}_{\hat{F}^{(\ell-1)}}\mathbf{U}_{*,i})+T^{-1/2}\text{abs}[\tilde{\Sigma}^{-1}_{J_i,J_i}\mathbf{X}'_{*,J_i}(\mathbb{P}_{F^0}-\mathbb{P}_{\hat{F}^{(\ell-1)}})\mathbf{U}_{*,i}]\\ &+T^{-1/2}\text{abs}[\tilde{\Sigma}^{-1}_{J_i,J_i}\mathbf{X}'_{*,J_i}\mathbb{M}_{\hat{F}^{(\ell-1)}}(\hat{F}^{(\ell-1)}-F^0\tilde{H})\tilde{H}^{-1}\lambda^0_i].\end{aligned}$$

Note that $\max_i||\tilde{\Sigma}^{-1}_{J_i,J_i}||_{\text{op}}\leq\bar{c}$ w.p.a.1 by the Lemma A.5. This, in conjunction with Lemma A.2(i)-(ii), implies that the first term on the RHS of (C.10) is uniformly $O_P(\log N)$. With $||\hat{F}^{(\ell-1)}-F^0\tilde{H}||_{\text{F}}/\sqrt{T}=O_P((\log N)T^{-1/2}\sqrt{K_J}+N^{-1/2})$,[1] we have $||\mathbb{P}_{F^0}-\mathbb{P}_{\hat{F}^{(\ell-1)}}||_{\text{op}}=O_P((\log N)T^{-1/2}\sqrt{K_J}+N^{-1/2})$. Note that Lemma A.4(ii) ensures $\max_{1\leq j\leq pN}||\mathbf{X}_{*,j}||/\sqrt{T}$ and $\max_{1\leq j\leq N}||\mathbf{U}_{*,j}||/\sqrt{T}$ are both bounded by an absolute constant. It follows that each entry of the second term on the RHS of (C.10) is $O_P(\log N\cdot\sqrt{K_J}+\sqrt{T/N})$. Similarly, each entry of the third term on the RHS is $O_P(\log N\cdot\sqrt{K_J}+\sqrt{T/N})$. These results, along with the fact that $\log N\cdot T^{-1/2}\sqrt{K_J}=o(\min_i\min_{k\in J_i}|B^0_{ki}|)$ and $N^{-1/2}=o_P(\min_i\min_{k\in J_i}|B^0_{ki}|)$ imply that $P(\mathcal{E}_{i,1})\to 1$.

Next, we consider $\mathcal{E}_{i,2}$. Similar to the analysis for $\mathcal{E}_{i,1}$, we can use Lemma A.5 (ii) to show that each entry of $T^{-1/2}(-\tilde{\Sigma}_{J^c_i,J_i}\tilde{\Sigma}^{-1}_{J_i,J_i}\cdot\mathbf{X}'_{*,J_i}+\mathbf{X}'_{*,J^c_i})\mathbb{M}_{\hat{F}^{(\ell-1)}}(\mathbf{U}_{*,i}+F^0\lambda^0_i)$ is $O_P(\log N\cdot\sqrt{K_J}+\sqrt{T/N})=o(\sqrt{T}\gamma_3)$. By the fact that $\gamma_3=o(\gamma_4)$, we have $P(\mathcal{E}_{i,2})\to 1$, as $(N,T)\to\infty$.

(ii) Conditional on the event $\{\hat{B}^{(\ell)}=_s B^0\}$, we can follow the proof of Lemma 1 in Zhao and Yu (2006) to establish the first order condition that

$$\tilde{\Sigma}_{J_i,J_i}(\hat{B}^{(\ell)}_{J_i,i}-B^0_{J_i,i})=\frac{1}{T}\mathbf{X}'_{*,J_i}\mathbb{M}_{\hat{F}^{(\ell-1)}}F^0\lambda^0_i+\frac{1}{T}\mathbf{X}'_{*,J_i}\mathbb{M}_{\hat{F}^{(\ell-1)}}\mathbf{U}_{*,i},$$

for $i\in[N]$. Note that $\sum_{i=1}^N|J_i|/N<C$, by Assumption A.5 (ii). It follows that

$$\begin{aligned}\frac{||\mathbf{X}(\hat{B}^{(\ell)}-B^0)||^2_{\text{F}}}{NT} &= \frac{1}{N}\sum_{i=1}^N\frac{||\mathbf{X}(\hat{B}^{(\ell)}_{*,i}-B^0_{*,i})||^2}{T}=\frac{1}{N}\sum_{i=1}^N\frac{1}{T}||\mathbf{X}_{*,J_i}(\hat{B}^{(\ell)}_{J_i,i}-B^0_{J_i,i})||^2\\ &= \frac{1}{N}\sum_{i=1}^N O_P[(\gamma_1\sqrt{K_J}+\gamma_2)^2|J_i|]=O_P[(\gamma_1\sqrt{K_J}+\gamma_2)^2].\end{aligned}$$

(iii) Note that $\mathbf{Y}-\mathbf{X}\hat{B}^{(\ell)}-F^0\Lambda^{0\prime}=\mathbf{U}-\mathbf{X}(\hat{B}^{(\ell)}-B^0)$. By the result in (ii) and Assumption A.3(i), the operator norm of $\mathbf{U}-\mathbf{X}(\hat{B}^{(\ell)}-B^0)$ is of the order $O_P(\gamma_1\sqrt{K_J}+\gamma_2)$. One can apply analysis similar to proof of Theorem 4.2 to obtain the desired result. ■

Proof of Theorem 4.5: Let $\hat{\Sigma}=\mathbf{X}'\mathbb{M}_{\hat{F}}\mathbf{X}/T$. From the proof of Theorem 2.4, we have that

$$\hat{\Sigma}_{J_i,J_i}(\hat{B}_{J_i,i}-B^0_{J_i,i})=\frac{1}{T}\mathbf{X}'_{*,J_i}\mathbb{M}_{\hat{F}}F^0\lambda^0_i+\frac{1}{T}\mathbf{X}'_{*,J_i}\mathbb{M}_{\hat{F}}\mathbf{U}_{*,i}-\gamma_4W^{(1,i)}\text{sgn}(B^0_{J_i,i}). \qquad (C.11)$$

Noting that the columns of $\hat{F}/\sqrt{T}$ are the first $\hat{R}$ eigenvectors of $\frac{1}{NT}\left(\mathbf{Y}-\mathbf{X}\hat{B}\right)\left(\mathbf{Y}-\mathbf{X}\hat{B}\right)'$,

[1] This claim holds for $\ell=1$ by Theorem 3.2. Given this claim, we will show below that $||\hat{F}^{(\ell)}-F^0\tilde{H}||_{\text{F}}/\sqrt{T}=O_P((\log N)T^{-1/2}\sqrt{K_J}+N^{-1/2})$.

we have

$$\hat{F}V_{NT} = \frac{1}{NT}\left(\mathbf{Y} - \mathbf{X}\hat{B}\right)\left(\mathbf{Y} - \mathbf{X}\hat{B}\right)'\hat{F} = \frac{1}{NT}\sum_{i=1}^{N}\left(\mathbf{Y}_{*,i} - \mathbf{X}_{*,J_i}\hat{B}_{J_i,i}\right)\left(\mathbf{Y}_{*,i} - \mathbf{X}_{*,J_i}\hat{B}_{J_i,i}\right)'\hat{F},$$

where V_{NT} is a diagonal matrix that consists of the $\hat{R}$ largest eigenvalues of the matrix $(NT)^{-1} \times \left(\mathbf{Y} - \mathbf{X}\hat{B}\right)\left(\mathbf{Y} - \mathbf{X}\hat{B}\right)'$, arranged in descending order along its diagonal line.

As the term $\gamma_4 W^{(1,i)}$ $\times \mathrm{sgn}(B^0_{J_i,i}) = o_p(T^{-1/2})$, we can follow the analysis of oracle least squares estimator to establish the asymptotic distribution of $\hat{B}_{J_i,i}$. By Proposition B.1 of the online supplement, we have

$$S_i(\hat{B}_{J_i,i} - B^0_{J_i,i}) = S_i[T^{-1}(\mathbf{X}'_{*,J_i}\mathbb{M}_{F^0}\mathbf{X}_{*,J_i})]^{-1}\frac{1}{T}\mathbf{X}'_{*,J_i}\mathbb{M}_{F^0}\mathbf{U}_{*,i} + o_P(T^{-1/2}).$$

It follows that

$$\begin{aligned}\sqrt{T}S_i(\hat{B}_{J_i,i} - B^0_{J_i,i}) &= \frac{1}{\sqrt{T}}S_i(\Sigma_{J_i,J_i})^{-1}(\mathbf{X}_{*,J_i} - F^0\Sigma_F^{-1}(\Sigma_{XF})'_{J_i,*})'\mathbf{U}_{*,i} + o_P(1)\\ &\equiv T^{-1/2}\sum_{t=1}^{T} z^*_{it}u_{it} + o_P(1),\end{aligned}$$

One can easily see that $\{z^*_{it}u_{it}\}$ is a martingale difference sequence. One can verify the conditions of central limit theorems for martingale difference sequence by straightforward calculation and establishes that $\sqrt{T}S_i(\check{B}_{J_i,i} - B^0_{J_i,i}) \xrightarrow{d} N(0, \sigma_i^2 S_i(\Sigma_{J_i,J_i})^{-1}S_i')$. ■

Bibliography

Ahn, S. C. and Horenstein, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica*, 81(3):1203–1227.

Alizadeh, S., Brandt, M. W., and Diebold, F. X. (2002). Range-based estimation of stochastic bolatility models. *The Journal of Finance*, 57(3):1047–1091.

Ando, T. and Bai, J. (2016). Panel data models with grouped factor structure under unknown group membership. *Journal of Applied Econometrics*, 31(1):163–191.

Ando, T. and Bai, J. (2017). Clustering huge number of financial time series: A panel data approach with high-dimensional predictors and factor structures. *Journal of the American Statistical Association*, 112(519):1182–1198.

Andrews, D. W. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica: Journal of the Econometric Society*, pages 821–856.

Arcand, J. L., Berkes, E., and Panizza, U. (2015). Too much finance? *Journal of Economic Growth*, 20(2):105–148.

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

Bai, J. (2009). Panel data mdels with interactive fixed effects. *Econometrica*, 77(4):1229–1279.

Bai, J. and Liao, Y. (2016). Efficient estimation of approximate factor models via penalized maximum likelihood. *Journal of Econometrics*, 191(1):1–18.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221.

Bai, Z. and Saranadasa, H. (1996). Effect of high dimension: an example of a two sample problem. *Statistica Sinica*, 6(2):311–329.

Barigozzi, M. and Brownlees, C. (2019). Nets: Network estimation for time series. *Journal of Applied Econometrics*, 34(3):347–364.

Barigozzi, M. and Hallin, M. (2017). A network analysis of the volatility of high dimensional financial series. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66(3):581–605.

Basu, S. and Michailidis, G. (2015). Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, 43(4):1535–1567.

Beck, T., Levine, R., and Levkov, A. (2010). Big bad banks? The winners and losers from bank deregulation in the United States. *The Journal of Finance*, 65(5):1637–1667.

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429.

Bernanke, B. S., Boivin, J., and Eliasz, P. (2005). Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach. *The Quarterly journal of economics*, 120(1):387–422.

Bernard, A. B., Jensen, J. B., Redding, S. J., and Schott, P. K. (2007). Firms in international trade. *Journal of Economic Perspectives*, 21(3):105–130.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732.

Billio, M., Getmansky, M., Lo, A. W., and Pelizzon, L. (2012). Econometric measures of connectedness and systemic risk in the finance and insurance sectors. *Journal of financial economics*, 104(3):535–559.

Bonhomme, S. and Manresa, E. (2015). Grouped patterns of heterogeneity in panel data. *Econometrica*, 83(3):1147–1184.

Browning, M. and Carro, J. (2007). Heterogeneity and microeconometrics modeling. *Econometric Society Monographs*, 43:47.

Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The annals of Statistics*, 35(6):2313–2351.

Caner, M. and Hansen, B. E. (2004). Instrumental variable estimation of a threshold model. *Econometric Theory*, 20(5):813–843.

Caner, M. and Kock, A. B. (2018). Asymptotically honest confidence regions for high dimensional parameters by the desparsified conservative lasso. *Journal of Econometrics*, 203(1):143–168.

Castagnetti, C., Rossi, E., and Trapani, L. (2015). Inference on factor structures in heterogeneous panels. *Journal of econometrics*, 184(1):145–157.

Cecchetti, S. G., Mohanty, M., and Zampolli, F. (2011). Achieving growth amid fiscal imbalances: the real effects of debt. In *Economic Symposium Conference Proceedings*, volume 352, pages 145–96. Federal Reserve Bank of Kansas City.

Chan, K.-S. (1993). Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model. *The annals of statistics*, 21(1):520–533.

Checherita-Westphal, C. and Rother, P. (2012). The impact of high government debt on economic growth and its channels: An empirical investigation for the euro area. *European economic review*, 56(7):1392–1405.

Chen, M., Fernández-Val, I., and Weidner, M. (2014). Nonlinear factor models for network and panel data. *working paper.*

Chen, S. X. and Qin, Y.-L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics*, 38(2):808–835.

Cheng, X. and Hansen, B. E. (2015). Forecasting with factor-augmented regression: A frequentist model averaging approach. *Journal of Econometrics*, 186(2):280–293.

Chernozhukov, V. and Hansen, C. (2008). Instrumental variable quantile regression: A robust inference approach. *Journal of Econometrics*, 142(1):379–398.

Chernozhukov, V., Hansen, C. B., Liao, Y., and Zhu, Y. (2019). Inference for heterogeneous effects using low-rank estimations. cemmap working paper CWP31/19, London.

Chudik, A., Mohaddes, K., Pesaran, M. H., and Raissi, M. (2017). Is there a debt-threshold effect on output growth? *Review of Economics and Statistics*, 99(1):135–150.

Chudik, A. and Pesaran, M. H. (2011). Infinite-dimensional VARs and factor models. *Journal of Econometrics*, 163(1):4–22.

Chudik, A., Pesaran, M. H., and Tosetti, E. (2011). *Weak and strong cross-section dependence and estimation of large panels.* Oxford University Press Oxford, UK.

Dang, V. A., Kim, M., and Shin, Y. (2012). Asymmetric capital structure adjustments: New evidence from dynamic panel threshold models. *Journal of Empirical Finance*, 19(4):465–482.

Demirer, M., Diebold, F. X., Liu, L., and Yilmaz, K. (2018). Estimating global bank network connectedness. *Journal of Applied Econometrics*, 33(1):1–15.

Diebold, F. X. and Yilmaz, K. (2014). On the network topology of variance decompositions: Measuring the connectedness of financial firms. *Journal of Econometrics*, 182(1):119–134.

Diebold, F. X. and Yilmaz, K. (2015). *Financial and macroeconomic connectedness: a network approach to measurement and monitoring.* Oxford University Press. Google-Books-ID: 5voGBgAAQBAJ.

Durlauf, S. N. (2001). Manifesto for a growth econometrics. *Journal of econometrics*, 100(1):65–69.

Durlauf, S. N. and Johnson, P. A. (1995). Multiple regimes and cross-country growth behaviour. *Journal of Applied Econometrics*, 10(4):365–384.

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56.

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32(3):928–961.

Fazzari, S., Hubbard, R. G., and Petersen, B. C. (1987). Financing constraints and corporate investment. Technical report, National Bureau of Economic Research.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2000). The generalized dynamic-factor model: Identification and estimation. *Review of Economics and statistics*, 82(4):540–554.

Garman, M. B. and Klass, M. J. (1980). On the estimation of security price volatilities from historical data. *Journal of business*, pages 67–78.

Geweke, J. (1977). The dynamic factor analysis of economic time series. *Latent Variables in Socio-Economic Models.*

Giannone, D., Reichlin, L., and Sala, L. (2004). Monetary policy in real time. *NBER macroeconomics annual*, 19:161–200.

Gonzalez, A., Teräsvirta, T., Van Dijk, D., and Yang, Y. (2017). Panel smooth transition regression models. *Working paper, Uppsala University.*

Guo, S., Wang, Y., and Yao, Q. (2016). High-dimensional and banded vector autoregressions. *Biometrika*, 103:889–903.

Hahn, J. and Kuersteiner, G. (2011). Bias reduction for dynamic nonlinear panel models with fixed effects. *Econometric Theory*, 27(6):1152–1191.

Hallin, M. and Liška, R. (2007). Determining the number of factors in the general dynamic factor model. *Journal of the American Statistical Association*, 102(478):603–617.

Han, F., Lu, H., and Liu, H. (2015). A direct estimation of high dimensional stationary vector autoregressions. *The Journal of Machine Learning Research*, 16(1):3115–3150.

Hansen, B. E. (1996). Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica*, 64(2):413–430.

Hansen, B. E. (1999). Threshold effects in non-dynamic panels: Estimation, testing, and inference. *Journal of Econometrics*, 93(2):345–368.

Hansen, B. E. (2000). Sample splitting and threshold estimation. *Econometrica*, 68(3):575–603.

Hansen, B. E. (2011). Threshold autoregression in economics. *Statistics and its Interface*, 4(2):123–127.

Haufe, S., Müller, K.-R., Nolte, G., and Krämer, N. (2010). Sparse causal discovery in multivariate time series. In *Causality: Objectives and Assessment*, pages 97–106.

Hautsch, N., Schaumburg, J., and Schienle, M. (2014). Financial network systemic risk contributions*. *Review of Finance*, 19(2):685–738.

Hsiao, C. (2014). *Analysis of Panel Data.* Cambridge University Press. Google-Books-ID: 7LIkBQAAQBAJ.

Huang, J., Ma, S., and Zhang, C.-H. (2008). Adaptive Lasso for sparse high-dimensional regression models. *Statistica Sinica*, pages 1603–1618.

Huang, W., Jin, S., and Su, L. (2020). Identifying latent grouped patterns in cointegrated panels. *Econometric Theory*, 36(3):410–456.

Hurn, S., Martin, V., Phillips, P. C. B., and Yu, J. (2019). *Financial econometric modeling.* Oxford University Press (Forthcoming).

Kasahara, H. and Shimotsu, K. (2009). Nonparametric identification of finite mixture models of dynamic discrete choices. *Econometrica*, 77(1):135–175.

Kim, J. and Pollard, D. (1990). Cube root asymptotics. *Annals of Statistics*, 18(1):191–219.

Kock, A. B. and Callot, L. (2015). Oracle inequalities for high dimensional vector autoregressions. *Journal of Econometrics*, 186(2):325–344.

Kourtellos, A., Stengos, T., and Tan, C. M. (2016). Structural threshold regression. *Econometric Theory*, 32(4):827–860.

Kremer, S., Bick, A., and Nautz, D. (2013). Inflation and growth: new evidence from a dynamic panel threshold analysis. *Empirical Economics*, 44(2):861–878.

Kroszner, R. S. and Strahan, P. E. (1999). What drives deregulation? economics and politics of the relaxation of bank branching restrictions*. *The Quarterly Journal of Economics*, 114(4):1437–1467.

Lam, C. and Fan, J. (2008). Profile-kernel likelihood inference with diverging number of parameters. *Annals of statistics*, 36(5):2232.

Law, S. H. and Singh, N. (2014). Does too much finance harm economic growth? *Journal of Banking & Finance*, 41:36–44.

Lee, N., Moon, H. R., and Weidner, M. (2012). Analysis of interactive fixed effects dynamic linear panel regression with measurement error. *Economics Letters*, 117(1):239–242.

Leeper, E. M., Sims, C. A., Zha, T., Hall, R. E., and Bernanke, B. S. (1996). What does monetary policy do? *Brookings Papers on Economic Activity*, 1996(2):1–78.

Levine, R. (2005). Chapter 12 Finance and Growth: Theory and Evidence. volume 1, pages 865–934. Elsevier.

Li, D. and Ling, S. (2012). On the least squares estimation of multiple-regime threshold autoregressive models. *Journal of Econometrics*, 167(1):240–253.

Li, D., Qian, J., and Su, L. (2016). Panel data models with interactive fixed effects and multiple structural breaks. *Journal of the American Statistical Association*, 111(516):1804–1819.

Lin, C.-C. and Ng, S. (2012). Estimation of panel data models with parameter heterogeneity when group membership is unknown. *Journal of Econometric Methods*, 1(1).

Liu, R., Shang, Z., Zhang, Y., and Zhou, Q. (2020). Identification and estimation in panel models with overspecified number of groups. *Journal of Econometrics*, 215(2):574–590.

Lu, X. and Su, L. (2016). Shrinkage estimation of dynamic panel data models with interactive fixed effects. *Journal of Econometrics*, 190(1):148–175.

Lu, X. and Su, L. (2017). Determining the number of groups in latent panel structures with an application to income and democracy. *Quantitative Economics*, 8(3):729–760.

Ludvigson, S. C. and Ng, S. (2007). The empirical risk–return relation: A factor analysis approach. *Journal of Financial Economics*, 83(1):171–222.

Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer Science & Business Media.

Ma, S., Lan, W., Su, L., and Tsai, C.-L. (2020). Testing Alphas in Conditional Time-Varying Factor Models With High-Dimensional Assets. *Journal of Business & Economic Statistics*, 38(1):214–227.

Mann, H. B. and Wald, A. (1943). On the statistical treatment of linear stochastic difference equations. *Econometrica, Journal of the Econometric Society*, pages 173–220.

Medeiros, M. C., Vasconcelos, G. F. R., Veiga, A., and Zilberman, E. (2019). Forecasting inflation in a data-rich environment: the benefits of machine learning methods. *Journal of Business & Economic Statistics*, pages 1–22.

Miao, K., Li, K., and Su, L. (2020a). Panel threshold with interactive fixed effects. *working paper*.

Miao, K., Su, L., and Wang, W. (2020b). Panel threshold regressions with latent group structures. *Journal of Econometrics*, 214(2):451–481.

Moon, H. R., Shum, M., and Weidner, M. (2018). Estimation of random coefficients logit demand models with interactive fixed effects. *Journal of Econometrics*, 206(2):613–644.

Moon, H. R. and Weidner, M. (2015). Linear regression for panel with unknown number of factors as interactive fixed effects. *Econometrica*, 83(4):1543–1579.

Moon, H. R. and Weidner, M. (2017). Dynamic linear panel regression models with interactive fixed effects. *Econometric Theory*, 33(1):158–195.

Moon, H. R. and Weidner, M. (2019). Nuclear norm regularized estimation of panel regression models. *arXiv:1810.10987*. arXiv: 1810.10987.

Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, 16(1):1–32.

Okui, R. and Wang, W. (2020). Heterogeneous structural breaks in panel data models. *Journal of Econometrics*.

Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics*, 92(4):1004–1016.

Pesaran, M. H. (2015). *Time Series and Panel Data Econometrics*. Oxford University Press.

Phillips, P. C. B. and Moon, H. R. (1999). Linear regression limit theory for nonstationary panel data. *Econometrica*, 67(5):1057–1111.

Potter, S. M. (1995). A nonlinear approach to US GNP. *Journal of Applied Econometrics*, 10(2):109–125.

Qian, J. and Su, L. (2016). Shrinkage estimation of common breaks in panel data models via adaptive group fused Lasso. *Journal of Econometrics*, 191(1):86–109.

Ramírez-Rondán, N. (2015). Maximum Likelihood Estimation of Dynamic Panel Threshold Models. Technical report, Peruvian Economic Association.

Reinhart, C. M. and Rogoff, K. S. (2010). Growth in a time of debt. *American Economic Review*, 100(2):573–578.

Sarafidis, V. and Weber, N. (2015). A partially heterogeneous framework for analyzing panel data. *Oxford Bulletin of Economics and Statistics*, 77(2):274–296.

Seo, M. H. and Linton, O. (2007). A smoothed least squares estimator for threshold regression models. *Journal of Econometrics*, 141(2):704–735.

Seo, M. H. and Shin, Y. (2016). Dynamic panels with threshold effect and endogeneity. *Journal of Econometrics*, 195(2):169–186.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica*, 48(1):1–48.

Sims, C. A. (1992). Interpreting the macroeconomic time series facts: The effects of monetary policy. *European economic review*, 36(5):975–1000.

Sims, C. A. (1993). A nine-variable probabilistic macroeconomic forecasting model. In *Business cycles, indicators, and forecasting*, pages 179–212. University of Chicago press.

Smeekes, S. and Wijler, E. (2018). Macroeconomic forecasting using penalized regression methods. *International journal of forecasting*, 34(3):408–430.

Spearot, A. C. (2012). Firm heterogeneity, new investment and acquisitions. *The Journal of Industrial Economics*, 60(1):1–45.

Stock, J. H. and Watson, M. W. (1999). Forecasting inflation. *Journal of Monetary Economics*, 44(2):293–335.

Stock, J. H. and Watson, M. W. (2002). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, 20(2):147–162.

Stock, J. H. and Watson, M. W. (2005). Implications of dynamic factor models for VAR analysis. Technical report, National Bureau of Economic Research.

Su, L. and Chen, Q. (2013). Testing homogeneity in panel data models with interactive fixed effects. *Econometric Theory*, 29(6):1079–1135.

Su, L. and Hoshino, T. (2016). Sieve instrumental variable quantile regression estimation of functional coefficient models. *Journal of Econometrics*, 191(1):231–254.

Su, L. and Ju, G. (2018). Identifying latent grouped patterns in panel data models with interactive fixed effects. *Journal of Econometrics*, 206(2):554–573.

Su, L., Shi, Z., and Phillips, P. C. B. (2016). Identifying latent structures in panel data. *Econometrica*, 84(6):2215–2264.

Su, L., Wang, X., and Jin, S. (2019). Sieve estimation of time-varying panel data models with latent structures. *Journal of Business & Economic Statistics*, 37(2):334–349.

Sun, Y. (2005). Estimation and inference in panel structure models. *Available at SSRN 794884*.

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tong, H. (1978). On a threshold model. In *Pattern Recognition and Signal Processing, ed. C. H. Chen*, 29, pages 575–586.

Uematsu, Y. and Tanaka, S. (2019). High-dimensional macroeconomic forecasting and variable selection via penalized regression. *The Econometrics Journal*, 22(1):34–56.

Vaart, A. v. d., Vaart, A. v. d., Vaart, A. W. v. d., and Wellner, J. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics.* Springer Science & Business Media. Google-Books-ID: OCenCW9qmp4C.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science, ser. Cambridge Series in Statistical and Probabilistic Mathematics.* Cambridge University Press.

Vogt, M. and Linton, O. (2017). Classification of non-parametric regression functions in longitudinal data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(1):5–27.

Vogt, M. and Linton, O. (2020). Multiscale clustering of nonparametric regression curves. *Journal of Econometrics*, 216(1):305–325.

Wang, W., Phillips, P. C. B., and Su, L. (2018). Homogeneity pursuit in panel data models: Theory and application. *Journal of Applied Econometrics*, 33(6):797–815.

Wang, W. and Su, L. (2020). Identifying latent group structures in nonlinear panels. *Journal of Econometrics, forthcoming.*

Wu, W. B. (2005). Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences*, 102(40):14150–14154.

Wu, W.-B. and Wu, Y. N. (2016). Performance bounds for parameter estimates of high-dimensional linear models with correlated errors. *Electronic Journal of Statistics*, 10(1):352–379.

Yu, P. and Phillips, P. C. B. (2018). Threshold regression with endogeneity. *Journal of Econometrics*, 203(1):50–68.

Yu, Y., Wang, T., and Samworth, R. J. (2015). A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323.

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine learning research*, 7(11):2541–2563.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.