8-2018

Nonstationary panels with unobserved heterogeneity

Wenxin HUANG

Nonstationary Panels with Unobserved Heterogeneity

Wenxin Huang

A DISSERTATION

IN

ECONOMICS

Presented to the Singapore Management University in Partial Fulfillment

of the Requirements for the Degree of PhD in Economics

2018

____

Supervisor of Dissertation

____

PhD in Economics, Programme Director

Nonstationary Panels with Unobserved Heterogeneity

by

Wenxin Huang

Submitted to School of Economics in Partial Fulfillment of
the Requirements for the Degree of Doctor of Philosophy in Economics

Dissertation Committee:

Liangjun Su (Supervisor/Co-chair)
Lee Kong Chian Professor of Economics
Singapore Management University

Peter C.B. Phillips (Supervisor/Co-chair)
Sterling Professor of Economics & Professor of Statistics
Yale University
Distinguished Term Professor of Economics
Singapore Management University

Sainan Jin
Professor of Economics
Singapore Management University

Qu Feng
Associate Professor of Economics
Nanyang Technological University

Singapore Management University
2018

Abstract

This dissertation develops several econometric techniques to address the unobserved heterogeneity in nonstationary panels, namely identifying latent group structures in cointegrated panels, studying nonstationary panels with both cross-sectional dependence and latent group structures, and estimating panel error-correction model with unobserved dynamic common factors.

Chapter 1 considers a panel cointegration model with latent group structures that allows for heterogeneous long-run relations across groups. We extend Su et al. (2013) classifier-Lasso (C-Lasso) method to the nonstationary panels and allow for the presence of endogeneity in both the stationary and nonstationary regressors in the model. In addition, we allow the dimension of the stationary regressors to diverge with the sample size. We show that we can identify the individuals' group membership and estimate the group-specific long-run cointegrated relationships simultaneously. We demonstrate the desirable property of uniform classification consistency and the oracle properties of both the C-Lasso estimators and their post-Lasso versions. The special case of dynamic penalized least squares is also studied. Simulations show superb finite sample performance in both classification and estimation. In an empirical application, we study the potential heterogeneous behavior in testing the validity of long-run PPP hypothesis in the post-Bretton Woods period from 1975-2014 covering 99 countries. We identify two groups in the period 1975-1998 and three ones in the period 1999-2014. The results confirm that at least some countries favor the long-run PPP hypothesis in the post-Bretton Woods period.

Chapter 2 proposes a novel approach, based on Lasso, to handle unobserved parameter heterogeneity and cross-sectional dependence in nonstationary panel models. We propose a penalized principal component method to jointly estimate group-specific long-run relationships, unobserved common factors and to identify unknown group membership. Our Lasso-type estimators are consistent and efficient. We provide a bias-correction procedure under which our estimators are centered around zero as both dimensions of the panel tend to infinity. We establish a mixed normal asymptotic distribution for our estimators, which permit inference using standard test statistics. Finally, we apply our approach to study the international R&D spillovers model with unobserved group patterns. The results shed

new light on growth convergence puzzle though the channel of technology diffusions.

Chapter 3 proposes a novel econometric model that accounts for both long-run and short-run co-movements in panel error correction models. By imposing latent group structures, we achieve efficient estimation for long-run cointegration vectors in the presence of unobserved heterogeneity. The short-run co-movements are driven by unobserved dynamic common factors, which can be consistently estimated by principal components. We propose a penalized generalized least squares method that jointly estimates long-run cointegration vectors and infers unobserved group structures. We establish asymptotic properties for two Lasso-type estimators. In an empirical application, we estimate long-run cointegration relationships between bid and ask quotes in stock market. We introduce a new measure for efficient price, which is weighted average of bid and ask quotes.

Acknowledgments

First and foremost I would like to express my sincere gratitude to my supervisors, Professors Liangjun Su and Peter C.B. Phillips, for all help and guidance during the past five years. Professor Su has taught me, both consciously and unconsciously, how good research is done. With his enthusiastic, his inspiration, and his great efforts to explain things clearly and accurately, he guide me into the world of econometric theory.

Professor Phillips is someone you will instantly love and never forget once you meet him. He is one of the smartest people I know. He has been always supportive and has encouraged me to pursue various projects without objection. Throughout my Ph.D. period, he provided sound advice, excellent teaching, and lots of good ideas. He make me believe in myself and I would have been lost without him.

I would also like to thank my committee members, professor Sainan Jin, professor Qu Feng. I want to thank you for letting my defense be an enjoyable moment, and for your brilliant comments and suggestions, thanks to you. I would especially like to thank professor Anthony Tay, and secretaries in the School of Economics at Singapore Management University. All of you have been there to support me in my Ph.D. life.

I wish to thank my friends and colleagues, Leyi Pan, Siyi Song, Ke Miao, Wuyi Wang, Xiaobin Liu, Yonghui Zhang, Yichong Zhang, Xin Zheng, and Yuan Zhuang, for helping me get through the tough time, and for all emotional support, camaraderie, entertainment, and caring they provided.

Lastly, and most importantly, I deeply thank my parents, Wenjun Li and Zhouchuan Huang for their endless love, unconditional support and timely encouragement. It was their love that raised me up again to more than I can be. This thesis is dedicated to my parents.

Contents

1 Identifying Latent Grouped Patterns in Cointegrated Panels

1.1 Introduction

Recently there has been a growing literature on large dimensional panel with latent group structures; see Lin and Ng (2012), Bonhomme and Manresa (2015) (BM hereafter), Sarafidis and Weber (2015), Ando and Bai (2016), Su et al. (2016) (SSP heterafter), Su and Ju (2017), Su et al. (2017), Lu and Su (2017), among others. In comparison with other approaches to model unobserved heterogeneity in panel data models, an important advantage of the latent group structures is that it offers a flexible way to modeling unobserved heterogeneity while maintaining certain degree of parsimony. Two popular methods have been proposed to identify the unknown group structures. One is based on the celebrated K-means clustering algorithm, and the other is based on the C-Lasso. For example, Lin and Ng (2012) and Sarafidis and Weber (2015) consider a heterogeneous linear regression panel data model where the slope coefficients exhibit an unknown group structure whereas BM consider a homogeneous linear panel data model where the additive fixed effects exhibit group structure. Both group of authors propose to apply the K-means clustering algorithm to achieve classification. Ando and Bai (2016) extend BM's approach to allow for group structure among the interactive fixed effects. Motivated by the sparse feature of the slope coefficients under latent group structures, SSP propose a novel variant of the Lasso procedure, i.e., classifier Lasso (C-Lasso), to achieve classification and estimation for both linear and nonlinear panel data models with or without endogeneity. Su and Ju (2017) extend SSP's C-Lasso to panel data models with interactive fixed effects; Su et al. (2017) consider C-Lasso-based sieve estimation of time-varying panel data models with latent structures; Lu and Su (2017) propose a sequential testing procedure to determine the unknown number of groups.

In this paper, we consider identifying the latent group structures in nonstationary panels where some regressors are generated from an integrated process. Despite the vast and diverse literature on nonstationary panels, most studies focus on panel unit root or cointegration tests with or without cross-sectional dependence and the literature on formal cointegration analysis is relatively sparse. Depending on whether the cointegrating relationship is allowed to be heterogeneous, one may consider either homogeneous or heterogeneous cointegrating relations. For example, Phillips and Moon (1999) consider general limit theory for both cases in large dimensional nonstationary panels; Groen and Kleibergen (2003) consider likelihood-based cointegration analysis for heterogeneous and homogeneous panel vector error-correction models; Kao and Chiang (2001) consider both dynamic OLS (DOLS) and fully modified OLS (FMOLS) estimation and inference in homogeneous cointegrated panels; Mark and Sul (2003) consider panel DOLS in homogeneous nonstationary panels; Bai et al. (2009) study homogeneous panel cointegration with global stochastic trends; Pedroni (2001) considers FMOLS for heterogeneous cointegrated panels. So the long-run cointegrating relationships can be assumed to be either homogeneous or heterogeneous and we face a trade-off between assuming heterogeneous long-run relationships, which is surely robust and perhaps also close to the reality, and estimating a common or at least an average long-run relationship, which offers efficiency in estimation and inference if the underlying homogeneous assumption is correct.

Despite the different treatments on the long-run relationships, the short-run dynamics, the individual intercepts, or the individual time trends, if exist, are commonly assumed to be heterogeneous across individuals. In this paper, we shall maintain the individual heterogeneity assumption on the individual effects and short-run dynamics and take an intermediate approach to model the long-run relationship. We propose a panel cointegration model with latent group structures where the long-run relationships are homogeneous within a group and heterogeneous across different groups, and the short run dynamics are allowed to be completely heterogeneous. The key issue is that the individual group membership is unknown and has to be estimated from the data together with the other parameters in the model. We extend SSP's C-Lasso method to the nonstationary panel framework. We consider the SSP's C-Lasso method rather than the K-means clustering algorithm for two reasons. First, the C-Lasso method has computational advantage over the K-means clustering algorithm. As SSP argue, the C-Lasso problem can be transformed into a sequence of convex problems to be solved easily while the K-means procedure is NP hard and tends to be much more computationally involved than the C-Lasso method. Second, the asymptotic theory for the C-Lasso method is well understood for stationary panels. It is natural to extend the theory to nonstationary panels. We will propose a C-Lasso-based penalized least squares (PLS) procedure to identify the unknown group structures and estimate the other parameters in the model.

Nevertheless, the extension of the asymptotic theory from stationary panels to nonstationary panels is technically challenging for several reasons. First, there is a lack of certain uniform convergence results in the nonstationary panel literature. It is well known that both the K-means clustering algorithm and the C-Lasso method enjoy certain oracle properties, which means the resulting estimators are as asymptotically efficient as if the latent group structures were known. But the establishment of such oracle properties rely on the application of certain exponential inequalities that are available for weakly dependent data as in stationary panels but not available for strongly dependent data as in nonstationary panels. To achieve the extension, we first need to establish some uniform convergence results associated with the nonstationary I(1) variables. Second, we allow for both stationary and nonstationary regressors in our cointegration models. Even though the number of nonstationary regressors is assumed to be fixed, we allow the dimension of stationary regressors to grow with the sample size at a controllable rate. The latter is very important for us to explore the idea of DOLS and develop a panel dynamic PLS procedure. Even though the growing dimension of the stationary regressors does not affect the convergence rate of the estimators of the long-run relationships, it complicates the asymptotic analysis in several places.

We assume that the number of groups is known and study the asymptotic properties of the PLS estimators. We first establish the preliminary rates of convergence for the coefficient estimators and show that, as expected, the long-run parameters can be estimated consistently at a faster rate than the short-run parameters. Given these preliminary consistency rates, we establish the uniform classification consistency of the C-Lasso method, which essentially means that all parameters within a group can be classified into the same group with probability approaching 1 (w.p.a.1) and all individuals that are classified into the same group indeed belong to the same group w.p.a.1. Such a uniform classification consistency lays down the foundation for the study of the asymptotic distributions of the PLS estimators. We show that both the C-Lasso estimators of the long-run parameters and their post-Lasso versions enjoy the asymptotic oracle properties and derive the asymptotic distribution under the joint limit theory.[1] We show that such presence of endogeneity in both nonstationary and stationary regressors does not cause the inconsistency of the long-run parameter estimators but does yield asymptotic bias in the estimators of both the short-run and long-run parameters. To remove the asymptotic bias in the estimation of the long-run parameters, we explore the idea of DOLS in the time series framework and propose a C-Lasso-based dynamic PLS procedure. When the number of groups is unknown, we propose an information criterion to determine the number of groups. Simulations show superb finite sample performance of the information criterion and C-Lasso-based PLS procedure.

As an empirical illustration, we apply our method to re-examine the validity of long-run PPP in the post-Bretton Woods period from 1975-2014 for a panel of 99 countries. Due to the establishment of the European Union in 1999, we divide the period into two parts 1975-1998 and 1999-2014. Then we estimate the long-run group-specific relationships by the dynamic PLS method. In general, we observe heterogeneous behavior on the long-run relation between nominal exchange rate and aggregate price ratio. We find two groups in the 1975-1999 subsample, with one group of countries in favor of the validity of PPP and the other group against the PPP hypothesis. In the 1999-2014 subsample, we identify three groups and significant evidence supporting the long-run PPP hypothesis in one group. There are more countries in this group in favor of the validity of the long-run PPP hypothesis in this period. We explain these results by the “Revived Bretton Woods system” (also called as Bretton Woods II in the literature) from 2000, see Dooley et al. (2004). These results confirm the belief that at least some selected group of countries obey the long-run PPP rule in the post-Bretton Woods period.

The rest of this paper is organized as follows. We introduce the cointegrated panel data model with latent group structures and propose a C-Lasso-based PLS estimation proce-

[1] Most asymptotic theories in the panel cointegration analysis have been established under the sequential limit theory. A few exceptions include Phillips and Moon (1999), Sun (2004), and Bai and Ng (2010).

dure in Section 2. Section 3 introduces the main assumptions for our asymptotic analysis. We study the asymptotic properties of the PLS estimators and propose an information criterion to determine the number of groups in Section 4. Section 5 reports Monte Carlo simulation results. Section 6 applies the dynamic PLS method to testing the long-run PPP hypothesis. Section 7 concludes. We relegate the proofs of the main results to Appendix A and those of technical lemmas to the online supplementary material.

NOTATION. For any real matrix A, we write the transpose A', the Frobenius norm $||A||$, the spectral norm $||A||_{\text{sp}}$, and the Moore-Penrose inverse as A^{+}. When A is symmetric, we use $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ to denote its largest and smallest eigenvalues, respectively. I_a and $\mathbf{0}_{a\times b}$ denote the $a \times a$ identity matrix and $a \times b$ matrices of zeros, and $\mathbf{1}\{\cdot\}$ is the usual indicator function. The operator $\xrightarrow{P}$ denotes convergence in probability, $\Rightarrow$ weak convergence, a.s. almost surely, and plim probability limit. We use $(N, T) \to \infty$ to signify that N and T pass jointly to infinity.

1.2 Model and Estimation

In this section we introduce the panel cointegration model with latent group structures and then propose a C-Lasso-based penalized least squares method to estimate the model.

1.2.1 Panel cointegration model with latent group structures

The dependent variable y_{it} is measured for individuals $i = 1, 2, ..., N$ over time $t = 1, 2, ..., T$. We suppose that the nonstationary I(1) variables y_{it} and $x_{1,it}$ are generated according to the following heterogeneous panel cointegration model

$$\begin{cases} y_{it} = \mu_i + \beta_{1,i}' x_{1,it} + \beta_{2,i}' x_{2,it} + u_{it} \\ x_{1,it} = x_{1,it-1} + \varepsilon_{1,it}, \end{cases} , \tag{1}$$

where μ_i is the unobserved individual fixed effect, $x_{1,it}$ is a $p_1 \times 1$ vector of nonstationary regressors of order one (I(1)) process) for all i, $x_{2,it}$ is a $p_2 \times 1$ vector of stationary regressors (I(0) process) for all i, u_{it} is the idiosyncratic error term with mean zero and finite long-run variance, $\varepsilon_{1,it}$ is also assumed to have zero mean and finite long-run variance, and $\beta_{1,i}$ and $\beta_{2,i}$ are p_1 and p_2 dimensional slope coefficients, respectively.

We assume that p_1 is fixed but allow p_2 to diverge to infinity at certain rate. The latter is very important because we will extend our theory to the panel DOLS framework. In this case, the first equation in (1) becomes

$$y_{it} = \mu_i + \beta'_{1,i} x_{1,it} + \sum_{j=-\bar{p}_2}^{\bar{p}_2} \gamma'_{i,j} \Delta x_{1,i,t+j} + v^{\dagger}_{it}, \tag{2}$$

where $\Delta x_{1,it} = x_{1,it} - x_{1,i,t-1}$, $x_{2,it}$ only contains the lags and leads of $\Delta x_{1,it} : x_{2,it} = (\Delta x'_{1,i,t-\bar{p}_2}, ..., \Delta x'_{1,i,t+\bar{p}_2})'$, $\beta_{2,i} = (\gamma'_{i,-\bar{p}_2}, ..., \gamma'_{i,\bar{p}_2})'$, $p_2 = (2\bar{p}_2 + 1)p_1$, $\bar{p}_2$ is divergent with T, and $v^{\dagger}_{it}$ is the new error term that typically contains some approximation errors.

In the literature on nonstationary panels, $\beta_{1,i}$, which stands for the long-run cointegrating relationship, can be either homogeneous or heterogeneous, whereas $\beta_{2,i}$, which represents the short-run dynamics, is allowed to be heterogeneous across all individuals in almost all studies. In this paper we maintain the heterogeneity assumption on $\beta_{2,i}$'s but follow the lead of SSP and assume that $\beta_{1,i}$'s are heterogeneous across groups and homogeneous within a group. We allow the true values of $\beta_{1,i}$, denoted as $\beta^0_{1,i}$, to follow a grouped pattern of the general form

$$\beta^0_{1,i} = \begin{cases} \alpha^0_1 & \text{if } i \in G^0_1 \\ \vdots & \vdots \\ \alpha^0_K & \text{if } i \in G^0_K \end{cases}, \tag{3}$$

where $\alpha^0_j \neq \alpha^0_k$ for any $j \neq k$, $\cup_{k=1}^K G^0_k = \{1, 2, \ldots N\}$, and $G^0_k \cap G^0_j = \varnothing$ for any $j \neq k$. For now, we assume that the number of groups, K, is known and fixed. Let $\boldsymbol{\alpha} \equiv (\alpha_1, \ldots, \alpha_K)$, $\boldsymbol{\beta_1} \equiv (\beta_{1,1}, \ldots, \beta_{1,N})$, and $\boldsymbol{\beta_2} \equiv (\beta_{2,1}, \ldots, \beta_{2,N})$. We denote their true values as $\boldsymbol{\alpha^0}$, $\boldsymbol{\beta^0_1}$, and $\boldsymbol{\beta^0_2}$, respectively. We also use $\beta^0_{2,i}$ and α^0_k to denote the true coefficients of $\beta_{2,i}$ and α_k. We will use $N_k \equiv \# G^0_k$ to denote the cardinality of the set G^0_k.

We are interested in identifying each individual's group membership and estimating the long-run cointegrating group-specific coefficients, α_k, $k = 1, ..., K$.

Even though we focus only on the linear cointegrating model in this paper, the theory that we are developing is quite different from that in SSP for three main reasons. First, the presence of nonstationary regressors substantially complicates the asymptotic analysis. In particular, we need to establish some uniform convergence rates that are not available in the nonstationary panel literature. Second, the increasing dimension of the stationary regressors in the model also complicates the issue. Third, we allow for endogeneity in both $x_{1,it}$ and $x_{2,it}$. In the time series framework, it is well known that the endogeneity of either the I(1) or I(0) regressor does not cause the inconsistency of the OLS estimator of the long-run relationship. We will show that similar phenomenon occurs in the panel setup.

1.2.2 Penalized least squares estimation

Without imposing the latent group structures in (3), we can estimate $\beta_{1,i}$ and $\beta_{2,i}$ in (1) by using the fixed effects estimator. In this case, we consider the within-group transformation

$$\tilde{y}_{it} = \beta_{1,i}'\tilde{x}_{1,it} + \beta_{2,i}'\tilde{x}_{2,it} + \tilde{u}_{it}, \tag{4}$$

or in vector-matrix form

$$\tilde{y}_i = \tilde{x}_{1,i}\beta_{1,i} + \tilde{x}_{2,i}\beta_{2,i} + \tilde{u}_i, \tag{5}$$

where $\tilde{y}_i = (\tilde{y}_{i1}, ..., \tilde{y}_{iT})'$, $\tilde{y}_{it} = y_{it} - \bar{y}_i$, $\bar{y}_i = \frac{1}{T}\sum_{t=1}^T y_{it}$, and $\tilde{x}_{1,it}$, $\tilde{x}_{2,it}$, $\tilde{u}_{it}$, $\bar{x}_{1,i}$, $\bar{x}_{2,i}$, $\bar{u}_i$, $\tilde{x}_{1,i}$, $\tilde{x}_{2,i}$, and $\tilde{u}_i$ are analogously defined. The FE estimators $\tilde{\beta}_{1,i}$ and $\tilde{\beta}_{2,i}$ are obtained as the minimizers of the following least squares criterion function

$$Q_{NT}(\boldsymbol{\beta_1}, \boldsymbol{\beta}_2) = \frac{1}{NT^2}\sum_{i=1}^N \|\tilde{y}_i - \tilde{x}_{1,i}\beta_{1,i} - \tilde{x}_{2,i}\beta_{2,i}\|^2 = \frac{1}{NT^2}\sum_{i=1}^N \|\tilde{y}_i - \tilde{x}_i\beta_i\|^2, \tag{6}$$

where $\beta_i = (\beta_{1,i}', \beta_{2,i}')'$ and $\tilde{x}_i = (\tilde{x}_{1,i}, \tilde{x}_{2,i})$ has a typical row $\tilde{x}_{it}' = (\tilde{x}_{1,it}', \tilde{x}_{2,it}')$. Let $\tilde{\beta}_i = (\tilde{\beta}_{1,i}', \tilde{\beta}_{2,i}')'$. Then $\tilde{\beta}_i = (\tilde{x}_i'\tilde{x}_i)^{-1}(\tilde{x}_i'\tilde{y}_i)$ for each i.

To explore the latent group structure of $\beta_{1,i}$'s in (3), we propose to estimate $\boldsymbol{\beta_1}$, $\boldsymbol{\beta_2}$, and $\boldsymbol{\alpha}$ by minimizing the following C-Lasso-based penalized least squares (PLS) criterion function

$$Q_{NT,\lambda}^K(\boldsymbol{\beta_1}, \boldsymbol{\beta}_2, \boldsymbol{\alpha}) = Q_{NT}(\boldsymbol{\beta_1}, \boldsymbol{\beta}_2) + \frac{\lambda}{N}\sum_{i=1}^N (\tilde{\sigma}_i)^{2-K}\prod_{k=1}^K \left\| \hat{Q}_{1i}(\beta_{1,i} - \alpha_k) \right\|, \tag{7}$$

where $\lambda = \lambda(N,T)$ is a tuning parameter, $\tilde{\sigma}_i^2 = \frac{1}{T}\sum_{t=1}^T(\tilde{y}_{it} - \tilde{\beta}_i'\tilde{x}_{it})^2$, and $\hat{Q}_{1i} = \frac{1}{T^2}\sum_{t=1}^T \tilde{x}_{1,it}\tilde{x}_{1,it}'$. When $\tilde{\sigma}_i$ and $\hat{Q}_{1i}$ are replaced by 1 and I_{p_1}, respectively, the penalty term in (7) reduces to that in SSP. Here, we introduce these two terms into the penalty to ensure the scale-invariant property of the penalized estimators.

Minimizing the above objective function yields the C-Lasso-based PLS estimates $\hat{\boldsymbol{\beta}}_1$, $\hat{\boldsymbol{\beta}}_2$, and $\hat{\boldsymbol{\alpha}}$. Let $\hat{\beta}_{1,i}$ and $\hat{\alpha}_k$ denote the i^{th} and k^{th} columns of $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\alpha}}$, respectively, i.e., $\hat{\boldsymbol{\beta}}_1 \equiv (\hat{\beta}_{1,1}, ..., \hat{\beta}_{1,N})$ and $\hat{\boldsymbol{\alpha}} \equiv (\hat{\alpha}_1, ..., \hat{\alpha}_K)$. We will study the asymptotic properties of the C-Lasso estimators below.

1.3 Notations and Assumptions

In this section, we spell out the main notations and assumptions that are needed for the study of the asymptotic properties of our estimators.

Since we include the fixed effects μ_i in (1) and assume covariance-stationarity of $x_{2,it}$. We assume without loss of generality that $x_{2,it}$ has zero mean.[2] Let $\varepsilon_{it} = \left(u_{it}, \varepsilon_{1,it}', \varepsilon_{2,it}'\right)'$ where $\varepsilon_{2,it} = x_{2,it}$. The long-run covariance matrix of $\{\varepsilon_{it}\}$ is given by

$$\Omega_i = \sum_{j=-\infty}^{\infty} E(\varepsilon_{ij}\varepsilon_{i0}') = \begin{pmatrix} \Omega_{00,i} & \Omega_{01,i} & \Omega_{02,i} \\ \Omega_{10,i} & \Omega_{11,i} & \Omega_{12,i} \\ \Omega_{20,i} & \Omega_{21,i} & \Omega_{22,i} \end{pmatrix},$$

where, e.g., $\Omega_{00,i} = \sum_{j=-\infty}^{\infty} E(u_{ij}u_{i0}')$, $\Omega_{01,i} = \sum_{j=-\infty}^{\infty} E(u_{ij}\varepsilon_{1,i0}')$, and $\Omega_{02,i} = \sum_{j=-\infty}^{\infty} E(u_{ij}\varepsilon_{2,i0}')$. Following the literature on nonstationary panels, we will make the following decomposition

$$\Omega_i = \Sigma_i + \Lambda_i + \Lambda_i',$$

where $\Sigma_i = E\left(\varepsilon_{it}\varepsilon_{it}'\right)$ denotes the short-run variance of $\{\varepsilon_{it}\}$ and $\Lambda_i = \sum_{j=1}^{\infty} E(\varepsilon_{ij}\varepsilon_{i0}')$. We partition Σ_i and Λ_i conformably with ε_{it} and Ω_i :

$$\Sigma_i = \begin{pmatrix} \Sigma_{00,i} & \Sigma_{01,i} & \Sigma_{02,i} \\ \Sigma_{10,i} & \Sigma_{11,i} & \Sigma_{12,i} \\ \Sigma_{20,i} & \Sigma_{21,i} & \Sigma_{22,i} \end{pmatrix} \text{ and } \Lambda_i = \begin{pmatrix} \Lambda_{00,i} & \Lambda_{01,i} & \Lambda_{02,i} \\ \Lambda_{10,i} & \Lambda_{11,i} & \Lambda_{12,i} \\ \Lambda_{20,i} & \Lambda_{21,i} & \Lambda_{22,i} \end{pmatrix}.$$

Let $\Delta_i = \Sigma_i + \Lambda_i$ denote the one-sided long-run covariance of $\{\varepsilon_{it}\}$. Let $p = 1 + p_1 + p_2$ denote the dimension of ε_{it}. Let S_0, S_1, and S_2 denote respectively the $1 \times p$, $p_1 \times p$, and $p_2 \times p$ selection matrices such that $S_0\varepsilon_{it} = u_{it}$ and $S_\ell\varepsilon_{it} = \varepsilon_{\ell,it}$ for $\ell = 1, 2$.

[2] If $E(x_{2,it}) = \nu_{2i} \neq 0$, we can rewrite the first equation in (1) as

$$y_{it} = \mu_i^* + \beta_{1,i}' x_{1,it} + \beta_{2,i}' x_{2,it}^* + u_{it},$$

where $x_{2,it}^* = x_{2,it} - \nu_{2i}$ has zero mean and $\mu_i^* = \mu_i + \beta_{2,i}'\nu_{2i}$.

Let $\max_i = \max_{1 \leqslant i \leqslant N}$ and $\max_t = \max_{1 \leqslant t \leqslant T}$ unless otherwise stated. Define $\min_i$ and $\min_t$ analogously. We make the following assumptions.

Assumption A.1 (i) For each i, $\{\varepsilon_{it}, t \geqslant 0\}$ is a linear process such that

$$\varepsilon_{it} = \psi_i(L) e_{it} = \sum_{j=0}^{\infty} \psi_{ij} e_{i,t-j},$$

where $\{e_{it}\}$ is an independent process with zero mean and variance-covariance matrix I_p. Each element of e_{it} has finite $2(q+\epsilon)$ moments that are bounded uniformly in (i,t), where $q > 4$ and ϵ is an arbitrarily small positive number.

(ii) $\max_i \sum_{j=0}^{\infty} j^2 \|S\psi_{ij}\| < \infty$ for any selection matrix S that selects any finite (non-divergent) number of rows in ψ_{ij}.

(iii) For each i, $\{\varepsilon_{it}, t \geqslant 0\}$ is a strong mixing process with mixing coefficients $\alpha_i(t)$ satisfying $\max_i \alpha_i(\tau) \leqslant c_\alpha \rho^\tau$ for some $c_\alpha < \infty$ and $\rho \in (0,1)$.

(iv) $\{\varepsilon_{it}, t \geqslant 0\}$ are independent across i.

Assumption A.2 (i) There exists a constant $\underline{c}_{11}$ such that $\liminf_{T\to\infty} \lambda_{\min}\left(\frac{b_T}{T^2}\sum_{t=1}^{T} \tilde{x}_{1,it}\tilde{x}_{1,it}'\right) \geqslant \underline{c}_{11} > 0$ almost surely (a.s.), where b_T diverges to infinity as $T \to \infty$ slowly.

(ii) There exists a constant $\bar{c}_{\Omega_{11}}$ such that $\max_i \lambda_{\max}(\Omega_{11,i}) \leqslant \bar{c}_{\Omega_{11}} < \infty$

(iii) There exist constants $\underline{c}_{22}$ and $\bar{c}_{22}$ such that $0 < \underline{c}_{22} \leqslant \min_i \lambda_{\min}(\Sigma_{22,i}) \leqslant \max_i \lambda_{\max}(\Sigma_{22,i}) \leqslant \bar{c}_{22} < \infty$.

(iv) Let $\Sigma_{0.2,i}^* = \Sigma_{00,i} - \Sigma_{02,i}\Sigma_{22,i}^{-1}\Sigma_{20,i}$. There exist constants $\underline{c}_{00}$ and $\bar{c}_{00}$ such that $0 < \underline{c}_{00} \leqslant \min_i \Sigma_{0.2,i}^* \leqslant \max_i \Sigma_{00,i} \leqslant \bar{c}_{00} < \infty$.

Assumption A.3 (i) For each $k = 1, ..., K$, $N_k/N \to \tau_k \in (0,1)$ as $N \to \infty$.

(ii) $\min_{1 \leqslant k \neq j \leqslant K} \left\|\alpha_k^0 - \alpha_j^0\right\| \geqslant \underline{c}_\alpha$ for some fixed $\underline{c}_\alpha > 0$.

(iii) As $(N,T) \to \infty$, $N/T^2 \to c_1 \in [0,\infty)$, $T/N^2 \to c_2 \in [0,\infty)$, and $p_2^3 T^{-1} (\log T)^6 \to c_2 \in [0,\infty)$.

(iv) As $(N,T) \to \infty$, $\lambda b_T \log\log T \to 0$, $\lambda T N^{-\frac{1}{q}} b_T^{-(K+1)}/\log T \to \infty$, $b_T^{K+1} N^{1/q} T^{-1} \log T \to 0$, $b_T N^{2/q} T^{-1/2}/\log T \to 0$, and $b_T p_2^{1/2} N^{1/q} T^{-1/2} \log T = O(1)$.

Assumption A.1(i)-(ii) imposes that the innovation process $\{\varepsilon_{it}\}$ is a linear process that exhibits certain moment and summability conditions. When p_2 is fixed, the selection matrix S is not needed. In our asymptotic analysis, we will frequently call upon the

Beveridge and Nelson (1981) BN decomposition:

$$\varepsilon_{it} = \psi_i(1) e_{it} + \breve{e}_{i,t-1} - \breve{e}_{it}, \tag{8}$$

where $\psi_i(1) = \sum_{j=0}^{\infty} \psi_{ij}$, $\breve{e}_{it} = \sum_{j=0}^{\infty} \breve{\psi}_{ij} e_{i,t-j}$, and $\breve{\psi}_{ij} = \sum_{k=j+1}^{\infty} \psi_{ik}$. Following Phillips and Solo (1992) (p.989), Assumption A.1(i)-(ii) ensures that

$$\max_i \max_t E \|S\breve{e}_{it}\|^{2q} < \infty$$

for any selection matrix S such that $S\breve{e}_{it}$ selects only a fixed number of elements in $\breve{e}_{it}$. For example, $S = (S_0', S_1')'$ selects the first $1+p_1$ elements $\breve{e}_{it}$ that corresponds to $\left(u_{it}, \varepsilon_{1,it}'\right)'$. Assumption A.1(iii) assumes that $\{\varepsilon_{it}, t \geqslant 0\}$ is a strong mixing process for the convenience of using some Bernstein-type exponential inequality that is available for strong mixing processes. Davidson (1994) (Chapter 14.4) provides some sufficient conditions to verify that a linear process of the type in Assumption A.1(i) is strong mixing. The geometric mixing rate can be relaxed to being algebraic with a little bit more involved notation in the proofs. Here we follow SSP and assume the geometric mixing rate condition for simplicity. By White (2001) (Theorem 7.18), Assumption A.1(i)-(iii) is far more sufficient to ensure the functional central limit theorem (FCLT) holds for $\{S\varepsilon_{it}, t \geqslant 0\}$ for each i provided its long-run variance-covariance matrix is positive definite. Assumption A.1(iv) imposes cross-sectional independence, as was done in the early literature on panel cointegration analyses (see, e.g., Phillips and Moon (1999); Kao and Chiang (2001); Mark and Sul (2003)). We do not relax such an assumption in this paper because even under this restrictive assumption, the rigorous asymptotic analysis is already extremely involved.

Assumption A.2(i) requires that $\hat{Q}_{1i} \equiv \frac{1}{T^2} \sum_{t=1}^{T} \tilde{x}_{1,it} \tilde{x}_{1,it}'$ is well behaved uniformly in i. For each i, we can readily apply the results in Park and Phillips (1988) and show that

$$\hat{Q}_{1i} \Rightarrow \int_0^1 \tilde{B}_{1,i}(r) \tilde{B}_{1,i}(r)' dr$$

where $\tilde{B}_{1,i} = B_{1,i} - \int_0^1 B_{1,i}(r) dr$, and $B_{1,i}$ is a p_1-dimensional Brownian motion with covariance $\Omega_{00,i}$. In this case, as long as $\Omega_{00,i}$ is positive definite, we can ensure that $\hat{Q}_{1i}$ is asymptotically nonsingular for each i. For our asymptotic analysis, we require that both the maximum and minimum eigenvalues of $\hat{Q}_{1i}$ are well behaved uniformly in i. For the maximum eigenvalue, we can call upon the usual law of iterated logarithm (LIL) and show that

$$\limsup_{T\to\infty} \lambda_{\max}(\hat{Q}_{1i}/(2\log\log T)) < \left(\frac{1}{2}+\epsilon\right)\bar{c}_{\Omega_{11}}\text{a.s.}$$

where ϵ is an arbitrarily small positive number and $\bar{c}_{\Omega_{11}}$ is a constant defined in Assumption A.2(ii). For the minimum eigenvalues, we conjecture that one can call upon the "other" or Chung-type LIL (see, e.g., Donsker and Varadhan (1977)) and show that Assumption A.2(i) holds with $b_T = \log\log T$. But the rigorous justification is beyond the scope of this paper. See Lai and Wei (1982), Phillips (1996) and Bai (2004) who apply similar conditions in their asymptotic analyses. Assumption A.2(ii)-(iii) imposes some conditions on the eigenvalues of nonstochastic square matrices. Assumption A.2(iv) is imposed to ensure nondegenerate limiting distribution. Given Assumption A.2(iii), it implicitly implies that $\Sigma_{20,i}'\Sigma_{20,i}$ is bounded away from the infinity and thus restrict the degree of endogeneity in the stationary regressors.

Assumption A.3(i)-(ii) is commonly assumed in the panel literature with latent group structures; see, e.g., Bonhomme and Manresa (2015), Ando and Bai (2016), Su et al. (2016), Lu and Su (2017), and Su and Ju (2017). In particular, Assumption A.3(ii) requires the separability of the group-specific parameters. Assumption A.3(iii) imposes conditions on N, T, and p_2. Note that we do not require $N = o(T)$ as in most studies on nonstationary panels under the joint limit theory (see, e.g., Phillips and Moon (1999); Bai and Ng (2010)). The last condition in Assumption A.3(iii) is analogous to the condition $p_2^3 T^{-1} = o(1)$ in the time series framework (e.g., Saikkonen (1991)). Assumption A.3(iv) looks quite complicated but can be simplified in various cases. First, if N and T pass to infinity at the same rate, which appears plausible in most macro applications, it reduces to:

Assumption A.3(iv*) As $(N,T)\to\infty$, $\lambda b_T \log\log T \to 0$, $\lambda T^{1-\frac{1}{q}} b_T^{-(K+1)}/\log T \to \infty$, $b_T^{K+1} T^{\frac{1}{q}-1}\log T \to 0$, $b_T T^{\frac{2}{q}-\frac{1}{2}}/\log T \to 0$, and $b_T p_2^{1/2} T^{\frac{1}{q}-\frac{1}{2}}\log T = O(1)$.

Second, if Assumption A.2(i) is satisfied with $b_T = \log\log T$, a sufficient condition for Assumption A.3(iv*) to hold is as follows:

Assumption A.3(iv)** As $(N,T)\to\infty$, $\lambda(\log\log T)^2 \to 0$, and $\lambda T^{1-\frac{1}{q}}/(\log T)^2 \to \infty$.

Here, we use the fact that $q > 4$ so that the third through last conditions in Assumption A.3(iv*) become redundant under Assumption A.3(iii). Then we can find a large range of values for λ satisfying Assumption A.3(iv**). It is sufficient to have

$$\lambda \propto T^{-\alpha} \text{ for } \alpha \in \left(0, \ \frac{q-1}{q}\right).$$

When q is sufficiently large (e.g., the tails of the error terms decay exponentially fast), the upper bound for α is arbitrarily close to 1. If we only require $q > 4$, then it is fine to choose $\lambda \propto T^{-3/4}$.

1.4 Asymptotic Properties

In this section, we first find the preliminary rates of convergence for the coefficient estimators and prove classification consistency. Then we study the oracle properties of C-Lasso estimators and their post-Lasso versions. The special case of panel dynamic PLS is also considered and a BIC-type information criterion is proposed to determine the number of groups.

1.4.1 Preliminary rates of consistency

Let $\beta_i^* = (\beta_{1,i}^{0\prime}, \beta_{2,i}^{*\prime})'$, where $\beta_{2,i}^* = \beta_{2,i}^0 + \Sigma_{22,i}^{-1}\Sigma_{20,i}$. The following theorem establishes the preliminary rates of consistency for both $\hat{\beta}_i$ and $\hat{\alpha}_k$.

Theorem 1 *Suppose that Assumptions A.1-A.3 hold. Then*

(i) $||\hat{\beta}_{1,i} - \beta_{1,i}^0|| = O_P(T^{-1} + \lambda)$ *and* $||\hat{\beta}_{2,i} - \beta_{2,i}^*|| = O_P(p_2^{1/2}(T^{-1/2} + \lambda))$ *for* $i = 1, ..., N$,

(ii) $\frac{1}{N}\sum_{i=1}^N ||\hat{\beta}_{1,i} - \beta_{1,i}^0||^2 = O_P(b_T^2 T^{-2})$ *and* $\frac{1}{N}\sum_{i=1}^N ||\hat{\beta}_{2,i} - \beta_{2,i}^*||^2 = O_P(p_2 T^{-1})$,

(iii) $(\hat{\alpha}_{(1)}, ..., \hat{\alpha}_{(K)}) - (\alpha_1^0, ..., \alpha_K^0) = O_P(b_T T^{-1})$ *where* $(\hat{\alpha}_{(1)}, ..., \hat{\alpha}_{(K)})$ *is a suitable permutation of* $(\hat{\alpha}_1, ..., \hat{\alpha}_K)$.

Theorems 1(i) and (ii) establish the point-wise and mean-square convergence of $\hat{\beta}_i = (\hat{\beta}_{1i}', \hat{\beta}_{2,i}')'$, respectively; Theorem 1(iii) indicates that $\hat{\alpha}_1, ..., \hat{\alpha}_K$, consistently estimate the true group-specific coefficients, $\alpha_1^0, ..., \alpha_K^0$, subject to suitable permutation. We summarize some interesting findings. First, despite the presence of endogeneity in both the nonstationary and stationary regressors, we can estimate the true coefficients ($\beta^0_{1,i}$) of the nonstationary regressors consistently. Second, when $\Sigma_{20,i}$ is nonzero, we cannot estimate the true coefficients ($\beta^0_{2,i}$) of the stationary regressors consistently. Instead, $\hat{\beta}_{2,i}$ is consistent with the pseudo true value $\beta^*_{2,i} = \beta^0_{2,i} + \Sigma^{-1}_{22,i}\Sigma_{20,i}$, where $\Sigma^{-1}_{22,i}\Sigma_{20,i}$ signifies the endogeneity bias. Third, the effect of increasing dimension (p_2) appears in the rates of convergence for $\hat{\beta}_{2,i}$ but not in those for $\hat{\beta}_{1,i}$. Apparently, $\hat{\beta}_{1,i}$'s converge to their true values faster than $\hat{\beta}_{2,i}$'s to their pseudo-true values. Fourth, as in SSP, the pointwise convergence of $\hat{\beta}_i$ depends on λ while the mean square convergence of $\{\hat{\beta}_i\}$ and the convergence of $\hat{\alpha}_k$'s do not. As we have shown in the proof, the convergence of $\hat{\alpha}_k$ only depends on the mean square convergence of $\{\hat{\beta}_{1,i}\}$.

For notational simplicity, hereafter we will write $\hat{\alpha}_{(k)}$ as $\hat{\alpha}_k$. We define the estimated groups

$$\hat{G}_k = \{i \in \{1, 2, ..., N\} : \hat{\beta}_{1,i} = \hat{\alpha}_k\} \text{ for } k = 1, ..., K.$$

To study the classification consistency, we need to establish the uniform consistency of $\hat{\beta}_{1,i}$ and $\hat{\beta}_{2,i}$. This is reported in the next theorem.

Theorem 2 *Suppose that Assumptions A.1-A.3 hold. Then for any fixed* $c > 0$,

(i) $P(\max_{1\leqslant i\leqslant N} ||\hat{\beta}_{1,i} - \beta^0_{1,i}|| \geqslant c b_T a_{1NT}) = o\left(N^{-1}\right)$,

(ii) $P(\max_{1\leqslant i\leqslant N} ||\hat{\beta}_{2,i} - \beta^*_{2,i}|| \geqslant c p_2^{1/2} a_{2NT}) = o\left(N^{-1}\right)$,

where $a_{1NT} = T^{-1}N^{1/q}\left(\log T\right)^{(1+\epsilon)/2}$ *for some arbitrarily small* $\epsilon > 0$, *and* $a_{2NT} = T^{-1/2}\left(\log T\right)^3$.

The uniform convergence rate of $\hat{\beta}_{1,i}$ is not affected by p_2 but is slower than the time series convergence rate T^{-1}. The higher q is (which means the higher order moments the error terms exhibit), the closer a_{1NT} is to T^{-1}. When the error terms have exponentially decaying tails as assumed in Bonhomme and Manresa (2015), we can make a_{1NT} arbitrarily close to T^{-1} subject to the logarithm factor.

1.4.2 Classification consistency

To study the classification consistency, we follow SSP and define the following two sequences of events

$$\hat{E}_{kNT,i} = \{i \notin \hat{G}_k | i \in G_k^0\} \quad \text{and} \quad \hat{F}_{kNT,i} = \{i \notin G_k^0 | i \in \hat{G}_k\}$$

where $i = 1, ..., N$ and $k = 1, ...K$. Let $\hat{E}_{kNT} = \cup_{i \in \hat{G}_k} \hat{E}_{kNTi}$ and $\hat{F}_{kNT} = \cup_{i \in \hat{G}_k} \hat{F}_{kNTi}$. $\hat{E}_{kNT}$ denotes the error event of not classifying an element of G_k^0 into estimated group $\hat{G}_k$; and $\hat{F}_{kNT}$ denotes the error event of classifying an element that does not belong to G_k^0 into the estimated group $\hat{G}_k$. Following SSP, we say that a classification method is *individually consistent* if $P(\hat{E}_{kNT,i}) \to 0$ and $P(\hat{F}_{kNT,i}) \to 0$ as $(N, T) \to \infty$ for each $i \in G_k^0$ and $k = 1, ..., K$, and it is *uniformly consistent* if $P(\cup_{k=1}^K \hat{E}_{kNT}) \to 0$ and $P(\cup_{k=1}^K \hat{F}_{kNT}) \to 0$ as $(N, T) \to \infty$.

The following theorem establishes the uniform classification consistency.

Theorem 3 *Suppose that Assumptions A.1-A.3 hold. Then as* $(N, T) \to \infty$

(i) $P(\cup_{k=1}^K \hat{E}_{kNT}) \leqslant \sum_{k=1}^K P(\hat{E}_{kNT}) \to 0,$

(ii) $P(\cup_{k=1}^K \hat{F}_{kNT}) \leqslant \sum_{k=1}^K P(\hat{F}_{kNT}) \to 0.$

Theorem 3 implies that all individuals within certain group, say G_k^0, can be simultaneously correctly classified into the same group (denoted as $\hat{G}_k$) w.p.a.1. Conversely, all individuals that are classified into the same group, say $\hat{G}_k$, simultaneously correctly belong to the same group (G_k^0) w.p.a.1. The result implies that in large samples, we can virtually take the estimated group as the true group. In particular, let $\hat{N}_k = \#\hat{G}_k$. One can easily show that $P(\hat{G}_k = G_k^0) \to 1$ so that $P(\hat{N}_k = N_k) \to 1$.

1.4.3 Oracle properties and post-Lasso estimators

To study the oracle property of the C-Lasso-based PLS estimators, we add some notations:

$$\begin{aligned}
\mathbb{Q}_{(k)} &\equiv \lim_{N_k\to\infty} \frac{1}{6N_k} \sum_{i\in G_k^0} S_1 \psi_i\left(1\right) \psi_i\left(1\right)' S_1' = \lim_{N_k\to\infty} \frac{1}{6N_k} \sum_{i\in G_k^0} \Omega_{11,i}, \\
\mathbb{B}_{k,NT} &\equiv \mathbb{B}_{1k,NT} + \mathbb{B}_{2k,NT}, \\
\mathbb{B}_{1k,NT} &= \frac{1}{\sqrt{N_k}} \sum_{i\in G_k^0} S_1 \sum_{r=0}^{\infty} \sum_{s=0}^{\infty} \psi_{i,s+r} \psi_{i,s}' s_i, \\
\mathbb{B}_{2k,NT} &= \frac{-1}{\sqrt{N_k}} \frac{T+1}{2T} \sum_{i\in G_k^0} S_1 \psi_i\left(1\right) \psi_i\left(1\right)' s_i, \\
\mathbb{V}_{(k)} &\equiv \lim_{N_k\to\infty} \frac{1}{N_k} \sum_{i\in G_k^0} \left(\frac{1}{6} s_i \Omega_i s_i' S_1 \Omega_i S_1' - \frac{1}{12} \left(s_i \Omega_i S_1' \otimes S_1 \Omega_i s_i' \right) K_{p,1} \right), \\
\mathbb{V}_{22,i} &= \left(\Sigma_{22,i}^{-1} J_{1,i} \otimes J_{2,i} \right) V_i^0 \left(J_{1,i}' \Sigma_{22,i}^{-1} \otimes J_{2,i}' \right),
\end{aligned}$$

where $s_i = S_0' - S_2' \Sigma_{22,i}^{-1} \Sigma_{20,i}$, $J_{1,i} = \left(\mathbf{0}_{p_2\times 1}, \mathbf{0}_{p_2\times p_1}, I_{p_2} \right)$, $J_{2,i} = \left(1, \mathbf{0}_{1\times p_1}, -\Sigma_{20,i}' \Sigma_{22,i}^{-1} \right)$, and $V_i^0 = \lim_{T\to\infty} \text{Var}(T^{-1/2} \sum_{t=1}^{T} \text{vec}(\varepsilon_{it}\varepsilon_{it}' - \Sigma_i))$.

The following theorem reports the asymptotic properties of $\hat{\alpha}_k$ and $\hat{\beta}_{2,i}$.

Theorem 4 *Suppose that Assumptions A.1-A.3 hold. Let $\mathbb{S}_2$ denote an $l \times p_2$ selection matrix such that $\mathbb{S}_2 \beta_{2,i}$ selects only l elements in $\beta_{2,i}$, where l is a fixed integer that does not grow with (N, T). Then*

(i) $\sqrt{N_k} T (\hat{\alpha}_k - \alpha_k^0) - \mathbb{Q}_{(k)}^{-1} \mathbb{B}_{k,NT} \Rightarrow N(0, \mathbb{Q}_{(k)}^{-1} \mathbb{V}_{(k)} \mathbb{Q}_{(k)}^{-1})$ as $(N, T) \to \infty$ for $k = 1, ..., K$,

(ii) $\sqrt{T} \mathbb{S}_2 \left(\hat{\beta}_{2,i} - \beta_{2,i}^* \right) \Rightarrow N\left(0, \mathbb{S}_2 \mathbb{V}_{22,i} \mathbb{S}_2'\right)$ as $T \to \infty$ for each $i = 1, ..., N$.

To understand the above results, we consider the case where the group membership is known. In this case, the oracle estimators of α_k and $\beta_{2,i}$ are respectively given by

$$\begin{aligned}
\hat{\alpha}_k^{\text{oracle}} &= \left(\sum_{i\in G_k^0} \tilde{x}_{1,i}' M_{2,i} \tilde{x}_{1,i} \right)^{-1} \sum_{i\in G_k^0} \tilde{x}_{1,i}' M_{2,i} \tilde{y}_i \text{ for } k = 1, ..., K, \\
\hat{\beta}_{2,i}^{\text{oracle}} &= \left(\tilde{x}_{2,i}' \tilde{x}_{2,i} \right)^{-1} \tilde{x}_{2,i}' (\tilde{y}_i - \tilde{x}_{1,i} \hat{\alpha}_k^{\text{oracle}}) \text{ for } i \in G_k^0,
\end{aligned}$$

where $M_{2,i} = I_T - \tilde{x}_{2,i} \left(\tilde{x}_{2,i}' \tilde{x}_{2,i} \right)^{-1} \tilde{x}_{2,i}'$. One can readily show that $\hat{\alpha}_k$ shares the same asymptotic bias and variance as $\hat{\alpha}_k^{\text{oracle}}$, and similarly, $\hat{\beta}_{2,i}$ shares the same asymptotic bias and variance as $\hat{\beta}_{2,i}^{\text{oracle}}$. In this case, we say that our C-Lasso estimators $\hat{\alpha}_k$ and $\hat{\beta}_{2,i}$ are asymptotically oracally efficient. As expected, $\hat{\alpha}_k$ may have asymptotic bias of order $O\left(T^{-1}\right)$ in the presence of endogeneity, but it converges to its true value at the usual $\sqrt{N_k}T$-rate after removing the bias term.

A close examination of the asymptotic bias of $\hat{\alpha}_k$ indicates that $\mathbb{B}_{k,NT}$ can be rewritten as the summation of two terms, $\mathbb{B}_{1k,NT}$ and $\mathbb{B}_{2k,NT}$. $\mathbb{B}_{1k,NT}$ appears even without the within-group transformation as in Phillips and Moon (1999); $\mathbb{B}_{2k,NT}$ is simply due to the time-demeaning operator. As mentioned above, we allow for both sources of endogeneity. When $\Sigma_{20,i} \neq 0$, we have contemporaneous correlation between the stationary regressor $x_{2,it}$ and the error term u_{it} in the cointegrating regression model. When $S_1 \sum_{r=0}^{\infty} \sum_{s=0}^{\infty} \psi_{i,s+r} \psi_{i,s}' S_0' \neq 0$ or $S_1 \psi_i \left(1\right) \psi_i \left(1\right)' S_0' \neq 0$, we allow the correlation of u_{it} with some leads or current values of $\varepsilon_{1,it}$. When both types of correlations vanish, $\mathbb{B}_{k,NT} = 0$, so that there is no endogeneity bias in this special case.

Note that we specify a selection matrix $\mathbb{S}_2$ in Theorem 4(ii) that is not needed if p_2 is fixed. When p_2 diverges to infinity, we cannot derive the asymptotic normality of $\hat{\beta}_{2,i}$ directly. Instead, we follow the literature on inferences with a diverging number of parameters (e.g., Fan and Peng, 2004; Lam and Fan, 2008; Lu and Su, 2015; Qian and Su, 2016a and 2016b) and prove the asymptotic normality for any arbitrary finite linear combinations of elements of $\hat{\beta}_{2,i}$.

Given the estimated groups, $\{\hat{G}_k, k = 1, ..., K\}$, we can obtain the post-Lasso estimators of α_k and $\beta_{2,i}$ as

$$\begin{aligned}
\hat{\alpha}_k^{\text{post}} &= \left(\sum_{i \in \hat{G}_k} \tilde{x}_{1,i}' M_{2,i} \tilde{x}_{1,i}\right)^{-1} \sum_{i \in \hat{G}_k} \tilde{x}_{1,i}' M_{2,i} \tilde{y}_i \text{ for } k = 1, ..., K, \\
\hat{\beta}_{2,i}^{\text{post}} &= \left(\tilde{x}_{2,i}' \tilde{x}_{2,i}\right)^{-1} \tilde{x}_{2,i}' (\tilde{y}_i - \tilde{x}_{1,i} \hat{\alpha}_k^{\text{post}}) \text{ for } i \in \hat{G}_k.
\end{aligned}$$

We show in the proof of Theorem 4 that the C-Lasso estimators $\hat{\alpha}_k$ and $\hat{\beta}_{2,i}$ are asymptotically equivalent to their post-Lasso versions $\hat{\alpha}_k^{\text{post}}$ and $\hat{\beta}_{2,i}^{\text{post}}$, respectively. The following theorem reports the limiting distributions of $\hat{\alpha}_k^{\text{post}}$ and $\hat{\beta}_{2,i}^{\text{post}}$.

Theorem 5 *Suppose that Assumptions A.1-A.3 hold. Then*

(i) $\sqrt{N_k}T(\hat{\alpha}_k^{post} - \alpha_k^0) - \mathbb{Q}_{(k)}^{-1} \mathbb{B}_{k,NT} \Rightarrow N(0, \mathbb{Q}_{(k)}^{-1} \mathbb{V}_{(k)} \mathbb{Q}_{(k)}^{-1})$ *for* $k = 1, ..., K$,

(ii) $\sqrt{T}\mathbb{S}_2(\hat{\beta}_{2,i}^{post} - \beta_{2,i}^*) \Rightarrow N(0, \mathbb{S}_2\mathbb{V}_{22,i}\mathbb{S}_2')$ *for* $i = 1, ..., N$,

where $\mathbb{Q}_{(k)}$, $\mathbb{B}_{k,NT}$, $\mathbb{V}_{(k)}$, *and* $\mathbb{V}_{22,i}$ *are as defined before Theorem 4 and* $\mathbb{S}_2$ *is as defined in Theorem 4.*

Given the asymptotic results in Theorems 4 and 5, one can make inference as if the true group membership is known. Despite the asymptotic equivalence of the C-Lasso estimators and their post-Lasso versions, it is well known that the post-Lasso estimators tend to have smaller finite sample bias in simulations and are thus recommended for practical uses. Despite this, in order to make inference on the long-run cointegrating relationship, we have to remove the bias. In principle, one can consider either the panel DOLS or fully modified OLS (FMOLS) method as in Kao and Chiang (2001) and Mark and Sul (2003) based on the estimated groups. The procedure is standard and thus omitted. Alternatively, we can consider the use of DOLS idea in the C-Lasso procedure, which yields the C-Lasso-based dynamic PLS (DPLS) estimation procedure. See the next subsection for details.

1.4.4 The case of dynamic PLS

In this subsection, we focus on the dynamic PLS estimation of the panel cointegration model with latent group structures. We show that the results in Theorems 4 and 5 continue to be valid with little modification.

For notational clarity, we now assume that $\{y_{it,} x_{1it}\}$ are generated by

$$\begin{cases} y_{it} = \mu_i + \beta_{1,i}' x_{1,it} + u_{it} \\ x_{1,it} = x_{1,it-1} + \varepsilon_{1,it} \end{cases}, \tag{9}$$

where μ_i, u_{it}, and $\varepsilon_{1,it}$ are defined as before, and $\beta_{1,i}$'s exhibit the latent structures in (3).

To consider the panel DPLS estimation method, we follow Saikkonen (1991) and Stock and Watson (1993) and make the following assumption.

Assumption A.4. (i) The process $\{u_{it}\}$ can be projected on to $\{\varepsilon_{1,it}\}$ as follows: $u_{it} = \sum_{j=-\infty}^{\infty} \gamma_{ij}\varepsilon_{1,i,t+j} + v_{it}$, where $\sum_{j=-\infty}^{\infty} \|\gamma_{ij}\| < \infty$, v_{it} is error term with mean zero and finite $2q^{\text{th}}$ moment where $q > 4$, and v_{it} and $\varepsilon_{1,it}$ are uncorrelated for all lags and leads.

(ii) As $(N, T) \to \infty$, there exists $a > 1/2$ such that $T^a \sum_{|j|>\bar{p}_2} ||\gamma_{ij}|| \to 0$, $N^{1/2}T^{1/2-a} \to 0$, and $N^{1/2}\bar{p}_2 T^{-a} \to 0$.

Assumption A.4(i) ensures that $E(\varepsilon_{1,it}v_{it+k}) = 0$ for $k = 0, \pm1, \pm2, \ldots$ and Assumption A.4(ii) ensures that the values of $\varepsilon_{1,it}$ in the very remote past and future have only negligible impact on u_{it}. Therefore we can truncate the leads and lags and run the following DOLS regression model

$$y_{it} = \mu_i + \beta_{1,i}' x_{1,it} + \sum_{j=-\bar{p}_2}^{\bar{p}_2} \gamma_{ij}' \Delta x_{1,i,t+j} + v_{it}^{\dagger}, \tag{10}$$

where $v_{it}^{\dagger} = v_{it}^{a} + v_{it}$, and $v_{it}^{a} = \sum_{|j|>\bar{p}_2} \gamma_{ij}' \Delta x_{1,i,t+j}$ signifies the approximation/truncation error. Let $x_{2,it}$ denote a collection of the lags and leads of $\Delta x_{1,it}$: $x_{2,it} = (\Delta x_{1,i,t-\bar{p}_2}', ..., \Delta x_{1,i,t+\bar{p}_2}')'$. Let $\beta_{2,i} = (\gamma_{i,-\bar{p}_2}', ..., \gamma_{i\bar{p}_2}')'$ and $p_2 = (2\bar{p}_2 + 1)p_1$. After the within-group transformation, we have the following model

$$\begin{aligned} \tilde{y}_{it} &= \beta_{1,i}' \tilde{x}_{1,it} + \sum_{j=-\bar{p}_2}^{\bar{p}_2} \gamma_{ij}' \widetilde{\Delta x}_{1,i,t+j} + \tilde{v}_{it}^{\dagger} \\ &= \beta_{1,i}' \tilde{x}_{1,it} + \beta_{2,i}' \tilde{x}_{2,it} + \tilde{v}_{it}^{\dagger}, \end{aligned} \tag{11}$$

where $\tilde{v}_{it}^{\dagger} = v_{it}^{\dagger} - \bar{v}_{i}^{\dagger}$, $\bar{v}_{i}^{\dagger} = \frac{1}{T-2\bar{p}_2} \sum_{t=\bar{p}_2+1}^{T-\bar{p}_2} v_{it}^{\dagger}$, and $\tilde{y}_{it}$ and $\tilde{x}_{2,i}$ are analogously defined.

As before, we can continue to consider the C-Lasso-based PLS regression and obtain the Lasso estimators of $\beta_{1,i}$, $\beta_{2,i}$, and α_k. We denote these estimators as $\hat{\beta}_{1,i}^{D}$, $\hat{\beta}_{2,i}^{D}$, and $\hat{\alpha}_{k}^{D}$, where D abbreviates dynamic PLS (DPLS). Let $\hat{G}_k$ denote the estimated group as before. The corresponding post-Lasso estimators of α_k and $\beta_{2,i}$ take the form

$$\begin{aligned} \hat{\alpha}_{k}^{D,\,\text{post}} &= \left(\sum_{i \in \hat{G}_k} \tilde{x}_{1,i}' M_{2,i} \tilde{x}_{1,i} \right)^{-1} \sum_{i \in \hat{G}_k} \tilde{x}_{1,i}' M_{2,i} \tilde{y}_i \text{ for } k = 1, ..., K, \\ \hat{\beta}_{2,i}^{D,\,\text{post}} &= \left(\tilde{x}_{2,i}' \tilde{x}_{2,i} \right)^{-1} \tilde{x}_{2,i}' \left(\tilde{y}_i - \tilde{x}_{1,i} \hat{\alpha}_{k}^{D,\,\text{post}} \right) \text{ for } i \in \hat{G}_k, \end{aligned}$$

where $\tilde{x}_{1,i} = \left(\tilde{x}_{1,i,\bar{p}_2+1}, ..., \tilde{x}_{1,i,T-\bar{p}_2} \right)'$, $\tilde{y}_i$ and $\tilde{x}_{2,i}$ are analogously defined, and $M_{2,i} = I_{T-2\bar{p}_2} - \tilde{x}_{2,i} \left(\tilde{x}_{2,i}' \tilde{x}_{2,i} \right)^{-1} \tilde{x}_{2,i}'$.

The following theorem shows the asymptotic properties of $\hat{\alpha}_{k}^{D,\,\text{post}}$ and $\hat{\beta}_{2,i}^{D,\,\text{post}}$ where expressions for both $\mathbb{V}_{(k)}$ and $\mathbb{V}_{22,i}$ are greatly simplified.

Theorem 6 *Suppose that Assumptions A.1, A.2(i)-(iii) and A.3-A.4 hold. Suppose that there exists a constant* $\underline{c}_{00}$ *such that* $\min_{1\leqslant i\leqslant N}\Sigma_{00,i}\geqslant\underline{c}_{00}>0$. *Then*

(i)$\sqrt{N_k}T(\hat{\alpha}_k^{D,post}-\alpha_k^0)\Rightarrow N(0,\mathbb{Q}_{(k)}^{-1}\mathbb{V}_{(k)}^{+}\mathbb{Q}_{(k)}^{-1})$ *for* $k=1,...,K$,

(ii) $\sqrt{T}\mathbb{S}_2(\hat{\beta}_{2,i}^{D,post}-\beta_{2,i}^0)\Rightarrow N\left(0,\mathbb{S}_2\mathbb{V}_{22,i}\mathbb{S}_2'\right)$ *for* $i=1,...,N$,

where $\mathbb{Q}_{(k)}\equiv\lim_{N_k\to\infty}\frac{1}{6N_k}\sum_{i\in G_k^0}\Omega_{11,i}$, $\mathbb{V}_{(k)}^{+}\equiv\lim_{N_k\to\infty}\frac{1}{N_k}\sum_{i\in G_k^0}\frac{1}{6}\Omega_{00,i}^{+}\Omega_{11,i}$, $\Omega_{00,i}^{+}=\Omega_{00,i}-\Omega_{01,i}\Omega_{11,i}^{-1}\Omega_{10,i}$ *and* $\mathbb{V}_{22,i}=\Sigma_{22,i}^{-1}V_{22,i}\Sigma_{22,i}^{-1}$ *with* $V_{22,i}=\lim_{T\to\infty}Var(T^{-1/2}\sum_{t=1}^{T}x_{2,it}u_{it})$.

Even though we have not stated in the above theorem, $\hat{\alpha}_k^D$ and $\hat{\beta}_{2,i}^D$ are asymptotically equivalent to $\hat{\alpha}_k^{D,\text{post}}$ and $\hat{\beta}_{2,i}^{D,\text{post}}$, respectively. Thus both C-Lasso-based DPLS estimators and their post-Lasso versions have asymptotic normal distributions and are asymptotically oracally efficient. One can readily construct the usual t-statistics and F-statistics to make inference. For example, to make inference on the group-specific long-run cointegrating relationship, we can estimate $\mathbb{Q}_{(k)}$ and $\mathbb{V}_{(k)}$ respectively by[3]

$$\hat{\mathbb{Q}}_{(k)}=\frac{1}{\hat{N}_kT^2}\sum_{i\in\hat{G}_k}\tilde{x}_{1,i}'M_{2,i}\tilde{x}_{1,i}\text{ and }\hat{\mathbb{V}}_{(k)}^{+}\equiv\frac{1}{\hat{N}_k}\sum_{i\in\hat{G}_k}\frac{1}{6}\hat{\Omega}_{00,i}^{+}\hat{\Omega}_{11,i},$$

where $\hat{\Omega}_{00,i}$ and $\hat{\Omega}_{11,i}$ denote the HAC estimator of the long-run variance-covariance components $\Omega_{00,i}$ and $\Omega_{11,i}$ in Ω_i. In practice, we recommend the use of $\hat{\alpha}_k^{D,\text{post}}$ and $\hat{\beta}_{2,i}^{D,\text{post}}$ because the post-Lasso estimators typically outperform the C-lasso ones.

1.4.5 The case of incidental time trends

Our cointegrated panel model can be extended to models with both individual fixed effects and incidental time trends,

$$y_{it}=\mu_i+\rho_it+\beta_{1,i}x_{1,it}+\beta_{2,i}x_{2,it}+u_{it},\quad i=1,...,N\text{ and }t=1,...,T,\tag{12}$$

where the incidental time trend ρ_it is introduced, the other variables are defined as above. In the original case, we can eliminate the individual fixed effects μ_i in (2.1) via the within-group transformation,

$$\tilde{y}_{it}=\beta_{1,i}'\tilde{x}_{1,it}+\beta_{2,i}\tilde{x}_{2,it}+\tilde{u}_{it},$$

[3] Alternatively, we can consistently estimate $\mathbb{Q}_{(k)}$ by $\tilde{\mathbb{Q}}_{(k)}=\frac{1}{\hat{N}_kT^2}\sum_{i\in\hat{G}_k}\tilde{x}_{1,i}'\tilde{x}_{1,i}$.

where $\tilde{y}_{it} = y_{it} - \frac{1}{T}\sum_{t=1}^{T} y_{it}$, $\tilde{x}_{it}$ and $\tilde{u}_{it}$ are analogously defined. The asymptotic properties are built on the demeaned data. When incidental time trends are allowed, we consider the detrended data on above model (4.4) to eliminate both individual fixed effects and incidental time trends. Then we have

$$\dot{y}_{it} = \beta_{1,i}'\dot{x}_{1,it} + \beta_{2,i}\dot{x}_{2,it} + \dot{u}_{it}, \tag{13}$$

where $\dot{y}_{it} = y_{it} - \sum_{s=1}^{T} y_{is}g_s' \left(\sum_{s=1}^{T} g_s g_s'\right)^{-1} g_t$, $g_t = (1,t)'$, and $\dot{x}_{1,it}$, $\dot{x}_{2,it}$ and $\dot{u}_{it}$ are analogously defined. Since Phillips and Hansen (1990) and Hansen (1992) (p.91) have shown that the estimation of $\beta_{1,i}$ is invariant to the presence of incidental time trends. Thus we can simply apply the same estimation procedure in Section 2.2 with dotted variables. And the asymptotic theorem can be modified with the detrended data. Given above model (4.4), we note that the incidental time trends come from random walk with drift, such that

$$x_{1,it} = \alpha_i + x_{1,it-1} + \varepsilon_{1,it} = \alpha_i t + \sum_{s=1}^{t} \varepsilon_{1,is} = x_{1,i0} + \alpha_i t + x_{i,t}^0$$

where $x_{1,it}^0 = \sum_{s=1}^{t} \varepsilon_{1,is}$ is purely random walk process. Define $\kappa_T = diag(1, T^{-1})$ and $g(r) = (1,r)'$. Let $t = [Tr]$, then as $T \to \infty$, $\kappa_T g_t \to g(r)$ uniformly in $r \in [0,1]$. By functional central limit theorem and continuous mapping theorem, we have

$$\begin{aligned}
\frac{1}{\sqrt{T}}\dot{x}_{1,it} =& \frac{1}{\sqrt{T}}\left(x_{1,it} - \sum_{t=1}^{T} x_{1,it}g_t' \left(\sum_{s=1}^{T} g_s g_s'\right)^{-1} g_t\right) \\
=& \frac{1}{\sqrt{T}}\left(x_{1,it}^0 - \sum_{s=1}^{T} x_{1,it}^0 g_t' \left(\sum_{s=1}^{T} g_s g_s'\right)^{-1} g_t\right) \\
=& \frac{x_{1,it}^0}{\sqrt{T}} - \frac{1}{T}\sum_{t=1}^{T} \frac{x_{1,it}^0}{\sqrt{T}}\kappa_T g_t \left(\frac{1}{T}\sum_{t=1}^{T} \kappa_T g_t g_t' \kappa_T\right)^{-1} \kappa_T g_T \\
\Rightarrow& B_{1i}(r) - \int B_{\varepsilon_{1i}}(r) g(r)' dr \left(\int g(r)g(r)'\right)^{-1} g(r) := B_{1i}^{\tau}(r)
\end{aligned}$$

where $B_{1i}^{\tau}(r)$ is a randomly scaled detrended Brownian motion and are independent across individual i. Following the analysis in Section 4, we can show Theorem 4.1-Theorem 4.3 hold under detrended data. And the limiting distribution in Theorem 4.4-4.6 can be modified under different asymptotic moments on $\mathbb{Q}_k$ and $\mathbb{V}_k$ upon detrended Brownian motion.

1.5 Monte Carlo Simulation

In this section, we evaluate the finite-sample performance of both PLS-based and DPLS-based C-lasso estimates and their post-Lasso versions.

1.5.1 Data generating processes

We consider three data generating processes (DGPs). The observations in DGP1-DGP3 are drawn from three groups with $N_1 : N_2 : N_3 = 0.3 : 0.4 : 0.3$. And that in DGP4-DGP5 are drawn from the same group structures in empirical applications, where DGP4 is drawn from two groups with $N_1 : N_2 = 0.9 : 0.1$ and DGP5 is drawn from three groups with $N_1 : N_2 : N_3 = 0.5 : 0.3 : 0.2$ There are four combinations of the sample sizes with $N = 50$, 100 and $T = 40$, 80.

DGP 1 (Strictly Exogenous Nonstationary Regressors) The observations (y_{it}, x_{it}') are generated from the following cointegrated panel,

$$\begin{cases} y_{it} = \mu_i + \beta_i^{0\prime} x_{it} + u_{it} = \mu_i + \beta_{1,i}^{0\prime} x_{1,it} + u_{it} \\ x_{1,it} = x_{1,it-1} + \varepsilon_{1,it} \end{cases}$$

where $\mu_i \sim$ IID $N(0,1)$, $x_{it} = x_{1,it}$ is a 2×1 vector, $\varepsilon_{it} = (u_{it},\, \varepsilon_{1,it}')'$ follows a multivariate standard normal distribution, and $\beta_i^0 = \beta_{1,i}^0$ exhibits the group-structure in (3) for $K = 3$ and

$$(\alpha_1^0, \alpha_2^0, \alpha_3^0) = \left(\begin{pmatrix} 0.4 \\ 1.6 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1.6 \\ 0.4 \end{pmatrix} \right).$$

DGP 2 (Weakly Dependent Nonstationary Regressors) The observations (y_{it}, x_{it}') are generated via (5.1). Note that $S_0 \varepsilon_{it} = u_{it}, S_1 \varepsilon_{it} = \varepsilon_{1,it}$,and $S_\mu \varepsilon_{i1} = \mu_i$,where $S_0 = (1,0,0,0)\,, S_1 = (0,1,1,0)$, and $S_\mu = (0,0,0,1)$. ε_{it} are generated from a linear process: $\varepsilon_{it} = \sum_{j=1}^{\infty} \psi_{ij} e_{i,t-j}$, where e_{it} are IID $N\left(0, I_4\right), \psi_{ij} = 0.5 \cdot j^{-3.5} \cdot \Omega_1^{1/2}$, and $\Omega_1^{1/2}$ is the symmetric square root of $\Omega_1 \equiv \begin{pmatrix} 1 & 0.3 & 0.2 & 0 \\ 0.3 & 1 & 0.2 & 0.2 \\ 0.2 & 0.3 & 1 & 0.2 \\ 0 & 0.2 & 0.2 & 1 \end{pmatrix}$.

DGP 3 (Weakly Dependent Nonstationary and Stationary Regressors) The observations (y_{it}, x'_{it}) are generated from the following cointegrated panel,

$$\begin{cases} y_{it} = \mu_i + \beta_i^{0\prime} x_{it} + u_{it} = \mu_i + \beta_{1,i}^{0\prime} x_{1,it} + \beta_{2i}^{0} x_{2,it} + u_{it} \\ x_{1,it} = x_{1,it-1} + \varepsilon_{1,it} \end{cases}$$

where $\mu_i \sim$ IID $N(0,1)$, $x_{1,it}$ is a 2×1 vector, $\beta_{1,i}^0$ exhibits the group structures and preserves the setting in DGP1 and $x_{2,it} = \varepsilon_{2,it}$ only contains one stationary regressor. Note that $S_0 \varepsilon_{it} = u_{it}, S_1 \varepsilon_{it} = \varepsilon_{1,it}, S_2 \varepsilon_{it} = \varepsilon_{2,it}$,and $S_\mu \varepsilon_{i1} = \mu_i$,where $S_0 = (1,0,0,0,0)\,, S_1 = (0,1,1,0,0)$, $S_2 = (0,0,0,1,0)$ and $S_\mu = (0,0,0,0,1)\,.$ ε_{it} are generated from a linear process: $\varepsilon_{it} = \sum_{j=1}^{\infty} \psi_{ij} e_{i,t-j}$, where e_{it} are IID $N\left(0, I_6\right)$, $\psi_{ij} = 0.5 \cdot j^{-3.5} \cdot \Omega_2^{1/2}$, and $\Omega_2^{1/2}$ is the symmetric square root of $\Omega_2 \equiv$ $\begin{pmatrix} 1 & 0.3 & 0.2 & 0.2 & 0 \\ 0.3 & 1 & 0.2 & 0 & 0.2 \\ 0.2 & 0.2 & 1 & 0 & 0.2 \\ 0.2 & 0 & 0 & 1 & 0.2 \\ 0 & 0.2 & 0.2 & 0.2 & 1 \end{pmatrix}$. The coefficients of the stationary regressors are heterogeneous across all i such that $\beta_{2,i} \sim$ IID $N(c, 1)$ with $c = 0.5$.

DGP 4 (Mimic Empirical Applications Table Panel A) The observations (y_{it}, x'_{it}) are generated via (5.1), where $x_{it} = x_{1,it}$ contains one nonstationary regressor. Note that $S_0 \varepsilon_{it} = u_{it}, S_1 \varepsilon_{it} = \varepsilon_{1,it}$,and $S_\mu \varepsilon_{i1} = \mu_i$,where $S_0 = (1,0,0)\,, S_1 = (0,1,0)$, and $S_\mu = (0,0,1)$. ε_{it} are generated from a linear process: $\varepsilon_{it} = \sum_{j=1}^{\infty} \psi_{ij} e_{i,t-j}$, where e_{it} are IID $N\left(0, I_3\right)$, $\psi_{ij} = 0.5 \cdot j^{-3.5} \cdot \Omega_1^{1/2}$, and $\Omega_1^{1/2}$ is the symmetric square root of $\Omega_1 \equiv \begin{pmatrix} 1 & 0.3 & 0 \\ 0.3 & 1 & 0.2 \\ 0 & 0.2 & 1 \end{pmatrix}$. And $\beta_i^0 = \beta_{1,i}^0$ exhibits the group-structure in (3) for $K = 2$ and has the similar estimates in Table Panel A $(\alpha_1^0, \alpha_2^0) = (0.9, -0.7)\,.$

DGP 5 (Mimic Empirical Applications Table Panel B)The observations (y_{it}, x'_{it}) are generated via (5.1). The innovation processes are generated from via the same processes in DGP4 and $\beta_i^0 = \beta_{1,i}^0$ exhibits the group-structure in (3) for $K = 3$ and has the similar estimates in Table Panel B $(\alpha_1^0, \alpha_2^0, \alpha_3^0) = (0.9, 0.2, -0.6)\,.$

In all cases, the number of replications is 10000.

1.5.2 Classification and estimation

We assume that the number of groups is known and examine the performance of classification and estimation. The results from Appendix show that the number of groups can be nearly perfectly selected by the information criterion.

For classification, we consider the PLS-based C-Lasso classification results for DGPs 1,2,4 and 5, and both the PLS- and DPLS-based C-Lasso classification results for DGP 3. For the DPLS-based classification in DGP 3, we introduce the lags and leads of $x_{1,it}$ in our penalized estimation by setting $\bar{p}_2$ to be approximately $[T^{1/4}]$. More precisely, we set $\bar{p}_2 = 2$ and 3 for $T = 40$ and 80, respectively. We follow Section 4.2 and define two types of average classification errors: $\bar{P}(\hat{E}) = \frac{1}{N}\sum_{i=1}^{N} \hat{P}(\cup_{k=1}^{K} \hat{E}_{kNT,i})$ and $\bar{P}(\hat{F}) = \frac{1}{N}\sum_{i=1}^{N} \hat{P}(\cup_{k=1}^{K} \hat{F}_{kNT,i})$ where $\hat{P}$ is the empirical mean over 500 replications. Table 1 reports the classification errors for a variety of choices of the tuning parameter λ or equivalently c_λ. We summarize some important findings from Table 1. First, both types of classification errors vary over c_λ. The smaller value of c_λ, the smaller percentage of the classification error. This means that a larger value of penalty term tends to lead to a higher rate of misclassification. Second, as T increases, the percentage of classification errors drops significantly. In fact, when T is 80, we have less than 1% individuals misclassified across the board in all cases. Third, for DGP 3, the performance of the DPLS-based C-Lasso classification is not as good as that of the PLS-based C-Lasso estimation. Despite this fact, the former performance becomes acceptable when $T = 80$ for all choices of c_λ.

For the estimation, we consider both the C-Lasso estimates and its post-Lasso versions. Specifically, for all DGPs we consider the PLS-based C-Lasso estimates, the OLS-based post-Lasso estimates, the DOLS-based post-Lasso estimates, and the oracle estimates which are obtained by using the true group structure. For DGP 3, we also consider the DPLS-based C-Lasso estimates, its post-Lasso version, and the oracle estimates. For all DOLS-based estimates, we set $\bar{p}_2$ as above. We report the bias, root-mean-square error (RMSE), and coverage probability of the two-sided nominal 95% confidence interval for the estimate $\hat{\beta}_{1,i}(1)$ of the first parameter $\beta_{1,i}(1)$ in $\beta_{1,i}$ for each DGP in Tables 2-4, where all criteria are averaged over different groups and across 10,000 replications. For example, we calculate the RMSE of $\hat{\beta}_{1,i}(1)$'s as $\frac{1}{N}\sum_{k=1}^{K_0} N_k$RMSE($\hat{\alpha}_{k,1}$) with $\hat{\alpha}_{k,1}$ denot-

Table 1: Empirical classification errors in percentage

	c_λ		0.1		0.2	
	N	T	$\bar{P}(\hat{E})$	$\bar{P}(\hat{F})$	$\bar{P}(\hat{E})$	$\bar{P}(\hat{F})$
DGP1	50	40	0.212	0.221	0.515	0.410
	50	80	0.000	0.000	0.001	0.001
	100	40	0.218	0.226	0.475	0.384
	100	80	0.000	0.000	0.001	0.001
DGP2	50	40	0.483	0.506	0.875	0.728
	50	80	0.000	0.000	0.003	0.002
	100	40	0.500	0.518	0.796	0.667
	100	80	0.000	0.000	0.004	0.003
DGP3	50	40	0.535	0.563	0.799	0.684
(PLS)	50	80	0.001	0.001	0.005	0.004
	100	40	0.532	0.562	0.745	0.640
	100	80	0.000	0.000	0.003	0.002
DGP3	50	40	6.337	5.630	12.255	9.700
(DPLS)	50	80	0.038	0.031	0.186	0.141
	100	40	6.027	5.432	11.453	9.138
	100	80	0.033	0.026	0.157	0.120
DGP4	50	40	1.234	0.834	0.821	0.543
	50	80	0.014	0.008	0.004	0.002
	100	40	1.225	0.823	0.801	0.527
	100	80	0.011	0.007	0.004	0.003
DGP5	50	40	0.000	0.000	0.040	0.004
	50	80	0.000	0.000	0.000	0.000
	100	40	0.000	0.000	0.032	0.004
	100	80	0.000	0.000	0.001	0.000

ing the first element in $\hat{\alpha}_k$ for one replication and then average them across all replications for each case.

Table 2 reports the estimation results for DGPs 1-2, Table 3 reports the estimation results for DGP 3 based on both the PLS and DPLS methods, and Table 4 reports the estimation results for DGPs 4-5. These tables reveal some general patterns. First, the bias and RMSE of the C-Lasso estimates and their post-Lasso versions always decrease as either N or T increases, and they decrease faster when T increases than when N increases. This is as expected due to faster convergence rate of the estimates along the time dimension than along the cross-sectional dimension. Second, when there is no endogeneity issue in DGP1, the finite sample performance of the post-Lasso (OLS) estimates is close to that of the oracle ones and dominates that of the DOLS-based post-Lasso estimates. This indicates that the DOLS may hurt in finite samples when there is no endogeneity issue in the model. Third, when endogeneity is present in DGPs 2-5, the post-Lasso (DOLS) estimators are distinctly superior to the C-Lasso and post-Lasso (OLS) ones for all cases and their performance is very close to that of the oracle one. Fourth, for DGP 3 the DPLS-based C-Lasso estimates outperform the PLS-based C-Lasso estimates to a great margin, but the post-Lasso estimates are not quite distinct from each other in terms of bias and RMSE. Fifth, the coverage probabilities of the post-Lasso (DOLS-based) estimates are generally quite close to the specified level (95%) in all cases (except for DGP 1 in the absence of endogeneity). For DGP3 the coverage probabilities of DPLS-based C-Lasso estimates are closer to the specified level compared to those of the PLS-based C-Lasso estimates. These two facts suggest that DOLS bias-correction yields good coverage probability when endogeneity is present. Lastly, in general the post-Lasso DOLS estimates outperform the C-Lasso estimates (except for DGP 1 in the absence of endogeneity) and thus are recommended for practical use.

1.6 Application: Testing the PPP hypothesis

In this section we apply our method to reinvestigate the purchasing power parity (PPP) hypothesis in international economics.

Table 2: Correct specification, RMSEs, Biases and Coverage probabilities for various estimates

	c_λ	0.1			0.2		
(N,T)		RMSE	Bias	Coverage %	RMSE	Bias	Coverage %
DGP1	(PLS)						
(50,40)	C-lasso	0.0180	0.0001	92.11	0.0174	0.0001	93.05
	Post-lasso$^{(OLS)}$	0.0173	0.0000	93.28	0.0173	0.0001	93.24
	Post-lasso$^{(DOLS)}$	0.0226	0.0000	84.60	0.0226	0.0000	84.58
	Oracle	0.0172	0.0001	93.30	0.0172	0.0001	93.30
(50,80)	C-lasso	0.0083	0.0001	93.09	0.0082	0.0001	93.51
	Post-lasso$^{(OLS)}$	0.0082	0.0001	93.55	0.0082	0.0001	93.55
	Post-lasso$^{(DOLS)}$	0.0091	0.0001	90.27	0.0091	0.0001	90.27
	Oracle	0.0082	0.0001	93.55	0.0082	0.0001	93.55
(100,40)	C-lasso	0.0126	0.0001	92.70	0.0122	0.0001	93.75
	Post-lasso$^{(OLS)}$	0.0120	0.0001	94.04	0.0121	0.0001	94.01
	Post-lasso$^{(DOLS)}$	0.0155	0.0001	85.75	0.0155	0.0001	85.75
	Oracle	0.0120	0.0001	94.08	0.0120	0.0001	94.08
(100,80)	C-lasso	0.0058	0.0001	93.82	0.0056	0.0000	94.42
	Post-lasso$^{(OLS)}$	0.0056	0.0000	94.42	0.0056	0.0000	94.42
	Post-lasso$^{(DOLS)}$	0.0063	0.0001	91.57	0.0063	0.0001	91.57
	Oracle	0.0056	0.0000	94.42	0.0056	0.0000	94.42
DGP2	(PLS)						
(50,40)	C-lasso	0.0312	0.0247	81.20	0.0287	0.0223	85.67
	Post-lasso$^{(OLS)}$	0.0276	0.0211	87.49	0.0276	0.0211	87.47
	Post-lasso$^{(DOLS)}$	0.0215	0.0001	94.68	0.0215	0.0001	94.72
	Oracle	0.0215	0.0000	94.73	0.0215	0.0000	94.73
(50,80)	C-lasso	0.0147	0.0117	71.52	0.0138	0.0107	75.74
	Post-lasso$^{(OLS)}$	0.0135	0.0105	76.98	0.0135	0.0105	76.98
	Post-lasso$^{(DOLS)}$	0.0088	0.0000	94.15	0.0088	0.0000	94.15
	Oracle	0.0088	0.0000	94.15	0.0088	0.0000	94.15
(100,40)	C-lasso	0.0277	0.0243	65.88	0.0252	0.0218	73.51
	Post-lasso$^{(OLS)}$	0.0240	0.0206	77.47	0.0240	0.0205	77.62
	Post-lasso$^{(DOLS)}$	0.0148	0.0002	95.77	0.0148	0.0001	95.82
	Oracle	0.0148	0.0001	95.85	0.0148	0.0001	95.85
(100,80)	C-lasso	0.0131	0.0115	52.70	0.0120	0.0105	59.63
	Post-lasso$^{(OLS)}$	0.0117	0.0101	62.01	0.0117	0.0101	62.00
	Post-lasso$^{(DOLS)}$	0.0060	0.0001	95.27	0.0060	0.0001	95.26
	Oracle	0.0060	0.0001	95.27	0.0060	0.0001	95.27

Table 3: Correct specification, RMSEs, Biases and Coverage probabilities for various estimates

	c_λ		0.1			0.2	
(N,T)		RMSE	Bias	Coverage %	RMSE	Bias	Coverage %
DGP3	(PLS)						
(50,40)	C-lasso	0.0304	0.0234	83.31	0.0275	0.0206	88.14
	Post-lasso$^{\text{(OLS)}}$	0.0318	0.0193	81.70	0.0318	0.0193	81.74
	Post-lasso$^{\text{(DOLS)}}$	0.0215	0.0000	94.95	0.0215	0.0000	94.90
	Oracle	0.0214	0.0000	95.02	0.0214	0.0000	95.02
(50,80)	C-lasso	0.0134	0.0103	76.45	0.0126	0.0094	80.41
	Post-lasso$^{\text{(OLS)}}$	0.0156	0.0091	71.00	0.0156	0.0091	71.00
	Post-lasso$^{\text{(DOLS)}}$	0.0086	0.0000	94.28	0.0086	0.0000	94.29
	Oracle	0.0086	0.0000	94.29	0.0086	0.0000	94.29
(100,40)	C-lasso	0.0265	0.0228	70.64	0.0237	0.0200	78.75
	Post-lasso$^{\text{(OLS)}}$	0.0254	0.0184	75.26	0.0254	0.0184	75.25
	Post-lasso$^{\text{(DOLS)}}$	0.0148	0.0000	95.99	0.0148	-0.0001	96.02
	Oracle	0.0147	-0.0001	96.11	0.0147	-0.0001	96.11
(100,80)	C-lasso	0.0118	0.0101	60.45	0.0108	0.0091	67.20
	Post-lasso$^{\text{(OLS)}}$	0.0121	0.0088	63.49	0.0121	0.0088	63.49
	Post-lasso$^{\text{(DOLS)}}$	0.0060	0.0000	95.01	0.0060	0.0000	95.01
	Oracle	0.0060	0.0000	95.01	0.0060	0.0000	95.01
DGP3	(DPLS)						
(50,40)	C-lasso	0.0234	0.0001	93.12	0.0232	0.0000	93.31
	Post-lasso	0.0222	0.0000	94.39	0.0227	0.0000	93.91
	Oracle	0.0214	0.0000	95.02	0.0214	0.0000	95.02
(50,80)	C-lasso	0.0088	0.0000	93.63	0.0087	0.0000	94.24
	Post-lasso	0.0086	0.0000	94.29	0.0086	0.0000	94.28
	Oracle	0.0086	0.0000	94.29	0.0086	0.0000	94.29
(100,40)	C-lasso	0.0166	-0.0002	94.28	0.0162	0.0000	94.67
	Post-lasso	0.0156	-0.0005	95.53	0.0157	-0.0001	95.24
	Oracle	0.0150	-0.0005	96.11	0.0150	-0.0005	96.11
(100,80)	C-lasso	0.0061	0.0000	94.49	0.0060	0.0000	95.11
	Post-lasso	0.0059	0.0000	95.16	0.0060	0.0000	95.11
	Oracle	0.0059	0.0000	95.16	0.0059	0.0000	95.16

Table 4: Correct specification, RMSEs, Biases and Coverage probabilities for various estimates

	c_λ		0.1			0.2	
(N,T)		RMSE	Bias	Coverage %	RMSE	Bias	Coverage %
DGP4	(PLS)						
(50,40)	C-lasso	0.0313	0.0256	68.18	0.0290	0.0233	73.88
	Post-lasso$^{(\mathrm{OLS})}$	0.0282	0.0221	76.03	0.0285	0.0226	76.03
	Post-lasso$^{(\mathrm{DOLS})}$	0.0191	-0.0006	93.24	0.0188	-0.0001	93.57
	Oracle	0.0188	0.0001	93.70	0.0188	0.0001	93.70
(50,80)	C-lasso	0.0148	0.0122	63.67	0.0140	0.0114	68.02
	Post-lasso$^{(\mathrm{OLS})}$	0.0139	0.0112	68.76	0.0139	0.0112	68.76
	Post-lasso$^{(\mathrm{DOLS})}$	0.0081	0.0000	94.05	0.0081	0.0000	94.06
	Oracle	0.0081	0.0000	94.06	0.0081	0.0000	94.06
(100,40)	C-lasso	0.0283	0.0253	45.97	0.0259	0.0229	53.83
	Post-lasso$^{(\mathrm{OLS})}$	0.0248	0.0216	58.16	0.0252	0.0221	58.16
	Post-lasso$^{(\mathrm{DOLS})}$	0.0132	-0.0007	93.84	0.0130	-0.0002	94.22
	Oracle	0.0130	0.0000	94.32	0.0130	0.0000	94.32
(100,80)	C-lasso	0.0135	0.0121	40.06	0.0126	0.0113	46.00
	Post-lasso$^{(\mathrm{OLS})}$	0.0124	0.0110	47.67	0.0124	0.0110	47.67
	Post-lasso$^{(\mathrm{DOLS})}$	0.0057	0.0000	94.48	0.0057	0.0000	94.49
	Oracle	0.0057	0.0000	94.49	0.0057	0.0000	94.49
DGP5	(PLS)						
(50,40)	C-lasso	0.0264	0.0228	51.59	0.0263	0.0226	52.22
	Post-lasso$^{(\mathrm{OLS})}$	0.0263	0.0226	52.19	0.0263	0.0226	52.21
	Post-lasso$^{(\mathrm{DOLS})}$	0.0139	0.0001	94.18	0.0139	0.0001	94.18
	Oracle	0.0139	0.0001	94.18	0.0139	0.0001	94.18
(50,80)	C-lasso	0.0128	0.0111	44.66	0.0128	0.0110	44.90
	Post-lasso$^{(\mathrm{OLS})}$	0.0128	0.0110	44.89	0.0128	0.0110	44.89
	Post-lasso$^{(\mathrm{DOLS})}$	0.0061	-0.0001	94.31	0.0061	-0.0001	94.31
	Oracle	0.0061	-0.0001	94.31	0.0061	-0.0001	94.31
(100,40)	C-lasso	0.0244	0.0224	23.77	0.0242	0.0223	24.31
	Post-lasso$^{(\mathrm{OLS})}$	0.0243	0.0223	24.27	0.0243	0.0223	24.27
	Post-lasso$^{(\mathrm{DOLS})}$	0.0097	0.0000	94.31	0.0097	0.0000	94.31
	Oracle	0.0097	0.0000	94.31	0.0097	0.0000	94.31
(100,80)	C-lasso	0.0119	0.0110	18.03	0.0119	0.0109	18.40
	Post-lasso$^{(\mathrm{OLS})}$	0.0119	0.0109	18.40	0.0119	0.0109	18.40
	Post-lasso$^{(\mathrm{DOLS})}$	0.0043	0.0000	94.45	0.0043	0.0000	94.45
	Oracle	0.0043	0.0000	94.45	0.0043	0.0000	94.45

1.6.1 PPP hypothesis

PPP assumes that in the absence of transaction costs and trade barriers, a basket of identical goods will have the same price in different markets when the prices are expressed in the same currency. Unlike the law of one price for one particular good, PPP is built on a “basket of goods”, indicating that nominal exchange rate is adjusted by general price index for the purpose of international comparison. The long-run equilibrium of PPP was broadly accepted in the post-war period before the breakdown of the Bretton Woods system in the early 1970s. In the post-Bretton Woods period, most applied work fails to support the validity of the long-run PPP; see, e.g., Frenkel (1981) and Adler and Lehmann (1983). Some researchers attribute this to the low power of time series unit root tests when T is short and advocate the use of panel unit tests. Indeed, some panel unit root testing results favor the PPP hypothesis in the post-Bretton Woods period; see, e.g., Oh (1996) and Papell (1997). Even so, the empirical findings are still mixed. There remain two main issues in testing the validity of PPP hypothesis by using panel data. One is the sample selection issue and the other is the unobserved heterogeneity issue. Our cointegrated panel model with latent group structures can provide a data-driven method to address these two issues simultaneously and is expected to offer some new insight into the PPP hypothesis.

1.6.2 Model and data

The PPP hypothesis has two versions: strong and weak. We first consider the strong PPP hypothesis. Denote the domestic price index as P_{it}, the corresponding foreign price index as P_{jt}, and E_{it} as nominal exchange rate. If the strong PPP hypothesis holds, we have the equation $E_{it} = \frac{P_{it}}{P_{jt}}$ where we have suppress the dependence of E_{it} on j which is typically fixed in panel studies. In the logarithmic form, we have $e_{it} = p_{it} - p_{jt}$, where $e_{it} = \log(E_{it})$, $p_{it} = \log(P_{it})$, and $p_{jt} = \log(P_{jt})$. Previous panel unit root tests are built on the equation

$$e_{it} = (p_{it} - p_{jt}) + u_{it}, \tag{14}$$

where u_{it} stands for real exchange rate. The rejection of the null hypothesis that the processes $\{u_{it}, t \geqslant 1\}$ are all nonstationary is regarded as evidence in favor of the validity of the long-run PPP or mean-reversion of real exchange rate. The most important assumption in the strong PPP hypothesis is that there exists a one-to-one relationship between the nominal exchange rates and aggregate price ratios. In practice the movements may not be directly proportional, leading to the cointegrating slopes deviating away from unity. Pedroni (2004) modifies (14) by allowing for heterogeneous coefficients across individuals and estimating the following long-run PPP hypothesis in weak version

$$e_{it} = \mu_i + \beta_i(p_{it} - p_{jt}) + u_{it} = \mu_i + \beta_i \Delta p_{ij,t} + u_{it}, \tag{15}$$

where β_i is allowed to vary across countries and is expected to be positive, $\Delta p_{ij,t} = \beta_i(p_{it} - p_{jt})$, and μ_i is the unobserved fixed effect for country i.

In our weak PPP model, we assume that β_i exhibits the latent group structures studied in this paper. By pooling the slope coefficients within a group together, we can obtain more efficient estimates than those obtained from a fully heterogeneous cointegrated panel model. In addition, since our C-Lasso method is a data-driven method, we do not manually assign different countries to different groups, which alleviates the sample selection problem.

We obtain monthly and quarterly data of nominal exchange rate and consumer price index (CPI) from Jan. 1975 to Jul. 2014 covering 99 countries from International Financial Statistics. Here, we use CPI to represent the general price index. We choose the time span from 1975 to 2014 to cover the post-Bretton Woods period. Given the fact that Euro dollar was introduced to the global financial markets as an accounting currency on 1 January 1999, we consider two subsamples. We obtain a balanced panel with 67 countries in the period 1975-1998 and another balanced panel with 99 countries in the period 1999-2014. For the quarterly data, we have 91 time series periods in 1975Q.1-1998.Q4 and 55 times series periods in 1999.Q1-2014.Q2. For the monthly data, we have 283 time series periods in period 1975.M1-1998.M12 and 172 times series periods in 1999.M1-2014.M7.

1.6.3 Group and estimation results

In this section, we present the classification and estimation results for the quarterly data. The results for the monthly data are relegated to the online supplementary appendix.

Table 5: The information criterion for different numbers of groups (quarterly data)

	From 1975.Q1-1998.Q4				From 1999.Q1-2014.Q2			
K/c_λ	0.025	0.05	0.10	0.20	0.025	0.05	0.10	0.20
1	-0.7503	-0.7503	-0.7503	-0.7503	-0.2074	-0.2074	-0.2074	-0.2074
2	-1.1262	**-1.1262**	**-1.1262**	**-1.0716**	-0.4719	-0.4730	-0.4902	**-0.4836**
3	**-1.1622**	-0.7961	-1.0956	-0.7135	**-0.5230**	**-0.5319**	**-0.5268**	-0.4418
4	-0.7719	-0.7507	-0.7507	-1.0596	-0.5037	-0.4994	-0.4958	-0.3815
5	-0.7233	-0.7203	-0.6750	-0.6750	-0.4789	-0.4749	-0.3499	-0.2093
6	-0.6946	-0.6405	-0.6005	-0.6844	-0.4454	-0.4358	-0.3566	-0.1720

We determine the number of groups by using the information criterion (IC) proposed in Section 4.5. Table 5 reports the information criterion with different tuning parameter values: $\lambda = c_\lambda \times T^{-3/4}$ where $c_\lambda = 0.025, 0.05, 0.1,$ and 0.2. Obviously, IC is robust to the choice of tuning parameters. Following the majority rule, we decide to select $K = 2$ groups for the period 1975.Q1-1998.Q4 and $K = 3$ groups for the period 1999.Q1-2014.Q2. Note that the IC is minimized at $c_\lambda = 0.1$ and 0.05 for the first and second subsamples respectively. We will choose $c_\lambda = 0.1$ and 0.05 for these two subsamples respectively and report the estimation results.

Table 6 reports the DPLS estimation results for the subsamples 1975.Q1-1998.Q4 and 1999.Q1-2014.Q2 by using $c_\lambda = 0.1$ and 0.05, respectively. We summarize some important findings from Table 6. First, the group-specific estimates vary a lot across groups, which indicates strong unobserved heterogeneities in both subsamples. Second, both C-Lasso estimate and its post-Lasso one for Group 1 are reasonably close to the unit in both the first and second subsamples, which lends some positive support to the weak form long-run PPP hypothesis. But the estimates in Group 2 in either subsamples suggest negative long-run relationship between the price index difference and the exchange rate, which contradicts the long-run PPP hypothesis. The estimate for Group 3 in the second subsample is positive and quite small in comparison with the unity, which suggests quite weak proportional relation between the change in general price index and that of exchange rate. Third, similar results are also observed for the monthly data, and the long-run relation between nominal exchange rate and general price index presents similar patterns in either

Table 6: Estimation results for the quarterly data

Panel A: From 1975.Q1-1998.Q4							
	Pool	Group 1			Group 2		
	DOLS	C-Lasso	post-Lasso		C-Lasso	post-Lasso	
β_i	0.7465	0.8609	0.8608		-0.7007	-0.6992	
	(0.0207)	(0.0190)	(0.0190)		(0.0857)	(0.0857)	
Panel B: From 1999.Q1-2014.Q2							
	Pool	Group 1		Group 2		Group 3	
	DOLS	C-Lasso	post-Lasso	C-Lasso	post-Lasso	C-Lasso	post-Lasso
β_i	0.3623	0.8667	0.8681	-0.5732	-0.5775	0.1986	0.1960
	(0.0184)	(0.0189)	(0.0189)	(0.0227)	(0.0228)	(0.0296)	(0.0296)

subsample period. This indicates the robustness of our findings.

Table 7 summarizes the group classification results for the two subsamples; see also Figure 1 for the classification results for the second subsample. Interestingly, we find that the majority of the countries in the first subsample are classified into Group 1, which indicates the long-run PPP holds for most countries in the period 1975.Q1-1998.Q4. During this time span, we have only 68 countries in the dataset and some developing countries like Argentina, Brazil, and Russia are excluded from our subsample due to the fact that they have experienced hyperinflation. For the second subsample, we find even more interesting results. Figure 1 suggests that those countries that support the long-run PPP equilibrium are mainly located in Europe, Africa, middle East, and north American. The members of Group 1 suggest a polarization of economic development. Further, we observe that most countries in Groups 2 and 3 are either fast-growing or middle-income countries (e.g., South Korea, Singapore and Brazil) in the last decades in East-Asia and South America. It confirms the Balassa-Samuelson effect, where the productivity differentials are one of the most important factors behind the PPP deviation, see Balassa (1964) and Samuelson (1964). In this case, countries with rapidly expanding economies should tend to have more rapidly appreciating exchange rate. In general, our results suggest heterogeneous behavior in the long-run PPP hypothesis.

Table 7: Classification results for the quarterly data

Panel A: From 1975.Q1-1998.Q4				
Group 1 ($N_1 = 62$)				
Algeria	Australia	**Austria**	**Bahrain**	**Belgium**
Bolivia	**Botswana**	**Canada**	**Colombia**	**Costa Rica**
Cyprus	**Denmark**	Dominican	**Egypt**	El Salvador
Finland	**France**	**Ghana**	**Greece**	Guatemala
Honduras	**Hungary**	Iceland	**India**	**Indonesia**
Iran	Ireland	**Israel**	**Italy**	**Ivory Coast**
Jamaica	**Japan**	**Jordan**	**Kenya**	**South Korea**
Luxembourg	**Malta**	**Mauritius**	**Mexico**	**Morocco**
Nepal	**Netherlands**	New Zealand	**Nigeria**	**Norway**
Pakistan	**Paraguay**	**Philippines**	**Portugal**	**Singapore**
South Africa	**Spain**	**Sri Lanka**	**Sudan**	**Sweden**
Switzerland	Tanzania	**Thailand**	**Trinidad and Tobago**	**Turkey**
Uruguay	**Venezuela**			
Group 2 ($N_2 = 5$)				
Ecuador	**Kuwait**	Malaysia	**Myanmar**	Saudi Arabia
Panel B: From 1999.Q1-2014.Q2				
Group 1 ($N_1 = 49$)				
Angola	**Argentina**	**Austria**	**Bangladesh**	**Belgium**
Botswana	Brunei	**Canada**	**Costa Rica**	**Denmark**
Dominican	**Europe**	**Finland**	**France**	**Germany**
Ghana	**Honduras**	**Iceland**	**Iran**	**Italy**
Jamaica	**Japan**	**Jordan**	**Luxembourg**	**Malawi**
Mexico	**Mongolia**	**Morocco**	**Mozambique**	**Netherlands**
Nigeria	**Norway**	**Pakistan**	**Romania**	**Saudi Arbia**
Sri Lanka	**Sudan**	**Sweden**	**Switzerland**	**Tanzania**
Trinidad and Tobago	**Tunisia**	**Turkey**	**Uganda**	**United Kingdom**
Ukraine	**Venezuela**	**Viet Nam**	Zambia	
Group 2 ($N_2 = 23$)				
Albania	**Armenia**	Australia	Bolivia	**Brazil**
Bulgaria	**Colombia**	**Congo**	**Croatia**	El Salvador
Georgia	**Hungary**	**Ireland**	**Ivory Coast**	**Kuwait**
Latvia	**Macau**	**Moldova**	New Zealand	**Peru**
Philippines	**Spain**	**Thailand**		
Group 3 ($N_3 = 27$)				
Algeria	Cambodia	**Czech Republic**	Egypt	**Guatemala**
Hong Kong	India	**Indonesia**	**Israel**	**Kazakhstan**
Kenya	**South Korea**	**Kyrgyzstan**	**Laos**	Lithuania
Macedonia	**Malaysia**	Mauritius	**Myanmar**	Nepal
Paraguay	**Poland**	**Portugal**	**Russia**	**Singapore**
South Africa	Uruguay			

Note: Countries in bold denote coincidences of the classification results based on the monthly and quarterly datasets.

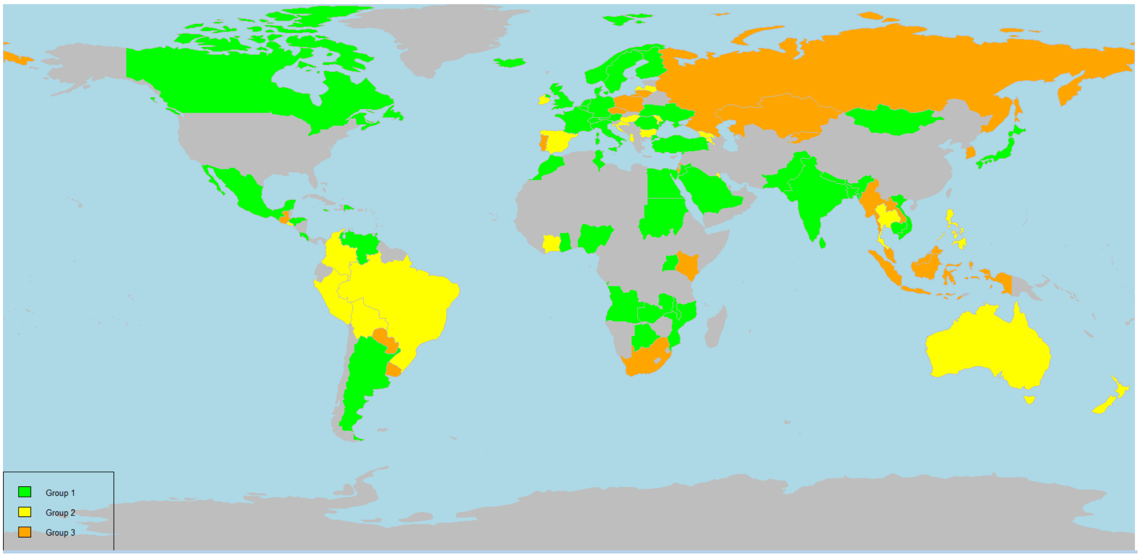


Figure 1: The geographic features of countries in the three groups in subsample 2 (1999-2014)

1.7 Conclusion

In this paper we propose a C-Lasso-based PLS procedure to estimate a cointegrated panel with latent group structures on the long-run cointegrating relationships. We allow for completely heterogeneous short-run dynamics but assume that long-run relationships are homogeneous within a group and heterogeneous across different groups. Our method can determine the individual's group membership consistently and estimate the parameters efficiently. To remove the asymptotic bias in the estimators of the long-run parameters, we also consider the dynamic PLS procedure. Simulation results confirm the asymptotic studies. An application to testing the validity of the long-run PPP hypothesis suggests strong evidence of latent group structures.

There are several interesting topics for further research. First, we do not allow for cross-sectional dependence in our model. In macro-econometrics, cross-sectional dependence is frequently modelled via the multi-factor error structure (Pesaran, 2006) or interactive fixed effects (Bai, 2009). Depending on whether we allow for unit-root behavior in the factors, different methods can be called upon (see, e.g., Bai and Ng, 2004; Bai and Kao, 2006; Bai et al., 2009; Bai and Ng, 2010). But this certainly complicates the asymptotic analysis and deserves a separate treatment. Second, when the dimension of the nonstationary variables is higher than 2, multiple cointegrating relationships may exist. It is worthwhile to consider the panel vector error-correction model or likelihood-based panel cointegration analysis in this case. We leave these topics for future research.

2 Nonstationary Panel Models with Latent Group Structures and Cross-sectional Dependence

2.1 Introduction

Nonstationary panel models have been extensively used in empirical analyses. Their asymptotic properties are well explored in classical settings, such as the assumptions of common coefficients and independence across individuals. Although these assumptions offer efficient estimation and simplify asymptotic theory, they are often hard to meet in real-world economic problems. In one case, researchers often face the issue of unobserved parameter heterogeneity that figures within models, including the "convergence clubs" (Durlauf and Johnson (1995) and Quah (1997), the relation between income and democracy (Acemoglu et al. (2008)), and the "resources curse" (Van der Ploeg (2011)). In another case, globalization and international spillovers raise to a new challenge–cross-sectional dependence. In general, these two features can substantially complicate asymptotic theory and statistical inference in nonstationary panels. The goal of this paper is to simultaneously study the unobserved parameter heterogeneity and cross-sectional dependence in nonstationary panel models. In the meanwhile, we seek to maintain simple and efficient estimation.

In this paper, we consider a nonstationary panel model with latent group structures and unobserved common factors. Specifically, we first assume that the long-run relationships are heterogeneous across groups and homogeneous within a group. The unobserved structures offer flexible parameter settings and remain efficiency from pooling within a group. Moreover, there are economic intuitions for considering group patterns on the long-run relationships. For example, the long-run equilibriums in growth regression share some common features in subsamples, such as developing and developed countries, but reveal distinct patterns across subsamples. Second, we employ factor structures to model cross-sectional dependence. In our model, we typically consider both stationary and nonstationary common factors. For example, an oil price shock and a global technology trend both affect countries' GDP level. A stock market shock and a macro-growth trend both affect security prices. In general, our framework allow us to fit more complex features to the data in empirical applications.

A distinctive feature of our estimation is purely data-driven. We propose a penalized principal component (PPC, hereafter) method, which can be regarded as iterative procedures between penalized regression and principal component analysis. These estimation procedures provide three Lasso-type estimators for long-run relationships, namely C-Lasso, post-Lasso and Cup-Lasso (continuous-updated) estimators, and principal component estimators for unobserved common factors. We take advantages of a growing literature on *Classifier-Lasso* techniques (see Su et al. (2016), Qian and Su (2016) and Su et al. (2017)) to build reliable computation algorithm in the presence of unobserved common factors.

Our theoretical results are concerned with developing a limit theory for our Lasso-type estimators. We first establish preliminary rates of convergence for the group-specific estimators and unobserved common factors. Next, we show classification consistency, which indicates that all individuals are classified into correct group with a probability approaching one (w.p.a.1). Moreover, we find that our Lasso-type estimators have non-negligible biases, which come from two sources. The first bias is commonly noted in nonstationary panels due to weakly dependent error processes (see Phillips and Moon (1999)). The second part comes from unobserved stationary common factors. We can further show that the stationary common factors complicates both asymptotic biases and covariance structures but don't affect consistency of the long-run estimators. In post-Lasso and Cup-Lasso estimators, we employ a fully modified procedure for bias-correction, proposed by Phillips and Hansen (1990). Therefore, our estimators are centered around zero and achieve the usual $\sqrt{NT}$ consistency in homogeneous nonstationary panel models. Furthermore, we show the oracle property for our estimators, which are asymptotically equivalent to the corresponding infeasible estimators, obtained by knowing the exact individuals' group membership. Lastly, we establish a mixed normal asymptotic distribution for our estimators. Thus, the usual t, Wald and F statistics can be used for inference.

In above analyses, we assume the number of groups and that of common factors are known. In practice, we propose three information criteria to determine the number of groups, the number of stationary common factors and the number of nonstationary common factors, respectively. We demonstrate that these information criteria can select the correct number of groups and common factors w.p.a.1. In terms of simulation, we show good finite sample performance for estimation and classification.

We illustrate the use of our methods by studying the heterogeneous behavior in international R&D spillover model. Similar to Coe and Helpman (1995), we regress the total factor productivity (TFP, hereafter) on domestic R&D capital stock and foreign R&D capital stock. In existing work, there are two limitations in econometric methods. First, an important assumption underlying the original work is that all countries obey a common linear specification. However, cross-countries productivity behavior typically reaches multiple steady states. In addition, recent work suggests two types of R&D spillovers–positive technology spillovers and negative market rivalry effects (see Bloom et al. (2013)). Therefore, a natural solution is to allow the parameters vary across countries and to reveal different spillover patterns. Second, variables like TFP and R&D stocks apparently share some common patterns, such as global technology trends and financial crisis shocks. In such cases, their regressions are misspecified and lead to inconsistent estimates. Therefore, our econometric method yields a direct solution for these two problems, first, to introduce the latent group structures in parameters of interests and, second, to estimate unobserved common patterns directly from data.

In the empirical application, we first confirm positive technology spillovers in pooled sample after considering one common trend. Moreover, the group-specific estimates identify heterogeneous spillover patterns across countries. These results first indicates the existence of two types R&D spillovers–positive technology spillovers and negative market rivalry effects in country level. Based on the group patterns, countries are classified into three groups–“Convergence”, “Divergence”, and “Balance”. The major sources of technology changes in “Convergence” group come from positive technology diffusions. As a result, the catch-up effects through the channel of technology diffusion favor growth convergence hypothesis. Conversely, when market rivalry effects dominate technology spillovers, we observe an overall negative R&D spillovers. For these countries, their technology growth rely on domestic innovations and exhibit divergence behavior. As a result, we explain the growth convergence puzzle by heterogeneous behavior in R&D spillovers.

A key contribution of this paper is offering a practical approach that accommodates the unobserved heterogeneity and cross-sectional dependence in nonstationary panels. We provide consistent and efficient estimators to the group-specific long-run relationships even when individuals' membership were unknown. The estimation method is similar to that proposed by Su et al. (2016) (SSP, hereafter), but simultaneously accounts for cross-sectional dependence. Recently several papers account for the unobserved heterogeneity in large dimensional panel models by clustering and grouping, such as grouped fixed effects (Bonhomme and Manresa (2015)), structure breaks (Qian and Su (2016)), grouped factor models (Ando and Bai (2016)). Almost all methods focus on stationary panel models. Although Huang et al. (2017) consider the latent group patterns in cointegrated panels, they do not allow for cross-sectional dependence. Relative to existing work, we establish formal conditions and asymptotic properties under which the Lasso-type estimators perform well in the sense of consistent estimates, good classification results and faster convergence rates.

Our theoretical results also contributes to two strands of literature on cointegrated panels and factor models. It is noted that the average and common long-run estimators permit a normal asymptotic distributions. But the heterogeneous long-run estimators have nonstandard asymptotics (see Phillips and Moon (1999), Kao and Chiang (2001), and Pedroni (2004)). In this context, we maintain the simplicity of normal distribution with unobserved parameter heterogeneity. In addition, there is a growing literature using factor models to capture cross-sectional dependence under large N and large T settings (see Bai and Ng (2002), Bai (2004), Phillips and Sul (2003), Pesaran (2006), Bai (2009), and Moon and Weidner (2017)). Compared to existing work, we formally study the presence of both stationary and nonstationary common factors. In addition, an appealing feature of our econometric theory is to build a linkage between time series asymptotics for integrated processes and advanced panel techniques. In this sense, the asymptotic results allow for more general forms of panel data features and time series properties both in dependent variables and common factors.

Our empirical work speaks to a long literature on growth convergence, in particular the analyses of global technology diffusions (see Barro and Sala-i Martin (1992), Quah (1996), Eaton and Kortum (2002), and Griffith et al. (2004)). For example, Barro and Sala-i Martin (1997) theoretically confirm a uniform convergence behavior across economies based on technology diffusions. However, Quah (1996) believes that economic structures varies in many –explicable and inexplicable–ways across countries. In addition, he argues the uniform convergence results may come from the misleading statistical implications of nonstationary time series. In this context, our empirical results reconcile these two controversial arguments on growth convergence. We find that some countries exhibit convergence but others fail, which is determined by the dominated effects in R&D spillovers.

This paper is structured as follows. Section 2 introduces a nonstationary panel model with latent group structures and cross-sectional dependence and proposes a penalized principal component method for estimation. Section 3 explains main assumptions and establishes asymptotic properties of the three Lasso-type estimators. Section 4 reports simulation results. Section 5 studies the heterogeneous behavior of international R&D spillovers. Section 6 concludes. All proofs are relegated to the appendix.

NOTATION. Hereafter, we write the integral $\int_0^1 W(s)ds$ as $\int W$ and define $\Omega^{1/2}$ to be any matrix such that $\Omega = (\Omega^{1/2})(\Omega^{1/2})'$, and $BM(\Omega)$ to denote Brownian motion with the covariance matrix Ω. For any $m \times n$ real matrix A, we write the Frobenius norm $\|A\|$, the spectral norm $\|A\|_{sp}$, the transpose A'. The operator $\xrightarrow{p}$ denotes convergence in probability, $\Rightarrow$ weak convergence, $a.s.$ almost surely, and $[x]$ the largest integer less than or equal to x. When A is symmetric, we use $\mu_{\max}(A)$ and $\mu_{\min}(A)$ to denote its largest and smallest eigenvalues, respectively. Let $M < \infty$ be a generic positive number, not depending on T or N. We also define the matrix that projects onto orthogonal space of A as $M_A = I_T - A(AA')^{-1}A'$. Let $0_{p\times 1}$ denote a $p \times 1$ vector of zeros and $\mathbf{1}\{\cdot\}$ the indicator function. We use "p.d." and "p.s.d." to abbreviate "positive definite" and "positive semidefinite", respectively. Unless indicated explicitly, we use $(N,T) \to \infty$ to stand for that N and T pass jointly to infinity.

2.2 Model and Estimation

In this section, we first introduce a nonstationary panel model with latent group structures and cross-sectional dependence. Then we propose a penalized principal component method to estimate the model.

2.2.1 Nonstationary panel model with latent group structures and cross-sectional dependence

The generating processes of (y_{it}, x_{it}) are as follows

$$\begin{cases} y_{it} = \beta_i^{0\prime} x_{it} + e_{it} \\ x_{it} = x_{it-1} + \varepsilon_{it}, \end{cases} \tag{16}$$

where y_{it} is a scalar, x_{it} is a $p \times 1$ vector of nonstationary regressors of order one (I(1) process) for all i, e_{it} is an error term and assumed to be cross-sectionally dependent due to unobserved common factors, ε_{it} is assumed to have zero mean and finite variance, and β_i^0 is a $p \times 1$ vector of unknown long-run cointegration relationships. We assume that β_i are heterogeneous across groups and homogeneous within a group. And we denote the true values of β_i as β_i^0, to follow the latent group structures, such that

$$\beta_i^0 = \begin{cases} \alpha_1^0 & \text{if } i \in G_1^0 \\ \vdots & \vdots \\ \alpha_K^0 & \text{if } i \in G_K^0 \end{cases}, \tag{17}$$

where $\alpha_j^0 \neq \alpha_k^0$ for any $j \neq k$, $\bigcup_{k=1}^K G_k^0 = \{1, 2, \dots N\}$, and $G_k^0 \bigcap G_j^0 = \varnothing$ for any $j \neq k$. Let $N_k = \#G_k$ denote the cardinality of the set G_k^0. For the moment, we assume that the number of group K is known and fixed but each individual's group membership is unknown. We propose a information criterion to determine the number of groups in Section 3.6.

Since e_{it} is cross-sectionally dependent, we impose a multi-factor structure on e_{it}. That is,

$$e_{it} = \lambda_i^{0\prime} f_t^0 + u_{it} = \lambda_{1i}^{0\prime} f_{1t}^0 + \lambda_{2i}^{0\prime} f_{2t}^0 + u_{it},$$

where f_t^0 is an $r \times 1$ vector of unobserved common factors that contains an $r_1 \times 1$ vector of nonstationary factors f_{1t}^0 of order one (I(1) process) and an $r_2 \times 1$ vector of stationary factors f_{2t}^0 (I(0) process), λ_i is an $r \times 1$ vector of factor loadings and u_{it} is the idiosyncratic component of e_{it} with zero mean and long-run variance, which is assumed to cross-sectionally independent. We emphasize that cross-sectional dependence only comes from common factors f_t such that e_{it} and e_{jt} are correlated due to common factors f_t in the form of $E(e_{it}e_{jt}) = \lambda_i' E(f_t f_t')\lambda_j \neq 0$.

If f_t only contains stationary factors, in some cases we can still obtain consistent estimators of β_i by a penalized least squares (PLS, hereafter) method when ignoring cross-sectional dependence (see Huang et al. (2017)). However, if there are serial correlations between dependent variable x_{it} and unobserved common factors f_t, ignoring those factors f_t yields biased inference for β_i. Furthermore, the unobserved nonstationary factors lead to inconsistency due to a spurious regression. In general, we fail to obtain consistent and unbiased group-specific estimators by the PLS-based method in nonstationary panel models with cross-sectional dependence.

Now we incorporate the multi-factor error structure to the first equation of (16) as follows

$$y_{it} = \beta_i^{0\prime} x_{it} + \lambda_i^{0\prime} f_t^0 + u_{it}. \tag{18}$$

Our estimation procedures are performed on model (18) by the penalized principal component method, proposed in Section 2.2. Let

$$\boldsymbol{\alpha} \equiv (\alpha_1, ..., \alpha_{K_0}), \quad \boldsymbol{\beta} \equiv (\beta_1, ..., \beta_N), \quad \Lambda = (\lambda_1, ..., \lambda_N)', \quad \text{and} \quad f = (f_1, ..., f_T)'.$$

The true values of $\boldsymbol{\alpha}, \boldsymbol{\beta}, \Lambda$ and f are denoted as $\boldsymbol{\alpha}^0, \boldsymbol{\beta}^0, \Lambda^0$, and f^0, respectively. We also use $\alpha_k^0, \beta_i^0, \lambda_i^0$ and f_t^0 denote the true value of $\alpha_k, \beta_i, \lambda_i$ and f_t. Our interest is to infer each individual's group identity and obtain consistent estimators of both group-specific long-run relationships α_k and unobserved common factors f_t.

2.2.2 Penalized principal component estimation

In this section, we propose an iterative PPC-based procedure to jointly estimate the long-run relationships β_i, unobserved common factors f_t and to identify group membership. Here, we rewrite model (18) in vector form,

$$y_i = x_i\beta_i^0 + f^0\lambda_i^0 + u_i = x_i\beta_i^0 + f_1^0\lambda_{1i}^0 + f_2^0\lambda_{2i}^0 + u_i, \tag{19}$$

where $f^0 = (f_1^0, f_2^0)$, $\lambda_i^0 = (\lambda_{1i}^{0\prime}, \lambda_{2i}^{0\prime})'$, $y_i = (y_{i1}, ..., y_{iT})'$, x_i, f_1^0, f_2^0, and u_i are analogously defined. As we discuss in Section 2.1, we can still obtain consistent estimates of β_i when ignoring unobserved stationary common factors. The principal component estimators of β_i and f_1^0 are obtained from the following least objective function

$$\text{SSR}(\beta_i, f_1, \Lambda_1) = \sum_{i=1}^{N}(y_i - x_i\beta_i - f_1\lambda_{1i})'(y_i - x_i\beta_i - f_1\lambda_{1i}), \tag{20}$$

subject to the constraint $\frac{f_1'f_1}{T^2} = I_{r_1}$ and $\Lambda_1'\Lambda_1$ being diagonal. Define the projection matrix $M_{f_1} = I_T - P_{f_1} = I_T - \frac{f_1f_1'}{T^2}$. We can obtain the least squares estimator of β_i for each given f_1 is

$$\hat{\beta}_i = (x_i'M_{f_1}x_i)^{-1}x_i'M_{f_1}y_i, \text{ for } i = 1, ..., N.$$

Given β_i, the variable $e_i = y_i - x_i\beta_i = f\lambda_i + u_i$ has a pure factor structure. Let $e = (e_1, e_2, ..., e_N)$, a $T \times N$ matrix and $\Lambda_1 = (\lambda_{11}, ..., \lambda_{1N})'$ a $N \times r_1$ matrix. We can obtain the least squares objective function for f_1, such that $\text{tr}\left[(e - f_1\Lambda_1')(e - f_1\Lambda_1')'\right]$. By Bai (2009), we concentrate out Λ_1 by its least square estimator, such that $\Lambda_1 = e'f_1(f_1'f_1)^{-1} = e'f_1/T^2$. The objective function (20) becomes

$$\text{tr}\left(e'M_{f_1}e\right) = \text{tr}(e'e) - \text{tr}\left(f_1'ee'f_1/T^2\right).$$

The final least squares estimator $(\hat{\beta}, \hat{f}_1)$ is the solution of the set of nonlinear equations,

$$\hat{\beta}_i = \left(x_i'M_{\hat{f}_1}x_i\right)^{-1}\left(x_i'M_{\hat{f}_1}y_i\right), \tag{21}$$

$$\hat{f}_1V_{1,NT} = \left[\frac{1}{NT^2}\sum_{i=1}^{N}\left(y_i - x_i\hat{\beta}_i\right)\left(y_i - x_i\hat{\beta}_i\right)'\right]\hat{f}_1, \tag{22}$$

where $M_{\hat{f}_1} = I_T - \frac{1}{T^2}\hat{f}_1\hat{f}_1'$, $\frac{1}{T^2}\hat{f}_1'\hat{f}_1 = I_{r_1}$, and $V_{1,NT}$ is a diagonal matrix consisting of the r_1 largest eigenvalues of the matrix inside the brackets, arranged in decreasing order. Based on (21) and (22), we can further show that $\hat{\Lambda}_1'\hat{\Lambda}_1$ is a diagonal matrix with descending diagonal elements as follows,

$$\frac{1}{N}\hat{\Lambda}_1'\hat{\Lambda}_1 = T^{-2}\hat{f}_1'\left(\frac{1}{NT^2}\sum_{i=1}^{N}\left(y_i - x_i\hat{\beta}_i\right)\left(y_i - x_i\hat{\beta}_i\right)'\hat{f}_1\right) = \left(T^{-2}\hat{f}_1'\hat{f}_1\right)V_{1,NT} = V_{1,NT}.$$

Given the initial estimates of β_i and f_1 obtained from (21) and (22), we propose a penalized principal component method to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, where $\boldsymbol{\beta}$ exhibits the latent group structures. The PPC criterion function is given by

$$Q_{NT}^{\lambda,K}(\boldsymbol{\beta},\boldsymbol{\alpha},f_1) = Q_{NT}(\boldsymbol{\beta},f_1) + \frac{\lambda}{N}\sum_{i=1}^{N}\prod_{k=1}^{K}\|\beta_i - \alpha_k\| \tag{23}$$

where $Q_{NT}(\boldsymbol{\beta},f_1) = \frac{1}{NT^2}\sum_{i=1}^{N}(y_i - x_i\beta_i)' M_{f_1}(y_i - x_i\beta_i)$, $\lambda = \lambda(N,T)$ is the tuning parameter. Minimizing the PPC criterion function in (23) produces the *Classifier-Lasso* (C-Lasso, hereafter) estimators of β_i and α_k, respectively. Then we update the estimates of the nonstationary common factors f_1 as follows

$$\hat{f}_1 V_{1,NT} = \left[\frac{1}{NT^2}\sum_{k=1}^{K}\sum_{i\in\hat{G}_k}(y_i - x_i\hat{\alpha}_k)(y_i - x_i\hat{\alpha}_k)'\right]\hat{f}_1. \tag{24}$$

with the identification restrictions: $\frac{1}{T^2}\hat{f}_1'\hat{f}_1 = I_{r_1}$ and $V_{1,NT}$ is a diagonal matrix with descending diagonal elements. Since we allow for both stationary and nonstationary common factors, we minimize the following equation to obtain consistent estimates of stationary common factors f_2,

$$\hat{f}_2 V_{2,NT} = \left[\frac{1}{NT}\sum_{k=1}^{K}\sum_{i\in\hat{G}_k}(y_i - x_i\hat{\alpha}_k - \hat{f}_1\hat{\lambda}_{1i})(y_i - x_i\hat{\alpha}_k - \hat{f}_1\hat{\lambda}_{1i})'\right]\hat{f}_2. \tag{25}$$

with the identification restrictions: $\frac{1}{T}\hat{f}_2'\hat{f}_2 = I_{r_2}$ and $V_{2,NT}$ is a diagonal matrix with descending diagonal elements. After obtaining the estimates of f_2, we apply bias-correction in post-Lasso estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. The biases emerge from the unobserved stationary common factors, endogeneity, and serial correlation issues from weakly dependent error terms.

Now we summarize the estimation procedures in the PPC-based estimation method. We first obtain the prior estimates of $\hat{\beta}_i$ and $\hat{f}_1$ by solving equations (21) and (22). Second, we minimize the above PPC criterion function (23), which produces the C-Lasso estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$. Third, with C-Lasso estimates of α, we update the estimates of nonstationary common factor f_1 by (24) and estimate stationary common factors f_2 by (25). Fourth, we apply bias-correction by a fully modified procedure in the post-Lasso estimator of $\boldsymbol{\alpha}$, which is explained in Section 3.4. We iterate steps 2–4 until achieving convergence to obtain the Cup-Lasso estimators. Our estimators, which we refer to as “C-Lasso”, “post-Lasso”, and “Cup-Lasso”, are based on the optimal group of cross-sectional individuals, according to the PPC criterion function. The triplet $(\hat{\beta}, \hat{\alpha}, \hat{f}_1)$ jointly minimizes the objective function (23). Let $\hat{\beta}_i$ and $\hat{\alpha}_k$ denote the i^{th} and k^{th} columns of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$, respectively, i.e., $\hat{\boldsymbol{\beta}} \equiv (\hat{\beta}_1, ..., \hat{\beta}_N)$ and $\hat{\boldsymbol{\alpha}} \equiv (\hat{\alpha}_1, ..., \hat{\alpha}_K)$. We study asymptotic properties of the C-Lasso, post-Lasso and Cup-Lasso estimators below.

2.3 Asymptotic Theory

2.3.1 Main assumptions

In this subsection, we introduce main assumptions that are needed to study the asymptotic properties of our estimators $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\alpha}}$ and $\hat{f}_1$.

Let $Q_{ixx}(f_1) = \frac{1}{T^2} x_i' M_{f_1} x_i$, $Q_1(f_1)$ =diag$(Q_{1,xx}, ..., Q_{N,xx})$, and

$$Q_2(f_1) = \begin{pmatrix} \frac{1}{NT^2} x_1' M_{f_1} x_1 a_{11} & \frac{1}{NT^2} x_1' M_{f_1} x_2 a_{12} & \cdots & \frac{1}{NT^2} x_1' M_{f_1} x_N a_{1N} \\ \frac{1}{NT^2} x_2' M_{f_1} x_1 a_{21} & \frac{1}{NT^2} x_2' M_{f_1} x_2 a_{22} & \cdots & \frac{1}{NT^2} x_2' M_{f_1} x_N a_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{NT^2} x_N' M_{f_1} x_1 a_{N1} & \frac{1}{NT^2} x_N' M_{f_1} x_2 a_{N2} & \cdots & \frac{1}{NT^2} x_N' M_{f_1} x_N a_{NN} \end{pmatrix},$$

where f_1 satisfies $\frac{1}{T^2} f_1' f_1 = I_{r_1}$. Note that $Q_2(f_1)$ is an $Np \times Np$ matrix. Let $w_{it} = (u_{it}, \varepsilon_{it}', \Delta f_{1t}^{0\prime}, f_{2t}^{0\prime},)'$. and $\mathcal{C} = \sigma(\Lambda, f)$ is the sigma algebra generated by factors and factor loadings. Let M be a generic constant that can vary across lines.

We make the following assumptions on $\{w_{it}\}$ and $\{\lambda_i\}$.

Assumption 3.1 *(i) For each i, $\{w_{it}, t \geqslant 1\}$ is a linear process: $w_{it} = \phi_i(L) v_{it} = \sum_{j=0}^{\infty} \phi_{ij} v_{i,t-j}$, where $v_{it} = (v_{it}^{u}, v_{it}^{\varepsilon\prime}, v_t^{f_1\prime}, v_t^{f_2\prime})'$ is a $(1 + p + r_1 + r_2) \times 1$ vector sequence of i.i.d. random variables over t with zero mean and variance matrix I_{1+p+r}; $\max_{1 \leqslant i \leqslant N} E(\|v_{it}\|^{2q+\epsilon}) < M$, where $q > 4$ and ϵ is an arbitrarily small positive constant; v_{it}^{u}, v_{it}^{ε}, $v_t^{f_1}$, and $v_t^{f_2}$ are mutually independent, and $(v_{it}^{u}, v_{it}^{\varepsilon\prime})'$ are independent across i.*

(ii) $\max_{1 \leqslant i \leqslant N} \sum_{j=0}^{\infty} j^k \|\phi_{ij}\| < \infty$ and $|\phi_i(1)| \neq 0$ for some $k \geqslant 2$.

(iii) u_{it} and ε_{it} are cross-sectionally independent conditional on $\mathcal{C}$.

(iv) λ_i is independent of v_{jt} for all i, j,and t.

Following Phillips and Solo (1992), we assume that $\{w_{it}\} = \{w_{it}, t \geqslant 1\}$ is a linear process in Assumption 3.1(i). For latter reference, we partition $\phi_i(L)$ conformably with

w_{it} as follows:

$$\phi_i(L) = \begin{pmatrix} \phi_i^{uu}(L) & \phi_i^{u\varepsilon}(L) & \phi_i^{uf_1}(L) & \phi_i^{uf_2}(L) \\ \phi_i^{\varepsilon u}(L) & \phi_i^{\varepsilon\varepsilon}(L) & \phi_i^{\varepsilon f_1}(L) & \phi_i^{\varepsilon f_2}(L) \\ \phi^{f_1 u}(L) & \phi^{f_1\varepsilon}(L) & \phi^{f_1 f_1}(L) & \phi^{f_1 f_2}(L) \\ \phi^{f_2 u}(L) & \phi^{f_2\varepsilon}(L) & \phi^{f_2 f_1}(L) & \phi^{f_2 f_2}(L) \end{pmatrix} = \begin{pmatrix} \phi_i^{uu}(L) & \phi_i^{u\varepsilon}(L) & \phi_i^{uf_1}(L) & 0 \\ \phi_i^{\varepsilon u}(L) & \phi_i^{\varepsilon\varepsilon}(L) & \phi_i^{\varepsilon f_1}(L) & \phi_i^{\varepsilon f_2}(L) \\ 0 & 0 & \phi^{f_1 f_1}(L) & \phi^{f_1 f_2}(L) \\ 0 & 0 & \phi^{f_2 f_1}(L) & \phi^{f_2 f_2}(L) \end{pmatrix}. \tag{26}$$

Since both nonstationary and stationary common factors do not depend on i, we have $\phi^{f_1 u}(L) = \phi^{f_1 \varepsilon}(L) = \phi^{f_2 u}(L) = \phi^{f_2 \varepsilon}(L) = 0$. Moreover, we assume that $\phi_i^{uf_2}(L) = 0$. This assumption indicates that there exists no serial correlation or contemporaneous correlation between the regression error u_{it} and unobserved stationary common factors f_{2t}^0, and it ensures consistency for our initial estimators. The finite $2q + \varepsilon$ moments for $q > 4$ ensure the validity of the law of large numbers (LLN) and functional central limit theory (FCLT) for the weakly dependent linear process $\{w_{it}\}$. We frequently apply the Beveridge and Nelson (BN) decomposition as follows

$$w_{it} = \phi_i(1)v_{it} + \tilde{w}_{it-1} - \tilde{w}_{it},$$

where $\tilde{w}_{it} = \sum_{j=0}^{\infty} \tilde{\phi}_{ij} v_{i,t-j}$ and $\tilde{\phi}_{ij} = \sum_{s=j+1}^{\infty} \phi_{is}$. Assumption 3.1(ii) gives the summability conditions on the coefficients matrix ϕ_{ij}. By Lemma (BN) in Phillips and Solo (1992), we have $\sum_{j=1}^{\infty} j^k \|\phi_{ij}\|^k < \infty \rightarrow \sum_{j=0}^{\infty} \|\tilde{\phi}_{ij}\|^k < \infty$, which implies that $\tilde{w}_{it}$ has Wold decomposition and behaves like a stationary process. Specifically, we have $\sum_{j=0}^{\infty} \left\|\tilde{\phi}_{ij}\right\|^2 < \infty$ under $\sum_{j=1}^{\infty} j^{1/2}\|\phi_{ij}\| < \infty$. The suitable choice of k ensures the finite kth moment of $\tilde{w}_{it}$. In our case, we need strong conditions to ensure the uniform behavior across i. The second part of Assumption 3.1(ii) rules out potential cointegration relationships among x_{it} and f_{1t}^0. Assumption 3.1(iii) emphasizes that the cross-sectional dependence only comes from the unobserved common factors. Assumption 3.1(iv) ensures that the factor loadings are independent of the generalization of the error processes both over t and across i.

By Assumption 3.1, we have the multivariate invariance principle for the partial sum process of w_{it}. That is,

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{[Tr]} w_{it} \Rightarrow B_i(r) \equiv BM_i(\Omega_i) \text{ as } T \rightarrow \infty \text{ for all } i,$$

where $B_i = (B_{1i}, B_{2i}', B_3', B_4')'$ is a $(1+p+r_1+r_2)\times 1$ vector of Brownian motions with long-run covariance matrix Ω_i. We can also define the temporal variance $\Sigma_i = E(w_{i0}w_{i0}')$ and the one-sided long-run covariance matrix $\Delta_i = \sum_{j=0}^{\infty} E(w_{i0}w_{ij}') = \Gamma_i + \Sigma_i$ of $\{w_{it}\}$, where Ω_i has the following partition

$$\Omega_i = \sum_{j=-\infty}^{\infty} E(w_{ij}w_{i0}') = \Gamma_i' + \Gamma_i + \Sigma_i = \begin{pmatrix} \Omega_{11,i} & \Omega_{12,i} & \Omega_{13,i} & \Omega_{14,i} \\ \Omega_{21,i} & \Omega_{22,i} & \Omega_{23,i} & \Omega_{24,i} \\ \Omega_{31,i} & \Omega_{32,i} & \Omega_{33} & \Omega_{34} \\ \Omega_{41,i} & \Omega_{42,i} & \Omega_{43} & \Omega_{44} \end{pmatrix}.$$

Let S_1, S_2, S_3, and S_4 denote respectively the $1\times(1+p+r)$, $p\times(1+p+r)$, $r_1\times(1+p+r)$ and $r_2\times(1+p+r)$, selection matrices such that $S_1 w_{it} = u_{it}$, $S_2 w_{it} = \varepsilon_{it}$, $S_3 w_{it} = \Delta f_{1t}^0$, and $S_4 w_{it} = f_{2t}^0$.

Assumption 3.2 *(i) As $N\to\infty$, $\frac{1}{N}\Lambda^{0\prime}\Lambda^0 \xrightarrow{p} \Sigma_\lambda > 0$. $\max_{1\leqslant i\leqslant N} E\|\lambda_i^0\|^{2q} \leqslant M$ for some $q\geqslant 4$ and $\Lambda_1^{0\prime}\Lambda_2^0 = O_P\left(N^{1/2}\right)$.*

(ii) $E\|\Delta f_{1t}^0\|^{2q+\epsilon} \leqslant M$ and $E\|f_{2t}^0\|^{2q+\epsilon} \leqslant M$ for some $\epsilon > 0$, $q \geqslant 4$ and for all t. As $T\to\infty$, $\frac{1}{T^2}\sum_{t=1}^{T} f_{1t}^0 f_{1t}^{0\prime} \xrightarrow{d} \int B_3 B_3$ and $\frac{1}{T}\sum_{t=1}^{T} f_{2t}^0 f_{2t}^{0\prime} \xrightarrow{p} \Sigma_{44} > 0$, where B_3 is a r_1-vector of Brownian motions with long-run covariance matrix $\Omega_{33} > 0$.

(iii) Let $\gamma_N(s,t) = E(\frac{1}{N}\sum_{i=1}^{N} u_{it}u_{is})$ and $\xi_{st} = \frac{1}{N}\sum_{i=1}^{N} u_{it}u_{is} - E(\frac{1}{N}\sum_{i=1}^{N} u_{it}u_{is})$. Then $\max_{1\leqslant s,t\leqslant T} N^2 \times E|\xi_{st}|^4 \leqslant M$ and $T^{-1}\sum_{s=1}^{T}\sum_{t=1}^{T}\|\gamma_N(s,t)\|^2 \leqslant M$.

(iv) There exists a constant $\rho_{\min} > 0$ such that $P\left(\min_{1\leqslant i\leqslant N}\inf_{f_1}\mu_{\min}\left(Q_1(f_1) - 2Q_2(f_1)\right) \geqslant c\rho_{\min}\right) = 1 - o(N^{-1})$, where the $\inf$ is taken respect to f_1 such that $\frac{1}{T^2}f_1'f_1 = I_{r_1}$.

Assumption 3.2(i)-(iii) imposes standard moment conditions in factor literature; see, e.g., Bai and Ng (2002), Bai (2004). The last condition in Assumption 3.2(i) indicates that the stationary factor loadings and the nonstationary factor loadings can be only weakly correlated, which will greatly facilitate the derivation. Assumption 3.2(iii) imposes conditions on error processes $\{u_{it}\}$, which are adapted from Bai (2003) and allow for weak forms of cross-sectional and serial dependence in error processes. Assumption 3.2(iv) assumes $Q_1(f_1) - 2Q_2(f_1)$ is positive definite in the limit across i when f_1 satisfies the restriction $\frac{1}{T^2}f_1'f_1 = I_{r_1}$. This assumption is the identification condition for β_i, which is related to ASSUMPTION A in Bai (2009) (p.1241). Since f_1 is to be estimated, the identification condition for β_i is imposed on the set of f_1 satisfying the restriction $\frac{1}{T^2}f_1'f_1 = I_{r_1}$.

Assumption 3.3 *(i) For each* $k = 1, ..., K_0$, $N_k/N \to \tau_k \in (0,1)$ *as* $N \to \infty$.

(ii) $\min_{1\leqslant k\neq j\leqslant K} \left\|\alpha_k^0 - \alpha_j^0\right\| \geqslant \underline{c}_\alpha$ *for some fixed* $\underline{c}_\alpha > 0$.

(iii) As $(N,T) \to \infty$, $N/T^2 \to c_1 \in [0,\infty)$, $T/N^2 \to c_2 \in [0,\infty)$.

(iv) As $(N,T) \to \infty$, $\lambda d_T \to 0$, $\lambda T N^{-1/q} d_T^{-2}/\left(\log T\right)^{1+\epsilon} \to \infty$, *and* $d_T^2 N^{1/q} T^{-1}\left(\log T\right)^{1+\epsilon} \to 0$.

Assumption 3.3(i)-(ii) are borrowed from SSP. Assumption 3.3(i) implies that each group has an asymptotically non-negligible number of individuals as $N \to \infty$ and Assumption 3.3(ii) requires the separability of group-specific parameters. Similar conditions are assumed in panel literature with latent group patterns, e.g., Bonhomme and Manresa (2015), Ando and Bai (2016), Su and Ju (2017). Assumption 3.3(iii)-(iv) imposes conditions to control the relative rates at which N and T pass to infinity. Note that N can pass to infinity at a faster or slower rate than T. The involving of d_T is due to the law of iterated logarithm, such that $d_T = O(\log\log T)$. One can verify that the range of values for λ to satisfy Assumption 3.3(iv) is $\lambda \propto T^{-\alpha}$ for $\alpha \in (0, \frac{q-1}{q})$.

2.3.2 Preliminary rates of convergence

Let $\hat{b}_i = \hat{\beta}_i - \beta_i^0$, $\delta_{NT} = \min(\sqrt{N}, T)$, $C_{NT} = \min(\sqrt{N}, \sqrt{T})$, $\eta_{NT}^2 = \frac{1}{N}\sum_{i=1}^N \left\|\hat{b}_i\right\|^2$, and $H_1 = (\frac{1}{N}\Lambda_1^{0\prime}\Lambda_1^0)(\frac{1}{T^2}f_1^{0\prime}\hat{f}_1) \times V_{1,NT}^{-1}$. We establish the consistency of $\hat{\beta}_i$ and $\hat{f}_1$ by the following theorem.

Theorem 7 *Suppose that Assumptions 3.1-3.2 hold. Then*

(i) $\frac{1}{N}\sum_{i=1}^N \left(\hat{\beta}_i - \beta_i^0\right)' \frac{1}{T^2} x_i' M_{\hat{f}_1} x_i \left(\hat{\beta}_i - \beta_i^0\right) = o_P(1)$,

(ii) $\left\|P_{\hat{f}_1} - P_{f_1^0}\right\| = o_P(1)$,

(iii) $\frac{1}{N}\sum_{i=1}^N \|\hat{\beta}_i - \beta_i^0\|^2 = o_P(1)$,

(iv) $\frac{1}{T}\|\hat{f}_1 - f_1^0 H_1\| = O_P(\eta_{NT}) + \frac{1}{\sqrt{T}} O_P(C_{NT}^{-1})$.

Theorem 7(i) establishes the weighted mean square consistency of $\{\hat{\beta}_i\}$. Theorem7(ii) shows that the space spanned by the columns of $\hat{f}_1$ and f_1^0 are asymptotically the same. Given the weighted mean square consistency and Assumption 3.2(iv), we can further establish the non-weighted mean square consistency of β_i in Theorem7(iii). As expected, Theorem 7(iv) indicates that the true factor f_1^0 can only be identified up to a nonsingular rotation matrix. Compared to Bai and Ng (2004) and Bai et al. (2009), our results allow for both heterogeneous slope coefficients and unobserved stationary and nonstationary common factors.

The following theorem establishes the rate of convergence for the individual and group-specific estimators and the estimated factors as well.

Theorem 8 *Suppose that Assumptions 3.1-3.2 hold. Then*

(i) $\frac{1}{N}\sum_{i=1}^{N}\|\hat{\beta}_i-\beta_i^0\|^2=O_P(d_TT^{-2})$,

(ii) $\hat{\beta}_i-\beta_i^0=O_P(d_T^{1/2}T^{-1}+\lambda)$ *for* $i=1,2,...,N$,

(iii) $(\hat{\alpha}_{(1)},...,\hat{\alpha}_{(K)})-(\alpha_1^0,...,\alpha_K^0)=O_P(d_TT^{-1})$ *for some suitable permutation* $(\hat{\alpha}_{(1)},...,\hat{\alpha}_{(K)})$ *of* $(\hat{\alpha}_1,...,\hat{\alpha}_K)$,

(iv) $T^{-1}\|\hat{f}_1-f_1^0H_1\|^2=O_P(N^{-1}+d_T^2T^{-1})$.

Theorem 8(i)-(ii) establishes the mean-square and point-wise convergence of the slope coefficients β_i, respectively. The usual super consistency of nonstationary estimators $\hat{\beta}_i$ is preserved if $\lambda=O(T^{-1})$ despite the fact that we ignore unobserved stationary common factors and allow for correlations between u_{it} and (x_{it},f_{1t}^0). Theorem 8(iii) indicates that the group-specific parameters, $\alpha_1^0,...,\alpha_{K_0}^0$, can be consistently estimated. Theorem 8(iv) updates the convergence rate of the unobserved nonstationary factors in Theorem 7(iv).

For notational simplicity, hereafter we simply write $\hat{\alpha}_k$ for $\hat{\alpha}_{(k)}$ as the consistent estimator of α_k^0's. Let $\hat{G}_k=\{i\in\{1,2,...,N\}:\hat{\beta}_i=\hat{\alpha}_k\}$ for $k=1,...,K$. Let $\hat{G}_0$ denote the group of individuals in $\{1,2,...,N\}$ that are not classified into any of the K groups.

2.3.3 Classification consistency

In this subsection, we study the classification consistency. Define

$$\hat{E}_{kNT,i}=\{i\notin\hat{G}_k|i\in G_k^0\}\quad\text{and}\quad\hat{F}_{kNT,i}=\{i\notin G_k^0|i\in\hat{G}_k\},$$

where $i=1,...,N$ and $k=1,...K^0$. Let $\hat{E}_{kNT}=\cup_{i\in\hat{G}_k}\hat{E}_{kNTi}$ and $\hat{F}_{kNT}=\cup_{i\in\hat{G}_k}\hat{F}_{kNTi}$. The events $\hat{E}_{kNT}$ and $\hat{F}_{kNT}$ mimic Type I and Type II errors in statistical tests. Following SSP, we say that a classification method is individual consistent if $P(\hat{E}_{kNT,i})\to 0$ as $(N,T)\to\infty$ for each $i\in G_k^0$ and $k=1,...,K$, and $P(\hat{F}_{kNT,i})\to 0$ as $(N,T)\to\infty$

for each $i \in G_k^0$ and $k = 1, ..., K$. It is uniformly consistent if $P(\cup_{k=1}^{K} \hat{E}_{kNT}) \to 0$ and $P(\cup_{k=1}^{K} \hat{F}_{kNT}) \to 0$ as $(N, T) \to \infty$.

The following theorem establishes the uniform classification consistency.

Theorem 9 *Suppose that Assumptions 3.1-3.3 hold. Then*

(i) $P(\cup_{k=1}^{K_0} \hat{E}_{kNT}) \leqslant \sum_{k=1}^{K_0} P(\hat{E}_{kNT}) \to 0$ *as* $(N, T) \to \infty$,

(ii) $P(\cup_{k=1}^{K_0} \hat{F}_{kNT}) \leqslant \sum_{k=1}^{K_0} P(\hat{F}_{kNT}) \to 0$ *as* $(N, T) \to \infty$.

Theorem 9 implies the uniform classification consistency– all individuals within a certain group, say G_k^0, can be simultaneously correctly classified into the same group (denoted $\hat{G}_k$) w.p.a.1. Conversely, all individuals that are classified into the same group, say $\hat{G}_k$, simultaneously belong to the same group (G_k^0) w.p.a.1.

2.3.4 Oracle properties, post-Lasso and Cup-Lasso estimators

In this subsection, we study oracle properties of PPC-based estimators. To proceed, we add some notations. For $k = 1, ..., K$, we define

$$U_{kNT} = \frac{1}{\sqrt{N_k T}} \sum_{i \in G_k^0} x_i M_{f_1^0} \left(\left(u_i + f_2^0 \lambda_{2i}^0\right) - \frac{1}{N} \sum_{j=1}^{N} \left(u_j + f_2^0 \lambda_{2j}^0\right) a_{ij} \right),$$

$$B_{kNT,1} = \frac{1}{\sqrt{N_k T}} \sum_{i \in G_k^0} \left(\sum_{t=1}^{T} \sum_{s=1}^{T} \left[\mathbf{1}\left\{t = s\right\} - \varkappa_{ts} \mathbf{1}\left\{s \leqslant t\right\}\right] \right) \Delta_{21,i},$$

$$B_{kNT,2} = \frac{1}{\sqrt{N_k T}} \sum_{i \in G_k^0} E_{\mathcal{C}}\left(x_i\right)' M_{f_1^0} f_2^0 \left(\lambda_{2i}^0 - \frac{1}{N} \sum_{j=1}^{N} \lambda_{2j}^0 a_{ij} \right),$$

$$\begin{aligned} V_{kNT} =& \frac{1}{\sqrt{N_k T}} \sum_{i \in G_k^0} S^{\varepsilon} \phi_i^{\dagger}(1) \sum_{t=1}^{T} \sum_{s=1}^{T} \left\{ \bar{\varkappa}_{ts} \left(V_{it}^{u\varepsilon} v_{is}^{u\varepsilon\prime}\right) - \left[\mathbf{1}\left\{t = s\right\} - \varkappa_{ts} \mathbf{1}\left\{s \leqslant t\right\}\right] I_{1+p} \right\} \phi_i^{\dagger}(1)' S^{u\prime} \\ &+ \frac{1}{\sqrt{N_k T}} \sum_{i=1}^{N} \left\{ E_{\mathcal{C}}\left(x_i'\right) \mathbf{1}\left\{i \in G_k^0\right\} - \frac{1}{N} \sum_{j \in G_k^0} a_{ij} E_{\mathcal{C}}(x_j') \right\} M_{f_1^0} u_i \\ &+ \frac{1}{\sqrt{N_k T}} \sum_{i \in G_k^0} [x_i - E_{\mathcal{C}}\left(x_i\right)]' M_{f_1^0} f_2^0 \lambda_{2i}^0, \end{aligned}$$

where $\varkappa_{ts} = f_{1t}^{0\prime} (f_1^{0\prime} f_1^0)^{-1} f_{1s}^0$, $\bar{\varkappa}_{ts} = \mathbf{1}\left\{t = s\right\} - \varkappa_{ts}$, $\mathcal{C} = \sigma\left(\Lambda^0, f^0\right)$, $E_{\mathcal{C}}\left(\cdot\right) = E_{\mathcal{C}}\left(\cdot | \mathcal{C}\right)$,

$$\phi_i^{\dagger}\left(L\right) = \begin{pmatrix} \phi_i^{u\dagger}\left(L\right) \\ \phi_i^{\varepsilon\dagger}\left(L\right) \end{pmatrix} = \begin{pmatrix} \phi_i^{uu}\left(L\right) & \phi_i^{u\varepsilon}\left(L\right) \\ \phi_i^{\varepsilon u}\left(L\right) & \phi_i^{\varepsilon\varepsilon}\left(L\right) \end{pmatrix}, \ S^u = \left(1, 0_{1 \times p}\right), \text{ and } S^{\varepsilon} = \left(0_{p \times 1}, I_p\right).$$

Let

$$Q_{1NT} = \text{diag}\left(\frac{1}{N_1T^2}\sum_{i\in G_1^0} x_i' M_{f_1^0} x_i, \ldots, \frac{1}{N_K T^2}\sum_{i\in G_K^0} x_i' M_{f_1^0} x_i\right),$$

$$Q_{2NT} = \begin{pmatrix} Q_{2NT,11} & \cdots & Q_{2NT,1K} \\ \vdots & \ddots & \vdots \\ Q_{2NT,K1} & \cdots & Q_{2NT,KK} \end{pmatrix}$$

where $Q_{2NT,kl} = \frac{1}{N_k NT^2}\sum_{i\in G_k^0}\sum_{j\in G_l^0} x_i' M_{f_1^0} x_j a_{ij}$ for $k, l = 1, ..., K$. Let

$$Q_{NT} = Q_{1NT} - Q_{2NT} \text{ and } Q_0 = \begin{pmatrix} Q_{1,1} - Q_{2,11} & -Q_{2,12} & \cdots & -Q_{2,1K} \\ -Q_{2NT,21} & Q_{1,2} - Q_{2,22} & \cdots & -Q_{2,2K} \\ \vdots & \vdots & \ddots & \vdots \\ -Q_{2,K1} & -Q_{2,K2} & \cdots & Q_{1NT,K} - Q_{2,KK} \end{pmatrix},$$

where

$$Q_{1,k} = \lim_{N\to\infty}\frac{1}{N_k}\sum_{i\in G_k^0} E_{\mathcal{C}}\left(\int \tilde{B}_{2i}\tilde{B}_{2i}'\right),$$

$$Q_{2,kl} = \lim_{N\to\infty}\frac{1}{NN_k}\sum_{i\in G_k^0}\sum_{j\in G_l^0} a_{ij} E_{\mathcal{C}}\left(\int \tilde{B}_{2,i}\tilde{B}_{2,j}\right),$$

$$\tilde{B}_{2i} = B_{2,i} - \int B_{2,i}B_3'\left(\int B_3 B_3'\right)^{-1} B_3.$$

Let $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, ..., \hat{\alpha}_K)$. Let $U_{NT} = (U_{1NT}', \ldots, U_{KNT}')'$, $B_{NT} = (B_{1NT}', \ldots, B_{KNT}')'$, $V_{NT} = (V_{1NT}', \ldots, V_{KNT}')'$ and $B_{kNT} = B_{kNT,1} + B_{kNT,2}$. The following theorem reports the Bahadur-type representation and asymptotic distribution of $\text{vec}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^0)$.

Theorem 10 *Suppose that assumptions 3.1-3.3 hold and $\sqrt{N} = o(T)$. Let $\hat{\alpha}_k$ be obtained by solving (23). Then*

(i) $\sqrt{N}T vec(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^0) = \sqrt{D_N} Q_{NT}^{-1} U_{NT} + o_P(1) = \sqrt{D_N} Q_{NT}^{-1}(V_{NT} + B_{NT}) + o_P(1),$

(ii) $\sqrt{N}T vec(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^0) - \sqrt{D_N} Q_{NT}^{-1} B_{NT} \Rightarrow MN(0, \lim_{N\to\infty} D_N Q_0^{-1}\Omega_0 Q_0^{-1})$ *as* $(N,T)\to\infty,$

where $D_N = diag\left(\frac{N}{N_1}, ..., \frac{N}{N_K}\right)$, $\Omega_0 = \lim_{(N,T)\to\infty}\Omega_{NT}$, *and* $\Omega_{NT} = Var(V_{NT}|\mathcal{C})$.

Theorem 10 indicates that V_{NT} and B_{NT} are associated with the asymptotic variance and bias of $\hat{\alpha}_k$'s, respectively. Note that $B_{kNT} = B_{kNT,1} + B_{kNT,2}$, which indicates two sources of biases. The appearance of $B_{kNT,1}$ results from the correlation between

(x_{it}, f_{1t}) and u_{it} and the serial correlation among the innovation process $\{w_{it}\}$. Apparently, the presence of unobserved nonstationary factors f_{1t}^0 complicates the formula for $B_{kNT,1}$ through the term $\varkappa_{ts}(= f_{1t}^{0\prime}(f_1^{0\prime}f_1^0)^{-1}f_{1s}^0)$. The second source of asymptotic bias is due to unobserved stationary factors f_{2t}^0 so that $B_{kNT,2} = 0$ if f_{2t}^0 is absent from the model. In the special case where neither f_{1t}^0 nor f_{2t}^0 is present in the model, we have $B_{kNT} = B_{kNT,1} = \frac{1}{\sqrt{N_k}} \sum_{i \in G_k^0} \Delta_{21,i}$. This is the usual bias term for panel cointegration regression that is associated with the one-sided long-run covariance; see Phillips (1995) and Phillips and Moon (1999). Note that the ith element of V_{NT} is independent across i conditional on $\mathcal{C}$ and $E_{\mathcal{C}}(V_{NT}) = 0$. This makes it possible for us to derive a version of conditional CLT for V_{NT} and establish the limiting distribution of our estimators $\hat{\boldsymbol{\alpha}}$ in 10(ii).

As we show in the proof of Theorem 10, the asymptotic bias term B_{NT} is of $O(\sqrt{N_k})$, which implies T-consistency of the C-Lasso estimators $\hat{\alpha}_k$. In order to obtain the$\sqrt{N}T$-convergence rate, we call upon various procedures to remove the asymptotic bias by constructing consistent estimates of B_{NT}.

The fully modified procedure

In this subsection, we first obtain the estimates of unobserved stationary factors f_{2t}^0 from (25). Then we employ a fully modified procedure of Phillips and Hansen (1990) and Phillips (1995) to make bias-corrections for endogeneity and serial correlation. Below we consider the three types of bias-corrected estimators: the bias-corrected post-Lasso estimator $\hat{\alpha}_{\hat{G}_k}^{bc}$, the fully-modified post-Lasso estimator $\hat{\alpha}_{\hat{G}_k}^{fm}$, and the fully-modified Cup-Lasso estimator $\hat{\alpha}_{\hat{G}_k}^{cup}$.

Following Phillips (1995), we first construct consistent time series estimators of the long-run covariance matrix Ω_i and the one-sided long-run covariance matrix Δ_i by

$$\hat{\Omega}_i = \sum_{j=-T+1}^{T-1} \omega\left(\frac{j}{H}\right) \hat{\Gamma}_i(j), \text{ and } \hat{\Delta}_i = \sum_{j=0}^{T-1} \omega\left(\frac{j}{H}\right) \hat{\Gamma}_i(j),$$

where $\omega(\cdot)$ is a kernel function, H is the bandwidth parameter, and $\hat{\Gamma}_i(j) = \frac{1}{T}\sum_{t=1}^{T-j} \hat{w}_{it+j}\hat{w}_{it}'$ with $\hat{w}_{it} = (\hat{u}_{it}, \Delta x_{it}', \Delta\hat{f}_{1t}', \hat{f}_{2t}')'$. We partition $\hat{\Omega}_i$ and $\hat{\Delta}_i$ conformably with Ω_i.

We make the following assumption on the kernel function and bandwidth.

Assumption 3.4 *(i) The kernel function* $\omega(\cdot)$*:* $R \to [-1, 1]$ *is a twice continuously differentiable symmetric function such that* $\int_{-\infty}^{\infty} \omega(x)^2 dx \leqslant \infty$*,* $\omega(0) = 1$*,* $\omega(x) = 0$ *for* $|x| \geqslant 1$*, and* $\lim_{|x| \to 1} \omega(x)/(1 - |x|)^q = c > 0$ *for some* $q \in (0, \infty)$*.*

(ii) As $(N, T) \to \infty$*,* $N/H^{2q} \to 0$ and $H/T \to 0$.

We modify the variable y_{it} with the follow transformation to correct the endogeneity

$$\hat{y}_{it}^{+} = y_{it} - \hat{\Omega}_{12,i}\hat{\Omega}_{22,i}^{-1}\Delta x_{it}. \tag{27}$$

This would lead to the modified equation $\hat{y}_{it}^{+} = \beta_i^{0\prime} x_{it} + \lambda_{1i}^{0\prime} f_{1t}^{0} + \lambda_{2i}^{0\prime} f_{2t}^{0} + \hat{u}_{it}^{+}$, where $\hat{u}_{it}^{+} = u_{it} - \hat{\Omega}_{12,i}\hat{\Omega}_{22i}^{-1}\Delta x_{it}$. Define

$$\hat{\Delta}_{12,i}^{+} = \hat{\Delta}_{12,i} - \hat{\Omega}_{12,i}\hat{\Omega}_{22i}^{-1}\hat{\Delta}_{22,i}. \tag{28}$$

By Phillips (1995), (27) and (28) give correction for endogeneity and serial correlation, respectively.

Therefore, we can obtain the bias-correction post-Lasso estimator $\hat{\alpha}_{\hat{G}}^{bc}$, fully modified post-Lasso estimator $\hat{\alpha}_{\hat{G}_k}^{fm}$, and the updated estimators of $\hat{f}_1$ and $\hat{f}_2$ by iteratively solving (29)-(32), such that

$$\text{vec}\left(\hat{\alpha}_{\hat{G}}^{bc}\right) = \text{vec}\left(\hat{\alpha}\right) - \frac{1}{\sqrt{NT}}\sqrt{D_N} Q_{NT}^{-1}\left(\hat{B}_{NT,1} + \hat{B}_{NT,2}\right), \tag{29}$$

$$\hat{\alpha}_{\hat{G}_k}^{fm} = \left(\sum_{i \in \hat{G}_k} x_i' M_{\hat{f}_1} x_i\right)^{-1} \left\{\sum_{i \in \hat{G}_k} x_i' M_{\hat{f}_1} \hat{y}_i^{+} - T\sqrt{N_k}\left(\hat{B}_{kNT,1}^{+} + \hat{B}_{kNT,2}\right)\right\}, \tag{30}$$

$$\hat{f}_1 V_{1,NT} = \left[\frac{1}{NT^2}\sum_{k=1}^{K}\sum_{i \in \hat{G}_k}(\hat{y}_i - x_i\hat{\alpha}_{\hat{G}_k}^{fm})(\hat{y}_i - x_i\hat{\alpha}_{\hat{G}_k}^{fm})'\right]\hat{f}_1, \tag{31}$$

$$\hat{f}_2 V_{2,NT} = \left[\frac{1}{NT}\sum_{k=1}^{K}\sum_{i \in \hat{G}_k}(\hat{y}_i - x_i\hat{\alpha}_{\hat{G}_k}^{fm} - \hat{f}_1\hat{\lambda}_{1i})(\hat{y}_i - x_i\hat{\alpha}_{\hat{G}_k}^{fm} - \hat{f}_1\hat{\lambda}_{1i})'\right]\hat{f}_2, \tag{32}$$

where $\hat{B}_{kNT,1} = \frac{1}{\sqrt{N_k}T}\sum_{i \in \hat{G}_k}\left(\sum_{t=1}^{T}\sum_{s=1}^{t}\hat{\hat{\varkappa}}_{ts}\right)\hat{\Delta}_{21,i}$, $\hat{B}_{kNT,1}^{+} = \frac{1}{\sqrt{N_k}T}\sum_{i \in \hat{G}_k}\left(\sum_{t=1}^{T}\sum_{s=1}^{t}\hat{\hat{\varkappa}}_{ts}\right)\hat{\Delta}_{21,i}^{+}$, $\hat{B}_{kNT,2} = \frac{1}{\sqrt{N_k}T}\sum_{i \in \hat{G}_k}\left(\sum_{t=1}^{T}\sum_{s=1}^{t}\hat{\hat{\varkappa}}_{ts}\right)\hat{\Delta}_{24,i}\hat{\hat{\lambda}}_{2i}$, $\hat{\hat{\varkappa}}_{ts} = \mathbf{1}\{t = s\} - \hat{\varkappa}_{ts}$,$\hat{\varkappa}_{ts} = \hat{f}_{1t}'(\hat{f}_1'\hat{f}_1)^{-1}\hat{f}_{1s} = \hat{f}_{1t}'\hat{f}_{1s}/T^2$ and $\hat{\hat{\lambda}}_{2i} = \hat{\lambda}_{2i} - \frac{1}{N}\sum_{j=1}^{N}\hat{\lambda}_{2j}\hat{a}_{ij}$. We obtain the fully modified Cup-Lasso estimators $\hat{\alpha}_{\hat{G}_k}^{cup}$ by iteratively solving (23), and (30)-(32), where we update the group classification results in each iteration.

Let $\hat{\boldsymbol{\alpha}}_{\hat{G}}^{fm} = (\hat{\alpha}_{\hat{G}_1}^{fm}, ..., \hat{\alpha}_{\hat{G}_K}^{fm})$ and $\hat{\boldsymbol{\alpha}}_{\hat{G}}^{cup} = (\hat{\alpha}_{\hat{G}_1}^{cup}, ..., \hat{\alpha}_{\hat{G}_K}^{cup})$. We establish the limiting distribution of the bias-correction post-Lasso estimators $\hat{\alpha}_{\hat{G}}^{bc}$, the fully modified post-Lasso estimators $\hat{\boldsymbol{\alpha}}_{\hat{G}}^{fm}$ and the Cup-Lasso estimators $\hat{\boldsymbol{\alpha}}_{\hat{G}}^{cup}$ by the following theorem.

Theorem 11 *Suppose that assumptions 3.1-3.4 hold. Let $\hat{\boldsymbol{\alpha}}_{\hat{G}}^{bc}$ be obtained by iteratively solving (29), (31)-(32), $\hat{\boldsymbol{\alpha}}_{\hat{G}}^{fm}$ be obtained by iteratively solving (30)-(32) and $\hat{\boldsymbol{\alpha}}_{\hat{G}}^{cup}$ be obtained by iteratively solving (23) and (30)-(32). As $(N,T) \to \infty$ with $\sqrt{N} = o\,(T)$, we have*

(i) $\sqrt{N}T vec(\hat{\boldsymbol{\alpha}}_{\hat{G}}^{bc} - \boldsymbol{\alpha}^0) \Rightarrow MN(0, \lim_{N\to\infty} D_N Q_0^{-1}\Omega_0 Q_0^{-1})$,

(ii) $\sqrt{N}T vec(\hat{\boldsymbol{\alpha}}_{\hat{G}}^{fm} - \boldsymbol{\alpha}^0) \Rightarrow MN(0, \lim_{N\to\infty} D_N Q_0^{-1}\Omega_0^{+} Q_0^{-1})$,

(iii)$\sqrt{N}T vec(\hat{\boldsymbol{\alpha}}_{\hat{G}}^{cup} - \boldsymbol{\alpha}^0) \Rightarrow MN(0, \lim_{N\to\infty} D_N Q_0^{-1}\Omega_0^{+} Q_0^{-1})$,

where $\Omega_0^{+} = \lim_{N,T\to\infty} \Omega_{NT}^{+}$ *and* $\Omega_{NT}^{+} = Var\big(V_{NT}^{+}|\mathcal{C}\big)$.

All three types of estimators achieve $\sqrt{N}T$ consistency and have a mixed normal asymptotic distribution. One can construct the asymptotic *t*-tests and Wald-tests as usual provided one can obtain consistent estimates of Q_0 and Ω_0^{+}. The procedure is standard given the estimated group structure.

2.3.5 Estimating the number of unobserved factors

In the previous subsections, we assume that the numbers of nonstationary and stationary factors, r_1 and r_2, are known. In this subsection, we introduce two information criteria to determine the number of unobserved factors before the PPC estimation procedure. Let r_1 denote a generic number of nonstationary factors. Let r denote a generic total number of nonstationary and stationary factors. We now use r_1^0 and r^0 to denote their true values, which are assumed to be bounded above by a finite integer $r_{\max}$.

Bai et al. (2009) find that it is not necessary to distinguish I(0) and I(1) factors when one tries to determine the total number of factors based on the first differenced model. After the first differencing, (18) takes the form

$$\Delta y_{it} = \beta_i^{0\prime}\Delta x_{it} + \lambda_i^{0\prime}\Delta f_t^0 + \Delta u_{it},\ t = 2, ..., T, \tag{33}$$

where e.g., $\Delta y_{it} = y_{it} - y_{i,t-1}$. Since the true dimension r^0 is unknown, we start with a model with r unobservable common factors. We now write the factors and factor loadings respectively as f_t^r and λ_i^r, where the superscript r highs the dimension of the underlying factors or factor loadings. Let $G^r \equiv \Delta f^r$ be a matrix of $(T-1) \times r$ unobserved differenced factors with a typical row given by $(G_t^r)' \equiv (\Delta f_t^r)'$. We consider the minimization problem

$$\left\{\hat{G}^r, \hat{\Lambda}^r\right\} = \arg\min_{\Lambda^r, G^r} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=2}^{T} (\Delta y_{it} - \hat{\beta}_i' \Delta x_{it} - \lambda_i^{r\prime} G_t^r)^2,$$

$$\text{s.t. } G^{r\prime} G^r / T = I_r \text{ and } \Lambda^{r\prime} \Lambda^r \text{ is diagonal,}$$

where $\hat{G}^r = (\hat{G}_2^{r\prime}, ..., \hat{G}_T^{r\prime})'$, $\hat{\Lambda}^r = (\hat{\lambda}_1^{r\prime}, ..., \hat{\lambda}_N^{r\prime})'$, and $\hat{\beta}_i$'s are obtained from the model with $r_1 = r_{\max}$ nonstationary factors. It is easy to show that $\hat{\beta}_i$'s are T-consistent, which suffices for our purpose. It is well known that given $\hat{G}^r$, we can solve $\hat{\Lambda}^r$ from the least squares regression as a function of $\hat{G}^r$. But we will suppress the dependence of $\hat{\Lambda}^r$ on $\hat{G}^r$ and define $V_1(r, \hat{G}^r) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=2}^{T} (\Delta y_{it} - \hat{\beta}_i' \Delta x_{it} - \hat{\lambda}_i^{r\prime} \hat{G}_t^r)^2$. Following Bai and Ng (2002), we consider the following information criterion

$$IC_1(r) = \log V_1(r, \hat{G}^r) + r g_1(N, T), \tag{34}$$

where $g_1(N,T)$ is a penalty function. Let $\hat{r} = \arg\min_{0 \leqslant r \leqslant r_{\max}} IC_1(r)$. We add the following assumption.

Assumption 3.5 *As* $(N,T) \to \infty$, $g_1(N,T) \to 0$ *and* $C_{NT}^2 g_1(N,T) \to \infty$, *where* $C_{NT} = \min(\sqrt{N}, \sqrt{T})$.

Assumption 3.5 is common in the literature. It requires that $g_1(N,T)$ pass to zero at certain rate so that both over- and under-fitted models can be eliminated asymptotically.

The following theorem demonstrates that we can apply $IC_1(r)$ to consistently estimate r^0.

Theorem 12 *Suppose that Assumptions 3.1-3.3 and 3.5 hold. Then* $P(\hat{r} = r^0) \to 1$ *as* $(N,T) \to \infty$.

Theorem 12 indicates that we can determine the total number of factors r^0 consistently by minimizing $IC_1(r)$.

As we have discussed in Section 3.4, ignoring the unobserved stationary factors will not affect the consistency of slope coefficient estimator, but generate a bias term that is asymptotically non-negligible. For this reason, it is important to distinguish between nonstationary and stationary factors. Fortunately, it is possible to estimate the number of unobserved nonstationary factors, r_1^0, consistently based on the level data. Once we obtain the consistent estimate of r_1^0, we can also obtain the consistent estimate of the number of unobserved stationary factors, r_2^0, based on Theorem 12.

Let $f_1^{r_1}$ be a matrix of $T \times r_1$ nonstationary factors and $\lambda_{1i}^{r_1}$ be an $r_1 \times 1$ vector of nonstationary factor loadings. Given the preliminary T-consistent estimators $\hat{\beta}_i$'s, we consider the following minimization problem

$$\left\{\hat{f}_1^{r_1}, \hat{\Lambda}^{r_1}\right\} = \arg \min_{\Lambda^{r_1}, f_t^{r_1}} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - \hat{\beta}_i' x_{it} - \lambda_i^{r_1\prime} f_{1t}^{r_1})^2,$$

$$\text{s.t. } f_1^{r_1\prime} f_1^{r_1} / T^2 = I_{r_1} \text{ and } \Lambda^{r_1\prime}\Lambda^{r_1} \text{ is diagonal.}$$

Given $\hat{f}_1^{r_1} = (\hat{f}_{11}^{r_1\prime}, ..., \hat{f}_{1T}^{r_1\prime})'$, we can solve $\hat{\Lambda}^{r_1} = (\hat{\lambda}_{11}^{r_1\prime}, ..., \hat{\lambda}_{1N}^{r_1\prime})'$ as a function of $\hat{f}_1^{r_1}$ through the least squares regression. But we suppress the dependence of $\hat{\Lambda}^{r_1}$ on $\hat{f}_1^{r_1}$ and define

$$V_2(r_1, \hat{f}_1^{r_1}) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - \hat{\beta}_i' x_{it} - \hat{\lambda}_i^{r_1\prime} \hat{f}_{1t}^{r_1})^2.$$

We consider the information criterion:

$$IC_2(r_1) = \log V_2(r_1, \hat{f}_1^{r_1}) + r_1 g_2(N, T), \tag{35}$$

where $g_2(N,T)$ is a penalty function. Let $\hat{r}_1 = \arg\min_{0 \leqslant r_1 \leqslant r_{\max}} IC_2(r_1)$. We add the following condition.

Assumption 3.6 *As* $(N,T) \to \infty$, $g_2(N,T)\frac{\log\log(T)}{T} \to 0$ *and* $g_2(N,T) \to \infty$.

Apparently, the conditions on $g_2(N,T)$ are quite different from the conventional conditions for the penalty function used in information criteria in the stationary framework (e.g., $g_1(N,T)$ in Assumption 3.5). In particular, we now require that $g_2(N,T)$ diverge to infinity rather than converge to zero as in Assumption 3.5. The intuition is that the mean squared residual, $V_2(r_1, \hat{f}_1^{r_1})$, does not have a finite probability limit when the number of nonstationary common factors is under-specified. In fact, we can show that $\frac{\log\log T}{T} V_2(r_1, \hat{f}_1^{r_1})$ converges in probability to a positive constant when $0 \leqslant r_1 < r_1^0$. On the other hand, we have $V_2(r_1, \hat{f}_1^{r_1}) - V_2(r_1^0, \hat{f}_1^{r_1^0}) = O_P(1)$ when $r_1 > r_1^0$.

The following theorem suggests that the use of $IC_2(r_1)$ helps to determine r_1^0 consistently.

Theorem 13 *Suppose that Assumptions 3.1-3.3 and 3.6 hold. Then* $P(\hat{r}_1 = r_1^0) \to 1$ *as* $(N, T) \to \infty$.

In the simulations and applications below, we simply follow Bai and Ng (2002) and Bai (2004) and set

$$g_1(N,T) = \frac{N+T}{NT} \ln\left(C_{NT}^2\right) \text{ and } g_2(N,T) = \alpha_T g_1(N,T),$$

where $\alpha_T = \frac{T}{4\log\log(T)}$. We first estimate the total number of unobserved factors by $\hat{r}$ based on the first-differenced model, and then estimate the number of unobserved non-stationary factors by $\hat{r}_1$ based on the level model. The estimator of r_2^0 is then given by $\hat{r}_2 \equiv \hat{r} - \hat{r}_1$.

2.3.6 Determination of the number of groups

In this subsection, we propose a BIC-type information criterion to determine the number of groups, K. We assume that the true number of group, K_0, is bounded from above by a finite integer $K_{\max}$. We now consider the PPC criterion function

$$Q_{NT,\lambda}^{K}(\boldsymbol{\beta}, \boldsymbol{\alpha}, f_1) = Q_{NT}(\boldsymbol{\beta}, f_1) + \frac{\lambda}{N}\sum_{i=1}^{N}\prod_{k=1}^{K}\|\beta_i - \alpha_k\|,$$

where $1 \leqslant K \leqslant K_{\max}$. By minimizing the above criterion function, we obtain the estimates $\hat{\beta}_i(K,\lambda)$, $\hat{\alpha}_k(K,\lambda)$, $\hat{\lambda}_{1i}(K,\lambda)$ and $\hat{f}_{1t}(K,\lambda)$ of $\beta_i^0, \alpha_k^0, \lambda_i^0$ and f_{1t}^0, where we make by the $\hat{\beta}_i, \hat{\alpha}_k, \hat{\lambda}_{1i}$ and $\hat{f}_{1t}$ on (K,λ) explicit. Let $\hat{G}_k(K,\lambda) = \{i \in \{1,2,...,N\} : \hat{\beta}_i(K,\lambda) = \hat{\alpha}_k(K,\lambda)\}$ for $k = 1,...,K$, and $\hat{G}(K,\lambda) = \{\hat{G}_1(K,\lambda),...,\hat{G}_K(K,\lambda)\}$. Let $\hat{\alpha}_{\hat{G}_k(K,\lambda)}^{cup}$ denote the Cup-Lasso estimate of α_k^0. Define

$$V_3(K) = \frac{1}{NT}\sum_{k=1}^{K}\sum_{i\in\hat{G}_k(K,\lambda)}\sum_{t=1}^{T}\left[y_{it} - \hat{\alpha}_{\hat{G}_k(K,\lambda)}^{cup\prime}x_{it} - \hat{\lambda}_{1i}(K,\lambda)'\hat{f}_{1t}(K,\lambda)\right]^2.$$

Following SSP and Lu and Su (2017), we consider the following information criterion

$$IC_3(K,\lambda) = \log V_3(K) + pKg_3(N,T), \tag{36}$$

where $g_3(N,T)$ is a penalty function. Let $\hat{K}(\lambda) = \arg\min_{1 \leqslant K \leqslant K_{\max}} GIC(K,\lambda)$.

Let $\mathcal{G}^{(K)} = (G_{K,1},...,G_{K,K})$ be any K-partition of the set of individual index $\{1,2,...,N\}$. Define $\hat{\sigma}^2_{\mathcal{G}^{(K)}} = \frac{1}{NT}\sum_{k=1}^{K}\sum_{i \in \hat{G}_{K,k}}\sum_{t=1}^{T}[y_{it} - \hat{\alpha}^{cup\prime}_{\hat{G}_{K,k}} x_{it} - \hat{\lambda}_i(K,\lambda)'\hat{f}_t(K,\lambda)]^2$, where $\hat{\alpha}^{cup}_{\hat{G}_{K,k}}$ is analogously defined as $\hat{\alpha}^{cup}_{\hat{G}_k(K,\lambda)}$ with $\hat{G}_k(K,\lambda)$ being replaced by $G_{K,k}$. Let σ_0^2 =plim$_{(N,T)\to\infty} \frac{1}{NT}\sum_{i=1}^{N}\sum_{i \in G_k^0}\sum_{t=1}^{T}[y_{it} - \alpha_k^{0\prime} x_{it} - \lambda_i^{0\prime} f_t^0]^2$. Define

$$\nu_{NT} = \begin{cases} (NT)^{-1/2} \text{ when there is no unobserved common factor,} \\ \delta_{NT}^{-1} \text{ when there are only unobserved nonstationary common factors,} \\ C_{NT}^{-1} \text{ when there are both unobserved nonstationary and stationary common factors.} \end{cases}$$

ν_{NT} indicates the effect of estimating the nonstationary panel on the use of $IC_3(K,\lambda)$ under different scenarios.

We add the following assumption.

Assumption 3.7 (i) *As* $(N,T) \to \infty$, $\min_{1 \leqslant K < K_0} \inf_{G^{(K)} \in \mathcal{G}_K} \hat{\sigma}^2_{\mathcal{G}^{(K)}} \xrightarrow{p} \underline{\sigma}^2 > \sigma_0^2$.

(ii) As $(N,T) \to \infty$, $g_3(N,T) \to 0$ and $g_3(N,T)/\nu_{NT}^2 \to \infty$.

Assumption 3.7(i) requires that all under-fitted models yield asymptotic mean square errors larger than σ_0^2, which is delivered by the true model. Assumption 3.7(ii) imposes the typical conditions on the penalty function $g_3(N,T)$: it cannot shrink to zero either too fast or too slowly.

The following theorem justifies the validity of using IC_3 to determine the number of groups.

Theorem 14 *Suppose that Assumption 3.1-3.4 and 3.7 hold. Then* $P(\hat{K}(\lambda) = K_0) \to 1$ *as* $(N,T) \to \infty$.

Theorem 14 indicates that as long as λ satisfies Assumption 3.3(iv) and $g_3(N,T)$ satisfies Assumption 3.7(ii), we have $\inf_{1 \leqslant K \leqslant K_{\max}, K \neq K_0} IC_3(K,\lambda) > IC_3(K_0,\lambda)$ as $(N,T) \to \infty$. Consequently, the minimizer of $IC_3(K,\lambda)$ with respect to K equals K_0 w.p.a.1 for a variety choices of λ.

2.4 Monte Carlo Simulations

In this section, we evaluate the finite sample performance of the C-Lasso, bias-corrected post-Lasso, fully-modified post-Lasso and Cup-Lasso estimators and that of the information criteria for determining the number of groups and the number of common factors.

2.4.1 Data generating processes

We consider four data generating processes (DGPs) that cover the cases of both stationary and nonstationary unobserved common factors. Throughout these DGPs, the observations in each DGP are drawn from three groups with $N_1 : N_2 : N_3 = 0.3 : 0.4 : 0.3$. There are four combinations of the sample sizes with $N = 50, 100$ and $T = 40, 80$.

DGP1 (Strictly exogenous nonstationary regressors and unobserved stationary common factors) The observations (y_{it}, x_{it}') are generated from the following model,

$$\begin{cases} y_{it} = \beta_i' x_{it} + c_2(\lambda_{2i}' f_{2t}) + u_{it} \\ x_{it} = x_{it-1} + \varepsilon_{it} \end{cases} \tag{37}$$

where $x_{it} = (x_{1it}, x_{2it})'$ is a 2×1 vector of nonstationary regressors, f_{2t} is a 2×1 vector of stationary common factors. Let $w_{it} = (u_{it}, \varepsilon_{it}', f_{2t}')' \sim$ i.i.d. $N(0, I_5)$. The factor loadings λ_{2i} are i.i.d. $N((1,1)', I_2)$ for $i = 1, ..., N$. We use c to control the importance of unobserved common factors and let $c_2 = 0.5$. The long-run slope coefficients β_i exhibits the group-structure in (17) for $K = 3$ and the true values are

$$(\alpha_1^0, \alpha_2^0, \alpha_3^0) = \left(\begin{pmatrix} 0.4 \\ 1.6 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1.6 \\ 0.4 \end{pmatrix} \right).$$

DGP2 (Weakly dependent nonstationary regressors and unobserved nonstationary common factors) The observations $(y_{it}, x_{it}', f_{1t}')$ are generated from the following model,

$$\begin{cases} y_{it} = \beta_i^{0\prime} x_{it} + c_1(\lambda_{1i}' f_{1,t}) + u_{it} \\ x_{it} = x_{it-1} + \varepsilon_{it} \\ f_{1t} = f_{1t-1} + \nu_t \end{cases} \tag{38}$$

where $x_{it} = (x_{1it}, x_{2it})'$ is a 2×1 vector of nonstationary regressors, f_{1t} is a 2×1 vector of nonstationary common factors. The idiosyncratic errors $w_{it} = (u_{it}, \varepsilon_{it}', \Delta f_{1t}')'$ are generated from a linear process: $w_{it} = \sum_{j=0}^{\infty} \psi_{ij} v_{i,t-j}$, where v_{it} are i.i.d. $N(0, I_5), \psi_{ij} = j^{-3.5} * \Omega^{1/2}$, $\Omega^{1/2}$ is the symmetric square root of Ω, where $\Omega_{lm} = 0.2$ for $l \neq m$, $\Omega_{ll} = 1$ for $l = 2, 3, 4, 5$ and $\Omega_{11} = 0.25$. Let $c_1 = 1$. The factor loadings of nonstationary common factors are i.i.d. $\lambda_{1i} \sim N((1, 1)', I_2)$. The true coefficients of β_i are the same in DGP1.

DGP3 (Weakly dependent nonstationary regressors and mixed unobserved stationary and nonstationary common factors) The observations $(y_{it}, x'_{it}, f'_{1t})$ are generated from the following model,

$$\begin{cases} y_{it} = \beta'_i x_{it} + c_1(\lambda'_{1i} f_{1t}) + c_2(\lambda'_{2i} f_{2t}) + u_{it} \\ x_{it} = x_{it-1} + \varepsilon_{it} \\ f_{1t} = f_{1t-1} + \nu_t \end{cases} \tag{39}$$

where $x_{it} = (x_{1it}, x_{2it})'$ is a 2×1 vector of nonstationary regressors, f_{1t} is a 2×1 vector of nonstationary common factors, and f_{2t} contains one stationary common factors. The idiosyncratic errors $w_{it} = (u_{it}, \varepsilon'_{it}, \Delta f'_{1t}, f'_{2t})'$ are generated from a linear process: $w_{it} = \sum_{j=0}^{\infty} \psi_{ij} v_{i,t-j}$ where v_{it} are i.i.d. $N(0, I_6), \psi_{ij} = j^{-3.5} * \Omega^{1/2}$, $\Omega^{1/2}$ is the symmetric square root of Ω where $\Omega_{lm} = 0.2$ for $l \neq m$, $\Omega_{11} = 0.25$, and $\Omega_{ll} = 1$ for $l = 2, ..., 6$. Let $c_1 = 1$ and $c_2 = 0.5$. The factor loadings $\lambda_i = (\lambda'_{1i}, \lambda'_{2i})'$ are i.i.d. $\lambda_{1i} \sim N((1, 1, 1)', I_3)$. The true coefficients of β_i are the same in DGP1.

DGP4 (Weakly dependent nonstationary regressors and mixed unobserved stationary and nonstationary common factors) The settings of DGP4 is the same with those of DGP3, except for allowing weakly correlation among factor loadings $\lambda_i \sim i.i.d.\ N((1, 1, 1)', I_3 * \Omega_2)$, where $\Omega_{2,lm} = 2/\sqrt{N}$ for $l \neq m$.

In all cases, the number of replications is 500.

2.4.2 Estimate number of unobserved factors

In this subsection, we assess the performance of two information criteria proposed in Section 3.5 before determining the number of group and PPC-based estimation procedure. We choice the BIC-type penalty function $g_1(N, T) = \frac{N+T}{NT} \log(\min(N, T))$ to determine the total number of unobserved factors and $g_2(N, T) = \frac{T}{4 \log(\log(T))} \times g_1(N, T)$

Table 8: Frequency for selecting $r = 1, 2, ..., 5$ total factors and $r_1 = 0, 1, ..., 4$ nonstationary factors

			Differenced Data					Level Data				
	N	T	$r = 1$	$r = 2$	$r = 3$	$r = 4$	$r = 5$	$r_1 = 0$	$r_1 = 1$	$r_1 = 2$	$r_1 = 3$	$r_1 = 4$
DGP1	50	40	0	1.000	0	0	0	1.000	0	0	0	0
	50	80	0	1.000	0	0	0	1.000	0	0	0	0
	100	40	0	1.000	0	0	0	0.998	0.002	0	0	0
	100	80	0	1.000	0	0	0	1.000	0	0	0	0
DGP2	50	40	0	1.000	0	0	0	0	0.004	0.964	0.032	0
	50	80	0	1.000	0	0	0	0.004	0.016	0.976	0.004	0
	100	40	0	1.000	0	0	0	0	0.002	0.958	0.040	0
	100	80	0	1.000	0	0	0	0	0.002	0.976	0.022	0
DGP3	50	40	0	0	1.000	0	0	0.018	0.088	0.894	0	0
	50	80	0	0	1.000	0	0	0.006	0.026	0.968	0	0
	100	40	0	0	1.000	0	0	0	0.008	0.972	0.020	0
	100	80	0	0	1.000	0	0	0	0.012	0.988	0	0
DGP4	50	40	0	0	0.998	0.002	0	0.002	0.060	0.938	0	0
	50	80	0	0	1.000	0	0	0.004	0.016	0.980	0	0
	100	40	0	0	1.000	0	0	0	0.012	0.988	0	0
	100	80	0	0	1.000	0	0	0	0.008	0.992	0	0

to determine the number of unobserved nonstationary factors. Based on 500 replications for each DGP, Table 1 displays the probability that a particular factor size from 0 to 5 is selected according to the information criteria proposed for both differenced data and level data. In differenced data, when $T = 40$, the probabilities are more than 99% in all cases and tend to unit when $T = 80$ for selecting the total number of unobserved factors. The information criterion for level data performs as good as that in difference data when $T = 80$. When T=40, the probabilities are at least 90% in all cases. The simulation results show that our two information criteria in both differenced data and level one works fairly well.

2.4.3 Determine the number of groups

The results from previous subsection show that the information criteria are useful even though we have no information of latent group structures. This section focuses on the performance of the information criterion for determining the number of groups, where we assume that the number of unobserved factors are known. Here the penalty function $\rho(N, T) = \frac{1}{3} \times \log(\min(N, T))/\min(N, T)$, which satisfies the two restrictions proposed in Theorem 3.9. Due to space limitations, we report outcomes under the tuning parameter $\lambda = c_\lambda \times T^{-3/4}$, where $c_\lambda = 0.1$. Based on 500 replications for each DGP, Table 2 displays the probability that a particular group size from 1 to 6 is selected according to

Table 9: Frequency for selecting K=1,2,...,6 groups

	N	T	1	2	3	4	5	6
DGP1	50	40	0	0	0.9860	0.0140	0	0
	50	80	0	0	0.9940	0.0060	0	0
	100	40	0	0	0.9700	0.0280	0	0.0020
	100	80	0	0	1.0000	0	0	0
DGP2	50	40	0	0	1.0000	0	0	0
	50	80	0	0	1.0000	0	0	0
	100	40	0	0	1.0000	0	0	0
	100	80	0	0	1.0000	0	0	0
DGP3	50	40	0	0	0.9760	0.0180	0.0060	0
	50	80	0	0	0.9980	0.0020	0	0
	100	40	0	0	0.9740	0.0240	0.0020	0
	100	80	0	0	1.0000	0	0	0
DGP4	50	40	0	0	0.9920	0.0060	0.0020	0
	50	80	0	0	1.0000	0	0	0
	100	40	0	0	0.9900	0.0100	0	0
	100	80	0	0	1.0000	0	0	0

the information criterion. The true number of group is 3. When $N = 50$ the probabilities are more than 99% in all cases and tend to unit when T=80.

2.4.4 Classification and point estimation

In this subsection, we test the performance of classification and estimation when we have prior knowledge of the number of groups and that of unobserved factors. Table 3 and Table 4 report classification and point estimation results from 500 replications for each DGP. As shown in Table 3 and Table 4, we set the tuning parameter in the objective function (23) $\lambda = c_\lambda \times T^{-3/4}$ and choose a sequence of increasing constants of $c_\lambda = (0.025, 0.05, 0.1, 0.2)$[4] to test the sensitivity of classification and estimation performance. Here we only report the performance results for the first coefficient

[4] Due to space limitation, we only report the results when $c_\lambda = (0.1, 0.2)$. The rest results are available upon request.

$\alpha_1 = \{\alpha_{1,k}\}_{k=1}^{K_0}$ in each model. In general, the outcomes are found robust over specified range of constants. Column 4 and 7 report the percentage of correct classification of the N units, calculated as $\frac{1}{N}\sum_{k=1}^{K_0}\sum_{i\in\hat{G}_k} 1\{\beta_i^0 = \alpha_k^0\}$, averaged over the 500 replications. Column 5-6 and 8-9 summarize the estimation performance, such as root-mean-squared error (hereafter, RMSE), and bias. For simplicity we define weighted average RMSE and bias, as $\frac{1}{N}\sum_{k=1}^{K_0} N_k$RMSE$(\hat{\alpha}_{1,k})$ with $\hat{\alpha}_{1,k}$ the same as bias. The estimate of the long-run covariance matrix is based on Fejer kernel with bandwidth set at 10. Results of other kernels (quadratic spectral kernel and Parzen kernel) are not reported, there are no essential differences for most cases. For comparison purpose, we report the results of corresponding statistics of the C-Lasso, bias-corrected post-Lasso, fully-modified post-Lasso, Cup-Lasso, and oracle estimators. The oracle estimator utilizes the exact group identity G_k^0, which is infeasible in practice.

For classification results, the correct classification percentage approaches 100% when T increases. The results with different c_λ's are quite similar, indicating the robustness of our algorithm to the choice of tuning parameter. In particular, we iteratively minimize the PPC objective function to obtain the Cup-Lasso estimators. The correct classification percentage is higher than that of C-Lasso and post-Lasso estimators in all cases. For estimation performance, the RMSE, bias, and coverage of post-Lasso and Cup-Lasso estimators approach that of oracle ones in DGP1. Since we only introduce stationary factors and strictly exogenous nonstationary regressors, there is no asymptotic bias coming from the endogeneity and serial correlation. The RMSE and coverage of C-Lasso estimators are poor due to ignoring the unobserved stationary factors in PPC-based estimation procedure. In DGP2 and DGP3, the performance of C-Lasso estimator is poorer due to the additional sources of non-negligible bias from the endogeneity and serial correlation. And we show that the fully modified procedure work better compared to direct bias-correction procedure. The performance of Cup-Lasso estimators is better than that of post-Lasso ones due to updated group classification results. In general, the finite sample performance of the Cup-Lasso estimators is close to that of the oracle ones, which empirically confirms oracle efficiency of the Cup-Lasso estimators. In practice, we recommend Cup-Lasso estimators for estimation and inference.

Table 10: Classification and point estimation of α_1 for DGP1 and DGP2

		c_λ	0.1			0.2		
N	T		% Correct specification	RMSE	Bias	% Correct specification	RMSE	Bias
DGP1								
50	40	C-Lasso	99.68	0.0137	0.0049	99.70	0.0130	0.0047
50	40	post^{bc}-Lasso	99.68	0.0130	0.0003	99.70	0.0129	0.0002
50	40	post^{fm}-Lasso	99.68	0.0129	0.0004	99.70	0.0128	0.0003
50	40	Cup-Lasso	99.68	0.0126	-0.0002	99.70	0.0126	-0.0002
50	40	Oracle	-	0.0126	-0.0002	-	0.0126	-0.0002
50	80	C-Lasso	100	0.0081	0.0031	100	0.0077	0.0028
50	80	post^{bc}-Lasso	100	0.0070	0.0003	100	0.0070	0.0003
50	80	post^{fm}-Lasso	100	0.0069	0.0004	100	0.0069	0.0004
50	80	Cup-Lasso	100	0.0069	0.0004	100	0.0069	0.0004
50	80	Oracle	-	0.0069	0.0001	-	0.0069	0.0001
100	40	C-Lasso	99.69	0.0109	0.0054	99.73	0.0101	0.0046
100	40	post^{bc}-Lasso	99.69	0.0091	0.0007	99.73	0.0087	0.0004
100	40	post^{fm}-Lasso	99.69	0.0090	0.0007	99.73	0.0086	0.0004
100	40	Cup-Lasso	99.69	0.0090	0.0007	99.73	0.0086	0.0004
100	40	Oracle	-	0.0087	-0.0001	-	0.0087	-0.0001
100	80	C-Lasso	100	0.0062	0.0032	99.99	0.0058	0.0029
100	80	post^{bc}-Lasso	100	0.0046	0.0005	99.99	0.0046	0.0005
100	80	post^{fm}-Lasso	100	0.0046	0.0005	99.99	0.0046	0.0005
100	80	Cup-Lasso	100	0.0046	0.0005	99.99	0.0046	0.0005
100	80	Oracle	-	0.0046	0.0004	-	0.0046	0.0004
DGP2								
50	40	C-Lasso	97.68	0.0654	0.0146	97.53	0.0743	0.0146
50	40	post^{bc}-Lasso	97.68	0.0405	0.0048	97.53	0.0430	0.0048
50	40	post^{fm}-Lasso	97.68	0.0405	0.0042	97.53	0.0430	0.0041
50	40	Cup-Lasso	100	0.0094	0.0004	100	0.0094	0.0004
50	40	Oracle	-	0.0094	0.0004	-	0.0094	0.0004
50	80	C-Lasso	99.21	0.0233	0.0047	99.19	0.0254	0.0047
50	80	post^{bc}-Lasso	99.21	0.0195	-0.0004	99.19	0.0195	-0.0007
50	80	post^{fm}-Lasso	99.21	0.0194	-0.0005	99.19	0.0194	-0.0009
50	80	Cup-Lasso	100	0.0047	-0.0001	100	0.0047	-0.0001
50	80	Oracle	-	0.0047	-0.0001	-	0.0047	-0.0001
100	40	C-Lasso	97.45	0.0500	0.0135	97.37	0.0543	0.0119
100	40	post^{bc}-Lasso	97.45	0.0601	-0.0011	97.37	0.0584	-0.0010
100	40	post^{fm}-Lasso	97.45	0.0601	-0.0016	97.37	0.0585	-0.0015
100	40	Cup-Lasso	100	0.0069	-0.0016	100	0.0069	-0.0016
100	40	Oracle	-	0.0069	-0.0016	-	0.0069	-0.0016
100	80	C-Lasso	99.25	0.0181	0.0061	99.23	0.0194	0.0057
100	80	post^{bc}-Lasso	99.25	0.0172	0.0012	99.23	0.0170	0.0010
100	80	post^{fm}-Lasso	99.25	0.0171	0.0010	99.23	0.0170	0.0010
100	80	Cup-Lasso	100	0.0032	-0.0001	100	0.0032	-0.0001
100	80	Oracle	-	0.0032	-0.0001	-	0.0032	-0.0001

Table 11: Classification and point estimation of α_1 for DGP3 and DGP4

		c_λ	0.1			0.2		
N	T		% Correct specification	RMSE	Bias	% Correct specification	RMSE	Bias
DGP3								
50	40	C-Lasso	96.97	0.0563	0.0118	96.87	0.0632	0.0101
50	40	post^{bc}-Lasso	96.97	0.0522	0.0029	96.87	0.0516	0.0022
50	40	post^{fm}-Lasso	96.97	0.0524	0.0023	96.87	0.0519	0.0016
50	40	Cup-Lasso	99.85	0.0145	0.0015	99.81	0.0146	0.0015
50	40	Oracle	-	0.0150	0.0014	-	0.0150	0.0014
50	80	C-Lasso	99.15	0.0297	0.0056	99.11	0.0327	0.0047
50	80	post^{bc}-Lasso	99.15	0.0275	0.0015	99.11	0.0265	0.0013
50	80	post^{fm}-Lasso	99.15	0.0274	0.0015	99.11	0.0265	0.0013
50	80	Cup-Lasso	100	0.0073	0.0010	100	0.0073	0.0010
50	80	Oracle	-	0.0073	0.0006	-	0.0073	0.0006
100	40	C-Lasso	98.65	0.0299	0.0119	98.43	0.0300	0.0110
100	40	post^{bc}-Lasso	98.65	0.0214	0.0028	98.43	0.0222	0.0035
100	40	post^{fm}-Lasso	98.65	0.0213	0.0023	98.43	0.0222	0.0031
100	40	Cup-Lasso	99.93	0.0108	0.0020	99.83	0.0110	0.0021
100	40	Oracle	-	0.0109	0.0018	-	0.0109	0.0018
100	80	C-Lasso	99.05	0.0194	0.0060	99.01	0.0208	0.0053
100	80	post^{bc}-Lasso	99.05	0.0181	0.0007	99.01	0.0183	0.0007
100	80	post^{fm}-Lasso	99.05	0.0180	0.0006	99.01	0.0182	0.0005
100	80	Cup-Lasso	100	0.0054	-0.0002	100	0.0054	-0.0002
100	80	Oracle	-	0.0054	-0.0003	-	0.0054	-0.0003
DGP4								
50	40	C-Lasso	96.92	0.0566	0.0110	96.77	0.0634	0.0099
50	40	post^{bc}-Lasso	96.92	0.0508	0.0018	96.77	0.0498	0.0008
50	40	post^{fm}-Lasso	96.92	0.0511	0.0013	96.77	0.0501	0.0008
50	40	Cup-Lasso	99.91	0.0130	0.0014	99.87	0.0130	0.0015
50	40	Oracle	-	0.0134	0.0014	-	0.0134	0.0014
50	80	C-Lasso	98.99	0.0299	0.0055	98.93	0.0331	0.0045
50	80	post^{bc}-Lasso	98.99	0.0277	0.0009	98.93	0.0263	0.0013
50	80	post^{fm}-Lasso	98.99	0.0277	0.0008	98.93	0.0263	0.0013
50	80	Cup-Lasso	100	0.0066	0.0010	100	0.0066	0.0010
50	80	Oracle	-	0.0065	0.0007	-	0.0065	0.0007
100	40	C-Lasso	98.77	0.0291	0.0123	98.53	0.0295	0.0113
100	40	post^{bc}-Lasso	98.77	0.0205	0.0032	98.53	0.0217	0.0037
100	40	post^{fm}-Lasso	98.77	0.0204	0.0027	98.53	0.0216	0.0032
100	40	Cup-Lasso	99.94	0.0102	0.0020	99.87	0.0103	0.0021
100	40	Oracle	-	0.0103	0.0017	-	0.0103	0.0017
100	80	C-Lasso	99.04	0.0197	0.0059	99.02	0.0211	0.0053
100	80	post^{bc}-Lasso	99.04	0.0181	0.0009	99.02	0.0183	0.0008
100	80	post^{fm}-Lasso	99.04	0.0180	0.0007	99.02	0.0183	0.0007
100	80	Cup-Lasso	100	0.0050	-0.0002	100	0.0050	-0.0002
100	80	Oracle	-	0.0050	-0.0002	-	0.0050	-0.0002

2.5 Empirical Application: Growth Convergence Puzzle

Many researchers have explored the behavior of economic growth across multiple countries. The main question in this literature is whether economies exhibit convergence. Based on our method, we study the heterogeneous behavior of convergence through the channel of technology diffusions. Our benchmark model is the international R&D spillovers model by Coe and Helpman (1995). Their work empirically identifies the positive technology spillovers based on country level data. Since technology change is the main source of economies' growth, the positive R&D spillovers are some forces toward growth convergence through the channel of technology catch-up effects. There are two potential problems in their work. First, they fail to identify two distinct types "spillovers" effects–positive technology spillovers and negative market rivalry effects, see Bloom et al. (2013). In addition, they haven't account for unobserved common patterns across countries, such as financial crisis shocks and technology trends.

Our PPC-based estimation method provides a purely data-driven approach to simultaneously identify the heterogeneous behavior in international R&D spillovers and consider the unobserved global patterns. In particular, we impose the latent group structures on the long-run relationships between technology changes, domestic R&D stock, and foreign R&D stock. Moreover, the heterogeneous long-run estimates on foreign R&D stock empirically indicate two directions of R&D spillovers–positive technology spillovers and negative market rivalry effects, which explain the economic convergence puzzle through the channel of technology growth.

2.5.1 International R&D spillover model

In the section, we introduce two linear specifications on the international R&D spillover model. Following standard growth literature, we define TFP as the Solow residual, which is often regarded as a measure of technology changes. That is,

$$\log(TFP) = \log(Y) - \theta \log(K) - (1-\theta)\log(L), \tag{40}$$

where Y is the final output, L is the labor force, K is the capital stock, and θ is the share of capital in GDP. On the one hand, domestic R&D investment is one of the main sources of technology changes by stimulating innovation. On the other hand, researchers believe that international trade in intermediate goods enables a country to gain access to all inputs available in the rest of the world. In this aspect, the foreign R&D stocks from a country's trading partners also affect this country's TFP by enhancing R&D transfer. In Coe and Helpman (1995), they empirically identify two sources of technology growth–innovation and catch-up effects by regressing following equation

$$\log(F_{it}) = \mu_i + \beta^d \log(s^d_{it}) + \beta^f \log(s^f_{it}) + u_{it},$$

where i is the country index, F is the total factor productivity, s^d is the real domestic R&D capital stock, s^f is the real foreign R&D capital stock. We follow their specification on the international R&D spillovers model and introduce unobserved common patterns, such that

$$\log(F_{it}) = \beta^d_i \log(s^d_{it}) + \beta^f_i \log(s^f_{it}) + \lambda_i' f_t + u_{it}, \tag{Eq1}$$

where f_t may stand for unobserved technology trends. The fixed effects μ_i are captured by the factor structure. Here we impose latent group structures on (β^d_i, β^f_i). In particular, the latent group structures on β^f_i allow us to study the two types of spillover effects–positive technology spillovers and negative market rivalry effects, respectively.

In addition, we consider logarithm of human capital (H) as an additional explanatory variable, see (Eq2)

$$\log(F_{it}) = \beta^d_i \log(s^d_{it}) + \beta^f_i \log(s^f_{it}) + \beta^h_i \log(h) + \lambda_i' f_t + u_{it}. \tag{Eq2}$$

The human capital accounts for innovation outside the R&D sector and other aspects of human capital not captured by formal R&D. Engelbrecht (1997) suggests that human capital is found to affect TFP directly as a factor of production, and as a channel for international technology diffusion associated with catch-up effects across countries.

We obtain Coe et al. (2009) (CH2009, hereafter) datasets from 1971-2004 for 24 OECD countries. The bilateral import weighted R&D S^{f-biw} from trading partners is a measure of foreign R&D stock. Human capital is measured by year of schooling. See Coe and Helpman's appendix for detailed definitions and constructions of these variables.

Table 12: The information criterion for K_0 (Eq1 & Eq2)

Eq1					
K/c_λ	0.05	0.1	0.2	0.4	0.6
$K=1$	-4.6315	-4.6584	-4.6812	-4.6834	-4.6794
$K=2$	-4.8073	-4.8760	-4.7356	-4.8319	-4.8332
$K=3$	**-5.0084**	**-5.0942**	**-5.2130**	**-5.2221**	**-5.0992**
$K=4$	-4.8985	-4.9708	-5.0092	-4.6353	-4.9279
$K=5$	-4.8598	-4.8240	-4.4272	-4.9821	-4.8042
$K=6$	-4.4159	-4.2700	-3.6774	-4.8858	-4.6118
Eq2					
K/c_λ	0.05	0.1	0.2	0.4	0.6
$K=1$	**-4.6011**	-4.6311	-4.6845	-4.6876	-4.6889
$K=2$	-4.5674	**-4.8101**	**-4.8693**	**-4.8138**	**-4.8127**
$K=3$	-3?9180	-4.2002	-4.7259	-4.7467	-4.7045
$K=4$	-2.8630	-3.5698	-4.0314	-4.2412	-4.2497
$K=5$	-2.2351	-4.0434	-1.9373	-3.5935	-4.0737
$K=6$	-2.7073	-3.6627	-3.1292	-3.7489	-2.6413

2.5.2 Estimation results

Before the PPC-based estimation procedure, we first employ information criteria in Section 3.5 to estimate the number of unobserved factors. We set penalty function as $g_1(N,T) = \frac{N+T}{NT}\log(\min(N,T))$ and $g_2(N,T) = \frac{T}{4\log\log T} \times g_1(N,T)$. The results for both differenced and level data indicate one unobserved nonstationary common factors. We fix $r = r_1 = 1$ in the determination of number groups and the PPC-based estimation procedure.

We set $\rho(N,T) = \frac{2}{3} \times \log(\min(N,T))/\min(N,T)$ and tuning variable $\lambda = c_\lambda \times T^{-3/4}$ where $c_\lambda = (0.05, 0.1, 0.2, 0.4, 0.6)$. Table 5 reports the information criterion as a function of the number of groups under these tuning parameters. The information criterion suggests three groups for (Eq1) and two groups for (Eq2). In our estimation, we first set the number of groups and then specify $c_\lambda = 0.2$, where the information criterion achieves the minimal values.

Table 6 reports the main results of pooled FMOLS and Cup-Lasso estimates with one unobserved nonstationary common factors, where we compare our results to CH2009. In (Eq1), we have two explanatory variables $(\log(S^d), \log(S^{f-biw}))$. First, we compare the result of CH2009 with the pooled FMOLS after controlling cross-sectional dependence. The coefficients of $\log(S^d)$ in CH2009 is qualitatively similar to our pooled FMOLS. In addition, we observe the slope coefficient of foreign R&D stock decrease more than half after incorporating one unobserved nonstationary common factors. Since the nonstationary common factor may stand for the unobserved global trends, it is reasonable that the direct spillovers effects decrease when the unobserved global technology patterns are taken into consideration. Second, we identify quite difference behavior in the group-specific Cup-Lasso estimates. The estimates of group 1 have the largest estimate on the domestic R&D stock but negative one on foreign R&D. For group 2 and group 3, they both have positive estimates on domestic R&D stock and foreign one. In particular, both estimates in group 2 are larger than that of group 3.

We summarize the estimation results into three aspects. First, we observe the negative market rivalry effects dominate technology spillovers for countries in group 1. Therefore, the technology changes of those countries rely mainly on innovations from domestic R&D stock. In addition, it implies that countries in group 1 don't favor convergence through the channel of technology changes. We call it as "Divergence"group. On the contrary, technology changes of countries in group 2 have balanced sources–innovation effects from domestic R&D stock and catch-up effects from technology spillovers. Moreover, the magnitudes of those estimates are similar. In this perspective, countries in group 2 favor growth convergence hypothesis. Here we refer it as "Balance"group. Lastly, the technology changes in group 3 are mainly determined by foreign R&D stock. They are classified as "Convergence"group.

In (Eq2), we introduce an additional regressor–human capital, which is regarded as a direct source of technology changes. Our results first confirm that human capital is the one of the main sources of productivity growth. In general, similar heterogeneous behavior preserves in (Eq2). First, we can still classify countries into two groups and define them as "Balance "and "Convergence ". For group 1, the innovations and catch-up effects have similar magnitudes. For group 2, referred as "Convergence ", they have

Table 13: PPC estimation results for (Eq1) and (Eq2)

Eq1					
Slope coefficients	Pooled	Pooled	Group 1	Group 2	Group 3
	CH2009	FM-OLS	Cup-Lasso	Cup-Lasso	Cup-Lasso
$\log(S^d)$	0.095***	0.090***	0.301***	0.107***	0.051***
	(0.0053)	(0.0134)	(0.0295)	(0.0248)	(0.0144)
$log(S^{f-biw})$	0.213***	0.092***	-0.124***	0.193***	0.148***
	(0.0136)	(0.0222)	(0.0306)	(0.0559)	(0.0342)
Eq2					
Slope coefficients	Pooled	Pooled	Group 1	Group 2	
	CH2009	FM-OLS	Post-Lasso	Post-Lasso	
$\log(S^d)$	0.098***	0.049***	0.071***	-0.098***	
	(0.0160)	(0.0163)	(0.0174)	(0.0270)	
$log(S^{f-biw})$	0.035***	0.132***	0.063**	0.323***	
	(0.0111)	(0.0316)	(0.0332)	(0.0398)	
$\log(h)$	0.725***	0.644***	0.638***	0.680***	
	(0.0870)	(0.1204)	(0.1302)	(0.1791)	

Note: *** 1% significant; ** 5% significant; * 10% significant.

Table 14: Group classification results of Eq1 and Eq2

Eq1				
Group 1 “Divergence” ($N_1 = 7$)				
Austria	Denmark	France	Germany	New Zealand
Norway	United States			
Group 2 “Balance” ($N_2 = 7$)				
Canada	Ireland	Israel	South Korea	Netherlands
Portugal	United Kingdom			
Group 3 “Convergence” ($N_3 = 10$)				
Australia	Belgium	Finland	Greece	Iceland
Italy	Japan	Spain	Sweden	Switzerland
Eq2				
Group 1 “Balance ” ($N_1 = 18$)				
Austria	Belgium	Finland	France	Germany
Iceland	Ireland	Israel	Italy	Japan
South Korea	Netherlands	New Zealand	Portugal	Spain
Sweden	Switzerland	United States		
Group 2 “Convergence” ($N_2 = 6$)				
Australia	Canada	Denmark	Greece	Norway
United Kingdom				

significant positive estimates on foreign R&D stocks.

The PPC-based estimation procedure simultaneously determine the group identities and estimate parameters. Table 7 reports the group classification results. We summarize several interesting findings. First, there are typically two types of countries in “Divergence” group–“Leader” and “Loser”. Countries like France, Germany, United States are already at technology frontiers in global, which own 61.1% proportion of global R&D stock. On the contrary, the rest countries in group 1 only accounts for 1.5% proportion of global R&D stock. Second, we notice that most OECD countries are classified into group 2 and group 3. It confirms the recent work of KELLER (2004) that the major sources of technical changes leading to productivity growth in OECD countries are not domestic, instead, they lie aboard through the channel of international technology diffusions. Furthermore, countries like Israel, South Korea and United Kingdom are classified into “Balance” group. Productivity growth for them relies on both innovation and catch-up effects.

Overall, we re-estimate Coe and Helpman's model by both pooled FMOLS and group-specific PPC-based method with one unobserved global trend. Our pooled FMOLS confirms the international R&D spillovers in global after considering unobserved global trend. In addition, our Cup-Lasso estimates show heterogeneous behavior in both innovations and catch-up effects. To the best of our knowledge, it is the first paper to empirically identifies two types of spillovers–positive technology spillovers and negative market rivalry effects in country level. Further, these facts empirically build a connection between the "Club convergence" theory (Quah (1996), Quah (1997)) and the conditional convergence model (Barro and Sala-i Martin (1997)). Consequently, economic growth patterns do vary across countries– some exhibit convergence pattern but some fail.

2.6 Conclusion

The main contribution of this paper is to propose a novel approach that handle the unobserved heterogeneity and cross-sectional dependence in nonstationary panel models. We assume that cross-sectional dependence is captured by unobserved common factors, which allow for stationary and nonstationary ones. In general, the penalized least square estimators are inconsistent due to a spurious regression problem induced by unobserved nonstationary factors. We propose a iterative penalized principal component method, which provides consistent and efficient estimators for long-run cointegration relationships under cross-sectional dependence. The Lasso-type estimators have a mixed normal asymptotic distribution after bias-correction. This property facilitates the use of usual t, Wald and F statistics for inference and hypothesis testing. In the empirical application, we offer an explanation for growth convergence puzzle through heterogeneous behavior in R&D spillovers.

3 Panel Error Correction Models with Unobserved Heterogeneity and Dynamic Common Factors

3.1 Introduction

It is widely observed that macroeconomic and financial panel data exhibits two stylized facts–co-movements and heterogeneity. The co-movements appear important enough to account for long-run cointegration relationships among variables and to capture short-run dynamics driven by unobserved stationary common factors. For example, monetary policies are always built on a principal belief of the long-run equilibrium between GDP, inflation and interest rate. In addition, central banks consider short-run impacts of common factors, such as oil price shocks or financial crises. Another challenge of panel methodology is to control for the unobserved heterogeneity in incidental parameters and structural parameters. This paper focuses on complications in the estimation of panel error correction models with both long-run and short-run co-movements. Besides, we allow for unobserved heterogeneity in long-run cointegration vectors. Furthermore, our method maintains several advantages of panel data, such as capturing the complexity of economic behavior, simplifying statistic inference and improving estimation efficiency.

In this paper, we consider a panel error-correction model with unobserved stationary common factors. In particular, we impose latent group structures on the long-run cointegration vectors. We first develop a novel econometric procedure that obtains efficient estimators for long-run cointegration vectors in the presence of unobserved heterogeneity. Then, we propose two Lasso-type estimators for long-run cointegration vectors–*Classifier-Lasso* and *post-Lasso*. Due to the stationary property of common factors, we can still obtain consistent estimators for long-run cointegration vectors when ignoring unobserved short-run co-movements. Afterward, we employ the principal components method to estimate unobserved stationary common factors. The presence of unobserved stationary common factors introduces a non-negligible bias in Lasso-type estimators and affects covariance structures in limiting distributions. The C-Lasso estimators are based on the minimization of a penalized generalized least squares (PGLS, hereafter) criterion function in which additive penalty terms (see Su et al. (2016), Li et al. (2016), and Qian and Su (2016)) are involved in parameters of slope coefficients.

We employ the GLS objective function for the sake of three advantages. First, the GLS estimator tends to be much more reliable than the maximum likelihood estimator (see Johansen (1991)) in small samples. Second, the GLS objective function is constructed on the triangular representation system with a prior identification condition $\beta_i = [I_r, b_i]$ (Phillips (1994)). This fact simplifies the statistic inference from a parametric approach. Third, the PGLS-based Lasso-type estimators achieve efficiency, compared to maximum likelihood ones. It reduces the number of long-run vectors from $N \times (J-r)J$ to $K \times (J-r)J$ in which the number of groups K is much smaller than that of cross-sectional units N.

Our theoretical results first show the Granger partial sum representation of nonstationary dependent vectors. Then we establish consistency for initial GLS estimators of long-run cointegration vectors in the presence of unobserved common factors. In addition, we show that the estimator of short-run adjustment matrix is inconsistent due to weak dependence in innovation processes. Furthermore, we can show that the spaces spanned by $\hat{F}$ and F^0 are asymptotically the same by similar arguments in Bai (2009). The initial results are obtained under large T asymptotics. In the asymptotic theory for the Lasso-type estimators, we establish the point consistency and classification consistency for the PGLS-based estimation. The latter property indicates that all individuals are classified into correct groups with a probability approaching to one (w.p.a.1). Moreover, we establish asymptotic normal distributions for our estimators. Throughout this paper, we assume the number of factors and the number of groups are known. In real economic applications, we can directly apply the information criteria proposed in Bai and Ng (2002) and Huang et al. (2017) to determine the number of factors and the number of groups, respectively. In terms of simulation, we show good finite sample performance for both estimation and classification.

Based on our method, we study price discovery in market microstructure model. We first propose a permanent-transitory (PT, hereafter) decomposition based on long-run cointegration vectors. Our PT decomposition suggests that efficient price is weighted average of bid and ask quotes. Then we apply our method to estimate the long-run cointegration relationships between bid and ask quotes by NYSE Trade and Quote (TAQ) data in year 2006. In general, our results suggests that the long-run equilibrium between bid and ask prices deviates from the one-to-one relationship, which is assumed by Hasbrouck (1995). Further, we observe asymmetric contributions of bid and ask quotes to price discovery in stock market. Consequently, the mid-point of bid and ask prices may not be a good proxy for efficient price in some cases.

The main contribution of this paper is to obtain efficient and simple estimators of long-run vectors with unobserved heterogeneity. Formally, we impose latent group structures on long-run cointegration vectors, where the long-run relationships are homogeneous within each group and heterogeneous across groups. The propensity to categorize objects is an integral part of human nature. For example, economists tend to analyze causal relationships among economic variables by subsamples, such as developing versus developed countries, and value versus growth stocks. These classifications are just some of the possible dichotomies in social economic problems. In the meantime, there are some unobserved patterns in economic data, which cannot be easily identified from observed characteristics. Therefore, the latent group structures, proposed by Su et al. (2016) allow the data speak freely to the unobserved heterogeneity in the parameters of slope coefficients and maintain simple and efficient estimators.

Second, our paper speak to a long literature on error-correction model since the seminal work by Engle and Granger (1987), such as Phillips's triangular system representation (Phillips (1991),Phillips (1994)) and Johansen's maximum likelihood method (Johansen (1988),Johansen (1991)). Recently, many work extends these two methods into panel framework under different restrictions (see Larsson et al. (2001), Groen and Kleibergen (2003), Larsson and Lyhagen (2007), Breitung (2005) and Brüggemann and Lütkepohl (2005))). Apart from two main approaches, Pesaran et al. (2004) and Dees et al. (2007) introduce a Global VAR model, which focuses on an individual-specific error correction model and allows for interdependence among individuals. Their estimators are inefficient and may be inconsistent under rank deficient case (see). In general, we obtain efficient estimators of the long-run cointegration vectors under more general settings.

Third, the detection of short-run co-movements sheds light on factor models. There is a growing list of econometric models employing multifactor structures in dependent variables and error terms. The factor-augmented VAR model (see Stock and Watson (2002) and Bernanke et al. (2005)) employs a small number of estimated factors to effectively summarize large amounts of information about the economy. Their methods provide a natural solution to the curse of dimensionality problem in VAR analyses and achieve better predication performance. Banerjee et al. (2017), and Barigozzi and Luciani (2017) extend the factor-augmented model by introducing error correction terms, which allow for nonstationary data. Compared to existing work, our primary interest is the estimation of long-run cointegration vectors under multifactor error structures. In the meanwhile, we can consistently estimate unobserved common factors by principal components (see Bai et al. (2009)).

This paper is structured as follows. Section 2 introduces a panel error correction model with stationary common factors and latent group structures and proposes a penalized generalized least squares objective function for estimation. Section 3 explains main assumptions and establishes asymptotic properties of the proposed regularized estimators. Section 4 reports simulation results. Section 5 studies efficient price in market microstructure model. Section 6 concludes. All proofs are relegated to the appendix.

3.2 Model and Estimation

3.2.1 Heterogeneous panel error correction model with dynamic common factors

In this section, we describe the main features of our panel error correction model with latent group structures and unobserved common factors. We consider a panel dataset consisting of N cross-section units (individuals) over T time periods. For each individual $i = 1, ..., N$, the $J-$dimensional y_{it} is generated as follows

$$y_{it} = s_{it} + \psi_i^{0\prime} f_t^0, \tag{41}$$

where s_{it} is a $J \times 1$ vector of unobserved idiosyncratic nonstationary component, f_t^0 is an $m \times 1$ vector of unobserved stationary common factors, and ψ_i^0 is an $m \times J$ matrix of factor loadings. For each individual $i = 1, ..., N$, we assume that s_{it} satisfies an error correction model, such that

$$\Delta s_{it} = \alpha_i^0 \beta_i^{0\prime} s_{it-1} + \sum_{l=1}^{p-1} \Gamma_{il}^0 \Delta s_{it-l} + \varepsilon_{it}, \tag{42}$$

where β_i^0 is a $J \times r$ matrix of long-run cointegration vectors, α_i^0 is a $J \times r$ matrix of adjustment parameters, r is the cointegration rank, Γ_{il}^0 is a $J \times J$ full rank matrix of short-run dynamics parameters, ε_{it} is the idiosyncratic error term with zero mean and finite variance. If $r = J$, s_{it} is a full rank $I(1)$ process. If $r = 0$, model (42) reduces to stationary VAR$(p-1)$ models for differenced data Δs_{it}. Since our focus is on the reduced rank case, it is appropriate to assume $1 \leqslant r < J$.

Combine (41) and (42), the observed process y_{it} follows a heterogeneous panel error correction model with unobserved dynamic common factors for all $i = 1, ..., N$,

$$\Delta y_{it} = \alpha_i^0 \beta_i{}^{0\prime} y_{it-1} + \sum_{l=1}^{p-1} \Gamma_{il}^0 \Delta y_{i,t-l} + \sum_{l=0}^{p} \Lambda_{il}^{0\prime} f_{t-l}^0 + \varepsilon_{it}. \tag{43}$$

Define the static factor $F_t^0 = [f_t^{0\prime}, f_{t-1}^{0\prime}, ..., f_{t-p}^{0\prime}]'$ and its factor loadings $\Lambda_i^0 = [\Lambda_{i0}^0, \Lambda_{i1}^0, ..., \Lambda_{ip}^0]'$. Then model (43) can be re-written as follows

$$\Delta y_{it} = \alpha_i^0 \beta_i^{0\prime} y_{it-1} + \sum_{l=1}^{p-1} \Gamma_{il}^0 \Delta y_{i,t-l} + \Lambda_i^{0\prime} F_t^0 + \varepsilon_{it} \tag{44}$$

where the dimension of static factor F_t is $M \times 1$ where $m \leqslant M \leqslant m(p+1)$. In this panel error correction model, β_i is a long-run cointegration matrix, which summarizes long-run co-movements among variables, and F_t is a vector of unobserved common factors, which stands for short-run co-movements across individuals.

In order to maintain efficiency, we assume that the long-run cointegration matrices β_i exhibit certain unobserved *group* patterns, where they are heterogeneous across groups and homogeneous within a group. Our interests are to obtain efficient estimators for long-run cointegration vectors, infer latent group identity and detect unobserved common factors F_t.

The estimation procedures are performed on model (44). Let

$$\text{vec}(\alpha) \equiv (\text{vec}(\alpha_1), ..., \ \text{vec}(\alpha_N)), \ \text{vec}(\beta) \equiv (\text{vec}(\beta_1), ..., \text{vec}(\beta_N)), \ \Lambda = (\Lambda_1, ..., \Lambda_N),$$

and $F = (F_1, ..., F_T)'$. The true values of $\alpha_i, \beta_i, \Lambda_i$, and F_t are denoted as $\alpha_i^0, \beta_i^0, \Lambda_i^0$ and F_t^0, respectively.

3.2.2 Estimation

In this section, we propose two Lasso-type estimators for the long-run cointegration matrix β_i. We first obtain consistent initial estimators, which is briefly discussed in the following section.

Initial estimation: Two-step GLS method

For simplicity, we assume that the number of lags $p = 1$ and the case when $p > 1$ can be easily extended. Thus, model (44) becomes

$$\Delta y_{it} = \alpha_i^0 \beta_i^{0\prime} y_{i,t-1} + \Lambda_i^{0\prime} F_t^0 + \varepsilon_{it} \tag{45}$$

where ε_{it} is assumed to be cross-sectionally independent but allowed to be weakly dependent over time, and F_t contains dynamics of unobserved common factors, which is also allowed to be weakly dependent. By Cheng and Phillips (2009), we note that the Johansen's maximum likelihood estimators have non-negligible biases in both long-run cointegration matrix β_i and short-run adjustment matrix α_i under weakly dependent innovation processes. But the consistency of β_i still holds.

At the first stage, we obtain preliminary estimators from the Johansen's optimization problem as follows

$$\begin{aligned}
\hat{\beta}_i^{ML} &= \arg\min_{\beta_i} \left| S_{00,i} - S_{01,i}\beta_i \left(\beta_i' S_{11,i}\beta_i\right)^{-1} \beta_i' S_{10,i} \right| \text{ for } i = 1, ..., N, \\
\hat{\alpha}_i^{ML} &= S_{01,i}\hat{\beta}_i^{ML} \left(\hat{\beta}_i^{ML\prime} S_{11,i}\hat{\beta}_i^{ML}\right)^{-1},
\end{aligned}$$

subject to the normalization conditions which ensure the uniqueness of $\hat{\beta}_i^{ML}$:

$$\hat{\beta}_i^{ML\prime} S_{11,i}\hat{\beta}_i^{ML} = I_r \text{ and } \hat{\beta}_i^{ML\prime} S_{10,i} S_{00,i}^{-1} S_{01,i}\hat{\beta}_i^{ML} = \hat{\Lambda}_i^r,$$

where $\hat{\Lambda}_i^r = \text{diag}\left(\hat{\lambda}_i^1, \hat{\lambda}_i^2, ..., \hat{\lambda}_i^r\right)$ and $\hat{\lambda}_i^1 \geqslant \cdots \geqslant \hat{\lambda}_i^r \geqslant 0$ are the first r ordered roots of the determinantal equation $\left|\Lambda_i^r S_{11,i} - S_{10,i} S_{00,i}^{-1} S_{01,i}\right| = 0$ for each i. The $S_{hl,i}$ are moment matrices of Δy_{it} ($h = 0$) and $y_{i,t-1}$ ($l = 1$), such as $S_{01,i} = \frac{1}{T}\sum_{t=1}^T \Delta y_{it} y_{i,t-1}'$ and $S_{11,i} = \frac{1}{T}\sum_{t=1}^T y_{i,t-1} y_{i,t-1}'$.

Then we consider the triangular system restriction, proposed by Phillips (1991). We assume that the leading $r \times r$ sub-matrix of the cointegration matrix β_i is an identity matrix, and takes the form $\beta_i = \left(I_r, b_i'\right)'$. Let $y_{it} = \left(y_{it}^{(1)\prime}, y_{it}^{(2)\prime}\right)'$ where $y_{it}^{(1)}$ and $y_{it}^{(2)}$ are the $r \times 1$ and $(J-r) \times 1$ sub-matrix of y_{it} respectively. Due to a non-negligible bias in α_i, we define $\tilde{\alpha}_i = \alpha_i^0 + \Sigma_{uv,i}\Sigma_{vv,i}^{-1}$ (see Theorem 3.1 for detail). The model (45) reduces to

$$\Delta y_{it} - \tilde{\alpha}_i y_{i,t-1}^{(1)} = \left(y_{i,t-1}^{(2)\prime} \otimes \tilde{\alpha}_i\right) \text{vec}(b_i^{0\prime}) + \Lambda_i^{0\prime} F_t^0 + \varepsilon_{it}^*. \tag{46}$$

where $\varepsilon_{it}^* = \varepsilon_{it} - \Sigma_{uv,i}\Sigma_{vv,i}^{-1}\beta_i^{0\prime} y_{i,t-1}$ is weakly stationary process. We have the infeasible LS and GLS estimators for the above model

$$\tilde{b}_i^{LS\prime} = \left(\tilde{\alpha}_i'\tilde{\alpha}_i\right)^{-1}\tilde{\alpha}_i'\left(\sum_{t=1}^{T}\left(\Delta y_{it} - \tilde{\alpha}_i y_{i,t-1}^{(1)}\right) y_{i,t-1}^{(2)\prime}\right)\left(\sum_{t=1}^{T} y_{i,t-1}^{(2)} y_{i,t-1}^{(2)\prime}\right)^{-1},$$

$$\tilde{b}_i^{GLS\prime} = \left(\tilde{\alpha}_i'\tilde{\Sigma}_i^{-1}\tilde{\alpha}_i\right)^{-1}\tilde{\alpha}_i'\tilde{\Sigma}_i^{-1}\left(\sum_{t=1}^{T}\left(\Delta y_{it} - \tilde{\alpha}_i y_{i,t-1}^{(1)}\right) y_{i,t-1}^{(2)\prime}\right)\left(\sum_{t=1}^{T} y_{i,t-1}^{(2)} y_{i,t-1}^{(2)\prime}\right)^{-1}.$$

The GLS estimator $\tilde{b}_i^{GLS}$ is the same as the two-stage estimator proposed by Breitung (2005), where he multiply $\left(\alpha_i^{0\prime}\Sigma_i^{-1}\alpha_i^0\right)^{-1}\alpha_i^{0\prime}\Sigma_i^{-1}$ on both sides of equation (46) under i.i.d. assumption. Similarly, we multiply $\left(\tilde{\alpha}_i'\tilde{\Sigma}_i^{-1}\tilde{\alpha}_i\right)^{-1}\tilde{\alpha}'\tilde{\Sigma}_i^{-1}$ on (46), such that

$$\Delta\tilde{y}_{it}^{+} - y_{it-1}^{(1)} = b_i^{0\prime} y_{i,t-1}^{(2)} + \left(\Lambda_i^0\tilde{\gamma}_i\right)' F_t^0 + \tilde{\varepsilon}_{it}^*, \tag{47}$$

where $\Delta\tilde{y}_{it}^{+} = \tilde{\gamma}_i'\Delta y_{it}$, $\tilde{\gamma}_i = \tilde{\Sigma}_i^{-1}\tilde{\alpha}_i\left(\tilde{\alpha}_i'\tilde{\Sigma}_i^{-1}\tilde{\alpha}_i\right)^{-1}$, and $\tilde{\varepsilon}_{it}^*$ is analogously defined.

The feasible estimators of b_i can be obtained from replacing $\tilde{\alpha}_i$ and $\tilde{\sigma}_i$ by its Johansen's estimates, such that $\hat{\Sigma}_i = \frac{1}{T}\sum_{t=1}^{T}\left(\Delta y_{it} - \hat{\alpha}_i^{ML}\hat{\beta}_i^{ML\prime}y_{i,t-1}\right)'\left(\Delta y_{it} - \hat{\alpha}_i^{ML}\hat{\beta}_i^{ML\prime}y_{i,t-1}\right)$ and $\hat{\alpha}_i = \hat{\alpha}_i^{ML}$. We can show that $\hat{\alpha}_i - \tilde{\alpha}_i = O_p(T^{-1/2})$ and $\hat{\Sigma}_i - \tilde{\Sigma}_i = O_p(T^{-1/2})$. Writing above model (47) in vector form with the estimated values, we have

$$\Delta\hat{Y}_i^{+} - Y_{i,-1}^{(1)} = Y_{i,-1}^{(2)} b_i^0 + F^0\tilde{\Lambda}_i^0 + \hat{\varepsilon}_i^*, \tag{48}$$

where $\Delta\hat{Y}_i^{+} = [\Delta\hat{y}_{i1}^{+}, ..., \Delta\hat{y}_{iT}^{+}]'$, $\Delta\hat{y}_{it}^{+} = \hat{\gamma}_i'\Delta y_{it}$, $\hat{\gamma}_i = \hat{\Sigma}_i^{-1}\hat{\alpha}_i\left(\hat{\alpha}_i'\hat{\Sigma}_i^{-1}\hat{\alpha}_i\right)^{-1}$ and $\hat{\varepsilon}_i^* = \varepsilon_i^*\tilde{\gamma}_i + \Delta Y_i\left(\hat{\gamma}_i - \tilde{\gamma}_i\right)$. We can obtain the feasible GLS estimator of $\hat{b}_i^{GLS}$ as follows

$$\hat{b}_i^{GLS} = \left(Y_{i,-1}^{(2)\prime}Y_{i,-1}^{(2)}\right)^{-1}\left(Y_{i,-1}^{(2)\prime}\left(\Delta\hat{Y}_i^{+} - Y_{i,-1}^{(1)}\right)\right). \tag{49}$$

The unobserved stationary common factors will not affect the consistency of the GLS estimator. Then, we can estimate the unobserved stationary common factors after obtaining the GLS estimator $\hat{b}_i^{GLS}$. By Bai and Ng (2002), we impose the usual normalization as follows

$$F'F/T = I_M \text{ and } vec(\tilde{\Lambda}^0)'vec(\tilde{\Lambda}^0) \text{ being diagonal,}$$

where $vec(\tilde{\Lambda}^0) = [vec(\tilde{\Lambda}_1^0)', ..., vec(\tilde{\Lambda}_N^0)']'$. These two conditions uniquely determine the $\tilde{\Lambda}^0$ and F. Then we obtain the consistent estimators of F by solving the following eigenvalue decomposition problem,

$$\left(\hat{F} \otimes I_r\right) V_{NT} = \left[\frac{1}{NT}\sum_{i=1}^{N} \text{vec}\left(\Delta\hat{Y}_i^{+} - Y_{i,-1}^{(1)} - Y_{i,-1}^{(2)}\hat{b}_i^{GLS}\right)\text{vec}\left(\Delta\hat{Y}_i^{+} - Y_{i,-1}^{(1)} - Y_{i,-1}^{(2)}\hat{b}_i^{GLS}\right)'\right]\left(\hat{F} \otimes I_r\right), \tag{50}$$

where V_{NT} is a diagonal matrix consisting of the $M*r$ largest eigenvalues of the matrix inside the brackets, arranged in decreasing order. Given b_i, we have ξ_i =vec$\left(\Delta\hat{Y}_i^{+} - Y_{i,-1}^{(1)} - Y_{i,-1}^{(2)}b_i^0\right) = (F \otimes I_r)vec(\tilde{\Lambda}_i)$+vec$(\hat{\varepsilon}_i^*)$ has a pure factor structure. Let $\xi = (\xi_1, \xi_2, ..., \xi_N)$, a $Tr \times N$ matrix. We obtain the least squares estimators of vec$(\hat{\Lambda}) = \frac{1}{T}\hat{\xi}'\left(\hat{F} \otimes I_r\right)$.

PGLS-based estimation

In this subsection, we propose the penalized generalized least squares (PGLS, hereafter) method to estimate the long-run cointegration matrix b_i and identify group membership for model (48). We denote the true values of b_i as b_i^0, to follow latent group structures, such that

$$b_i^0 = \begin{cases} B_1^0, & i \in G_1^0 \\ \vdots & \vdots \\ B_K^0, & i \in G_K^0 \end{cases},$$

where $B_j^0 \neq B_k^0$ for any $j \neq k$, $\bigcup_{k=1}^{K} G_k^0 = \{1, 2, \ldots N\}$, and $G_k^0 \bigcap G_j^0 = \varnothing$ for any $j \neq k$. Let $N_k = \#G_k$ denote the cardinality of the set G_k. For the moment, we assume that the number of groups K is known and fixed but each individual's group membership is unknown. Since b exhibits latent group structures, we propose the following PGLS criterion function to estimate b and B

$$\hat{Q}_{NT}^{K,\lambda}(b, B) = \hat{Q}_{NT}(b) + \frac{\lambda}{N}\sum_{i=1}^{N}\prod_{k=1}^{K}\left\|\text{vec}(b_i - B_k)\right\|, \tag{51}$$

where $\hat{Q}_{NT}(b) = \frac{1}{NT^2}\sum_{i=1}^{N}$ vec $\left(\Delta\hat{Y}_i^{+} - Y_{i,-1}^{(1)} - Y_{i,-1}^{(2)}b_i\right)'$ vec $\left(\Delta\hat{Y}_i^{+} - Y_{i,-1}^{(1)} - Y_{i,-1}^{(2)}b_i\right)$ is the GLS objective function, and $\lambda = \lambda(N, T)$ is the tuning parameter. Minimizing the

PGLS criterion function in (51) produces the *Classifier-Lasso* (C-Lasso, hereafter) estimators of b_i and B_k, respectively. Then we can update estimates of unobserved common factors by the following equation with $\hat{B}_k$ as follows

$$\left(\hat{F} \otimes I_r\right) V_{NT} = \left[\frac{1}{NT}\sum_{k=1}^{K}\sum_{i\in\hat{G}_k} \text{vec}\left(\Delta\hat{Y}_i^{+} - Y_{i,-1}^{(1)} - Y_{i,-1}^{(2)}\hat{B}_k\right) \text{vec}\left(\Delta\hat{Y}_i^{+} - Y_{i,-1}^{(1)} - Y_{i,-1}^{(2)}\hat{B}_k\right)'\right]\left(\hat{F} \otimes I_r\right). \tag{52}$$

where $F'F/T = I_M$ and V_{NT} is a diagonal matrix consisting of the $M * r$ largest eigenvalues of the matrix inside the brackets, arranged in decreasing order.

Now we summarize estimation procedures. First, we need to obtain prior estimates of $\hat{\alpha}_i$, $\hat{\beta}_i$ and $\hat{F}$ from the two-step GLS method from (49)-(50). Second, we minimize the above PGLS criterion function (51), which produces the corresponding C-Lasso estimates $\hat{b}$ and $\hat{B}$. Third, we update the estimates of adjustment matrix $\hat{\alpha}_i = S_{01,i}\hat{B}_k\left(\hat{B}_k' S_{11,i}\hat{B}_k\right)^{-1}$ and unobserved common factors F by (52). Forth, we apply bias-correction in post-Lasso estimator of B_k with group classification from C-Lasso estimators. At last, we iterate procedures 3-4 to obtain the post-Lasso estimates until achieving convergence. Our estimators, which we will refer to as "C-Lasso" and "post-Lasso" are based on the optimal group of cross-sectional individuals, according to the PGLS criterion function. We study asymptotic properties of the C-Lasso, and post-Lasso below.

3.3 Asymptotic Theory

3.3.1 Main Assumptions

In this subsection, we explain main assumptions needed in asymptotic properties. Without loss of generality, we assume that F_t^0 has zero mean. Let $w_{it} = (\varepsilon_{it}', F_t^{0\prime})'$ and $\mathcal{C} = \sigma\left(F, \Lambda\right)$ is the sigma algebra generate by F and Λ.

Assumption 3.1 *(i) For each i, $\{w_{it}\}_0^\infty$ is a linear process and $w_{it} = \phi_i(L)e_{i,t} = \sum_{j=0}^{\infty}\phi_{ij}e_{i,t-j}$, where e_{it} is a $(J+M)\times 1$ vector sequence of i.i.d random variables with zero mean and variance matrix I_{J+M} and $\max_{1\leqslant i\leqslant N}\sup_t E(\|e_{i,t}\|^{2q+\epsilon}) < \infty$, where $q > 4$ and ϵ is an arbitrarily small positive constant. Define $e_{it} = (e_{it}^{\varepsilon\prime}, e_t^{F\prime})'$. e_{it}^{ε} and e_t^{F} are mutually independent.*

(ii) $\max_{1\leqslant i\leqslant N}\sum_{j=0}^{\infty} j^k\|\phi_{ij}\| < \infty$ and $|\phi_i(L)| \neq 0$ for some $k \geqslant 2$.

(iii) Λ_i *is independent of* e_{jt} *for all* $i, j,$ *and* t.

(iv) ε_{it} *are cross-sectionally independent conditional on* $\mathcal{C}$.

Assumption 3.2 *(i) The determinant equation* $|I_J - (I_J + \alpha_i^0 \beta_i^{0\prime}) L| = 0$ *has roots on or outside the unit circle, i.e.* $|L| \geqslant 1$.

(ii) The $I_J + \alpha_i^0 \beta_i^{0\prime}$ *where* α_i *and* β_i *are* $J \times r$ *matrices of full column rank* r, $0<r \leqslant J$.

(iii) The matrix $R_i = I_r + \beta_i^{0\prime} \alpha_i^0$ *has eigenvalues within the unit circle.*

Assumption 3.3 *(i) For each* $k = 1, ..., K_0$, $N_k/N \to \tau_k \in (0,1)$ *as* $N \to \infty$.

(ii) $\min_{1 \leqslant k \neq j \leqslant K} \left\| B_k^0 - B_j^0 \right\| \geqslant \underline{c}_\alpha$ *for some fixed* $\underline{c}_\alpha > 0$.

(iii) As $(N, T) \to \infty$, $N/T^2 \to c \in [0, \infty)$, $T/N^2 \to c \in [0, \infty)$.

(iv) As $(N, T) \to \infty$, $\lambda \iota_T \to 0$, $\lambda T N^{-1/q} \iota_T^{-2} / (\log T)^{1+\epsilon} \to \infty$, *and* $\iota_T^2 N^{1/q} T^{-1} (\log T)^{1+\epsilon} \to$ 0.

Assumption 3.1 (i)-(ii) imposes that the innovation process $\{w_{it}\}$ is a linear process that exhibits certain moment and summability conditions. Let $S_1 = (I_{J\times J}, 0_{J\times M})$ and $S_2 = (0_{M\times J}, I_{M\times M})$ be $J \times (J+M)$ and $M \times (J+M)$ selection matrices such that $S_1 w_{it} = \varepsilon_{it}$ and $S_2 w_{it} = F_t^0$. By Phillips and Solo (1992) and Phillips and Moon (1999), the finite $2q+\varepsilon$ moments for $q > 4$ ensure the validity of the law of large numbers (LLN) and functional central limit theorem for partial sum processes of w_{it}. In our asymptotic analysis, we will frequently call upon the Beveridge and Nelson (1981) BN decomposition:

$$w_{it} = \phi_i(1) e_{it} + \tilde{w}_{i,t-1} - \tilde{w}_{it},$$

where $\tilde{w}_{it} = \sum_{j=0}^{\infty} \tilde{\phi}_{ij} e_{i,t-j}$ and $\tilde{\phi}_{ij} = \sum_{s=j+1}^{\infty} \phi_{is}$. Assumption 3.1 (ii) indicates that $\tilde{w}_{it}$ has Wold decomposition and behaves like a stationary process. By Phillips and Solo (1992, p973), we have $\sum_{j=0}^{\infty} \|\tilde{\phi}_{ij}\|^2 < \infty$ under $\sum_{j=1}^{\infty} j^{1/2} \|\tilde{\phi}_{ij}\| < \infty$. In our case, we need stronger conditions to ensure the uniform behavior across i. For latter reference, we partition $\phi_i(L)$ conformably with w_{it} as follows:

$$\phi_i(L) = \begin{pmatrix} \phi_i^{\varepsilon\varepsilon}(L) & \phi_i^{\varepsilon F}(L) \\ \phi_i^{F\varepsilon}(L) & \phi^{FF}(L) \end{pmatrix} = \begin{pmatrix} \phi_i^{\varepsilon\varepsilon}(L) & 0_{J\times M} \\ 0_{M\times J} & \phi^{FF}(L) \end{pmatrix}$$

Since stationary common factors do not dependent on i, we have $\phi_i^{F\varepsilon}(L) = 0$. We set $\phi_i^{\varepsilon F}(L) = 0$ to ensure that ε_{it} is cross-sectionally independent conditional on $\mathcal{C}$. (iii) ensures that factor loadings are independent of the generation of innovation processes both over time dimension and across cross-sectional units. (iv) emphasizes that the cross-sectional dependence only comes from unobserved common factors.

Assumption 3.2 gives conditions that are standard in error correction model with reduced rank restriction. Assumption 3.2 (iii) ensures that the matrix $\beta_i'\alpha_i$ has full rank for each individual i. (see Johansen (1988), Johansen (1995) and Phillips (1995)). Let $\alpha_{i\perp}$ and $\beta_{i\perp}$ be orthogonal complements to α_i and β_i, so that $\alpha_{i\perp}$ and $\beta_{i\perp}$ are full rank $J \times (J-r)$ matrices satisfying $\alpha_{i\perp}'\alpha_i = 0_{(J-r)\times r}$ and $\beta_{i\perp}'\beta_i = 0_{(J-r)\times r}$ respectively. Without loss of generality, we assume $\beta_{i\perp}'\beta_{i\perp} = I_{J-r}$ and $\alpha_{i\perp}'\alpha_{i\perp} = I_{J-r}$. Also, $[\alpha_i, \alpha_{i\perp}]$ and $[\beta_i, \beta_{i\perp}]$ are $J \times J$ non-singular matrices. The non-singularity of $\beta_i'\alpha_i$ implies the non-singularity of $\alpha_{i\perp}'\beta_{i\perp}$. Note that y_{it} are generated from the following equation

$$y_{it} = (I_J + \alpha_i^{0\prime}\beta_i^0)y_{i,t-1} + \Lambda_i^{0\prime}F_t^0 + \varepsilon_{it}.$$

Multiplying $\beta_i^{0\prime}$ and $\alpha_{i\perp}^{0\prime}$ on both sides of above equation, we have

Stationary components: $\beta_i^{0\prime}y_{it} = (I_r + \beta_i^{0\prime}\alpha_i^0)\beta_i'y_{i,t-1} + \beta_i^{0\prime}(\Lambda_i^{0\prime}F_t^0 + \varepsilon_{it}) = R_i(L)\beta_i^{0\prime}u_{it}$

Nonstationary components: $\alpha_{i\perp}^{0\prime}y_{it} = \alpha_{i\perp}^{0\prime}y_{i,t-1} + \alpha_{i\perp}^{0\prime}\left(\Lambda_i^{0\prime}F_t^0 + \varepsilon_{it}\right) = \displaystyle\sum_{s=1}^{t}\alpha_{i\perp}^{0\prime}u_{is} + \alpha_{i\perp}^{0\prime}y_{i0}$

where $u_{it} = \Lambda_i^{0\prime}F_t^0 + \varepsilon_{it}$ and $R_i(L) = \sum_{j=0}^{\infty} R_i^j = \sum_{j=0}^{\infty}(I_r + \beta_i^{0\prime}\alpha_i^0)^j$. Combing the above two components and the fact that $\beta_{i\perp}^0\left(\alpha_{i\perp}^{0\prime}\beta_{i\perp}^0\right)^{-1}\alpha_{i\perp}^{0\prime} + \alpha_i^0\left(\beta_i^{0\prime}\alpha_i^0\right)^{-1}\beta_i^{0\prime} = I_J$, it yields the following Granger partial sum representation

$$y_{it} = \beta_{i\perp}^0\left(\alpha_{i\perp}^{0\prime}\beta_{i\perp}^0\right)^{-1}\alpha_{i\perp}^{0\prime}\left(\sum_{s=1}^{t}u_{is} + y_{i0}\right) + \alpha_i^0\left(\beta_i^{0\prime}\alpha_i^0\right)^{-1}R_i(L)\beta_i^{0\prime}u_{it},$$

where $\text{rank}\left(\beta_{i\perp}^0\left(\alpha_{i\perp}^{0\prime}\beta_{i\perp}^0\right)^{-1}\alpha_{i\perp}^{0\prime}\right) = J - r$. Since $\beta_i = \begin{pmatrix} I_r & b_i' \end{pmatrix}'$, we have the sub-vector $y_{it}^{(2)}$ satisfies the following full rank partial sum process

$$y_{it}^{(2)} = b_{i\perp}^0\left(\alpha_{i\perp}^{0\prime}\beta_{i\perp}^0\right)^{-1}\alpha_{i\perp}^{0\prime}\left(\sum_{s=1}^{t}u_{is} + y_{i0}\right) + \alpha_{2i}^0\left(\beta_i^{0\prime}\alpha_i^0\right)^{-1}R_i(L)\beta_i^{0\prime}u_{it},$$

where $\left(b_{i\perp}^0\left(\alpha_{i\perp}^{0\prime}\beta_{i\perp}^0\right)^{-1}\alpha_{i\perp}^{0\prime}\right)$ is a $(J-r)\times(J-r)$ full rank matrix and $\alpha_i^0 = \begin{pmatrix}\alpha_{1i}^{0\prime} & \alpha_{2i}^{0\prime}\end{pmatrix}'$. Since $y_{it}^{(2)}$ is a full rank nonstationary process, we have the multivariate invariance principle for $y_{it}^{(2)}$, such that

$$\frac{1}{\sqrt{T}}y_{it}^{(2)} = b_{i\perp}^0\left(\alpha_{i\perp}^{0\prime}\beta_{i\perp}^0\right)^{-1}\alpha_{i\perp}^{0\prime}\frac{1}{\sqrt{T}}\left(\sum_{s=1}^{t}u_{is}\right) + \alpha_{2i}^0\left(\beta_i^{0\prime}\alpha_i^0\right)^{-1}\frac{1}{\sqrt{T}}v_{it} \Rightarrow b_{i\perp}^0\left(\alpha_{i\perp}^{0\prime}\beta_{i\perp}^0\right)^{-1}\alpha_{i\perp}^{0\prime}B_{ui},$$

where $v_{it} = R_i(L)\beta_i^{0\prime} u_{it}$, $B_{ui} = S_i B_i$, $S_i = S_1 + \Lambda_i^{0\prime} S_2$, and B_i is a $(J + M) \times 1$ vector of Brownian motions with long-run covariance matrix Ω_i. Define the temporal variance $\Sigma_i = E(w_{i0} w_{i0}')$ and the one-sided long-run covariance matrix $\Delta_i = \sum_{j=0}^{\infty} E(w_{i0} w_{ij}') = \Gamma_i + \Sigma_i$ of $\{w_{it}\}$, where Ω_i has the following partition

$$\Omega_i = \sum_{j=-\infty}^{\infty} E(w_{i0} w_{ij}') = \Gamma_i' + \Gamma_i + \Sigma_i = \begin{pmatrix} \Omega_{11,i} & 0_{J\times M} \\ 0_{M\times J} & \Omega_{22,i} \end{pmatrix}.$$

Assumption 3.3(i) implies that each group has an asymptotically non-negligible number of individuals as $N \to \infty$. Assumption 3.3(ii) is obtained from SSP, which requires the separability of the group-specific parameters. Similar conditions are assumed in the panel literature with latent group patterns (see Bonhomme and Manresa (2015) and Huang et al. (2017)). Assumption 3.3(iii)-(iv) impose conditions to control the rate of N and T passing to infinity, which is important to the proof of uniform classification consistency. In particular, we allow for N and T pass to infinity at the same rate. The involving of ι_T is due to the law of iterated logarithm, such that $\iota_T = O(\log\log T)$. Here we can show that the range of values for λ satisfying Assumption 3.3(iv) is $\lambda \propto T^{-\alpha}$ for $\alpha \in (0, \frac{q-1}{q})$.

3.3.2 Point estimation and classification consistency

Let $C_{NT} = \min(\sqrt{N}, \sqrt{T})$, $\eta_{NT}^2 = \frac{1}{N}\sum_{i=1}^{N} \left\|\text{vec}(\hat{b}_i - b_i^0)\right\|^2$, and $H = \left(\frac{vec(\Lambda^0)vec(\Lambda^0)'}{N}\right)\left(\frac{F'\hat{F}}{T} \otimes I_r\right) V_{NT}^{-1}$. The consistency of initial estimators $\hat{b}_i^{GLS}$ and the estimated factor $\hat{F}$ is ensured by the following theorem.

Theorem 15 *Suppose that Assumptions 3.1-3.2 hold. Then for some* $\iota_T \to \infty$,
(i) $\hat{\alpha}_i - \tilde{\alpha}_i = O_p(T^{-1/2})$, *for each* $i = 1, ..., N$
(ii) $\hat{b}_i^{GLS} - b_i = O_p(T^{-1})$, *for each* $i = 1, ..., N$,
(iii) $\frac{1}{\sqrt{T}} \left\|\left(\hat{F} \otimes I_r\right) - \left(F^0 \otimes I_r\right) H\right\| = O_p(C_{NT}^{-1}) + O_p(\sqrt{T}\iota_T \eta_{NT})$
where $\tilde{\alpha}_i = \alpha_i^0 + \Sigma_{uv,i}\Sigma_{vv,i}^{-1}$.

Remark: Theorem 3.1(i)-(ii) establish the point-wise consistency for estimators of short-run adjustment matrix α_i and long-run cointegration vector b_i. We summarize some key findings. First, the estimator of short-run adjustment matrix is inconsistent around the true value α_i^0 when $\Sigma_{uv,i} \neq 0$. Instead, $\hat{\alpha}_i$ is consistent with the pseudo true value $\tilde{\alpha}_i =$

$\alpha_i^0 + \Sigma_{uv,i}\Sigma_{vv,i}^{-1}$, where $\Sigma_{uv,i}$ comes from the serial correlation and endogeneity in the innovation processes of ε_{it} and F_t^0. When we have the i.i.d. assumption, (i) reduce to the case in Johansen (1991), where $\hat{\alpha}_i - \alpha_i^0 = O_p(T^{-1/2})$. Second, despite the weakly dependence, we can still obtain super-consistency for long-run cointegration vectors. This GLS estimator is similar to the two-step parametric estimator in Breitung (2005), where he focuses on the i.i.d. case. Based on the convergence rate of $\hat{b}_i^{GLS}$, we can show that the spaces spanned by the columns of $\hat{F}$ and F^0 are asymptotically the same.

Here we presents the preliminary rates of convergence for PGLS-based estimates $\hat{b}_i$ and $\hat{B}_k$.

Theorem 16 *Suppose that Assumptions 3.1-3.2 hold. Then for some $\iota_T \to \infty$,*

(i) vec$(\hat{b}_i - b_i^0) = O_p(T^{-1} + \lambda)$, for $i = 1, ..., N$.

(ii) $\frac{1}{N}\sum_{i=1}^{N}\left\|vec(\hat{b}_i - b_i^0)\right\|^2 = O_p(\iota_T^2 T^{-2})$ for $i = 1, ..., N$.

(iii) $(vec(\hat{B}_{(1)}), ..., vec(\hat{B}_{(K)})) - (vec(B_{(1)}^0), ..., vec(B_{(K)}^0)) = O_p(\iota_T T^{-1})$

where $(vec(B_{(1)}^0), ..., vec(B_{(K)}^0))$ is suitable permutation of $(vec(B_1^0), ..., vec(B_K^0))$.

Remark: Theorem 3.2(i)-(ii) establish the point-wise and mean-square consistency for the long-run cointegration matrix b_i.Theorem 3.2(iii) indicates that the estimator vec$(\hat{B}_k)$ consistently estimate the true group-specific coefficient vec(B_k^0). We note that the point-wise convergence rate of b_i depends on λ but mean-square convergence rate of b_i and the convergence rate of $\hat{B}_k$ doesn't.

For simplicity, we will write $\hat{B}_{(k)}$ as $\hat{B}_k$. Define the estimated group

$$\hat{G}_k = \left\{i \in \{1, 2, ..., N\} : \hat{b}_i = \hat{B}_k\right\} \text{ for } k = 1, ..., K.$$

For rigorous statement of classification consistency, we define the following sequences of events

$$\hat{E}_{kNT,i} = \{i \notin \hat{G}_k | i \in G_k^0\} \quad \text{and} \quad \hat{F}_{kNT,i} = \{i \notin G_k^0 | i \in \hat{G}_k\}$$

where $i = 1, ..., N$ and $k = 1, ...K$. Let $\hat{E}_{kNT} = \cup_{i \in \hat{G}_k} \hat{E}_{kNTi}$ and $\hat{F}_{kNT} = \cup_{i \in \hat{G}_k} \hat{F}_{kNTi}$. This definition is identical to that in Su et al. (2016). The events $\hat{E}_{kNT}$ and $\hat{F}_{kNT}$ mimic Type I and Type II errors in statistical tests: $\hat{E}_{kNT}$ denotes the error event of not classifying an element of G_k^0 into estimated group $\hat{G}_k$; and $\hat{F}_{kNT}$ denotes the error event of classifying an element that does not belong to G_k^0 into the estimated group $\hat{G}_k$. We adopt the following definition to investigate the asymptotic properties of classification.

Definition 3.1 *(Uniform consistency of classification) We say that a classification method is individual consistent if $P(\hat{E}_{kNT,i}) \to 0$ as $(N,T) \to \infty$ for each $i \in G_k^0$ and $k = 1, ..., K_0$, and $P(\hat{F}_{kNT,i}) \to 0$ as $(N,T) \to \infty$ for each $i \in G_k^0$ and $k = 1, ..., K_0$. It is uniformly consistent if $P(\cup_{k=1}^{K_0} \hat{E}_{kNT}) \to 0$ and $P(\cup_{k=1}^{K_0} \hat{F}_{kNT}) \to 0$ as $(N,T) \to \infty$.*

Theorem 17 *Suppose that Assumptions 3.1-3.3 hold. Then*
(i) $P(\cup_{k=1}^{K} \hat{E}_{kNT}) \leqslant \sum_{k=1}^{K} P(\hat{E}_{kNT}) \to 0$ as $(N,T) \to \infty$,
(ii) $P(\cup_{k=1}^{K} \hat{F}_{kNT}) \leqslant \sum_{k=1}^{K} P(\hat{F}_{kNT}) \to 0$ as $(N,T) \to \infty$.

Remark: Theorem 3.3 implies the uniform classification consistency– all individuals within a certain group, say G_k^0 can be simultaneously correctly classified into same group (denoted $\hat{G}_k$) w.p.a.1. This theorem has also been established in Huang, Phillips, Su (2017).

3.3.3 Asymptotic distribution and oracle property

In this subsection, we study the oracle properties of PGLS-based estimation method. Given the estimated group $\{\hat{G}_k, k = 1, ..., K\}$, we can readily pool the observations within each estimated group to obtain the post-Lasso estimator as follows

$$\text{vec}(\hat{B}_k^{post}) = \left(I_r \otimes \frac{1}{N_k T^2} \sum_{i \in \hat{G}_k} Y_{i,-1}^{(2)\prime} Y_{i,-1}^{(2)} \right)^{-1} \text{vec} \left(\frac{1}{N_k T^2} \sum_{i \in \hat{G}_k} Y_{i,-1}^{(2)\prime} \left(\Delta \hat{Y}_i^{+} - Y_{i,-1i}^{(1)} \right) \right)$$

for $k = 1, ..., K$. When the group identity for each individual is known, we have the oracle estimators

$$\text{vec}(\hat{B}_k^{oracle}) = \left(I_r \otimes \frac{1}{N_k T^2} \sum_{i \in \hat{G}_k} Y_{i,-1}^{(2)\prime} Y_{i,-1}^{(2)} \right)^{-1} \text{vec} \left(\frac{1}{N_k T^2} \sum_{i \in G_k^0} Y_{i,-1}^{(2)\prime} \left(\Delta \hat{Y}_i^{+} - Y_{i,-1}^{(1)} \right) \right).$$

The oracle property indicates that the Lasso-type estimators are asymptotically equivalent to the infeasible estimator $\text{vec}(\hat{B}_k^{oracle})$, which can be obtained only if one knows all individuals' group identity. In the following theorem, we establish the oracle property of the PGLS-based C-Lasso estimators and their post-Lasso version.

Theorem 18 *Suppose that Assumptions 3.1-3.3 hold. Then for $k = 1, ..., K$, as $(N, T) \to \infty$,*

(i) $\sqrt{N_k}T(vec(\hat{B}_k - B_k^0)) - \mathbb{Q}_k^{-1}B_{k,NT} \Rightarrow N(0, \mathbb{Q}_k^{-1}\mathbb{V}_k\mathbb{Q}_k^{-1})$,

(ii) $\sqrt{N_k}T(vec(\hat{B}_k^{post} - B_k^0)) - \mathbb{Q}_k^{-1}B_{k,NT} \Rightarrow N(0, \mathbb{Q}_k^{-1}\mathbb{V}_k\mathbb{Q}_k^{-1})$,

where $\mathbb{Q}_k = \lim_{N\to\infty} \frac{1}{N_k}\sum_{i\in G_k^0} C_{1i}E_{\mathcal{C}}\left(\int B_{ui}B_{ui}'\right)C_{1i}'$, $\mathbb{V}_k = \lim_{N\to\infty} Var_{\mathcal{C}}\left(V_{kNT}\right) = \Omega_k$, *and* $C_{1i} = b_{i,\perp}^0\left(\alpha_{i,\perp}^{0\prime}\beta_{i,\perp}^0\right)^{-1}\alpha_{i,\perp}^{0\prime}$.

Given the limiting distribution in Theorem 3.4, one can make inference as if the true group identity is known. Since there is a non-negligible bias $B_{k,NT}$, we can remove the bias by panel DOLS or fully modified OLS method in the post-Lasso estimators (see Mark and Sul (2003) and Phillips and Moon (1999)). Therefore, the post-Lasso estimators after bias-correction are recommended for practical uses.

3.4 Monte Carlo Simulations

In this section, we evaluate the finite sample performance of the C-Lasso and post-Lasso estimators under PGLS-based methods.

We consider three data generating processes that cover panel error correction models with different specifications. Throughout these DGPs, the observations in each DGP are drawn from two groups with the proportion $N_1 : N_2 = 0.4 : 0.6$. We try four combinations of the sample sizes with $N = 40, 80$ and $T = 50, 100$.

DGP1 (Three-variate Panel ECM(1) with no common factors) The observations y_{it} are generated according to

$$\Delta y_{it} = \alpha_i^0\beta_i^{0\prime}y_{it-1} + \varepsilon_{it},$$

where ε_{it} is i.i.d $N(0, 1)$ and the cointegration rank $r = 2$, the true coefficients are $\alpha_i^0 = \alpha^0 + 0.2 * N(0, 1)$ and $\beta_i^0 = [I_r, b_i^{0\prime}]'$ where

$$\alpha^0 = \begin{pmatrix} -0.3 & 0.3 \\ 0.3 & -0.2 \\ 0.2 & 0.3 \end{pmatrix}, (B_1^0, B_2^0) = \left(\begin{pmatrix} -1/3 & -1 \end{pmatrix}, \begin{pmatrix} -1 & -1/3 \end{pmatrix}\right),$$

and $b_i^0 = B_k^0$ for $i \in G_k^0$ and $k = 1, 2$.

DGP2 (Three-variate Panel ECM(1) with two common factors) The observations y_{it} are generated according to

$$\begin{cases} y_{it} = s_{it} + \psi_i' f_t^0 \\ \Delta s_{it} = \alpha_i^0 \beta_i^{0\prime} s_{i,t-1} + \varepsilon_{it} \end{cases}$$

where the cointegration rank, α_i and β_i are the same with DGP1. Let $\psi_i = (\psi_{1i}, \psi_{2i})$, $f_t = (f_{1t}, f_{2t})$. The variables ψ_{ji} and f_{jt} are all i.i.d. $N(0,1)$ for $j = 1,2$.

DGP3 (Three-variate Panel ECM(1) with two common factors and endogeneity) The observations y_{it} are generated from the same model in DGP2. The settings of r, α_i, β_i and ψ_i are the same in DGP2. We introduce linear processes in idiosyncratic errors $w_{it} = (\varepsilon_{it}', \Delta f_t')'$, such that $w_{it} = \sum_{j=0}^{\infty} \phi_j e_{i,t-j}$ where e_{it} are i.i.d. $N(0, I_5)$, $\phi_j = j^{-3.5} * \Omega^{1/2}$, $\Omega^{1/2}$ is the symmetric square root of Ω where $\Omega = \begin{pmatrix} \Omega^{\varepsilon} & 0_{3\times 2} \\ 0_{2\times 3} & \Omega^F \end{pmatrix}$, $\Omega_{lm}^{\varepsilon} = \Omega_{lm}^F = 0.2$ for all $l \neq m$ and $\Omega_{ll}^{\varepsilon} = \Omega_{ll}^F = 1$.

In all cases, the simulation results are based on 500 replications.

Table 1 show the finite sample performance of group classification and estimation. Now we assume that the number of groups, the number of factors, and the cointegration rank are known in assessing the estimation performance. We set $\lambda = c_\lambda T^{-3/4}$ where $c_\lambda = 0.1$ or 0.2. For classification, all DGPs achieve 95% correct specification rate for different values of c_λ. Moreover, as T increases, the percentage of correct specification increases significantly and is almost close to the perfect rate 100%. The simulation results confirm good classification performance.

For the estimation, we consider both the C-Lasso estimator and its post-Lasso version. We report the bias and root-mean-square-error (RMSE) for the estimate $\hat{b}_i(1)$, which is the first parameter in b_i^0 for each DGP, where all criteria are averaged over different groups and across 500 replications. For example, we calculate the RMSE of $\hat{b}_i(1)$'s as $\frac{1}{N}\sum_{k=1}^{K_0} N_k \text{RMSE}(\hat{B}_k(1))$ with $\hat{B}_k(1)$ denoting the first element in $\hat{B}_k$ for one replication and then average them across all replications for each case.

In general, the estimation performance reveals some general patterns. First, the bias and RMSE of the C-Lasso estimates and their post-Lasso versions always decreases as either N or T increases, and they decrease faster when T increases than when N increases.

Table 15: RMSEs and Biases for Lasso-types and oracle estimates under PGLS-based method

			$c^{\lambda} = 0.1$			$c^{\lambda} = 0.2$		
N	T		% Correct specification	RMSE	Bias	% Correct specification	RMSE	Bias
DGP1								
40	50	C-Lasso	98.43	0.0626	0.0502	98.42	0.0624	0.0499
		post-Lasso	98.43	0.0139	-0.0011	98.42	0.0139	-0.0011
		Oracle	-	0.0139	-0.0016	-	0.0139	-0.0016
40	100	C-Lasso	99.84	0.0289	0.0233	99.84	0.0289	0.0232
		post-Lasso	99.84	0.0069	-0.0003	99.84	0.0069	-0.0003
		Oracle	-	0.0069	-0.0003	-	0.0069	-0.0003
80	50	C-Lasso	98.35	0.0566	0.0489	98.35	0.0560	0.0481
		post-Lasso	98.35	0.0096	-0.0001	98.35	0.0096	-0.0001
		Oracle	-	0.0095	-0.0005	-	0.0095	-0.0005
80	100	C-Lasso	99.88	0.0284	0.0253	99.89	0.0254	0.0221
		post-Lasso	99.88	0.0044	0.0001	99.89	0.0044	0.0001
		Oracle	-	0.0044	0.0000	-	0.0044	0.0000
DGP2								
40	50	C-Lasso	95.59	0.1178	0.1022	95.58	0.1176	0.1018
		post-Lasso	95.59	0.0230	0.0066	95.58	0.0231	0.0066
		Oracle	-	0.0189	-0.0042	-	0.0189	-0.0042
40	100	C-Lasso	99.10	0.0552	0.0486	99.15	0.0545	0.0478
		post-Lasso	99.10	0.0091	0.0025	99.15	0.0091	0.0025
		Oracle	-	0.0079	-0.0011	-	0.0079	-0.0011
80	50	C-Lasso	95.49	0.1095	0.1017	95.48	0.1085	0.1005
		post-Lasso	95.49	0.0167	0.0075	95.48	0.0167	0.0075
		Oracle	-	0.0113	-0.0022	-	0.0113	-0.0022
80	100	C-Lasso	99.03	0.0522	0.0490	99.03	0.0491	0.0457
		post-Lasso	99.03	0.0063	0.0024	99.03	0.0063	0.0024
		Oracle	-	0.0048	-0.0004	-	0.0048	-0.0004
DGP3								
40	50	C-Lasso	97.58	0.0786	0.0636	97.58	0.0782	0.0631
		post-Lasso	97.58	0.0174	0.0058	97.58	0.0174	0.0058
		Oracle	-	0.0173	0.0005	-	0.0173	0.0005
40	100	C-Lasso	99.57	0.0374	0.0311	99.57	0.0365	0.0300
		post-Lasso	99.57	0.0077	0.0020	99.57	0.0077	0.0020
		Oracle	-	0.0072	0.0001	-	0.0072	0.0001
80	50	C-Lasso	97.65	0.0717	0.0628	97.66	0.0708	0.0615
		post-Lasso	97.65	0.0129	0.0058	97.66	0.0129	0.0058
		Oracle	-	0.0122	0.0002	-	0.0122	0.0002
80	100	C-Lasso	99.60	0.0358	0.0325	99.60	0.0321	0.0287
		post-Lasso	99.60	0.0055	0.0020	99.60	0.0055	0.0021
		Oracle	-	0.0050	0.0003	-	0.0050	0.0003

It confirms that the estimates have faster convergence rate along the time dimension than along the cross-sectional dimension. Second, the finite sample performance of the post-Lasso estimates is close to that of the oracle ones and dominates the C-Lasso estimates in DGP1 and DGP2. This is because we have iterated estimation procedure in post-Lasso and oracle estimators. In addition, when endogeneity is introduced in DGP3, we apply DOLS bias-correction in post-Lasso and oracle-estimators. The simulation results suggest that DOLS bias-correction works fine in error correction model. Based on above simulation results, we recommend the post-Lasso estimators with DOLS bias-correction for practical uses.

3.5 Efficient price in market microstructure model

There are two fundamental functions of market microstructure, viz., liquidity and price discovery. Popular liquidity and price discovery measures, such as the bid-ask spread, pricing error (Hasbrouck (1993)), information share (Hasbrouck (1995)), and Common factor (Harris et al. (2002)), reflect the adverse selection costs and inventory costs of market makers, and the mechanics of how new information impounding into security price. Therefore, understanding market microstructure is essential to asset pricing. To economize on the number of issues, one of the central concepts is the "efficient price", which is economically meaningful for asset pricing but generally unobservable.

In the literature, the efficient price is either measured by transaction price or the mid-point of bid-ask prices, see, e.g., Hasbrouck (1993), Hasbrouck (1995), and Hansen and Lunde (2006). The underlying assumption obeys the law of one price, by which the bid and ask quotes deviate from efficient price with transient errors. In this case, the bid-ask spread can only have "second-order" effect on the level of stock price and it is hard to be detected empirically. However, a number of studies identify the positive return-illiquidity relationship across stocks (see Amihud and Mendelson (1986), and Amihud (2002)). It implies potential deviations from the "absolute" version of the law of one price. Based on this argument, we believe some patterns of unobserved heterogeneity in the long-run equilibrium between bid and ask prices. This leads to the possibility that efficient price lies between the bid and ask prices but not exactly on the mid-point. The asymmetric contributions of bid and ask prices to efficient price have been identified in commodity market and foreign exchange market, see Figuerola-Ferretti and Gonzalo (2010) and Chen and Gau (2014).

In this paper, we suggest a new measure for implicit "efficient price", which is the weighted average of bid and ask prices. First, we summarize two existing methods for price discovery that are advocated by Hasbrouck (1995) and Gonzalo and Granger (1995). Second, we propose a permanent-transitory decomposition to derive efficient price and show the linkage between the two existing methods. Third, we discuss several advantages for our PT decomposition from economic intuition and econometric perspective. Fourth, we apply our method to the NYSE Trade and Quote (TAQ) data in the sample year 2006 and summarize some findings.

3.5.1 Permanent-Transitory decomposition

The basic belief in microstructure model is that pairs of prices for a single security are cointegrated, and contain a permanent component. Therefore, there exists an error correction model for the vector of the bid and ask prices, such that

$$\Delta p_{it} = \alpha_i \beta_i' p_{i,t-1} + \sum_{l=1}^{p-1} \Gamma_{il} \Delta p_{i,t-l} + e_{it},$$

where $p_{it} = [b_{it}, a_{it}]'$, b_{it} is the bid price, and a_{it} is the ask price. Most of microstructure models assume $\beta_i = [1, -1]'$ as a prior condition, for example Hasbrouck (1995) and Hansen and Lunde (2006). The bid and ask prices can be summarized as the efficient price plus a transitory component impounding various microstructure effects:

$$a_{it} = m_{it} + s_{it}^a, \text{ and } b_{it} = m_{it} - s_{it}^b$$

where m_{it} is the implicit "efficient price", s_{it}^a and s_{it}^b are pricing errors. From the above set-up, the mid-point of bid-ask prices is a good measure for efficient price. By Hasbrouck (2002), the efficient price follows a pure random walk

$$m_{it} = m_{i,t-1} + u_{it},$$

where u_{it} is uncorrelated over time, such that $E\left(u_{it}u_{is}\right) = 0$ for $t \neq s$. In addition, he assumes the pricing error $\left(s_{it}^a, s_{it}^b\right)$ is serially uncorrelated over time but may correlated with e_{it}. Intuitively, the random-walk variance σ_{ui}^2 can be recovered from the observed price changes. The information share measure of price discovery comes from the different variance contributions of price series.

As discussed earlier, much empirical work identifies "illiquidity premium" across stocks over time. It indicates that the mis-priced issue may be fundamental, where the bid and ask prices are cointegrated but they can deviate from the one-to-one long-run equilibrium. Based on this belief, we have

$$a_{it} = I^a m_{it} + s_{it}^a, \text{ and } b_{it} = I^b m_{it} - s_{it}^b$$

where $I^a \neq 1$ or $I^b \neq 1$ for some stocks. Naturally, the mid-point of bid and ask prices may not be a good measure of the efficient price.

Gonzalo and Granger (1995) propose a permanent-transitory (PT) decomposition as

$$p_{it} = \beta_{i\perp} m_{it}^{GG} + \alpha_i s_{it}^{GG}$$

where $m_{it}^{GG} = \left(\alpha_{i\perp}' \beta_{i\perp}\right)^{-1} \alpha_{i\perp}' p_{it}$ and $s_{it}^{GG} = \left(\beta_i' \alpha_i\right)^{-1} \beta_i' p_{it}$. Based on this PT decomposition, Harris et al. (2002) propose a new measure for price discovery, such that

$$\begin{aligned} &\text{GG decomposition:} \;\; p_{it} = m_{it}^{GG} + \alpha_i s_{it}^{GG}, \\ &\text{Efficient price:} \quad m_{it}^{GG} = \frac{\alpha_{i\perp}^a}{\alpha_{i\perp}^a + \alpha_{i\perp}^b} a_{it} + \frac{\alpha_{i\perp}^b}{\alpha_{i\perp}^a + \alpha_{i\perp}^b} b_{it}, \end{aligned}$$

where $\beta_i = [1, -1]$ is a prior condition. From the above equation, the "efficient price" m_{it}^{GG} is a linear function of observed prices, which may not be the mid-point of bid and ask prices. These weights are useful measures for the price discovery. However, there are two fundamental problems by using this GG decomposition. First, the bid and ask prices still obey the "absolute" version of the law of one price where $I^a = I^b = 1$ This implies that the bid-ask spread only contains the transitory component and has no predictive power on stock returns. Second, Hansen and Lunde (2006) summarize two important facts of the market microstructure noise: (1). the noise is correlated with the efficient price, and (2) the noise is time dependent. In the GG decomposition, the market microstructure noise is $\alpha_i s_{it}^{GG}$. By Granger partial sum representation ($p = 1$), we have

$$\alpha_i s_{it}^{GG} = \alpha_i \left(\beta_i' \alpha_i\right)^{-1} R_i(L) \beta_i' e_{it}.$$

It is obvious that this market microstructure noise is time dependent but orthogonal to the efficient price. Third, from the econometric perspective we show that the short-run adjustment matrix α_i cannot be consistently estimated under weakly dependent innovation processes. In general, the GG decomposition fails to give an unbiased and consistent estimator for the "efficient price" and may be contradicted to some empirical findings.

In this paper, we propose a new permanent-transitory decomposition method based on consistent and efficient estimator of the long-run equilibrium vector β_i.[5] Due to the fact

[5] Kasa (1992), and Johansen (Corollary 4.4 p.53, 1995) discuss this PT decomposition.

that $P_{\beta_{i\perp}} + P_{\beta_i} = I_2$, we can decompose p_{it} as follows

$$\beta-\text{decomposition: } p_{it} = \beta_{i\perp} m_{it}^{\beta} + \beta_i s_{it}^{\beta}$$

where $m_{it}^{\beta} = \left(\beta_{i\perp}'\beta_{i\perp}\right)^{-1}\beta_{i\perp}' p_{it}$ and $s_{it} = \left(\beta_i'\beta_i\right)^{-1}\beta_i' p_{it}$. There are several advantages of this decomposition. First, this decomposition satisfies the definition of permanent-transitory decomposition proposed by Quah (1992). Second, the "efficient price" is also the linear combination of observed prices. In the meanwhile, the accurate estimation of m_{it} only depends on the consistency of the long-run equilibrium vector β_i. Therefore, we can still obtain consistent measure for "efficient price" under weakly dependent innovation processes. Third, we can replace p_{it} in the GG and $\beta-$decompositions by its Granger partial sum representation

$$m_{it}^{GG} = \left(\alpha_{i\perp}'\beta_{i\perp}\right)^{-1}\alpha_{i\perp}'\left(\sum_{s=1}^{t} e_{is}\right), \text{ and } m_{it}^{\beta} = \left(\alpha_{i\perp}'\beta_{i\perp}\right)^{-1}\alpha_{i\perp}'\left(\sum_{s=1}^{t} e_{is}\right) + \varepsilon_{it}$$

where $\varepsilon_{it} = \left(\beta_{i\perp}'\beta_{i\perp}\right)^{-1}\beta_{i\perp}'\alpha_i\left(\beta_i'\alpha_i\right)^{-1}R_i(L)\beta_i' e_{it}$ is transient. We observe that our efficient price measure m_{it}^{β} has exactly the same permanent component with m_{it}^{GG}. When $\beta_i = [1, -1]$, we have m_{it}^{β} lies in the mid-point of bid and ask prices, such that

$$a_{it} = m_{it}^{\beta} + s_{it}^{\beta}, \text{ and } b_{it} = m_{it}^{\beta} - s_{it}^{\beta}.$$

Fourth, the market microstructure noise has two components–a permanent component from the deviation of one-to-one relationship and transitory component, such that

$$p_{it} - m_{it} = [\beta_{i\perp} - \iota] m_{it}^{\beta} + \beta_i s_{it}^{\beta},$$

where $\iota = [1, 1]'$. When the long-run cointegration relationship $\beta_i \neq [1, -1]'$, the microstructure noise will be correlated with efficient price, which is consistent with the empirical findings in Hansen and Lunde (2006). Further, when $\beta_i = [1, -1]$, we only have the orthogonal transitory component, which is uncorrelated with the efficient price. This is consistent with the general settings in the continuous time model for realized variance (see Aït-Sahalia et al. (2005)). In general, our $\beta-$decomposition maintains the key features in the Hasbrouck and GG decompositions. Also, the potential heterogeneity in long-run cointegration relationships directly determines asymmetric contributions of bid and ask prices to efficient price.

3.5.2 Empirical results

We estimate the parameters in (ECM) for each of the days individually, with the number of lags, $p = 1$ and $p = 5$.[6] The key parameter in our analysis is $\beta_i^{\perp}$, which shows how the efficient price is constructed by the linear combinations of observed prices. We normalize the long-run cointegration vector $\beta_i = [1, b_i]'$. We impose latent group structures on the long-run cointegration relationship b_i. In our empirical results, we set the number of group $K = 3, 4$ and the elements of $\beta_k^{\perp}$ correspond to bid and ask prices, such as $(\beta_{k,bid}^{\perp}, \beta_{k,ask}^{\perp},)$ for $k = 1, 2, 3$ or $k = 1, 2, 3, 4$.

Table 2 and Table 3 report a summary of our empirical results, which are averages, minimum, maximum and standard deviation of the daily estimates, b_k, and group percentages of the whole sample N_k/N. We summarize some interesting findings. First, the long-run cointegration relations $[1, b_k]$ deviate from the one-to-one relationship. Second, on average 95% of our sample has negative estimates on b_i, which indicates the efficient price lies between bid and ask prices. Similar results preserve when $K = 4$. In addition, we notice abnormal estimates in the Group 3 when $K = 3$ and Group 4 when $K = 4$, where the efficient price is out of the bid-ask region. However, the proportion of these stock is generally below 5% in our sample.

Figure 1 and Figure 2 show the contributions of bid and ask prices to efficient price based on the group-specific estimates of b_k for $K = 3$ and $K = 4$. In general, there are distinctive asymmetric patterns of the bid and ask prices to efficient price. The results suggest some forms of information asymmetry at one of the quoted price to price discovery. It empirically identifies that buyers and sellers respond differently to information and face asymmetric trading costs in stock market.

Table 16: Summary estimates of the long-run cointegration relationships when $K = 3$

	b_1	N_1/N	b_2	N_2/N	b_3	N_3/N	Number of Observations
Average	-1.220	0.400	-0.658	0.554	0.256	0.046	932.35
Max	-0.799	0.977	-0.063	0.830	0.678	0.132	1005
Min	-1.664	0.109	-0.906	0.017	-0.314	0.000	850
Std.	0.144	0.174	0.126	0.176	0.221	0.024	29.830

[6] The results of $p = 5$ are similar to those of $p = 1$ and available upon request.

Table 17: Summary estimates of the long-run cointegration relationships when $K = 4$

	b_1	N_1/N	b_2	N_2/N	b_3	N_3/N	b_4	N_4/N	Number of Observations
Average	-1.774	0.112	-0.936	0.661	-0.312	0.168	0.493	0.060	933.72
Max	-1.065	0.597	-0.565	0.847	0.284	0.670	1.330	0.133	1005
Min	-4.232	0.006	-1.720	0.014	-0.825	0.000	0.084	0.006	850
Std.	0.326	0.082	0.139	0.113	0.182	0.120	0.161	0.018	29.942

Figure 2: Weighted contribution of bid and ask price to efficient price when $K = 3$

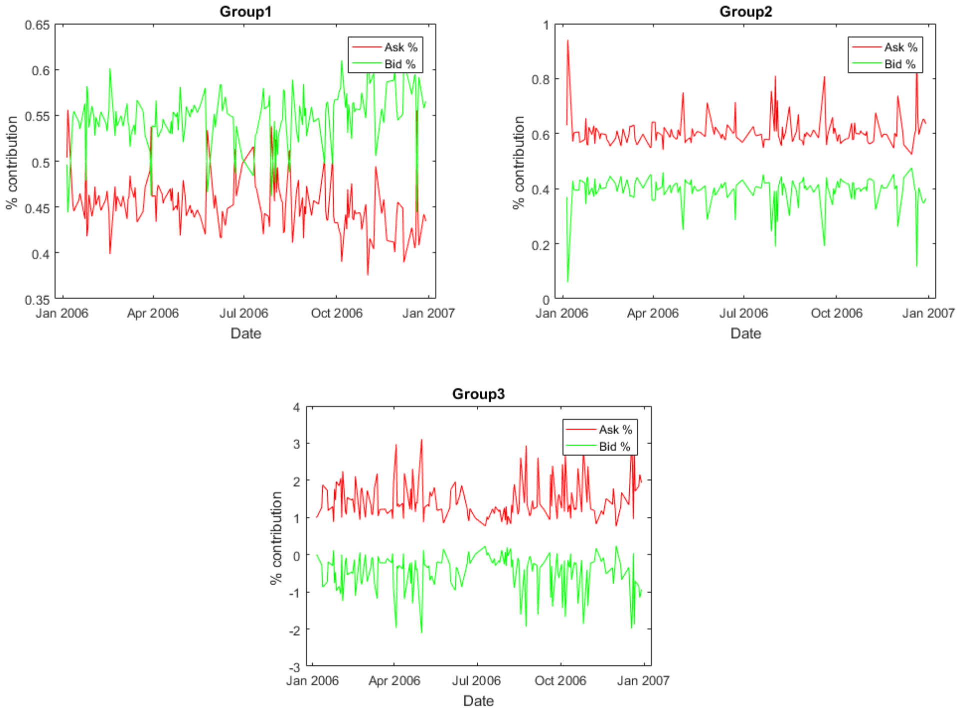
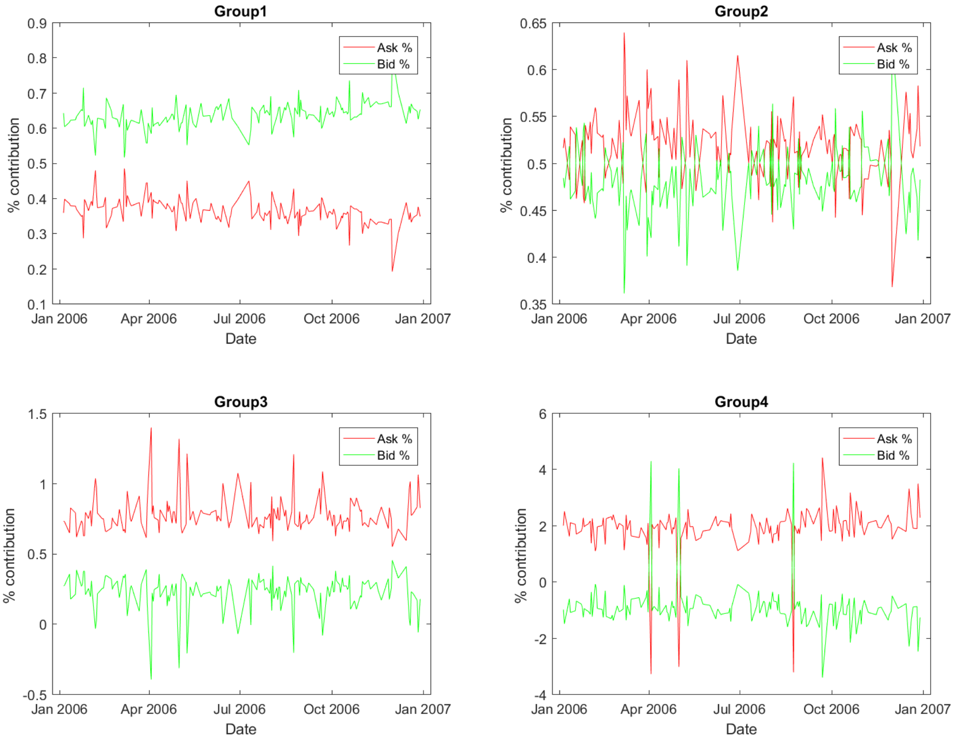


Figure 3: Weighted contribution of bid and ask price to efficient price when $K = 4$


Group1

Group2

Group3

Group4

3.6 Conclusion

In this paper, we propose a panel error correction model, where the dependent variable are generated from the combination of idiosyncratic nonstationary component and unobserved stationary common factors. The central result is to obtain consistent and efficient estimator for long-run cointegration vectors in the presence of unobserved heterogeneity. In addition, unlike the usual i.i.d. assumption in error-correction model, we allow for weak dependence in both error terms and unobserved stationary common factors. Furthermore, the unobserved common factors introduce interdependence among individuals. The estimation and group classification procedure is purely data-driven, based on the penalized generalized least squares method. We have proposed two Lasso-type estimators for long-run cointegration vectors. The asymptotic properties are derived and discussed in this paper.

In the empirical application, we employ our method on a large panel of TAQ data from sample year 2006. We propose a permanent-transitory decomposition on the vector of bid and ask prices. This decomposition yields a measure for "efficient price", which is weighted average of bid and ask prices. Compared to the mid-point assumption, we discuss several advantages for our measure. There are a number of aspects of our model that we have not fully developed in our empirical analysis and that are left for future research. First, we can quantify the permanent and transitory part in the bid-ask spread. It may explain contradicted views on the effects of bid-ask spread, Amihud and Mendelson (1986), and Vayanos (1998). Second, it is interesting to analyze the realized variance based on our estimates of efficient price.

References

ACEMOGLU, D., S. JOHNSON, J. A. ROBINSON, AND P. YARED (2008): “Income and democracy,” *American Economic Review*, 98, 808–42.

ADLER, M. AND B. LEHMANN (1983): “Deviations from purchasing power parity in the long run,” *The Journal of Finance*, 38, 1471–1487.

AÏT-SAHALIA, Y., P. A. MYKLAND, AND L. ZHANG (2005): “How often to sample a continuous-time process in the presence of market microstructure noise,” *The review of financial studies*, 18, 351–416.

AMIHUD, Y. (2002): “Illiquidity and stock returns: cross-section and time-series effects,” *Journal of financial markets*, 5, 31–56.

AMIHUD, Y. AND H. MENDELSON (1986): “Asset pricing and the bid-ask spread,” *Journal of financial Economics*, 17, 223–249.

ANDO, T. AND J. BAI (2016): “Panel data models with grouped factor structure under unknown group membership,” *Journal of Applied Econometrics*, 31, 163–191.

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

——— (2004): “Estimating cross-section common stochastic trends in nonstationary panel data,” *Journal of Econometrics*, 122, 137–183.

——— (2009): “Panel data models with interactive fixed effects,” *Econometrica*, 77, 1229–1279.

BAI, J., C. KAO, AND S. NG (2009): “Panel cointegration with global stochastic trends,” *Journal of Econometrics*, 149, 82–99.

BAI, J. AND S. NG (2002): “Determining the number of factors in approximate factor models,” *Econometrica*, 70, 191–221.

——— (2004): "A PANIC attack on unit roots and cointegration," *Econometrica*, 72, 1127–1177.

——— (2010): "Panel unit root tests with cross-section dependence: a further investigation," *Econometric Theory*, 26, 1088–1114.

BALASSA, B. (1964): "The purchasing-power parity doctrine: a reappraisal," *The Journal of Political Economy*, 584–596.

BANERJEE, A., M. MARCELLINO, AND I. MASTEN (2017): "Structural FECM: Cointegration in large-scale structural FAVAR models," *Journal of Applied Econometrics*.

BARIGOZZI, M. AND M. LUCIANI (2017): "Common factors, trends, and cycles in large datasets," *arXiv preprint arXiv:1709.01445*.

BARRO, R. J. AND X. SALA-I MARTIN (1992): "Convergence," *Journal of political Economy*, 100, 223–251.

——— (1997): "Technological diffusion, convergence, and growth," *Journal of Economic growth*, 2, 1–26.

BERNANKE, B. S., J. BOIVIN, AND P. ELIASZ (2005): "Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach," *The Quarterly journal of economics*, 120, 387–422.

BEVERIDGE, S. AND C. R. NELSON (1981): "A new approach to decomposition of economic time series into permanent and transitory components with particular attention to measurement of the business cycle," *Journal of Monetary economics*, 7, 151–174.

BLOOM, N., M. SCHANKERMAN, AND J. VAN REENEN (2013): "Identifying technology spillovers and product market rivalry," *Econometrica*, 81, 1347–1393.

BONHOMME, S. AND E. MANRESA (2015): "Grouped patterns of heterogeneity in panel data," *Econometrica*, 83, 1147–1184.

BREITUNG, J. (2005): "A parametric approach to the estimation of cointegration vectors in panel data," *Econometric Reviews*, 24, 151–173.

BRÜGGEMANN, R. AND H. LÜTKEPOHL (2005): "Practical Problems with Reduced-rank ML Estimators for Cointegration Parameters and a Simple Alternative," *Oxford Bulletin of Economics and Statistics*, 67, 673–690.

CHEN, Y.-L. AND Y.-F. GAU (2014): "Asymmetric responses of ask and bid quotes to information in the foreign exchange market," *Journal of Banking & Finance*, 38, 194–204.

CHENG, X. AND P. C. PHILLIPS (2009): "Semiparametric cointegrating rank selection," *The Econometrics Journal*, 12.

COE, D. T. AND E. HELPMAN (1995): "International r&d spillovers," *European economic review*, 39, 859–887.

COE, D. T., E. HELPMAN, AND A. W. HOFFMAISTER (2009): "International R&D spillovers and institutions," *European Economic Review*, 53, 723–741.

DAVIDSON, J. (1994): *Stochastic limit theory: An introduction for econometricians*, OUP Oxford.

DEES, S., F. D. MAURO, M. H. PESARAN, AND L. V. SMITH (2007): "Exploring the international linkages of the euro area: a global VAR analysis," *Journal of applied econometrics*, 22, 1–38.

DONSKER, M. AND S. VARADHAN (1977): "On laws of the iterated logarithm for local times," *Communications on Pure and Applied Mathematics*, 30, 707–753.

DOOLEY, M. P., D. FOLKERTS-LANDAU, AND P. GARBER (2004): "The revived bretton woods system," *International Journal of Finance & Economics*, 9, 307–313.

DURLAUF, S. N. AND P. A. JOHNSON (1995): "Multiple regimes and cross-country growth behaviour," *Journal of applied econometrics*, 10, 365–384.

EATON, J. AND S. KORTUM (2002): "Technology, geography, and trade," *Econometrica*, 70, 1741–1779.

ENGELBRECHT, H.-J. (1997): “International R&D spillovers, human capital and productivity in OECD economies: An empirical investigation,” *European Economic Review*, 41, 1479–1488.

ENGLE, R. F. AND C. W. GRANGER (1987): “Co-integration and error correction: representation, estimation, and testing,” *Econometrica: journal of the Econometric Society*, 251–276.

FIGUEROLA-FERRETTI, I. AND J. GONZALO (2010): “Modelling and measuring price discovery in commodity markets,” *Journal of Econometrics*, 158, 95–107.

FRENKEL, J. A. (1981): “The collapse of purchasing power parities during the 1970’s,” *European Economic Review*, 16, 145–165.

GONZALO, J. AND C. GRANGER (1995): “Estimation of common long-memory components in cointegrated systems,” *Journal of Business & Economic Statistics*, 13, 27–35.

GRIFFITH, R., S. REDDING, AND J. V. REENEN (2004): “Mapping the two faces of R&D: Productivity growth in a panel of OECD industries,” *Review of economics and statistics*, 86, 883–895.

GROEN, J. J. J. AND F. KLEIBERGEN (2003): “Likelihood-based cointegration analysis in panels of vector error-correction models,” *Journal of Business & Economic Statistics*, 21, 295–318.

HANSEN, B. E. (1992): “Efficient estimation and testing of cointegrating vectors in the presence of deterministic trends,” *Journal of Econometrics*, 53, 87–121.

HANSEN, P. R. AND A. LUNDE (2006): “Realized variance and market microstructure noise,” *Journal of Business & Economic Statistics*, 24, 127–161.

HASBROUCK, J. (1993): “Assessing the quality of a security market: A new approach to transaction-cost measurement,” *The Review of Financial Studies*, 6, 191–212.

——— (1995): “One security, many markets: Determining the contributions to price discovery,” *The journal of Finance*, 50, 1175–1199.

——— (2002): "Stalking the efficient price in market microstructure specifications: an overview," *Journal of Financial Markets*, 5, 329–339.

JOHANSEN, S. (1988): "Statistical analysis of cointegration vectors," *Journal of economic dynamics and control*, 12, 231–254.

——— (1991): "Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models," *Econometrica: Journal of the Econometric Society*, 1551–1580.

——— (1995): "Likelihood-based inference in cointegrated vector autoregressive models," *OUP Catalogue*.

KAO, C. AND M.-H. CHIANG (2001): "On the estimation and inference of a cointegrated regression in panel data," *Advances in econometrics*, 15, 179–222.

KASA, K. (1992): "Common stochastic trends in international stock markets," *Journal of monetary Economics*, 29, 95–124.

KELLER, W. (2004): "International Technology Diffusion," *Journal of Economic Literature*, 42, 752–782.

LAI, T. L. AND C. Z. WEI (1982): "Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems," *The Annals of Statistics*, 154–166.

LARSSON, R. AND J. LYHAGEN (2007): "Inference in Panel Cointegration Models with long panels," *Journal of Business & Economic Statistics*, 25, 473–483.

LARSSON, R., J. LYHAGEN, AND M. LÖTHGREN (2001): "Likelihood-based cointegration tests in heterogeneous panels," *The Econometrics Journal*, 109–142.

LI, D., J. QIAN, AND L. SU (2016): "Panel data models with interactive fixed effects and multiple structural breaks," *Journal of the American Statistical Association*, 111, 1804–1819.

LIN, C.-C. AND S. NG (2012): "Estimation of panel data models with parameter heterogeneity when group membership is unknown," *Journal of Econometric Methods*, 1, 42–55.

LU, X. AND L. SU (2017): "Determining the number of groups in latent panel structures with an application to income and democracy," *Quantitative Economics*, 8, 729–760.

MARK, N. C. AND D. SUL (2003): "Cointegration vector estimation by panel DOLS and long-run money demand," *Oxford Bulletin of Economics and statistics*, 65, 655–680.

MOON, H. R. AND M. WEIDNER (2017): "Dynamic linear panel regression models with interactive fixed effects," *Econometric Theory*, 33, 158–195.

OH, K.-Y. (1996): "Purchasing power parity and unit root tests using panel data," *Journal of International Money and Finance*, 15, 405–418.

PAPELL, D. H. (1997): "Searching for stationarity: Purchasing power parity under the current float," *Journal of international Economics*, 43, 313–332.

PARK, J. Y. AND P. C. PHILLIPS (1988): "Statistical inference in regressions with integrated processes: Part 1," *Econometric Theory*, 4, 468–497.

——— (1989): "Statistical inference in regressions with integrated processes: Part 2," *Econometric Theory*, 5, 95–131.

PEDRONI, P. (2001): "Fully modified OLS for heterogeneous cointegrated panels," in *Nonstationary panels, panel cointegration, and dynamic panels*, Emerald Group Publishing Limited, 93–130.

——— (2004): "Panel cointegration: asymptotic and finite sample properties of pooled time series tests with an application to the PPP hypothesis," *Econometric theory*, 20, 597–625.

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

PESARAN, M. H., T. SCHUERMANN, AND S. M. WEINER (2004): "Modeling regional interdependencies using a global error-correcting macroeconometric model," *Journal of Business & Economic Statistics*, 22, 129–162.

PHILLIPS, P. C. (1991): "Optimal inference in cointegrated systems," *Econometrica: Journal of the Econometric Society*, 283–306.

——— (1994): "Some exact distribution theory for maximum likelihood estimators of cointegrating coefficients in error correction models," *Econometrica*, 62, 73–93.

——— (1995): "Fully modified least squares and vector autoregression," *Econometrica: Journal of the Econometric Society*, 1023–1078.

——— (1996): "Econometric model determination," *Econometrica: Journal of the Econometric Society*, 763–812.

PHILLIPS, P. C. AND B. E. HANSEN (1990): "Statistical inference in instrumental variables regression with I (1) processes," *The Review of Economic Studies*, 57, 99–125.

PHILLIPS, P. C. AND H. R. MOON (1999): "Linear regression limit theory for nonstationary panel data," *Econometrica*, 67, 1057–1111.

PHILLIPS, P. C. AND V. SOLO (1992): "Asymptotics for linear processes," *The Annals of Statistics*, 971–1001.

PHILLIPS, P. C. AND D. SUL (2003): "Dynamic panel estimation and homogeneity testing under cross section dependence," *The Econometrics Journal*, 6, 217–259.

QIAN, J. AND L. SU (2016): "Shrinkage estimation of common breaks in panel data models via adaptive group fused lasso," *Journal of Econometrics*, 191, 86–109.

QUAH, D. (1992): "The relative importance of permanent and transitory components: identification and some theoretical bounds," *Econometrica: Journal of the Econometric Society*, 107–118.

QUAH, D. T. (1996): "Twin peaks: growth and convergence in models of distribution dynamics," *The economic journal*, 1045–1055.

——— (1997): "Empirics for growth and distribution: stratification, polarization, and convergence clubs," *Journal of economic growth*, 2, 27–59.

SAIKKONEN, P. (1991): "Asymptotically efficient estimation of cointegration regressions," *Econometric theory*, 7, 1–21.

SAMUELSON, P. A. (1964): "Theoretical notes on trade problems," *The Review of Economics and Statistics*, 145–154.

SARAFIDIS, V. AND N. WEBER (2015): "A partially heterogeneous framework for analyzing panel data," *Oxford Bulletin of Economics and Statistics*, 77, 274–296.

STOCK, J. H. AND M. W. WATSON (1993): "A simple estimator of cointegrating vectors in higher order integrated systems," *Econometrica: Journal of the Econometric Society*, 783–820.

——— (2002): "Forecasting using principal components from a large number of predictors," *Journal of the American statistical association*, 97, 1167–1179.

SU, L. AND G. JU (2017): "Identifying Latent Grouped Patterns in Panel Data Models with Interactive Fixed Effects," *Working paper*.

SU, L., Z. SHI, AND P. C. PHILLIPS (2016): "Identifying latent structures in panel data," *Econometrica*, 84, 2215–2264.

SU, L., Z. SHI, AND P. C. B. PHILLIPS (2013): "Identifying latent structures in panel data," .

SU, L., X. WANG, AND S. JIN (2017): "Sieve estimation of time-varying panel data models with latent structures," *Journal of Business & Economic Statistics*.

SUN, Y. (2004): "Estimation of the long-run average relationship in nonstationary panel time series," *Econometric Theory*, 20, 1227–1260.

VAN DER PLOEG, F. (2011): "Natural resources: curse or blessing?" *Journal of Economic Literature*, 49, 366–420.

VAYANOS, D. (1998): "Transaction costs and asset prices: A dynamic equilibrium model," *The Review of Financial Studies*, 11, 1–58.

WHITE, H. (2001): *Asymptotic theory for econometricians*, Academic press.

A Appendix to Chapter 1

In this appendix, we first state some technical lemmas that are used in the proofs of Theorems 1-6 and then prove these main results. The proofs of the technical lemmas are relegated to the online supplementary Appendix.

A.1 Proofs of the Main Results in Section 4

Let $x^0_{1,it} = \sum_{s=1}^t \varepsilon_{1,it}$. Noting that $x_{1,it} = x_{1,i0} + \sum_{s=1}^t \varepsilon_{1,it}$ and $\tilde{x}_{1i,t} = x_{1,it} - \frac{1}{T}\sum_{s=1}^T x_{1,is} = x^0_{1,it} - \frac{1}{T}\sum_{s=1}^T x^0_{1,is}$, the initial value $x_{1,i0}$ does not play a role in our analysis. Without loss of generality, we assume that $x_{1,i0} = 0$ and write $x_{1,it}$ for $\sum_{s=1}^t \varepsilon_{1,it}$ hereafter. Recall that

$$\begin{aligned}
\hat{Q}_{i,\tilde{x}\tilde{x}} &= \begin{pmatrix} \frac{1}{T^2}\sum_{t=1}^T \tilde{x}_{1,it}\tilde{x}'_{1,it} & \frac{1}{T^2}\sum_{t=1}^T \tilde{x}_{1,it}\tilde{x}'_{2,it} \\ \frac{1}{T^2}\sum_{t=1}^T \tilde{x}_{2,it}\tilde{x}'_{1,it} & \frac{1}{T}\sum_{t=1}^T \tilde{x}_{2,it}\tilde{x}'_{2,it} \end{pmatrix} = \begin{pmatrix} \hat{Q}_{i,\tilde{x}_1\tilde{x}_1} & \hat{Q}_{i,\tilde{x}_1\tilde{x}_2} \\ \hat{Q}_{i,\tilde{x}_2\tilde{x}_1} & \hat{Q}_{i,\tilde{x}_2\tilde{x}_2} \end{pmatrix}, \\
\hat{Q}_{i,\tilde{x}\tilde{u}} &= \begin{pmatrix} \frac{1}{T^2}\sum_{t=1}^T \tilde{x}_{1,it}\tilde{u}_{it} \\ \frac{1}{T^2}\sum_{t=1}^T \tilde{x}_{2,it}\tilde{u}_{it} \end{pmatrix} = \begin{pmatrix} \hat{Q}_{i,\tilde{x}_1\tilde{u}} \\ \hat{Q}_{i,\tilde{x}_2\tilde{u}} \end{pmatrix}, \\
\hat{Q}_{i,\tilde{x}\tilde{u}^*} &= \begin{pmatrix} \frac{1}{T^2}\sum_{t=1}^T \tilde{x}_{1,it}\tilde{u}^*_{it} \\ \frac{1}{T^2}\sum_{t=1}^T \tilde{x}_{2,it}\tilde{u}^*_{it} \end{pmatrix} = \begin{pmatrix} \hat{Q}_{i,\tilde{x}_1\tilde{u}^*} \\ \hat{Q}_{i,\tilde{x}_2\tilde{u}^*} \end{pmatrix},
\end{aligned}$$

where $\tilde{u}^*_{it} = \tilde{u}_{it} - \tilde{x}'_{2it}\Sigma^{-1}_{22,i}\Sigma_{20,i}$. Let $\tilde{x}_{1,i} = (\tilde{x}_{1,i1}, ..., \tilde{x}_{1,iN})'$. Define $\tilde{x}_{2,i}$, $\tilde{u}_i$, and $\tilde{u}^*_i$ analogously. Let $M_{\ell,i} = I_T - \tilde{x}_{\ell,i}(\tilde{x}'_{\ell,i}\tilde{x}_{\ell,i})^{-1}\tilde{x}'_{\ell,i}$ for $\ell = 1, 2$, where I_T is a $T \times T$ identity matrix. Recall that $D_T = \begin{pmatrix} I_{p_1} & 0 \\ 0 & \sqrt{T}I_{p_2} \end{pmatrix}$. We shall abbreviate $\hat{Q}_{i,\tilde{x}_1\tilde{x}_1}$ as $\hat{Q}_{1i}$ frequently for notational simplicity.

To prove the main results in the paper, we need the following lemmas.

Lemma 1 *Let $\mathbb{S} = (\mathbb{S}_1, \mathbb{S}_2)$ be a selection matrix, where $\mathbb{S}_1$ and $\mathbb{S}_2$ are $l \times p_1$ and $l \times p_2$ matrices, respectively, and l is a fixed integer. Suppose that Assumptions A.1-A.3 hold. Then for each $i = 1, ..., N$,*

(i) $\mathbb{S}D_T\hat{Q}_{i,\tilde{x}\tilde{x}}D_T\mathbb{S}' \Rightarrow \mathbb{S}\begin{pmatrix} \int_0^1 \tilde{B}_{1,i}\tilde{B}'_{1,i} & 0 \\ 0 & \Sigma_{22,i} \end{pmatrix}\mathbb{S}'$,

(ii) $T\hat{Q}_{i,\tilde{x}_1\tilde{u}^*} \Rightarrow \int_0^1 \tilde{B}_{1,i}dB'_{0,i} + \Delta_{10,i} - \left(\int_0^1 \tilde{B}_{1,i}dB'_{2,i} + \Delta_{12,i}\right)\Sigma^{-1}_{22,i}\Sigma_{20,i}$,

(iii) $T^{3/2}\mathbb{S}_2\hat{Q}_{i,\tilde{x}_2\tilde{u}^*} \Rightarrow \mathbb{S}_2\left(J_{1,i} \otimes J_{2,i}\right)N\left(0, V^0_i\right)$,

(iv) $T\left(\tilde{\beta}_{1,i} - \beta^0_{1,i}\right) \Rightarrow \left(\int_0^1 \tilde{B}_{1,i}\tilde{B}'_{1,i}\right)^{-1}\left[\int_0^1 \tilde{B}_{1,i}dB'_{0,i} + \Delta_{10,i} - \left(\int_0^1 \tilde{B}_{1,i}dB'_{2,i} + \Delta_{12,i}\right)\Sigma^{-1}_{22,i}\Sigma_{20,i}\right]$,

(v) $\sqrt{T}\mathbb{S}_2\left(\tilde{\beta}_{2,i}-\beta^*_{2,i}\right)\Rightarrow\mathbb{S}_2\left(\Sigma^{-1}_{22,i}J_{1,i}\otimes J_{2,i}\right)N\left(0,V^0_i\right)$,

where $\tilde{B}_{1,i}=B_{1,i}-\int_0^1 B_{1,i}(r)dr$, $\Delta_{10,i}=\Sigma_{10,i}+\Lambda_{10,i}$, $J_{1,i}=\left(\mathbf{0}_{p_2\times1},\mathbf{0}_{p_2\times p_1},I_{p_2}\right)$, $J_{2,i}=\left(1,\mathbf{0}_{1\times p_1},-\Sigma'_{20,i}\Sigma^{-1}_{22,i}\right)$, *and* $V^0_i=\lim_{T\to\infty}Var(T^{-1/2}\sum_{t=1}^T vec(\varepsilon_{it}\varepsilon'_{it}-\Sigma_i))$.

Lemma 2 *Suppose that Assumptions A.1-A.3 hold. Then for any fixed constant* $c>0$,

(i) $P\left(\max_{1\leqslant i\leqslant N}\frac{1}{T^2}\left\|\tilde{x}'_{1,i}\tilde{u}_i\right\|\geqslant ca_{1NT}\right)=o\left(N^{-1}\right)$,

(ii) $P\left(\max_{1\leqslant i\leqslant N}\left\|\frac{1}{T}\tilde{x}'_{2,i}\tilde{u}_i-\Sigma_{20,i}\right\|\geqslant cp_2^{1/2}a_{2NT}\right)=o\left(N^{-1}\right)$,

(iii) $P\left(\max_{1\leqslant i\leqslant N}\frac{1}{T^2}\left\|\tilde{x}'_{1,i}\tilde{x}_{2,i}\right\|\geqslant cp_2^{1/2}a_{1NT}\right)=o\left(N^{-1}\right)$,

(iv) $P\left(\max_{1\leqslant i\leqslant N}\left\|\frac{1}{T}\sum_{t=1}^T\tilde{x}_{2,it}\tilde{x}'_{2,it}-\Sigma_{22,i}\right\|\geqslant cp_2a_{2NT}\right)=o\left(N^{-1}\right)$,

(v) $P\left(\max_{1\leqslant i\leqslant N}\left\|\hat{Q}_{i,\tilde{x}_1\tilde{u}^*}\right\|\geqslant ca_{1NT}\right)=o\left(N^{-1}\right)$,

(vi) $P\left(\max_{1\leqslant i\leqslant N}\left\|T\hat{Q}_{i,\tilde{x}_2\tilde{u}^*}\right\|\geqslant cp_2^{1/2}a_{2NT}\right)=o\left(N^{-1}\right)$.

Lemma 3 *Suppose that Assumptions A.1-A.3 hold. Then*

(i) $\underset{T\to\infty}{\mathit{lim\ sup}}\lambda_{\max}\left(\frac{1}{2T^2\log\log T}\tilde{x}_{1,i}\tilde{x}'_{1,i}\right)\leqslant\left(\frac{1}{2}+c\right)\bar{c}_{\Omega_{11}}$ *a.s. for any fixed small constant* $c>0$,

(ii) $P\left(\min_{1\leqslant i\leqslant N}\lambda_{\min}(T\hat{Q}_{i,\tilde{x}_2\tilde{x}_2})\geqslant\underline{c}_{22}/2\right)=1-o\left(N^{-1}\right)$,

(iii) $P\left(\min_{1\leqslant i\leqslant N}\lambda_{\min}(D_T\hat{Q}_{i,\tilde{x}\tilde{x}}D_T)\geqslant\underline{c}_{11}/(2b_T)\right)=1-o\left(N^{-1}\right)$.

Lemma 4 *Suppose that Assumptions A.1-A.3 hold. Then for any constant* $c>0$,

(i) $P\left(\max_{1\leqslant i\leqslant N}\left\|\frac{1}{T^2}\tilde{x}'_{1,i}M_{2,i}\tilde{x}_{1,i}-\frac{1}{T^2}\tilde{x}'_{1,i}\tilde{x}_{1,i}\right\|>cb_T^{-1}\right)=o\left(N^{-1}\right)$,

(ii) $P\left(\max_{1\leqslant i\leqslant N}\left\|\frac{1}{T}\tilde{x}'_{2,i}M_{1,i}\tilde{x}_{2,i}-\Sigma_{22,i}\right\|>c\,p_2a_{2NT}\right)=o\left(N^{-1}\right)$,

(iii) $P\left(\max_{1\leqslant i\leqslant N}\left\|\frac{1}{T^2}\tilde{x}'_{1,i}M_{2,i}\tilde{u}^*_i\right\|>ca_{1NT}\right)=o\left(N^{-1}\right)$,

(iv) $P\left(\max_{1\leqslant i\leqslant N}\left\|\frac{1}{T}\tilde{x}'_{2,i}M_{1,i}\tilde{u}^*_i\right\|>cp_2^{1/2}a_{2NT}\right)=o\left(N^{-1}\right)$.

Lemma 5 *Suppose that Assumptions A.1-A.3 hold. Then for any* $\epsilon>0$,

(i) $P\left(\max_{1\leqslant i\leqslant N}\left\|\tilde{\beta}_{1,i}-\beta^0_{1,i}\right\|>c\,b_Ta_{1NT}\right)=o\left(N^{-1}\right)$,

(ii) $P\left(\max_{1\leqslant i\leqslant N}\left\|\tilde{\beta}_{2,i}-\beta^*_{2,i}\right\|>cp_2^{1/2}a_{2NT}\right)=o\left(N^{-1}\right)$,

(iii) $P\left(\max_{1\leqslant i\leqslant N}\left\|\tilde{\sigma}^2_i-\Sigma^*_{0.2,i}\right\|>\epsilon\right)=o\left(N^{-1}\right)$,

where recall that $\Sigma^*_{0.2,i}=\Sigma_{00,i}-\Sigma_{02,i}\Sigma^{-1}_{22,i}\Sigma_{20,i}$.

Lemma 6 *Suppose that Assumptions A.1-A.3 hold. Then*

(i) $\frac{1}{N}\sum_{i=1}^N\left\|\frac{1}{T^2}\tilde{x}'_{1,i}\tilde{u}^*_i\right\|^2=O_P\left(T^{-2}\right)$,

(ii) $\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{1}{T^{3/2}}\tilde{x}_{2,i}'\tilde{u}_i^*\right\|^2 = O_P\left(p_2T^{-2}\right),$

(iii) $\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{1}{T^2}\tilde{x}_{1,i}'\tilde{x}_{1,i}\right\|^2 = O_P\left(1\right),$

(iv) $\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{1}{T^2}\tilde{x}_{1,i}'\tilde{x}_{2,i}\right\|^2 = O_P\left(p_2T^{-2}\right)$

(v) $\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{1}{T^2}\tilde{x}_{1,i}'M_{2,i}\tilde{u}_i^*\right\|^2 = O_P\left(T^{-2}\right).$

To study the asymptotic distributions of the post-Lasso estimators $\hat{\alpha}_k^{\text{post}}$, we let $Q_{k,NT} = \frac{1}{N_kT^2}\sum_{i\in G_k^0}\tilde{x}_{1,i}' \times M_{2,i}\tilde{x}_{1,i}$ and $V_{k,NT} = \frac{1}{\sqrt{N_k}T}\sum_{i\in G_k^0}\tilde{x}_{1,i}'M_{2,i}\tilde{u}_i$ for $k = 1, ..., K$. We make the following decomposition for $V_{k,NT} = \frac{1}{\sqrt{N_k}T}\sum_{i\in G_k^0}\tilde{x}_{1,i}'M_{2,i}\tilde{u}_i$:

$$\begin{aligned}
V_{k,NT} &= \frac{1}{\sqrt{N_k}T}\sum_{i\in G_k^0}\tilde{x}_{1,i}'\left(\tilde{u}_i - \tilde{x}_{2,i}\Sigma_{22,i}^{-1}\Sigma_{20,i}\right) \\
&\quad + \frac{1}{\sqrt{N_k}T}\sum_{i\in G_k^0}\tilde{x}_{1,i}'\tilde{x}_{2,i}\Sigma_{22,i}^{-1}(\Sigma_{20,i} - \frac{1}{T}\tilde{x}_{2,i}'\tilde{u}_i) \\
&\quad + \frac{1}{\sqrt{N_k}T}\sum_{i\in G_k^0}\tilde{x}_{1,i}'\tilde{x}_{2,i}\left[\Sigma_{22,i}^{-1} - \left(\frac{1}{T}\tilde{x}_{2,i}'\tilde{x}_{2,i}\right)^{-1}\right]\Sigma_{20,i} \\
&\quad + \frac{1}{\sqrt{N_k}T}\sum_{i\in G_k^0}\tilde{x}_{1,i}'\tilde{x}_{2,i}\left[\Sigma_{22,i}^{-1} - \left(\frac{1}{T}\tilde{x}_{2,i}'\tilde{x}_{2,i}\right)^{-1}\right]\left(\frac{1}{T}\tilde{x}_{2,i}'\tilde{u}_i - \Sigma_{20,i}\right) \\
&\equiv V_{1k,NT} + V_{2k,NT} + V_{3k,NT} + V_{4k,NT}.
\end{aligned}$$

The following lemma studies the asymptotic properties of $Q_{k,NT}$, $V_{\ell k,NT}$ for $\ell = 1, 2, 3, 4,$ and $V_{k,NT}$.

Lemma 7 *Suppose that Assumptions A.1-A.3 hold. Then*

(i) $Q_{k,NT} \overset{P}{\to} \mathbb{Q}_{(k)},$

(ii) $V_{1k,NT} - \mathbb{B}_{k,NT} \Rightarrow N\left(0, \mathbb{V}_{(k)}\right),$

(iii) $V_{2k,NT} = o_P\left(1\right),$

(iv) $V_{3k,NT} = o_P\left(1\right)$

(v) $V_{4k,NT} = o_P\left(1\right),$

(vi) $V_{k,NT} - \mathbb{B}_{k,NT} \Rightarrow N\left(0, \mathbb{V}_{(k)}\right),$

where $\mathbb{Q}_{(k)}$, $\mathbb{B}_{k,NT}$, *and* $\mathbb{V}_{(k)}$ *are as defined before Theorem 4.*

To consider the DOLS estimator. Let $\tilde{v}_i^a = \left(\tilde{v}_{i,\bar{p}_2+1}^a, ..., \tilde{v}_{i,T-\bar{p}_2}^a\right)'$, $\tilde{v}_{it}^a = v_{it}^a - \frac{1}{T-2\bar{p}_2}\sum_{t=\bar{p}_2+1}^{T-\bar{p}_2} v_{it}^a$, where $v_{it}^a = \sum_{|j|\geqslant\bar{p}_2}\gamma_{i,j}'\Delta x_{1,i,t-j}$ signifies the approximation error. Adjust the definitions of $\tilde{x}_{1,i}$ and $M_{2,i}$ to use the time series observations $x_{\ell,i} = (x_{\ell,i,\bar{p}_2+1}, ..., x_{\ell,i,T-\bar{p}_2})'$, $\ell = 1, 2$, where recall that $x_{2,it} = (\Delta x'_{1,i,t-\bar{p}_2+1}, ..., \Delta x'_{1,i,t+\bar{p}_2})'$.

Lemma 8 *Let the conditions in Theorem 6 hold. Then* $\frac{1}{\sqrt{N_k}T}\sum_{i\in G_k^0} \tilde{x}'_{1,i} M_{2,i} \tilde{v}_i^a = o_P(1)$.

Proof of Theorem 1 (i) First, noting that $\beta_{2,i}$'s do not enter the penalty term in the PLS objective function in (7), we can concentrate them out to obtain the following objective function

$$Q^{K,c}_{NT,\lambda}(\boldsymbol{\beta}_1, \boldsymbol{\alpha}) = \frac{1}{N}\sum_{i=1}^{N} Q^c_{NT,i}(\beta_{1,i}) + \frac{\lambda}{N}\sum_{i=1}^{N}(\tilde{\sigma}_i)^{2-K}\prod_{k=1}^{K} \parallel \hat{Q}_{1i}(\beta_{1,i} - \alpha_k) \parallel . \tag{53}$$

where $Q^c_{NT,i}(\beta_{1,i}) = \frac{1}{T^2}\left\|M_{2,i}(\tilde{y}_i - \tilde{x}_{1,i}\beta_{1,i})\right\|^2$. Let $Q^{K,c}_{NTi,\lambda}(\beta_{1i}, \boldsymbol{\alpha}) = Q^c_{NT,i}(\beta_{1,i}) + \lambda(\tilde{\sigma}_i)^{2-K}\prod_{k=1}^{K}||\hat{Q}_{1i}(\beta_{1,i} - \alpha_k)||$. Then $Q^{K,c}_{NT,\lambda}(\boldsymbol{\beta}_1, \boldsymbol{\alpha}) = \frac{1}{N}\sum_{i=1}^{N} Q^{K,c}_{NTi,\lambda}(\beta_{1,i}, \alpha)$. Let $\hat{b}_{1,i} = \hat{\beta}_{1,i} - \beta^0_{1,i}$ and $\hat{b}_{2,i} = \hat{\beta}_{2,i} - \beta^*_{2,i}$. Noting that $M_{2,i}(\tilde{y}_i - \tilde{x}_{1,i}\beta_{1,i}) = M_{2,i}\left[\tilde{u}_i - \tilde{x}_{1,i}(\beta_{1,i} - \beta^0_{1,i})\right]$, we have

$$\begin{aligned} Q_{NT,i}(\hat{\beta}_{1,i}) - Q_{NT,i}(\beta^0_{1,i}) &= \frac{1}{T^2}\left\|M_{2,i}(\tilde{u}_i - \tilde{x}_{1,i}\hat{b}_{1,i})\right\|^2 - \frac{1}{T^2}\left\|M_{2,i}\tilde{u}_i\right\|^2 \\ &= \hat{b}'_{1,i}\breve{Q}_{i,\tilde{x}_1\tilde{x}_1}\hat{b}_{1,i} - 2\hat{b}'_{1,i}\breve{Q}_{i,\tilde{x}_1\tilde{u}}, \end{aligned} \tag{54}$$

where $\breve{Q}_{i,\tilde{x}_1\tilde{x}_1} = \frac{1}{T^2}\tilde{x}'_{1,i}M_{2,i}\tilde{x}_{1,i}$ and $\breve{Q}_{i,\tilde{x}_1\tilde{u}} = \frac{1}{T^2}\tilde{x}'_{1,i}M_{2,i}\tilde{u}_i$. By the triangle and reverse triangle inequalities, the fact that $||Ab|| \leqslant ||A||_{\text{sp}}||b||$ for conformable matrix A and vector b, we have

$$\begin{aligned} &\left|\prod_{k=1}^{K}||\hat{Q}_{1i}(\hat{\beta}_{1,i} - \alpha_k)|| - \prod_{k=1}^{K}||\hat{Q}_{1i}(\beta^0_{1,i} - \alpha_k)||\right| \\ \leqslant &\left|\prod_{k=1}^{K-1}||\hat{Q}_{1i}(\hat{\beta}_{1,i} - \alpha_k)||\{||\hat{Q}_{1i}(\hat{\beta}_{1,i} - \alpha_K)|| - ||\hat{Q}_{1i}(\beta^0_{1,i} - \alpha_K)||\}\right| \\ &+ \left|\prod_{k=1}^{K-2}||\hat{Q}_{1i}(\hat{\beta}_{1,i} - \alpha_k)||||\hat{Q}_{1i}(\beta^0_{1,i} - \alpha_K)||\{||\hat{Q}_{1i}(\hat{\beta}_{1,i} - \alpha_{K-1})|| - ||\hat{Q}_{1i}(\beta^0_{1i} - \alpha_{K-1})||\}\right| \\ &+ ... \\ &+ \left|\prod_{k=2}^{K}||\hat{Q}_{1i}(\beta^0_{1,i} - \alpha_k)||\{||\hat{Q}_{1i}(\hat{\beta}_{1,i} - \alpha_1)|| - ||\hat{Q}_{1i}(\beta^0_{1,i} - \alpha_1)||\}\right| \\ \leqslant &\hat{c}_{i,NT}(\boldsymbol{\alpha})||\hat{Q}_{1i}(\hat{\beta}_{1,i} - \beta^0_{1,i})|| \leqslant \hat{c}_{i,NT}(\boldsymbol{\alpha})||\hat{Q}_{1i}||_{\text{sp}}||\hat{b}_{1,i}||, \end{aligned} \tag{55}$$

where $\hat{c}_{i,NT}(\boldsymbol{\alpha}) = \prod_{k=1}^{K-1} \|\hat{Q}_{1i}(\hat{\beta}_{1,i} - \alpha_k)\| + \prod_{k=1}^{K-2} \|\hat{Q}_{1i}(\hat{\beta}_{1,i} - \alpha_k)\| \|\hat{Q}_{1i}(\beta^0_{1,i} - \alpha_K)\| + \ldots + | \prod_{k=2}^{K} \|\hat{Q}_{1i}(\beta^0_{1,i} - \alpha_k)\| = O_p(1)$ as $||\hat{Q}_{1i}||_{\text{sp}} = O_P(1)$. Since $\hat{\beta}_{1,i}$ minimize $Q^{K,c}_{NTi,\lambda}$, we have $Q^{K,c}_{NTi,\lambda}(\hat{\beta}_{1,i}, \hat{\alpha}) - Q^{K,c}_{NTi,\lambda}(\beta^0_{1,i}\hat{\alpha}) \leqslant 0$. Combining with (54)-(55), we have

$$\hat{b}'_{1,i}\breve{Q}_{i,\tilde{x}_1\tilde{x}_1}\hat{b}_{1,i} \leqslant 2\hat{b}'_{1,i}\breve{Q}_{i,\tilde{x}_1\tilde{u}} + \lambda(\tilde{\sigma}_i)^{2-K}\hat{c}_{i,NT}(\alpha)\|\hat{Q}_{1i}\|_{\text{sp}}\|\hat{b}_{1,i}\|.$$

Then

$$\underline{c}_{i,\tilde{x}_1\tilde{x}_1}\|\hat{b}_{1,i}\| \leqslant \|2\breve{Q}_{i,\tilde{x}_1\tilde{u}}\| + \lambda(\tilde{\sigma}_i)^{2-K}\hat{c}_{i,NT}(\hat{\alpha})\|\hat{Q}_{1i}\|_{\text{sp}}, \tag{56}$$

where $\underline{c}_{i,\tilde{x}_1\tilde{x}_1} = \lambda_{\min}(\breve{Q}_{i,\tilde{x}_1\tilde{x}_1}) = \lambda_{\min}(\hat{Q}_{i,\tilde{x}_1\tilde{x}_1} - T^{1/2}\hat{Q}_{i,\tilde{x}_1\tilde{x}_2}(T\hat{Q}_{i,\tilde{x}_2\tilde{x}_2})^{-1}T^{1/2}\hat{Q}_{i,\tilde{x}_2\tilde{x}_1}) \geqslant \lambda_{\min}(\hat{Q}_{i,\tilde{x}_1\tilde{x}_1}) - o_P(1)$ is bounded away from zero in probability by Lemma 1(i). In fact, we can apply Lemmas 2(iii)-(iv) and Assumptions A.2(i), A.2(iii), and A.3(iii)-(iv) and show that

$$P\left(\min_i b_T \underline{c}_{i,\tilde{x}_1\tilde{x}_1} \geqslant \underline{c}_{11}/2\right) = 1 - o\left(N^{-1}\right). \tag{57}$$

Then, by Lemmas 1(i), 2(iv), 5(iii), and Assumption A.2(iii),

$$\|\hat{b}_{1,i}\| \leqslant \underline{c}^{-1}_{i,\tilde{x}_1\tilde{x}_1}\left(2\|\breve{Q}_{i,\tilde{x}_1\tilde{u}}\| + \lambda(\tilde{\sigma}_i)^{2-K}\hat{c}_{i,NT}(\hat{\alpha})\|\hat{Q}_{1i}\|_{\text{sp}}\right) = O_P(T^{-1} + \lambda), \tag{58}$$

because

$$\begin{aligned}
\|\breve{Q}_{i,\tilde{x}_1\tilde{u}}\| &= \frac{1}{T^2}\left\|\tilde{x}'_{1,i}M_{2,i}\tilde{u}_i\right\| = \frac{1}{T^2}\left\|\tilde{x}'_{1,i}M_{2,i}\tilde{u}^*_i\right\| \\
&= \left\|\hat{Q}_{i,\tilde{x}_1\tilde{u}^*} - \hat{Q}_{i,\tilde{x}_1\tilde{x}_2}(\hat{Q}_{i,\tilde{x}_2\tilde{x}_2})^{-1}\hat{Q}_{i,\tilde{x}_1\tilde{u}^*}\right\| \\
&\leqslant \left\|\hat{Q}_{i,\tilde{x}_1\tilde{u}^*}\right\| + T^{-1}\left\|T\hat{Q}_{i,\tilde{x}_1\tilde{x}_2}\right\|\left\|T\hat{Q}_{i,\tilde{x}_2\tilde{u}^*}\right\|\left\|(T\hat{Q}_{i,\tilde{x}_2\tilde{x}_2})^{-1}\right\| = O_P\left(T^{-1}\right).
\end{aligned}$$

Now, noting that $\tilde{y}_i - \tilde{x}_{1,i}\hat{\beta}_{1,i} = \tilde{u}^*_i + \tilde{x}_{2,i}\beta^*_{2,i} - \tilde{x}_{1,i}\hat{b}_{1,i}$ and $\hat{\beta}_{2,i} = \left(\tilde{x}'_{2,i}\tilde{x}_{2,i}\right)^{-1}\tilde{x}'_{2,i}(\tilde{y}_i - \tilde{x}_{1,i}\hat{\beta}_{1,i}) = \beta^*_{2,i} + \left(\tilde{x}'_{2,i}\tilde{x}_{2,i}\right)^{-1}\tilde{x}'_{2,i}(\tilde{u}^*_i - \tilde{x}_{1,i}\hat{b}_{1,i})$, we have

$$\begin{aligned}
\left\|\hat{b}_{2,i}\right\| &= \left\|\hat{\beta}_{2,i} - \beta^*_{2,i}\right\| \leqslant \left\|\left(\frac{1}{T}\tilde{x}'_{2,i}\tilde{x}_{2,i}\right)^{-1}\right\|_{\text{sp}}\left\{\frac{1}{T}\left\|\tilde{x}'_{2,i}\tilde{u}^*_i\right\| + \frac{1}{T}\left\|\tilde{x}'_{2,i}\tilde{x}_{1,i}\right\|\left\|\hat{b}_{1,i}\right\|\right\} \\
&= O_P(1)\left\{O_P(p_2^{1/2}T^{-1/2}) + O_P(p_2^{1/2})O_P(T^{-1} + \lambda)\right\} = O_P(p_2^{1/2}(T^{-1/2} + \lambda)),
\end{aligned} \tag{59}$$

as we can readily show that $||(\frac{1}{T}\tilde{x}'_{2,i}\tilde{x}_{2,i})^{-1}||_{\text{sp}} = O_P(1)$ given Lemma 2(iv) and Assumption A.2(iii), and that $\frac{1}{T}\left\|\tilde{x}'_{2,i}\tilde{u}^*_i\right\| = O_P(p_2^{1/2}T^{-1/2})$ and $\frac{1}{T}\left\|\tilde{x}'_{2,i}\tilde{x}_{1,i}\right\| = O_P(p_2^{1/2})$ as in the proof of Lemma 1(i)-(iii).

(ii) By the Minkowski's inequality, as $(N,T)\to\infty$ we have

$$\begin{aligned}\hat{c}_{i,NT}(\boldsymbol{\alpha}) &\leqslant \prod_{k=1}^{K-1}\left\{\|\hat{Q}_{1i}(\hat{\beta}_{1,i}-\beta_{1,i}^0)\|+\|\hat{Q}_{1i}(\beta_{1,i}^0-\alpha_k)\|\right\} \\ &+\prod_{k=1}^{K-2}\left\{\|\hat{Q}_{1i}(\hat{\beta}_{1,i}-\beta_{1,i}^0)\|+\|\hat{Q}_{1i}(\beta_{1,i}^0-\alpha_k)\|\right\}\|\hat{Q}_{1i}(\beta_{1,i}^0-\alpha_K)\|+...+\prod_{k=2}^{K}\|\hat{Q}_{1i}(\beta_{1,i}^0-\alpha_k)\| \\ &=\sum_{s=0}^{K-1}\|\hat{Q}_{1i}(\hat{\beta}_{1,i}-\beta_{1,i}^0)\|^s\prod_{k=1}^{s}a_{ks}\|\hat{Q}_{1i}(\beta_{1,i}^0-\alpha_k)\|^{K-1-s} \\ &\leqslant C_{K,NT}(\alpha)\sum_{s=0}^{K-1}\|\hat{Q}_{1i}(\hat{\beta}_{1,i}-\beta_{1,i}^0)\|^s\leqslant C_{K,NT}(\alpha)(1+2\|\hat{Q}_{1i}\|_{\text{sp}}\|\hat{b}_{1,i}\|),\end{aligned} \tag{60}$$

where a_{ks}'s are finite integers and $C_{K,NT}(\boldsymbol{\alpha})=\max_i\max_{1\leqslant s\leqslant k\leqslant K-1}\prod_{k=1}^s a_{ks}\|\hat{Q}_{1i}(\beta_{1,i}^0-\alpha_k)\|^{K-1-s}=\max_{1\leqslant l\leqslant K}\max_{1\leqslant s\leqslant k\leqslant K-1}\prod_{k=1}^s a_{ks}\|\hat{Q}_{1i}(\alpha_l^0-\alpha_k)\|^{K-1-s}=O(1)$ as K is finite. Let $\hat{C}_K=C_{KNT}(\hat{\alpha})$. By Lemmas 3(i) and (iii) and Assumption A.3(iv) , $2\lambda\hat{C}_K(\tilde{\sigma}_i)^{2-K}\underline{c}_{i,\tilde{x}_1\tilde{x}_1}^{-1}\|\hat{Q}_{1i}\|_{\text{sp}}^2=O_P(\lambda b_T\log\log T)=o_P(1)$ uniformly in i. Combining (58) and (60) yields

$$\|\hat{b}_{1,i}\|\leqslant\frac{\underline{c}_{i,\tilde{x}_1\tilde{x}_1}^{-1}}{1-c_{NT}}\left\{\|2\breve{Q}_{i,\tilde{x}_1\tilde{u}}\|+\lambda\hat{C}_K(\tilde{\sigma}_i)^{2-K}\|\hat{Q}_{1i}\|_{\text{sp}}\right\},$$

where $c_{NT}=2\lambda\hat{C}_K\max_i(\tilde{\sigma}_i)^{2-K}\underline{c}_{i,\tilde{x}_1\tilde{x}_1}^{-1}\|\hat{Q}_{1i}\|_{\text{sp}}^2=o_P(1)$. Then by Lemmas 5(iii) and 6(v),

$$\frac{1}{N}\sum_{i=1}^N\|\hat{b}_{1,i}\|^2\leqslant(\frac{\hat{\underline{c}}_{\tilde{x}_1\tilde{x}_1}}{1-c_{NT}})^2\frac{1}{N}\sum_{i=1}^N[\|2\breve{Q}_{i,\tilde{x}_1\tilde{u}}\|+\lambda\hat{C}_K(\tilde{\sigma}_i)^{2-K}\|\hat{Q}_{1i}\|]^2=O_P(b_T^2(T^{-2}+\lambda^2)), \tag{61}$$

where $\hat{\underline{c}}_{\tilde{x}_1\tilde{x}_1}=\left[\min_i\underline{c}_{i,\tilde{x}_1\tilde{x}_1}\right]^{-1}=O_P(b_T)$ by (57).

To refine the result in (61), we shall prove that $\frac{1}{N}\sum_{i=1}^N||\hat{b}_{1,i}||^2=O_P(b_T^2T^{-2})$. Let $\boldsymbol{\beta}_1^0=\left(\beta_{1,1}^{0\prime},...,\beta_{1,N}^{0\prime}\right)'$ and $\boldsymbol{\beta}_1=\boldsymbol{\beta}_1^0+b_TT^{-1}\nu_{\mathbf{1}}$, where $\nu_1=(v_{1,1}',...,v_{1,N}')'$ and $\nu_{1,i}$ is a p_1-vector. We want to show that for any given $\epsilon^*>0$, there exists a large constant $L=L(\epsilon^*)$ such that, for sufficiently large N and T we have

$$P\left\{\inf_{N^{-1}\sum_{i=1}^N\|\nu_{1,i}\|^2=L}Q_{NT,\lambda}^{K,c}(\boldsymbol{\beta}_1^0+b_TT^{-1}\nu_{\mathbf{1}},\hat{\boldsymbol{\alpha}})>Q_{NT,\lambda}^{K,c}(\boldsymbol{\beta}_1^0,\boldsymbol{\alpha}^0)\right\}\geqslant 1-\epsilon^*. \tag{62}$$

This implies that w.p.a.1 there is a local minimum $\{\hat{\boldsymbol{\beta}}_1,\hat{\boldsymbol{\alpha}}\}$ such that $\frac{1}{N}\sum_{i=1}^N||\hat{b}_{1,i}||^2=O_p(b_T^2T^{-2})$ regardless of the property of $\hat{\boldsymbol{\alpha}}$. By (54), Lemma 3(iii), and the Cauchy-

Schwarz inequality, with probability $1-o\left(N^{-1}\right)$ we have

$$
\begin{aligned}
& T^2\left[Q_{NT,\lambda}^{K,c}(\boldsymbol{\beta}_1^0+b_T T^{-1}\boldsymbol{\nu}_{\mathbf{1}},\hat{\boldsymbol{\alpha}})-Q_{NT,\lambda}^{K,c}(\boldsymbol{\beta}_1^0,\boldsymbol{\alpha}^0)\right] \\
& =\frac{1}{N}\sum_{i=1}^N b_T^2\nu_{1,i}'\breve{Q}_{i,\tilde{x}_1\tilde{x}_1}\nu_{1,i}-\frac{2T}{N}\sum_{i=1}^N b_T\nu_{1,i}'\breve{Q}_{i,\tilde{x}_1\tilde{u}}+\frac{\lambda T^2}{N}\sum_{i=1}^N(\tilde{\sigma}_i)^{2-K}\prod_{k=1}^K\|\hat{Q}_{1i}(\beta_{1,i}^0+b_T T^{-1}\nu_{1,i}-\hat{\alpha}_k)\| \\
& \geqslant\left[\frac{1}{2}\underline{c}_{11}\frac{1}{N}\sum_{i=1}^N\|b_T\nu_{1,i}\|^2-2\left(\frac{1}{N}\sum_{i=1}^N\|b_T\nu_{1,i}\|^2\right)^{1/2}\left(\frac{T^2}{N}\sum_{i=1}^N\|\breve{Q}_{i,\tilde{x}_1\tilde{u}}\|^2\right)^{1/2}\right] \\
& \equiv D_{1NT}-D_{2NT},\ \text{say.}
\end{aligned}
$$

By Lemma 6(v), $\frac{T^2}{N}\sum_{i=1}^N\|\breve{Q}_{i,\tilde{x}_1\tilde{u}}\|^2=O_P(1)$. So D_{1NT} dominates D_{2NT} for sufficiently large L. That is, $T^2[Q_{NT,\lambda}^{K,c}(\boldsymbol{\beta}_1^0+b_T T^{-1}\nu_{\mathbf{1}},\hat{\boldsymbol{\alpha}})-Q_{NT,\lambda}^{K,c}(\boldsymbol{\beta}_1^0,\boldsymbol{\alpha}^0)]>0$ for sufficiently large L. Consequently, we must have $N^{-1}\sum_{i=1}^N\|\hat{b}_{1,i}\|^2=O_P(b_T^2T^{-2})$.

Note that $\left\|(\frac{1}{T}\tilde{x}_{2,i}'\tilde{x}_{2,i})^{-1}\right\|_{\text{sp}}=[\lambda_{\min}(\frac{1}{T}\tilde{x}_{2,i}'\tilde{x}_{2,i})]^{-1}$ and

$$
\min_i\lambda_{\min}(\frac{1}{T}\tilde{x}_{2,i}'\tilde{x}_{2,i})\geqslant\min_i\lambda_{\min}(\Sigma_{22,i})-\max_i\left\|\frac{1}{T}\tilde{x}_{2,i}'\tilde{x}_{2,i}-\Sigma_{22,i}\right\|\geqslant\frac{\underline{c}_{22}}{2}\text{ with probability }1-o\left(N^{-1}\right) \tag{63}
$$

by Lemma 2(iv) and Assumption A.2(iii). Then we have by (59), Lemmas 2(iii)-(iv) and 6(ii), and Assumptions A.2(iii) and A.3(iv) that

$$
\begin{aligned}
\frac{1}{N}\sum_{i=1}^N\left\|\hat{b}_{2,i}\right\|^2 &\leqslant 2\max_i\left\|\left(\frac{1}{T}\tilde{x}_{2,i}'\tilde{x}_{2,i}\right)^{-1}\right\|_{\text{sp}}^2\frac{1}{NT^2}\sum_{i=1}^N\left\{\left\|\tilde{x}_{2,i}'\tilde{u}_i^*\right\|^2+\left\|\tilde{x}_{2,i}'\tilde{x}_{1,i}\right\|^2\left\|\hat{b}_{1,i}\right\|^2\right\} \\
&\leqslant O_P(1)\left\{\frac{1}{NT^2}\sum_{i=1}^N\left\|\tilde{x}_{2,i}'\tilde{u}_i^*\right\|^2+\max_i\frac{1}{T^2}\left\|\tilde{x}_{2,i}'\tilde{x}_{1,i}\right\|^2\frac{1}{N}\sum_{i=1}^N\left\|\hat{b}_{1,i}\right\|^2\right\} \\
&= O_P\left(p_2T^{-1}\right)+O_P(p_2a_{1NT}^2)O_P(b_T^2T^{-2})=O_P\left(p_2T^{-1}\right).
\end{aligned}
$$

(iii) Let $P_{NT}(\boldsymbol{\beta}_1,\boldsymbol{\alpha})=\frac{1}{N}\sum_{i=1}^N\prod_{k=1}^K\|\beta_{1,i}-\alpha_k\|$. By (55) and (60), as $(N,T)\to\infty$,

$$
\begin{aligned}
|P_{NT}(\hat{\boldsymbol{\beta}}_1,\boldsymbol{\alpha})-P_{NT}(\boldsymbol{\beta}_1^0,\boldsymbol{\alpha})| &\leqslant C_{K,NT}(\boldsymbol{\alpha})\frac{1}{N}\sum_{i=1}^N\|\hat{b}_{1,i}\|+2C_{K,NT}(\boldsymbol{\alpha})\frac{1}{N}\sum_{i=1}^N\|\hat{b}_{1,i}\|^2 \\
&\leqslant C_{K,NT}(\boldsymbol{\alpha})\left(\frac{1}{N}\sum_{i=1}^N\|\hat{b}_{1,i}\|^2\right)^{1/2}+O_P(b_T^2T^{-2})=O_P(b_TT^{-1}).
\end{aligned} \tag{64}
$$

By (64), and the fact that $P_{NT}(\boldsymbol{\beta}_1^0,\boldsymbol{\alpha}^0)=0$ and that $P_{NT}(\hat{\boldsymbol{\beta}}_1,\hat{\boldsymbol{\alpha}})-P_{NT}(\hat{\boldsymbol{\beta}}_1,\boldsymbol{\alpha}^0)\leqslant 0$, we have

$$
\begin{aligned}
0 &\geqslant P_{NT}(\hat{\boldsymbol{\beta}}_1,\hat{\boldsymbol{\alpha}}) - P_{NT}(\hat{\boldsymbol{\beta}}_1,\boldsymbol{\alpha}^0) = P_{NT}(\boldsymbol{\beta}_1^0,\hat{\boldsymbol{\alpha}}) - P_{NT}(\boldsymbol{\beta}_1^0,\boldsymbol{\alpha}^0) + O_P(b_T T^{-1}) \\
&= \frac{1}{N}\sum_{i=1}^{N}\prod_{k=1}^{K}\|\beta_{1,i}^0 - \hat{\alpha}_k\| + O_P(b_T T^{-1}) \\
&= \frac{N_1}{N}\prod_{k=1}^{K}\|\hat{\alpha}_k - \alpha_1^0\| + \frac{N_2}{N}\prod_{k=1}^{K}\|\hat{\alpha}_k - \alpha_2^0\| + \ldots + \frac{N_K}{N}\prod_{k=1}^{K}\|\hat{\alpha}_k - \alpha_K^0\| + O_P(b_T T^{-1}).
\end{aligned}
\tag{65}
$$

By Assumption A.3(i), $N_k/N \to \tau_k \in (0,1)$ for each $k = 1,...K$. So (90) implies that $\prod_{k=1}^{K}\|\hat{\alpha}_k - \alpha_j^0\| = O_P(b_T T^{-1})$ for $j = 1,...,K$. It follows that $(\hat{\alpha}_{(1)},...,\hat{\alpha}_{(K)}) - (\alpha_1^0,...,\alpha_K^0) = O_P(b_T T^{-1})$. ■

Proof of Theorem 2. (i) By Lemma 3(i), $\limsup_{T\to\infty}\left\|\hat{Q}_{1i}\right\|_{\text{sp}} \leqslant 2\bar{c}_{\Omega_{11}}\log\log T$ a.s. By Lemma 3(iii), $P(\min_{1\leqslant i\leqslant N} b_T \underline{c}_{i,\tilde{x}_1\tilde{x}_1} \geqslant \underline{c}_{11}/2) = 1 - o\left(N^{-1}\right)$. By Lemma 5(iii) and Assumption A.2(iv), $P\left(\min_{1\leqslant i\leqslant N}\tilde{\sigma}_i^2 \geqslant \underline{c}_{00}/2\right) = 1 - o\left(N^{-1}\right)$. Noting that

$$
\left\|\breve{Q}_{i,\tilde{x}_1\tilde{u}}\right\|^2 \leqslant 2\left\|\hat{Q}_{i,\tilde{x}_1\tilde{u}^*}\right\|^2 + 2\left\|\hat{Q}_{i,\tilde{x}_1\tilde{x}_2}\right\|^2\left\|T\hat{Q}_{i,\tilde{x}_2\tilde{u}^*}\right\|^2\left\|(T\hat{Q}_{i,\tilde{x}_2\tilde{x}_2})^{-1}\right\|_{\text{sp}}^2,
$$

we can readily apply Lemma 2(iii)-(v) and Assumptions A.2(iii) and A.3(iii)-(iv) and show that $P(\max_i\left\|\breve{Q}_{i,\tilde{x}_1\tilde{u}}\right\| \geqslant ca_{1NT}) = o(N^{-1})$. Then by (56) and (60) we can show that $P(\max_{1\leqslant i\leqslant N}||\hat{b}_{1,i}|| \geqslant cb_T a_{1NT}) = o(N^{-1})$.

(ii) By (59) and (63), Lemma 2(vi), the result in part (i), and Assumption A.3(iii)-(iv)

$$
\begin{aligned}
& P\left(\max_i\left\|\hat{b}_{2,i}\right\| \geqslant cp_2^{1/2}a_{2NT}\right) \\
\leqslant\ & P\left(\max_i\left\|\left(\frac{1}{T}\tilde{x}_{2,i}'\tilde{x}_{2,i}\right)^{-1}\right\|_{\text{sp}}\frac{1}{T}\left\{\|\tilde{x}_{2,i}'\tilde{u}_i^*\| + \|\tilde{x}_{2,i}'\tilde{x}_{1,i}\|\left\|\hat{b}_{1,i}\right\|\right\} \geqslant cp_2^{1/2}a_{2NT}\right) \\
\leqslant\ & P\left(\max_i\frac{1}{T}\left(\|\tilde{x}_{2,i}'\tilde{u}_i^*\| + \|\tilde{x}_{2,i}'\tilde{x}_{1,i}\|\left\|\hat{b}_{1,i}\right\|\right) \geqslant cp_2^{1/2}a_{2NT}\underline{c}_{22}/2\right) \\
& + P\left(\min_i\lambda\left(\frac{1}{T}\tilde{x}_{2,i}'\tilde{x}_{2,i}\right) \leqslant \underline{c}_{22}/2\right) \\
\leqslant\ & P\left(\max_i\frac{1}{T}\|\tilde{x}_{2,i}'\tilde{u}_i^*\| \geqslant cp_2^{1/2}a_{2NT}\underline{c}_{22}/4\right) \\
& + P\left(\max_i\frac{1}{T}\|\tilde{x}_{2,i}'\tilde{x}_{1,i}\|\left\|\hat{b}_{1,i}\right\| \geqslant cp_2^{1/2}a_{2NT}\underline{c}_{22}/4\right) + o\left(N^{-1}\right) \\
=\ & o\left(N^{-1}\right),
\end{aligned}
$$

where we also use the fact $\max_i \frac{1}{T} \left\| \tilde{x}_{2,i}' \tilde{x}_{1,i} \right\| ||\hat{b}_{1,i}|| = o\left(Ta_{1NT}\right) o\left(b_T a_{1NT}\right) = o(p_2^{1/2} a_{2NT})$ with probability $1 - o\left(N^{-1}\right)$. ■

Proof of Theorem 3. We fix $k \in \{1, ..., K\}$. By the consistency of $\hat{\alpha}_k$ and $\hat{\beta}_{1,i}$, we have $\hat{\beta}_{1,i} - \hat{\alpha}_l \to \alpha_k^0 - \alpha_l^0 \neq 0$ for all $i \in G_k^0$ and $l \neq k$. It follows that w.p.a.1 $\left\| \hat{\beta}_{1,i} - \hat{\alpha}_l \right\| \neq 0$ for all $i \in G_k^0$ and $l \neq k$. Note that $\tilde{y}_{it} - \tilde{x}_{1,it}' \hat{\beta}_{1,i} - \tilde{x}_{2,it}' \hat{\beta}_{2,i} = \tilde{u}_{it}^* - \tilde{x}_{1,it}' \hat{b}_{1,i} - \tilde{x}_{2,it}' \hat{b}_{2,i}$.

Now, suppose that $\left\| \hat{\beta}_{1,i} - \hat{\alpha}_k \right\| \neq 0$ for some $i \in G_k^0$. Then the first order condition (with respect to $\beta_{1,i}$) for the minimization problem in (7) implies that

$$
\begin{aligned}
0 &= T \frac{\partial Q_{iNT,\lambda}^K (\hat{\boldsymbol{\beta}}_{\mathbf{1}}, \hat{\boldsymbol{\beta}}_{\mathbf{2}}, \hat{\boldsymbol{\alpha}})}{\partial \beta_{1,i}} \\
&= -2 \frac{1}{T} \sum_{t=1}^{T} \tilde{x}_{1,it} (\tilde{y}_{it} - \tilde{x}_{1,it}' \hat{\beta}_{1,i} - \tilde{x}_{2,it}' \hat{\beta}_{2,i}) + T\lambda (\tilde{\sigma}_i)^{2-K} \sum_{j=1}^{K} \hat{Q}_{1i} \hat{\varrho}_{ij} \prod_{l=1, l \neq j}^{K} \left\| \hat{Q}_{1i} (\hat{\beta}_{1,i} - \hat{\alpha}_l) \right\| \\
&= -\frac{2}{T} \sum_{t=1}^{T} \tilde{x}_{1,it} \tilde{u}_{it}^* + \left(2 + \frac{\lambda (\tilde{\sigma}_i)^{2-K} \hat{c}_{1,ik}}{\left\| \hat{Q}_{1i} (\hat{\beta}_{1,i} - \hat{\alpha}_k) \right\|} \hat{Q}_{1i} \right) T \hat{Q}_{1i} (\hat{\beta}_{1,i} - \hat{\alpha}_k) \\
&\quad + 2T \hat{Q}_{i,x_1 x_2} \hat{b}_{2,i} + 2T \hat{Q}_{1i} (\hat{\alpha}_k - \alpha_k^0) + T\lambda (\tilde{\sigma}_i)^{2-K} \sum_{j=1, j \neq k}^{K} \hat{Q}_{1i} \hat{\varrho}_{ij} \prod_{l=1, l \neq j}^{K} \left\| \hat{Q}_{1i} (\hat{\beta}_{1,i} - \hat{\alpha}_l) \right\| \\
&\equiv -\hat{B}_{i1} + \hat{B}_{i2} + \hat{B}_{i3} + \hat{B}_{i4} + \hat{B}_{i5}, \qquad (66)
\end{aligned}
$$

where $\hat{\varrho}_{ij} = \hat{Q}_{1i} (\hat{\beta}_{1,i} - \hat{\alpha}_j) / \left\| \hat{Q}_{1i} (\hat{\beta}_{1,i} - \hat{\alpha}_j) \right\|$ if $\left\| \hat{Q}_{1i} (\hat{\beta}_{1,i} - \hat{\alpha}_j) \right\| \neq 0$ and $\left\| \hat{\varrho}_{ij} \right\| \leqslant 1$ otherwise, $\hat{c}_{1,ik} = \prod_{l=1, l \neq k}^{K} \left\| \hat{Q}_{1i} (\hat{\beta}_{1,i} - \hat{\alpha}_l) \right\| \asymp c_{1,ik}^0 \equiv \prod_{l=1, l \neq k}^{K} \left\| \hat{Q}_{1i} (\alpha_k^0 - \alpha_l^0) \right\|$ for $i \in G_k^0$ by Theorem 1, where $a \asymp b$ signifies that a and b are of the same probability order.

By Theorem 2(ii), we can readily show that $P\left(\left\| \hat{\alpha}_k - \alpha_k^0 \right\| \geqslant c b_T a_{1NT} \right) = o\left(N^{-1}\right)$ for any fixed $c > 0$. This, in conjunction with Lemma 3(i) and Theorem 2(i)-(ii), implies that

$$
\left\| \hat{Q}_{1i} \right\|_{\text{sp}} \leqslant 2 \overline{c}_{\Omega_{11}} \log\log T \text{ and } c_k^0 \left(\underline{c}_{11} / b_T \right)^{K-1} \leqslant \hat{c}_{1,ik} \leqslant c_k^0 \left(2 \overline{c}_{\Omega_{11}} \log\log T \right)^{K-1} \text{ a.s.,} \qquad (67)
$$

where $c_k^0 \equiv \prod_{l=1, l \neq k}^{K} \left\| \alpha_k^0 - \alpha_l^0 \right\| > 0$ by Assumption A.3(ii). Then

$$
P\left(\max_{i \in G_k^0} \left\| \hat{B}_{i5} \right\| \geqslant CT\lambda \left(\log\log T \right)^K b_T a_{1NT} \right) = o\left(N^{-1}\right) \qquad (68)
$$

for some large constant $C > 0$. By Lemma 3(i) and Theorem 2(iii),

$$\begin{aligned}
& P\left(\max_{i\in G_k^0}\left\|\hat{B}_{i4}\right\| \geqslant Cb_T T a_{1NT}\log\log T\right) \\
\leqslant\ & P\left(\max_{i\in G_k^0}\left\|2\hat{Q}_{1i}(\hat{\alpha}_k-\alpha_k^0)\right\| \geqslant Cb_T a_{1NT}\log\log T,\ \max_{i\in G_k^0}\left\|\hat{Q}_{1i}\right\|_{\text{sp}} \leqslant 2\bar{c}_{\Omega_{11}}\log\log T\right) \\
& +P\left(\max_{i\in G_k^0}\left\|\hat{Q}_{1i}\right\|_{\text{sp}} \geqslant 2\bar{c}_{\Omega_{11}}\log\log T\right) \\
\leqslant\ & P\left(\max_{i\in G_k^0}\left\|\hat{\alpha}_k-\alpha_k^0\right\| \geqslant Cb_T a_{1NT}/(4\bar{c}_{\Omega_{11}})\right)+0 = o\left(N^{-1}\right) \qquad (69)
\end{aligned}$$

for any constant $C > 0$. By Lemma 2(iii) and Theorem 2(ii)

$$P\left(\max_{i\in G_k^0}\left\|\hat{B}_{i3}\right\| \geqslant CTb_T p_2 a_{1NT}a_{2NT}\right) = P\left(\max_{i\in G_k^0}\left\|2T\hat{Q}_{i,x_1x_2}\hat{b}_{2,i}\right\| \geqslant Cb_T p_2 a_{1NT}a_{2NT}\right) = o\left(N^{-1}\right). \qquad (70)$$

By Lemma 5(iii), Assumptions A.2(i) and A.2(iv), we have with probability $1-o\left(N^{-1}\right)$

$$\begin{aligned}
\left(\hat{Q}_{1i}(\hat{\beta}_{1,i}-\hat{\alpha}_k)\right)'\hat{B}_{i2} &= (\hat{\beta}_{1,i}-\hat{\alpha}_k)'\hat{Q}_{1i}\left(2+\frac{\lambda(\tilde{\sigma}_i)^{2-K}\hat{c}_{1,ik}}{\left\|\hat{Q}_{1i}(\hat{\beta}_{1,i}-\hat{\alpha}_k)\right\|}\hat{Q}_{1i}\right)T\hat{Q}_{1i}(\hat{\beta}_{1,i}-\hat{\alpha}_k) \\
&\geqslant T\lambda(\hat{\beta}_{1,i}-\hat{\alpha}_k)'\hat{Q}_{1i}\frac{(\tilde{\sigma}_i)^{2-K}\hat{c}_{1,ik}}{\left\|\hat{Q}_{1i}(\hat{\beta}_{1,i}-\hat{\alpha}_k)\right\|}\hat{Q}_{1i}\hat{Q}_{1i}(\hat{\beta}_{1,i}-\hat{\alpha}_k) \\
&\geqslant T\lambda b_T^{-1}\lambda_{\min}\left(b_T\hat{Q}_{1i}\right)(\tilde{\sigma}_i)^{2-K}\hat{c}_{1,ik}\left\|\hat{Q}_{1i}(\hat{\beta}_{1,i}-\hat{\alpha}_k)\right\| \\
&\geqslant \underline{c}_{11}^K c_k^0(2\bar{c}_{00})^{1-K/2}T\lambda b_T^{-K}\left\|\hat{Q}_{1i}(\hat{\beta}_{1,i}-\hat{\alpha}_k)\right\|. \qquad (71)
\end{aligned}$$

Define

$$\begin{aligned}
\Gamma_{kNT} \equiv\ & \left\{\underline{c}_{11}c_k^0/b_T \leqslant \min_{i\in G_k^0}\hat{c}_{1,ik} \leqslant \max_{i\in G_k^0}\hat{c}_{1,ik} \leqslant 2c_k^0\bar{c}_{\Omega_{11}}\log\log T\right\} \\
& \cap\left\{\max_{i\in G_k^0}\left\|\hat{B}_{i5}\right\| \leqslant CT\lambda\left(\log\log T\right)^K b_T a_{1NT}\right\}\cap\left\{\max_{i\in G_k^0}\left\|\hat{B}_{i4}\right\| \leqslant Cb_T T a_{1NT}\log\log T\right\} \\
& \cap\left\{\max_{i\in G_k^0}\left\|\hat{B}_{i3}\right\| \leqslant CTb_T p_2 a_{1NT}a_{2NT}\right\}.
\end{aligned}$$

Then $P\left(\Gamma_{kNT}\right) = 1-o\left(N^{-1}\right)$ by (67)-(95). Let Γ_{kNT}^c denote the complement of Γ_{kNT}.

Conditional on Γ_{kNT}, we have, uniformly in $i \in G_k^0$,

$$\begin{aligned}
&\left|\left(\hat{Q}_{1i}(\hat{\beta}_{1,i}-\hat{\alpha}_k)\right)'\left(\hat{B}_{i2}+\hat{B}_{i3}+\hat{B}_{i4}+\hat{B}_{i5}\right)\right| \\
\geqslant\ & \left|\left(\hat{Q}_{1i}(\hat{\beta}_{1,i}-\hat{\alpha}_k)\right)'\hat{B}_{i2}\right| - \left|\left(\hat{Q}_{1i}(\hat{\beta}_{1,i}-\hat{\alpha}_k)\right)'\left(\hat{B}_{i3}+\hat{B}_{i4}+\hat{B}_{i5}\right)\right| \\
\geqslant\ & \left\{\underline{c}_{11}^K c_k^0 (2\bar{c}_{00})^{1-K/2} T\lambda b_T^{-K} - C\left[Tb_T p_2 a_{1NT} a_{2NT} + b_T T a_{1NT}\log\log T + T\lambda\left(\log\log T\right)^K b_T a_{1NT}\right]\right\} \\
&\times \left\|\hat{Q}_{1i}(\hat{\beta}_{1,i}-\hat{\alpha}_k)\right\| \\
\geqslant\ & \frac{1}{2}\underline{c}_{11}^K c_k^0 (2\bar{c}_{00})^{1-K/2} T\lambda b_T^{-K}\left\|\hat{Q}_{1i}(\hat{\beta}_{1,i}-\hat{\alpha}_k)\right\| \text{ for sufficiently large } (N,T),
\end{aligned}$$

where the last equality follows because $Tb_T p_2 a_{1NT} a_{2NT} + b_T T a_{1NT}\log\log T + T\lambda\left(\log\log T\right)^K b_T a_{1NT} = o\left(T\lambda b_T^{-K}\right)$ by Assumption A.3(iv). It follows that for all $i \in G_k^0$,

$$\begin{aligned}
P(\hat{E}_{kNT,i}) &= P\left(i \notin \hat{G}_k | i \in G_k^0\right) = P\left(\hat{B}_{i1} = \hat{B}_{i2}+\hat{B}_{i3}+\hat{B}_{i4}+\hat{B}_{i5}\right) \\
&\leqslant P\left(\left\|\hat{Q}_{1i}(\hat{\beta}_{1,i}-\hat{\alpha}_k)\hat{B}_{i1}\right\| \geqslant \left\|\hat{Q}_{1i}(\hat{\beta}_{1,i}-\hat{\alpha}_k)\left(\hat{B}_{i2}+\hat{B}_{i3}+\hat{B}_{i4}+\hat{B}_{i5}\right)\right\|\right) \\
&\leqslant P\left(\left\|\hat{Q}_{1i}(\hat{\beta}_{1,i}-\hat{\alpha}_k)\hat{B}_{i1}\right\| \geqslant \left\|\hat{Q}_{1i}(\hat{\beta}_{1,i}-\hat{\alpha}_k)\left(\hat{B}_{i2}+\hat{B}_{i3}+\hat{B}_{i4}+\hat{B}_{i5}\right)\right\|, \Gamma_{kNT}\right) \\
&\quad + P\left(\Gamma_{kNT}^c\right) \\
&\leqslant P\left(\left\|\hat{B}_{i1}\right\| \geqslant \frac{1}{2}\underline{c}_{11}^K c_k^0 (2\bar{c}_{00})^{1-K/2} T\lambda b_T^{-K}\right) + o\left(N^{-1}\right) \\
&= o\left(1\right),
\end{aligned}$$

where the last line follows by the fact that $||\hat{B}_{i1}|| = O_P\left(1\right)$ by Lemma 1(ii) and that $T\lambda b_T^{-K} \to \infty$ under Assumption A.3(iv).

In addition, by Lemma 2(v) and the fact that $a_{1NT} = o\left(\lambda b_T^{-K}\right)$ under Assumption A.3(iv),

$$\begin{aligned}
P(\cup_{k=1}^K \hat{E}_{kNT}) &\leqslant \sum_{k=1}^K P(\hat{E}_{kNT}) \leqslant \sum_{k=1}^K \sum_{i \in G_k} P(\hat{E}_{kNT,i}) \\
&\leqslant \sum_{k=1}^K \sum_{i \in G_k} P\left(\left\|\hat{B}_{i1}\right\| \geqslant \frac{1}{2}\underline{c}_{11}^K c_k^0 (2\bar{c}_{00})^{2-K} T\lambda b_T^{-K}\right) + o\left(1\right) \\
&\leqslant N \max_{1 \leqslant i \leqslant N} P\left(\left\|\hat{B}_{i1}\right\| \geqslant \frac{1}{2}\underline{c}_{11}^K c_k^0 (2\bar{c}_{00})^{2-K} T\lambda b_T^{-K}\right) + o(1) = o\left(1\right).
\end{aligned} \tag{72}$$

We have completed the proof of Theorem 3(i).

Given (i), the proof of (ii) is similar to Theorem 4.2(ii) in SSP and thus omitted. ■

Proof of Theorem 4. We first write our mixed panel model in vector form: $\tilde{y}_i = \tilde{x}_{1,i}\beta_{1,i} + \tilde{x}_{2,i}\beta_{2,i} + \tilde{u}_i$, where $\tilde{x}_{l,i} = (\tilde{x}_{l,i1}, ..., \tilde{x}_{l,iT})'$ for $l = 1, 2$, and $\tilde{y}_i$ and $\tilde{u}_i$ are similarly defined. Recall that $M_{2,i} = I_T - \tilde{x}_{2,i}(\tilde{x}_{2,i}'\tilde{x}_{2,i})^{-1}\tilde{x}_{2,i}'$. Then we rewrite the objective function $Q_{NT,\lambda}^K(\boldsymbol{\beta_1}, \boldsymbol{\beta}_2, \boldsymbol{\alpha})$ as follows

$$Q_{NT,\lambda}^K(\boldsymbol{\beta_1}, \boldsymbol{\beta}_2, \boldsymbol{\alpha}) = Q_{NT}(\boldsymbol{\beta_1}, \boldsymbol{\beta}_2) + \frac{\lambda}{N}\sum_{i=1}^N (\tilde{\sigma}_i)^{2-K} \prod_{k=1}^K \|\hat{Q}_{1i}(\beta_{1,i} - \alpha_k)\|, \tag{73}$$

where

$$Q_{NT}(\boldsymbol{\beta_1}, \boldsymbol{\beta}_2) = \frac{1}{NT^2}\sum_{i=1}^N (\tilde{y}_i - \tilde{x}_{1,i}\beta_{1,i} - \tilde{x}_{2,i}\beta_{2,i})'(\tilde{y}_i - \tilde{x}_{1,i}\beta_{1,i} - \tilde{x}_{2,i}\beta_{2,i}). \tag{74}$$

The first order conditions are

$$\mathbf{0}_{p_1 \times 1} = \frac{-2}{T^2}\tilde{x}_{1,i}'(\tilde{y}_i - \tilde{x}_{1,i}\hat{\beta}_{1,i} - \tilde{x}_{2,i}\hat{\beta}_{2,i}) + \lambda(\tilde{\sigma}_i)^{2-K}\sum_{j=1}^K \hat{Q}_{1i}\hat{\varrho}_{ij} \prod_{l=1, l\neq j}^K \|\hat{Q}_{1i}(\hat{\beta}_{1,i} - \hat{\alpha}_l)\| \; \forall\, i = 1, ..., N, \tag{75}$$

$$\mathbf{0}_{p_2 \times 1} = \frac{-2}{T^2}\tilde{x}_{2,i}'(\tilde{y}_i - \tilde{x}_{1,i}\hat{\beta}_{1,i} - \tilde{x}_{2,i}\hat{\beta}_{2,i}) \; \forall\, i = 1, ..., N, \text{ and} \tag{76}$$

$$\mathbf{0}_{p_1 \times 1} = \frac{\lambda}{N}\sum_{i=1}^N (\tilde{\sigma}_i)^{2-K}\hat{Q}_{1i}\hat{\varrho}_{ik} \prod_{l=1, l\neq k}^K \|\hat{Q}_{1i}(\hat{\beta}_{1,i} - \hat{\alpha}_l)\| \; \forall\, k = 1, ..., K, \tag{77}$$

where $\hat{\varrho}_{ij}$ is defined after (66). Let $k \in \{1, ..., K\}$ be fixed. We observe that (a) $\|\hat{\beta}_{1,i} - \hat{\alpha}_k\| = 0$ for any $i \in \hat{G}_k$ by the definition of $\hat{G}_k$, and (b) $\hat{\beta}_{1,i} - \hat{\alpha}_l \xrightarrow{p} \alpha_k^0 - \alpha_l^0 \neq 0$ for any $i \in \hat{G}_k$ and $l \neq k$. It follows that $\|\hat{\varrho}_{ij}\| \leqslant \|1\|$ for any $i \in \hat{G}_k$ and $\hat{\varrho}_{ij} = \hat{Q}_{1i}(\hat{\alpha}_k - \hat{\alpha}_j)/\|\hat{Q}_{1i}(\hat{\alpha}_k - \hat{\alpha}_j)\|$ for any $i \in \hat{G}_k$ and $j \neq k$. Let $\hat{G}_0$ denote the set of unclassified individuals. Given Theorem 3, it is easy to show that $P(\#\hat{G}_0 > 0) = o(1)$. Noting that $\prod_{l=1}^K \|\hat{Q}_{1i}(\hat{\alpha}_k - \hat{\alpha}_l)\| = 0$ for any l, we have

$$\begin{aligned}
&\sum_{i\in\hat{G}_k}\sum_{j=1, j\neq k}^K (\tilde{\sigma}_i)^{2-K}\hat{Q}_{1i}\hat{\varrho}_{ij}\prod_{l=1, l\neq j}^K \|\hat{Q}_{1i}(\hat{\beta}_{1,i} - \hat{\alpha}_l)\| \\
=&\sum_{i\in\hat{G}_k}\sum_{j=1, j\neq k}^K (\tilde{\sigma}_i)^{2-K}\frac{\hat{Q}_{1i}^2(\hat{\alpha}_k - \hat{\alpha}_j)}{\|\hat{Q}_{1i}(\hat{\alpha}_k - \hat{\alpha}_j)\|}\prod_{l=1, l\neq j}^K \|\hat{Q}_{1i}(\hat{\alpha}_k - \hat{\alpha}_l)\| = \mathbf{0}_{p_1 \times 1}.
\end{aligned} \tag{78}$$

It follows that by (77) and (78)

$$\begin{aligned}
\mathbf{0}_{p_1\times 1} &= \sum_{i=1}^{N}(\tilde{\sigma}_i)^{2-K}\hat{Q}_{1i}\hat{\varrho}_{ik}\prod_{l=1,l\neq k}^{K}\|\hat{Q}_{1i}(\hat{\beta}_{1,i}-\hat{\alpha}_l)\| \\
&= \sum_{i\in\hat{G}_k}(\tilde{\sigma}_i)^{2-K}\hat{Q}_{1i}\hat{\varrho}_{ik}\prod_{l=1,l\neq k}^{K}\|\hat{Q}_{1i}(\hat{\alpha}_k-\hat{\alpha}_l)\| + \sum_{i\in\hat{G}_0}(\tilde{\sigma}_i)^{2-K}\hat{Q}_{1i}\hat{\varrho}_{ik}\prod_{l=1,l\neq k}^{K}\|\hat{Q}_{1i}(\hat{\beta}_{1,i}-\hat{\alpha}_l)\| \\
&\quad + \sum_{j=1,j\neq k}^{K}\sum_{i\in\hat{G}_j}(\tilde{\sigma}_i)^{2-K}\frac{\hat{Q}_{1i}^2(\hat{\alpha}_j-\hat{\alpha}_k)}{\|\hat{Q}_{1i}(\hat{\alpha}_j-\hat{\alpha}_k)\|}\prod_{l=1,l\neq j}^{K}\|\hat{Q}_{1i}(\hat{\alpha}_j-\hat{\alpha}_l)\| \\
&= \sum_{i\in\hat{G}_k}(\tilde{\sigma}_i)^{2-K}\hat{Q}_{1i}\hat{\varrho}_{ik}\prod_{l=1,l\neq k}^{K}\|\hat{Q}_{1i}(\hat{\alpha}_k-\hat{\alpha}_l)\| + \sum_{i\in\hat{G}_0}(\tilde{\sigma}_i)^{2-K}\hat{Q}_{1i}\hat{\varrho}_{ik}\prod_{l=1,l\neq k}^{K}\|\hat{Q}_{1i}(\hat{\beta}_{1,i}-\hat{\alpha}_l)\|.
\end{aligned} \tag{79}$$

Averaging both sides of (75) over $i\in\hat{G}_k$ and using (78) and (79), we have

$$\mathbf{0}_{p_1\times 1} = \frac{2}{N_kT^2}\sum_{i\in\hat{G}_k}\tilde{x}_{1,i}'(\tilde{y}_i-\tilde{x}_{1,i}\hat{\alpha}_k-\tilde{x}_{2,i}\hat{\beta}_{2,i})+\frac{\lambda}{N_k}\sum_{i\in\hat{G}_0}(\tilde{\sigma}_i)^{2-K}\hat{Q}_{1i}\hat{\varrho}_{ik}\prod_{l=1,l\neq k}^{K}\|\hat{Q}_{1i}(\hat{\beta}_{1,i}-\hat{\alpha}_l)\|. \tag{80}$$

Solving $\hat{\beta}_{2,i}$ from (76) as a function of $\hat{\beta}_{1,i}$ and replacing $\hat{\beta}_{1,i}$ by $\hat{\alpha}_k$ for $i\in\hat{G}_k$ yields

$$\hat{\beta}_{2,i} = \left(\tilde{x}_{2,i}'\tilde{x}_{2,i}\right)^{-1}\tilde{x}_{2,i}'(\tilde{y}_i-\tilde{x}_{1,i}\hat{\alpha}_k). \tag{81}$$

Plugging (81) into (80) yields

$$\begin{aligned}
\hat{\alpha}_k &= \left(\frac{1}{N_kT^2}\sum_{i\in\hat{G}_k}\tilde{x}_{1,i}'M_{2,i}\tilde{x}_{1,i}\right)^{-1}\frac{1}{N_kT^2}\sum_{i\in\hat{G}_k}\tilde{x}_{1,i}'M_{2,i}\tilde{y}_i \\
&\quad + \left(\frac{1}{N_kT^2}\sum_{i\in\hat{G}_k}\tilde{x}_{1,i}'M_{2,i}\tilde{x}_{1,i}\right)^{-1}\frac{\lambda}{2N_k}\sum_{i\in\hat{G}_0}(\tilde{\sigma}_i)^{2-K}\hat{e}_{ik}\prod_{l=1,l\neq k}^{K}\|\hat{Q}_{1i}(\hat{\beta}_{1,i}-\hat{\alpha}_l)\| \\
&\equiv \hat{\alpha}_k^{\text{post}}+\hat{R}_k,\ \text{say.}
\end{aligned}$$

Noting that $\hat{Q}_{1i}\hat{\varrho}_{ik}\prod_{l=1,l\neq k}^{K}\|\hat{Q}_{1i}(\hat{\beta}_{1,i}-\hat{\alpha}_l)\|\neq 0$ only if $i\in\hat{G}_0$ and by (72), we have that for any $\epsilon>0$

$$P\left(\sqrt{N}T||\hat{R}_k||\geqslant\epsilon\right)\leqslant\sum_{k=1}^{K}\sum_{i\in G_k^0}P(i\in\hat{G}_0|i\in G_k^0)\leqslant\sum_{k=1}^{K}\sum_{i\in G_k^0}P(i\notin\hat{G}_k|i\in G_k^0)=o(1).$$

That is, $\sqrt{NT}||\hat{R}_k|| = o_P(1)$ and $\hat{\alpha}_k$ is asymptotically equivalent to its post-Lasso estimator $\hat{\alpha}_{\hat{G}_k}$. Similarly, given the fast convergence rate of $\hat{\alpha}_{\hat{G}_k}$, $\hat{\beta}_{2,i}$ in (81) is also asymptotically equivalent to its post-Lasso version $\hat{\beta}_{2,i}^{\text{post}}$, where $\hat{\beta}_{2,i}^{\text{post}} = \left(\tilde{x}_{2,i}'\tilde{x}_{2,i}\right)^{-1}\tilde{x}_{2,i}'(\tilde{y}_i - \tilde{x}_{1,i}\hat{\alpha}_k^{\text{post}})$ for each $i \in \hat{G}_k$. We formally study the asymptotic properties of $\hat{\alpha}_k^{\text{post}}$ and $\hat{\beta}_{2,i}^{\text{post}}$ in the proof of Theorem 5 below. ■

Proof of Theorem 5. (i) Noting that $\tilde{y}_i = \tilde{x}_{1,i}\beta_{1,i}^0 + \tilde{x}_{2,i}\beta_{2,i}^0 + \tilde{u}_i$, we have

$$\sqrt{N_k}T(\hat{\alpha}_k^{\text{post}} - \alpha_k^0) = \hat{Q}_{(k)}^{-1}\hat{V}_{(k)} + \hat{Q}_{(k)}^{-1}\hat{R}_{(k)},$$

where $\hat{Q}_{(k)} = \frac{1}{N_kT^2}\sum_{i\in\hat{G}_k}\tilde{x}_{1,i}'M_{2,i}\tilde{x}_{1,i}$, $\hat{V}_{(k)} = \frac{1}{\sqrt{N_k}T}\sum_{i\in\hat{G}_k}\tilde{x}_{1,i}'M_{2,i}\tilde{u}_i$, and $\hat{R}_{(k)} = \frac{1}{\sqrt{N_k}T}\sum_{i\in\hat{G}_k}\tilde{x}_{1,i}'M_{2,i} \times\tilde{x}_{1,i}\left(\beta_{1,i}^0 - \alpha_k^0\right)$. Noting that $\mathbf{1}\{i \in \hat{G}_k\} = \mathbf{1}\{i \in G_k^0\} + \mathbf{1}\{i \in \hat{G}_k\backslash G_k^0\} - \mathbf{1}\{i \in G_k^0\backslash\hat{G}_k\}$, we have

$$\begin{aligned}\hat{Q}_{(k)} &= \frac{1}{N_kT^2}\sum_{i\in G_k^0}\tilde{x}_{1,i}'M_{2,i}\tilde{x}_{1,i} + \frac{1}{N_kT^2}\sum_{i\in\hat{G}_k\backslash G_k^0}\tilde{x}_{1,i}'M_{2,i}\tilde{x}_{1,i} - \frac{1}{N_kT^2}\sum_{i\in G_k^0\backslash\hat{G}_k}\tilde{x}_{1,i}'M_{2,i}\tilde{x}_{1,i}\\ &\equiv Q_{k,NT} + \hat{Q}_{1(k)} + \hat{Q}_{2(k)}, \text{ say.}\end{aligned}$$

By Theorem 3 $P(||\hat{Q}_{1(k)}|| \geqslant \epsilon N^{-1/2}T^{-1}) \leqslant P(\hat{F}_{kNT}) = o(1)$ and $P(||\hat{Q}_{2(k)}|| \geqslant \epsilon N^{-1/2}T^{-1}) \leqslant P(\hat{E}_{kNT}) = o(1)$ for any $\epsilon > 0$. It follows that $\hat{Q}_{(k)} = Q_{k,NT} + o_P\left(N^{-1/2}T^{-1}\right)$. Similarly, we can show that $\hat{V}_{(k)} = V_{k,NT} + o_P\left(N^{-1/2}T^{-1}\right)$ and $\hat{R}_{(k)} = o_P\left(N^{-1/2}T^{-1}\right)$, where $V_{k,NT} = \frac{1}{\sqrt{N_k}T}\sum_{i\in G_k^0}\tilde{x}_{1,i}'M_{2,i}\tilde{u}_i$. It follows that

$$\sqrt{N_k}T(\hat{\alpha}_k^{\text{post}} - \alpha_k^0) = Q_{k,NT}^{-1}V_{k,NT} + o_P(1).$$

Then the conclusion in (i) follows from Lemmas 7(i)-(vi).

(ii) Noting that $\hat{\beta}_{2,i}^{\text{post}} = \left(\tilde{x}_{2,i}'\tilde{x}_{2,i}\right)^{-1}\tilde{x}_{2,i}'(\tilde{y}_i - \tilde{x}_{1,i}\hat{\alpha}_k^{\text{post}})$ and $\tilde{y}_i = \tilde{x}_{1,i}\alpha_k^0 + \tilde{x}_{2,i}\beta_{2,i}^* + \tilde{u}_i^*$ for $i \in G_k^0$, we have for $i \in G_k^0$ and $l \times p_2$ selection matrix $\mathbb{S}_2$,

$$\begin{aligned}\sqrt{T}\mathbb{S}_2\left(\hat{\beta}_{2,i}^{\text{post}} - \beta_{2,i}^*\right) &= \mathbb{S}_2\left(\frac{1}{T}\tilde{x}_{2,i}'\tilde{x}_{2,i}\right)^{-1}\frac{1}{\sqrt{T}}\tilde{x}_{2,i}'\tilde{u}_i^* + \frac{1}{\sqrt{T}}\mathbb{S}_2\left(\frac{1}{T}\tilde{x}_{2,i}'\tilde{x}_{2,i}\right)^{-1}\frac{1}{T}\tilde{x}_{2,i}'\tilde{x}_{1,i}T\left(\alpha_k^0 - \hat{\alpha}_k^{\text{post}}\right)\\ &= \mathbb{S}_2\left(\frac{1}{T}\tilde{x}_{2,i}'\tilde{x}_{2,i}\right)^{-1}\frac{1}{\sqrt{T}}\tilde{x}_{2,i}'\tilde{u}_i^* + O_P\left(T^{-1/2}\right)\\ &\Rightarrow N\left(0, \mathbb{S}_2\mathbb{V}_{22,i}\mathbb{S}_2'\right)\end{aligned}$$

by (i) and Lemmas 1(i) and (iii). Here $\mathbb{V}_{22,i} = \left(\Sigma_{22,i}^{-1}J_{1,i} \otimes J_{2,i}\right)V_i^0\left(J_{1,i}'\Sigma_{22,i}^{-1} \otimes J_{2,i}'\right)$. ■

Proof of Theorem 6. (i) In vector form, we have the regression model:

$$\tilde{y}_i = \tilde{x}_{1,i}\beta_{1,i} + \tilde{x}_{2,i}\beta_{2,i} + \tilde{v}_i^{\dagger}, \tag{82}$$

where $\tilde{x}_{2,i} = (\tilde{x}_{2,i,\bar{p}_2+1}, ..., \tilde{x}_{2,i,T-\bar{p}_2})'$, $\tilde{x}_{2,it} = x_{2,it} - \frac{1}{T-2\bar{p}_2}\sum_{t=\bar{p}_2+1}^{T-\bar{p}_2} x_{2,it}$, $x_{2,it} = (\Delta x'_{1,i,t-\bar{p}_2}, ..., \Delta x'_{1,it}, ..., \Delta x'_{1,i,t+\bar{p}_2})'$, and $\tilde{x}_{1,i}$ and $\tilde{v}_i^{\dagger}$ are similarly defined. In particular, a typical element of $\tilde{v}_i$ is given by $\tilde{v}_{it}^{\dagger} = v_{it}^{\dagger} - \frac{1}{T-2\bar{p}_2}\sum_{t=\bar{p}_2+1}^{T-\bar{p}_2} v_{it}^{\dagger}$, where $v_{it}^{\dagger} = v_{it}^{a} + v_{it}$ and $v_{it}^{a} = \sum_{|j|\geqslant\bar{p}_2} \gamma'_{i,j}\Delta x_{1,i,t-j}$ signifies the approximation error.

Assumption A4 ensures the approximation error term v_{it}^{a} is asymptotically negligible in our asymptotic analysis. Following the proofs of Theorems 1-4, we can prove that the C-Lasso estimator $\hat{\alpha}_k^{D}$ of α_k is asymptotically equivalent to its post-Lasso version $\hat{\alpha}_k^{D,\text{post}}$, where

$$\hat{\alpha}_k^{D,\text{post}} = \left(\sum_{i\in\hat{G}_k} \tilde{x}'_{1,i}M_{2,i}\tilde{x}_{1,i}\right)^{-1} \sum_{i\in\hat{G}_k} \tilde{x}'_{1,i}M_{2,i}\tilde{y}_i.$$

As in the proof Theorem 5, we can show that

$$\sqrt{N_k}T(\hat{\alpha}_k^{\text{post}} - \alpha_k^0) = Q_{k,NT}^{-1}V_{k,NT} + o_P(1),$$

where $Q_{k,NT} = \frac{1}{N_kT^2}\sum_{i\in G_k^0} \tilde{x}'_{1,i}M_{2,i}\tilde{x}_{1,i}$ and $V_{k,NT} = \frac{1}{\sqrt{N_k}T}\sum_{i\in G_k^0} \tilde{x}'_{1,i}M_{2,i}\tilde{v}_i^{\dagger}$. Lemma 7(i) continues to apply: $Q_{k,NT} = \mathbb{Q}_{(k)} + o_P(1)$. Now

$$V_{k,NT} = \frac{1}{\sqrt{N_k}T}\sum_{i\in G_k^0} \tilde{x}'_{1,i}M_{2,i}\tilde{v}_i + \frac{1}{\sqrt{N_k}T}\sum_{i\in G_k^0} \tilde{x}'_{1,i}M_{2,i}\tilde{v}_i^{a} \equiv \mathcal{V}_{k,NT} + \mathcal{V}_{k,NT}^{a}, \text{ say.}$$

Lemma 7(ii)-(vi) continues to apply to $V_{k,NT}(1)$ with little modification. Now, v_{it} plays the role of u_{it}^{*} in the lemma. But since v_{it} is uncorrelated to all lags and leads of $\Delta x_{1,it} = \varepsilon_{1,it}$, s_i defined in Theorem 4 becomes $s_i = S'_0 - S'_2\Sigma_{22,i}^{-1}\Sigma_{20,i} = S'_0$ as $\Sigma_{20,i}$ is now zero. Then

$$\begin{aligned}
\mathbb{B}_{1k,NT} &= \frac{1}{\sqrt{N_k}}\sum_{i\in G_k^0} S_1\sum_{r=0}^{\infty}\sum_{s=0}^{\infty}\psi_{i,s+r}\psi'_{i,s}S'_0 = \frac{1}{\sqrt{N_k}}\sum_{i\in G_k^0}\sum_{t=0}^{\infty}E(\varepsilon_{1,it}v_{i0}) = 0,\\
\mathbb{B}_{2k,NT} &= \frac{-1}{\sqrt{N_k}}\frac{T+1}{2T}\sum_{i\in G_k^0} S_1\psi_i(1)\psi_i(1)'S_0 = \frac{-1}{\sqrt{N_k}}\frac{T+1}{2T}\sum_{i\in G_k^0}\sum_{t=-\infty}^{\infty}E(\varepsilon_{1,it}v_{i0}) = 0.
\end{aligned}$$

It follows that $\mathcal{V}_{k,NT} \Rightarrow N\left(0, \mathbb{V}_{(k)}^{+}\right)$, where $\mathbb{V}_{(k)}^{+} \equiv \lim_{N_k\to\infty}\frac{1}{N_k}\sum_{i\in G_k^0}\frac{1}{6}\Omega_{00,i}^{+}\Omega_{11,i}$,and $\Omega_{00,i}^{+} = \Omega_{00,i} - \Omega_{01,i}\Omega_{11,i}^{-1}\Omega_{10,i}$. In addition, $\mathcal{V}_{k,NT}^{a} = o_P(1)$ by Lemma 8. Consequently, $\sqrt{N_k}T(\hat{\alpha}_k^{\text{post}} - \alpha_k^0) \Rightarrow N(0, \mathbb{Q}_{(k)}^{-1}\mathbb{V}_{(k)}^{+}\mathbb{Q}_{(k)}^{-1})$.

(ii) This follows from Theorem 5(ii) and the fact that $\Sigma_{20,i} = 0$ so that $\beta^*_{2,i} = \beta^0_{2,i}$. ■

A.2 Practical estimation procedures

In this appendix, we describe details on the practical estimation procedure in the followings steps

- **Initial estimators:** Obtain consistent initial estimates of $\tilde{\beta}_{1,i}$ and $\tilde{\beta}_{2,i}$. For model (2.4), we can employ the least square estimators as the initial estimators.

- **Penalized Least Squares:** Solve the PLS problem

$$Q^K_{NT,\lambda}(\boldsymbol{\beta_1}, \boldsymbol{\beta}_2, \boldsymbol{\alpha}) = Q_{NT}(\boldsymbol{\beta_1}, \boldsymbol{\beta}_2) + \frac{\lambda}{N}\sum_{i=1}^{N}(\tilde{\sigma}_i)^{2-K}\prod_{k=1}^{K}\left\|\hat{Q}_{1i}(\beta_{1,i} - \alpha_k)\right\|.$$

  By Assumption A.3(iv), we set the tuning parameter $\lambda = c_\lambda T^{-3/4}$, where c_λ is a constant. Given the tuning parameter $\lambda(N, T)$, we can obtain C-Lasso estimator $\hat{\alpha}_k$ and identify the unknown group structures.

- **Post-Lasso estimator with bias-correction:** Given the estimated groups, $\{\hat{G}_k, k = 1, ..., K\}$, we can obtain the post-Lasso estimators of α_k and $\beta_{2,i}$ as

$$\hat{\alpha}^{\text{post}}_k = \left(\sum_{i\in\hat{G}_k}\tilde{x}'_{1,i}M_{2,i}\tilde{x}_{1,i}\right)^{-1}\sum_{i\in\hat{G}_k}\tilde{x}'_{1,i}M_{2,i}\tilde{y}_i \text{ for } k = 1, ..., K,$$
$$\hat{\beta}^{\text{post}}_{2,i} = \left(\tilde{x}'_{2,i}\tilde{x}_{2,i}\right)^{-1}\tilde{x}'_{2,i}(\tilde{y}_i - \tilde{x}_{1,i}\hat{\alpha}^{\text{post}}_k) \text{ for } i \in \hat{G}_k.$$

  In addition, we apply the dynamic OLS method in post-Lasso estimator for removing the bias. After bias-correction, we compute the standard errors by the following formulas

$$\hat{\mathbb{Q}}_{(k)} = \frac{1}{\hat{N}_k T^2}\sum_{i\in\hat{G}_k}\tilde{x}'_{1,i}M_{2,i}\tilde{x}_{1,i} \text{ and } \hat{\mathbb{V}}^+_{(k)} \equiv \frac{1}{\hat{N}_k}\sum_{i\in\hat{G}_k}\frac{1}{6}\hat{\Omega}^+_{00,i}\hat{\Omega}_{11,i},$$

  where $\hat{\Omega}_{00,i}$ and $\hat{\Omega}_{11,i}$ denote the HAC estimator of the long-run variance-covariance components $\Omega_{00,i}$ and $\Omega_{11,i}$ in Ω_i.

- **Determine the number of groups by BIC:** In practice, K is typically unknown. Let $\hat{\sigma}^2_{\hat{G}_k(K,\lambda)} = \frac{1}{NT}\sum_{k=1}^{K}\sum_{i\in\hat{G}_k(K,\lambda)}\sum_{t=1}^{T}[\hat{u}_{it}(k)]^2$, where $\hat{u}_{it}(k) = \tilde{y}_{it} - \tilde{x}'_{1,it}\hat{\alpha}^{\text{post}}_{\hat{G}_k(K,\lambda)} - \tilde{x}'_{2,it}\hat{\beta}^{\text{post}}_{2,i}(\hat{G}_k(K,\lambda))$ for $i \in \hat{G}_k(K,\lambda)$. We choose $\hat{K}$ to minimize the following information criterion:

$$IC(K,\lambda) = \ln[\hat{\sigma}^2_{\hat{G}_k(K,\lambda)}] + \rho_{NT}p_1K$$

By Assumption A.5(ii), we set tuning parameter $\rho_{NT} = \frac{1}{3}(NT)^{-1/3}$.

B Appendix to Chapter 2

In this appendix, we prove the main results, namely, Theorems 7-14 in the paper. The proofs of these results need some technical lemmas whose proofs are relegated to the online supplementary Appendix.

B.1 Proofs of the Main Results in Section 3

To proceed, we define some notations.

(i) Let $H_1 = \left(\frac{1}{N}\Lambda_1^{0\prime}\Lambda_1^0\right)\left(\frac{1}{T^2}f_1^{0\prime}\hat{f}_1\right)V_{1,NT}^{-1}$, $H_2 = \left(\frac{1}{N}\Lambda_2^{0\prime}\Lambda_2^0\right)\left(\frac{1}{T}f_2^{0\prime}\hat{f}_2\right)V_{2,NT}^{-1}$ and $a_{ij} = \lambda_{1i}^{0\prime}(\frac{\Lambda_1^{0\prime}\Lambda_1^0}{N})^{-1}\lambda_{1j}^0$.

(ii) Let $\mathbf{b} = (b_1, ..., b_N)$ and b =vec($\mathbf{b}$), where $b_i = \beta_i - \beta_i^0$ for $i = 1, ..., N$. Let $\hat{\boldsymbol{b}} = (\hat{b}_1, ..., \hat{b}_N)$ and $\hat{b}$ =vec($\hat{\boldsymbol{b}}$), where $\hat{b}_i = \hat{\beta}_i - \beta_i^0$.

(iii) Let $\eta_{NT}^2 = \frac{1}{N}\sum_{i=1}^N \|\hat{b}_i\|^2$, $\varrho_{NT}^2 = \frac{1}{K}\sum_{k=1}^K \left\|\hat{\alpha}_k - \alpha_k^0\right\|^2$, $C_{NT} = \min(\sqrt{N}, \sqrt{T})$, $\delta_{NT} = \min(\sqrt{N}, T)$, and $\psi_{NT} = N^{1/q}T^{-1}(\log T)^{1+\epsilon}$ for some $\epsilon > 0$.

(iv) Let $\hat{Q}_{i,xx} = \frac{1}{T^2}x_i'M_{\hat{f}_1}x_i$, $Q_{i,xx} = \frac{1}{T^2}x_i'M_{f_1}x_i$, and $Q_{i,xx}(f_1^0) = \frac{1}{T^2}x_i'M_{f_1^0}x_i$.

(v) Without loss of generality, we set $x_{i0} = 0$ throughout the proof of the main results and supplementary Appendix.

To prove Theorem 3.1, we need four lemmas.

Lemma A.1 *Suppose that Assumptions 3.1 hold. Then for each $i = 1, ..., N$,*

(i) $\frac{1}{T^2}x_i'M_{f_1^0}x_i \Rightarrow \int \tilde{B}_{2i}\tilde{B}_{2i}'$,

(ii) $\frac{1}{T}x_i'M_{f_1^0}u_i \Rightarrow \int (B_{2i} - \pi_i'B_3)\, dB_{1i} + (\Delta_{21,i} - \pi_i'\Delta_{31,i})$,

where $\tilde{B}_{2i} = B_{2i} - \int B_{2i}B_3' \left(\int B_3B_3'\right)^{-1} B_3$ and $\pi_i = \left(\int B_3B_3'\right)^{-1}\int B_3B_{2i}'$.

Lemma A.2 *Suppose that Assumptions 3.1-3.2 hold. Then for any fixed small constant $c \in (0, 1/2)$,*

(i) $\limsup_{T\to\infty}\mu_{\max}\left(\frac{W_i'W_i}{d_TT^2}\right) \leqslant (1+c)\rho_{\max}$ *a.s.*,

(ii) $\liminf_{T\to\infty}\mu_{\min}\left(\frac{d_TW_i'W_i}{T^2}\right) \geqslant c\rho_{\min}$ *a.s.*,

(iii) $\limsup_{T\to\infty}\mu_{\max}\left(\frac{x_i'M_{f_1^0}x_i}{d_TT^2}\right) \leqslant (1+c)\rho_{\max}$ *a.s.*,

(iv) $\liminf_{T\to\infty}\mu_{\min}\left(\frac{d_Tx_i'M_{f_1^0}x_i}{T^2}\right) \geqslant [(1+c)\rho_{\max}]^{-1}$ *a.s.*,

where $W_{it} = (x_{it}', f_{1t}^{0\prime})'$ and $W_i = (W_{i,1}, W_{i,2}, ..., W_{i,T})'$.

Lemma A.3 *Suppose that Assumptions 3.1-3.2 hold. Then*

(i) $\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{1}{T^2}x_i'M_{f_1^0}u_i\right\|^2 = O_P(d_T^2T^{-2})$,

(ii) $\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{1}{T^2}x_i'M_{f_1^0}u_i^*\right\|^2 = O_P(d_T^2T^{-2})$,

(iii) $\left\|\frac{1}{NT^2}\sum_{j=1}^{N}x_i'M_{f_1^0}u_ja_{ij}\right\| = O_P(d_TT^{-1})$,

(iv) $\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{1}{T^2}x_i'M_{f_1^0}x_i\right\| = O_P(d_T)$,

where f_1 *satisfies* $\frac{1}{T^2}f_1'f_1 = I_{r_1}$ *and* $u_i^* = u_i + f_2^0\lambda_{2i}^0$.

Lemma A.4 *Suppose that Assumptions 3.1-3.2 hold. Then*

(i) $\sup_{f_1}\sup_{N^{-1}\|\mathbf{b}\|^2\leqslant M}\left\|\frac{1}{NT^2}\sum_{i=1}^{N}b_i'x_i'M_{f_1}u_i^*\right\| = o_P(d_T^{-3})$,

(ii) $\sup_{f_1}\left\|\frac{1}{NT^2}\sum_{i=1}^{N}\lambda_{1i}^{0\prime}f_1^{0\prime}M_{f_1}u_i^*\right\| = o_P(d_T^{-3})$,

(iii) $\sup_{f_1}\left\|\frac{1}{NT^2}\sum_{i=1}^{N}u_i^{*\prime}P_{f_1}u_i^*\right\| = o_P(d_T^{-3})$,

where the sup *is taken with respect to* f_1 *such that* $\frac{f_1'f_1}{T^2} = I_{r_1}$ *and* u_i^* *are defined in Lemma A.3.*

Proof of Theorem 3.1. (i) Let $Q_{i,NT}(\beta_i, f_1) = \frac{1}{T^2}(y_i - x_i\beta_i)'M_{f_1}(y_i - x_i\beta_i)$ and $Q_{i,NT}^{K,\lambda}(\beta_i, \alpha, f_1) = Q_{i,NT}(\beta_i, f_1) + \lambda\prod_{k=1}^{K}\|\beta_i - \alpha_k\|$. Then $Q_{NT}^{K,\lambda}(\boldsymbol{\beta}, \boldsymbol{\alpha}, f_1) = \frac{1}{N}\sum_{i=1}^{N}Q_{i,NT}^{K,\lambda}(\beta_i, \alpha, f_1)$. Noting that $y_i - x_i\beta_i = -x_ib_i + \lambda_{1i}^{0\prime}f_1^{0\prime} + u_i^*$, we have

$$\begin{aligned}Q_{i,NT}(\beta_i, f_1) - Q_{i,NT}(\beta_i^0, f_1^0) =& \frac{1}{T^2}(b_i'x_i'M_{f_1}x_ib_i + \lambda_{1i}^{0\prime}f_1^{0\prime}M_{f_1}f_1^0\lambda_{1i}^0 - 2b_i'x_i'M_{f_1}f_1^0\lambda_{1i}^0)\\ &+ \frac{1}{T^2}(2\lambda_{1i}^{0\prime}f_1^{0\prime}M_{f_1}u_i^* - 2b_i'x_i'M_{f_1}u_i^*) - \frac{1}{T^2}u_i^{*\prime}(P_{f_1} - P_{f_1^0})u_i^*,\end{aligned} \tag{83}$$

where $u_i^* = u_i + f_2^0\lambda_{2i}^0$. Let $S_{i,NT}(\beta_i, f_1) = \frac{1}{T^2}\left(b_i'x_i'M_{f_1}x_ib_i + \lambda_{1i}^{0\prime}f_1^{0\prime}M_{f_1}f_1^0\lambda_{1i}^0 - 2b_i'x_i'M_{f_1}f_1^0\lambda_{1i}^0\right)$. Then we have

$$\begin{aligned}Q_{NT}(\boldsymbol{\beta}, f_1) - Q_{NT}(\boldsymbol{\beta}^0, f_1^0) =& \frac{1}{N}\sum_{i=1}^{N}S_{i,NT}(\beta_i, f_1)\\ &+ \frac{1}{NT^2}\sum_{i=1}^{N}\left(2\lambda_{1i}^{0\prime}f_1^{0\prime}M_{f_1}u_i^* - 2b_i'x_i'M_{f_1}u_i^* - u_i^{*\prime}(P_{f_1} - P_{f_1^0})u_i^*\right)\\ =& \frac{1}{N}\sum_{i=1}^{N}S_{i,NT}(\beta_i, f_1) + o_P(d_T^{-3}),\end{aligned} \tag{84}$$

where the last three terms on the right hand side of (84) are $o_P(d_T^{-3})$ uniformly in $\{b_i\}$ and f_1 such that $\frac{f_1'f_1}{T^2} = I_{r_1}$ and $\frac{1}{N}\sum_{i=1}^{N}\|b_i\|^2 \leqslant M$ by Lemma A.4(i)-(iii) and the fact that $\frac{1}{NT^2}\sum_{i=1}^N u_i^{*\prime}P_{f_1^0}u_i^* = o_P(d_T^{-3})$. Then we have

$$\begin{aligned} Q_{NT}^{K,\lambda}(\boldsymbol{\beta},\hat{\boldsymbol{\alpha}},f_1) - Q_{NT}^{K,\lambda}(\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0,f_1^0) =& \frac{1}{N}\sum_{i=1}^N [Q_{NT,i}(\beta_i,f_1) - Q_{NT,i}(\beta_i^0,f_1^0)] + \frac{\lambda}{N}\sum_{i=1}^N\prod_{k=1}^{K^0}\|\beta_i - \hat{\alpha}_k\| \\ \geqslant& S_{NT}(\boldsymbol{\beta},f_1) + o_P(d_T^{-3}). \end{aligned} \tag{85}$$

where $S_{NT}(\boldsymbol{\beta},f_1) = \frac{1}{N}\sum_{i=1}^N S_{i,NT}(\beta_i,f_1)$. Then by (84) and (85) and the fact that $Q_{NT}^{K,\lambda}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\alpha}},\hat{f}_1) - Q_{NT}^{K,\lambda}(\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0,f_1^0) \leqslant 0$, we have

$$S_{NT}(\hat{\boldsymbol{\beta}},\hat{f}_1) = \frac{1}{NT^2}\sum_{i=1}^N \left[\hat{b}_i'x_i'M_{\hat{f}_1}x_i\hat{b}_i + \lambda_{1i}^{0\prime}f_1^{0\prime}M_{\hat{f}_1}f_1^0\lambda_{1i}^0 - 2\hat{b}_i'x_i'M_{\hat{f}_1}f_1^0\lambda_{1i}^0\right] = o_P(d_T^{-3}). \tag{86}$$

Similarly, by (84), (85) and Lemma A.4(i)-(iii), we have

$$\begin{aligned} Q_{NT}^{K,\lambda}(\boldsymbol{\beta},\hat{\boldsymbol{\alpha}},\hat{f}_1) - Q_{NT}^{K,\lambda}(\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0,\hat{f}_1) =& \frac{1}{N}\sum_{i=1}^N [Q_{NT,i}(\beta_i,\hat{f}_1) - Q_{NT,i}(\beta_i^0,\hat{f}_1)] + \frac{\lambda}{N}\sum_{i=1}^N\prod_{k=1}^{K_0}\|\beta_i - \hat{\alpha}_k\| \\ \geqslant& \frac{1}{NT^2}\sum_{i=1}^N \left[b_i'x_i'M_{\hat{f}_1}x_ib_i - 2b_i'x_i'M_{\hat{f}_1}f_1^0\lambda_{1i}^0\right] + o_P(d_T^{-3}). \end{aligned} \tag{87}$$

This, in conjunction with the fact that $Q_{NT}^{K,\lambda}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\alpha}},\hat{f}_1) - Q_{NT}^{K,\lambda}(\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0,\hat{f}_1) \leqslant 0$, implies that

$$\frac{1}{NT^2}\sum_{i=1}^N \left[\hat{b}_i'x_i'M_{\hat{f}_1}x_i\hat{b}_i - 2\hat{b}_i'x_i'M_{\hat{f}_1}f_1^0\lambda_{1i}^0\right] \leqslant o_P(d_T^{-3}). \tag{88}$$

Combining (86) and (88) yields that

$$o_P(d_T^{-3}) = \frac{1}{NT^2}\lambda_{1i}^{0\prime}f_1^{0\prime}M_{\hat{f}_1}f_1^0\lambda_{1i}^0 = \text{tr}\left[\left(\frac{f_1^{0\prime}M_{\hat{f}_1}f_1^0}{T^2}\right)\left(\frac{\Lambda_1^{0\prime}\Lambda_1^0}{N}\right)\right] \geqslant \text{tr}\left(\frac{f_1^{0\prime}M_{\hat{f}_1}f_1^0}{T^2}\right)\mu_{\min}\left(\frac{\Lambda_1^{0\prime}\Lambda_1^0}{N}\right).$$

It follows that $\text{tr}\left(\frac{f_1^{0\prime}M_{\hat{f}_1}f_1^0}{T^2}\right) = o_P(d_T^{-3})$ as $\mu_{\min}\left(\frac{\Lambda_1^{0\prime}\Lambda_1^0}{N}\right)$ is bounded away from zero in probability by Assumption 3.2(i). As in Bai (2009, p.1265), this implies that

$$\frac{f_1^{0\prime}M_{\hat{f}_1}f_1^0}{T^2} = \frac{f_1^{0\prime}f_1^0}{T^2} - \frac{f_1^{0\prime}\hat{f}_1}{T^2}\frac{\hat{f}_1{}'f_1^0}{T^2} = o_P(d_T^{-3}), \tag{89}$$

and $\frac{f_1^{0\prime}\hat{f}_1}{T^2}$ is asymptotically invertible by the fact that $\frac{f_1^{0\prime}f_1^0}{T^2}$ is asymptotically invertible from Assumption 3.2(ii). (89) implies that $\frac{\hat{f}_1'P_{f_1^0}\hat{f}_1}{T^2} - I_{r_1} = o_P(d_T^{-3})$, which further implies that $\left\|P_{\hat{f}_1} - P_{f_1^0}\right\|^2 = 2\mathrm{tr}\left(I_{r_1} - \frac{\hat{f}_1{}' P_{f_1^0} \hat{f}_1}{T^2}\right) = o_P(d_T^{-3})$. By Cauchy-Schwarz inequality and (62),

$$o_P(d_T^{-3}) \geqslant \frac{1}{NT^2}\sum_{i=1}^{N} \hat{b}_i' x_i' M_{\hat{f}_1} x_i \hat{b}_i - 2\left\{\frac{1}{NT^2}\sum_{i=1}^{N} \hat{b}_i' x_i' M_{\hat{f}_1} x_i \hat{b}_i\right\}^{1/2} \left\{\frac{1}{NT^2}\lambda_{1i}^{0\prime} f_1^{0\prime} M_{\hat{f}_1} f_1^0 \lambda_{1i}^0\right\}^{1/2}. \tag{90}$$

This result, in conjunction with (64), implies that $\frac{1}{NT^2}\sum_{i=1}^{N} \hat{b}_i' x_i' M_{\hat{f}_1} x_i \hat{b}_i = o_P(d_T^{-3})$. So we have shown parts (i) and (ii) in the theorem.

(iii) By the results in parts (i) and (ii) and Lemma A.2(i) and (iv), we have

$$\begin{aligned}
o_P(d_T^{-3}) &= \frac{1}{N}\sum_{i=1}^{N} \hat{b}_i' \left(\frac{1}{T^2} x_i' M_{\hat{f}_1} x_i\right) \hat{b}_i \\
&= \frac{1}{N}\sum_{i=1}^{N} \hat{b}_i' \left(\frac{1}{T^2} x_i' M_{f_1^0} x_i\right) \hat{b}_i + \frac{1}{N}\sum_{i=1}^{N} \hat{b}_i' \left(\frac{1}{T^2} x_i' (M_{\hat{f}_1} - M_{f_1^0}) x_i\right) \hat{b}_i \\
&\geqslant \frac{1}{d_T} \min_{1\leqslant i\leqslant N} \mu_{\min}\left(\frac{d_T}{T^2} x_i' M_{f_1^0} x_i\right) \frac{1}{N}\sum_{i=1}^{N} \|\hat{b}_i\|^2 - \max_{1\leqslant i\leqslant N} \frac{\|x_i\|^2}{T^2} \|P_{f_1^0} - P_{\hat{f}_1}\| \frac{1}{N}\sum_{i=1}^{N} \|\hat{b}_i\|^2 \\
&\geqslant \frac{1}{d_T}\left(c\rho_{\min} - o_P(d_T^{-1})\right) \frac{1}{N}\sum_{i=1}^{N} \|\hat{b}_i\|^2,
\end{aligned}$$

where the second inequality follows from the fact that $\min_{1\leqslant i\leqslant N} \mu_{\min}\left(\frac{d_T}{T^2} x_i' M_{f_1^0} x_i\right) \geqslant c\rho_{\min} > 0$ a.s. by Lemma A.2(iv), and $\max_{1\leqslant i\leqslant N} \frac{\|x_i\|^2}{T^2} \leqslant \max_{1\leqslant i\leqslant N} d_T \mu_{\max}\left(\frac{x_i' x_i}{d_T T^2}\right) = O_P(d_T)$ by Lemma A.2(i). Then we have $\frac{1}{N}\sum_{i=1}^{N} \|\hat{b}_i\|^2 = o_P(d_T^{-2}) = o_P(1)$.

(iv) We want to establish the consistency of the estimated factor space $\hat{f}_1$, which extends the results of Bai and Ng (2004) and Bai (2009). Our model allows for the heterogeneous slope coefficients and unobserved stationary common factors. We estimate $\hat{f}_1$ from equation (24) in Section 2.2 as follows

$$\left[\frac{1}{NT^2}\sum_{i=1}^{N} (y_i - x_i\hat{\beta}_i)(y_i - x_i\hat{\beta}_i)'\right] \hat{f}_1 = \hat{f}_1 V_{1,NT}. \tag{91}$$

Combining (91) and the fact that $y_i - x_i\hat{\beta}_i = -x_i\hat{b}_i + f^0\lambda_i^0 + u_i = -x_i\hat{b}_i + f_1^0\lambda_{1i}^0 + f_2^0\lambda_{2i}^0 + u_i$, we have

$$\begin{aligned}
\hat{f}_1 V_{1,NT} =& \frac{1}{NT^2}\sum_{i=1}^{N} x_i\hat{b}_i\hat{b}_i' x_i'\hat{f}_1 - \frac{1}{NT^2}\sum_{i=1}^{N} x_i\hat{b}_i\lambda_i^{0\prime} f^{0\prime}\hat{f}_1 - \frac{1}{NT^2}\sum_{i=1}^{N} x_i\hat{b}_i u_i'\hat{f}_1 \\
&- \frac{1}{NT^2}\sum_{i=1}^{N} f^0\lambda_i^0\hat{b}_i' x_i'\hat{f}_1 - \frac{1}{NT^2}\sum_{i=1}^{N} u_i\hat{b}_i' x_i'\hat{f}_1 + \frac{1}{NT^2}\sum_{i=1}^{N} f^0\lambda_i^0 u_i'\hat{f}_1 \\
&+ \frac{1}{NT^2}\sum_{i=1}^{N} u_i\lambda_i^{0\prime} f^{0\prime}\hat{f}_1 + \frac{1}{NT^2}\sum_{i=1}^{N} u_i u_i'\hat{f}_1 + \frac{1}{NT^2}\sum_{i=1}^{N} f_2^0\lambda_{2i}^0\lambda_{2i}^{0\prime} f_2^{0\prime}\hat{f}_1 \\
&+ \frac{1}{NT^2}\sum_{i=1}^{N} f_1^0\lambda_{1i}^0\lambda_{2i}^{0\prime} f_2^{0\prime}\hat{f}_1 + \frac{1}{NT^2}\sum_{i=1}^{N} f_2^0\lambda_{2i}^0\lambda_{1i}^{0\prime} f_1^{0\prime}\hat{f}_1 + \frac{1}{NT^2}\sum_{i=1}^{N} f_1^0\lambda_{1i}^0\lambda_{1i}^{0\prime} f_1^{0\prime}\hat{f}_1 \\
\equiv& I_1 + ... + I_{11} + \frac{1}{NT^2}\sum_{i=1}^{N} f_1^0\lambda_{1i}^0\lambda_{1i}^{0\prime} f_1^{0\prime}\hat{f}_1, \text{ say.}
\end{aligned}$$

It follows that $\hat{f}_1 V_{1,NT} - f_1^0\left(\frac{\Lambda_1^{0\prime}\Lambda_1^0}{N}\right)\left(\frac{f_1^{0\prime}\hat{f}_1}{T^2}\right) = I_1 + ... + I_{11}$. Let $H_1 = \left(\frac{\Lambda_1^{0\prime}\Lambda_1^0}{N}\right)\left(\frac{f_1^{0\prime}\hat{f}_1}{T^2}\right)V_{1,NT}^{-1}$. Then it is easy to see that $H_1 = O_P(1)$, it is asymptotically nonsingular, and

$$\hat{f}_1 H_1^{-1} - f_1^0 = [I_1 + ... + I_{11}]\left(\frac{f_1^{0\prime}\hat{f}_1}{T^2}\right)^{-1}\left(\frac{\Lambda_1^{0\prime}\Lambda_1^0}{N}\right)^{-1}.$$

Note that

$$\frac{1}{T}\left\|\hat{f}_1 H^{-1} - f_1^0\right\| \leqslant \frac{1}{T}(\|I_1\| + ... + \|I_{11}\|)\left\|\left(\frac{f_1^{0\prime}\hat{f}_1}{T^2}\right)^{-1}\right\|\left\|\left(\frac{\Lambda_1^{0\prime}\Lambda_1^0}{N}\right)^{-1}\right\|.$$

It remains to analyze $\|I_l\|$ for $l = 1, 2, ..., 11$. For I_1, we have that by the result in (iii),

$$\begin{aligned}
\frac{1}{T}\|I_1\| = \frac{1}{T}\left\|\frac{1}{NT^2}\sum_{i=1}^{N} x_i\hat{b}_i\hat{b}_i' x_i'\hat{f}_1\right\| &\leqslant \frac{1}{N}\sum_{i=1}^{N}\frac{\|x_i\|}{T}\|\hat{b}_i\|^2\frac{\|x_i'\hat{f}_1\|}{T^2} \\
&\leqslant \max_{1\leqslant i\leqslant N}\frac{\|x_i\|^2}{T^2}\frac{\|\hat{f}_1\|}{T}\frac{1}{N}\sum_{i=1}^{N}\|\hat{b}_i\|^2 = O_P(d_T\eta_{NT}^2) = o_P(\eta_{NT}),
\end{aligned}$$

where we use the fact that $\max_{1\leqslant i\leqslant N}\frac{\|x_i\|^2}{T^2} = O_P(d_T)$ by Lemma A.2(i) and $\frac{\|\hat{f}_1\|}{T} \leqslant \sqrt{r_1}$. For I_2, we have

$$\begin{aligned}
\frac{1}{T}\|I_2\| = \frac{1}{T}\left\|\frac{1}{NT^2}\sum_{i=1}^{N} x_i\hat{b}_i\lambda_i^{0\prime} f^{0\prime}\hat{f}_1\right\| &\leqslant \frac{\|f^{0\prime}\hat{f}_1\|}{T^2}\max_{1\leqslant i\leqslant N}\frac{\|x_i\|}{T}\frac{1}{N}\sum_{i=1}^{N}\|\hat{b}_i\lambda_i^{0\prime}\| \\
\leqslant \frac{\|f^{0\prime}\hat{f}_1\|}{T^2}\max_{1\leqslant i\leqslant N}\frac{\|x_i\|}{T}&\left\{\frac{1}{N}\sum_{i=1}^{N}\|\hat{b}_i\|^2\right\}^{1/2}\left\{\frac{1}{N}\sum_{i=1}^{N}\|\lambda_i^0\|^2\right\}^{1/2} = O_P(\sqrt{d_T}\eta_{NT}),
\end{aligned}$$

where we use the fact that $\frac{\|f^{0\prime}\hat{f}_1\|}{T^2} = O_P(1)$ and $\frac{1}{N}\sum_{i=1}^N \|\lambda_i^0\|^2 = O_P\left(1\right)$ by Assumption 3.2(i). For I_3,

$$\frac{1}{T}\|I_3\| = \frac{1}{T}\left\|\frac{1}{NT^2}\sum_{i=1}^N x_i b_i u_i' \hat{f}_1\right\| \leqslant \max_{1\leqslant i\leqslant N}\frac{\|x_i\|}{T}\frac{1}{NT^2}\sum_{i=1}^N \|\hat{b}_i u_i'\hat{f}_1\|$$

$$\leqslant \frac{1}{\sqrt{T}}\frac{\|\hat{f}_1\|}{T}\max_{1\leqslant i\leqslant N}\frac{\|x_i\|}{T}\left\{\frac{1}{N}\sum_{i=1}^N\|\hat{b}_i\|^2\right\}^{1/2}\left\{\frac{1}{N}\sum_{i=1}^N\frac{\|u_i\|^2}{T}\right\}^{1/2} = O_P\left(\sqrt{\frac{d_T}{T}}\eta_{NT}\right),$$

where $\frac{1}{N}\sum_{i=1}^N\frac{\|u_i\|^2}{T} = O_P(1)$ by Assumption 3.1(i). For I_4,

$$\frac{1}{T}\left\|I_4\right\| = \left\|\frac{1}{NT^2}\sum_{i=1}^N f^0\lambda_i^0\hat{b}_i' x_i'\hat{f}_1\right\| \leqslant \frac{1}{N}\sum_{i=1}^N\frac{\|f^0\|}{T}\left\|\lambda_i^0\hat{b}_i'\right\|\left\|\frac{x_i'\hat{f}_1}{T^2}\right\|$$

$$\leqslant \frac{\|f^0\|}{T}\frac{\|\hat{f}_1\|}{T}\max_{1\leqslant i\leqslant N}\frac{\|x_i\|}{T}\left\{\frac{1}{N}\sum_{i=1}^N\|\hat{b}_i\|^2\right\}^{1/2}\left\{\frac{1}{N}\sum_{i=1}^N\|\lambda_i^0\|^2\right\}^{1/2} = O_P(\sqrt{d_T}\eta_{NT}).$$

where $\frac{\|f^0\|}{T} \leqslant \frac{\|f_1^0\|}{T} + \frac{1}{\sqrt{T}}\frac{\|f_2^0\|}{\sqrt{T}} = O_P(1)$. For I_5,

$$\frac{1}{T}\|I_5\| = \frac{1}{T}\left\|\frac{1}{NT^2}\sum_{i=1}^N u_i\hat{b}_i' x_i'\hat{f}_1\right\| \leqslant \max_{1\leqslant i\leqslant N}\frac{\|x_i'\hat{f}_1\|}{T^2}\frac{1}{NT}\sum_{i=1}^N\|u_i\hat{b}_i'\|$$

$$\leqslant \frac{1}{\sqrt{T}}\frac{\|\hat{f}_1\|}{T}\max_{1\leqslant i\leqslant N}\frac{\|x_i\|}{T}\left\{\frac{1}{N}\sum_{i=1}^N\frac{\|u_i\|^2}{T}\right\}^{1/2}\left\{\frac{1}{N}\sum_{i=1}^N\|\hat{b}_i\|\right\}^{1/2} = O_P\left(\sqrt{\frac{d_T}{T}}\eta_{NT}\right).$$

For I_6,

$$\frac{1}{T}\|I_6\| =\frac{1}{T}\left\|\frac{1}{NT^2}\sum_{i=1}^N f^0\lambda_i^0 u_i'\hat{f}_1\right\| = \frac{1}{T}\left\|\frac{1}{NT^2}f^0\Lambda^{0\prime}u\hat{f}_1\right\|$$

$$\leqslant \frac{1}{\sqrt{NT}}\left(\frac{1}{T}\left\|\hat{f}_1\right\|\right)\left(\frac{1}{T}\left\|f^0\right\|\right)\frac{1}{\sqrt{NT}}\|\Lambda^{0\prime}u\| = O_P(T^{-1/2}N^{-1/2}),$$

where $u = (u_1, ..., u_N)'$ and we have used the fact that $\frac{1}{NT}\left\|\Lambda^{0\prime}u\right\|^2 = O_P(1)$ by Assumption 3.2(iii). Analogously, we can show that $\frac{1}{T}\|I_7\| = O_P(T^{-1/2}N^{-1/2})$. For I_8,

$$\frac{1}{T^2}\|I_8\|^2 = \frac{1}{T^2}\left\|\frac{1}{NT^2}\sum_{i=1}^N u_i u_i'\hat{f}_1\right\|^2 = \frac{1}{T^2}\left\|\frac{1}{NT^2}u'u\hat{f}_1\right\|^2$$

$$\leqslant 2\sum_{t=1}^T\left\|T^{-3}\sum_{s=1}^T\gamma_N(s,t)\hat{f}_{1s}'\right\|^2 + 2\sum_{t=1}^T\left\|T^{-3}\sum_{s=1}^T\xi_{st}\hat{f}_{1s}'\right\|^2 \equiv 2\left(\|I_8(a)\| + \|I_8(b)\|\right),$$

where $\gamma_N(s,t)$ and ξ_{st} are defined in Assumption 3.2(iii). For $I_8(a)$,

$$\|I_8(a)\|^2 \leqslant \sum_{t=1}^T\left\|T^{-3}\sum_{s=1}^T\gamma_N(s,t)\hat{f}_{1s}'\right\|^2 \leqslant T^{-3}\left(T^{-2}\sum_{s=1}^T\|\hat{f}_{1s}\|^2\right)\left(T^{-1}\sum_{t=1}^T\sum_{s=1}^T\|\gamma_N(s,t)\|^2\right)$$

$$= O_P(T^{-3}),$$

where $T^{-1}\sum_{s=1}^{T}\sum_{t=1}^{T}\|\gamma_N(s,t)\|^2 \leqslant M$ by Assumption 3.2(iii) (see also Lemma 1(i) in Bai and Ng (2002)). For $I_8(b)$,

$$\begin{aligned}\|I_8(b)\| &= \sum_{t=1}^{T}\left\|T^{-3}\sum_{s=1}^{T}\xi_{st}\hat{f}_{1s}'\right\|^2 \leqslant T^{-2}N^{-1}\left(T^{-2}\sum_{s=1}^{T}\left\|\hat{f}_{1s}\right\|^2\right)\left(T^{-2}N\sum_{t=1}^{T}\sum_{s=1}^{T}\|\xi_{st}\|^2\right)\\ &= O_P(T^{-2}N^{-1}),\end{aligned}$$

where we use the fact that $E(\|\xi_{st}\|^2) \leqslant N^{-2}M$ under Assumption 3.2(iii). Then we have $\frac{1}{T}\|I_8\| = O_P(N^{-1/2}T^{-1} + T^{-3/2})$. For $\|I_9\|$,

$$\frac{1}{T}\|I_9\| = \frac{1}{T}\left\|\frac{1}{NT^2}f_2^0\Lambda_2^{0\prime}\Lambda_2^0 f_2^{0\prime}\hat{f}_1\right\| \leqslant \frac{1}{T}\frac{\|f_2^0\|^2}{T}\frac{\|\hat{f}_1\|}{T}\left\|\frac{\Lambda_2^{0\prime}\Lambda_2^0}{N}\right\| = O_P(T^{-1}).$$

For $\|I_{10}\|$,

$$\frac{1}{T}\|I_{10}\| = \frac{1}{T}\left\|\frac{1}{NT^2}f_1^0\Lambda_1^{0\prime}\Lambda_2^0 f_2^{0\prime}\hat{f}_1\right\| \leqslant \frac{1}{\sqrt{NT}}\frac{\|f_1^0\|}{T}\frac{\|f_2^0\|}{\sqrt{T}}\frac{\|\hat{f}_1\|}{T}\frac{\|\Lambda_1^{0\prime}\Lambda_2^0\|}{\sqrt{N}} = O_P((NT)^{-1/2}),$$

where $\frac{\Lambda_1^{0\prime}\Lambda_2^0}{\sqrt{N}} = O_P(1)$ by Assumption 3.2(i). Analogously, we have $\frac{1}{T}\|I_{11}\| = O_P((NT)^{-1/2})$. In sum, we have shown that

$$\frac{1}{T}\left\|\hat{f}_1 H_1^{-1} - f_1^0\right\| = O_P(\sqrt{d_T}\eta_{NT}) + \frac{1}{\sqrt{T}}O_P(C_{NT}^{-1}).$$

Then (iv) follows. ■

To prove Theorem 3.2, we need the following two lemmas.

Lemma A.5 *Suppose that Assumptions 3.1-3.2 hold. Then*

(i) $\frac{1}{T}f_1^{0\prime}(\hat{f}_1 - f_1^0 H_1) = O_P(T\sqrt{d_T}\eta_{NT} + \delta_{NT}^{-1})$,

(ii) $\frac{1}{T}\hat{f}_1'(\hat{f}_1 - f_1^0 H_1) = O_P(T\sqrt{d_T}\eta_{NT} + \delta_{NT}^{-1})$,

(iii) $\|P_{\hat{f}_1} - P_{f_1^0}\|^2 = O_P(\sqrt{d_T}\eta_{NT} + T^{-1}\delta_{NT}^{-1})$,

(iv) $\frac{1}{T}u_k^{*\prime}\left(\hat{f}_1 H_1^{-1} - f_1^0\right) = O_P(\sqrt{Td_T}\eta_{NT} + \delta_{NT}^{-1})$ for each $k = 1, ..., N$.

Lemma A.6 *Suppose that Assumptions 3.1-3.2 hold. Let* $R_{1i} = \frac{1}{T^2}x_i'(P_{f_1^0} - P_{\hat{f}_1})u_i^*$, $R_{2i} = \frac{1}{T^2}x_i'M_{\hat{f}_1}f_1^0\lambda_{1i}^0 - \frac{1}{NT^2}\sum_{j=1}^{N}x_i'M_{\hat{f}_1}x_j a_{ij}\hat{b}_j + \frac{1}{NT^2}\sum_{j=1}^{N}a_{ij}x_i'M_{\hat{f}_1}u_j$, $R_{3i} = \frac{1}{NT^2}\sum_{j=1}^{N}a_{ij}x_i'(P_{f_1^0} - P_{\hat{f}_1})u_j$, *and* $R_{4i} = \frac{1}{T^2}x_i'M_{f_1^0}u_i^* - \frac{1}{NT^2}\sum_{j=1}^{N}a_{ij}x_i'M_{f_1^0}u_j$. *Then*

(i) $R_{1i} = O_P(\varsigma_{1NT})$ *for each* $i = 1, ..., N$, *and* $N^{-1}\sum_{i=1}^{N}\|R_{1i}\|^2 = O_P(\varsigma_{1NT}^2)$,

(ii) $R_{2i} = O_P(\varsigma_{2NT})$ *for each* $i = 1, ..., N$, *and* $N^{-1}\sum_{i=1}^{N}\|R_{2i}\|^2 = O_P(\varsigma_{2NT}^2)$,

(iii) $R_{3i} = O_P(\varsigma_{3NT})$ *for each* $i = 1, ..., N$, *and* $N^{-1}\sum_{i=1}^N \|R_{3i}\|^2 = O_P(\varsigma_{3NT}^2)$,

(iv) $R_{4i} = O_P(T^{-1})$ *for each* $i = 1, ..., N$, *and* $N^{-1}\sum_{i=1}^N \|R_{4i}\|^2 = O_P(T^{-2})$,

where $\varsigma_{1NT} = T^{-1/2}\sqrt{d_T}\eta_{NT} + d_T\eta_{NT}^2 + T^{-1}C_{NT}^{-1}$, $\varsigma_{2NT} = T^{-1}\sqrt{d_T}\eta_{NT} + d_T\eta_{NT}^2 + T^{-1}\delta_{NT}^{-1}$, *and* $\varsigma_{3NT} = T^{-1/2}d_T^{1/4}\eta_{NT}^{1/2} + T^{-1}\delta_{NT}^{-1/2}$.

Proof of Theorem 3.2. (i) Based on the sub-differential calculus, a necessary condition for $\hat{\beta}_i, \hat{\alpha}_k$, and $\hat{f}_1$ to minimize the objective function (2.8) is the for each $i = 1, ..., N$, $0_{p\times 1}$ belongs to the sub-differential of $Q_{NT,\lambda}^K(\beta, \alpha, f_1)$ with respect to β_i (resp. α_k) evaluated at $\{\hat{\beta}_i\}$, $\{\hat{\alpha}_k\}$ and $\hat{f}_1$. That is, for each $i = 1, ..., N$ and $k = 1, ..., K$, we have

$$0_{p\times 1} = -\frac{2}{T^2}x_i'M_{\hat{f}_1}(y_i - x_i\hat{\beta}_i) + \lambda\sum_{j=1}^K \hat{e}_{ij}\prod_{l=1,l\neq j}^K \|\hat{\beta}_i - \hat{\alpha}_l\|, \tag{92}$$

where $\hat{e}_{ij} = \frac{\hat{\beta}_i - \hat{\alpha}_j}{\|\hat{\beta}_i - \hat{\alpha}_j\|}$ if $\|\hat{\beta}_i - \hat{\alpha}_j\| \neq 0$ and $\|\hat{e}_{ij}\| \leqslant 1$ if $\|\hat{\beta}_i - \hat{\alpha}_j\| = 0$. Noting that $y_i = x_i\beta_i^0 + \hat{f}_1H_1^{-1}\lambda_{1i}^0 + u_i^* + (f_1^0 - \hat{f}_1H_1^{-1})\lambda_{1i}^0$, (92) implies that

$$\hat{Q}_{i,xx}\hat{b}_i = \frac{1}{T^2}x_i'M_{\hat{f}_1}u_i^* + \frac{1}{T^2}x_i'M_{\hat{f}_1}f_1^0\lambda_{1i}^0 - \frac{\lambda}{2}\sum_{j=1}^{K_0}\hat{e}_{ij}\prod_{l=1,l\neq j}^K \|\hat{\beta}_i - \hat{\alpha}_l\|, \tag{93}$$

which can be rewritten as

$$\hat{Q}_{i,xx}\hat{b}_i = \frac{1}{NT^2}\sum_{j=1}^N x_i'M_{\hat{f}_1}x_ja_{ij}\hat{b}_j + R_i, \tag{94}$$

where $R_i = R_{1i} + R_{2i} - R_{3i} + R_{4i} - R_{5i}$, R_{1i}, R_{2i}, R_{3i} and R_{4i} are defined in the statement of Lemma A.6, and $R_{5i} = \frac{\lambda}{2}\sum_{j=1}^K \hat{e}_{ij}\prod_{l=1,l\neq j}^K \|\hat{\beta}_i - \hat{\alpha}_l\|$. By Lemma A.6(i)-(iv), we have that $\sum_{l=1}^4 \frac{1}{N}\sum_{i=1}^N \|R_{li}\|^2 = O_P(T^{-1}d_T^{1/2}\eta_{NT} + d_T^2\eta_{NT}^4 + T^{-2}C_{NT}^{-2} + T^{-2}\delta_{NT}^{-1} + T^{-2}) = O_P(T^{-1}d_T^{1/2}\eta_{NT} + d_T^2\eta_{NT}^4 + T^{-2})$. In addition, we can show that $\frac{1}{N}\sum_{i=1}^N \|R_{5i}\|^2 = O_P(\lambda^2)$. It follows that $\frac{1}{N}\sum_{i=1}^N \|R_i\|^2 = O_P(T^{-1}d_T^{1/2}\eta_{NT} + d_T^2\eta_{NT}^4 + T^{-2} + \lambda^2)$.

Let $\hat{Q}_1 = diag(\hat{Q}_{1,xx}, ..., \hat{Q}_{N,xx})$ and $\hat{Q}_2$ as an $Np \times Np$ matrix with typical blocks $\frac{1}{NT^2}x_i'M_{\hat{f}_1}x_ja_{ij}$, such that

$$\hat{Q}_2 = \begin{pmatrix} \frac{1}{NT^2}x_1'M_{\hat{f}_1}x_1a_{11} & \frac{1}{NT^2}x_1'M_{\hat{f}_1}x_2a_{12} & \cdots & \frac{1}{NT^2}x_1'M_{\hat{f}_1}x_Na_{1N} \\ \frac{1}{NT^2}x_2'M_{\hat{f}_1}x_1a_{21} & \frac{1}{NT^2}x_2'M_{\hat{f}_1}x_2a_{22} & \cdots & \frac{1}{NT^2}x_2'M_{\hat{f}_1}x_Na_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{NT^2}x_N'M_{\hat{f}_1}x_1a_{N1} & \frac{1}{NT^2}x_N'M_{\hat{f}_1}x_2a_{N2} & \cdots & \frac{1}{NT^2}x_N'M_{\hat{f}_1}x_Na_{NN} \end{pmatrix}.$$

Let $R = (R_1', ..., R_N')'$. Then (94) implies that $(\hat{Q}_1 - \hat{Q}_2)\hat{\mathbf{b}} = R$. It follows that

$$\|R\|^2 = \text{tr}(\hat{\boldsymbol{b}}'(\hat{Q}_1 - \hat{Q}_2)'(\hat{Q}_1 - \hat{Q}_2)\hat{\boldsymbol{b}}) \geqslant \|\hat{\boldsymbol{b}}\|^2 \left[\mu_{\min}\left(\hat{Q}_1 - \hat{Q}_2\right)\right]^2.$$

By Assumption 3.2(v), we have that $\mu_{\min}\left(\hat{Q}_1 - \hat{Q}_2\right) \geqslant \rho_{\min}/2 > 0$ w.p.a.1. Then we have $\frac{1}{N}\|\hat{\boldsymbol{b}}\|^2 \leqslant \frac{\rho_{\min}^2}{4N}\sum_{i=1}^N \|R_i\|^2 = O_P(T^{-1}d_T^{1/2}\eta_{NT} + d_T^2\eta_{NT}^4 + T^{-2} + \lambda^2) = O_P\left(d_T T^{-2} + \lambda^2\right)$. Consequently, we prove the means square convergence rate of C-Lasso estimators that $\frac{1}{N}\sum_{i=1}^N \|\hat{b}_i\|^2 = O_P\left(d_T T^{-2} + \lambda^2\right)$.

Next, we want to strengthen the last result to a stronger one: $\frac{1}{N}\sum_{i=1}^N \|\hat{b}_i\|^2 = O_P(d_T T^{-2})$. Let $\boldsymbol{\beta} = \boldsymbol{\beta}^0 + d_T T^{-1}\boldsymbol{v}$, where $\boldsymbol{v} = (v_1, ..., v_N)$ is a $p \times N$ matrix. Let v =vec$(\boldsymbol{v})$. We wan to show that for any given $\epsilon^* > 0$, there exists a large constant $L = L(\epsilon^*)$ such that for sufficiently large N and T we have

$$P\left\{\inf_{\frac{1}{N}\sum_{i=1}^N \|v_i\|^2 = L} Q_{NT}^{\lambda,K}(\boldsymbol{\beta} + d_T^{1/2}T^{-1}v, \hat{\boldsymbol{\alpha}}, \hat{f}_1) > Q_{NT}^{\lambda,K}(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0, \hat{f}_1)\right\} \geqslant 1 - \epsilon^*.$$

Regardless the property of $\hat{f}_1$ and $\hat{\alpha}$, this implies that w.p.a.1 there is a local minimum $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, ..., \hat{\beta}_N)$ such that $\frac{1}{N}\sum_{i=1}^N \|\hat{b}_i\|^2 = O_P(d_T T^{-2})$. Note that

$$\begin{aligned}
&T^2\left[Q_{NT}^{\lambda,K}(\boldsymbol{\beta} + d_T^{1/2}T^{-1}\boldsymbol{v}, \hat{\boldsymbol{\alpha}}, \hat{f}_1) - Q_{NT}^{\lambda,K}(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0, \hat{f}_1)\right] \\
&\geqslant \frac{d_T^{1/2}}{N}\sum_{i=1}^N \left(\frac{d_T^{1/2}}{T^2}v_i'x_i'M_{\hat{f}_1}x_i v_i - \frac{2}{T}v_i'x_i'M_{\hat{f}_1}(f_1^0 - \hat{f}_1 H_1)\lambda_{1i}^0 - \frac{2}{T}v_i'x_i'M_{\hat{f}_1}u_i^*\right) \\
&= \frac{d_T}{N}\sum_{i=1}^N \frac{1}{T^2}v_i'x_i'M_{\hat{f}_1}x_i v_i \\
&\quad - \frac{2d_T^{1/2}}{N}\sum_{i=1}^N v_i'\left\{TR_{2i} + \frac{1}{T}x_i'M_{\hat{f}_1}u_i^* + \frac{1}{NT}\sum_{j=1}^N a_{ij}x_i'M_{\hat{f}_1}x_j\hat{b}_j - \frac{1}{NT}\sum_{j=1}^N a_{ij}x_i'M_{\hat{f}_1}u_j\right\} \\
&\equiv D_{1NT} - 2D_{2NT}, \text{ say.}
\end{aligned}$$

where $R_{2i} = \frac{1}{T^2}x_i'M_{\hat{f}_1}f_1^0\lambda_{1i}^0 - \frac{1}{NT^2}\sum_{j=1}^N x_i'M_{\hat{f}_1}x_j a_{ij}\hat{b}_j + \frac{1}{NT^2}\sum_{j=1}^N a_{ij}x_i'M_{\hat{f}_1}u_j$ as defined in Lemma A.6. By Assumption 3.2(v) and Lemma A.5(iii), $D_{1NT} = \frac{d_T}{N}v'\hat{Q}_1 v \geqslant$

$d_T\mu_{\min}\left(\hat{Q}_1\right)N^{-1}\left\|\boldsymbol{v}\right\|^2 \geqslant d_T\rho_{\min}N^{-1}\left\|\boldsymbol{v}\right\|^2/2$ w.p.a.1. By Lemmas A.6(i)-(ii) and A.5(iii),

$$\begin{aligned}
\frac{T^2}{d_T N}\sum_{i=1}^N \|R_{2i}\|^2 &= \frac{T^2}{d_T}O_P(T^{-2}d_T\eta_{NT}^2 + d_T^2\eta_{NT}^4 + T^{-2}\delta_{NT}^{-2}) = o_P(1),\\
\frac{1}{d_T NT^2}\sum_{i=1}^N \|x_i' M_{\hat{f}_1}u_i^*\|^2 &\leqslant \frac{2T^2}{d_T N}\sum_{i=1}^N \left\|\frac{1}{T^2}x_i'(M_{\hat{f}_1} - M_{f_1^0})u_i^*\right\|^2 + \frac{2}{d_T N}\sum_{i=1}^N \left\|\frac{1}{T}x_i' M_{f_1^0}u_i^*\right\|^2\\
&= \frac{T^2}{d_T}O_P(T^{-1}d_T\eta_{NT}^2 + d_T^2\eta_{NT}^4 + T^{-2}C_{NT}^{-2}) + \frac{1}{d_T}O_P(1) = o_P(1).
\end{aligned}$$

Next, we have

$$\begin{aligned}
\frac{1}{d_T N^3 T^2}\sum_{i=1}^N\sum_{j=1}^N \|a_{ij}x_i' M_{\hat{f}_1}x_j\hat{b}_j\|^2 &\leqslant \frac{1}{d_T}\frac{1}{N^3T^2}\sum_{i=1}^N\sum_{j=1}^N \|a_{ij}\|^2\left\|x_i' M_{\hat{f}_1}x_j\hat{b}_j\right\|^2\\
&\leqslant \frac{T^2}{N}\left[\mu_{\min}\left(\frac{\Lambda_1^{0\prime}\Lambda_1^0}{N}\right)\right]^{-2}\left\{\max_{1\leqslant j\leqslant N}\frac{1}{d_T T^2}\|x_j\|^2\right\}\max_{1\leqslant j\leqslant N}\left\|\lambda_{1j}^0\right\|^2\\
&\times\left\{\frac{1}{NT^2}\sum_{i=1}^N\left\|\lambda_{1i}^0\right\|^2\|x_i\|^2\right\}\left\{\frac{1}{N}\sum_{j=1}^N\left\|\hat{b}_j\right\|^2\right\}\\
&= \frac{T^2}{N}O_P(1)\,O_P(1)\,o_P\left(N^{1/q}\right)O_P(1)\,O_P\left(d_T T^{-2} + \lambda^2\right) = o_P(1).
\end{aligned}$$

where we use the fact that $\max_{1\leqslant j\leqslant N}\frac{1}{d_T T^2}\|x_j\|^2 = O_P(1)$ by Lemma A.2(i), $\max_{1\leqslant j\leqslant N}\left\|\lambda_{1j}^0\right\|^2 = o_P\left(N^{1/q}\right)$ by Assumption 3.2(i) and Markov inequality, and $\frac{1}{NT^2}\sum_{i=1}^N\left\|\lambda_{1i}^0\right\|^2\|x_i\|^2 = O_P(1)$ by Markov inequality and $\frac{1}{N}\sum_{j=1}^N\left\|\hat{b}_j\right\|^2 = O_P\left(d_T T^{-2} + \lambda^2\right)$. Similarly, we have by Lemma A.5(iii),

$$\begin{aligned}
&\frac{1}{d_T N^3 T^2}\sum_{i=1}^N\sum_{j=1}^N \|a_{ij}x_i' M_{\hat{f}_1}u_j\|^2\\
&\leqslant \frac{1}{d_T}\frac{1}{N^3T^2}\sum_{i=1}^N\sum_{j=1}^N \|a_{ij}\|^2\left\|x_i' M_{\hat{f}_1}u_j\right\|^2\\
&\leqslant \frac{1}{d_T}\left[\mu_{\min}\left(\frac{\Lambda_1^{0\prime}\Lambda_1^0}{N}\right)\right]^{-2}\frac{2}{N^3T^2}\sum_{i=1}^N\sum_{j=1}^N\left\|\lambda_{1i}^0\right\|^2\left\|\lambda_{1j}^0\right\|^2\left\{\left\|x_i'(M_{\hat{f}_1} - M_{f_1^0})u_j\right\|^2 + \left\|x_i' M_{f_1^0}u_j\right\|^2\right\}\\
&= \frac{1}{d_T}O_P\left(N^{-1}Td_T(\sqrt{d_T}\eta_{NT} + \delta_{NT}^{-1}) + 1\right) = o_P(1).
\end{aligned}$$

It follows that

$$\begin{aligned}|D_{2NT}| &\leqslant d_T \left\{ \frac{1}{N} \sum_{i=1}^{N} \|v_i\|^2 \right\}^{1/2} \left\{ \left(\frac{T^2}{d_T N} \sum_{i=1}^{N} \|\bar{R}_{2i}\|^2 \right)^{1/2} + \left(\frac{1}{d_T N T^2} \sum_{i=1}^{N} \|x_i' M_{\hat{f}_1} u_i^*\|^2 \right)^{1/2} \right. \\ &\quad \left. + \left(\frac{1}{d_T N^3 T^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \|a_{ij} x_i' M_{\hat{f}_1} \hat{b}_j\|^2 \right)^{1/2} + \left(\frac{1}{d_T N^3 T^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \|a_{ij} x_i' M_{\hat{f}_1} u_j\|^2 \right)^{1/2} \right\} \\ &= d_T N^{-1/2} \|\boldsymbol{v}\| \, o_P(1).\end{aligned}$$

Then D_{1NT} dominates D_{2NT} for sufficiently large L. That is $T^2[Q_{NT}^{\lambda,K}(\boldsymbol{\beta}+ d_T^{1/2}T^{-1}\boldsymbol{v}, \hat{\boldsymbol{\alpha}}, \hat{f}_1) - Q_{NT}^{\lambda,K}(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0, \hat{f}_1)] > 0$ for sufficiently large L. Consequently, the result in (i) follows.

(ii) We study the probability bound for each term on the right hand side of (**??**). For the first term, we have by Lemma A.6(i)

$$\begin{aligned}\left\| \frac{1}{T^2} x_i' M_{\hat{f}_1} u_i^* \right\| &\leqslant \left\| \frac{1}{T^2} x_i' M_{f_1^0} u_i^* \right\| + \left\| \frac{1}{T^2} x_i' (M_{\hat{f}_1} - M_{f_1^0}) u_i^* \right\| \\ &= O_P(T^{-1}) + O_P(T^{-1/2}\sqrt{d_T}\eta_{NT} + d_T\eta_{NT}^2 + T^{-1}C_{NT}^{-1}) = O_P(d_T T^{-1}).\end{aligned} \tag{95}$$

For the second term, we can readily apply Lemmas A.6(ii), A.5(iii) and A.3(iii), and Theorem 8(i) to obtain

$$\begin{aligned}\left\| \frac{1}{T^2} x_i' M_{\hat{f}_1} f_1^0 \lambda_{1i}^0 \right\| &\leqslant \|R_{2i}\| + \left\| \frac{1}{NT^2} \sum_{j=1}^{N} x_i' M_{\hat{f}_1} x_j \hat{b}_j a_{ij} \right\| + \left\| \frac{1}{NT^2} \sum_{j=1}^{N} x_i' M_{\hat{f}_1} u_j a_{ij} \right\| \\ &= O_P(T^{-1}\sqrt{d_T}\eta_{NT} + d_T\eta_{NT}^2 + T^{-1}\delta_{NT}^{-1}) + O_P(\eta_{NT}) + O_P(d_T T^{-1}) = O_P(d_T T^{-1}).\end{aligned} \tag{96}$$

The third term is $O_P(\lambda)$. By Lemma A.5(iii), $\mu_{\min}\left(\frac{1}{T^2} x_i' M_{\hat{f}_1} x_i\right) = \mu_{\min}\left(\frac{1}{T^2} x_i' M_{f_1^0} x_i\right) + o_P(1)$. Noting that $\left(\frac{1}{T^2} x_i' M_{f_1^0} x_i\right)^{-1}$ is the principal $p \times p$ submatrix of $\left(\frac{1}{T^2} W_i' W_i\right)^{-1}$, $\mu_{\min}\left(\frac{1}{T^2} x_i' M_{f_1^0} x_i\right) \geqslant \mu_{\min}\left(\frac{1}{T^2} W_i' W_i\right)$, and the last object is bounded away from zero w.p.a.1. It follows that $\hat{b}_i = O_P(d_T T^{-1} + \lambda)$ for $i = 1, 2, ..., N$.

(iii) Let $P_{NT}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \frac{1}{N}\sum_{i=1}^{N} \prod_{k=1}^{K} \|\beta_i - \alpha_k\|$ and $\hat{c}_{iNT}(\alpha) = \prod_{k=1}^{K-1} \|\hat{\beta}_i - \alpha_k\| + \prod_{k=1}^{K-2} \|\hat{\beta}_i - \alpha_k\| \times \|\beta_i^0 - \alpha_K\| + ... + \prod_{k=2}^{K} \|\beta_i^0 - \alpha_k\|$. By SSP, we have that as $(N, T) \to \infty$, $\left| \prod_{k=1}^{K} \left\| \hat{\beta}_i - \alpha_k \right\| - \prod_{k=1}^{K} \|\beta_i^0 - \alpha_k\| \right| \leqslant \hat{c}_{iNT}(\alpha) \|\hat{\beta}_i - \beta_i^0\|$, where $\hat{c}_{iNT}(\alpha) \leqslant C_{KNT}(\alpha)(1 + 2\|\hat{\beta}_i - \beta_i^0\|)$ and $C_{KNT}(\alpha) = \max_{1 \leqslant i \leqslant N} \max_{1 \leqslant s \leqslant k \leqslant K-1} \prod_{k=1}^{s} c_{ks} \|\beta_i^0 - \alpha_k\|^{K-1-s} =$

$\max_{1\leqslant l\leqslant K}\max_{1\leqslant s\leqslant k\leqslant K_0-1}\prod_{k=1}^{s}c_{ks}\|\alpha_l^0-\alpha_k\|^{K-1-s}=O(1)$ with c_{ks} being finite integers. It follows that as $(N,T)\to\infty$

$$\begin{aligned}|P_{NT}(\hat{\boldsymbol{\beta}},\boldsymbol{\alpha})-P_{NT}(\boldsymbol{\beta}^0,\boldsymbol{\alpha})| &\leqslant C_{KNT}(\alpha)\frac{1}{N}\sum_{i=1}^{N}\|\hat{b}_i\|+2C_{KNT}(\alpha)\frac{1}{N}\sum_{i=1}^{N}\|\hat{b}_i\|^2\\ &\leqslant C_{KNT}(\alpha)\left\{\frac{1}{N}\sum_{i=1}^{N}\|\hat{b}_i\|^2\right\}^{1/2}+O_P(d_TT^{-2})=O_P(d_T^{1/2}T^{-1}).\end{aligned} \tag{97}$$

By (97) and the fact that $P_{NT}(\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0)=0$ and that $P_{NT}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\alpha}})-P_{NT}(\hat{\boldsymbol{\beta}},\boldsymbol{\alpha}^0)\leqslant 0.$ we have

$$\begin{aligned}0 &\geqslant P_{NT}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\alpha}})-P_{NT}(\hat{\boldsymbol{\beta}},\boldsymbol{\alpha}^0)=P_{NT}(\boldsymbol{\beta}^0,\hat{\boldsymbol{\alpha}})-P_{NT}(\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0)+O_P(d_T^{1/2}T^{-1})\\ &=\frac{1}{N}\sum_{i=1}^{N}\prod_{k=1}^{K}\|\beta_i^0-\hat{\alpha}_k\|+O_P(d_T^{1/2}T^{-1})\\ &=\frac{N_1}{N}\prod_{k=1}^{K}\|\hat{\alpha}_k-\alpha_1^0\|+\frac{N_2}{N}\prod_{k=1}^{K}\|\hat{\alpha}_k-\alpha_2^0\|+...+\frac{N_K}{N}\prod_{k=1}^{K}\|\hat{\alpha}_k-\alpha_K^0\|+O_P(d_T^{1/2}T^{-1})\end{aligned} \tag{98}$$

By Assumption 3.3(i), $N_k/N\to\tau_k\in(0,1)$ for each $k=1,...K$. So (98) implies that $\prod_{k=1}^{K}\|\hat{\alpha}_k-\alpha_l^0\|=O_P(d_T^{1/2}T^{-1})$ for $l=1,...K$. It follows that $(\hat{\alpha}_{(1)},...,\hat{\alpha}_{(K)})-(\alpha_1^0,...,\alpha_K^0)=O_P(d_T^{1/2}T^{-1})$.

(iv) By Theorem 7(iv) and Theorem 8(i), we have $\frac{1}{T}\|\hat{f}_1-f_1^0H_1\|^2=O_P(Td_T\eta_{NT}^2+C_{NT}^{-2})=O_P(d_T^2T^{-1}+N^{-1})$.■

To prove Theorem 3.3, we need the following two lemmas.

Lemma A.7 *Suppose that Assumptions 3.1-3.3 hold. Then for any $c>0$,*

(i) $P\left(\max_{1\leqslant i\leqslant N}\left\|\frac{1}{T^2}x_i'u_i^*\right\|>c\psi_{NT}\right)=o(N^{-1})$,

(ii) $P\left(\max_{1\leqslant i\leqslant N}\left\|\frac{1}{T^2}x_i'M_{f_1^0}u_i^*\right\|>cd_T\psi_{NT}\right)=o(N^{-1})$.

Lemma A.8 *Suppose that Assumptions 3.1-3.3 hold. Then for any $c>0$,*

(i) $P\left(\max_{1\leqslant i\leqslant N}\|R_{1i}\|>c\left(d_T\eta_{NT}+T^{-1/2}d_T^{1/2}C_{NT}^{-1}\right)\left(\psi_{NT}+T^{-1/2}(\log T)^3\right)\right)=o(N^{-1})$,

(ii) $P\left(\max_{1\leqslant i\leqslant N}\|R_{2i}\|>cd_T^{1/2}N^{(1/2q)}\varsigma_{2NT}\right)=o(N^{-1})$,

(iii) $P\left(\max_{1\leqslant i\leqslant N}\|R_{3i}\| > cd_T^{1/2}N^{(1/2q)}\varsigma_{2NT}\right) = o(N^{-1})$,

(iv) $P\left(\max_{1\leqslant i\leqslant N}\|R_{4i}\| > c(d_T + N^{(1/2q)})\psi_{NT}\right) = o(N^{-1})$,

(v) $P\left(\max_{1\leqslant i\leqslant N}\left\|\hat{\beta}_i - \beta_i^0\right\| > c\left(N^{(1/2q)}\psi_{NT} + \lambda(\log T)^{\epsilon/2}\right)\right) = o(N^{-1})$ *for any* $\epsilon > 0$,

(vi) $P\left(\frac{1}{N}\sum_{i=1}^N\left\|\hat{\beta}_i - \beta_i^0\right\|^2 > cd_T^2\psi_{NT}^2\right) = o(N^{-1})$ *for any* $\epsilon > 0$,

(vii) $P\left(\max_{1\leqslant i\leqslant N}\left\|\frac{1}{T^2}x_i'M_{\hat{f}_1}f_1^0\lambda_{1i}^0\right\| > cN^{1/2q}(d_T\eta_{NT} + T^{-1/2}d_T^{1/2}C_{NT}^{-1})\right) = o(N^{-1})$.

Proof of Theorem 3.3. (i) Fix $k \in \{1, ..., K\}$. By the consistency of $\hat{\alpha}_k$ and $\hat{\beta}_i$, we have $\hat{\beta}_i - \hat{\alpha}_l \xrightarrow{p} \alpha_k^0 - \alpha_l^0 \neq 0$ for all $i \in G_k^0$ and $l \neq k$. Now, suppose that $\|\hat{\beta}_i - \hat{\alpha}_k\| \neq 0$ for some $i \in G_k^0$. Then the first order condition (with respect to β_i) for the minimization of the objective function (2.8) implies that

$$\begin{aligned}0_{p\times 1} = &-\frac{2}{T}x_i'M_{f_1^0}u_i^* + \frac{2}{T}x_i'(M_{f_1^0} - M_{\hat{f}_1})u_i^* - \frac{2}{T}x_i'M_{\hat{f}_1}f_1^0\lambda_{1i}^0 + \frac{2}{T^2}x_i'M_{\hat{f}_1}x_iT(\hat{\alpha}_k - \alpha_k^0)\\ &+\left(\frac{2}{T^2}x_i'M_{\hat{f}_1}x_i + \frac{\lambda\hat{c}_{ki}}{\|\hat{\beta}_i - \hat{\alpha}_k\|}I_p\right)T(\hat{\beta}_i - \hat{\alpha}_k) + T\lambda\sum_{j=1,j\neq k}^K\hat{e}_{ij}\prod_{l=1,l\neq j}^K\|\hat{\beta}_i - \hat{\alpha}_l\|\\ &\equiv -\hat{A}_{1i} + \hat{A}_{2i} - \hat{A}_{3i} + \hat{A}_{4i} + \hat{A}_{5i} + \hat{A}_{6i},\ \text{say,}\end{aligned}$$

where $\hat{e}_{ij}$ are defined in the proof of Theorem 3.2(i), $\hat{c}_{ki} = \prod_{l=1,l\neq k}^K\|\hat{\beta}_i - \hat{\alpha}_l\| \xrightarrow{p} c_k^0 \equiv \prod_{l=1,l\neq k}^K\|\alpha_k^0 - \alpha_l^0\| > 0$ for $i \in G_k^0$ by Assumption 3.3(ii). Let $\Psi_{NT} = N^{1/(2q)}\psi_{NT} + \lambda(\log T)^{\epsilon/2}$. Let c denote a generic constant that may vary across lines. By Lemma A.8(v)-(vi), we have

$$P\left(\max_{i\in G_k^0}\left\|\hat{\beta}_i - \beta_i^0\right\| > c\Psi_{NT}\right) = o(N^{-1}) \text{ and } P\left(\frac{1}{N}\sum_{i=1}^N\left\|\hat{\beta}_i - \beta_i^0\right\|^2 > cd_T^2\psi_{NT}^2\right) = o(N^{-1}). \tag{99}$$

This, in conjunction with the proof of Theorem 3.2(i)-(iii), implies that

$$P(\|\hat{\alpha}_k - \alpha_k^0\| > cd_T\psi_{NT}) = o(N^{-1}),\ \text{and } P(\max_{i\in G_k^0}\left|\hat{c}_{ki} - c_k^0\right| \geqslant c_k^0/2) = o(N^{-1}). \tag{100}$$

By (99)-(100) and the fact that $\max_{i\in G_k^0}\frac{1}{T^2}x_i'M_{\hat{f}_1}x_i \leqslant cd_T\rho_{\max}$ a.s.,

we have $P\left(\max_{i\in G_k^0}\left\|\hat{A}_{4i}\right\| > cd_T^2T\psi_{NT}\right) = o(N^{-1})$ and $P\left(\max_{i\in G_k^0}\left\|\hat{A}_{6i}\right\| > c\lambda T\Psi_{NT}\right) =$

$o(N^{-1})$. By Lemma A.7(ii) and Lemma A.8(i),(iii), we can directly claim that

$$P\left(\max_{i\in G_k^0}\|\hat{A}_{1i}\| > cTb_T\psi_{NT}\right) = o(N^{-1}),$$
$$P\left(\max_{i\in G_k^0}\|\hat{A}_{2i}\| > c\left(Td_T\eta_{NT} + T^{1/2}d_T^{1/2}C_{NT}^{-1}\right)\left(\psi_{NT} + T^{-1/2}(\log T)^3\right)\right) = o(N^{-1}), \text{ and}$$
$$P\left(\max_{i\in G_k^0}\|\hat{A}_{3i}\| > cN^{1/2q}(Td_T\eta_{NT} + T^{1/2}d_T^{1/2}C_{NT}^{-1})\right) = o(N^{-1}).$$

For $\hat{A}_{5i}$, we have

$$\begin{aligned}(\hat{\beta}_i - \hat{\alpha}_k)'\hat{A}_{5i} &= (\hat{\beta}_i - \hat{\alpha}_k)'\left(\frac{2}{T^2}x_i'M_{\hat{f}_1}x_i + \frac{\lambda\hat{c}_{ki}}{\|\hat{\beta}_i - \hat{\alpha}_k\|}I_p\right)T(\hat{\beta}_i - \hat{\alpha}_k)\\ &\geqslant 2\hat{Q}_{i,xx}T\|\hat{\beta}_i - \hat{\alpha}_k\|^2 + T\lambda\hat{c}_{ki}\|\hat{\beta}_i - \hat{\alpha}_k\| \geqslant cT\lambda c_k^0\|\hat{\beta}_i - \hat{\alpha}_k\|.\end{aligned}$$

Combing above results together, it follows that $P(\Xi_{k,NT}) = 1 - o(N^{-1})$, where

$$\begin{aligned}\Xi_{k,NT} = &\left\{\max_{i\in G_k^0}\|\hat{A}_{2i}\| < c\left(Td_T\eta_{NT} + T^{1/2}d_T^{1/2}C_{NT}^{-1}\right)\left(\psi_{NT} + T^{-1/2}(\log T)^3\right)\right\}\\ &\cap\left\{\max_{i\in G_k^0}\|\hat{A}_{3i}\| < cN^{1/2q}(Td_T\eta_{NT} + T^{1/2}d_T^{1/2}C_{NT}^{-1})\right\}\cap\left\{\max_{i\in G_k^0}\left|\hat{c}_{ki} - c_k^0\right| < c_k^0/2\right\}\\ &\cap\left\{\max_{i\in G_k^0}\left\|\hat{A}_{4i}\right\| < cd_T^2T\psi_{NT}\right\}\cap\left\{\max_{i\in G_k^0}\left\|\hat{A}_{6i}\right\| < c\lambda T\Psi_{NT}\right\}.\end{aligned}$$

Then conditional on Ξ_{kNT}, we have that uniformly in $i \in G_k^0$,

$$\begin{aligned}&\left|(\hat{\beta}_i - \hat{\alpha}_k)'(\hat{A}_{2i} + \hat{A}_{3i} + \hat{A}_{4i} + \hat{A}_{5i} + \hat{A}_{6i})\right|\\ \geqslant&\left|(\hat{\beta}_i - \hat{\alpha}_k)'\hat{A}_{5i}\right| - \left|(\hat{\beta}_i - \hat{\alpha}_k)'(\hat{A}_{2i} + \hat{A}_{3i} + \hat{A}_{4i} + \hat{A}_{6i})\right|\\ \geqslant&\left\{cT\lambda c_k^0 - c\left(N^{1/2q}\left(Td_T^{1/2}\eta_{NT} + T^{1/2}d_T^{1/2}C_{NT}^{-1}\right) + Td_T^2\psi_{NT} + \lambda T\Psi_{NT}\right)\right\}\|\hat{\beta}_i - \hat{\alpha}_k\|\\ \geqslant& cT\lambda c_k^0\|\hat{\beta}_i - \hat{\alpha}_k\|/2,\end{aligned}$$

where the last inequality follows by the fact that $N^{1/2q}\left(Td_T^{1/2}\eta_{NT} + T^{1/2}d_T^{1/2}C_{NT}^{-1}\right) + Td_T^2\psi_{NT} + \lambda T\Psi_{NT} = o(T\lambda)$ for sufficiently large (N, T) by Assumption 3.3(iv). It follows that

$$\begin{aligned}P(\hat{E}_{kNT,i}) = P(i \notin \hat{G}_k|i \in G_k^0) &= P(\hat{A}_{1i} = \hat{A}_{2i} + \hat{A}_{3i} + \hat{A}_{4i} + \hat{A}_{5i} + \hat{A}_{6i})\\ &\leqslant P\left(|(\hat{\beta}_i - \hat{\alpha}_k)'\hat{A}_{1i}| \geqslant |(\hat{\beta}_i - \hat{\alpha}_k)'\hat{A}_{5i} - (\hat{\beta}_i - \hat{\alpha}_k)'(\hat{A}_{2i} + \hat{A}_{3i} + \hat{A}_{4i} + \hat{A}_{6i})\right)\\ &\leqslant P(\|\hat{A}_{i1}\| \geqslant cT\lambda c_k^0/4, \Xi_{kNT}) + o(N^{-1}) \to 0 \quad\text{as}\quad (N, T) \to \infty,\end{aligned}$$

where the last inequality follows because that $T\lambda \gg Tb_T\psi_{NT}$ by Assumption 3.3(iv). Consequently, we can conclude that w.p.a.1 $\hat{\beta}_i - \hat{\alpha}_k$ must be in position where $\|\beta_i - \alpha_k\|$ is not differentiable with respect to β_i for any $i \in G_k^0$. That is $P(\|\hat{\beta}_i - \hat{\alpha}_k\| = 0 | i \in G_k^0) = 1 - o(N^{-1})$ as $(N, T) \to \infty$.

For uniform consistency, we have that

$$P(\cup_{k=1}^K \hat{E}_{kNT}) \leqslant \sum_{k=1}^K P(\hat{E}_{kNT}) \leqslant \sum_{k=1}^K \sum_{i \in G_k^0} P(\hat{E}_{kNT,i})$$
$$\leqslant N \max_{1 \leqslant i \leqslant N} P(\|\hat{A}_{i1}\| \geqslant cT\lambda c_k^0/4) + o(1) \to 0 \quad \text{as} \quad (N, T) \to \infty.$$

This completes the proof of (i). Then the proof of (ii) directly follows SSP and thus omitted. ■

To prove Theorem 3.4, we need the following two lemmas.

Lemma A.9 *Suppose that Assumptions 3.1-3.3 hold and* $\sqrt{N} = o\left(T\right)$. *Then for any* $k = 1, ..., K$,

(i) $\frac{1}{N_k T^2} \sum_{i \in \hat{G}_k} x_i' M_{\hat{f}_1} f_1^0 \lambda_{1i}^0 = \frac{1}{N_k T^2} \sum_{i \in \hat{G}_k} \frac{1}{N} \sum_{j=1}^N x_i' M_{\hat{f}_1} x_j a_{ij} \hat{b}_j - \frac{1}{N_k T^2} \sum_{i \in \hat{G}_k} \frac{1}{N} \sum_{j=1}^N a_{ij} x_i' M_{\hat{f}_1} u_j - \frac{1}{N_k T^2} \sum_{i \in \hat{G}_k} \frac{1}{N} \sum_{j=1}^N a_{ij} x_i' M_{\hat{f}_1} f_2^0 \lambda_{2j}^0 + o_P(N^{-1/2} T^{-1})$,

(ii) $\frac{1}{N_k T^2} \sum_{i \in \hat{G}_k} x_i' M_{\hat{f}_1} x_i = \frac{1}{N_k T^2} \sum_{i \in G_k^0} x_i' M_{f_1^0} x_i + o_P(1)$,

(iii) $\frac{1}{\sqrt{N_k} T} \sum_{i \in \hat{G}_k} x_i' M_{\hat{f}_1} \left(u_i^* - \frac{1}{N} \sum_{j=1}^N u_j^* a_{ij} \right) = U_{kNT} + o_P(1)$,

(iv) $\frac{1}{N_k T^2} \sum_{i \in \hat{G}_k} \frac{1}{N} \sum_{j \in \hat{G}_l} x_i' M_{\hat{f}_1} x_j a_{ij} = \frac{1}{N_k T^2} \sum_{i \in G_k^0} \frac{1}{N} \sum_{j \in G_l^0} x_i' M_{f_1^0} x_j a_{ij} + o_P(1)$.

Lemma A.10 *Suppose that Assumptions 3.1-3.3 hold and* $\sqrt{N} = o\left(T\right)$. *Then*

(i) $Q_{NT} \xrightarrow{d} Q_0$,

(ii) $U_{kNT} = V_{kNT} + B_{kNT} + o_P(1)$,

(iii) $V_{NT} \xrightarrow{d} N\left(0, \Omega_0\right)$ conditional on $\mathcal{C}$,

where $\Omega_0 = \lim_{N,T \to \infty} \Omega_{NT}$.

Proof of Theorem 3.4. (i) To study of the oracle property of the C-Lasso estimator, we invoke the sub-differential calculus. A necessary and sufficient condition for $\{\hat{\beta}_i\}$ and $\{\hat{\alpha}_k\}$ to minimize the objective function in (23) is that for each $i = 1, ..., N$ (resp. $k = 1, ..., K$), $0_{p \times 1}$ belongs to the sub-differential of $Q_{NT,\lambda}^K(\boldsymbol{\beta}, \boldsymbol{\alpha}, \hat{f}_1)$ with respect to β_i (resp. α_k) evaluated at $\{\hat{\beta}_i\}$ and $\{\hat{\alpha}_k\}$. That is, for each $i = 1, ..., N$ and $k = 1, ..., K$, we have

$$0_{p\times 1} = -\frac{2}{NT^2} x_i' M_{\hat{f}_1}(y_i - x_i\hat{\beta}_i) + \frac{\lambda}{N}\sum_{j=1}^{K} \hat{e}_{ij} \prod_{l=1,l\neq j}^{K} \|\hat{\beta}_i - \hat{\alpha}_l\|, \tag{101}$$

$$0_{p\times 1} = \frac{\lambda}{N}\sum_{i=1}^{N} \hat{e}_{ik} \prod_{l=1,l\neq k}^{K} \|\hat{\beta}_i - \hat{\alpha}_l\|, \tag{102}$$

where $\hat{e}_{ij} = \frac{\hat{\beta}_i - \hat{\alpha}_j}{\|\hat{\beta}_i - \hat{\alpha}_j\|}$ if $\|\hat{\beta}_i - \hat{\alpha}_j\| \neq 0$ and $\|\hat{e}_{ij}\| \leqslant 1$ if $\|\hat{\beta}_i - \hat{\alpha}_j\| = 0$. First we observe that $\|\hat{\beta}_i - \hat{\alpha}_k\| = 0$ for any $i \in \hat{G}_k$ by the definition of $\hat{G}_k$, then $\hat{\beta}_i - \hat{\alpha}_l \to \alpha_k^0 - \alpha_l^0 \neq 0$ for any $i \in \hat{G}_k$ and $l \neq k$ by Assumption 3.3(ii). It follows that $\|\hat{e}_{ik}\| \leqslant 1$ for any $i \in \hat{G}_k$ and $\hat{e}_{ij} = \frac{\hat{\beta}_i - \hat{\alpha}_j}{\|\hat{\beta}_i - \hat{\alpha}_j\|} = \frac{\hat{\alpha}_k - \hat{\alpha}_j}{\|\hat{\alpha}_k - \hat{\alpha}_j\|}$ w.p.a.1 for any $i \in \hat{G}_k$ and $j \neq k$. This further implies that w.p.a.1

$$\sum_{i\in\hat{G}_k}\sum_{j=1,j\neq k}^{K} \hat{e}_{ij} \prod_{l=1,l\neq j}^{K} \|\hat{\beta}_i - \hat{\alpha}_l\| = \sum_{i\in\hat{G}_k}\sum_{j=1,j\neq k}^{K} \frac{\hat{\alpha}_k - \hat{\alpha}_j}{\|\hat{\alpha}_k - \hat{\alpha}_j\|} \prod_{l=1,l\neq j}^{K} \|\hat{\alpha}_k - \hat{\alpha}_l\| = 0_{p\times 1},$$

and

$$\begin{aligned}
0_{p\times 1} &= \sum_{i=1}^{N} \hat{e}_{ik} \prod_{l=1,l\neq k}^{K} \|\hat{\beta}_i - \hat{\alpha}_l\| \\
&= \sum_{i\in\hat{G}_k} \hat{e}_{ik} \prod_{l=1,l\neq k}^{K} \|\hat{\alpha}_k - \hat{\alpha}_l\| + \sum_{i\in\hat{G}_0} \hat{e}_{ik} \prod_{l=1,l\neq k}^{K} \|\hat{\beta}_i - \hat{\alpha}_l\| + \sum_{j=1,j\neq k}^{K}\sum_{i\in\hat{G}_j} \hat{e}_{ik} \prod_{l=1,l\neq k}^{K} \|\hat{\alpha}_j - \hat{\alpha}_l\| \\
&= \sum_{i\in\hat{G}_k} \hat{e}_{ik} \prod_{l=1,l\neq k}^{K} \|\hat{\alpha}_k - \hat{\alpha}_l\| + \sum_{i\in\hat{G}_0} \hat{e}_{ik} \prod_{l=1,l\neq k}^{K} \|\hat{\beta}_i - \hat{\alpha}_l\|.
\end{aligned} \tag{103}$$

Then by (101), (102) and (103) we have

$$\frac{2}{N_k T^2}\sum_{i\in\hat{G}_k} x_i' M_{\hat{f}_1}(y_i - x_i\hat{\alpha}_k) + \frac{\lambda}{N}\sum_{i\in\hat{G}_0} \hat{e}_{ik} \prod_{l=1,l\neq k}^{K} \|\hat{\beta}_i - \hat{\alpha}_l\| = 0_{p\times 1}. \tag{104}$$

Noting that $1\{i \in \hat{G}_k\} = 1\{i \in G_k^0\} + 1\{i \in \hat{G}_k \setminus G_k^0\} - 1\{i \in G_k^0 \setminus \hat{G}_k\}$ and $y_i = x_i\alpha_k^0 + f_1^0\lambda_{1i}^0 + u_i^*$ when $i \in G_k^0$, we have

$$\begin{aligned}
\frac{1}{N_k T^2}\sum_{i\in\hat{G}_k} x_i M_{\hat{f}_1} y_i =& \frac{1}{N_k T^2}\sum_{i\in\hat{G}_k} x_i' M_{\hat{f}_1} x_i \beta_i^0 + \frac{1}{N_k T^2}\sum_{i\in\hat{G}_k} x_i' M_{\hat{f}_1} f_1^0 \lambda_{1i}^0 + \frac{1}{N_k T^2}\sum_{i\in\hat{G}_k} x_i' M_{\hat{f}_1} u_i^* \\
=& \frac{1}{N_k T^2}\sum_{i\in G_k^0} x_i' M_{\hat{f}_1} x_i \alpha_k^0 + \frac{1}{N_k T^2}\sum_{i\in\hat{G}_k\setminus G_k^0} x_i' M_{\hat{f}_1} x_i \beta_i^0 - \frac{1}{N_k T^2}\sum_{i\in G_k^0\setminus\hat{G}_k} x_i' M_{\hat{f}_1} x_i \alpha_k^0 \\
&+ \frac{1}{N_k T^2}\sum_{i\in\hat{G}_k} x_i' M_{\hat{f}_1} f_1^0 \lambda_{1i}^0 + \frac{1}{N_k T^2}\sum_{i\in\hat{G}_k} x_i' M_{\hat{f}_1} (u_i + f_2^0 \lambda_{2i}^0).
\end{aligned} \tag{105}$$

Combining (104) and (105) yields

$$\frac{1}{N_kT^2}\sum_{i\in\hat{G}_k} x_i'M_{\hat{f}_1}x_i(\hat{\alpha}_k-\alpha_k^0)=\frac{1}{N_kT^2}\sum_{i\in\hat{G}_k} x_i'M_{\hat{f}_1}f_1^0\lambda_{1i}^0+\frac{1}{N_kT^2}\sum_{i\in\hat{G}_k} x_i'M_{\hat{f}_1}\left(u_i+f_2^0\lambda_{2i}^0\right)$$
$$+\hat{C}_{1k}-\hat{C}_{2k}+\hat{C}_{3k}, \tag{106}$$

where $\hat{C}_{1k}=\frac{1}{N_kT^2}\sum_{i\in\hat{G}_k\backslash G_k^0} x_i'M_{\hat{f}_1}x_i\beta_i^0$, $\hat{C}_{2k}=\frac{1}{N_kT^2}\sum_{i\in G_k^0\backslash\hat{G}_k} x_i'M_{\hat{f}_1}x_i\alpha_k^0$, and $\hat{C}_{3k}=\frac{\lambda}{2N_k}\sum_{i\in\hat{G}_0}\hat{e}_{ik}\times\prod_{l=1,l\neq k}^K\|\hat{\beta}_i-\hat{\alpha}_l\|$. By Theorem 3.3 and Lemmas S1.11-S1.12 in SSP, we have $P(N^{1/2}T\|\hat{C}_{1k}\|\geqslant\epsilon)\leqslant P(\hat{F}_{kNT})\to 0$, $P(N^{1/2}T\|\hat{C}_{2k}\|\geqslant\epsilon)\leqslant P(\hat{E}_{kNT})\to 0$, and $P(N^{1/2}T\|\hat{C}_{3k}\|\geqslant\epsilon)\leqslant\sum_{k=1}^K\sum_{i\in G_k^0}P(i\in\hat{G}_0|i\in G_k^0)\leqslant\sum_{k=1}^K\sum_{i\in G_k^0}P(\hat{E}_{kNT,i})=o(1)$. It follows that $\|\hat{C}_{1k}-\hat{C}_{2k}+\hat{C}_{3k}\|=o_P(N^{-1/2}T^{-1})$. By Lemma A.9 (i), we have as $\frac{\sqrt{N}}{T}\to 0$

$$\frac{1}{N_kT^2}\sum_{i\in\hat{G}_k} x_i'M_{\hat{f}_1}f_1^0\lambda_{1i}^0=\frac{1}{N_kT^2}\sum_{i\in\hat{G}_k}\frac{1}{N}\sum_{j=1}^N x_i'M_{\hat{f}_1}x_ja_{ij}\hat{b}_j-\frac{1}{N_kT^2}\sum_{i\in\hat{G}_k}\frac{1}{N}\sum_{j=1}^N a_{ij}x_i'M_{\hat{f}_1}u_j$$
$$-\frac{1}{N_kT^2}\sum_{i\in\hat{G}_k}\frac{1}{N}\sum_{j=1}^N a_{ij}x_i'M_{\hat{f}_1}f_2^0\lambda_{2j}^0+o_P(N^{-1/2}T^{-1}). \tag{107}$$

In addition,

$$\frac{1}{N_kT^2}\sum_{i\in\hat{G}_k}\frac{1}{N}\sum_{j=1}^N x_i'M_{\hat{f}_1}x_ja_{ij}\hat{b}_j=\frac{1}{N_kT^2}\sum_{i\in\hat{G}_k}\frac{1}{N}\sum_{l=1}^K\sum_{j\in\hat{G}_l} x_i'M_{\hat{f}_1}x_ja_{ij}\left(\hat{\alpha}_l-\alpha_l^0\right)+o_P(N^{-1/2}T^{-1}) \tag{108}$$

by Theorem 3.3. Let $\hat{Q}_{1NT}$ =diag$\left(\frac{1}{N_1T^2}\sum_{i\in\hat{G}_1} x_i'M_{\hat{f}_1}x_i,\ldots,\frac{1}{N_KT^2}\sum_{i\in\hat{G}_K} x_i'M_{\hat{f}_1}x_i\right)$ and $\hat{Q}_{2NT}$ is a $Kp\times Kp$ matrix with typical blocks $\frac{1}{NN_kT}\sum_{i\in\hat{G}_k}\sum_{j\in\hat{G}_l}a_{ij}x_i'M_{\hat{f}_1}x_j$ such that

$$\hat{Q}_{2NT}=\begin{pmatrix}\frac{1}{NN_1T^2}\sum_{i\in\hat{G}_1}\sum_{j\in\hat{G}_1}a_{ij}x_i'M_{\hat{f}_1}x_j, & \cdots & \frac{1}{NN_1T^2}\sum_{i\in\hat{G}_1}\sum_{j\in\hat{G}_K}a_{ij}x_i'M_{\hat{f}_1}x_j\\ \frac{1}{NN_2T^2}\sum_{i\in\hat{G}_2}\sum_{j\in\hat{G}_1}a_{ij}x_i'M_{\hat{f}_1}x_j, & \cdots & \frac{1}{NN_2T^2}\sum_{i\in\hat{G}_2}\sum_{j\in\hat{G}_K}a_{ij}x_i'M_{\hat{f}_1}x_j,\\ \vdots & \ddots & \vdots\\ \frac{1}{NN_KT^2}\sum_{i\in\hat{G}_K}\sum_{j\in\hat{G}_1}a_{ij}x_i'M_{\hat{f}_1}x_j, & \cdots & \frac{1}{NN_KT^2}\sum_{i\in\hat{G}_K}\sum_{j\in\hat{G}_K}a_{ij}x_i'M_{\hat{f}_1}x_j\end{pmatrix}.$$

Combining (106)-(108), we have

$$\sqrt{N}T\text{vec}(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^0)=(\hat{Q}_{1NT}-\hat{Q}_{2NT})^{-1}\sqrt{D_N}\hat{U}_{NT}+o_P(1),$$

where the kth element of $\hat{U}_{NT}$ is

$$\hat{U}_{kNT}=\frac{1}{\sqrt{N_k}T}\sum_{i\in\hat{G}_k}x_i'M_{\hat{f}_1}\left(\left(u_i+f_2^0\lambda_{2i}^0\right)-\frac{1}{N}\sum_{j=1}^N a_{ij}\left(u_j+f_2^0\lambda_{2j}^0\right)\right)$$

and D_N =diag$(\frac{N}{N_1},...,\frac{N}{N_K})$. By Lemma A.9(ii)-(iv), we have that $\hat{Q}_{1NT} - \hat{Q}_{2NT} = Q_{NT} + o_P(1)$, $\hat{U}_{NT} = U_{NT} + o_P(1)$, where U_{NT} and Q_{NT} are defined in Theorem 3.4. Then we have $\sqrt{N}T\text{vec}(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^0) = Q_{NT}^{-1}\sqrt{D_N}U_{NT}+o_P(1)$. By Lemma A.10(ii), we have $U_{kNT} - B_{kNT,1} - B_{kNT,2} = V_{kNT} + o_P(1)$, where V_{kNT} and $B_{kNT} = B_{kNT,1} + B_{kNT,2}$ are defined in Theorem 3.4. Thus,

$$\sqrt{N}T\text{vec}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^0) = Q_{NT}^{-1}\sqrt{D_N}\left(V_{NT} + B_{NT}\right) + o_P(1), \tag{109}$$

where $V_{NT} = (V_{1NT},...,V_{KNT})$ and $B_{NT} = (B_{1NT},...,B_{KNT})$. This completes the proof of Theorem 3.4.

(ii) By Lemma A.10 (i) and (iii), we have

$$Q_{NT} \xrightarrow{d} Q_0 \text{ and } V_{NT} \xrightarrow{d} N(0, \Omega_0) \text{ conditional } \mathcal{C}. \tag{110}$$

Combining (109) and (110), we have

$$\sqrt{N}Tvec(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^0) - \sqrt{D_N}Q_{NT}^{-1}B_{NT} \xrightarrow{d} MN(0, \lim_{N\to\infty} D_N Q_0^{-1}\Omega_0 Q_0^{-1}).$$

■

To prove Theorem 3.5, we need the following lemma.

Lemma A.11 *Suppose that Assumptions 3.1-3.3 hold and* $\sqrt{N} = o\left(T\right)$*. Then as* $(N,T) \to \infty$,

(i) $\frac{1}{\sqrt{T}}\|\hat{f}_1\hat{\lambda}_{1i} - f_1^0\lambda_{1i}^0\| = O_P(\sqrt{d_T T}\eta_{NT}) + O_P(C_{NT}^{-1})$,

(ii) $\frac{1}{\sqrt{T}}\|\hat{f}_2 - f_2^0 H_2\| = O_P(C_{NT}^{-1})$

(iii) $\frac{1}{\sqrt{N_k}}\sum_{i\in\hat{G}_k}\left(\hat{\lambda}_{2i} - H_2^{-1}\lambda_{2i}^0\right) = o_P(1)$,

(iv) $\frac{1}{\sqrt{T}}\left\|\hat{f}_2\hat{\lambda}_{2i} - f_2^0\lambda_{2i}^0\right\| = O_P(C_{NT}^{-1})$,

(v) $\frac{1}{\sqrt{N_k}}\sum_{i\in\hat{G}_k}(\hat{\Delta}_{21,i} - \Delta_{21,i}) = o_P(1)$,

(vi) $\frac{\sqrt{N_k}}{T}\sum_{t=1}^T\sum_{s=1}^T\left(\hat{\varkappa}_{ts} - \varkappa_{ts}\right)\mathbf{1}\left\{s\leqslant t\right\} = o_P(1)$,

(vii) $\frac{1}{\sqrt{N_k}}\sum_{i\in G_k^0}(\hat{\Delta}_{24,i}\hat{\bar{\lambda}}_{2i} - \Delta_{24,i}\bar{\lambda}_{2i}^0) = o_P(1)$,

(viii) $\frac{1}{\sqrt{N_k}T}\sum_{i\in G_k^0}\sum_{t=1}^T\sum_{s=1}^T\left[\hat{\varkappa}_{ts}\mathbf{1}\left\{s\leqslant t\right\}\hat{\Delta}_{24,i}\hat{\bar{\lambda}}_{2i} - \varkappa_{ts}\mathbf{1}\left\{s\leqslant t\right\}\Delta_{24,i}\bar{\lambda}_{2i}^0\right] = o_P(1)$.

where $\bar{\lambda}_{2i}^0 = \lambda_{2i}^0 - \frac{1}{N}\sum_{j=1}^N\lambda_{2j}^0 a_{ij}$.

Proof of Theorem 3.5. (i) We first consider the bias-correction post-Lasso estimators $\operatorname{vec}(\hat{\boldsymbol{\alpha}}_{\hat{G}}^{bc})$. By construction and Theorem 3.4, we have

$$\begin{aligned}\sqrt{NT}\operatorname{vec}\left(\hat{\boldsymbol{\alpha}}_{\hat{G}}^{bc}-\boldsymbol{\alpha}^0\right) &= \sqrt{NT}\operatorname{vec}\left(\hat{\boldsymbol{\alpha}}_{\hat{G}}^{bc}-\hat{\boldsymbol{\alpha}}\right)+\sqrt{NT}\operatorname{vec}\left(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^0\right)\\ &= \sqrt{D_N}Q_{NT}^{-1}V_{NT}+\sqrt{D_N}\left[Q_{NT}^{-1}\left(B_{NT,1}+B_{NT,2}\right)-\hat{Q}_{NT}^{-1}\left(\hat{B}_{NT,1}+\hat{B}_{NT,2}\right)\right]\\ &\quad +o_P(1).\end{aligned}$$

It suffices to show the $\sqrt{NT}\operatorname{vec}\big(\hat{\boldsymbol{\alpha}}_{\hat{G}}^{bc}-\boldsymbol{\alpha}^0\big)=\sqrt{D_N}Q_{NT}^{-1}V_{NT}+o_P(1)$ by (i1) $\hat{Q}_{1NT}-\hat{Q}_{2NT}=Q_{NT}+o_P(1)$, (i2) $\hat{B}_{NT,1}=B_{NT,1}+o_P(1)$, and (i3) $\hat{B}_{NT,2}=B_{NT,2}+o_P(1)$. (i1) holds by Lemma A.9 (i) and (iv). For (i2), it suffices to show that $\hat{B}_{kNT,1}-B_{kNT,1}=o_P\left(1\right)$ for $k=1,...,K$. By Theorem 3.3 and using arguments as used in the proof of Lemma A.9(ii), we can readily show that $\hat{B}_{kNT,1}=\tilde{B}_{kNT,1}+o_P(1)$, where $\tilde{B}_{kNT,1}=\frac{1}{\sqrt{N_k}}\sum_{i\in G_k^0}\hat{\Delta}_{21,i}-\frac{1}{\sqrt{N_k}T}\sum_{i\in G_k^0}\sum_{t=1}^T\sum_{s=1}^T\hat{\varkappa}_{ts}\mathbf{1}\left\{s\leqslant t\right\}\hat{\Delta}_{21,i}$. It follows that

$$\begin{aligned}&\hat{B}_{kNT,1}-B_{kNT,1}\\ =&\frac{1}{\sqrt{N_k}}\sum_{i\in G_k^0}(\hat{\Delta}_{21,i}-\Delta_{21,i})-\frac{1}{\sqrt{N_k}T}\sum_{i\in G_k^0}\sum_{t=1}^T\sum_{s=1}^T\left[\hat{\varkappa}_{ts}\mathbf{1}\left\{s\leqslant t\right\}\hat{\Delta}_{21,i}-\varkappa_{ts}\mathbf{1}\left\{s\leqslant t\right\}\Delta_{21,i}\right]\\ &+o_P(1)\\ =&\frac{1}{\sqrt{N_k}}\sum_{i\in G_k^0}(\hat{\Delta}_{21,i}-\Delta_{21,i})-\frac{1}{T}\sum_{t=1}^T\sum_{s=1}^T\hat{\varkappa}_{ts}\mathbf{1}\left\{s\leqslant t\right\}\left(\frac{1}{\sqrt{N_k}}\sum_{i\in G_k^0}(\hat{\Delta}_{21,i}-\Delta_{21,i})\right)\\ &-\frac{\sqrt{N_k}}{T}\sum_{t=1}^T\sum_{s=1}^T\left(\hat{\varkappa}_{ts}-\varkappa_{ts}\right)\mathbf{1}\left\{s\leqslant t\right\}\left(\frac{1}{N_k}\sum_{i\in G_k^0}\Delta_{21,i}\right)+o_P(1)\\ \equiv&B_{kNT,1}\left(1\right)+B_{kNT,1}\left(2\right)+B_{kNT,1}\left(3\right)+o_P(1),\ \text{say,}\end{aligned}$$

We can prove $\hat{B}_{kNT,1}=B_{kNT,1}+o_P(1)$ by showing that $B_{kNT,1}\left(l\right)=o_P\left(1\right)$ for $l=1,2,3$. Noting that $\left|\frac{1}{T}\sum_{t=1}^T\sum_{s=1}^T\hat{\varkappa}_{ts}\mathbf{1}\left\{s\leqslant t\right\}\right|\leqslant\frac{1}{T^3}\sum_{t=1}^T\sum_{s=1}^T\left\|\hat{f}_{1t}\right\|\left\|\hat{f}_{1s}\right\|=O_P\left(1\right)$ and $\frac{1}{N_k}\sum_{i\in G_k^0}\Delta_{21,i}=O_P\left(1\right)$, these results would follow by Lemma A.11(v)-(vi). To show (i3), we first observe that

$$\begin{aligned}B_{kNT,2}=&\frac{1}{\sqrt{N_k}T}\sum_{i\in G_k^0}\mathbb{E}\left(x_i'|\mathcal{C}\right)M_{f_1^0}f_2^0\left(\lambda_{2i}^0-\frac{1}{N}\sum_{j=1}^N\lambda_{2j}^0a_{ij}\right)\\ =&\frac{1}{\sqrt{N_k}T}\sum_{i\in G_k^0}\mathbb{E}\left(x_i'|\mathcal{C}\right)f_2^0\bar{\lambda}_{2i}^0-\frac{1}{\sqrt{N_k}T}\sum_{i\in G_k^0}\mathbb{E}\left(x_i'|\mathcal{C}\right)P_{f_1^0}f_2^0\bar{\lambda}_{2i}^0\equiv B_{kNT,21}-B_{kNT,22},\ \text{say,}\end{aligned}$$

where $\bar{\lambda}_{2i}^0 = \lambda_{2i}^0 - \frac{1}{N}\sum_{j=1}^N \lambda_{2j}^0 a_{ij}$. Let $\phi^{f_2,f_1f_2} = (\phi^{f_2f_1}(L), \phi^{f_2f_2}(L))$, $\phi_i^{\varepsilon,f_1f_2} = (\phi_i^{\varepsilon f_1}(L), \phi_i^{\varepsilon f_2}(L)) = (\phi^{\varepsilon f_1}(L), \phi^{\varepsilon f_2}(L))$, and $v_t^{f_1f_2} = (v_t^{f_1\prime}, v_t^{f_1\prime})'$. Note that $\varepsilon_{it} = w_{it}^{\varepsilon} = \phi_i^{\varepsilon u}(L) v_{it}^u + \phi_i^{\varepsilon\varepsilon}(L) v_{it}^{\varepsilon} + \phi^{\varepsilon f_1}(L) v_t^{f_1} + \phi^{\varepsilon f_2}(L) v_t^{f_2}$. By the BN decomposition and the independence of $\{v_{it}^{u\varepsilon}\}$ and $\{v_s^{f_1f_2}\}$, we have

$$\begin{aligned} f_{2t}^0 =& S_4 w_{it} = \phi^{f_2f_1}(L)v_t^{f_1} + \phi^{f_2f_2}(L)v_t^{f_2} = \phi^{f_2,f_1f_2}(L)v_t^{f_1f_2} \\ =& \phi^{f_2,f_1f_2}(1)v_t^{f_1f_2} + S_4\tilde{w}_{it-1} - S_4\tilde{w}_{it}, \\ E_{\mathcal{C}}(x_{it}) =& \mathbb{E}_{\mathcal{C}}\left(S_2\sum_{m=1}^t w_{im}\right) = \sum_{m=1}^t \left(\phi_i^{\varepsilon f_1}(L) v_m^{f_1} + \phi_i^{\varepsilon f_2}(L) v_m^{f_2}\right) = \phi^{\varepsilon,f_1f_2}(L) V_t^{f_1f_2} \\ =& \phi_i^{\varepsilon,f_1f_2}(1) V_t^{f_1f_2} + S_2\mathbb{E}_{\mathcal{C}}(\tilde{w}_{i0} - \tilde{w}_{it}). \end{aligned}$$

where $V_t^{f_1f_2} = (V_t^{f_1\prime}, V_t^{f_2\prime})' = \left(\sum_{m=1}^t v_m^{f_1\prime}, \sum_{m=1}^t v_m^{f_2\prime}\right)'$, w_{it} and $\tilde{w}_{it}$ are defined in Assumption 3.1. Let $B_{kNT,21}^* = \frac{1}{\sqrt{N_k}}\sum_{i\in G_k^0} S_2 \sum_{r=0}^{\infty}\sum_{l=0}^{\infty} \phi_{i,l+r}\phi_{i,l}' S_4' \bar{\lambda}_{2i}^0$. It follows that

$$\begin{aligned} & B_{kNT,21} - B_{kNT,21}^* \\ =& \frac{1}{\sqrt{N_k}}\sum_{i\in G_k^0}\frac{1}{T}\sum_{t=1}^T \phi_i^{\varepsilon,f_1f_2}(L) V_t^{f_1f_2} v_t^{f_1f_2\prime}\phi^{f_2,f_1f_2}(L)'\bar{\lambda}_{2i}^0 - \frac{1}{\sqrt{N_k}}\sum_{i\in G_k^0} S_2\sum_{r=0}^{\infty}\sum_{l=0}^{\infty}\phi_{i,l+r}\phi_{i,l}' S_4\bar{\lambda}_{2i}^0 \\ =& \frac{1}{\sqrt{N_k}}\sum_{i\in G_k^0}\frac{1}{T}\sum_{t=1}^T \phi^{\varepsilon,f_1f_2}(1)(V_t^{f_1f_2}v_t^{f_1f_2\prime} - I_r)\phi^{f_2,f_1f_2}(1)'\bar{\lambda}_{2i}^0 \\ & + \frac{1}{\sqrt{N_k}}\sum_{i\in G_k^0} S_2\left\{\frac{1}{T}\sum_{t=1}^{T-1}\left(\mathbb{E}_{\mathcal{C}}(w_{it+1})\tilde{w}_{it}' - \sum_{l=0}^{\infty}\phi_{i,l+1}\phi_{i,l}'\right) S_4'\bar{\lambda}_{2i}^0 - \frac{1}{T}\sum_{l=0}^{\infty}\phi_{i,l+1}\phi_{i,l}' S_4'\bar{\lambda}_{2i}^0\right. \\ & - \frac{1}{T}\sum_{t=1}^T\left(\mathbb{E}_{\mathcal{C}}(\tilde{w}_{i0}) v_t^{f_1f_2\prime}\phi^{f_2,f_1f_2}(1)' - \tilde{\phi}_{i,0}\phi_i(1)' S_4'\right)\bar{\lambda}_{2i}^0 + \frac{1}{T}\sum_{t=1}^T \mathbb{E}_{\mathcal{C}}(\tilde{w}_{it}) v_t^{f_1f_2\prime}\phi^{f_2,f_1f_2}(1)'\bar{\lambda}_{2i}^0 \\ & \left. -\frac{1}{T}E_{\mathcal{C}}\left(\sum_{t=1}^T w_{it}\right)\tilde{w}_{iT}' S_4'\bar{\lambda}_{2i}^0 + \frac{1}{T}E_{\mathcal{C}}(w_{i1})\tilde{w}_{i0}' S_4'\bar{\lambda}_{2i}^0\right\} \\ \equiv& \frac{1}{\sqrt{N_k}}\sum_{i\in G_k^0} Q_{iT}^{f_2} + \frac{1}{\sqrt{N_k}}\sum_{i\in G_k^0} S_2\left\{R_{iT,1}^{f_2} + R_{iT,2}^{f_2} + R_{iT,3}^{f_2} + R_{iT,4}^{f_2} + R_{iT,5}^{f_2} + R_{iT,6}^{f_2}\right\} S_4'\bar{\lambda}_{2i}^0, \end{aligned}$$

where we use the fact that $\phi_i^{\varepsilon,f_1f_2}(1)\phi^{f_2,f_1f_2}(1)' = S_2\phi_i(1)\phi_i(1)' S_4'$ by construction and that $\sum_{r=0}^{\infty}\sum_{l=0}^{\infty}\phi_{i,l+r}\phi_{i,l}' = \phi_i(1)\phi_i(1)' - \sum_{l=0}^{\infty}\phi_{i,l+1}\phi_{i,l}' + \tilde{\phi}_{i,0}\phi_i(1)'$. Following the proof of Lemma A.7 in Huang et al. (2017), we can show that $\frac{1}{\sqrt{N_k}}\sum_{i\in G_k^0} S_2 R_{iT,l}^{f_2} S_4'\bar{\lambda}_{2i}^0 = o_P(1)$ for $l = 1, 2, ..., 6$ and $\frac{1}{\sqrt{N_k}}\sum_{i\in G_k^0}\mathbb{E}(Q_{iT}^{f_2}) = 0$. It follows that $B_{kNT,21} = B_{kNT,21}^* + o_P(1) = \frac{1}{\sqrt{N_k}}\sum_{i\in G_k^0}\Delta_{24,i}\bar{\lambda}_{2i}^0 + o_P(1)$. Analogously, we have $B_{kNT,22} = B_{kNT,22}^* + o_P(1)$, where $B^*_{kNT,22} = \frac{1}{\sqrt{N_k}} \sum_{i \in G_k^0} \frac{1}{T} \sum_{t=1}^{T} \sum_{s=1}^{T} \varkappa_{ts} 1 \left\{ s \leqslant t \right\} \times S_2 \sum_{r=0}^{\infty} \sum_{l=0}^{\infty} \phi_{i,l+r} \phi'_{i,l} S'_4 \bar{\lambda}^0_{2i}$. Let $B^*_{kNT,2} = B^*_{kNT,21} - B^*_{kNT,22}$. Then

$$\begin{aligned} B^*_{kNT,2} =& \frac{1}{\sqrt{N_k}} \sum_{i \in G_k^0} \frac{1}{T} \sum_{t=1}^{T} \sum_{s=1}^{T} \left(\mathbf{1}\{s=t\} - \varkappa_{ts} \mathbf{1} \left\{ s \leqslant t \right\} \right) S_2 \sum_{r=0}^{\infty} \sum_{l=0}^{\infty} \phi_{i,l+r} \phi'_{i,l} S'_4 \bar{\lambda}^0_{2i} \\ =& \frac{1}{T} \sum_{t=1}^{T} \sum_{s=1}^{t} \bar{\varkappa}_{ts} \sum_{r=0}^{\infty} \sum_{l=0}^{\infty} \left(\phi^{\varepsilon f_1}_{l+r} \phi^{f_2 f_1}_{l} + \phi^{\varepsilon f_2}_{l+r} \phi^{f_2 f_2}_{l} \right) \frac{1}{\sqrt{N_k}} \sum_{i \in G_k^0} \bar{\lambda}^0_{2i} \\ =& \frac{1}{T} \sum_{t=1}^{T} \sum_{s=1}^{t} \bar{\varkappa}_{ts} \frac{1}{\sqrt{N_k}} \sum_{i \in G_k^0} \Delta_{24,i} \bar{\lambda}^0_{2i}. \end{aligned}$$

By Theorem 3.3 and using arguments as used in the proof of Lemma A.9(ii), we can readily show that $\hat{B}_{kNT,2} = \tilde{B}_{kNT,2} + o_P(1)$, where $\tilde{B}_{kNT,2} = \frac{1}{\sqrt{N_k}} \sum_{i \in G_k^0} \frac{1}{T} \sum_{t=1}^{T} \sum_{s=1}^{t} \hat{\varkappa}_{ts} \hat{\Delta}_{24,i} \hat{\bar{\lambda}}_{2i}$. Thus we can prove that $\hat{B}_{NT,2} = B_{NT,2} + o_P(1)$ by showing $\tilde{B}_{kNT,2} = B^*_{kNT,2} + o_P(1)$ for $k = 1, ..., K$. Note that

$$\begin{aligned} & \tilde{B}_{kNT,2} - B^*_{kNT,2} \\ = & \frac{1}{\sqrt{N_k}} \sum_{i \in G_k^0} (\hat{\Delta}_{24,i} \hat{\bar{\lambda}}_{2i} - \Delta_{24,i} \bar{\lambda}^0_{2i}) - \frac{1}{\sqrt{N_k} T} \sum_{i \in G_k^0} \sum_{t=1}^{T} \sum_{s=1}^{T} \left[\hat{\varkappa}_{ts} \mathbf{1} \left\{ s \leqslant t \right\} \hat{\Delta}_{24,i} \hat{\bar{\lambda}}_{2i} - \varkappa_{ts} \mathbf{1} \left\{ s \leqslant t \right\} \Delta_{24,i} \bar{\lambda}^0_{2i} \right] \\ = & o_P(1) - o_P(1) = o_P(1) \end{aligned}$$

by Lemma A.11(vii)-(viii). Consequently, $\hat{B}_{kNT,2} - B_{kNT,2} = o_P(1)$.

In sum, we have $\sqrt{N} T \text{vec}\left(\hat{\boldsymbol{\alpha}}^{bc}_{\hat{G}} - \boldsymbol{\alpha}^0 \right) = \sqrt{D_N} Q^{-1}_{NT} V_{NT} + o_P(1)$.

(ii) For the fully-modified post-Lasso estimators $\hat{\alpha}^{fm}_{\hat{G}_k}$, we first consider the asymptotic distribution for the infeasible version of fully modified post-Lasso estimator $\tilde{\alpha}^{fm}_{G_k}$. Noting that $y_i^+ = x_i \alpha^0_k + f^0_1 \lambda^0_{1i} + f^0_2 \lambda^0_{2i} + u_i^+$, by (106) and (107) and Theorem 3.3, we have

$$\begin{aligned} \frac{1}{N_k T^2} \sum_{i \in \hat{G}_k} x'_i M_{\hat{f}_1} x_i (\tilde{\alpha}^{fm}_{G_k} - \alpha^0_k) =& \frac{1}{N_k T^2} \sum_{i \in G_k^0} x'_i M_{\hat{f}_1} \left(u_i^+ + f^0_2 \lambda^0_{2i} \right) + \frac{1}{N_k T^2} \sum_{i \in \hat{G}_k} x'_i M_{\hat{f}_1} f^0_1 \lambda^0_{1i} \\ & - \frac{1}{\sqrt{N_k} T} B^+_{kNT,1} - \frac{1}{\sqrt{N_k} T} B_{kNT,2} + o_P(N^{-1/2} T^{-1}). \end{aligned} \quad (111)$$

Combing (108), (111) and Lemma A.9(i) yields

$$\begin{aligned}
&\frac{1}{N_kT^2}\sum_{i\in\hat{G}_k} x_i'M_{\hat{f}_1}x_i(\tilde{\alpha}_{G_k}^{fm}-\alpha_k^0)-\frac{1}{N_kT^2}\sum_{i\in\hat{G}_k}\frac{1}{N}\sum_{j=1}^{N}x_i'M_{\hat{f}_1}x_ja_{ij}\hat{b}_j\\
=&\frac{1}{N_kT^2}\sum_{i\in G_k^0}x_i'M_{f_1^0}\left(u_i^+-\frac{1}{N}\sum_{j=1}^{N}u_j^+a_{ij}\right)+\frac{1}{N_kT^2}\sum_{i\in G_k^0}x_i'M_{f_1^0}f_2^0\left(\lambda_{2i}^0-\frac{1}{N}\sum_{j=1}^{N}\lambda_{2j}^0a_{ij}\right)\\
&-\frac{1}{\sqrt{N_k}T}B_{kNT,1}^+-\frac{1}{\sqrt{N_k}T}B_{kNT,2}+o_P(N^{-1/2}T^{-1})
\end{aligned}$$

By (108) and Lemma A.10 (i)-(iii), we have

$$\begin{aligned}
\sqrt{N}T\text{vec}(\tilde{\boldsymbol{\alpha}}_G^{fm}-\boldsymbol{\alpha}^0)&=(\hat{Q}_{1NT}-\hat{Q}_{2NT})^{-1}\sqrt{D_N}\left(\left(U_{NT}^{u+}+U_{NT}^{f_2}\right)-B_{NT,1}^+-B_{NT,2}\right)+o_P(1)\\
&=\sqrt{D_N}Q_{NT}^{-1}V_{NT}^++o_P(1)
\end{aligned}$$

where

$$\begin{aligned}
U_{k,NT}^{u+}=&\frac{1}{\sqrt{N_k}T}\sum_{i\in G_k^0}x_i'M_{f_1^0}\left(u_i^+-\frac{1}{N}\sum_{j=1}^{N}a_{ij}u_j^+\right),\\
V_{kNT,1}^+=&\frac{1}{\sqrt{N_k}T}\sum_{i\in G_k^0}S^{\varepsilon}\phi_i^{\dagger}(1)\sum_{t=1}^{T}\sum_{s=1}^{T}\left\{\bar{\varkappa}_{ts}\left(V_{it}^{u\varepsilon}v_{is}^{u\varepsilon,+\prime}\right)-\left[\mathbf{1}\left\{t=s\right\}-\varkappa_{ts}\mathbf{1}\left\{s\leqslant t\right\}\right]I_{1+p}\right\}\phi_i^{\dagger}(1)'S^{u\prime},\\
V_{kNT,2}^+=&\frac{1}{\sqrt{N_k}}\sum_{i=1}^{N}\left\{\frac{1}{T}E\left(x_i'|\mathcal{C}\right)\mathbf{1}\left\{i\in G_k^0\right\}-\frac{1}{N}\sum_{j\in G_k^0}a_{ij}\frac{1}{T}E(x_j'|\mathcal{C})\right\}M_{f_1^0}u_i^+,
\end{aligned}$$

and $U_{k,NT}^+=U_{k,NT}^{u+}+U_{k,NT}^{f_2}$ and $V_{kNT}^+=V_{kNT,1}^++V_{kNT,2}^++V_{kNT,3}$ are the kth block-element of U_{NT}^+ and V_{NT}^+, respectively. We have a new error process $w_{it}^+=(u_{it}^+,\Delta x_{it}',\Delta f_{1t}',f_{2t}',)'$ whose partial sum satisfies the multivariate invariance principle:$\frac{1}{\sqrt{T}}\sum_{t=1}^{[T\cdot]}w_{it}^+\Rightarrow B_i^+=BM(\Omega_i^+)$. Following the proof of Lemma A.10(iii) (see also Theorem 9 in Phillips and Moon), we can show that $V_{NT}^+\xrightarrow{d}N(0,\Omega_0^+)$ condition on $\mathcal{C}$ where $\Omega_0^+=\lim_{N,T\to\infty}\Omega_{NT}^+$ and $\Omega_{NT}^+=$Var$\left(V_{NT}^+|\mathcal{C}\right)$. Then we have

$$\sqrt{N}T\text{vec}(\tilde{\boldsymbol{\alpha}}_G^{fm}-\boldsymbol{\alpha}^0)\xrightarrow{d}MN(0,\lim_{N\to\infty}D_NQ_0^{-1}\Omega_0^+Q_0^{-1}).$$

Next, we show that $\hat{\boldsymbol{\alpha}}_G^{fm}$ is asymptotically equivalent to $\tilde{\boldsymbol{\alpha}}_G^{fm}$ by showing that $\sqrt{N}T(\hat{\boldsymbol{\alpha}}_G^{fm}-\tilde{\boldsymbol{\alpha}}_G^{fm})=o_P\left(1\right)$. Note that

$$\sqrt{N}T(\hat{\boldsymbol{\alpha}}_G^{fm}-\tilde{\boldsymbol{\alpha}}_G^{fm})=\sqrt{D_N}\left[(\hat{Q}_{1NT}-\hat{Q}_{2NT})^{-1}\left(\hat{U}_{NT}^++\hat{B}_{NT,1}^++\hat{B}_{NT,2}\right)-Q_{NT}^{-1}\left(U_{NT}^++B_{NT,1}^++B_{NT,2}\right)\right].$$

Then it suffices to show (ii1) $\hat{Q}_{1NT}-\hat{Q}_{2NT}=Q_{NT}+o_P(1)$, (ii2) $\hat{B}^{+}_{NT,1}=B^{+}_{NT,1}+o_P(1)$,(ii3) $\hat{U}^{+}_{NT}=U^{+}_{NT}+o_P(1)$, and (ii4) $\hat{B}_{NT,2}=B_{NT,2}+o_P(1)$. In the proof of bias-correction post-Lasso estimators, we have already prove (ii1) and (ii4). For (ii2), we can apply analogous arguments as used in the proof of Lemma A.11(v) to prove that $\mathbb{E}_{\mathcal{C}}\left\|\frac{1}{\sqrt{N_k}}\sum_{i\in\hat{G}_k}(\hat{\Omega}_i-\Omega_i)\right\|=O_P(\frac{H}{T}+\frac{N}{H^{2q}})=o_P(1)$. Since $\Delta^{+}_{lm,i}=\Delta_{lm,i}-\Omega_{lm,i}\Omega^{-1}_{mi}\Delta_{m,i}$, this implies that $\left\|\frac{1}{\sqrt{N_k}}\sum_{i\in\hat{G}_k}(\hat{\Delta}^{+}_{21,i}-\Delta^{+}_{21,i})\right\|^2=o_P(1)$. The latter further implies that $\hat{B}^{+}_{NT,1}=B^{+}_{NT,1}+o_P(1)$. For (ii3) we can apply Theorem 3 to show that

$$
\begin{aligned}
&\hat{U}^{+}_{kNT}-U^{+}_{kNT}\\
=&\hat{U}^{u+}_{kNT}-\tilde{U}^{u+}_{kNT}+\tilde{U}^{u+}_{kNT}-U^{u+}_{kNT}\\
=&\frac{1}{\sqrt{N_k}T}\sum_{i\in\hat{G}_k}x_i'M_{\hat{f}_1}\left(\hat{u}^{+}_i-\frac{1}{N}\sum_{j=1}^{N}a_{ij}\hat{u}^{+}_j\right)-\frac{1}{\sqrt{N_k}T}\sum_{i\in\hat{G}_k}x_i'M_{\hat{f}_1}\left(u^{+}_i-\frac{1}{N}\sum_{j=1}^{N}a_{ij}u^{+}_j\right)+o_P(1)\\
=&\frac{1}{\sqrt{N_k}T}\sum_{i\in G^0_k}x_i'M_{\hat{f}_1}\left(\hat{u}^{+}_i-u^{+}_i\right)-\frac{1}{\sqrt{N_k}NT}\sum_{i\in G^0_k}\sum_{j=1}^{N}x_i'M_{\hat{f}_1}\left(\hat{u}^{+}_j-u^{+}_j\right)a_{ij}+o_P(1)\\
=&\frac{1}{\sqrt{N_k}T}\sum_{i\in G^0_k}x_i'\Delta x_i\left(\Omega_{12,i}\Omega^{-1}_{22i}-\hat{\Omega}_{12,i}\hat{\Omega}^{-1}_{22i}\right)-\frac{1}{\sqrt{N_k}T}\sum_{i\in G^0_k}x_i'P_{\hat{f}_1}\Delta x_i\left(\Omega_{12,i}\Omega^{-1}_{22i}-\hat{\Omega}_{12,i}\hat{\Omega}^{-1}_{22i}\right)\\
&-\frac{1}{\sqrt{N_k}NT}\sum_{i\in G^0_k}\sum_{j=1}^{N}x_i'M_{\hat{f}_1}\Delta x_j\left(\Omega_{12,j}\Omega^{-1}_{22j}-\hat{\Omega}_{12,j}\hat{\Omega}^{-1}_{22j}\right)a_{ij}+o_P(1)\\
\equiv&UU_1+UU_2+UU_3+o_P(1),
\end{aligned}
$$

where $\tilde{U}^{u+}_{kNT}=\frac{1}{\sqrt{N_k}T}\sum_{i\in G^0_k}x_i'M_{\hat{f}_1}\left(u^{+}_i-\frac{1}{N}\sum_{j=1}^{N}a_{ij}u^{+}_j\right)$ and $\tilde{U}^{u+}_{kNT}-U^{u+}_{kNT}=o_P(1)$ by Lemma A.9(iii). Following the proof of Lemma A.11(v), we can show that $UU_l=o_P(1)$ for $l=1$, 2, 3. The (ii3) follows. This completes the proof of (ii).

(iii) The proof is analogous to that of (ii) and thus omitted. ■

To prove Theorems 3.6-3.7, we need the following two lemmas.

Lemma A.12 *Suppose that Assumptions 3.1-3.3 and 3.5 hold. Then*

(i) For any $1\leqslant r\leqslant r^0$, $V_1(r,\hat{G}^r)-V_1(r,G^0H^r)=O_P(C^{-1}_{NT})$,

(ii) For each r with $0\leqslant r<r^0$, *there exist a positive number* c_r *such that* $\operatorname{plim}_{(N,T)\to\infty}\inf(V_1(r,G^0H^r)-V_1(r^0,G^0))=c_r$,

(iii) For any fixed r, with $r^0\leqslant r\leqslant r_{\max}$, $V_1(r,\hat{G}^r)-V_1(r^0,\hat{G}^{r^0})=O_P(C^{-2}_{NT})$,

where $V_1(r, G^0 H^r)$ is defined analogously to $V_1(r, \hat{G}^r)$ with $\hat{G}^r$ replaced by $G^0 H^r$, $H^r = (N^{-1}\Lambda^{0\prime}\Lambda^0)(T^{-1}G^{0\prime}\hat{G}^r)$, and $G^0 = \Delta f^0$.

Lemma A.13 *Suppose that Assumptions 3.1-3.3 and 3.6 hold. Then*

(i) For any $1 \leqslant r_1 \leqslant r_1^0$, $V_2(r_1, \hat{f}_1^{r_1}) - V_2(r_1, f_1^0 H_1^{r_1}) = O_P(\sqrt{T})$,

(ii) For any $1 \leqslant r_1 < r_1^0$, we have $\text{plim}_{(N,T)\to\infty} \inf d_T T^{-1}[V_2(r_1, f_1^0 H_1^{r_1}) - V_2(r_1, f_1^0)] = d_{r_1}$ for some $d_{r_1} > 0$,

(iii) For any $r_1^0 \leqslant r_1 \leqslant r_{\max}$, $V_2(r_1, \hat{f}^{r_1}) - V_2(r_1^0, \hat{f}^{r_1^0}) = O_P(1)$,

where $V_2(r_1, f_1^0 H_1^{r_1})$ is defined analogously to $V_2(r_1, \hat{f}_1^{r_1})$ with $\hat{f}_1^{r_1}$ replaced by $f_1^0 H_1^{r_1}$, and $H_1^{r_1} = (N^{-1}\Lambda^{0\prime}\Lambda^0) \times (T^{-2} f^{0\prime}\hat{f}^{r_1})$.

Proof of Theorem 3.6. Noting that $IC_1(r) - IC_1(r^0) = V_1(r, \hat{G}^r) - V_1(r^0, \hat{G}^{r^0}) - (r^0 - r)g_1(N,T)$, it suffices to show that $P\left(V_1(r, \hat{G}^r) - V_1(r^0, \hat{G}^{r^0}) < (r^0 - r)g_1(N,T)\right) \to 0$ as $(N,T) \to \infty$ when $r \neq r^0$. We consider the under- and over-fitted models, respectively. When $0 \leqslant r < r^0$, we make the following decomposition:

$$\begin{aligned} V_1(r, \hat{G}^r) - V_1(r^0, \hat{G}^{r^0}) =& [V_1(r, \hat{G}^r) - V(r, G^0 H^r)] + [V_1(r, G^0 H^r) - V_1(r^0, G^0 H^{r^0})] \\ &+ [V_1(r^0, G^0 H^{r^0}) - V_1(r^0, \hat{G}^{r^0})]. \end{aligned}$$

Lemma A.12(i) implies that the first and third terms on the right hand side of the last displayed equation are both $O_P(C_{NT}^{-1})$. Noting that $V_1(r^0, G^0 H^{r^0}) = V_1(r^0, G^0)$, the second term has a positive probability limit c_r when $r < r^0$ by Lemma A.12(ii). It follows that $P(IC_1(r) < IC_1(r^0)) \to 0$ as $g_1(N,T) \to 0$ as $(N,T) \to \infty$ under Assumption 3.5.

Now, we consider the case where $r^0 < r \leqslant r_{\max}$, Note that $C_{NT}^2\left(V_1(r, \hat{G}^r) - V_1(r^0, \hat{G}^{r^0})\right) = O_P(1)$ and $C_{NT}^2(r - r^0)g_1(N,T) > C_{NT}^2 g_1(N,T) \to \infty$ by Lemma A.12(iii) and Assumption 3.5, we have $P(IC_1(r) < IC_1(r^0)) = P(V_1(r, \hat{G}^r) - V_1(r^0, \hat{G}^{r^0}) < (r^0 - r)g_1(N,T)) \to 0$ as $(N,T) \to \infty$. ■

Proof of Theorem 3.7. Noting that $IC_2(r_1) - IC_2(r_1^0) = V_2(r_1, \hat{f}_1^{r_1}) - V_2(r_1^0, \hat{f}_1^{r_1^0}) - (r_1^0 - r_1)g_2(N,T)$, it suffices to show that $P\left(V_2(r_1, \hat{f}_1^{r_1}) - V_2(r_1^0, \hat{f}_1^{r_1^0}) < (r_1^0 - r_1)g_2(N,T)\right) \to$

0 as $(N,T)\to\infty$ when $r\neq r^0$. First, when $r_1<r_1^0$, we consider the decomposition:

$$\begin{aligned}V_2(r_1,\hat{f}_1^{r_1})-V_2(r_1^0,\hat{f}_1^{r_1^0}) =& \left[V_2(r_1,\hat{f}^{r_1})-V_2(r_1,f_1^0H_1^{r_1})\right]+\left[V_2(r_1,f_1^0H_1^{r_1})-V_2(r_1^0,f_1^0H_1^{r_1^0})\right]\\ &+\left[V(r_1^0,f_1^0H_1^{r_1^0})-V(r_1^0,\hat{f}_1^{r_1^0})\right]\\ \equiv& DD_1+DD_2+DD_3,\text{ say.}\end{aligned}$$

By Lemma A.13), $DD_1=O_P(T^{1/2})$, DD_2 is of exact probability order $O_P(T/\log\log T)$, and $DD_3=O_P\left(1\right)$. It follows that

$$P(IC_2(r_1)<IC_2(r_1^0))=P\left(V_2(r_1,\hat{f}_1^{r_1})-V_2(r_1^0,\hat{f}_1^{r_1^0})<(r_1^0-r_1)g_2(N,T)\right)\to 0$$

as $g_2(N,T)\log\log T/T\to 0$ under Assumption 3.6.

Next, for $r_1>r_1^0$, we have $V(r_1,\hat{f}_1^{r_1})-V(r_1^0,\hat{f}_1^{r_1^0})=O_P(1)$ for $r_1>r_1^0$ by Lemma A.13(iii), and $(r_1-r_1^0)g_2(N,T)\to\infty$ by Assumption 3.6. This implies that

$$P(IC_2(r_1)-IC_2(r_1^0)<0)=P(V_2(r_1,\hat{f}_1^{r_1})-V_2(r_1^0,\hat{f}_1^{r_1^0})<(r_1^0-r_1)g_2(N,T))\to 0.$$

as $N,T\to\infty$. ■

To prove Theorem 3.8, we need the following lemma.

Lemma A.14 *Suppose that Assumptions 3.1-3.3 and 3.7 hold. Then*

$$\max_{K_0\leqslant K\leqslant K_{\max}}\left|\hat{\sigma}^2_{G_{(K,\lambda)}}-\hat{\sigma}^2_{\hat{G}(K_0,\lambda)}\right|=O_P(\nu_{NT}^2)$$

where $\hat{\sigma}^2_{G_{(K,\lambda)}}=\frac{1}{NT}\sum_{k=1}^K\sum_{i\in\hat{G}_k(K,\lambda)}\sum_{t=1}^T[y_{it}-\hat{\alpha}^{cup\prime}_{\hat{G}_k(K,\lambda)}x_{it}-\hat{\lambda}_{1i}(K,\lambda)'\hat{f}_{1t}(K,\lambda)]^2$ *and* ν_{NT} *is defined in Section 3.6.*

Proof of Theorem 3.8. First, we can show that

$$\begin{aligned}IC_3(K_0,\lambda)&=\ln[V_3(K_0)]+pK_0g_3(N,T)\\ &=\ln\frac{1}{NT}\sum_{k=1}^{K_0}\sum_{i\in\hat{G}_k(K_0,\lambda)}\sum_{t=1}^{T}\left[y_{it}-\hat{\alpha}^{fm\prime}_{\hat{G}_k(K_0,\lambda)}x_{it}-\hat{\lambda}_{1i}(K_0,\lambda)'\hat{f}_{1t}(K_0,\lambda)\right]^2+o(1)\xrightarrow{p}\ln(\sigma_0^2).\end{aligned}$$

We consider the cases of under- and over-fitted models separately. When $1 \leqslant K < K_0$, we have

$$\begin{aligned}V_3(K) =&\frac{1}{NT}\sum_{k=1}^{K}\sum_{i\in\hat{G}_k(K_0,\lambda)}\sum_{t=1}^{T}\left[y_{it}-\hat{\alpha}^{fm\prime}_{\hat{G}_k(K,\lambda)}x_{it}-\hat{\lambda}_{1i}(K,\lambda)'\hat{f}_{1t}(K,\lambda)\right]^2\\ \geqslant&\min_{1\leqslant K<K_0}\inf_{G^{(K)}\in G_K}\frac{1}{NT}\sum_{k=1}^{K}\sum_{i\in G_{K,k}}\sum_{t=1}^{T}\left[y_{it}-\hat{\alpha}^{fm\prime}_{\hat{G}_k(K,\lambda)}x_{it}-\hat{\lambda}_{1i}(K,\lambda)'\hat{f}_{1t}(K,\lambda)\right]^2\\ =&\min_{1\leqslant K<K_0}\inf_{G^{(K)}\in G_K}\hat{\sigma}^2_{G^{(K)}}.\end{aligned}$$

By Assumption 3.7 and Slutsky's Lemma, we can demonstrate

$$\min_{1\leqslant K<K_0} IC_3(K,\lambda)\geqslant\min_{1\leqslant K<K_0}\inf_{G^{(K)}\in G_K}\ln(\hat{\sigma}^2_{G^{(K)}})+pKg_3(N,T)\xrightarrow{p}\ln(\underline{\sigma}^2)>\ln(\sigma_0^2).$$

It follows that $P(\min_{1\leqslant K<K_0} IC_3(K,\lambda)>IC_3(K_0,\lambda))\to 1$.

When $K_0 < K \leqslant K_{\max}$, we can show that $NT[\hat{\sigma}^2_{\hat{G}(K,\lambda)}-\hat{\sigma}^2_{\hat{G}(K_0,\lambda)}]=O_P(1)$ when there is no unobserved common factor and no endogeneity in x_{it}, $\delta^2_{NT}[\hat{\sigma}^2_{\hat{G}(K,\lambda)}-\hat{\sigma}^2_{\hat{G}(K_0,\lambda)}]=O_P(1)$ when there are only unobserved nonstationary common factors and $C^2_{NT}[\hat{\sigma}^2_{\hat{G}(K,\lambda)}-\hat{\sigma}^2_{\hat{G}(K_0,\lambda)}]=O_P(1)$ when there are both nonstationary and stationary common factors. Then by Lemma 14,

$$\begin{aligned}&P\left(\min_{K\in K^+} IC_3(K,\lambda)>IC_3(K_0,\lambda)\right)\\ =&P\left(\min_{K\in K^+}\nu_{NT}^{-2}\ln\left(\hat{\sigma}^2_{\hat{G}(K,\lambda)}/\hat{\sigma}^2_{\hat{G}(K_0,\lambda)}\right)+\nu_{NT}^{-2}g_3(N,T)(K-K_0)>0\right)\\ \approx&P\left(\min_{K\in K^+}\nu_{NT}^{-2}\left(\hat{\sigma}^2_{\hat{G}(K,\lambda)}-\hat{\sigma}^2_{\hat{G}(K_0,\lambda)}\right)/\hat{\sigma}^2_{\hat{G}(K_0,\lambda)}+\nu_{NT}^{-2}g_3(N,T)(K-K_0)>0\right)\\ \to&1\quad\text{as}\quad(N,T)\to\infty.\ \blacksquare\end{aligned}$$

C Appendix to Chapter 3

This appendix provides proofs for main theorems and ancillary results in the above paper. The proof relies on some technical lemmas whose proofs are given in the Supplemental Appendix.

NOTATIONS: (i) $\hat{\gamma}_i = \hat{\Sigma}_i^{-1}\hat{\alpha}_i \left(\hat{\alpha}_i'\hat{\Sigma}_i^{-1}\hat{\alpha}_i\right)^{-1}$, and $\tilde{\gamma}_i = \tilde{\Sigma}_i^{-1}\tilde{\alpha}_i \left(\tilde{\alpha}_i'\tilde{\Sigma}_i^{-1}\tilde{\alpha}_i\right)^{-1}$

(ii) $u_{it} = \Lambda_i^{0\prime} F_t^0 + \varepsilon_{it}$, and $v_{it} = R_i(L)\beta_i^{0\prime} u_{it}$

(ii) $\delta_i = b_i - b_i^0$, and $\hat{\delta}_i = \hat{b}_i - b_i^0$.

(iii) Let $\eta_{NT}^2 = \frac{1}{N}\sum_{i=1}^N \left\| vec(\hat{b}_i - b_i^0)\right\|^2$, $C_{NT} = \min\left(\sqrt{N}, \sqrt{T}\right)$, $\iota_T = O\left(\log\log T\right)$, and $\psi_{NT} = N^{1/q}T^{-1}\left(\log T\right)^{1+\epsilon}$ for some $\epsilon > 0$.

(iv) Without loss of generality, we assume that $y_{i0} = 0$ in all proofs.

C.1 Proofs of the Main Results in Section 3

Lemma B.1 *Suppose that Assumptions 3.1-3.2 hold. Then for each* $i = 1, ..., N$,

(i) $\frac{1}{T^2}\sum_{t=1}^T y_{i,t-1}^{(2)} y_{i,t-1}^{(2)\prime} \Rightarrow b_{i,\perp}^0 \left(\alpha_{i,\perp}^{0\prime}\beta_{i,\perp}^0\right)^{-1} \alpha_{i,\perp}^{0\prime} \left(\int B_{ui}B_{ui}'\right) \alpha_{i,\perp}^0 \left(\beta_{i,\perp}^{0\prime}\alpha_{i,\perp}^0\right)^{-1} b_{i,\perp}^{0\prime}$,

(ii) $\frac{1}{T}\sum_{t=1}^T y_{i,t-1}^{(2)}\left(\varepsilon_{it} + \Lambda_i^{0\prime}F_t^0\right)' \Rightarrow b_{i,\perp}^0 \left(\alpha_{i,\perp}^{0\prime}\beta_{i,\perp}^0\right)^{-1} \alpha_{i,\perp}^{0\prime} \left(\int B_{ui}dB_{ui}' + \Gamma_{uu,i}\right) + \alpha_{2i}^0 \left(\beta_i^{0\prime}\alpha_i^0\right)^{-1}\Sigma_{vu,i}$,

(iii) $\frac{1}{T}\sum_{t=1}^T y_{i,t-1}^{(2)}\left(u_{it} - \Sigma_{uv,i}\Sigma_{vv,i}^{-1}v_{i,t-1}\right)'$
$\Rightarrow b_{i,\perp}^0 \left(\alpha_{i,\perp}^{0\prime}\beta_{i,\perp}^0\right)^{-1} \alpha_{i,\perp}^{0\prime} \left\{\int B_{ui}dB_{ui}' \left(I_J + \Sigma_{uv,i}\Sigma_{vv,i}^{-1}\left(\beta_i^{0\prime}\alpha_i^0\right)^{-1}\beta_i^{0\prime}\right)' + \Gamma_{uu,i} + \Delta_{uv,i}\Sigma_{vv,i}^{-1}\Sigma_{vu,i}\right\}$,

(iv) $\frac{1}{T}\sum_{t=1}^T y_{i,t-1}^{(2)}\Delta y_{it}' \Rightarrow b_{i,\perp}^0 \left(\alpha_{i,\perp}^{0\prime}\beta_{i,\perp}^0\right)^{-1} \alpha_{i,\perp}^{0\prime} \left(\int B_{ui}dB_{ui}' \left(I_J - \beta_i^0\left(\alpha_i^{0\prime}\beta_i^0\right)^{-1}\alpha_i^{0\prime}\right) + \Gamma_{uu,i} + \Delta_{uv,i}\right)$
$+\alpha_{2i}^0\left(\beta_i^{0\prime}\alpha_i^0\right)^{-1}\left(\Sigma_{vu,i} + \Sigma_{vv,i}\alpha_i^{0\prime}\right)$. where $\Gamma_{uu,i} = S_i\Gamma_i S_i'$ and $\Delta_{uv,i} = \sum_{h=0}^\infty \mathbb{E}\left(u_{it}v_{it+h}'\right)$.

Lemma B.2 *Suppose that Assumptions 3.1-3.2 hold. Then for any fixed small constant* $c > 0$*,*

(i) $\limsup_{T\to\infty}\mu_{\max}\left(\frac{Y_{i,-1}^{(2)\prime}Y_{i,-1}^{(2)}}{\iota_T T^2}\right) \leqslant c\rho_{\max}$ *a.s.,*

(ii) $\liminf_{T\to\infty}\mu_{\min}\left(\frac{\iota_T Y_{i,-1}^{(2)\prime}Y_{i,-1}^{(2)}}{T^2}\right) \geqslant c\rho_{\min}$ *a.s.*

Proof for Theorem 3.1. (i) By Lemma 3.2 in Cheng and Phillips (2009), the Johansen's maximum likelihood procedure still provides consistent estimator for the long-run cointegration vector β_i and $\hat{\beta}_i - \beta_i^0 = O_p(T^{-1})$. It follows that

$$\begin{aligned}
\hat{\alpha}_i =& S_{01,i}\hat{\beta}_i \left(\hat{\beta}_i' S_{11,i}\hat{\beta}_i\right)^{-1} \\
=& \left\{ \left(\frac{1}{T}\sum_{t=1}^{T} \Delta y_{it} y_{i,t-1}' \beta_i^0\right) + \left(\frac{1}{T}\sum_{t=1}^{T} \Delta y_{it} y_{i,t-1}' \left(\hat{\beta}_i - \beta_i^0\right)\right) \right\} \left(\frac{1}{T}\sum_{t=1}^{T} \hat{\beta}_i' y_{i,t-1} y_{i,t-1}' \hat{\beta}_i\right)^{-1} \\
=& \left(\frac{1}{T}\sum_{t=1}^{T} \Delta y_{it} y_{i,t-1}' \beta_i^0\right) \left(\frac{1}{T}\sum_{t=1}^{T} \beta_i^{0\prime} y_{i,t-1} y_{i,t-1}' \beta_i^0\right)^{-1} + O_p(T^{-1}) \\
=& \alpha_i^0 + \left(\frac{1}{T}\sum_{t=1}^{T} u_{it} v_{it-1}'\right) \left(\frac{1}{T}\sum_{t=1}^{T} v_{it-1} v_{it-1}'\right)^{-1} + O_p(T^{-1}) \xrightarrow{p} \alpha_i^0 + \Sigma_{uv,i}\Sigma_{vv,i}^{-1} \equiv \tilde{\alpha}_i
\end{aligned}$$

where $\frac{1}{T}\sum_{t=1}^{T} v_{it-1} v_{it-1}' \xrightarrow{p} E\left(v_{it-1} v_{it-1}'\right) \equiv \Sigma_{vv,i}^{-1}$,and $\frac{1}{T}\sum_{t=1}^{T} u_{it} v_{it-1}' \xrightarrow{p} E\left(u_{it} v_{it-1}'\right) \equiv \Sigma_{uv,i} \neq 0$ due to weak dependence in innovation processes w_{it}. Similarly, we have $\hat{\Sigma}_i = \frac{1}{T}\sum_{t=1}^{T} \left(\Delta y_{it} - \tilde{\alpha}_i \beta_i^{0\prime} y_{i,t-1}\right)\left(\Delta y_{it} - \tilde{\alpha}_i \beta_i^{0\prime} y_{i,t-1}\right)' + O_p(T^{-1/2}) \xrightarrow{p} \tilde{\Sigma}_i$. (i) directly follows.

(ii) For $\hat{b}_i^{GLS}$, we have

$$\begin{aligned}
\hat{b}_i^{GLS} &= \left(\sum_{t=1}^{T} y_{i,t-1}^{(2)} y_{i,t-1}^{(2)\prime}\right)^{-1} \left(\sum_{t=1}^{T} y_{i,t-1}^{(2)} \left(\Delta y_{it} - \tilde{\alpha}_i y_{i,t-1}^{(1)}\right)' \hat{\gamma}_i\right) \\
&= b_i^0 + \left(\frac{1}{T^2}\sum_{t=1}^{T} y_{i,t-1}^{(2)} y_{i,t-1}^{(2)\prime}\right)^{-1} \left\{\frac{1}{T^2}\sum_{t=1}^{T} y_{i,t-1}^{(2)} \left(\left(u_{it} - \Sigma_{uv,i}\Sigma_{vv,i}^{-1} v_{i,t-1}\right)' \tilde{\gamma}_i + \Delta y_{it}' \left(\hat{\gamma}_i - \tilde{\gamma}_i\right)\right)\right\}.
\end{aligned}$$

By part (i) and Lemma B.1(iv), we have $\hat{\gamma}_i - \tilde{\gamma}_i = \hat{\Sigma}_i^{-1}\hat{\alpha}_i \left(\hat{\alpha}_i' \hat{\Sigma}_i^{-1} \hat{\alpha}_i\right)^{-1} - \tilde{\Sigma}_i^{-1}\tilde{\alpha}_i \left(\tilde{\alpha}_i' \tilde{\Sigma}_i^{-1} \tilde{\alpha}_i\right)^{-1} = O_p(T^{-1/2})$. and $\frac{1}{T}\sum_{t=1}^{T} y_{i,t-1}^{(2)} \Delta y_{it}' = O_p(1)$. It follows that $\frac{1}{T}\sum_{t=1}^{T} y_{i,t-1}^{(2)} \Delta y_{it}' \left[\hat{\gamma}_i - \tilde{\gamma}_i\right] = O_p(T^{-1/2})$. Then by Lemma B.1 (i)-(iii), we have

$$\begin{aligned}
&\frac{1}{T^2}\sum_{t=1}^{T} y_{i,t-1}^{(2)} y_{i,t-1}^{(2)\prime} \Rightarrow b_{i,\perp}^0 \left(\alpha_{i,\perp}^{0\prime}\beta_{i,\perp}^0\right)^{-1} \alpha_{i,\perp}^{0\prime} \left(\int B_{ui} B_{ui}'\right) \alpha_{i,\perp}^0 \left(\beta_{i,\perp}^{0\prime}\alpha_{i,\perp}^0\right)^{-1} b_{i,\perp}^{0\prime}, \text{and} \\
&\frac{1}{T}\sum_{t=1}^{T} y_{i,t-1}^{(2)} \left(u_{it} - \Sigma_{uv,i}\Sigma_{vv,i}^{-1} v_{i,t-1}\right)' \\
\Rightarrow& b_{i,\perp}^0 \left(\alpha_{i,\perp}^{0\prime}\beta_{i,\perp}^0\right)^{-1} \alpha_{i,\perp}^{0\prime} \left\{\int B_{ui} dB_{ui}' \left(I_J + \Sigma_{uv,i}\Sigma_{vv,i}^{-1} \left(\beta_i^{0\prime}\alpha_i^0\right)^{-1} \beta_i^{0\prime}\right)' + \Gamma_{uu,i} + \Delta_{uv,i}\Sigma_{vv,i}^{-1}\Sigma_{vu,i}\right\},
\end{aligned}$$

Combining above two results, it directly yields $T\left(\hat{b}_i^{GLS} - b_i^0\right) = O_p(1)$.

(iii) Combine (2.10) and the fact that

$$\Delta\hat{Y}_i^{+} - Y_{i,-1}^{(1)} - Y_{i,-1}^{(2)}\hat{b}_i = -Y_{i,-1}^{(2)}\hat{\delta}_i + F^0\tilde{\Lambda}_i^0 + \hat{\varepsilon}_i^{*},$$

where $\hat{\varepsilon}_i^* = \varepsilon_i^*\tilde{\gamma}_i + \Delta Y_i\left(\hat{\gamma}_i - \tilde{\gamma}_i\right)$, $\varepsilon_{it}^* = \varepsilon_{it} - \Sigma_{uv,i}\Sigma_{vv,i}^{-1}v_{i,t-1}$ and $\hat{\delta}_i = \hat{b}_i - b_i^0$. Note that $\hat{\varepsilon}_i^*$ is weakly dependent stationary process. We have

$$\begin{aligned}
&\left(\hat{F}\otimes I_r\right)V_{NT} \\
&= \frac{1}{NT}\sum_{i=1}^{N}(Y_{i,-1}^{(2)}\otimes I_r)\text{vec}(\hat{\delta}_i)\text{vec}(\hat{\delta}_i)'\left(Y_{i,-1}^{(2)\prime}\hat{F}\otimes I_r\right) - \frac{1}{NT}\sum_{i=1}^{N}\text{vec}(\hat{\varepsilon}_i^*)\text{vec}(\hat{\delta}_i)'\left(Y_{i,-1}^{(2)\prime}\hat{F}\otimes I_r\right) \\
&- \frac{1}{NT}\sum_{i=1}^{N}\left(F^0\otimes I_r\right)\text{vec}(\tilde{\Lambda}_i^0)\text{vec}(\hat{\delta}_i)'\left(Y_{i,-1}^{(2)\prime}\hat{F}\otimes I_r\right) - \frac{1}{NT}\sum_{i=1}^{N}\left(Y_{i,-1}^{(2)\prime}\otimes I_r\right)\text{vec}(\hat{\delta}_i)\text{vec}(\hat{\varepsilon}_i^*)'\left(\hat{F}\otimes I_r\right) \\
&- \frac{1}{NT}\sum_{i=1}^{N}\left(Y_{i,-1}^{(2)\prime}\otimes I_r\right)\text{vec}(\hat{\delta}_i)\text{vec}(\tilde{\Lambda}_i^{0\prime})\left(F^{0\prime}\hat{F}\otimes I_r\right) + \frac{1}{NT}\sum_{i=1}^{N}\text{vec}(\hat{\varepsilon}_i^*)\text{vec}(\hat{\varepsilon}_i^*)'\left(\hat{F}\otimes I_r\right) \\
&+ \frac{1}{NT}\sum_{i=1}^{N}\text{vec}(\hat{\varepsilon}_i^*)\text{vec}(\tilde{\Lambda}_i^0)'\left(F^{0\prime}\hat{F}\otimes I_r\right) + \frac{1}{NT}\sum_{i=1}^{N}\left(F^0\otimes I_r\right)\text{vec}(\tilde{\Lambda}_i^0)\text{vec}(\hat{\varepsilon}_i^*)'\left(\hat{F}\otimes I_r\right) \\
&+ \frac{1}{NT}\sum_{i=1}^{N}\left(F^0\otimes I_r\right)\text{vec}(\tilde{\Lambda}_i^0)\text{vec}(\tilde{\Lambda}_i^0)'\left(F^{0\prime}\hat{F}\otimes I_r\right) \\
\equiv& I_1 + I_2 + ... + I_8 + \frac{1}{NT}\sum_{i=1}^{N}\left(F^0\otimes I_r\right)\text{vec}(\tilde{\Lambda}_i^0)\text{vec}(\tilde{\Lambda}_i^0)'\left(F^{0\prime}\hat{F}\otimes I_r\right),\ \text{say.}
\end{aligned}$$

It follows that

$$\left(\hat{F}\otimes I_r\right)V_{NT} - \left(F^0\otimes I_r\right)\left(\frac{1}{N}\sum_{i=1}^{N}\text{vec}(\tilde{\Lambda}_i^0)\text{vec}(\tilde{\Lambda}_i^0)'\right)\left(\frac{F^{0\prime}\hat{F}}{T}\otimes I_r\right) = I_1 + I_2 + ... + I_8.$$

Define $H = \left(\frac{1}{N}\sum_{i=1}^{N}\text{vec}(\tilde{\Lambda}_i^0)\text{vec}(\tilde{\Lambda}_i^0)'\right)\left(\frac{F^{0\prime}\hat{F}}{T}\otimes I_r\right)V_{NT}^{-1} = O_p(1)$ and H is asymptotically nonsingular. Note that

$$\frac{1}{\sqrt{T}}\left\|\left(\hat{F}\otimes I_r\right)H^{-1} - \left(F^0\otimes I_r\right)\right\| \leqslant \frac{1}{\sqrt{T}}\left(\|I_1\| + \|I_2\| + ... + \|I_8\|\right)\|G\|$$

where $G = \left(\frac{F^{0\prime}\hat{F}}{T}\otimes I_r\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\text{vec}(\tilde{\Lambda}_i^0)\text{vec}(\tilde{\Lambda}_i^0)'\right)^{-1} = O_p(1)$. It remains to show the properties of $\|I_l\|$ for $l = 1, 2, ..., 8$. For I_1,

$$\frac{1}{\sqrt{T}}\|I_1\| \leqslant T\max_{1\leqslant i\leqslant N}\frac{\left\|Y_{i,-1}^{(2)}\right\|^2}{T^2}\frac{\left\|\hat{F}\right\|}{\sqrt{T}}\frac{1}{N}\sum_{i=1}^{N}\left\|\text{vec}(\hat{\delta}_i)\right\|^2 = O_p(T\iota_T\eta_{NT}^2)$$

where $\max_{1\leqslant i\leqslant N}\frac{\left\|Y_{i,-1}^{(2)}\right\|^2}{T^2} \leqslant \mu_{\max}\left(\frac{1}{T^2}\sum_{t=1}^{T}y_{i,t-1}^{(2)}y_{i,t-1}^{(2)\prime}\right) = O_p(\iota_T)$ by Lemma B.2(i) and $\frac{\|\hat{F}\|}{\sqrt{T}} \leqslant \sqrt{r}$. For I_2,

$$\frac{1}{\sqrt{T}}\|I_2\| \leqslant \sqrt{T}\frac{\left\|\hat{F}\right\|}{\sqrt{T}}\max_{1\leqslant i\leqslant N}\frac{\left\|Y_{i,-1}^{(2)}\right\|}{T}\left\{\frac{1}{N}\sum_{i=1}^{N}\frac{\|\text{vec}(\hat{\varepsilon}_i^*)\|^2}{T}\right\}^{1/2}\left\{\frac{1}{N}\sum_{i=1}^{N}\left\|\text{vec}(\hat{\delta}_i)\right\|^2\right\}^{1/2} = O_p(\sqrt{T\iota_T}\eta_{NT})$$

where $\frac{1}{N}\sum_{i=1}^{N}\frac{\|vec(\hat{\varepsilon}_i^*)\|^2}{T}=O_p(1)$ due to stationarity of $\hat{\varepsilon}_i^*$. For I_3,

$$\begin{aligned}\frac{1}{\sqrt{T}}\left\|I_3\right\| &\leqslant \sqrt{T}\frac{\|F^0\|}{\sqrt{T}}\frac{\left\|\hat{F}\right\|}{\sqrt{T}}\max_{1\leqslant i\leqslant N}\frac{\left\|Y_{i,-1}^{(2)}\right\|}{T}\left\{\frac{1}{N}\sum_{i=1}^{N}\left\|\text{vec}(\tilde{\Lambda}_i^0)\right\|^2\right\}^{1/2}\left\{\frac{1}{N}\sum_{i=1}^{N}\left\|\text{vec}(\hat{\delta}_i)\right\|^2\right\}^{1/2}\\ &=O_p(\sqrt{T\iota_T}\eta_{NT})\end{aligned}$$

Analogously, we can show that $\frac{1}{\sqrt{T}}\left\|I_4\right\|=O_p(\sqrt{T\iota_T}\eta_{NT})$ and $\frac{1}{\sqrt{T}}\left\|I_5\right\|=O_p(\sqrt{T\iota_T}\eta_{NT})$. The rest of three terms are directly obtained from Bai and Ng (2002) such that $\frac{1}{\sqrt{T}}\left\|I_l\right\|=O_p(C_{NT}^{-1})$ for $l=6,7,8$. Thus, we directly have

$$\frac{1}{\sqrt{T}}\left\|\left(\hat{F}\otimes I_r\right)H^{-1}-\left(F^0\otimes I_r\right)\right\|=O_p(\sqrt{T\iota_T}\eta_{NT})+O_p(C_{NT}^{-1}).$$

■

Proof of Theorem 3.2. Let $\hat{Q}_{iNT}(b_i)=\frac{1}{T^2}\text{vec}\left(\Delta\hat{Y}_i^+-Y_{i,-1}^{(1)}-Y_{i,-1}^{(2)}b_i\right)'\text{vec}\left(\Delta\hat{Y}_i^+-Y_{i,-1}^{(1)}-Y_{i,-1}^{(2)}b_i\right)$, $\hat{Q}_{iNT}^{\lambda,K}(b_i,B_k)=\hat{Q}_{iNT}(b_i)+\lambda\prod_{k=1}^{K}\left\|\text{vec}\left(b_i-B_k\right)\right\|$, and $\hat{Q}_{NT}^{\lambda,K}(b,B)=\frac{1}{N}\sum_{i=1}^{N}\hat{Q}_{iNT}^{\lambda,K}(b_i,B_k)$. Note that

$$\begin{aligned}&\hat{Q}_{iNT}(b_i)-\hat{Q}_{iNT}(b_i^0)\\ =&\frac{1}{T^2}\text{vec}(\delta_i)'\left(I_r\otimes Y_{i,-1}^{(2)\prime}Y_{i,-1}^{(2)}\right)\text{vec}(\delta_i)-\frac{2}{T^2}\text{vec}(\delta_i)'\text{vec}\left(Y_{i,-1}^{(2)\prime}\left(\left(F^0\Lambda_i^0+\varepsilon_i^*\right)\tilde{\gamma}_i+\Delta Y_i\left(\hat{\gamma}_i-\tilde{\gamma}_i\right)\right)\right)\end{aligned} \tag{A.1}$$

where $\varepsilon_i^*=(\varepsilon_{iT}^{*\prime},...,\varepsilon_{i1}^{*\prime})'$ and $\varepsilon_{it}^*=\varepsilon_{it}-\Sigma_{uv,i}\Sigma_{vv,i}^{-1}v_{i,t-1}$. By the triangle, reverse triangle inequalities Su et al. (2016) ((A.6), p.2246), we have

$$\left|\prod_{k=1}^{K}\left\|\text{vec}\left(\hat{b}_i-B_k\right)\right\|-\prod_{k=1}^{K}\left\|\text{vec}\left(b_i^0-B_k\right)\right\|\right|\leqslant c_{i,NT}(\hat{\mathbf{B}})\left\|\text{vec}\left(\delta_i\right)\right\| \tag{A.2}$$

where $c_{i,NT}(\hat{\mathbf{B}})=\prod_{k=1}^{K-1}\|\text{vec}\left(\hat{b}_i-B_k\right)\|+\prod_{k=1}^{K-2}\|\text{vec}\left(\hat{b}_i-B_k\right)\|\|\text{vec}(b_i^0-B_K)\|+$ ...$+\prod_{k=2}^{K}\|\text{vec}(b_i^0-B_k)\|=O_p(1)$. Based on the fact that $\hat{Q}_{iNT}^{\lambda,K}(\hat{b}_i,\hat{B}_k)\leqslant\hat{Q}_{iNT}^{\lambda,K}(b_i^0,B_k^0)$, we have the basic inequality

$$\begin{aligned}&\left\|\text{vec}\left(\hat{\delta}_i\right)'\left(I_r\otimes\frac{2}{T^2}\sum_{t=1}^{T}y_{i,t-1}^{(2)}y_{i,t-1}^{(2)\prime}\right)\text{vec}\left(\hat{\delta}_i\right)\right\|\\ \leqslant&2\left\|\text{vec}\left(\hat{\delta}_i\right)\right\|\left(\left\|\text{vec}\left(\frac{1}{T^2}\sum_{t=1}^{T}y_{i,t-1}^{(2)}\left(\left(\Lambda_i^{0\prime}F_t^0+\varepsilon_{it}^*\right)'\tilde{\gamma}_i+\Delta y_{it}'(\hat{\gamma}_i-\tilde{\gamma}_i)\right)\right)\right\|+\lambda c_{i,NT}(\hat{\mathbf{B}})\right).\end{aligned}$$

Note that

$$\left\|\text{vec}\left(\hat{\delta}_i\right)'\left(I_r\otimes\frac{2}{T^2}\sum_{t=1}^{T} y_{i,t-1}^{(2)}y_{i,t-1}^{(2)\prime}\right)vec\left(\hat{\delta}_i\right)\right\| \geqslant \left\|\text{vec}\left(\hat{\delta}_i\right)\right\|^2\mu_{\min}\left(\frac{1}{T^2}\sum_{t=1}^{T} y_{i,t-1}^{(2)}y_{i,t-1}^{(2)\prime}\right).$$

Define $\rho_{i,\min}=\mu_{\min}\left(\frac{1}{T^2}\sum_{t=1}^{T} y_{i,t-1}^{(2)}y_{i,t-1}^{(2)\prime}\right)$ is bounded away from zero in probability by Lemma B.1 (i) and Lemma B.2(ii). Then by Assumption 3.2 and Lemma B.2(ii),

$$\begin{aligned}\left\|\text{vec}\left(\hat{\delta}_i\right)\right\| \leqslant&\rho_{i,\min}^{-1}\left(2\left\|\text{vec}\left(\frac{1}{T^2}\sum_{t=1}^{T} y_{i,t-1}^{(2)}\left(\left(\Lambda_i^{0\prime}F_t^0+\varepsilon_{it}^*\right)'\tilde{\gamma}_i+\Delta y_{it}'(\hat{\gamma}_i-\tilde{\gamma}_i)\right)\right)\right\|+\lambda c_{i,NT}(\hat{\mathbf{B}})\right)\\ =&O_p(T^{-1}+\lambda),\end{aligned}\tag{A.3}$$

since $\frac{1}{T^2}\sum_{t=1}^{T} y_{i,t-1}^{(2)}\left(\left(\Lambda_i^{0\prime}F_t^0+\varepsilon_{it}^*\right)'\tilde{\gamma}_i\right)=O_p(T^{-1})$ by Lemma B.1(iii) and $\hat{\gamma}_i-\tilde{\gamma}_i=O_p(T^{-1/2})$ by Theorem 3.1(i).

(ii) By the Minkowski's inequality, as $(N,T)\to\infty$ we have

$$\begin{aligned}c_{i,NT}(\hat{\mathbf{B}})\leqslant&\prod_{k=1}^{K-1}\left\{\|\text{vec}(\hat{b}_i-b_i^0)\|+\|\text{vec}(b_i^0-B_k)\|\right\}\\ &+\prod_{k=1}^{K-2}\left\{\|\text{vec}(\hat{b}_i-b_i^0)\|+\|\text{vec}(b_i^0-B_k)\|\right\}\|\text{vec}(b_i^0-B_K)\|+...+\prod_{k=2}^{K}\|\text{vec}(b_i^0-B_k)\|\\ =&\sum_{s=0}^{K-1}\|\text{vec}(\hat{b}_i-b_i^0)\|^s\prod_{k=1}^{s}a_{ks}\|\text{vec}(b_i^0-B_k)\|^{K-1-s}\\ \leqslant&C_{K,NT}(\hat{\mathbf{B}})\sum_{s=0}^{K-1}\|\text{vec}(\hat{b}_i-b_i^0)\|^s\leqslant C_{K,NT}(B)(1+2\|\text{vec}(\hat{\delta}_i)\|)\end{aligned}\tag{A.4}$$

where a_{ks}'s are finite integers and

$C_{K,NT}(\hat{\mathbf{B}})=\max_{1\leqslant i\leqslant N}\max_{1\leqslant s\leqslant k\leqslant K-1}\prod_{k=1}^{s}a_{ks}\left\|\text{vec}(B_l^0-B_k^0)\right\|^{K-1-s}=O(1)$ by Assumption 3.3(ii) as K is finite. Let $\hat{C}_K=C_{K,NT}(\hat{\mathbf{B}})$. Combining (A.3) and (A.4) yields

$$\left\|\text{vec}(\hat{\delta}_i)\right\|\leqslant\left(\frac{\rho_{i,\min}^{-1}}{1-c_{NT}}\right)\left\{2\left\|\text{vec}\left(\frac{1}{T^2}\sum_{t=1}^{T} y_{i,t-1}^{(2)}\left(\left(\Lambda_i^{0\prime}F_t^0+\varepsilon_{it}^*\right)'\tilde{\gamma}_i+\Delta y_{it}'(\hat{\gamma}_i-\tilde{\gamma}_i)\right)\right)\right\|+\lambda\hat{C}_K\right\}$$

where $c_{NT}=2\lambda\hat{C}_K\max_{1\leqslant i\leqslant N}\rho_{i,\min}^{-1}=O_p(\lambda\iota_T)=o_p(1)$ by Assumption 3.3(iii) and

$\rho_{\min} = \min_{1\leqslant i\leqslant N} \rho_{i,\min} = O_p(\iota_T)$ by Lemma B.2(ii). Then we have

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\left\|\text{vec}(\hat{\delta}_i)\right\|^2 \\
&\leqslant \left(\frac{\rho_{\min}^{-1}}{1-c_{NT}}\right)^2 \frac{1}{N}\sum_{i=1}^{N}\left(2\left\|\text{vec}\left(\frac{1}{T^2}\sum_{t=1}^{T} y_{i,t-1}^{(2)}\left(\left(\Lambda_i^{0\prime}F_t^0+\varepsilon_{it}^*\right)'\tilde{\gamma}_i+\Delta y_{it}'(\hat{\gamma}_i-\tilde{\gamma}_i)\right)\right)\right\|+\lambda\hat{C}_K\right)^2 \\
&= O_p\left(\iota_T^2(T^{-2}+\lambda^2)\right). \qquad\qquad (A.5)
\end{aligned}$$

Now we show that $\frac{1}{N}\sum_{i=1}^{N}\left\|\text{vec}(\hat{\delta}_i)\right\|^2 = O_p(\iota_T^2 T^{-2})$. Let vec$(\hat{b}_i)$ =vec$(b_i^0)+\iota_T T^{-1}\nu_i$, where $\nu = (\nu_1, ..., \nu_N)$ is a $(J-r)r\times N$ matrix. We want to show that for any given $\epsilon^* > 0$, there exists a large constant $L = L(\epsilon^*)$ such that, for sufficiently large N and T we have

$$P\left\{\inf_{\frac{1}{N}\sum_{i=1}^{N}\|\nu_i\|^2=L} Q_{NT}^{\lambda,K}(\text{vec}(b^0)+\iota_T T^{-1}\nu,\hat{B}) > Q_{NT}^{\lambda,K}(\text{vec}(b^0),B^0)\right\} \geqslant 1-\epsilon^*.$$

This implies that w.p.a.1 there is a local minimum $\{\hat{\mathbf{b}},\hat{\mathbf{B}}\}$ such that $\frac{1}{N}\sum_{i=1}^{N}\|\text{vec}(\hat{\delta}_i)\|^2 = O_p(\iota_T^2T^{-2})$ regardless of the property of $\hat{\mathbf{B}}$. By (A.1) and the Cauchy-Schwarz inequality, with probability approach to $1-o(N^{-1})$ we have

$$\begin{aligned}
&T^2\left[Q_{NT}^{\lambda,K}(\text{vec}(b^0)+\iota_T T^{-1}\nu,\hat{B}) - Q_{NT}^{\lambda,K}(\text{vec}(b^0),B^0)\right] \\
=&\frac{1}{N}\sum_{i=1}^{N}\iota_T^2\nu_i'\left(\frac{1}{T^2}Y_{i,-1}^{(2)\prime}Y_{i,-1}^{(2)}\otimes I_r\right)\nu_i - \frac{2T}{N}\sum_{i=1}^{N}\iota_T\nu_i'\text{vec}\left(\frac{1}{T^2}Y_{i,-1}^{(2)\prime}\left(F^0\Lambda_i^0+\varepsilon_i^*\right)\tilde{\gamma}_i\right) \\
&-\frac{2T}{N}\sum_{i=1}^{N}\iota_T\nu_i'\text{vec}\left(\frac{1}{T^2}Y_{i,-1}^{(2)\prime}\Delta Y_i\left(\hat{\gamma}_i-\tilde{\gamma}_i\right)\right)+\frac{\lambda T^2}{N}\sum_{i=1}^{N}\prod_{k=1}^{K}\left\|vec\left(b_i^0-\hat{B}_k\right)+\iota_T T^{-1}\nu_i\right\| \\
\geqslant&\iota_T\left(\frac{1}{N}\sum_{i=1}^{N}\|\nu_i\|^2\right)\min_{1\leqslant i\leqslant N}\mu_{\min}\left(\frac{\iota_T Y_{i,-1}^{(2)\prime}Y_{i,-1}^{(2)}}{T^2}\right) \\
&-\iota_T\left(\frac{1}{N}\sum_{i=1}^{N}\|\nu_i\|^2\right)^{1/2}\left(\frac{T^2}{N}\sum_{i=1}^{N}\left\|\text{vec}\left(\frac{1}{T^2}Y_{i,-1}^{(2)\prime}\left(F^0\Lambda_i^0+\varepsilon_i^*\right)\tilde{\gamma}_i\right)\right\|^2\right)^{1/2} \\
&-\iota_T\left(\frac{1}{N}\sum_{i=1}^{N}\|\nu_i\|^2\right)^{1/2}\left(\frac{T^2}{N}\sum_{i=1}^{N}\left\|\text{vec}\left(\frac{1}{T^2}Y_{i,-1}^{(2)\prime}\Delta Y_i\left(\hat{\gamma}_i-\tilde{\gamma}_i\right)\right)\right\|^2\right)^{1/2} \\
\equiv&\iota_T(D_{1NT}-D_{2NT}-D_{3NT}), \text{ say.}
\end{aligned}$$

By Lemma B.1(iii) and Lemma B.2(ii), we have $\frac{T^2}{N}\sum_{i=1}^{N}\left\|\text{vec}\left(\frac{1}{T^2}Y_{i,-1}^{(2)\prime}\left(F^0\Lambda_i^0+\varepsilon_i^*\right)\tilde{\gamma}_i\right)\right\|^2 = O_p(1)$ and $\min_{1\leqslant i\leqslant N}\mu_{\min}\left(\frac{\iota_T Y_{i,-1}^{(2)\prime}Y_{i,-1}^{(2)}}{T^2}\right) = O_p(1)$. Note that $\frac{T^2}{N}\sum_{i=1}^{N}\left\|vec\left(\frac{1}{T^2}Y_{i,-1}^{(2)\prime}\Delta Y_i\left(\hat{\gamma}_i-\tilde{\gamma}_i\right)\right)\right\|^2 =$

$o_p(1)$ since $(\hat{\gamma}_i - \tilde{\gamma}_i) = O_p(T^{-1/2})$. Then we have $D_{3NT} = o_p(D_{2NT})$. So D_{1NT} dominates $(D_{2NT} + D_{3NT})$ for sufficiently large L. That is,

$T^2 \left[Q_{NT}^{\lambda,K}(\text{vec}(b^0) + \iota_T T^{-1}\nu, \hat{B}) - Q_{NT}^{\lambda,K}(\text{vec}(b^0), B^0) \right] > 0$ for sufficiently large L. Consequently, we must have $\frac{1}{N}\sum_{i=1}^N \|\text{vec}(\hat{\delta}_i)\|^2 = O_p(\iota_T^2 T^{-2})$.

(iii) Let $P_{NT}(\mathbf{b}, \mathbf{B}) = \frac{1}{N}\sum_{i=1}^N \prod_{k=1}^K \|\text{vec}(b_i - B_k)\|$. By (A.2) and (A.4), we have that as $(N,T) \to \infty$,

$$\begin{aligned} |P_{NT}(\hat{\mathbf{b}}, \mathbf{B}) - P_{NT}(\mathbf{b}^0, \mathbf{B})| &\leqslant C_{KNT}(\mathbf{B}) \frac{1}{N}\sum_{i=1}^N \|\text{vec}(\hat{\delta}_i)\| + 2C_{KNT}(\mathbf{B})\frac{1}{N}\sum_{i=1}^N \|\text{vec}(\hat{\delta}_i)\|^2 \\ &\leqslant C_{KNT}(\mathbf{B}) \left\{ \frac{1}{N}\sum_{i=1}^N \|\text{vec}(\hat{\delta}_i)\|^2 \right\}^{1/2} + O_p(\iota_T^2 T^{-2}) = O_p(\iota_T T^{-1}). \end{aligned} \tag{A.6}$$

By (A.6) and the fact that $P_{NT}(\mathbf{b}^0, \mathbf{B}^0) = 0$ and that $P_{NT}(\hat{\mathbf{b}}, \hat{\mathbf{B}}) - P_{NT}(\hat{\mathbf{b}}, \mathbf{B}^0) \leqslant 0$. we have

$$\begin{aligned} 0 &\geqslant P_{NT}(\hat{\mathbf{b}}, \hat{\mathbf{B}}) - P_{NT}(\hat{\mathbf{b}}, \mathbf{B}^0) = P_{NT}(\mathbf{b}^0, \hat{\mathbf{B}}) - P_{NT}(\mathbf{b}^0, \mathbf{B}^0) + O_p(\iota_T T^{-1}) \\ &= \frac{1}{N}\sum_{i=1}^N \prod_{k=1}^K \|\text{vec}(b_i^0 - \hat{B}_k)\| + O_p(\iota_T T^{-1}) \\ &= \frac{N_1}{N}\prod_{k=1}^K \|\text{vec}(\hat{B}_k - B_1^0)\| + ... + \frac{N_K}{N}\prod_{k=1}^K \|\text{vec}(\hat{B}_k - B_K^0)\| + O_p(\iota_T T^{-1}) \end{aligned} \tag{A.7}$$

By Assumption 3.3(i), $N_k/N \to \tau_k \in (0,1)$ for each $k = 1, ...K$. So (A.7) implies that $\prod_{k=1}^K \|\text{vec}(\hat{B}_k - B_l^0)\| = O_p(\iota_T T^{-1})$ for $l = 1, ...K$. It follows that $(\text{vec}(\hat{B}_1), ..., \text{vec}(\hat{B}_k)) - (\text{vec}(B_1^0), ..., \text{vec}(B_K^0)) = O_p(\iota_T T^{-1})$. ■

Lemma B.3 *Suppose that Assumptions 3.1-3.3 hold. Then*

(i) $P\left(\max_{1\leqslant i\leqslant N}\left\|\text{vec}\left(\frac{1}{T^2}\sum_{t=1}^T y_{i,t-1}^{(2)}(\Lambda_i^{0\prime}F_t^0 + \varepsilon_{it}^*)'\right)\right\| > c\psi_{NT}\right) = o(N^{-1})$.

(ii) $P\left(\max_{1\leqslant i\leqslant N}\left\|\text{vec}\left(\hat{b}_i - b_i^0\right)\right\| > c\Psi_{NT}\right) = o(N^{-1})$

where $\Psi_{NT} = (\iota_T\psi_{NT} + \lambda)(\ln T)^\epsilon$.

Proof of Theorem 3.3. We fix $k \in \{1, ..., K\}$. By the consistency of $\hat{B}_k$ and $\hat{b}_i$, we have $\text{vec}(\hat{b}_i - \hat{B}_k) \to \text{vec}(B_k^0 - B_l^0) \neq 0$ for all $i \in G_k^0$ and $k \neq l$. It follows that w.p.a.1 $\left\|\text{vec}(\hat{b}_i - \hat{B}_l)\right\| \neq 0$ for all $i \in G_k^0$ and $l \neq k$. And $\hat{c}_{ki} = \prod_{l=1,l\neq k}^K \left\|\text{vec}(\hat{b}_i - \hat{B}_k)\right\| \xrightarrow{p}$

$c_k^0 \equiv \prod_{l=1,l\neq k}^K \|\text{vec}(B_k^0 - B_l^0)\| \geqslant c_B^{k-1} > 0$ for $i \in G_k^0$ by Assumption 3.3(ii). Now suppose that $\left\|\text{vec}(\hat{b}_i - \hat{B}_k)\right\| \neq 0$ for some $i \in G_k^0$. Then the first-order condition (with respect to b_i) for the minimization problem in (2.11) yields that

$$\begin{aligned}
0_{(J-r)r\times 1} =& T\frac{\partial Q_{iNT}^{\lambda,K}(\hat{b}_i, \hat{B}_k)}{\partial b_i} \\
=& -\frac{2}{T}\left(I_r \otimes Y_{i,-1}^{(2)\prime}\right)\text{vec}\left(\Delta\hat{Y}_i^{+} - Y_{i,-1}^{(1)} - Y_{i,-1}^{(2)}\hat{b}_i\right) + T\lambda\sum_{j=1}^K \hat{e}_{ij}\prod_{k=1}^K\left\|vec(\hat{b}_i - \hat{B}_k)\right\| \\
=& -\text{vec}\left(\frac{2}{T}\sum_{t=1}^T y_{i,t-1}^{(2)}\left(\left(\Lambda_i^{0\prime}F_t^0 + \varepsilon_{it}^*\right)^{\prime}\tilde{\gamma}_i + \Delta y_{it}^{\prime}(\hat{\gamma}_i - \tilde{\gamma}_i)\right)\right) \\
&+ \left(I_r \otimes \frac{2}{T^2}\sum_{t=1}^T y_{i,t-1}^{(2)}y_{i,t-1}^{(2)\prime}\right)T\text{vec}\left(\hat{B}_k - B_k^0\right) \\
&+ \left[\left(I_r \otimes \frac{2}{T^2}\sum_{t=1}^T y_{i,t-1}^{(2)}y_{i,t-1}^{(2)\prime}\right) + \frac{\lambda\hat{c}_{ik}}{\left\|\text{vec}(\hat{b}_i - \hat{B}_k)\right\|}\right]T\text{vec}\left(\hat{b}_i - \hat{B}_k\right) \\
&+ T\lambda\sum_{j=1,j\neq k}^K \hat{e}_{ij}\prod_{l=1,l\neq j}^K\left\|\text{vec}(\hat{b}_i - \hat{B}_k)\right\| \equiv -\hat{A}_{1i} + \hat{A}_{2i} + \hat{A}_{3i} + \hat{A}_{4i}, \text{ say.}
\end{aligned}$$

where $\hat{e}_{ij} = \frac{\text{vec}(\hat{b}_i - \hat{B}_j)}{\left\|\text{vec}(\hat{b}_i - \hat{B}_j)\right\|}$ if $\left\|\text{vec}(\hat{b}_i - \hat{B}_j)\right\| \neq 0$, $\|\hat{e}_{ij}\| \leqslant 1$ if $\left\|\text{vec}(\hat{b}_i - \hat{B}_j)\right\| = 0$. Let c denote a generic constant that may vary across lines. By Theorem 3.3(ii) and Lemma B.3(i), we can readily show

$$P\left(\max_{i\in G_k^0}\left\|\text{vec}\left(\hat{B}_k - B_k^0\right)\right\| \geqslant c\iota_T\psi_{NT}\right).$$

By Lemma B.2(i) and the matrix norm of Kronecker product, we have

$$\left\|\frac{1}{T^2}\sum_{t=1}^T y_{i,t-1}^{(2)}y_{i,t-1}^{(2)\prime} \otimes I_r\right\|_{sp} = \left\|\frac{1}{T^2}\sum_{t=1}^T y_{i,t-1}^{(2)}y_{i,t-1}^{(2)\prime}\right\|_{sp} \leqslant c\rho_{\max}\iota_T, \text{a.s.}$$

This, in conjunction with Lemma B.3(ii), yields

$$P\left(\max_{i\in G_k^0}\left\|\hat{A}_{2i}\right\| \geqslant c\iota_T^2 T\psi_{NT}\right) = o(N^{-1}) \text{ and } P\left(\max_{i\in G_k^0}\left\|\hat{A}_{4i}\right\| \geqslant cT\lambda\Psi_{NT}\right) = o(N^{-1}).$$

By lemma B.3(i), we directly have that $P\left(\max_{i\in G_k^0}\left\|\hat{A}_{1i}\right\| \geqslant c\iota_T T\psi_{NT}\right) = o(N^{-1})$. For

$\hat{A}_{3i}$,

$$\begin{aligned}
&\text{vec}\left(\hat{b}_i - \hat{B}_k\right)' \hat{A}_{3i} \\
=&T\text{vec}\left(\hat{b}_i - \hat{B}_k\right)' \left[\left(\frac{2}{T^2}\sum_{t=1}^{T} y_{i,t-1}^{(2)} y_{i,t-1}^{(2)\prime} \otimes I_r\right) + \frac{\lambda \hat{c}_{ik}}{\left\|\text{vec}(\hat{b}_i - \hat{B}_k)\right\|}\right] \text{vec}\left(\hat{b}_i - \hat{B}_k\right)' \\
\geqslant&\left(\frac{2}{T^2}\sum_{t=1}^{T} y_{i,t-1}^{(2)} y_{i,t-1}^{(2)\prime} \otimes I_r\right) \left\|\text{vec}(\hat{b}_i - \hat{B}_k)\right\|^2 + \lambda T \hat{c}_{ik} \left\|\text{vec}(\hat{b}_i - \hat{B}_k)\right\| \\
\geqslant&\lambda T \hat{c}_{ik} \left\|\text{vec}(\hat{b}_i - \hat{B}_k)\right\|
\end{aligned}$$

Combing above results together, it follows that $P(\Xi_{kNT}) = 1 - o(N^{-1})$, where

$$\Xi_{kNT} = \left\{\max_{i \in G_k^0} \|A_{2i}\| < c\iota_T^2 T \psi_{NT}\right\} \cap \left\{\max_{i \in G_k^0} \|A_{4i}\| < cT\lambda\Psi_{NT}\right\} \cap \left\{\max_{i \in G_k^0} \left|\hat{c}_{ki} - c_k^0\right| < c_k^0/2\right\}$$

Then conditional on Ξ_{kNT}, we have that uniformly in $i \in G_k^0$,

$$\begin{aligned}
&\left\|\text{vec}\left(\hat{b}_i - \hat{B}_k\right)' (\hat{A}_{2i} + \hat{A}_{3i} + \hat{A}_{4i})\right\| \\
\geqslant&\left\|\text{vec}\left(\hat{b}_i - \hat{B}_k\right)' \hat{A}_{3i}\right\| - \left\|\text{vec}\left(\hat{b}_i - \hat{B}_k\right)' (\hat{A}_{2i} + \hat{A}_{4i})\right\| \\
\geqslant&c\left\{\lambda T c_k^0/2 - \iota_T^2 T \psi_{NT} - T\lambda\Psi_{NT}\right\} \left\|vec\left(\hat{b}_i - \hat{B}_k\right)\right\| \\
\geqslant&c\lambda T c_k^0 \left\|vec\left(\hat{b}_i - \hat{B}_k\right)\right\|/4, \text{for sufficient large } (N, T),
\end{aligned}$$

where the last inequality follows by the fact that $(\iota_T^2 T\psi_{NT} - T\lambda\Psi_{NT}) = o(\lambda T)$ by Assumption 3.3(iv). It follows that

$$\begin{aligned}
P(\hat{E}_{kNT,i}) = P(i \notin \hat{G}_k | i \in G_k^0) &= P(\hat{A}_{1i} = \hat{A}_{2i} + \hat{A}_{3i} + \hat{A}_{4i}) \\
&\leqslant P\left(\left\|\text{vec}\left(\hat{b}_i - \hat{B}_k\right)' \hat{A}_{1i}\right\| \geqslant \left\|\text{vec}\left(\hat{b}_i - \hat{B}_k\right)' \left[\hat{A}_{3i} - (\hat{A}_{2i} + \hat{A}_{4i})\right]\right\|\right) \\
&\leqslant P\left(\left\|\hat{A}_{1i}\right\| \geqslant cT\lambda c_k^0/4, \Xi_{kNT}\right) + o(N^{-1}) \to 0, \text{ as } (N, T) \to \infty,
\end{aligned}$$

where the last inequality follows because that $T\lambda c_k^0 \gg \iota_T T \psi_{NT}$ and that $T\lambda c_k^0 \to \infty$ by Assumption 3.3(iv). Consequently, we can conclude that w.p.a.1 $\hat{b}_i - \hat{B}_k$ must be in position where $\|\text{vec}(b_i - B_k)\|$ is not differentiable with respect to b_i for any $i \in G_k^0$. That is $P\left(\left\|\text{vec}(\hat{b}_i - \hat{B}_k)\right\| = 0 | i \in G_k^0\right) = 1 - o(N^{-1})$ as $(N, T) \to \infty$.

For uniform consistency, we have that

$$P\left(\cup_{k=1}^{K}\hat{E}_{kNT}\right) \leqslant \sum_{k=1}^{K} P\left(\hat{E}_{kNT}\right) \leqslant \sum_{k=1}^{K}\sum_{i\in G_k^0} P\left(\hat{E}_{kNT,i}\right)$$
$$\leqslant N \max_{1\leqslant i\leqslant N} P\left(\left\|\hat{A}_{1i}\right\| \geqslant cT\lambda c_k^0/4\right) + o(1) \to 0 \text{ as } (N,T) \to \infty,$$

where the last inequality follows by Lemma B.3(i). This completes the proof of (i). Then the proof of (ii) directly follows SSP and thus omitted. ■

Lemma B.4 *Suppose that Assumptions 3.1-3.3 hold. Then*

(i) $Q_{k,NT} \xrightarrow{p} \mathbb{Q}_k$*,*

(ii) $V_{k,NT} - B_{k,NT} \Rightarrow N(0, \mathbb{V}_k)$*,*

where $\mathbb{Q}_k = \lim_{N\to\infty} \frac{1}{N_k}\sum_{i\in G_k^0} C_{1i}E_{\mathcal{C}}\left(\int B_{ui}B_{ui}'\right)C_{1i}'$*,* $\mathbb{V}_k = \lim_{N\to\infty}\mathit{Var}_{\mathcal{C}}\left(V_{kNT}\right) = \Omega_k$*, and* $C_{1i} = b_{i,\perp}^0\left(\alpha_{i,\perp}^{0\prime}\beta_{i,\perp}^0\right)^{-1}\alpha_{i,\perp}^{0\prime}$*.*

Proof of Theorem 3.4 A necessary and sufficient conditions for $\hat{b}_i$ and $\hat{B}_k$ to minimize the objective function in (2.11) is that for each $i = 1, ..., N$ (resp. $k = 1, ..., K$), $0_{(J-r)r\times 1}$ belongs to the sub-differential of $Q_{NT}^{K,\lambda}(b, B)$ with respect to b_i (resp. B_k) evaluated at $\hat{b}_i$ and $\hat{B}_k$. The first order conditions are

$$0_{(J-r)r\times 1} = -\frac{2}{NT^2}\left(I_r \otimes Y_{i,-1}^{(2)\prime}\right)\text{vec}\left(\Delta\hat{Y}_i^{+} - Y_{i,-1}^{(1)} - Y_{i,-1}^{(2)}\hat{b}_i\right) + \frac{\lambda}{N}\sum_{j=1}^{K}\hat{e}_{ij}\prod_{l=1,l\neq j}^{K}\left\|\text{vec}(\hat{b}_i - \hat{B}_l)\right\|,$$
$$0_{(J-r)r\times 1} = \frac{\lambda}{N}\sum_{i=1}^{N}\hat{e}_{ik}\prod_{l=1,l\neq k}^{K}\left\|\text{vec}(\hat{b}_i - \hat{B}_l)\right\|,$$

where $\hat{e}_{ij}$ is defined in Theorem 3.3. First, we observe that $\left\|\text{vec}(\hat{b}_i - \hat{B}_k)\right\| = 0$ for any $i \in \hat{G}_k$ by the definition of $\hat{G}_k$.And $\hat{b}_i - \hat{B}_j \xrightarrow{p} \hat{B}_k - \hat{B}_l \neq 0$ for any $i \in \hat{G}_k$ and $l \neq k$ by Assumption 3.3(ii). It follows that $\|\hat{e}_{ik}\| \leqslant 1$ for any $i \in \hat{G}_k$ and $\hat{e}_{ij} = \frac{\text{vec}(\hat{b}_i - \hat{B}_j)}{\left\|\text{vec}(\hat{b}_i - \hat{B}_j)\right\|} = \frac{\text{vec}(\hat{B}_k - \hat{B}_l)}{\left\|\text{vec}(\hat{B}_k - \hat{B}_l)\right\|}$ w.p.a.1 for any $i \in \hat{G}_k$ and $l \neq k$. This further implies that w.p.a.1

$$\sum_{i\in\hat{G}_k}\sum_{j=1,j\neq k}^{K}\hat{e}_{ij}\prod_{l=1,l\neq j}^{K}\left\|\text{vec}(\hat{b}_i - \hat{B}_l)\right\| = \sum_{i\in\hat{G}_k}\sum_{j=1,j\neq k}^{K}\frac{\text{vec}(\hat{B}_k - \hat{B}_j)}{\left\|\text{vec}(\hat{B}_k - \hat{B}_j)\right\|}\prod_{l=1,l\neq j}^{K}\left\|\text{vec}(\hat{B}_k - \hat{B}_l)\right\| = 0_{(J-r)r\times 1},$$

and

$$
\begin{aligned}
&0_{(J-r)r\times 1}\\
&=\sum_{i=1}^{N}\hat{e}_{ik}\prod_{l=1,l\neq k}^{K}\left\|\text{vec}(\hat{b}_i-\hat{B}_l)\right\|\\
&=\sum_{i\in\hat{G}_k}\hat{e}_{ik}\prod_{l=1,l\neq k}^{K}\left\|\text{vec}(\hat{B}_k-\hat{B}_l)\right\|+\sum_{i\in\hat{G}_0}\hat{e}_{ik}\prod_{l=1,l\neq k}^{K}\left\|\text{vec}(\hat{b}_i-\hat{B}_l)\right\|+\sum_{j=1,j\neq k}^{K}\sum_{i\in\hat{G}_j}\hat{e}_{ik}\prod_{l=1,l\neq k}^{K}\left\|\text{vec}(\hat{B}_j-\hat{B}_l)\right\|\\
&=\sum_{i\in\hat{G}_k}\hat{e}_{ik}\prod_{l=1,l\neq k}^{K}\left\|\text{vec}(\hat{B}_k-\hat{B}_l)\right\|+\sum_{i\in\hat{G}_0}\hat{e}_{ik}\prod_{l=1,l\neq k}^{K}\left\|\text{vec}(\hat{b}_i-\hat{B}_l)\right\|.
\end{aligned}
$$

Combining above results, we have

$$
\frac{2}{N_kT^2}\sum_{i\in\hat{G}_k}\left(I_r\otimes Y_{i,-1}^{(2)\prime}\right)\text{vec}\left(\Delta\hat{Y}_i^{+}-Y_{i,-1}^{(1)}-Y_{i,-1}^{(2)}\hat{B}_k\right)+\frac{\lambda}{N_k}\sum_{i\in\hat{G}_0}\hat{e}_{ik}\prod_{l=1,l\neq k}^{K}\left\|\text{vec}(\hat{b}_i-\hat{B}_l)\right\|=0_{(J-r)r\times 1}
$$

It yields that the C-Lasso estimators follow

$$
\begin{aligned}
\text{vec}(\hat{B}_k)=&\left(I_r\otimes\frac{1}{N_kT^2}\sum_{i\in\hat{G}_k}Y_{i,-1}^{(2)\prime}Y_{i,-1}^{(2)}\right)^{-1}\text{vec}\left(\frac{1}{N_kT^2}\sum_{i\in\hat{G}_k}Y_{i,-1}^{(2)\prime}\left(\Delta\hat{Y}_i^{+}-Y_{i,-1}^{(1)}\right)\right)\\
&+\left(I_r\otimes\frac{1}{N_kT^2}\sum_{i\in\hat{G}_k}Y_{i,-1}^{(2)\prime}Y_{i,-1}^{(2)}\right)^{-1}\frac{\lambda}{2N_k}\sum_{i\in\hat{G}_0}\hat{e}_{ik}\prod_{l=1,l\neq k}^{K}\left\|\text{vec}(\hat{b}_i-\hat{B}_l)\right\|\\
&\equiv vec(\hat{B}_k^{post})+\hat{\mathcal{R}}_{1Nk},\ \text{say.}
\end{aligned}
$$

where $\text{vec}(\hat{B}_k^{post})=\left(\frac{1}{N_kT^2}\sum_{i\in\hat{G}_k}Y_{i,-1}^{(2)\prime}Y_{i,-1}^{(2)}\otimes I_r\right)^{-1}vec\left(\frac{1}{N_kT^2}\sum_{i\in\hat{G}_k}Y_{i,-1}^{(2)\prime}\left(\Delta\hat{Y}_i^{+}-Y_{i,-1}^{(1)}\right)\right)$. By Theorem 3.3, we readily show that

$$
P\left(\sqrt{N_k}T\|\hat{\mathcal{R}}_{1Nk}\|\geqslant\epsilon\right)\leqslant\sum_{k=1}^{K}\sum_{i\in G_k^0}P(i\in\hat{G}_0|i\in G_k^0)\leqslant\sum_{k=1}^{K}\sum_{i\ni G_k^0}P(i\notin\hat{G}_k|i\in G_k^0)=o(1)
$$

It follows that

$$
\sqrt{N_k}T\text{vec}(\hat{B}_k-B_k^0)=\sqrt{N_k}T\text{vec}(\hat{B}_k^{post}-B_k^0)+o_p(1)
$$

This implies that $\hat{B}_k$ is asymptotically equivalent to its post-Lasso estimator $\hat{B}_k^{post}$. Then we formally study the asymptotic distribution of $\hat{B}_k^{post}$ in the proof of Theorem (ii) below.

(ii) Note that

$$
\sqrt{N_k}T\text{vec}\left(\hat{B}_k^{post}-B_k^0\right)=\hat{Q}_k^{-1}\left(\hat{V}_k+\hat{\mathcal{R}}_{2Nk}+\hat{\mathcal{R}}_{3Nk}\right)
$$

where

$$\begin{aligned}\hat{Q}_k =& \frac{1}{N_k T^2} \sum_{i \in \hat{G}_k} \left(I_r \otimes Y_{i,-1}^{(2)\prime} Y_{i,-1}^{(2)} \right), \; \hat{V}_k = \frac{1}{\sqrt{N_k} T} \sum_{i \in \hat{G}_k} \text{vec} \left(Y_{i,-1}^{(2)\prime} \left(F^0 \Lambda_i^0 + \varepsilon_i^* \right) \tilde{\gamma}_i \right), \\ \hat{\mathcal{R}}_{2Nk} =& \frac{1}{\sqrt{N_k} T} \sum_{i \in \hat{G}_k} \text{vec} \left(Y_{i,-1}^{(2)\prime} \Delta Y_i \left(\hat{\gamma}_i - \tilde{\gamma}_i \right) \right), \\ \hat{\mathcal{R}}_{3Nk} =& \frac{1}{N_k T^2} \sum_{i \in \hat{G}_k} \left(I_r \otimes Y_{i,-1}^{(2)\prime} Y_{i,-1}^{(2)} \right) \text{vec} \left(b_i^0 - B_k^0 \right).\end{aligned}$$

It follows that

$$\begin{aligned}\hat{Q}_k =& \frac{1}{N_k T^2} \sum_{i \in G_k^0} \left(I_r \otimes Y_{i,-1}^{(2)\prime} Y_{i,-1}^{(2)} \right) + \frac{1}{N_k T^2} \sum_{i \in \hat{G}_k \backslash G_k^0} \left(I_r \otimes Y_{i,-1}^{(2)\prime} Y_{i,-1}^{(2)} \right) \\ & - \frac{1}{N_k T^2} \sum_{i \in G_k^0 \backslash \hat{G}_k} \left(I_r \otimes Y_{i,-1}^{(2)\prime} Y_{i,-1}^{(2)} \right) \\ \equiv & Q_{k,NT} + \hat{Q}_{1k} - \hat{Q}_{2k}, \text{ say.}\end{aligned}$$

by the fact that $1\{i \in \hat{G}_k\} = 1\{i \in G_k^0\} + 1\{i \in \hat{G}_k \backslash G_k^0\} - 1\{i \in G_k^0 \backslash \hat{G}_k\}$. By Theorem 3.3, we can show that $P\left(\|\hat{Q}_{2k}\| \geqslant \epsilon N^{-1/2} T^{-1} \right) \leqslant P(\hat{E}_{kNT}) = o(N^{-1})$ and $P\left(\|\hat{Q}_{1k}\| \geqslant \epsilon N^{-1/2} T^{-1} \right) \leqslant P(\hat{F}_{kNT}) = o(N)$ for any $\epsilon > 0$. It follows that $\hat{Q}_k = Q_{k,NT} + o_p(N^{-1/2} T^{-1})$. Analogously, we can show that $P\left(\sqrt{N} T \|\hat{\mathcal{R}}_{jNk}\| \geqslant \epsilon \right) = o(1)$ for $j = 1, 2$ and $\hat{V}_k = V_{k,NT} + o_p(N^{-1/2} T^{-1})$ by Theorem 3.3. Thus we have

$$\sqrt{N_k} T \text{vec}(\hat{B}_k^{post} - B_k^0) = Q_{k,NT}^{-1} V_{k,NT} + o_p(1)$$

where $V_{k,NT} = \frac{1}{\sqrt{N_k} T} \sum_{i \in G_k^0} \text{vec}\left(Y_{i,-1}^{(2)\prime} \left(F^0 \Lambda_i^0 + \varepsilon_i^* \right) \tilde{\gamma}_i \right)$. By Lemma B.4(i)-(ii), it follows that $V_{k,NT} - B_{k,NT} \Rightarrow N(0, \mathbb{V}_k)$ and $Q_{k,NT} \xrightarrow{p} \mathbb{Q}_k$. By Slutsky's theorem, we directly have $\sqrt{N_k} T \text{vec}(\hat{B}_k^{post} - B_k^0) - \mathbb{Q}_k^{-1} B_{k,NT} \Rightarrow N(0, \mathbb{Q}_k^{-1} \mathbb{V}_k \mathbb{Q}_k^{-1\prime})$.■