11-2016

Identifying latent structures in panel data

Liangjun SU
Singapore Management University, ljsu@smu.edu.sg

Zhentao SHI
Chinese University of Hong Kong

Peter C. B. PHILLIPS
Singapore Management University, peterphillips@smu.edu.sg

Published in Econmetrica, November 2016, vol. 84 (6), 2215-2264.
http://doi.org/10.3982/ECTA12560

Identifying Latent Structures in Panel Data*

Liangjun Su[a], Zhentao Shi[b], and Peter C. B. Phillips[c]
[a] *School of Economics, Singapore Management University*
[b] *Department of Economics, Yale University*
[c] *Yale University, University of Auckland*
University of Southampton & Singapore Management University

June 10, 2014

Abstract

This paper provides a novel mechanism for identifying and estimating latent group structures in panel data using penalized regression techniques. We focus on linear models where the slope parameters are heterogeneous across groups but homogenous within a group and the group membership is unknown. Two approaches are considered – penalized least squares (PLS) for models without endogenous regressors, and penalized GMM (PGMM) for models with endogeneity. In both cases we develop a new variant of Lasso called classifier-Lasso (C-Lasso) that serves to shrink individual coefficients to the unknown group-specific coefficients. C-Lasso achieves simultaneous classification and consistent estimation in a single step and the classification exhibits the desirable property of uniform consistency. For PLS estimation C-Lasso also achieves the oracle property so that group-specific parameter estimators are asymptotically equivalent to infeasible estimators that use individual group identity information. For PGMM estimation the oracle property of C-Lasso is preserved in some special cases. Simulations demonstrate good finite-sample performance of the approach both in classification and estimation. An empirical application investigating the determinants of cross-country savings rates finds two latent groups among 56 countries, providing empirical confirmation that higher savings rates go in hand with higher income growth.

JEL Classification: C33, C36, C38, C51

*The authors thank Stéphane Bonhomme, Xiaohong Chen, and Cheng Hsiao for discussions on the subject matter and comments on the paper. Su acknowledges support from the Singapore Ministry of Education for Academic Research Fund under grant number MOE2012-T2-2-021. Phillips acknowledges NSF support under Grant Nos. SES-0956687 and SES-1285258. Address Correspondence to: Liangjun Su, School of Economics, Singapore Management University, 90 Stamford Road, Singapore 178903; E-mail: ljsu@smu.edu.sg, Phone: +65 6828 0386.

Key Words: Classification; Cluster analysis; Convergence club; Dynamic panel; Group Lasso; High dimensionality; Oracle property; Panel structure model; Parameter heterogeneity; Penalized least squares; Penalized GMM

1 Introduction

Panel data models are widely used in empirical analysis in many disciplines across the social and medical sciences. The capacity to store and retrieve vast electronic datasets on individual behavior over time has made these models a particularly prominent research vehicle in economics and finance. Such data usually cover individual units sampled from different backgrounds and with different individual characteristics so that an abiding feature of the data is its heterogeneity, much of which is simply unobserved. Neglecting latent heterogeneity in the data can lead to many difficulties, including inconsistent estimation and misleading inference, as is well explained in the literature (e.g., Hsiao, 2003, Chapter 6). It is therefore widely acknowledged that an important feature of good empirical modeling is to control for heterogeneity in the data as well as for potential heterogeneity in the response mechanisms that figure within the model. Since heterogeneity is a latent feature of the data and its extent is unknown a priori, respecting the potential influence of heterogeneity on model specification is a serious challenge in empirical research. Even in the simplest linear panel data models the challenge is manifest and clearly stated: do we allow for heterogeneous slope coefficients in regression as well as heterogeneous error variances?

While it may be clearly stated, this challenge to the empirical researcher is by no means easily addressed. While allowing for cross-sectional slope heterogeneity in regression may help to avert misspecification bias, it also sacrifices the power of cross section averaging in the estimation of response patterns that may be common across individuals, or more subtly, certain groups of individuals in the panel. In the absence of prior information on such grouping and with data where every new individual to the panel may bring new idiosyncratic elements to be explained, the challenge is demanding and almost universally relevant.

Traditional panel data models frequently deal with this challenge by avoidance. Complete slope homogeneity is assumed for certain specified common parameters in the panel. Under this assumption, the regression parameters are the same across individuals and unobserved heterogeneity is modeled through individual-specific effects which are either fixed or random and (typically) enter the model additively. This approach is an exemplar of a convenient assumption that facilitates estimation and inference.

The cross section homogeneity assumption has been frequently questioned and rejected in empirical studies. The following is only a partial list of work where homogeneity has been found to fail. Burnside (1996) rejects slope homogeneity in the production function of US manufacturing firms; Hsiao and Tahmiscioglu (1997) find parameter heterogeneity in investment functions

using the U.S. firm level panel data; Lee, Pesaran, and Smith (1997) find that the convergence rates of per capita output to the steady state level are heterogeneous across countries; Durlauf, Kourtellos, and Minkin (2001) find substantial country-specific heterogeneity in the parameters in Solow growth model that is associated with differences in initial income; Phillips and Sul (2007a) provide a new approach to testing for economic growth convergence under heterogeneous technology and explore these differences in the Penn World Table; Browning and Carro (2007) present a selective overview on heterogeneity in microeconometric modelling and find that there is more heterogeneity than econometricians usually allow for; Browning and Carro (2010) document heterogeneity in a dynamic discrete choice panel data model for consumer milk-type choices where heterogeneity occurs in both the levels parameter and the state dependence parameter; Browning and Carro (2014) show that individual unemployment dynamics are heterogenous even within a homogeneous group of Danish workers in terms of their observed characteristics; Su and Chen (2013) reject the null of slope homogeneity in an economic growth model for OECD countries even after they control for unobserved heterogeneity through interactive fixed effects.

Despite general agreement that slope heterogeneity is endemic in empirical work with panels, few methods are available to allow for heterogeneity in the slope parameters when the extent of the heterogeneity is unknown. In the following discussion we group the methods that are available into two broad categories and consider the different approaches pursued within them. In the first category, complete slope heterogeneity is assumed and regression coefficients are taken as differing across individuals. Several approaches are adopted in the literature. Perhaps the most common method is to use a random coefficient structure in which the parameters are assumed to be independent draws from a common distribution – see Hsiao and Pesaran (2008) for an overview of the approach. The random coefficient model allows for estimation of the mean coefficient effect but is uninformative about responses at the disaggregate level, thereby missing what is often the object of interest. A second approach uses Bayesian methods to shrink the individual slope estimates towards the overall mean – see Maddala, Trost, Li, and Joutz (1997). This approach is based on the presumption that the slope parameters, while not precisely the same, are sufficiently similar to warrant shrinkage toward the mean – a presumption that may be questionable in some empirical applications. A third approach is to parameterize individual slope coefficients as a function of observed characteristics – see Durlauf, Kourtellos, and Minkin (2001) and Browning, Ejrnæs, and Alvarez (2010). Apparently, this approach depends crucially on the specification of the functional coefficient and is subject to potential misspecification problems. A fourth approach is to estimate the individual slope coefficients using heterogenous time series regressions for each individual, which is only feasible in systems where the time dimension T is large. Even in this case, there is a considerable debate on the options: whether to pool the data and obtain a single estimate for the whole sample, whether to estimate the equations separately for each individual, and whether to rely on the average response from individual time series regressions – see Pesaran

and Smith (1995), Baltagi and Griffin (1997), Hsiao, Pesaran, and Tahmiscioglu (1999), Pesaran, Shin and Smith (1999), and the survey by Baltagi, Bresson, and Pirotte (2008).

The second category takes a totally different viewpoint on the nature of the heterogeneity in panels. In place of complete slope homogeneity or heterogeneity an intermediate approach is adopted in which the panel structure models individuals as belonging to a number of homogeneous groups or clubs within a broadly heterogeneous population. In this framework, the regression parameters are the same within each group but differ across groups. Two essential questions remain: how to determine the unknown number of groups (dubbed convergence clubs in the economic growth literature); and how to identify the individuals belonging to each group. These are longstanding questions of statistical classification in panel data. No completely satisfactory solution has yet been found, although various approaches have been adopted in empirical research. For instance, Bester and Hansen (2013) consider a panel structure model where individuals are grouped according to some external classification, geographic location, or observable explanatory variables; Bai and Ando (2013) consider a multifactor asset-pricing model where there exist group-specific pervasive factors influencing a subset of assets and the group membership is assumed to be known. So the group structure is *completely known* to the researcher, an approach that is common in practical work because of its convenience. In the economic growth literature, for example, countries are often classified according to continental location or economic development levels, which both lead to determinate group structures. In spite of its convenience, this approach to panel inference is inevitably misleading when the number of groups and individual identities are incorrectly classified.

Several approaches have been proposed to determine an *unknown* group structure in modeling unobserved slope heterogeneity in panels. The first approach is to apply finite mixture models that do not assume a known group structure. For example, Sun (2005) considers a *parametric* finite mixture panel data model by employing a multinomial logistic regression to model membership probabilities. Sun's model comprises a heterogenous linear panel regression model that relates the response variable to explanatory variables and a logistic regression that identifies individual memberships. In a related thematic, Kasahara and Shimotsu (2009) and Browning and Carro (2011) study identification in discrete choice panel data models for a fixed number of groups using *nonparametric* discrete mixture distributions. The second approach is based on the K-means algorithm in statistical cluster analysis. Lin and Ng (2012) and Sarafidis and Weber (2011) propose to modify the K-means algorithm to perform conditional clustering to estimate linear panel structure models but no asymptotic properties of that procedure or the estimators are derived. Bonhomme and Manresa (2014) introduce time-varying grouped patterns of heterogeneity in linear panel data models, propose two classification algorithms that are also closely related to the K-means algorithm, and study the asymptotic properties of the resulting estimators. Ando and Bai (2013) consider SCAD estimation of panel data models with unobserved group factor

structures. Lin and Ng (2012), Bonhomme and Manresa (2014), and Ando and Bai (2013) all assume that N and T pass to infinity jointly. Lin and Ng (2012) propose another method to estimate a panel structure model by turning the problem of parameter heterogeneity into the estimation of a panel threshold model with an unknown threshold value and using the individual time series estimates of the parameters to form threshold variables. Phillips and Sul (2007) develop an algorithm for determining group clusters that relies on the estimation of evaporating trend functions to determine convergence clusters. Again, joint limits as $(N,T) \to \infty$ are used in the development of the asymptotic theory.

The present paper proposes a new method for econometric estimation and inference in panel models when the slope parameters are heterogenous across groups, individual group membership is unknown, and classification is to be determined empirically. Our modeling strategy therefore falls within the second category discussed above. It is an automated data-determined procedure and does not require the specification of any modeling mechanism for the unknown group structure. The approach we suggest involves a new variant of Lasso (Tibshirani, 1996) technology that is designed to classify parametric slope coefficients in a heterogeneous panel model into a group structure in which both the groups and the elements in the groups are data-determined. Like Lin and Ng (2012), Bonhomme and Manresa (2014) and Phillips and Sul (2007), we assume that $(N,T) \to \infty$ jointly (Phillips and Moon, 1999). But in our asymptotic theory T can pass to infinity at a very slow rate, even a slowly varying rate such as $O\left((\ln N)^{1+\epsilon}\right)$ for any $\epsilon > 0$ in the case of uniformly bounded regressors, thereby opening up empirical applications of the method to short wide panels. The methods proposed here have several novel aspects in relation to earlier research and they contribute to both the Lasso and econometric classification literatures in various ways, which we outline in the following paragraphs.

First, our approach is motivated by one of the key features of Lasso technology that enables the method to deliver simultaneous variable selection and estimation in a single step. This advantage is particularly useful when the set of unknown parameters is potentially very large but may also embody certain *sparse* features. In a typical panel model structure, the *effective* number of unknown slope parameters $\{\beta_i,\, i = 1, ..., N\}$ is not of order $O(N)$ as it would be if these parameters were all incidental, but rather of some order $O(K_0)$, where K_0 denotes the number of unknown groups within which the slope coefficients are homogeneous. Moreover, when the number of groups is finite, K_0 is fixed and so the order of unknown coefficients is then $O(1)$ as $(N,T) \to \infty$. Hence, in many empirical applications the set of unknown slope parameters in a panel structure model surely exhibits the desirable sparsity feature, making the use of Lasso technology highly appealing.

Second, the procedures developed in the present paper contribute to the fused Lasso literature in which sparsity arises because some parameters take the same value. The fused Lasso was proposed by Tibshirani, Saunders, Rosset, Zhu, and Knight (2005) and was designed for problems

with features that can be ordered in some meaningful way (e.g., in time series regression where the time periods have natural ordering). The method cannot be used to classify individuals into different groups because there is no natural ordering across individuals and so a different algorithm to locate common individuals is required. The present paper develops a *new* variant of the Lasso method that does not rely on the order of individuals in the data and which therefore contributes to the fused Lasso technology.

Third, standard Lasso technology involves an additive penalty term to the least-squares, GMM, or log-likelihood objective function and when multiple penalty terms are needed, they also enter the objective function *additively*. To achieve simultaneous group classification and estimation in a single step our variant of Lasso involves N *additive* penalty terms, each of which takes a *multiplicative* form as a product of K_0 penalty terms. To the best of our knowledge, this paper is the first to propose a mixed additive-multiplicative penalty form that can serve as an engine for simultaneous classification and estimation. The method works by using each of the K_0 penalty terms in the *multiplicative* expression to shrink the individual-level slope parameter vectors to a particular *unknown* group-level parameter vector, thereby producing a joint shrinkage process. This process is distinct from the prototypical Lasso method that shrinks an individual parameter to zero and the group Lasso method that shrinks a parameter vector to a vector of zeros (see Yuan and Lin, 2006). To emphasize its role as a classifier and for future reference, we describe our new Lasso method as the *classifier-Lasso* or *C-Lasso*.

Fourth, we develop a limit theory for the C-Lasso that demonstrates its capacity to achieve simultaneous classification and consistent estimation in a single step. As mentioned in the Abstract, the paper develops two classes of estimators for panel structure models – penalized least squares (PLS) and penalized GMM (PGMM). The former is applicable to panel models without endogenous regressors and with or without dynamic structures, while the latter is applicable to panel models with endogeneity or dynamic structures. In either case, we show uniform classification consistency in the sense that all individuals belonging to a certain group can be classified into the same group correctly uniformly over both individuals and group identities with probability approaching one (w.p.a.1). Conversely, all individuals that are classified into a certain group belong to the same group uniformly over both individuals and group identities w.p.a.1. Under some regularity conditions, such a uniform result allows us to establish an *oracle* property of the PLS estimator that it is asymptotically equivalent to the corresponding infeasible estimator of the group-specific parameter vector that is obtained by knowing all individual group identities. Note that traditional Lasso only possesses the selection consistency and oracle property under the so-called restrictive *irrepresentable condition*. This shortcoming of Lasso motivated Zou (2006) to propose the *adaptive Lasso* that possesses these attractive properties.[1] Unfortunately, our

[1] Other methods that possess the selection consistency and oracle property include the Bridge and SCAD (smoothly clipped absolute deviation) procedures; see Knight and Fu (2000) and Fan and Li (2001).

PGMM estimator generally does not have the oracle property despite the uniform selection consistency of the C-Lasso. The uniform classification consistency also allows us to develop a limit theory for post-C-Lasso estimators that are obtained by pooling all individuals in an estimated group to estimate the group-specific parameters.

Fifth, C-Lasso enables empirical researchers to study panel structures without *a priori* knowledge of the number of groups, without the need to specify any ancillary regression models to model individual group identities, and with no need to make any distributional assumptions. When the number K_0 of groups is unknown, a BIC-type information criterion is proposed to determine the number of groups and it is shown that this procedure selects the correct number of groups consistently. The same information criterion can also be used to determine a data-driven tuning parameter for the PLS or PGMM estimation.

Sixth, while the focus of the present paper is on linear panel data modeling, the methodology developed here can be extended to nonlinear models such as discrete choice models, to semiparametric and nonparametric models, to models where only a subset of parameters are allowed to be group-specific, and to models where one considers group-specific effects along the time dimension. Extension to panel data models with interactive-fixed effects is also possible and is presently under way.

We envisage a large number of potential empirical applications of the C-Lasso approach within economics and finance and more broadly across the social and business sciences. The following list provides three distinct areas of application in international macroeconomics, microeconometrics, and nonstationary panel econometrics.

1. Economic Growth Convergence: Much of the recent literature on economic growth addresses sources of possible heterogeneity, including the occurrence of multiple steady states and history-dependence in growth trajectories - see Deissenberg, Feichtinger, Semmler, and Wirl (2004) and Durlauf, Johnson, and Temple (2005) and Eberhardt and Teal (2011) for overviews of the relevant growth theory and empirics. Contingent upon historical conditions economic systems may converge towards distinct steady states, the empirical manifestation of which are the so-called convergence clubs that occur in cross-country growth studies. In an application to cross-country growth, Phillips and Sul (2007a) evaluated evidence in support of panel data growth clustering, locating three convergence clubs and one divergent group among 88 countries in the Penn World Tables in terms of real per capita GDP over the period 1960-1996. Their methodology involved a stepwise algorithm with multi-level decision making to isolate the convergence clubs. The panel structure framework suggested in the present paper is a natural setting to consider growth convergence and the C-Lasso procedure provides a one step classifier and estimation approach with no sequential decision making. The method can also be used to isolate convergence clubs and remaining divergent elements in the panel.

2. Subsample Studies of Stability: Much empirical research is concerned with studying

the stability of certain regression coefficients over subsamples of the data. In this work, the whole sample is split into multiple subsamples and regression relationships are checked for coefficient stability. The groupings may be arbitrarily selected or may be determined by covariates or thresholds, each of which may have a significant impact on the findings. For example, in order to test whether financing constraints affect investment decisions, Fazzari, Hubbard, and Petersen (1988) divided a sample of firms into multiple groups based on empirical proxies such as the dividend-income ratio. Similarly, in testing whether liquidity constraints affect consumption decisions in PSID data, Zeldes (1989) uses two different wealth-to-income ratios as prescribed variables to divide the sample into subsamples. Sample splitting techniques of this type are inevitably vulnerable to the choice of prescribed driver variables. The methodology of the present paper does not require driver variables or thresholds to determine regression stability.

3. Panel Unit Root Grouping: Several approaches are available for testing the presence of unit roots in panel data. Two popular tests in applications are the Levin, Lin, and Chu (2002) and Im, Pesaran, and Shin (2003) tests. Levin, Lin, and Chu (2002) devise an adjusted t-test for a unit root for various panel data models, assuming that all individuals (countries, regions, industries, etc.) have the same autoregressive (AR) coefficients while permitting individual specific effects as well as dynamic heterogeneity across individuals. Im, Pesaran, and Shin (2003) propose a test based on the average of the augmented Dickey-Fuller statistics computed for each individual series in heterogenous panels. Both tests rule out the possibility that some individual series have a unit root while others do not - precisely the empirical possibility that many argue is the most relevant in practical work (e.g., Maddala and Kim, 1998). Our methodology is designed to directly address this possibility and can be used to classify a subgroup of unit-root processes in the panel from a wider class of stationary and nonstationary processes.

The rest of the paper is organized as follows. We study the C-Lasso PLS estimation and inference of panel structure models in Section 2. PGMM estimation and inference is addressed in Section 3. Section 4 reports Monte Carlo simulation findings. We apply our method to study the determinants of cross-country savings rates in Section 5. Final remarks are contained in Section 6. Proofs of the main results in the body of the paper are given in Appendices A and B. The supplementary Appendices C and D provide primitive conditions for some high level conditions that are used in the body of the paper and bias correction for the C-Lasso estimates, respectively.

NOTATION. Throughout the paper we adopt the following notation. For an $m \times n$ real matrix A, we write the transpose A', the Frobenius norm $\|A\|$ $(\equiv [\mathrm{tr}\,(AA')]^{1/2})$, and the Moore-Penrose inverse as A^{+}. When A is symmetric, we use $\mu_{\max}(A)$ and $\mu_{\min}(A)$ to denote the largest and smallest eigenvalues, respectively. I_p and $\mathbf{0}_{p\times 1}$ denote the $p \times p$ identity matrix and $p \times 1$ vector of zeros. $\mathbf{1}\{\cdot\}$ denotes the indicator function and "p.d." abbreviates "positive definite". The operator $\xrightarrow{P}$ denotes convergence in probability, $\xrightarrow{D}$ convergence in distribution, and plim probability limit. We use $(N,T) \to \infty$ to signify that N and T pass jointly to infinity.

2 Penalized Least Squares Estimation

This section considers panel structure models without endogeneity. It is convenient to assume first that the number of groups is known and later consider the determination of the number of unknown groups.

2.1 Panel Structure Models

The dependent variable y_{it} is measured for individual $i = 1, ..., N$ over time $t = 1, ..., T$. The generating mechanism is the panel structure model

$$y_{it} = \beta_i^{0\prime} x_{it} + \mu_i + u_{it} \tag{2.1}$$

where x_{it} is a $p \times 1$ vector of exogenous or predetermined variables, μ_i is an individual fixed effect that may be correlated with some components of x_{it}, u_{it} is the idiosyncratic error term with zero mean, and β_i^0 is a $p \times 1$ vector of slope parameters such that

$$\beta_i^0 = \begin{cases} \alpha_1^0 & \text{if } i \in G_1^0 \\ \vdots & \vdots \\ \alpha_{K_0}^0 & \text{if } i \in G_{K_0}^0 \end{cases}. \tag{2.2}$$

Here $\alpha_j^0 \neq \alpha_k^0$ for any $j \neq k$, $\cup_{k=1}^{K_0} G_k^0 = \{1, 2, ..., N\}$, and $G_k^0 \cap G_j^0 = \varnothing$ for any $j \neq k$. Let $N_k = \#G_k^0$ denote the cardinality of the set G_k^0. For the moment, we assume that the number K_0 of groups is known and fixed but that each individual's group membership is unknown. In addition, following Sun (2005) and Lin and Ng (2012), we implicitly assume that individual group membership does not vary over time. Let

$$\boldsymbol{\alpha} \equiv (\alpha_1, ..., \alpha_{K_0}) \text{ and } \boldsymbol{\beta} \equiv (\beta_1, ..., \beta_N). \tag{2.3}$$

Let $\mathcal{B}_i$ denote the parameter space of β_i.[2] We assume that $\mathcal{B}_i$ are compact uniformly in i and denote the true values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ as $\boldsymbol{\alpha}^0$ and $\boldsymbol{\beta}^0$, respectively. We are interested in developing econometric methods to infer each individual's group identity and to estimate the $p \times K_0$ matrix $\boldsymbol{\alpha}^0$ of group-specific coefficients.

2.2 Penalized Least Squares Estimation of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$

Our starting point is to develop PLS estimation of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ when the elements of x_{it} are either strictly exogenous or predetermined so that least squares criteria are appropriate. We first apply

[2] When the β_i's are group-specific, we can also regard the respective parameter spaces $\mathcal{B}_i$ to be group-specific.

ordinary least squares (OLS) regression, minimizing the following objective function[3]

$$Q_{0,NT}(\boldsymbol{\beta},\boldsymbol{\mu}) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it}-\beta_i' x_{it}-\mu_i\right)^2,$$

where $\boldsymbol{\mu}=(\mu_1,\mu_2,...,\mu_N)'$. Since the individual effects μ_i are not of primary interest, we concentrate them out and obtain the following concentrated function

$$Q_{1,NT}(\boldsymbol{\beta}) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\tilde{y}_{it}-\beta_i'\tilde{x}_{it}\right)^2,$$

giving the OLS estimates $\hat{\beta}_i^{OLS} = \left(\frac{1}{T}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{x}_{it}'\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{y}_{it}\right)^{-1}$, where $\tilde{x}_{it} = x_{it} - T^{-1}\sum_{t=1}^{T}x_{it}$ and $\tilde{y}_{it} = y_{it} - T^{-1}\sum_{t=1}^{T}y_{it}$.

Motivated by the literature on group Lasso (e.g., Yuan and Lin, 2006), we next propose to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ by minimizing the following PLS criterion function

$$Q_{1NT,\lambda_1}^{(K_0)}(\boldsymbol{\beta},\boldsymbol{\alpha}) = Q_{1,NT}(\boldsymbol{\beta}) + \frac{\lambda_1}{N}\sum_{i=1}^{N}\Pi_{k=1}^{K_0}\left\|\beta_i-\alpha_k\right\|, \tag{2.4}$$

where $\lambda_1 = \lambda_{1NT}$ is a tuning parameter. Minimizing the above criterion function produces *classifier-Lasso* (C-Lasso) estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$ of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, respectively. 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., $\hat{\boldsymbol{\alpha}} \equiv (\hat{\alpha}_1,...,\hat{\alpha}_K)$ and $\hat{\boldsymbol{\beta}} \equiv (\hat{\beta}_1,...,\hat{\beta}_N)$.

The penalty term in (2.4) takes a novel mixed *additive-multiplication* form that does not appear in the literature. Traditionally Lasso includes an additive penalty term to the least-squares, GMM, or log-likelihood objective function. When multiple penalty terms are needed, they also enter the objective function additively. In contrast, the C-Lasso method has N additive terms, each of which takes a multiplicative form as the product of K_0 separate penalties. Each of the K_0 penalty terms in the multiplicative expression shrinks the individual-level slope parameter vector β_i to a particular *unknown* group-level parameter vector α_k. This approach differs from the prototypical Lasso method of Tibshirani (1996) that shrinks a parameter to zero as well as the group Lasso method of Yuan and Lin (2006) that shrinks a parameter vector to a vector of zeros.

Note that the objective function in (2.4) is not convex in $\boldsymbol{\beta}$ even though it is (conditionally) convex in α_k when one fixes α_j for $j \neq k$. In Section 4.2 we propose an iterative algorithm to obtain the estimates $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$.

[3] If β_i's are identical across i, the approach will yield the well known within-group (WG) estimator or least squares dummy variable (LSDV) estimator, or fixed effects Guassian maximum likelihood estimator (MLE) in the literature; see, e.g., Kiviet (1995), Hahn and Kuersteiner (2002), and Alvarez and Arellano (2003). As will be clear, this appraoch can be easily extended to nonlinear panel data models.

2.3 Preliminary Rates of Convergence for Coefficient Estimates

We first present sufficient conditions to ensure the consistency of $(\hat{\boldsymbol{\beta}},\ \hat{\boldsymbol{\alpha}})$. Let $\tilde{u}_{it} = u_{it} - T^{-1}\sum_{t=1}^{T} u_{it}$, $\hat{Q}_{i,\tilde{x}\tilde{x}} = \frac{1}{T}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{x}_{it}'$, and $\hat{Q}_{i,\tilde{x}\tilde{u}} = \frac{1}{T}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{u}_{it}$. We make the following assumption.

ASSUMPTION A1. (i) $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{u}_{it} = O_P(1)$ *for each* $i = 1, ..., N$.

(ii) $\hat{Q}_{i,\tilde{x}\tilde{x}} \xrightarrow{P} Q_{i,\tilde{x}\tilde{x}} > 0$ *for each* $i = 1, ..., N$. *There exists a constant* $\underline{c}_{\tilde{x}\tilde{x}}$ *such that* $\underline{\lim}_{(N,T)\to\infty} \min_{1\le i\le N}\mu_{\min}(\hat{Q}_{i,\tilde{x}\tilde{x}}) \ge \underline{c}_{\tilde{x}\tilde{x}} > 0$.

(iii) $\frac{1}{N}\sum_{i=1}^{N}\left\|\hat{Q}_{i,\tilde{x}\tilde{u}}\right\|^2 = O_P\left(T^{-1}\right)$.

(iv) $N_k/N \to \tau_k \in (0,1)$ *for each* $k = 1, ..., K_0$ *as* $N \to \infty$.

(v) $\lambda_1 \to 0$ *as* $(N,T) \to \infty$.

Assumption A1(i) is rather weak and will be satisfied in most (stable) large dimensional linear panel data models without endogeneity. Sufficient conditions for A1(i) to hold are $\frac{1}{\sqrt{T}}\sum_{t=1}^{T} x_{it}u_{it}$, $\frac{1}{\sqrt{T}}\sum_{t=1}^{T} u_{it}$, and $\frac{1}{T}\sum_{t=1}^{T} x_{it} = O_P(1)$ for $i = 1, ..., N$. More primitive conditions for A1(i) to hold include $\mathbb{E}(u_{it}) = 0$, $\mathbb{E}(x_{it}u_{it}) = 0$ and suitable moment and weak dependence conditions on the process $\{(x_{it}, u_{it}), t \ge 1\}$ that ensure CLT validity. Note that we do not require that the panel model be dynamically correctly specified in the sense that $\mathbb{E}(u_{it}|\mathcal{F}_{i,t-1}) = 0$ where $\mathcal{F}_{i,t-1}$ is the sigma-field generated by $(x_{it}, x_{i,t-1}, u_{i,t-1}, ...)$. Instead, we allow both conditional heteroskedasticity and serial correlation in $\{u_{it}, t \ge 1\}$.

A1(ii) contains two parts, the first part being standard and the second part being a high-level condition. Appendix C.1 gives primitive conditions to ensure the second part. Intuitively, these conditions impose some restrictions on the moments of x_{it}, the dependence structure on the processes $\{x_{it}, t \ge 1\}$, and the relative rates at which N and T pass to infinity. More specifically, under suitable weak dependence conditions, if $\|x_{it}\|^2$ exhibits only $2q$-th finite moments for some $q > 1$, then we need a stringent (lower rate) condition on the expansion of T, viz., $T/N^{\epsilon} \to c \in (0, \infty]$ for some $\epsilon > 1/(2q-1)$. On the other hand, if $\|x_{it}\|^2$ has finite exponential moments with an index parameter γ as specified in Assumption C1(iv), then only $T/(\ln N)^{(1+\gamma)/\gamma} \to \infty$ is required for sufficiency. In the extreme case, if x_{it} is uniformly bounded (i.e., $\gamma = \infty$), it simply suffices that $T/\ln N \to \infty$.

A1(iii) can be easily verified via the Markov inequality. A1(iv) implies that each group has an asymptotically non-negligible membership number of individuals as $N \to \infty$. This assumption can be relaxed at the cost of more lengthy arguments, in which case the estimates of α_k^0, $k = 1, ..., K_0$, will exhibit different convergence rates. A1(v) implies that the penalty term cannot be too large.

The following theorem establishes the consistency of the PLS estimates $\{\hat{\beta}_i\}$ and $\{\hat{\alpha}_k\}$.

Theorem 2.1 *Suppose that Assumption A1 holds. Then*

(i) $\hat{\beta}_i - \beta_i^0 = O_P\left(T^{-1/2} + \lambda_1\right)$ *for* $i = 1, 2, ..., N$,

(ii) $\frac{1}{N}\sum_{i=1}^{N}\left\|\hat{\beta}_i-\beta_i^0\right\|^2=O_P\left(T^{-1}\right)$,
(iii) $\left(\hat{\alpha}_{(1)},...,\hat{\alpha}_{(K_0)}\right)-\left(\alpha_1^0,...,\alpha_{K_0}^0\right)=O_P\left(T^{-1/2}\right)$
where $\left(\hat{\alpha}_{(1)},...,\hat{\alpha}_{(K_0)}\right)$ *is a suitable permutation of* $(\hat{\alpha}_1,...,\hat{\alpha}_{K_0})$.

REMARK 1. Parts (i) and (ii) of Theorem 2.1 establish the pointwise and mean-square convergence of $\hat{\beta}_i$. Part (iii) of Theorem 2.1 indicates that the group-specific parameters $\alpha_1^0,...,\alpha_K^0$ can also be estimated consistently by $\hat{\alpha}_1,...,\hat{\alpha}_{K_0}$ subject to permutation. As expected and consonant with other Lasso limit theory, the pointwise convergence rate of $\hat{\beta}_i$ depends on the rate at which the tuning parameter λ_1 converges to zero. Somewhat unexpectedly, this requirement is not the case either for mean-square convergence of $\hat{\beta}_i$ or convergence of $\hat{\alpha}_k$. Apparently if $\lambda_1=O\left(T^{-1/2}\right)$, we get the usual $\sqrt{T}$-convergence rate for the $\hat{\beta}_i$.

For notational simplicity, hereafter we simply write $\hat{\alpha}_k$ for $\hat{\alpha}_{(k)}$ as the consistent estimator of α_k^0, and define

$$\hat{G}_k=\left\{i\in\{1,2,...,N\}:\hat{\beta}_i=\hat{\alpha}_k\right\}\text{ for }k=1,...,K_0. \tag{2.5}$$

2.4 Classification Consistency

This section studies classification consistency. Roughly speaking, a classification method is consistent if it classifies each individual to the correct group w.p.a.1. For a rigorous statement of this property we define the following sequences of events

$$\hat{E}_{kNT,i}=\left\{i\notin\hat{G}_k\mid i\in G_k^0\right\}\text{ and }\hat{F}_{kNT,i}=\left\{i\notin G_k^0\mid i\in\hat{G}_k\right\}, \tag{2.6}$$

where $i=1,...,N$ and $k=1,...,K_0$. Let $\hat{E}_{kNT}=\cup_{i\in G_k^0}\hat{E}_{kNT,i}$ and $\hat{F}_{kNT}=\cup_{i\in\hat{G}_k}\hat{F}_{kNT,i}$. The events $\hat{E}_{kNT}$ and $\hat{F}_{kNT}$ mimic Type I and II errors in statistical tests: $\hat{E}_{kNT}$ denotes the error event of not classifying an element of G_k^0 into the 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 $\hat{G}_k$. To achieve uniform consistency in estimation both error types must be controlled. We use the following definition.

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

To establish consistency of the PLS classifier we add the following assumption.

ASSUMPTION A2. (i) $T\lambda_1\to\infty$ *and* $T\lambda_1^4\to c_0\in[0,\infty)$ *as* $(N,T)\to\infty$.
(ii) *For any* $c>0$, $N\max_{1\le i\le N}P\left(\left\|T^{-1}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{u}_{it}\right\|\ge c\sqrt{\lambda_1}\right)\to 0$ *as* $(N,T)\to\infty$.

Assumption A2(i) is required for *individual consistency* of the PLS classifier. Assumption A2(ii) is a high level assumption that ensures the *uniform consistency* of the classifier. In Appendix C, we verify this condition for strong mixing processes with geometric decay rates under certain moment conditions. In particular, if (a) $\|x_{it}\|$, $|u_{it}|$, and $\|x_{it}u_{it}\|$ have finite $2q^{\text{th}}$ moments, then A2(ii) will be satisfied provided

$$\lambda_1 \gg \max\{T^{-1}\ln N,\ T^{-2}(NT)^{1/q}(\ln T)^4(\ln N)^2\}; \tag{2.7}$$

(b) if $\|x_{it}\|$, $|u_{it}|$, and $\|x_{it}u_{it}\|$ have exponential moments with an index parameter γ, then A2(ii) will be satisfied provided

$$\lambda_1 \gg \max\{T^{-1}\ln N,\ T^{-2}[\ln(NT)]^{2(1+\gamma)/\gamma}\}. \tag{2.8}$$

In either case, we need $T\lambda_1 \gg \ln N$. If $T \propto N^{\epsilon_1}$ for some $\epsilon_1 > 1/(q-1)$ in case (a) and $T \propto N^{\epsilon_2}$ for some $\epsilon_2 > 0$ in case (b), then we can easily verify that $T\lambda_1 \gg \ln N$ would also be sufficient to ensure A2(ii). Combining this requirement with A2(i) suggests that under certain conditions on the moments and on the related rates at which N and T pass to infinity, it suffices to require that

$$\lambda_1 \propto T^{-a} \text{ for any } a \in [1/4, 1). \tag{2.9}$$

The following theorem establishes uniform consistency for the PLS classifier.

Theorem 2.2 *Suppose that Assumptions A1-A2 hold. Then*
(i) $P\left(\cup_{k=1}^{K_0}\hat{E}_{kNT}\right) \leq \sum_{k=1}^{K_0} P(\hat{E}_{kNT}) \to 0$ *as* $(N,T) \to \infty$,
(ii) $P\left(\cup_{k=1}^{K_0}\hat{F}_{kNT}\right) \leq \sum_{k=1}^{K_0} P(\hat{F}_{kNT}) \to 0$ *as* $(N,T) \to \infty$.

REMARK 2. Theorem 2.2 implies that 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 correctly belong to the same group (G_k^0) w.p.a.1. Let $\hat{G}_0$ denote the group of individuals in $\{1, 2, ..., N\}$ that are not classified into any of the K_0 groups, i.e., $\hat{G}_0 = \{1, 2, ..., N\} \setminus (\cup_{k=1}^{K_0}\hat{G}_k)$. Define the events $\hat{H}_{iNT} = \{i \in \hat{G}_0\}$. Theorem 2.2(i) implies that $P(\cup_{1\leq i\leq N}\hat{H}_{iNT}) \leq \sum_{k=1}^{K_0} P(\hat{E}_{kNT}) \to 0$. That is, all individuals can be classified into one of the K_0 groups w.p.a.1. Nevertheless, when T is not large, it is possible for a small percentage of individuals to be left unclassified if we stick with the classification method defined in (2.5). To ensure that all individuals are classified into one of the K_0 groups in finite samples, one need only slightly modify the classifier to achieve it. In particular, we classify $i \in \hat{G}_k$ if $\hat{\beta}_i = \hat{\alpha}_k$ for some $k = 1, ..., K_0$, and $i \in \hat{G}_l$ for some $l = 1, ..., K_0$ if

$$\left\|\hat{\beta}_i - \hat{\alpha}_l\right\| = \min\left\{\left\|\hat{\beta}_i - \hat{\alpha}_1\right\|, ..., \left\|\hat{\beta}_i - \hat{\alpha}_K\right\|\right\} \text{ and } \sum_{k=1}^{K_0}\mathbf{1}\left\{\hat{\beta}_i = \hat{\alpha}_k\right\} = 0.$$

Since the event $\sum_{k=1}^{K_0} \mathbf{1}\{\hat{\beta}_i = \hat{\alpha}_k\} = 0$ occurs with probability tending to zero uniformly in i, we can ignore it in large samples in subsequent theoretical analysis and restrict our attention to the previous classification rule in (2.5) to avoid confusion. That is, $\hat{G}_k = \{i \in \{1, ..., N\} : \hat{\beta}_i = \hat{\alpha}_k\}$ for $k = 1, ..., K_0$.

Let $\hat{N}_k = \sum_{i=1}^{N} \mathbf{1}\{i \in \hat{G}_k\}$. The following corollary indicates that we can estimate the number of individuals within each group consistently.

Corollary 2.3 *Suppose that Assumptions A1-A2 hold. Then* $\hat{N}_k - N_k = o_P(1)$ *for* $k = 1, ..., K_0$.

2.5 The Oracle Property and Asymptotic Properties of Post-Lasso

To establish the oracle property of the PLS estimates $\{\hat{\alpha}_k\}$, we add the following assumption.

ASSUMPTION A3. (i) *For each* $k = 1, ..., K_0$, $\bar{\Phi}_k \equiv \frac{1}{N_k T} \sum_{i \in G_k^0} \sum_{t=1}^{T} \tilde{x}_{it} \tilde{x}_{it}' \xrightarrow{P} \Phi_k > 0$ *as* $(N, T) \to \infty$.

(ii) *For each* $k = 1, ..., K_0$, $\frac{1}{\sqrt{N_k T}} \sum_{i \in G_k^0} \sum_{t=1}^{T} \tilde{x}_{it} \tilde{u}_{it} - \mathbb{B}_{kNT} \xrightarrow{D} N(0, \Psi_k)$ *as* $(N, T) \to \infty$ *where* $\mathbb{B}_{kNT} = \frac{1}{\sqrt{N_k T}} \sum_{i \in G_k^0} \sum_{t=1}^{T} \mathbb{E}(x_{it} \tilde{u}_{it})$ *is either* 0 *or* $O(\sqrt{N_k/T})$ *depending on whether* x_{it} *is strictly exogenous.*

Assumption A3 is a convenient high level condition. It can be verified under various commonly occurring primitive conditions. For example, if (a) $\{(x_{it}, u_{it})\}$ is a stationary strong mixing process with a geometric mixing rate along the time dimension and is independently and identically distributed (IID) along the cross section dimension for all individuals within the same group G_k^0, (b) x_{it} and $x_{it} u_{it}$ have finite two-plus moments, and (c) $\mathbb{E}(x_{it} \tilde{u}_{it}) = 0$ and $\mathbb{E}(u_{it}) = 0$, then A3 is satisfied with $\mathbb{B}_{kNT} = 0$, $\Phi_k = \text{Var}(x_{it})$, and $\Psi_k = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \sum_{s=1}^{T} \mathbb{E}(x_{it} x_{is}' u_{it} u_{is})$ for any $i \in G_k^0$. Apparently, condition (c) rules out the case of dynamic panel data models. If x_{it} contains lagged dependent variables (e.g., $y_{i,t-1}$), it is well known that the fixed effects within-group (WG) estimator has asymptotic bias of order $O(1/T)$ in homogeneous dynamic panel data models. This suggests that $\mathbb{B}_{kNT} = O(\sqrt{N_k/T})$ in dynamic panel data models and bias correction is required for statistical inference unless T passes to infinity faster than N_k. Matters of bias correction and some explicit formulae in this case are discussed below in Remark 5 and Appendix D.1.

The following theorem gives the oracle property of the Lasso estimator $\{\hat{\alpha}_k\}$.

Theorem 2.4 *Suppose that Assumptions A1-A3 hold. Then* $\sqrt{N_k T}(\hat{\alpha}_k - \alpha_k^0) - \bar{\Phi}_k^{-1} \mathbb{B}_{kNT} \xrightarrow{D} N(0, \Phi_k^{-1} \Psi_k \Phi_k^{-1})$ *for* $k = 1, ..., K_0$.

REMARK 3. If each individual's group membership is known, the WG estimator of α_k^0 is $\bar{\alpha}_k = \left(\sum_{i \in G_k^0} \sum_{t=1}^{T} \tilde{x}_{it} \tilde{x}_{it}'\right)^{-1} \sum_{i \in G_k^0} \sum_{t=1}^{T} \tilde{x}_{it} \tilde{y}_{it}$, and then $\sqrt{N_k T}(\bar{\alpha}_k - \alpha_k^0) - \bar{\Phi}_k^{-1} \mathbb{B}_{kNT} \xrightarrow{D} N\left(0, \Phi_k^{-1} \Psi_k \Phi_k^{-1}\right)$ under Assumption A3. Theorem 2.4 indicates that the PLS estimator $\hat{\alpha}_k$ achieves the same limit distribution as this oracle WG estimator with knowledge of the exact

membership of each individual. In this sense, we say that the PLS estimators $\{\hat{\alpha}_k\}$ have the asymptotic oracle property. In the Appendix, we prove the above theorem by inspection of the Karush-Kuhn-Tucker (KKT) optimality conditions for minimizing the objective function in (2.4) based on subdifferential calculus (e.g., Bertsekas, 1995, Appendix B.5). We then show that $\sqrt{N_k T}\left(\hat{\alpha}_k - \alpha_k^0\right) = \sqrt{N_k T}(\hat{\alpha}_{\hat{G}_k} - \alpha_k^0) + o_P(1)$, where $\hat{\alpha}_{\hat{G}_k}$ is the post-Lasso estimator of α_k^0 given by

$$\hat{\alpha}_{\hat{G}_k} = \left(\sum_{i \in \hat{G}_k} \sum_{t=1}^{T} \tilde{x}_{it} \tilde{x}_{it}' \right)^{-1} \sum_{i \in \hat{G}_k} \sum_{t=1}^{T} \tilde{x}_{it} \tilde{y}_{it}. \tag{2.10}$$

The following theorem reports the asymptotic distribution of $\hat{\alpha}_{\hat{G}_k}$.

Theorem 2.5 *Suppose that Assumptions A1-A3 hold. Then* $\sqrt{N_k T}(\hat{\alpha}_{\hat{G}_k} - \alpha_k^0) - \bar{\Phi}_k^{-1} \mathbb{B}_{kNT} \xrightarrow{D} N(0,\, \Phi_k^{-1} \Psi_k \Phi_k^{-1})$ *for* $k = 1, ..., K_0$.

REMARK 4. The proof of the above theorem is based on the uniform classification consistency results in Theorem 2.2. In a totally different framework, Belloni and Chernozhukov (2013) study post-Lasso estimators which apply OLS to the model selected by first-step penalized estimators and show that the post-Lasso estimators perform at least as well as Lasso in terms of rate of convergence and have the advantage of having a smaller bias. It would also be interesting to compare the high-order asymptotic properties of $\hat{\alpha}_k$ and $\hat{\alpha}_{\hat{G}_k}$ given that they share the same first-order asymptotic distribution. But that analysis goes beyond the scope of the current paper. We do compare the performance of the post-Lasso estimators and the C-Lasso estimators in simulations reported below.

REMARK 5. As mentioned above, $\mathbb{B}_{kNT} = 0$ in Assumption A3(ii) under strict exogeneity. In the case of dynamic panel data models, we have to obtain a consistent estimate of $b_{kNT} \equiv \bar{\Phi}_k^{-1} \mathbb{B}_{kNT}$ in order to perform inference. Various methods have been proposed to estimate b_{kNT} in the literature under conditions that are typically simpler than the latent structure model considered here. These methods generally involve first stage consistent estimates that are subsequently plugged-into analytic formulae for the asymptotic bias function to achieve the correction. For example, Kiviet (1995) and Hahn and Kuersteiner (2002) derived bias formulae for the WG estimator of a common autoregressive coefficient in first-order autoregressive (AR(1)) panel data models with exogenous regressors and propose ways to correct the bias such as the use of plug-in corrections. Phillips and Sul (2007b) provide explicit asymptotic bias formulae for linear dynamic panel regression estimators where the models may or may not exhibit unit roots, incidental trends, exogenous regressors, and cross section dependence, all of which lead to different formulae. Lee (2012) considers bias correction for WG estimators in higher-order autoregressive models with exogenous regressors where the lag order is possibly misspecified. Other methods, such as median unbiased estimation, indirect inference (Gourieroux, Phillips, and Yu, 2010), and

X-differencing (Han, Phillips, and Sul, 2014) have been used in dynamic panel data models to avoid bias problems. To conserve space, we refer the readers directly to those papers for details of these particular formulae and the correction procedures employed. In the present case, since the formula for $b_{kNT} \equiv \bar{\Phi}_k^{-1}\mathbb{B}_{kNT}$ is known and can be explicitly represented in cases such as the presence of lagged dependent variables in x_{it}, we can also use a plug-in estimator to achieve bias correction. The approach is similar to that proposed in Hahn and Kuersteiner (2002) and recently reviewed in Moon, Perron, and Phillips (2014). However, in the present model the bias term $\mathbb{B}_{kNT} = \frac{1}{\sqrt{N_k T}} \sum_{i \in G_k^0} \sum_{t=1}^{T} \mathbb{E}\left(x_{it}\tilde{u}_{it}\right)$ inevitably reflects the latent structure of the model and thereby involves further complications. For instance, in the panel AR(1) model there is no longer a single common AR coefficient as in Hahn and Kuersteiner (2002). Implementation therefore requires plug-in estimates of each of the common autoregressive coefficients that appear in the group structures $\{G_k^0\}_{k=1}^{K_0}$. It follows that consistent group structure estimation by $\{\hat{G}_k\}_{k=1}^{K_0}$ is necessary for the plug-in mechanism to be feasible. To fix ideas, suppose the model (2.1) has the panel AR(1) form

$$y_{it} = \beta_i^0 y_{it-1} + \mu_i + u_{it}, \ \ \left|\beta_i^0\right| < 1 \text{ for all } i, \ u_{it} \sim iid\left(0, \sigma^2\right) \tag{2.11}$$

with latent structure (2.2) giving $\beta_i^0 = \alpha_k^0$ for $i \in G_k^0$. Since $\mathbb{E}\left(u_{i,t-1-j}u_{is}\right) = \sigma^2 \mathbf{1}\left\{t = s+1+j\right\}$, we have for $i \in G_k^0$

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}\left(y_{it-1}\tilde{u}_{it}\right) &= -T^{-1}\sum_{t,s=1}^{T} \mathbb{E}\left(y_{it-1}u_{is}\right) = -T^{-1}\sum_{t,s=1}^{T}\sum_{j=0}^{\infty}\left(\alpha_k^0\right)^j \mathbb{E}\left(u_{i,t-1-j}u_{is}\right) \\
&= -\sigma^2\frac{1}{T}\sum_{s=1}^{T}\sum_{j=0}^{T-s-1}\left(\alpha_k^0\right)^j = -\sigma^2\frac{1}{T}\sum_{s=1}^{T}\frac{1-\left(\alpha_k^0\right)^{T-s}}{1-\alpha_k^0} \\
&= -\frac{\sigma^2}{1-\alpha_k^0} + \frac{\sigma^2}{1-\alpha_k^0}\frac{1}{T}\sum_{s=1}^{T}\left(\alpha_k^0\right)^{T-s} = -\frac{\sigma^2}{1-\alpha_k^0} + \frac{\sigma^2}{1-\alpha_k^0}\frac{1}{T}\frac{1-\left(\alpha_k^0\right)^T}{1-\alpha_k^0},
\end{aligned}$$

so that

$$\mathbb{B}_{kNT} = \sqrt{\frac{1}{N_k T}}\sum_{i \in G_k^0}\sum_{t=1}^{T}\mathbb{E}\left(y_{it-1}\tilde{u}_{it}\right) = -\sqrt{\frac{N_k}{T}}\frac{\sigma^2}{1-\alpha_k^0} + O\left(\frac{1}{\sqrt{N_k T}}\right).$$

Further, as $(N_k, T) \to \infty$ we have

$$\bar{\Phi}_k \equiv \frac{1}{N_k T}\sum_{i \in G_k^0}\sum_{t=1}^{T}\tilde{y}_{it-1}^2 \to_{a.s} \mathbb{E}\left(y_{it-1}^2 \mathbf{1}\left\{i \in G_k^0\right\}\right) = \frac{\sigma^2}{1-\left(\alpha_k^0\right)^2} = \Phi_k,$$

so that

$$\begin{aligned}\sqrt{N_kT}\left(\hat{\alpha}_{\hat{G}_k}-\alpha_k^0\right)-\bar{\Phi}_k^{-1}\mathbb{B}_{kNT} &= \sqrt{N_kT}\left(\hat{\alpha}_{\hat{G}_k}-\alpha_k^0\right)+\sqrt{\frac{N_k}{T}}\frac{1-\left(\alpha_k^0\right)^2}{1-\alpha_k^0}+O\left(\frac{1}{\sqrt{N_kT}}\right)\\ &= \sqrt{N_kT}\left(\hat{\alpha}_{\hat{G}_k}-\alpha_k^0+\frac{1+\alpha_k^0}{T}\right)+o_p(1)\\ &\xrightarrow{D} N(0,1-\left(\alpha_k^0\right)^2),\end{aligned} \tag{2.12}$$

since $\Psi_k=\sigma^4/[1-\left(\alpha_k^0\right)^2]$ here. As in Hahn and Kuersteiner (2002), (2.12) suggests a simple bias correction within $\hat{G}_k$, viz.,

$$\tilde{\alpha}_{\hat{G}_k}=\hat{\alpha}_{\hat{G}_k}+\frac{1+\hat{\alpha}_{\hat{G}_k}}{T}=\frac{T+1}{T}\hat{\alpha}_{\hat{G}_k}+\frac{1}{T},\ k=1,....,K_0, \tag{2.13}$$

giving bias corrected estimators for the latent structure panel AR(1) model (2.11). Of course, formula (2.13) gives appropriate bias correction only in the stationary case where $\left|\alpha_k^0\right|<1$ for all k. For the general case, see the supplementary Appendix D.1 for the bias correction.

2.6 Determination of the Number of Groups

In practice, the exact number K_0 of groups is typically unknown. We assume that the true number of groups is bounded from above by a finite integer $K_{\max}$ and study the determination of the number of groups via some information criterion. Consider the following PLS criterion

$$Q_{1NT,\lambda_1}^{(K)}(\boldsymbol{\beta},\boldsymbol{\alpha})=Q_{1,NT}(\boldsymbol{\beta})+\frac{\lambda_1}{N}\sum_{i=1}^{N}\Pi_{k=1}^{K}\left\|\beta_i-\alpha_k\right\|, \tag{2.14}$$

where $1\leq K\leq K_{\max}$. By minimizing the objective function (2.14), we obtain the C-Lasso estimates $\{\hat{\beta}_i(K,\lambda_1),\hat{\alpha}_k(K,\lambda_1)\}$ of $\{\beta_i,\alpha_k\}$, where we make the dependence of $\hat{\beta}_i$ and $\hat{\alpha}_k$ on (K,λ_1) explicit. As above, we can classify individual i into group $\hat{G}_k(K,\lambda_1)$ if and only if $\hat{\beta}_i(K,\lambda_1)=\hat{\alpha}_k(K,\lambda_1)$, i.e.,

$$\hat{G}_k(K,\lambda_1)=\left\{i\in\{1,2,...,N\}:\hat{\beta}_i(K,\lambda_1)=\hat{\alpha}_k(K,\lambda_1)\right\}\text{ for }k=1,...,K. \tag{2.15}$$

Let $\hat{G}(K,\lambda_1)=\{\hat{G}_1(K,\lambda_1),...,\hat{G}_K(K,\lambda_1)\}$. Based on (2.15), define the post-Lasso estimate of α_k^0 by

$$\hat{\alpha}_{\hat{G}_k(K,\lambda_1)}=\left(\sum_{i\in\hat{G}_k(K,\lambda_1)}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{x}_{it}'\right)^{+}\sum_{i\in\hat{G}_k(K,\lambda_1)}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{y}_{it}. \tag{2.16}$$

where A^+ denotes the Moore-Penrose inverse of A. Let $\hat{\sigma}^2_{\hat{G}(K,\lambda_1)}=\frac{1}{NT}\sum_{k=1}^{K}\sum_{i\in\hat{G}_k(K,\lambda_1)}\sum_{t=1}^{T}[\tilde{y}_{it}-\hat{\alpha}'_{\hat{G}_k(K,\lambda_1)}\tilde{x}_{it}]^2$. We propose to select the number of groups by choosing K to minimize the following information criterion:

$$IC_1(K,\lambda_1)=\ln\left[\hat{\sigma}^2_{\hat{G}(K,\lambda_1)}\right]+\rho_{1NT}pK, \tag{2.17}$$

where ρ_{1NT} is a tuning parameter. Similar information criteria are used to choose tuning parameters by Wang, Li, and Tsai (2007), Liao (2013), and Lu and Su (2013) for shrinkage estimation in various contexts and have been found to work satisfactorily.

We proceed to describe the asymptotic properties of (2.17). First, some notation. Let $\mathcal{K} = \{1, 2, ..., K_{\max}\}$. We divide $\mathcal{K}$ into three subsets $\mathcal{K}_0$, $\mathcal{K}_-$ and $\mathcal{K}_+$ as follows

$$\mathcal{K}_0 = \{K \in \mathcal{K} : K = K_0\}, \ \mathcal{K}_- = \{K \in \mathcal{K} : K < K_0\}, \text{ and } \mathcal{K}_+ = \{K \in \mathcal{K} : K > K_0\}.$$

The sets $\mathcal{K}_0$, $\mathcal{K}_-$ and $\mathcal{K}_+$ denote subsets of $\mathcal{K}$ in which true, under-, and over-fitted models are produced. Let $G^{(K)} = (G_{K,1}, ..., G_{K,K})$ be any K-partition of the set of individual indices $\{1, 2, ..., N\}$. Let $\mathcal{G}_K$ denote the collection of such partitions. Let $\hat{\sigma}^2_{G^{(K)}} = \frac{1}{NT}\sum_{k=1}^{K}\sum_{i\in G_{K,k}}\sum_{t=1}^{T}[\tilde{y}_{it} - \hat{\alpha}'_{G_{K,k}}\tilde{x}_{it}]^2$, where $\hat{\alpha}_{G_{K,k}} = \left(\sum_{i\in G_{K,k}}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{x}'_{it}\right)^{+}\sum_{i\in G_{K,k}}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{y}_{it}$. The following assumptions are useful in the asymptotic development.

ASSUMPTION A4. *As* $(N,T) \to \infty$, $\min_{1\le K<K_0}\inf_{G^{(K)}\in\mathcal{G}_K}\hat{\sigma}^2_{G^{(K)}} \xrightarrow{P} \underline{\sigma}^2 > \sigma_0^2$, *where* $\sigma_0^2 = p\lim_{(N,T)\to\infty}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{u}_{it}^2$.

ASSUMPTION A5. *As* $(N,T) \to \infty$, $\rho_{1NT} \to 0$ *and* $\rho_{1NT}\delta_{NT}^2 \to \infty$ *where* $\delta_{NT} = N^{1/2}T^{1/2}$ *if* $\mathbb{B}_{kNT} = 0$ and $\min(N^{1/2}T^{1/2}, T)$ otherwise.

Assumption A4 is intuitively clear and applies under primitive conditions in a variety of models, such as panel autoregressions. It requires that all under-fitted models yield asymptotic mean square errors that are larger than σ_0^2, which is delivered by the true model. A5 reflects the usual conditions for the consistency of model selection. The penalty coefficient ρ_{1NT} cannot shrink to zero either too fast or too slowly.

The following theorem justifies the use of (2.17) as a selector criterion for K.

Theorem 2.6 *Suppose that Assumptions A1-A5 hold. Then*

$$P\left(\inf_{K\in\mathcal{K}_-\cup\mathcal{K}_+} IC_1(K,\lambda_1) > IC_1(K_0,\lambda_1)\right) \to 1 \text{ as } (N,T)\to\infty.$$

REMARK 6. Let $K(\lambda_1) = \arg\min_{1\le K\le K_{\max}} IC_1(K,\lambda_1)$. As Theorem 2.6 indicates, as long as λ_1 satisfies Assumptions A1(v) and A2, we have $P(K(\lambda_1) = K_0) \to 1$ as $(N,T)\to\infty$. Consequently, the minimizer of $IC_1(K,\lambda_1)$ with respect to K is equal to K_0 w.p.a.1 for a variety of choices of λ_1. In practice, it is desirable to have a data-driven method to choose the tuning parameter λ_1. For this purpose, define

$$IC_1^*(\lambda_1) = IC_1(K(\lambda_1),\lambda_1).$$

The tuning parameter can then be chosen as $\hat{\lambda}_1 = \arg\min_{\lambda_1\in\Lambda_1} IC_1^*(\lambda_1)$, where $\Lambda_1 = \{\lambda_1 : \lambda_1 \propto T^{-a} \text{ for any } a \in [1/4, 1)\}$ provided some conditions on the moments of $\|x_{it}\|, |u_{it}|$ and $\|x_{it}u_{it}\|$ and on the relative rates at which N and T pass to infinity are satisfied – see the remark after Assumption A2.

2.7 Extensions

Several major extensions of the C-Lasso methodology to other models and contexts are worth mentioning. We discuss four possibilities below.

1. Mixed Panel Structure Models: Consider the case where some of the parameters in β_i^0 are common across all individuals whereas others are group-specific. Write $\beta_i^0 = (\beta_{i(1)}^{0\prime}, \beta_{i(2)}^{0\prime})'$ where $\beta_{i1}^0 = \beta_{(1)}^0$ for all $i = 1, ..., N$. Partition x_{it} conformably as $x_{it} = (x_{it(1)}', x_{it(2)}')'$. The panel structure becomes

$$y_{it} = \beta_{(1)}^{0\prime} x_{it(1)} + \beta_{i(2)}^{0\prime} x_{it(2)} + \mu_i + u_{it}, \tag{2.18}$$

where $\beta_{i(2)}^0 = \alpha_k^0$ if $i \in G_k^0$ where $k = 1, ..., K_0$ and $G_1^0, ..., G_{K_0}^0$ form a partition for $\{1, 2, ..., N\}$. The model (2.18) is closely related to the model studied by Pesaran, Shin, and Smith (1999) in which long-run coefficients are constrained to be identical across individuals while short-run coefficients may be heterogenous. In this case, the PLS objective function becomes

$$Q_{1NT,\lambda_1}^{(K_0)}\left(\beta_{(1)}, \boldsymbol{\beta}_{(2)}, \boldsymbol{\alpha}\right) = Q_{1,NT}\left(\beta_{(1)}, \boldsymbol{\beta}_{(2)}\right) + \frac{\lambda_1}{N}\sum_{i=1}^{N} \Pi_{k=1}^{K_0} \left\|\beta_{i(2)} - \alpha_k\right\|, \tag{2.19}$$

where $Q_{1,NT}(\beta_{(1)}, \boldsymbol{\beta}_{(2)}) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\tilde{y}_{it} - \beta_{(1)}'\tilde{x}_{it(1)} - \beta_{i(2)}'\tilde{x}_{it(2)}\right)^2$, $\boldsymbol{\beta}_{(2)} = (\beta_{1(2)}, ..., \beta_{N(2)})$, and $\tilde{x}_{it(r)} = x_{it(r)} - T^{-1}\sum_{s=1}^{T} x_{is(r)}$ for $r = 1, 2$. Our previous analysis can now be followed to establish uniform consistency for the classifier and the oracle property for the resulting estimators of $\beta_{(1)}^0$ and α_k^0's.

2. Nonlinear Panel Data Models: Bester and Hansen (2013) consider estimation of nonlinear panel data models with common and group-specific parameters where the group structure is completely known, e.g., based on some external classification or geographic location. They provide conditions under which their group effects estimators of the common parameter are asymptotically unbiased. To fix ideas, consider minimizing the following objective function

$$Q_{1,NT}\left(\theta, \boldsymbol{\mu}\right) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\varphi\left(w_{it}, \theta, \mu_i\right), \tag{2.20}$$

where θ is a finite dimensional common parameter, $\boldsymbol{\mu}$=$(\mu_1, ..., \mu_N)$, $\varphi = -\ln f$, and $f\left(w_{it}, \theta, \mu_i\right)$ is the density function of w_{it} with respect to some measure. Here the μ_i denote time invariant individual-specific effects that are held constant according to an *observed* group structure: $\mu_i^0 = \alpha_k^0$ if $i \in G_k^0$ where $k = 1, ..., K_0$ and $\left\{G_1^0, ..., G_{K_0}^0\right\}$ forms a partition for $\{1, 2, ..., N\}$.[4] Interestingly, the PLS C-Lasso method can be extended to study such nonlinear panel data models straightforwardly without the need to know each individual's group membership. The PLS

[4] In traditional nonlinear panel data models, the individual effect μ_i is a scalar, but our theory allows it to be a vector. The α_k^0's are referred to as the group (fixed) effects in the literature.

objective function here takes the form

$$Q_{1NT,\lambda_1}^{(K_0)}\left(\theta,\boldsymbol{\mu},\boldsymbol{\alpha}\right)=Q_{1,NT}\left(\theta,\boldsymbol{\mu}\right)+\frac{\lambda_1}{N}\sum_{i=1}^{N}\Pi_{k=1}^{K_0}\left\|\mu_i-\alpha_k\right\|. \tag{2.21}$$

One can readily modify our numerical algorithm to estimate both the common parameter θ^0 and the group-specific parameters $\left\{\alpha_k^0\right\}$. The uniform consistency of the C-Lasso classifier and the oracle properties of the parametric estimates can also be established.

3. Group Patterns of Heterogeneity: Bonhomme and Manresa (2014) consider a linear panel data model with grouped patterns of heterogeneity that take the following form

$$y_{it}=\theta^{0\prime}x_{it}+\mu_{g_it}+u_{it}, \tag{2.22}$$

where the group membership variables $g_i\in\{1,...,K_0\}$ map individual units into groups. They propose to estimate the group membership along with the common parameter θ^0 in the model based on some variants of the K-means algorithm and establish the asymptotic distributions for the resulting estimators. In view of the fact that μ_{g_it} has a factor structure $\mu_{g_it}=\lambda_i'f_t$ where $f_t=(\mu_{1t},...,\mu_{K_0t})'$, $\lambda_i=(0,...1,...0)'$ with 1 in the k^{th} position if $i\in G_k^0$ for $k=1,...,K_0$ and zeros elsewhere, we may embed (2.22) in the more general model

$$y_{it}=\theta^{0\prime}x_{it}+\lambda_i^{0\prime}f_t^0+u_{it}, \tag{2.23}$$

where $\lambda_i^0=\alpha_k^0$ if $i\in G_k^0$ where $k=1,...,K_0$ and $\left\{G_1^0,...,G_{K_0}^0\right\}$ forms a partition for $\{1,2,...,N\}$. In the economic growth literature, f_t represents unobserved global shocks to the economy, and λ_i^0 the marginal effects of the shocks to country i's economic growth. It is sensible to assume that the marginal effects are identical for countries that exhibit similar features. To estimate (2.23) with the unknown group structure, we propose a two-step approach. In the first step, we follow Bai (2009) and obtain the Gaussian quasi-maximum likelihood estimates $\breve{\theta}$, $\breve{\lambda}_i$, and $\breve{f}_t$ of θ^0, λ_i^0, and f_t^0 under the identification restrictions that $T^{-1}\sum_{t=1}^{T}f_tf_t'=I_{K_0}$ and $N^{-1}\sum_{i=1}^{N}\lambda_i\lambda_i'$ is diagonal. In the second step, we consider the following regression

$$y_{it}=\theta^{0\prime}x_{it}+\lambda_i^{0\prime}\breve{f}_t+u_{it}, \tag{2.24}$$

by imposing the unknown group structure: $\lambda_i^0=\alpha_k^0$ if $i\in G_k^0$ where $k=1,...,K_0$. The PLS objective function is similar to that in (2.19). In this framework, we can readily show that C-Lasso yields uniform consistency for the classification and the oracle properties of the estimators of θ^0 and α_k^0 just as if we were able to observe the exact group structure.

4. Granger-causality, Unit Roots, and Cointegration in Heterogenous Panels: The C-Lasso methodology can also be extended to analyze Granger-causality, unit roots, and cointegration in heterogenous panels. In Granger-causality analysis we may consider either completely

homogenous or completely heterogenous relationships. The former may produce misleading conclusions if the causal or non-causal relationship is heterogeneous; the latter may yield imprecise estimates and low power in hypothesis testing. An intermediate specification is to allow the relationship to be group-specific. Similar remarks hold for panel unit root and cointegration tests – see Breitung and Pesaran (2008) for an overview on this. As usual in nonstationary settings, careful attention must be given to allow for different convergence rates for different parameters in such systems (Phillips and Moon, 1999).

The C-Lasso approach is also well suited to testing for structural change in heterogeneous panel data models, to nonparametric and semiparametric panel data models, and to models with heterogeneous parametric or nonparametric time trends (e.g., Kneip, Sickles, and Song 2012, Zhang, Su, and Phillips 2012). We can expect C-Lasso to deliver substantial efficiency gains in some of these cases where there is only partial heterogeneity in the structure. These and other applications of the methodology will be examined in separate studies.

3 Penalized GMM Estimation of Panel Structure Models

This section considers penalized GMM estimation of panel structure models when some regressors are lagged dependent variables or endogenous. As before, we first assume that the number of groups is known and then consider the determination of the number of groups when that information is unknown.

3.1 Penalized GMM Estimation of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$

We consider the first differenced system

$$\Delta y_{it} = \beta_i^{0\prime} \Delta x_{it} + \Delta u_{it}, \tag{3.1}$$

where, e.g., $\Delta y_{it} = y_{it} - y_{i,t-1}$ for $t = 1, ..., T$ and $i = 1, ..., N$, and we assume that we have observations on y_{i0} and x_{i0}. Let z_{it} be a $d \times 1$ vector of instruments for Δx_{it} where $d \geq p$. Define $\Delta y_i = (\Delta y_{i1}, ..., \Delta y_{iT})'$, with similar definitions for Δx_i and Δu_i.

We propose to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ by minimizing the following penalized GMM (PGMM) criterion function[5]

$$Q_{2NT,\lambda_2}^{(K_0)}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = Q_{2,NT}(\boldsymbol{\beta}) + \frac{\lambda_2}{N} \sum_{i=1}^{N} \Pi_{k=1}^{K_0} \|\beta_i - \alpha_k\|, \tag{3.2}$$

[5] We were unable to establish asymptotic theory for the case where the criterion $Q_{2,NT}(\boldsymbol{\beta})$ is replaced by the fully pooled criterion $\tilde{Q}_{2,NT}(\boldsymbol{\beta}) = \left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} z_{it}(\Delta y_{it} - \beta_i'\Delta x_{it})\right]' W_{NT} \left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} z_{it}(\Delta y_{it} - \beta_i'\Delta x_{it})\right]$, where W_{NT} is a $d \times d$ symmetric p.d. matrix. Use of the criterion $Q_{2,NT}(\boldsymbol{\beta})$ means that the PGMM estimator has the oracle property only in some specical cases.

where $Q_{2,NT}(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{T}\sum_{t=1}^{T} z_{it}\left(\Delta y_{it} - \beta_i'\Delta x_{it}\right)\right]' W_{iNT}\left[\frac{1}{T}\sum_{t=1}^{T} z_{it}\left(\Delta y_{it} - \beta_i'\Delta x_{it}\right)\right]$, W_{iNT} is a $d \times d$ matrix that is p.d. asymptotically and $\lambda_2 = \lambda_{2NT}$ is a tuning parameter. Minimizing the above criterion function produces the PGMM estimates $\tilde{\boldsymbol{\alpha}}$ and $\tilde{\boldsymbol{\beta}}$. Let $\tilde{\beta}_i$ and $\tilde{\alpha}_k$ denote the i^{th} and k^{th} columns of $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\alpha}}$, respectively, so that $\tilde{\boldsymbol{\alpha}} \equiv (\tilde{\alpha}_1, ..., \tilde{\alpha}_{K_0})$ and $\tilde{\boldsymbol{\beta}} \equiv (\tilde{\beta}_1, ..., \tilde{\beta}_N)$.

As before, the objective function in (3.2) is convex in α_k but not in $\boldsymbol{\beta}$ when one fixes α_j for $j \neq k$. With minor modifications, the numerical algorithm described in Section 4.2 can be used to obtain the estimates $\tilde{\boldsymbol{\alpha}}$ and $\tilde{\boldsymbol{\beta}}$.

3.2 Preliminary Rates of Convergence for Coefficient Estimates

We first present sufficient conditions to ensure the consistency of $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}})$. Let $\tilde{Q}_{i,z\Delta x} = \frac{1}{T}\sum_{t=1}^{T} z_{it} \times (\Delta x_{it})'$, $\tilde{Q}_{i,z\Delta y} = \frac{1}{T}\sum_{t=1}^{T} z_{it}\Delta y_{it}$, $\bar{Q}_{i,z\Delta x} = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[z_{it}(\Delta x_{it})']$, and $\bar{Q}_{i,z\Delta y} = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[z_{it}\Delta y_{it}]$. Let $\xi_{it} = (\Delta y_{it}, (\Delta x_{it})', z_{it}')'$, $\rho(\xi_{it}, \beta) = z_{it}\left(\Delta y_{it} - \beta'\Delta x_{it}\right)$, and $\bar{\rho}_{i,T}(\beta) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\{\rho(\xi_{it}, \beta) - \mathbb{E}[\rho(\xi_{it}, \beta)]\}$. We make the following assumption.

ASSUMPTION B1. (i) $\mathbb{E}\left[\rho\left(\xi_{it}, \beta_i^0\right)\right] = 0$ *for each* $i = 1, ..., N$ *and* $t = 1, ..., T$.

(ii) $\sup_{\beta\in\mathcal{B}_i}\left\|\bar{\rho}_{i,T}(\beta)\right\| = O_P(1)$ *and* $\frac{1}{N}\sum_{i=1}^{N}\left\|\bar{\rho}_{i,T}(\beta_i)\right\|^2 = O_P(1)$ *for any* $\beta_i \in \mathcal{B}_i$ *and* $i = 1, ..., N$.

(iii) $\tilde{Q}_{i,z\Delta x} = \bar{Q}_{i,z\Delta x} + o_P(1)$ *for each* $i = 1, ..., N$ *and* $\liminf_{(N,T)\to\infty}\min_{1\leq i\leq N}\mu_{\min}(\bar{Q}_{i,z\Delta x}'\bar{Q}_{i,z\Delta x}) = \underline{c}_{\bar{Q}} > 0$.

(iv) *There exist nonrandom matrices* W_i *such that* $\max_{1\leq i\leq N}\|W_{iNT} - W_i\| = o_P(1)$ *and* $\liminf_{N\to\infty}\min_{1\leq i\leq N}\mu_{\min}(W_i) = \underline{c}_W > 0$.

(v) $N_k/N \to \tau_k \in (0,1)$ *for each* $k = 1, ..., K_0$ *as* $N \to \infty$.

(vi) $\lambda_2 \to 0$ *as* $(N,T) \to \infty$.

Assumption B1(i) specifies moment conditions to identify β_i^0. B1(ii) is a high level condition because we do not specify the data structure (or instruments) along with either the cross section or time series dimension. Its first part can generally be verified by applying Donsker's theorem to specific cases. For example, if there exists $\mathcal{F}_{it}$, a σ-field, such that $\{\xi_{it}, \mathcal{F}_{it}\}$ is a stationary ergodic adapted mixingale with size -1 (e.g., White, 2001, pp. 124-125), and $\text{Var}\left(\omega'\bar{\rho}_{i,T}(\beta_i)\right) \to \omega'\Sigma_i\omega \in (0, \infty)$ as $T \to \infty$ for some p.d. matrix Σ_i and any $\omega \in \mathbb{R}^d$ with $\|\omega\| = 1$, then $\bar{\rho}_{i,T}(\beta_i) \xrightarrow{d} N(0, \Sigma_i)$ and the first part of B1(ii) follows. In conjunction with B1(i), B1(iii) provides a rank condition for the identification of β_i^0. It may also be used to establish the mean square convergence of $\tilde{\beta}_i$ as it implicitly requires that $\bar{Q}_{i,z\Delta x}$ is of full rank uniformly in i. B1(iv) is automatically satisfied if one sets $W_{iNT} = I_d$, the $d \times d$ identity matrix. Conditions B1(v)-(vi) parallel the earlier conditions A1(iv)-(v).

Theorem 3.1 *If Assumption B1 holds, then*

(i) $\tilde{\beta}_i - \beta_i^0 = O_P\left(T^{-1/2} + \lambda_2\right)$ *for* $i = 1, ..., N$,

(ii) $\frac{1}{N}\sum_{i=1}^{N}\left\|\tilde{\beta}_i-\beta_i^0\right\|^2=O_P\left(T^{-1}\right)$,
(iii) $\left(\tilde{\alpha}_{(1)},...,\tilde{\alpha}_{(K_0)}\right)-\left(\alpha_1^0,...,\alpha_{K_0}^0\right)=O_P\left(T^{-1/2}\right)$,
where $\left(\tilde{\alpha}_{(1)},...,\tilde{\alpha}_{(K_0)}\right)$ *is a suitable permutation of* $(\tilde{\alpha}_1,...,\tilde{\alpha}_{K_0})$.

REMARK 7. Parts (i) and (ii) of Theorem 3.1 establish the pointwise and mean-square convergence of $\hat{\beta}_i$. Part (iii) indicates that the group-specific parameters $\left\{\alpha_1^0,...,\alpha_K^0\right\}$ can also be estimated consistently by $\{\tilde{\alpha}_1,...,\tilde{\alpha}_{K_0}\}$ subject to permutation. For notational simplicity, hereafter we simply write $\tilde{\alpha}_k$ for $\tilde{\alpha}_{(k)}$ as the consistent estimator of α_k^0, and define

$$\tilde{G}_k=\left\{i\in\{1,2,...,N\}:\tilde{\beta}_i=\tilde{\alpha}_k\right\}\text{ for }k=1,...,K_0. \tag{3.3}$$

3.3 Classification Consistency

Define the following sequences of events:

$$\tilde{E}_{kNT,i}=\left\{i\notin\tilde{G}_k\mid i\in G_k^0\right\}\text{ and }\tilde{F}_{kNT,i}=\left\{i\notin G_k^0\mid i\in\tilde{G}_k\right\}, \tag{3.4}$$

where $i=1,...,N$ and $k=1,...,K_0$. Let $\tilde{E}_{kNT}=\cup_{i\in G_k^0}\tilde{E}_{kNT,i}$ and $\tilde{F}_{kNT}=\cup_{i\in\tilde{G}_k}\tilde{F}_{kNT,i}$. We add the following assumption.

ASSUMPTION B2. (i) $T\lambda_2\to\infty$ *and* $T\lambda_2^4\to c_0\in[0,\infty)$ *as* $(N,T)\to\infty$.
(ii) *For any* $c>0$, $N\max_{1\le i\le N}P\left(\left\|T^{-1}\sum_{t=1}^{T}z_{it}\Delta u_{it}\right\|\ge c\sqrt{\lambda_2}\right)\to 0$ *as* $(N,T)\to\infty$.

Assumptions B2(i)-(ii) parallel A2(i)-(ii). Like the case of A2(ii), one can also verify B2(ii) under some primitive conditions on the process $\{z_{it}\Delta u_{it},t\ge 1\}$. The required moment conditions are now imposed on $\|z_{it}\Delta u_{it}\|$. Following the remark after Assumption A2, for a large range of moment conditions on $\|z_{it}\Delta u_{it}\|$ and the relative rates at which N and T pass to infinity, it suffices to require that

$$\lambda_2\propto T^{-a}\text{ for any }a\in[1/4,1). \tag{3.5}$$

Uniform consistency of the classification is established in the next theorem.

Theorem 3.2 *If Assumptions B1-B2 hold, then*
(i) $P\left(\cup_{k=1}^{K_0}\tilde{E}_{kNT}\right)\le\sum_{k=1}^{K_0}P(\tilde{E}_{kNT})\to 0$ *as* $(N,T)\to\infty$,
(ii) $P\left(\cup_{k=1}^{K_0}\tilde{F}_{kNT}\right)\le\sum_{k=1}^{K_0}P(\tilde{F}_{kNT})\to 0$ *as* $(N,T)\to\infty$.

REMARK 8. Remark 2 also holds for the above theorem with obvious modifications. In particular, let $\tilde{G}_0$ denote the group of individuals in $\{1,2,...,N\}$ that are not classified into any of the K_0 groups, i.e., $\tilde{G}_0=\{1,2,...,N\}\setminus(\cup_{k=1}^{K_0}\tilde{G}_k)$. Define the events $\tilde{H}_{iNT}=\{i\in\tilde{G}_0\}$. Theorem 3.2(i) implies that $P(\cup_{1\le i\le N}\tilde{H}_{iNT})\le\sum_{k=1}^{K_0}P(\tilde{E}_{kNT})\to 0$. That is, all individuals can be classified into one of the K_0 groups w.p.a.1.

Let $\tilde{N}_k=\sum_{i=1}^{N}\mathbf{1}\{i\in\tilde{G}_k\}$. Following the proof of Corollary 2.3, one can also prove that $\tilde{N}_k$ consistently estimates N_k.

Corollary 3.3 *Suppose that Assumptions B1-B2 hold. Then* $\tilde{N}_k - N_k = o_P(1)$.

3.4 Improved Convergence and Asymptotic Properties of Post-Lasso

To obtain an improved rate of convergence for $\{\tilde{\alpha}_k\}$ we provide more specific conditions with the following assumption.

ASSUMPTION B3. (i) *For each* $k = 1, ..., K_0$, $\frac{1}{N_k}\sum_{i\in G_k^0}\left\|\tilde{Q}_{i,z\Delta x} - \bar{Q}_{i,z\Delta x}\right\|^2 = o_P(1)$ and $W_{iNT} \xrightarrow{P} W_i > 0$ for $i \in G_k^0$.

(ii) *For each* $k = 1, ..., K_0$, $\bar{A}_k \equiv \frac{1}{N_k}\sum_{i\in G_k^0}\bar{Q}'_{i,z\Delta x}W_i\bar{Q}_{i,z\Delta x} \to A_k > 0$ *as* $(N,T) \to \infty$.

(iii) *For each* $k = 1, ..., K_0$, $\frac{1}{\sqrt{N_kT}}\sum_{i\in G_k^0}\tilde{Q}'_{i,z\Delta x}W_{iNT}\sum_{t=1}^T z_{it}\Delta u_{it} - B_{kNT} \xrightarrow{D} N(0, C_k)$ *as* $(N,T) \to \infty$.

Assumptions B3(i)-(iii) can be verified under various primitive conditions. For example, B3(i) can be verified by the Markov inequality under (standard) conditions that (a) $\mathbb{E}\|z_{it}(\Delta x_{it})'\|^{2+\sigma} > 0$ for some $\sigma > 0$ and (b) $\{(\Delta x_{it}, z_{it}, \Delta u_{it}), t \geq 1\}$ is strong mixing for each i with mixing coefficients $\alpha_i(\tau)$ that satisfy $\frac{1}{N_k}\sum_{i\in G_k^0}\sum_{\tau=1}^\infty \alpha_i(\tau)^{(2+\sigma)/\sigma} < \infty$. If, in addition, (c) $\{(\Delta x_{it}, z_{it})\}$ is also stationary along the time dimension and IID along the individual dimension for all individuals within the same group G_k^0, and (d) $W_i = W$ for all $i \in G_k^0$, then B3(ii) is satisfied with $A_k = \{\mathbb{E}[z_{it}(\Delta x_{it})']\}'W\mathbb{E}[z_{it}(\Delta x_{it})']$ for any $i \in G_k^0$. To verify B3(iii), for simplicity we assume that $W_{iNT} = I_d$ and make the following decomposition

$$
\begin{aligned}
&\frac{1}{\sqrt{N_kT}}\sum_{i\in G_k^0}\tilde{Q}'_{i,z\Delta x}\sum_{t=1}^T z_{it}\Delta u_{it} \\
= \;& \frac{1}{N_k^{1/2}T^{3/2}}\sum_{i\in G_k^0}\sum_{s=1}^T\sum_{t=1}^T \mathbb{E}\left(\Delta x_{is}z'_{is}z_{it}\Delta u_{it}\right) \\
&+\frac{1}{N_k^{1/2}T^{3/2}}\sum_{i\in G_k^0}\sum_{s=1}^T\sum_{t=1}^T \mathbb{E}\left(\Delta x_{is}z'_{is}\right)z_{it}\Delta u_{it} \\
&+\frac{1}{N_k^{1/2}T^{3/2}}\sum_{i\in G_k^0}\sum_{s=1}^T\sum_{t=1}^T \left\{\left[\Delta x_{is}z'_{is} - \mathbb{E}\left(\Delta x_{is}z'_{is}\right)\right]z_{it}\Delta u_{it} - \mathbb{E}\left(\Delta x_{is}z'_{is}z_{it}\Delta u_{it}\right)\right\} \\
\equiv \;& B_{kNT} + V_{kNT} + R_{kNT}, \text{ say,}
\end{aligned}
\tag{3.6}
$$

where B_{kNT} and V_{kNT} contributes to the asymptotic bias and variance, respectively, and R_{kNT} is a term that is asymptotically negligible under suitable conditions. Then B3(iii) will be satisfied with $W_{iNT} = I_d$ if $V_{kNT} = \frac{1}{N_k^{1/2}T^{1/2}}\sum_{i\in G_k^0}\sum_{t=1}^T \bar{Q}'_{i,z\Delta x}z_{it}\Delta u_{it} \xrightarrow{d} N(0, C_k)$ and $R_{kNT} = o_P(1)$, both of which can be verified by strengthening the conditions in (a)-(c). Note that $\bar{A}_k^{-1}B_{kNT}$ signifies the asymptotic bias of $\tilde{\alpha}_k$, which may not be vanishing asymptotically but can be corrected; see

Appendix D.2.[6]

The following theorem establishes the asymptotic distribution of the C-Lasso estimators $\{\tilde{\alpha}_k\}$.

Theorem 3.4 *Suppose that Assumptions B1-B3 hold. Then* $\sqrt{N_kT}\left(\tilde{\alpha}_k - \alpha_k^0\right) - \bar{A}_k^{-1}B_{kNT} \xrightarrow{D} N(0,\, A_k^{-1}C_kA_k^{-1})$ *for* $k = 1, ..., K_0$.

REMARK 9. In contrast to the PLS case, the PGMM estimators $\{\tilde{\alpha}_k\}$ may fail to possess the oracle property. If the group identities were known in advance, one could obtain the GMM estimate $\breve{\alpha}_k$ of α_k^0 by minimizing the following objective function

$$\tilde{Q}_{NT}\left(\alpha_k\right) = \left[\frac{1}{N_kT}\sum_{i\in G_k^0}\sum_{t=1}^{T} z_{it}\left(\Delta y_{it} - \alpha_k'\Delta x_{it}\right)\right]' W_{NT}^{(k)} \left[\frac{1}{N_kT}\sum_{i\in G_k^0}\sum_{t=1}^{T} z_{it}\left(\Delta y_{it} - \alpha_k'\Delta x_{it}\right)\right], \tag{3.7}$$

where for each $k = 1, ..., K_0$, $W_{NT}^{(k)}$ is a $d \times d$ symmetric positive definite matrix. Let $Q_{z\Delta x,NT}^{(k)} = \frac{1}{N_kT}\sum_{i\in G_k^0}\sum_{t=1}^{T} z_{it}\left(\Delta x_{it}\right)'$ and $Q_{z\Delta y,NT}^{(k)} = \frac{1}{NT}\sum_{i\in G_k^0}\sum_{t=1}^{T} z_{it}\Delta y_{it}$. Then $\breve{\alpha}_k = [Q_{z\Delta x,NT}^{(k)\prime} W_{NT}^{(k)} Q_{z\Delta x,NT}^{(k)}]^{-1}\, Q_{z\Delta x,NT}^{(k)\prime} W_{NT}^{(k)} Q_{z\Delta y,NT}^{(k)}$. We can readily show that the asymptotic distribution of $\breve{\alpha}_k$ is typically different from that of $\tilde{\alpha}_k$ under some regularity conditions. See also the remark after Theorem 3.5 below.

When the individuals have group identities that are unknown, we can replace G_k^0 by its C-Lasso estimate $\tilde{G}_k$ in the GMM objective function (3.7) and obtain the post-Lasso GMM estimator of α_k^0 given by

$$\tilde{\alpha}_{\tilde{G}_k} = \left[\tilde{Q}_{z\Delta x}^{(k)\prime} W_{NT}^{(k)} \tilde{Q}_{z\Delta x}^{(k)}\right]^{-1} \tilde{Q}_{z\Delta x}^{(k)\prime} W_{NT}^{(k)} \tilde{Q}_{z\Delta y}^{(k)}$$

where $\tilde{Q}_{z\Delta x}^{(k)} = \frac{1}{N_kT}\sum_{i\in\tilde{G}_k}\sum_{t=1}^{T} z_{it}\left(\Delta x_{it}\right)'$ and $\tilde{Q}_{z\Delta y}^{(k)} = \frac{1}{NT}\sum_{i\in\tilde{G}_k}\sum_{t=1}^{T} z_{it}\Delta y_{it}$. To study the asymptotic normality of $\tilde{\alpha}_{\tilde{G}_k}$, we add the following assumption.

ASSUMPTION B4. (i) *For each* $k = 1, ..., K_0$, $W_{NT}^{(k)} \xrightarrow{P} W^{(k)} > 0$ as $(N, T) \to \infty$.

(ii) $Q_{z\Delta x,NT}^{(k)} \xrightarrow{P} Q_{z\Delta x}^{(k)}$ *where* $Q_{z\Delta x}^{(k)}$ *has rank* p.

(iii) $\frac{1}{\sqrt{N_kT}}\sum_{i\in G_k^0}\sum_{t=1}^{T} z_{it}\Delta u_{it} \xrightarrow{D} N\left(0, V_k\right)$.

Assumption B4 is standard in the literature on GMM estimation. The assumption can be verified under various primitive conditions that allow for both conditional heteroskedasticity and serial correlation in $\{z_{it}\Delta u_{it}\}$. The following theorem establishes the asymptotic normality of $\{\tilde{\alpha}_{\tilde{G}_k}\}$.

[6] If Conditions (a)-(b) after Assumption B3 are satisfied and $E\left\|z_{it}\Delta u_{it}\right\|^{2+\sigma} > 0$, one can simply apply Davydov's inequality to obtain $\|B_{kNT}\| = \|E\left(\mathcal{B}_{kNT}\right)\| \leq \frac{1}{T\sqrt{N_kT}}\sum_{i\in G_k^0}\sum_{t=1}^{T}\sum_{s=1}^{T}\left\|E\left[\Delta x_{is}z_{is}'z_{it}\Delta u_{it}\right]\right\| = O\left((N/T)^{-1/2}\right)$, which is $o(1)$ if $T \gg N$ and usually not asymptocially negligible otherwise. For general choices of W_{iNT}, it may be difficult to verify Assumption B3(iii).

Theorem 3.5 *Suppose that Assumptions B1-B4 hold. Then* $\sqrt{N_k T}(\tilde{\alpha}_{\tilde{G}_k} - \alpha_k^0) \xrightarrow{D} N(0, \Omega_k)$ *where* $\Omega_k = \left[Q_{z\Delta x}^{(k)\prime} W^{(k)} Q_{z\Delta x}^{(k)}\right]^{-1} Q_{z\Delta x}^{(k)\prime} W^{(k)} V_k W^{(k)} Q_{z\Delta x}^{(k)} \left[Q_{z\Delta x}^{(k)\prime} W^{(k)} Q_{z\Delta x}^{(k)}\right]^{-1}$ *and* $k = 1, ..., K_0$.

REMARK 10. As in the proof of Theorem 2.5, one can apply Theorem 3.2 and demonstrate that

$$\sqrt{N_k T}\left(\tilde{\alpha}_{\tilde{G}_k} - \alpha_k^0\right) = \sqrt{N_k T}\left(\breve{\alpha}_k - \alpha_k^0\right) + o_P(1).$$

That is, the post-Lasso GMM estimator $\tilde{\alpha}_{\tilde{G}_k}$ is asymptotically equivalent to the infeasible estimate $\breve{\alpha}_k$ which an oracle could obtain with knowledge of each individual's group identity. To obtain the most efficient estimator among the class of GMM estimators based on the moment conditions specified in Assumption B1(i), one can set $W_{NT}^{(k)}$ to be a consistent estimator of V_k^{-1}. The procedure is standard and we omit the details for brevity.

REMARK 11. If $W_{iNT} = W_{NT}^{(k)}$, $\bar{Q}_{i,z\Delta x} = Q_{z\Delta x}^{(k)}$ for each $i \in G_k^0$ in Assumptions B3(i)-(ii), and $B_{kNT} = 0$ in Assumption B3(iii), then $A_k = Q_{z\Delta x}^{(k)\prime} W^{(k)} Q_{z\Delta x}^{(k)}$, $C_k = Q_{z\Delta x}^{(k)\prime} W^{(k)} \Omega_k W^{(k)} Q_{z\Delta x}^{(k)}$, and $\sqrt{N_k T}\left(\tilde{\alpha}_k - \alpha_k^0\right) \xrightarrow{D} N(0, \Omega_k)$. That is, in this special case, the C-Lasso estimator $\tilde{\alpha}_k$ also has the oracle property. But as remarked before, $B_{kNT} = 0$ would typically require $T \gg N$, a condition that we do not usually want to impose. For this reason, we recommend the post-Lasso estimator $\tilde{\alpha}_{\tilde{G}_k}$ for the general case.[7]

3.5 Determination of the Number of Groups

When the true number of groups K_0 is unknown, we continue to assume that it is bounded from above by a finite integer $K_{\max}$. We consider the following PGMM criterion function

$$Q_{2NT,\lambda_2}^{(K)}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = Q_{2,NT}(\boldsymbol{\beta}) + \frac{\lambda_2}{N} \sum_{i=1}^{N} \Pi_{k=1}^{K} \|\beta_i - \alpha_k\|, \tag{3.8}$$

where $1 \leq K \leq K_{\max}$. Minimizing the above objective function, we obtain the C-Lasso estimates $\left\{\tilde{\beta}_i(K, \lambda_2), \tilde{\alpha}_k(K, \lambda_2)\right\}$ of $\{\beta_i, \alpha_k\}$, where we make the dependence of $\tilde{\beta}_i$ and $\tilde{\alpha}_k$ on (K, λ_2) explicit. As above, we classify individual i into group $\tilde{G}_k(K, \lambda_2)$ if and only if $\tilde{\beta}_i(K, \lambda_2) = \tilde{\alpha}_k(K, \lambda_2)$, i.e.,

$$\tilde{G}_k(K, \lambda_2) = \left\{i \in \{1, 2, ..., N\} : \tilde{\beta}_i(K, \lambda_2) = \tilde{\alpha}_k(K, \lambda_2)\right\} \text{ for } k = 1, ..., K. \tag{3.9}$$

Let $\tilde{G}(K, \lambda_1) = \{\tilde{G}_1(K, \lambda_1), ..., \tilde{G}_K(K, \lambda_1)\}$. Based on (3.9), we define the post-Lasso GMM estimate of α_k^0 by

$$\tilde{\alpha}_{\tilde{G}_k(K,\lambda_2)} = \left[\tilde{Q}_{z\Delta x}^{(K,k)\prime} W_{NT}^{(k)} \tilde{Q}_{z\Delta x}^{(K,k)}\right]^{+} \tilde{Q}_{z\Delta x}^{(K,k)\prime} W_{NT}^{(k)} \tilde{Q}_{z\Delta y}^{(K,k)}, \tag{3.10}$$

[7] Of course one cannot choose W_{iNT} to be group-specific (i.e., $W_{NT}^{(k)}$) because we do not know the group structure.

where $\tilde{Q}_{z\Delta x}^{(K,k)} = \frac{1}{N_k T}\sum_{i\in\tilde{G}_k(K,\lambda_2)}\sum_{t=1}^{T} z_{it}\left(\Delta x_{it}\right)'$, $\tilde{Q}_{z\Delta y}^{(K,k)} = \frac{1}{N_k T}\sum_{i\in\tilde{G}_k(K,\lambda_2)}\sum_{t=1}^{T} z_{it}\Delta y_{it}$, and $W_{NT}^{(k)}$ is defined as before but with $k = 1, 2, ..., K$.

Let $\tilde{\sigma}^2_{\tilde{G}(K,\lambda_2)} = \frac{1}{NT}\sum_{k=1}^{K}\sum_{i\in\tilde{G}_k(K,\lambda_2)}\sum_{t=1}^{T}[\Delta y_{it} - \tilde{\alpha}'_{\tilde{G}_k(K,\lambda_1)}\Delta x_{it}]^2$. We propose to select K to minimize the following information criterion:

$$IC_2\left(K,\lambda_2\right) = \ln\left[\tilde{\sigma}^2_{\tilde{G}(K,\lambda_2)}\right] + \rho_{2NT}pK,$$

where ρ_{2NT} is a tuning parameter. As before, for any $G^{(K)} = (G_{K,1}, ..., G_{K,K}) \in \mathcal{G}_K$, define $\tilde{\sigma}^2_{G^{(K)}} = \frac{1}{NT}\sum_{k=1}^{K}\sum_{i\in G_{K,k}}\sum_{t=1}^{T}[\Delta y_{it} - \tilde{\alpha}'_{G_{K,k}}\Delta x_{it}]^2$, where $\tilde{\alpha}_{G_{K,k}}$ is analogously defined as $\tilde{\alpha}_{\tilde{G}_k(K,\lambda_2)}$ with $\tilde{G}_k\left(K,\lambda_2\right)$ being replaced by $G_{K,k}$.

To proceed, we add the following two assumptions.

ASSUMPTION B5. *As* $(N,T)\to\infty$, $\min_{1\le K<K_0}\inf_{G^{(K)}\in\mathcal{G}_K}\tilde{\sigma}^2_{G^{(K)}} \xrightarrow{P} \underline{\sigma}^2_{\Delta u} > \sigma^2_{\Delta u}$, *where* $\sigma^2_{\Delta u} = \operatorname{plim}_{(N,T)\to\infty}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\Delta u_{it}\right)^2$.

ASSUMPTION B6. *As* $(N,T)\to\infty$, $\rho_{2NT}\to 0$ *and* $\rho_{2NT}NT\to\infty$.

Assumptions B5-B6 parallel earlier Assumptions A4-A5. The following theorem proves consistency of this choice of K as the minimizer of $IC_2\left(K,\lambda_2\right)$ with respect to K.

Theorem 3.6 *Suppose that Assumptions B1-B2 and B4-B6 hold. Then*

$$P\left(\inf_{K\in\mathcal{K}_-\cup\mathcal{K}_+} IC_2\left(K,\lambda_2\right) > IC_2\left(K_0,\lambda_2\right)\right)\to 1 \text{ as } (N,T)\to\infty.$$

REMARK 12. The remark after Theorem 2.6 also holds here after obvious modifications. To obtain a data-driven choice of the tuning parameter λ_2, define

$$K\left(\lambda_2\right) = \arg\min_K IC_2\left(K,\lambda_2\right) \text{ and } IC_2^*\left(\lambda_2\right) = IC_2\left(K\left(\lambda_2\right),\lambda_2\right).$$

We can select the tuning parameter as $\hat{\lambda}_2 = \arg\min_{\lambda_2\in\Lambda_2} IC_2^*\left(\lambda_2\right)$, where $\Lambda_2 = \{\lambda_2 \propto T^{-a}$ for some $a\in[1/4,1)\}$ provided some conditions on the moments of $\|z_{it}\Delta u_{it}\|$ and on the relative rates at which N and T pass to infinity are satisfied. See the remarks after Assumptions A2 and B2.

4 Simulation

In this section, we evaluate the finite-sample performance of the C-Lasso and the post-Lasso estimates.

4.1 Data Generating Processes

We consider three data generating processes (DGPs) that cover static and dynamic panels. Throughout these DGPs, the fixed effect μ_i and the idiosyncratic error u_{it} follow the standard normal distribution and are mutually independent all across i and t. The observations in each DGP are drawn from three groups with the proportion $N_1 : N_2 : N_3 = 0.3 : 0.3 : 0.4$. We try six combinations of the sample sizes with $N = 100$, 200 and $T = 10$, 20, 40.

DGP 1 (Static panel with two exogenous regressors.) The observations (y_{it}, x_{it}) are generated from the panel structure model (2.1) where $x_{it} = (x_{it1}, x_{it2})'$, $x_{it1} = 0.2\mu_i + e_{it1}$, $x_{it2} = 0.2\mu_i + e_{it2}$, and e_{it1} and e_{it2} are both IID $N(0,1)$ and mutually independent. The true coefficients are

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

DGP 2 (Static panel with endogeneity.) We maintain the panel structure model (2.1) with two regressors in x_{it}. $x_{it2} \sim N(0,1)$ is independent of the idiosyncratic shock u_{it} while x_{it1} is generated from the following underlying reduced-form equation: $x_{it1} = 0.2\mu_i + 0.5z_{it1} + 0.5z_{it2} + 0.5e_{it}$, where z_{it1} and z_{it2}, the two excluded instrumental variables, are each IID $N(0,1)$, mutually independent, and independent of u_{it} and e_{it}. Endogeneity arises since the reduced-form error term e_{it} and the structural-equation idiosyncratic shock u_{it} follow a bivariate normal distribution:

$$\begin{pmatrix} u_{it} \\ e_{it} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix} \right).$$

The econometrician observes (y_{it}, x_{it}, z_{it}) with $x_{it} = (x_{it1}, x_{it2})'$ and $z_{it} = (z_{it1}, z_{it2})'$. The true coefficients are

$$\left(\alpha_1^0, \alpha_2^0, \alpha_3^0\right) = \left(\begin{pmatrix} 0.2 \\ 1.8 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1.8 \\ 0.2 \end{pmatrix} \right).$$

We set the gaps between the groups of the coefficients larger than those in DGP1 to compensate for the weaker signal strength caused by instrumentation.

DGP 3 (Panel AR(1) with two exogenous regressors.) The model is

$$y_{it} = \beta_{i1}^0 y_{i,t-1} + \beta_{i2}^0 x_{it2} + \beta_{i3}^0 x_{it3} + \mu_i(1 - \beta_{i1}^0) + u_{it},$$

where x_{it2} and x_{it3} are two exogenous regressors. They follow the standard normal distributions, mutually independent, and are independent of the error term. For each i, the initial

value is $y_{i0} = \beta_{i2}^0 x_{i02} + \beta_{i3}^0 x_{i03} + \mu_i + u_{i0}$ so that the i-th time series is strictly stationary with mean μ_i. The true coefficients are

$$\left(\alpha_1^0, \alpha_2^0, \alpha_3^0\right) = \left(\begin{pmatrix} 0.8 \\ 0.4 \\ 0.4 \end{pmatrix}, \begin{pmatrix} 0.6 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 0.4 \\ 1.6 \\ 1.6 \end{pmatrix} \right).$$

The choices of the lag term coefficients represent strong, moderate, and weak persistence, respectively. The choices of the coefficients of the exogenous regressors balance the different signal strength that stems from the dynamic structure.

4.2 Numerical Algorithm

The numerical operation of C-Lasso is high-dimensional. Here we propose an iterative algorithm to obtain the PLS estimates $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ in Section 2. A similar algorithm applies for PGMM estimation.

1. Start with arbitrary initial values $\hat{\boldsymbol{\alpha}}^{(0)} = (\hat{\alpha}_1^{(0)}, ..., \hat{\alpha}_{K_0}^{(0)})$ and $\hat{\boldsymbol{\beta}}^{(0)} = (\hat{\beta}_1^{(0)}, ..., \hat{\beta}_N^{(0)})$ such that $\sum_{i=1}^N ||\hat{\beta}_i^{(0)} - \hat{\alpha}_k^{(0)}|| \neq 0$ for each $k = 2, ..., K_0$.[8]

2. Having obtained $\hat{\boldsymbol{\alpha}}^{(r-1)} \equiv (\hat{\alpha}_1^{(r-1)}, ..., \hat{\alpha}_{K_0}^{(r-1)})$ and $\hat{\boldsymbol{\beta}}^{(r-1)} \equiv (\hat{\beta}_1^{(r-1)}, ..., \hat{\beta}_N^{(r-1)})$, in step $r \geq 1$, we first choose $(\boldsymbol{\beta}, \alpha_1)$ to minimize

$$Q_{K_0 NT}^{(r,1)}(\boldsymbol{\beta}, \alpha_1) = Q_{1,NT}(\boldsymbol{\beta}) + \frac{\lambda_1}{N} \sum_{i=1}^N \|\beta_i - \alpha_1\| \, \Pi_{k \neq 1}^{K_0} \left\| \hat{\beta}_i^{(r-1)} - \hat{\alpha}_k^{(r-1)} \right\|,$$

and obtain the updated estimate $(\hat{\boldsymbol{\beta}}^{(r,1)}, \hat{\alpha}_1^{(r)})$ of $(\boldsymbol{\beta}, \alpha_1)$. Next choose $(\boldsymbol{\beta}, \alpha_2)$ to minimize

$$Q_{K_0 NT}^{(r,2)}(\boldsymbol{\beta}, \alpha_2) = Q_{1,NT}(\boldsymbol{\beta}) + \frac{\lambda_1}{N} \sum_{i=1}^N \|\beta_i - \alpha_2\| \left\| \hat{\beta}_i^{(r,1)} - \hat{\alpha}_1^{(r)} \right\| \Pi_{k \neq 1,2}^{K_0} \left\| \hat{\beta}_i^{(r-1)} - \hat{\alpha}_k^{(r-1)} \right\|$$

to obtain the updated estimate $(\hat{\boldsymbol{\beta}}^{(r,2)}, \hat{\alpha}_2^{(r)})$ of $(\boldsymbol{\beta}, \alpha_2)$. Repeat this procedure until $(\boldsymbol{\beta}, \alpha_{K_0})$ is chosen to minimize

$$Q_{K_0 NT}^{(r,K_0)}(\boldsymbol{\beta}, \alpha_{K_0}) = Q_{1,NT}(\boldsymbol{\beta}) + \frac{\lambda_1}{N} \sum_{i=1}^N \|\beta_i - \alpha_{K_0}\| \, \Pi_{k=1}^{K_0 - 1} \left\| \hat{\beta}_i^{(r,k)} - \hat{\alpha}_k^{(r)} \right\|$$

[8] Under the condition that T diverges to the infinity, we can obtain the preliminary consistent estimate $\hat{\beta}_i^{(0)}$ as $\hat{\beta}_i^{OLS}$. In the simulations, we always set $\hat{\alpha}_k^{(0)} = \tilde{\alpha}_k^{(0)} = 0$ and $\{\hat{\beta}_i^{(0)}\}_{i=1}^N$ or $\{\tilde{\beta}_i^{(0)}\}_{i=1}^N$ to be the within-group estimates. We experimented with $\hat{\beta}_i^{(0)} = \tilde{\beta}_i^{(0)} = 1$ for all i and $\hat{\alpha}_k^{(0)} = \tilde{\alpha}_k^{(0)} = 0$ for all k. The latter choice delivers similar classification and estimation results. This suggests that the algorithm is insensitive to the initial value under sensible choices, although the high-dimensionality hinders a straightforward visualization of the shapes of the objective functions against the parameters.

to obtain the updated estimate $(\hat{\boldsymbol{\beta}}^{(r,K_0)},\ \hat{\alpha}_{K_0}^{(r)})$ of $(\boldsymbol{\beta},\alpha_{K_0})$. Let $\hat{\boldsymbol{\beta}}^{(r)}=\hat{\boldsymbol{\beta}}^{(r,K_0)}$ and $\hat{\boldsymbol{\alpha}}^{(r)}=(\hat{\alpha}_1^{(r)},...,\hat{\alpha}_{K_0}^{(r)})$.

3. Repeat step 2 until a convergence criterion is met, e.g., when

$$\frac{\sum_{i=1}^N\left\|\hat{\beta}_i^{(r)}-\hat{\beta}_i^{(r-1)}\right\|^2}{\sum_{i=1}^N\left\|\hat{\beta}_i^{(r-1)}\right\|^2+0.0001}<\epsilon_{tol}\text{ and }\frac{\sum_{k=1}^{K_0}\left\|\hat{\alpha}_k^{(r)}-\hat{\alpha}_k^{(r-1)}\right\|^2}{\sum_{k=1}^{K_0}\left\|\hat{\alpha}_k^{(r-1)}\right\|^2+0.0001}<\epsilon_{tol},$$

where ϵ_{tol} is some prescribed tolerance level (e.g., 0.0001). Define the final iterative estimate of $\boldsymbol{\alpha}$ as $\hat{\boldsymbol{\alpha}}=(\hat{\alpha}_1^{(R)},...,\hat{\alpha}_{K_0}^{(R)})$ for sufficiently large R such that the convergence criterion is met. The final iterative estimate of $\boldsymbol{\beta}$ is defined as $\hat{\boldsymbol{\beta}}=(\hat{\beta}_1,...,\hat{\beta}_N)$ where

$$\begin{aligned}\hat{\beta}_i &= \sum_{k=1}^{K_0}\hat{\alpha}_k^{(R)}\mathbf{1}\left\{\hat{\beta}_i^{(R,l)}=\hat{\alpha}_k^{(R)}\text{ for some }l=1,...,K_0\right\}\\ &\quad+\hat{\beta}_i^{(R,K_0)}\left[1-\sum_{k=1}^{K_0}\mathbf{1}\left\{\hat{\beta}_i^{(R,l)}=\hat{\alpha}_k^{(R)}\text{ for some }l=1,...,K_0\right\}\right]\end{aligned}\tag{4.1}$$

where $\hat{\beta}_i^{(R,l)}$ denotes the i^{th} column of $\hat{\boldsymbol{\beta}}^{(R,l)}$ for $l=1,2,...,K$. Intuitively, individual i is classified to group $\hat{G}_k$ if $\hat{\beta}_i^{(R,l)}=\hat{\alpha}_k^{(R)}$ for some $l=1,...,K_0$; otherwise it is left unclassified so that $\hat{\beta}_i$ is defined as $\hat{\beta}_i^{(R,K_0)}$.

Obviously, each iteration step r has K_0 substeps and we can use $r.k$ to denote substep k within step r. Note the objective function $Q_{K_0NT}^{(r,k)}(\boldsymbol{\beta},\alpha_k)$ is convex in $(\boldsymbol{\beta},\alpha_k)$ in each substep $r.k$. So the above iteration procedure has fast implementation in practice. Moreover, in view of the fact that

$$Q_{1NT,\lambda_1}^{(K_0)}(\hat{\boldsymbol{\beta}}^{(r-1)},\hat{\boldsymbol{\alpha}}^{(r-1)})\geq Q_{K_0NT}^{(r,1)}(\hat{\boldsymbol{\beta}}^{(r,1)},\hat{\alpha}_1^{(r)})\geq\cdots\geq Q_{K_0NT}^{(r,1)}(\hat{\boldsymbol{\beta}}^{(r,K_0)},\hat{\alpha}_K^{(r)})=Q_{1NT,\lambda_1}^{(K_0)}(\hat{\boldsymbol{\beta}}^{(r)},\hat{\boldsymbol{\alpha}}^{(r)}),$$

the convergence of $(\hat{\boldsymbol{\beta}}^{(r)},\hat{\boldsymbol{\alpha}}^{(r)})$ is readily established and simulations confirm that convergence is rapid, usually occurring after just a few iterations.

We will estimate the parameters in DGP 1 with PLS, in DGP 2 with PGMM, and in DGP 3 with both PLS and PGMM. The bias is corrected via the one-sided kernel as discussed in Appendix D.1 and D.2 with a tuning parameter $M_T=2\times\lfloor T^{1/4}\rfloor$, where $\lfloor a\rfloor$ denotes the integer part of a real number a. In DGP 3 PGMM uses $(y_{i,t-2},y_{i,t-3},\Delta x_{it2},\Delta x_{it3})$ as the instruments for $(\Delta y_{i,t-1},\Delta x_{it2},\Delta x_{it3})$ in the first-differenced model.

4.3 Determination of the Number of Groups

Since classification consistency and the oracle property both hinge on the correct number of groups, our first simulation exercise is designed to assess how well the proposed information

Table 1: Frequency of selecting $K = 1, 2, ..., 6$ groups

	N	T	1	2	3	4	5	6
DGP1	100	10	0.000	0.040	0.718	0.230	0.012	0.000
PLS	100	20	0.000	0.002	0.994	0.004	0.000	0.000
	100	40	0.000	0.000	1.000	0.000	0.000	0.000
	200	10	0.000	0.000	0.428	0.468	0.104	0.000
	200	20	0.000	0.000	0.982	0.018	0.000	0.000
	200	40	0.000	0.000	1.000	0.000	0.000	0.000
DGP2	100	10	0.000	0.448	0.518	0.032	0.002	0.000
PGMM	100	20	0.000	0.006	0.914	0.076	0.004	0.000
	100	40	0.000	0.000	0.992	0.008	0.000	0.000
	200	10	0.000	0.244	0.736	0.020	0.000	0.000
	200	20	0.000	0.000	0.962	0.034	0.002	0.002
	200	40	0.000	0.000	0.988	0.012	0.000	0.000
DGP3	100	10	0.000	0.472	0.518	0.010	0.000	0.000
PLS	100	20	0.000	0.098	0.902	0.000	0.000	0.000
	100	40	0.000	0.000	1.000	0.000	0.000	0.000
	200	10	0.000	0.090	0.856	0.050	0.004	0.000
	200	20	0.000	0.002	0.996	0.002	0.000	0.000
	200	40	0.000	0.000	1.000	0.000	0.000	0.000
DGP3	100	10	0.000	0.242	0.614	0.136	0.008	0.000
PGMM	100	20	0.000	0.008	0.908	0.076	0.008	0.000
	100	40	0.000	0.000	0.996	0.004	0.000	0.000
	200	10	0.000	0.078	0.754	0.150	0.018	0.000
	200	20	0.000	0.000	0.908	0.090	0.002	0.000
	200	40	0.000	0.000	0.998	0.002	0.000	0.000

criteria in Sections 2.6 and 3.5 perform in selecting the number of groups. Asymptotically, all sequences ρ_{1NT} work if they satisfy Assumption A5, and so do the sequences ρ_{2NT} if these satisfy Assumption B6. In practice, the choice of ρ_{jNT}, $(j = 1, 2)$ can be crucial. Our findings indicate that use of the Bayesian information criterion (BIC) $\rho_{jNT} = (NT)^{-1} \ln(NT)$ is too small for group number selection. We experimented with alternatives and found that $\rho_{jNT} = \frac{2}{3}(NT)^{-1/2}$ $(j = 1, 2)$ works fairly well for the determination of the number of groups and this setting is used throughout the simulations as well as the empirical application.

Based on 500 replications for each DGP, Table 1 displays the empirical probability that a particular group size from 1 to 6 is selected according to the information criteria. Due to space limitations, we report outcomes under the tuning parameter $\lambda_j = 1 \times s_Y^2 T^{-1/2}$ for $j = 1, 2$, where s_Y^2 is the sample variance of $\tilde{y}_{it}$ for PLS or the sample variance of Δy_{it} for PGMM. The results are found to be robust for a reasonable range of values of the tuning parameter, as will be seen in the the following subsection on point estimation and in the empirical application. Recall that the true number is 3. When $T = 10$, the correct choice probabilities vary across the three DGPs and the two penalized methods. These probabilities rise to more than 90% in all cases when $T = 20$ and tend to unity when $T = 40$. Some intuitive graphics demonstrating how well the information criteria work in these simulations can be found in the supplemental Appendix E.

Table 2: Results of Classification

	C_λ		0.2		0.4		0.8		1.6		3.2	
	N	T	$\bar{P}(\hat{E})$	$\bar{P}(\hat{F})$	$\bar{P}(\hat{E})$	$\bar{P}(\hat{F})$	$\bar{P}(\hat{E})$	$\bar{P}(\hat{F})$	$\bar{P}(\hat{E})$	$\bar{P}(\hat{F})$	$\bar{P}(\hat{E})$	$\bar{P}(\hat{F})$
DGP1	100	10	0.1805	0.0901	0.1899	0.0954	0.2236	0.1115	0.2777	0.1305	0.4216	0.1897
PLS	100	20	0.0593	0.0289	0.0585	0.0292	0.0576	0.0290	0.0805	0.0396	0.1304	0.0598
	100	40	0.0103	0.0049	0.0098	0.0046	0.0093	0.0045	0.0094	0.0048	0.0149	0.0070
	200	10	0.1691	0.0848	0.1771	0.0894	0.2097	0.1054	0.2766	0.1322	0.3976	0.1746
	200	20	0.0586	0.0284	0.0556	0.0275	0.0552	0.0277	0.0719	0.0362	0.1338	0.0613
	200	40	0.0092	0.0044	0.0083	0.0040	0.0081	0.0039	0.0078	0.0040	0.0141	0.0066
DGP2	100	10	0.2082	0.0993	0.2001	0.0974	0.2024	0.1004	0.2145	0.1076	0.2527	0.1274
PGMM	100	20	0.1027	0.0485	0.0958	0.0462	0.0888	0.0437	0.0878	0.0440	0.0996	0.0504
	100	40	0.0321	0.0152	0.0307	0.0147	0.0266	0.0130	0.0230	0.0115	0.0227	0.0116
	200	10	0.2037	0.0980	0.1982	0.0971	0.1968	0.0984	0.2113	0.1071	0.2482	0.1257
	200	20	0.1020	0.0483	0.0942	0.0456	0.0872	0.0432	0.0841	0.0424	0.0942	0.0480
	200	40	0.0332	0.0158	0.0299	0.0144	0.0266	0.0130	0.0222	0.0111	0.0212	0.0109
DGP3	100	10	0.2063	0.1038	0.1839	0.0908	0.1913	0.0937	0.2305	0.1092	0.4058	0.1715
PLS	100	20	0.1000	0.0501	0.0826	0.0404	0.0750	0.0357	0.0800	0.0391	0.1968	0.0886
	100	40	0.0277	0.0137	0.0222	0.0106	0.0183	0.0085	0.0158	0.0072	0.0373	0.0177
	200	10	0.2025	0.1026	0.1714	0.0853	0.1709	0.0844	0.2079	0.0998	0.3539	0.1498
	200	20	0.0983	0.0490	0.0794	0.0386	0.0703	0.0333	0.0716	0.0347	0.1451	0.0657
	200	40	0.0255	0.0126	0.0209	0.0100	0.0173	0.0080	0.0151	0.0069	0.0220	0.0103
DGP3	100	10	0.3173	0.1566	0.2991	0.1482	0.2924	0.1437	0.3016	0.1471	0.3379	0.1650
PGMM	100	20	0.1688	0.0833	0.1525	0.0753	0.1405	0.0683	0.1335	0.0629	0.1422	0.0665
	100	40	0.0729	0.0355	0.059	0.029	0.0495	0.0239	0.0436	0.0203	0.0421	0.0189
	200	10	0.3151	0.1557	0.2919	0.1449	0.2789	0.1381	0.2876	0.1415	0.3243	0.1597
	200	20	0.1714	0.0847	0.1503	0.0745	0.1345	0.0655	0.1288	0.0609	0.1363	0.0638
	200	40	0.0731	0.0356	0.0575	0.0284	0.0486	0.0236	0.0426	0.0199	0.0406	0.0183

4.4 Classification and Point Estimation

The results from the previous section show that the information criteria are useful when it is now known *a priori* how many groups exist in the panel. This section now focuses on classification and estimation performance under the true number of groups. Here the tuning parameter λ_j is set to be $C_{\lambda_j} s_Y^2 T^{-1/2}$ for $j = 1, 2$, where C_{λ_j} is a sequence of geometrically increasing constants. Five values $\{0.2, 0.4, 0.8, 1.6, 3.2\}$ are used for $C_\lambda = C_{\lambda_j}$.

Table 2 reports the classification results from 500 replications. As discussed in Remark 2, we classify all observations into the group whose $\hat{\alpha}_k$ is the closest to $\hat{\beta}_i$. We summarize the pointwise classification error using averages over $i = 1, \ldots, N$, as there is no space to report results for each individual. The values reported in the table are the means of the *average* classification errors $\bar{P}(\hat{E}) = \frac{1}{N}\sum_{i=1}^{N} \hat{P}(\hat{E}_{kNT,i})$ and $\bar{P}(\hat{F}) = \frac{1}{N}\sum_{i=1}^{N} \hat{P}(\hat{F}_{kNT,i})$ where $\hat{P}$ denotes the empirical mean over the replications.

Table 2 shows that the classification errors quickly shrink towards 0 as T increases. The results are not sensitive to the choice of the tuning parameter via C_λ. In particular, when $T = 40$ the PLS classification errors $\bar{P}(\hat{E})$ and $\bar{P}(\hat{F})$ typically take on values 0.5–3%, and PGMM classification errors are also small. In DGP 3 PLS appears to be more accurate than PGMM.

Table 3: Estimation of $\beta_{i,1}$ in DGP 1 by PLS

		C_λ	0.2		0.4		0.8		1.6		3.2	
N	T		RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias
100	10	C-Lasso	0.1010	0.0364	0.1116	0.0364	0.1303	0.0293	0.1780	-0.0150	0.3206	-0.0968
		Post-lasso	0.0907	0.0282	0.1035	0.0293	0.1274	0.0254	0.1788	-0.0162	0.3216	-0.0984
		Oracle	0.0583	-0.0033	0.0583	-0.0033	0.0583	-0.0033	0.0583	-0.0033	0.0583	-0.0033
100	20	C-Lasso	0.0590	0.0154	0.0560	0.0183	0.0507	0.0154	0.0690	0.0054	0.0856	0.0012
		Post-lasso	0.0450	0.0066	0.0467	0.0092	0.0470	0.0090	0.0687	0.0038	0.0846	0.0012
		Oracle	0.0399	-0.0021	0.0399	-0.0021	0.0399	-0.0021	0.0399	-0.0021	0.0399	-0.0021
100	40	C-Lasso	0.0347	0.0096	0.0348	0.0047	0.0305	0.0053	0.0301	0.0023	0.0347	0.0011
		Post-lasso	0.0292	0.0012	0.0293	0.0002	0.0291	0.0010	0.0290	0.0008	0.0337	0.0010
		Oracle	0.0281	-0.0010	0.0281	-0.0010	0.0281	-0.0010	0.0281	-0.0010	0.0281	-0.0010
200	10	C-Lasso	0.0767	0.0312	0.0856	0.0319	0.1017	0.0256	0.1457	-0.0004	0.3127	-0.0985
		Post-lasso	0.0630	0.0225	0.0759	0.0237	0.0963	0.0210	0.1441	-0.0009	0.3137	-0.1001
		Oracle	0.0410	0.0019	0.0410	0.0019	0.0410	0.0019	0.0410	0.0019	0.0410	0.0019
200	20	C-Lasso	0.0491	0.0152	0.0424	0.0151	0.0366	0.0137	0.0501	0.0102	0.0930	-0.0032
		Post-lasso	0.0320	0.0056	0.0327	0.0067	0.0329	0.0077	0.0473	0.0089	0.0916	-0.0031
		Oracle	0.0280	0.0007	0.0280	0.0007	0.0280	0.0007	0.0280	0.0007	0.0280	0.0007
200	40	C-Lasso	0.0276	0.0122	0.0259	0.0048	0.0222	0.0062	0.0210	0.0036	0.0233	0.0016
		Post-lasso	0.0204	0.0023	0.0203	0.0012	0.0202	0.0018	0.0204	0.0021	0.0222	0.0016
		Oracle	0.0193	0.0004	0.0193	0.0004	0.0193	0.0004	0.0193	0.0004	0.0193	0.0004

We next discuss point estimation. Tables 3–6 show the root-mean-squared error (RMSE) and the bias of the estimates of the first element $\beta_{i,1}$ in β_i in each model.[9] Since each DGP has three groups of different coefficients, the outcomes of the coefficient estimation are not directly comparable across groups. For brevity we weight the RMSEs and the biases by their proportion in the population. For example, $\text{RMSE}(\hat{\beta}_1)$ is calculated as $\frac{1}{N}\sum_{k=1}^{K_0} N_k \text{RMSE}(\hat{\alpha}_{k,1})$ with $\hat{\alpha}_{k,1}$ being the first element in $\hat{\alpha}_k$, and so is the bias.

The findings in the tables reveal the following general pattern. First, the RMSEs and biases of the estimators shrink toward zero when T increases and N remains fixed. Second, post-Lasso generally outperforms C-Lasso. Third, bias correction works in the right direction. The finite-sample performance of the post-Lasso PLS is close to that of the oracle estimator, which demonstrates the practical relevance of the oracle property. The RMSE of post-Lasso generally remains the smallest in comparison with C-Lasso and bias-corrected C-Lasso in PGMM, in which the oracle property is missing. Based on these findings we recommend the post-Lasso estimator for practical use.

5 Empirical Application

Understanding the disparate savings behavior across countries is a longstanding research interest in development economics. Theoretical advances and empirical studies have accumulated over

[9] Results for estimation of the other coefficients are available upon request.

Table 4: Estimation of $\beta_{i,1}$ in DGP 2 by PGMM

		C_λ	0.2		0.4		0.8		1.6		3.2	
N	T		RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias
100	10	C-Lasso	0.1906	0.1093	0.1907	0.1242	0.2018	0.1388	0.2096	0.1490	0.2220	0.1581
		Post-lasso	0.1416	0.0152	0.1368	0.0251	0.1413	0.0325	0.1421	0.0381	0.1533	0.0443
		C-Lasso BC	0.1603	0.0684	0.1586	0.0811	0.1679	0.0928	0.1737	0.1009	0.1858	0.1085
		Oracle	0.0993	-0.0001	0.0993	-0.0001	0.0993	-0.0001	0.0993	-0.0001	0.0993	-0.0001
100	20	C-Lasso	0.1179	0.0560	0.1176	0.0683	0.1182	0.0799	0.1239	0.0898	0.1321	0.0985
		Post-lasso	0.0838	0.0138	0.0815	0.0181	0.0810	0.0200	0.0826	0.0212	0.0871	0.0216
		C-Lasso BC	0.0986	0.0374	0.0978	0.0464	0.0986	0.0539	0.1021	0.0600	0.1083	0.0652
		Oracle	0.0680	-0.0004	0.0680	-0.0004	0.0680	-0.0004	0.0680	-0.0004	0.0680	-0.0004
100	40	C-Lasso	0.0712	0.0400	0.0754	0.0422	0.0761	0.0464	0.0753	0.0504	0.0772	0.0557
		Post-lasso	0.0519	0.0136	0.0522	0.0129	0.0519	0.0122	0.0516	0.0112	0.0522	0.0108
		C-Lasso BC	0.0614	0.0274	0.0632	0.0282	0.0637	0.0301	0.0634	0.0317	0.0645	0.0343
		Oracle	0.0492	0.0007	0.0492	0.0007	0.0492	0.0007	0.0492	0.0007	0.0492	0.0007
200	10	C-Lasso	0.1606	0.1139	0.1726	0.1285	0.1797	0.1424	0.1897	0.1525	0.1989	0.1585
		Post-lasso	0.0963	0.0230	0.1034	0.0282	0.1063	0.0371	0.1117	0.0417	0.1201	0.0436
		C-Lasso BC	0.1255	0.0739	0.1355	0.0843	0.1415	0.0961	0.1497	0.1038	0.1575	0.1078
		Oracle	0.0687	0.0007	0.0687	0.0007	0.0687	0.0007	0.0687	0.0007	0.0687	0.0007
200	20	C-Lasso	0.0961	0.0588	0.1000	0.0708	0.1029	0.0820	0.1071	0.0902	0.1118	0.0949
		Post-lasso	0.0572	0.0169	0.0581	0.0207	0.0578	0.0225	0.0582	0.0220	0.0601	0.0197
		C-Lasso BC	0.0755	0.0410	0.0784	0.0495	0.0805	0.0566	0.0829	0.0610	0.0859	0.0628
		Oracle	0.0501	-0.0007	0.0501	-0.0007	0.0501	-0.0007	0.0501	-0.0007	0.0501	-0.0007
200	40	C-Lasso	0.0642	0.0386	0.0627	0.0411	0.0649	0.0443	0.0636	0.0486	0.0661	0.0539
		Post-lasso	0.0411	0.0106	0.0377	0.0097	0.0374	0.0084	0.0370	0.0075	0.0373	0.0072
		C-Lasso BC	0.0513	0.0250	0.0490	0.0258	0.0495	0.0269	0.0489	0.0286	0.0501	0.0313
		Oracle	0.0346	0.0006	0.0346	0.0006	0.0346	0.0006	0.0346	0.0006	0.0346	0.0006

Table 5: Estimation of $\beta_{i,1}$ in DGP 3 by PLS

		C_λ	0.2		0.4		0.8		1.6		3.2	
N	T		RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias
100	10	C-Lasso	0.1331	-0.1216	0.1264	-0.1143	0.1189	-0.1028	0.1120	-0.0858	0.1557	-0.0561
		Post-lasso	0.1011	-0.0863	0.1041	-0.0897	0.1059	-0.0866	0.1077	-0.0784	0.1573	-0.0560
		C-Lasso BC	0.1220	-0.1088	0.1157	-0.1022	0.1088	-0.0909	0.1033	-0.0740	0.1532	-0.0443
		Post-Lasso BC	0.0922	-0.0745	0.0949	-0.0782	0.0971	-0.0751	0.0998	-0.0667	0.1548	-0.0441
		Oracle	0.0928	-0.0855	0.0928	-0.0855	0.0928	-0.0855	0.0928	-0.0855	0.0928	-0.0855
100	20	C-Lasso	0.0782	-0.0711	0.0740	-0.0670	0.0671	-0.0603	0.0580	-0.0505	0.0711	-0.0254
		Post-lasso	0.0539	-0.0431	0.0558	-0.0471	0.0558	-0.0482	0.0529	-0.0444	0.0713	-0.0233
		C-Lasso BC	0.0723	-0.0643	0.0682	-0.0605	0.0614	-0.0540	0.0527	-0.0443	0.0691	-0.0191
		Post-Lasso BC	0.0494	-0.0368	0.0508	-0.0410	0.0507	-0.0421	0.0479	-0.0382	0.0694	-0.0170
		Oracle	0.0527	-0.0469	0.0527	-0.0469	0.0527	-0.0469	0.0527	-0.0469	0.0527	-0.0469
100	40	C-Lasso	0.0428	-0.0372	0.0405	-0.0351	0.0363	-0.0310	0.0321	-0.0270	0.0315	-0.0213
		Post-lasso	0.0289	-0.0224	0.0295	-0.0236	0.0297	-0.0241	0.0293	-0.0238	0.0313	-0.0204
		C-Lasso BC	0.0401	-0.0339	0.0378	-0.0319	0.0336	-0.0279	0.0295	-0.0239	0.0294	-0.0182
		Post-Lasso BC	0.0266	-0.0193	0.0272	-0.0206	0.0273	-0.0210	0.0269	-0.0207	0.0294	-0.0173
		Oracle	0.0285	-0.0236	0.0285	-0.0236	0.0285	-0.0236	0.0285	-0.0236	0.0285	-0.0236
200	10	C-Lasso	0.1297	-0.1235	0.1218	-0.1154	0.1113	-0.1040	0.0976	-0.0855	0.1241	-0.0532
		Post-lasso	0.0941	-0.0859	0.0971	-0.0900	0.0952	-0.0865	0.0899	-0.0761	0.1244	-0.0520
		C-Lasso BC	0.1180	-0.1106	0.1105	-0.1032	0.1004	-0.0921	0.0874	-0.0736	0.1201	-0.0412
		Post-Lasso BC	0.0847	-0.0741	0.0872	-0.0785	0.0854	-0.0751	0.0807	-0.0644	0.1206	-0.0400
		Oracle	0.0898	-0.0859	0.0898	-0.0859	0.0898	-0.0859	0.0898	-0.0859	0.0898	-0.0859
200	20	C-Lasso	0.0748	-0.0703	0.0705	-0.0661	0.0634	-0.0595	0.0541	-0.0501	0.0540	-0.0331
		Post-lasso	0.0491	-0.0418	0.0512	-0.0462	0.0517	-0.0474	0.0484	-0.0441	0.0538	-0.0312
		C-Lasso BC	0.0687	-0.0636	0.0645	-0.0596	0.0575	-0.0532	0.0484	-0.0439	0.0507	-0.0268
		Post-Lasso BC	0.0444	-0.0356	0.0460	-0.0400	0.0463	-0.0413	0.0430	-0.0379	0.0507	-0.0249
		Oracle	0.0492	-0.0460	0.0492	-0.0460	0.0492	-0.0460	0.0492	-0.0460	0.0492	-0.0460
200	40	C-Lasso	0.0399	-0.0364	0.0377	-0.0346	0.0335	-0.0305	0.0295	-0.0265	0.0267	-0.0221
		Post-lasso	0.0259	-0.0216	0.0266	-0.0230	0.0268	-0.0234	0.0266	-0.0233	0.0264	-0.0212
		C-Lasso BC	0.0370	-0.0332	0.0348	-0.0314	0.0307	-0.0274	0.0267	-0.0234	0.0243	-0.0190
		Post-Lasso BC	0.0234	-0.0185	0.0241	-0.0199	0.0242	-0.0203	0.0240	-0.0202	0.0240	-0.0181
		Oracle	0.0261	-0.0231	0.0261	-0.0231	0.0261	-0.0231	0.0261	-0.0231	0.0261	-0.0231

Table 6: Estimation of $\beta_{i,1}$ in DGP 3 by PGMM

		C_λ	0.2		0.4		0.8		1.6		3.2	
N	T		RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias
100	10	C-Lasso	0.1823	-0.1065	0.1892	-0.1241	0.1980	-0.1417	0.2090	-0.1627	0.2271	-0.1817
		Post-lasso	0.1304	-0.0352	0.1231	-0.0331	0.1161	-0.0311	0.1137	-0.0352	0.1202	-0.0427
		C-Lasso BC	0.1494	-0.0698	0.1509	-0.0800	0.1516	-0.0897	0.1572	-0.1047	0.1729	-0.1206
		Oracle	0.0664	-0.0013	0.0664	-0.0013	0.0664	-0.0013	0.0664	-0.0013	0.0664	-0.0013
100	20	C-Lasso	0.0808	-0.0319	0.0858	-0.0478	0.0974	-0.0687	0.1114	-0.0888	0.1247	-0.1035
		Post-lasso	0.0584	-0.0010	0.0565	-0.0031	0.0546	-0.0068	0.0538	-0.0109	0.0554	-0.0138
		C-Lasso BC	0.0678	-0.0175	0.0690	-0.0275	0.0739	-0.0411	0.0814	-0.0548	0.0904	-0.0648
		Oracle	0.0399	-0.0027	0.0399	-0.0027	0.0399	-0.0027	0.0399	-0.0027	0.0399	-0.0027
100	40	C-Lasso	0.0442	-0.0126	0.0447	-0.0198	0.0519	-0.0329	0.0646	-0.0491	0.0742	-0.0606
		Post-lasso	0.0356	0.0025	0.0334	0.0006	0.0327	-0.0018	0.0325	-0.0037	0.0320	-0.0046
		C-Lasso BC	0.0395	-0.0047	0.0384	-0.0094	0.0406	-0.0173	0.0459	-0.0268	0.0507	-0.0333
		Oracle	0.0274	-0.0011	0.0274	-0.0011	0.0274	-0.0011	0.0274	-0.0011	0.0274	-0.0011
200	10	C-Lasso	0.1666	-0.0979	0.1711	-0.1168	0.1788	-0.1386	0.1916	-0.1582	0.2059	-0.1783
		Post-lasso	0.1062	-0.0297	0.0972	-0.0275	0.0912	-0.0276	0.0909	-0.0312	0.0915	-0.0380
		C-Lasso BC	0.1324	-0.0640	0.1305	-0.0743	0.1327	-0.0879	0.1408	-0.1018	0.1497	-0.1171
		Oracle	0.0476	-0.0009	0.0476	-0.0009	0.0476	-0.0009	0.0476	-0.0009	0.0476	-0.0009
200	20	C-Lasso	0.0764	-0.0326	0.0800	-0.0487	0.0910	-0.0700	0.1056	-0.0903	0.1167	-0.1039
		Post-lasso	0.0463	-0.0021	0.0417	-0.0037	0.0408	-0.0075	0.0401	-0.0116	0.0401	-0.0143
		C-Lasso BC	0.0612	-0.0191	0.0603	-0.0289	0.0657	-0.0428	0.0737	-0.0564	0.0809	-0.0657
		Oracle	0.0287	-0.0010	0.0287	-0.0010	0.0287	-0.0010	0.0287	-0.0010	0.0287	-0.0010
200	40	C-Lasso	0.0395	-0.0138	0.0395	-0.0214	0.0466	-0.0348	0.0591	-0.0511	0.0689	-0.0621
		Post-lasso	0.0269	0.0011	0.0235	-0.0007	0.0233	-0.0028	0.0231	-0.0049	0.0227	-0.0055
		C-Lasso BC	0.0320	-0.0066	0.0304	-0.0114	0.0333	-0.0194	0.0392	-0.0289	0.0441	-0.0349
		Oracle	0.0192	-0.0010	0.0192	-0.0010	0.0192	-0.0010	0.0192	-0.0010	0.0192	-0.0010

Table 7: Summary statistics for the savings data set

	mean	median	s.e.	min	max
Savings rate	22.099	20.790	8.833	-3.207	53.434
Inflation rate	7.724	4.853	15.342	-3.846	293.679
Real interest rate	7.422	5.927	10.062	-63.761	93.915
Per capita GDP growth rate	2.855	2.971	3.865	-17.545	14.060

many years; see Feldstein (1980), Deaton (1990), Edwards (1996) Bosworth, Collins, and Reinhart (1999), Rodrik (2000), and Li, Zhang, and Zhang (2007), among many others. Empirical research in this area typically employs standard panel data methods to handle heterogeneity or relies on prior information to categorize countries into groups. Classification criteria vary from geographic locations to the notion of developed countries versus developing countries (Loayza, Schmidt-Hebbel and Servén, 2000). This section applies the methodology developed in the present paper to revisit this empirical problem.

5.1 Model and Data

Following Edwards (1996), we consider the following simple regression model

$$S_{it} = \beta_{1i} S_{i,t-1} + \beta_{2i} I_{it} + \beta_{3i} R_{it} + \beta_{4i} G_{it} + \mu_i + u_{it}, \tag{5.1}$$

where S_{it} is the ratio of savings to GDP, I_{it} is the CPI-based inflation rate, R_{it} is the real interest rate, G_{it} is the per capita GDP growth rate, μ_i is a fixed effect, and u_{it} is an idiosyncratic error term. Inflation characterizes the degree of the macroeconomic stability and the real interest rate reflects the price of money. The relationship between the savings rate and GDP growth rate is well documented, with the latter being found to Granger-cause the former (Carroll and Weil, 1994). A lagged dependent variable is added to the specification to capture persistence of the savings rate.

Data are obtained from the widely used World Development Indicators, a comprehensive dataset compiled by the World Bank.[10] We extract all countries for which there is complete information for all the variables in (5.1). For many countries the time series of real interest rates are often short in comparison with the other variables. Using the time span 1995–2010, we were able to construct a balanced panel of 57 countries, each consisting of 15 time series observations. After removing one outlier,[11] in Table 7 we report the basic descriptive statistics for the remaining 56 countries. As is apparent, there is substantial heterogeneity across countries in all these major macroeconomic indicators. Finding supporting evidence of within group homogeneity is therefore particularly important in supporting the use of panel data pooling techniques.

[10] See http://data.worldbank.org/data-catalog/world-development-indicators.

[11] Bulgaria's 1997 economic collapse produced hyperinflation in the CPI that significantly pulls up the overall mean and the standard deviation. We therefore removed Bulgaria as an outlier from the sample.

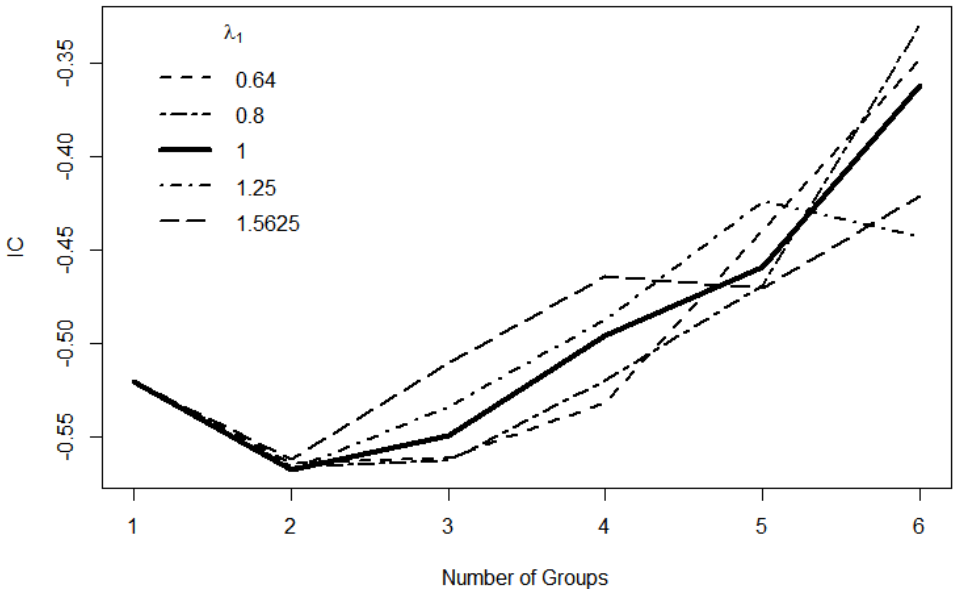

Figure 1: Information criterion as a function of the number of groups under different tuning parameters

Unlike standard FE estimation, the coefficients estimated by PLS as in (2.4) are scale-invariant to neither the dependent variable nor the explanatory variables, due to the presence of the penalty term. For this real data problem, we therefore normalized the data and modified the penalty to enforce scale-invariance. First, after demeaning we standardize each explanatory variable, dividing by the within-country standard deviation so that the standard deviation is unity for each transformed explanatory variable in each country. The transformation makes the coefficients comparable: they can be interpreted as the *ceteris paribus* effect of a one-standard-deviation change of that explanatory variable on the dependent variable. Second, we modify (2.4) to be

$$Q_{1,NT}(\boldsymbol{\beta}) + \frac{\lambda_1}{N}\sum_{i=1}^{N}(\hat{\sigma}_i)^{2-K_0}\,\Pi_{k=1}^{K_0}\|\beta_i - \alpha_k\|. \tag{5.2}$$

where $\hat{\sigma}_i = \left(T^{-1}\sum_{t=1}^{T}\tilde{y}_{it}^2\right)^{1/2}$. The estimate from the above criterion function is scale-invariant to the dependent variable. It is easy to show that the asymptotic theory established earlier continues to hold under these modifications.

Table 8: Estimation results

Slope coefficients	Common	Group 1		Group 2	
	FE	C-Lasso	post-Lasso	C-Lasso	post-Lasso
β_1	0.6203^{***}	0.5746^{***}	0.5652^{***}	0.5715^{***}	0.5813^{***}
	(0.1330)	(0.1059)	(0.1080)	(0.1090)	(0.1051)
β_2	0.0303	-0.1166^{**}	-0.1392^{**}	0.2437^{***}	0.2874^{***}
	(0.0484)	(0.0541)	(0.0517)	(0.0553)	(0.0545)
β_3	0.0068	-0.1039^{**}	-0.0832^{*}	0.1182^{**}	0.1398^{***}
	(0.0432)	(0.0491)	(0.0492)	(0.0459)	(0.0444)
β_4	0.1880^{***}	0.2834^{***}	0.2685^{***}	0.0767	0.0898^{*}
	(0.0450)	(0.0479)	(0.0459)	(0.0477)	(0.0465)

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

5.2 Estimation

Following the practice in Section 5.3 we set $\rho_{1NT} = \frac{2}{3}(NT)^{-1/2}$ and $c_{\lambda_1} = 1$ in the tuning parameter $\lambda_1 = c_{\lambda_1} T^{-1/2}$ in (5.2). We also tried other settings ($c_{\lambda_1} = 0.64$, 0.8, 1.25, and 1.5625) to examine sensitivity of the results to this scaling parameter. Figure 1 plots the information criterion as a function of the number of groups under these tuning parameters. The information criterion suggests two groups for all the tuning parameters under investigation, and it achieves the minimal value when $c_{\lambda_1} = 1$. Based on this choice of tuning parameter, the members in each group are:

- Group 1 (36 countries): Armenia, Australia, Bangladesh, Bolivia, Botswana, Cape Verde, China, Costa Rica, Czech, Guatemala, Honduras, Hungary, Indonesia, Israel, Italy, Japan, Jordan, Latvia, Malawi, Malaysia, Mauritius, Mexico, Mongolia, Panama, Paraguay, Philippines, Romania, Russian, South Africa, Sri Lanka, Switzerland, Syrian, Thailand, Uganda, Ukraine, United Kingdom;

- Group 2 (20 countries): Bahamas, Belarus, Canada, Dominican, Egypt, Guyana, Iceland, India, Kenya, South Korea, Lithuania, Malta, Netherlands, Papua New Guinea, Peru, Singapore, Swaziland, Tanzania, United States, Uruguay.

Here the data determine the group identities. Interestingly, some geographic features are still salient. For example, we observe the dominance of Asian countries in Group 1. Group 1 accommodates 13 Asian countries whereas Group 2 contains only 3. Except South Korea and the city state Singapore, Group 1 includes all Eastern Asian and Southeastern Asian countries in our sample (China, Japan, Indonesia, Malaysia, Philippines, and Thailand).

Table 8 reports the results for the PLS-based C-Lasso and post-Lasso estimation, in comparison with those for the single-group FE estimation. The estimates are bias-corrected and the standard errors (in parentheses) are calculated based on the asymptotic variance-covariance

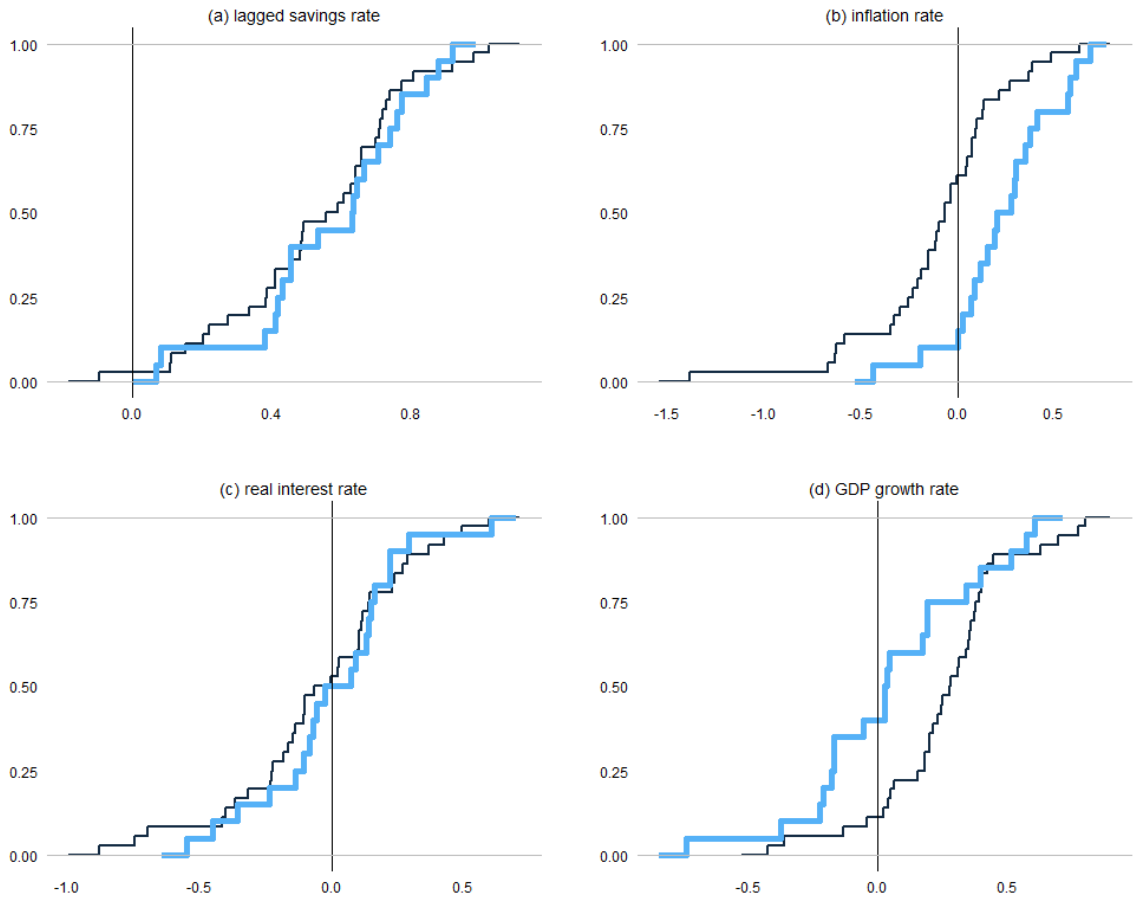

Figure 2: Empirical distribution functions of the time series estimates of regression coefficients for the two estimated groups (thin line: Group 1; thick line: Group 2)

formula. Compared with Edwards (1996), the FE results re-confirm the significance of lagged savings and GDP growth rate as well as the insignificance of inflation and interest rates in the determination of savings rate. This result also lends support to the *conventional wisdom* that across countries higher saving rates tend to go hand in hand with higher income growth (e.g., Loayza, Schmidt-Hebbel and Servén, 2000). The C-Lasso and post-Lasso estimates deliver some interesting findings. First, the coefficients of the inflation rate and the real interest rate become significant in both groups but have opposite signs, which lead to insignificant effects in pooled FE estimation. Second, the coefficient of the GDP growth rate is significant in Group 1 at the 1% level and in Group 2 at the 10% level, which suggests that conventional wisdom is universally relevant and applies both within and across groups.

Figure 2 plots the empirical distribution functions (EDFs) for the time series estimates of the four slope coefficients based on the two estimated groups. The thin and thick lines associate with Groups 1 and 2, respectively. Whilst the time series regression estimates are not precise with only 15 observations for each country, the general pattern in Figure 2 is clearly evident. In the top-left panel for the coefficients of the lagged savings rate, almost all countries exhibit positive coefficient estimates, and the two EDFs are close to each other. Similar remarks also hold for the real interest rate. On the other hand, empirical outcomes for both the inflation rate and GDP growth rates are different. The top-right panel shows that roughly 2/3 of the countries in Group 1 have negative estimates for the inflation rate coefficient in comparison with only 10% of the countries in Group 2; moreover, the Group 2 estimates appear to first-order stochastically dominate those of Group 1. In addition, the bottom-right panel reveals that the GDP growth rate for countries in Group 1 tends to have a larger effect on the savings rate on average than that for countries in Group 2. In sum, the EDF graphics shown in Figure 2 suggest that inflation and GDP growth are the main variables separating the two groups.

6 Conclusion

This paper's main contribution is a novel approach to identifying and estimating latent group structures in panel data. Our work has focussed on linear panel data models where the slope parameters are heterogenous across groups but homogenous within a group and the group identity is unknown, a setting that encompasses many different empirical applications. We have developed panel PLS and PGMM classification and estimation methods. Both these classification methods enjoy the desirable property of uniform consistency. The PLS method has the advantage of possessing the oracle property whereas the PGMM method typically does not. Post-Lasso estimates are also studied and a BIC-type information criterion is proposed to determine the number of groups. These techniques combine to provide a systematic approach to classifying and estimating panel models with unknown homogeneous groups and heterogeneity across groups. Simulations show that the approach has good finite sample performance and can be readily implemented in practical work. Our empirical work on the determinants of cross-country savings rates finds strong evidence that the slope coefficients are heterogeneous and can be conveniently classified into two distinct groups, reinforcing conventional wisdom that higher saving rates go in hand with higher income growth.

APPENDIX

A Proof of the Results in Section 2

Proof of Theorem 2.1. Let $Q_{1NT,i}(\beta_i) = \frac{1}{T}\sum_{t=1}^{T}\left(\tilde{y}_{it} - \beta_i'\tilde{x}_{it}\right)^2$ and $Q_{1iNT,\lambda_1}^{(K_0)}(\beta_i, \boldsymbol{\alpha}) = Q_{1NT,i}(\beta_i)$ $+\lambda_1 \Pi_{k=1}^{K_0}\|\beta_i - \alpha_k\|$. Let $b_i = \beta_i - \beta_i^0$ and $\hat{b}_i = \hat{\beta}_i - \beta_i^0$. Note that

$$Q_{1NT,i}(\beta_i) - Q_{1NT,i}\left(\beta_i^0\right) = \frac{1}{T}\sum_{t=1}^{T}\left(\tilde{u}_{it} - b_i'\tilde{x}_{it}\right)^2 - \frac{1}{T}\sum_{t=1}^{T}\tilde{u}_{it}^2 = b_i'\hat{Q}_{i,\tilde{x}\tilde{x}}b_i - 2b_i'\hat{Q}_{i,\tilde{x}\tilde{u}}. \tag{A.1}$$

By the triangle and reverse triangle inequalities,

$$\begin{aligned}
& \left|\Pi_{k=1}^{K_0}\left\|\hat{\beta}_i - \alpha_k\right\| - \Pi_{k=1}^{K_0}\left\|\beta_i^0 - \alpha_k\right\|\right| \\
\leq\ & \left|\Pi_{k=1}^{K_0-1}\left\|\hat{\beta}_i - \alpha_k\right\|\left\{\left\|\hat{\beta}_i - \alpha_{K_0}\right\| - \left\|\beta_i^0 - \alpha_{K_0}\right\|\right\}\right| \\
& + \left|\Pi_{k=1}^{K_0-2}\left\|\hat{\beta}_i - \alpha_k\right\|\left\|\beta_i^0 - \alpha_{K_0}\right\|\left\{\left\|\hat{\beta}_i - \alpha_{K_0-1}\right\| - \left\|\beta_i^0 - \alpha_{K_0-1}\right\|\right\}\right| \\
& + ... \\
& + \left|\Pi_{k=2}^{K_0}\left\|\beta_i^0 - \alpha_k\right\|\left\{\left\|\hat{\beta}_i - \alpha_1\right\| - \left\|\beta_i^0 - \alpha_1\right\|\right\}\right| \\
\leq\ & \hat{c}_{iNT}(\boldsymbol{\alpha})\left\|\hat{\beta}_i - \beta_i^0\right\|
\end{aligned} \tag{A.2}$$

where $\hat{c}_{iNT}(\boldsymbol{\alpha}) = \Pi_{k=1}^{K_0-1}\left\|\hat{\beta}_i - \alpha_k\right\| + \Pi_{k=1}^{K_0-2}\left\|\hat{\beta}_i - \alpha_k\right\|\left\|\beta_i^0 - \alpha_{K_0}\right\| + ... + \Pi_{k=2}^{K_0}\left\|\beta_i^0 - \alpha_k\right\| = O_P(1)$. By (A.1)-(A.2) and the fact that $Q_{1iNT,\lambda_1}^{(K_0)}(\hat{\beta}_i, \hat{\boldsymbol{\alpha}}) - Q_{1iNT,\lambda_1}^{(K_0)}\left(\beta_i^0, \hat{\boldsymbol{\alpha}}\right) \leq 0$, we have $\underline{c}_{i,\tilde{x}\tilde{x}}\left\|\hat{b}_i\right\|^2 \leq \left(\left\|2\hat{Q}_{i,\tilde{x}\tilde{u}}\right\| + \hat{c}_{iNT}(\hat{\boldsymbol{\alpha}})\lambda_1\right)\left\|\hat{b}_i\right\|$ where $\underline{c}_{i,\tilde{x}\tilde{x}} = \mu_{\min}(\hat{Q}_{i,\tilde{x}\tilde{x}})$. Then, by Assumptions A1(i)-(ii)

$$\left\|\hat{b}_i\right\| \leq \underline{c}_{i,\tilde{x}\tilde{x}}^{-1}\left(2\left\|\hat{Q}_{i,\tilde{x}\tilde{u}}\right\| + \hat{c}_{iNT}(\hat{\boldsymbol{\alpha}})\lambda_1\right) = O_P\left(T^{-1/2} + \lambda_1\right). \tag{A.3}$$

(ii) By Minkowski's inequality and the result in (i), as $(N,T) \to \infty$,

$$\begin{aligned}
\hat{c}_{iNT}(\boldsymbol{\alpha}) &\leq \Pi_{k=1}^{K_0-1}\left\{\left\|\hat{\beta}_i - \beta_i^0\right\| + \left\|\beta_i^0 - \alpha_k\right\|\right\} + \Pi_{k=1}^{K_0-2}\left\{\left\|\hat{\beta}_i - \beta_i^0\right\| + \left\|\beta_i^0 - \alpha_k\right\|\right\}\left\|\beta_i^0 - \alpha_{K_0}\right\| \\
&\quad + ... + \Pi_{k=2}^{K_0}\left\|\beta_i^0 - \alpha_k\right\| \\
&= \sum_{s=0}^{K_0-1}\left\|\hat{\beta}_i - \beta_i^0\right\|^s \Pi_{k=1}^{s} a_{ks}\left\|\beta_i^0 - \alpha_k\right\|^{K_0-1-s} \\
&\leq C_{K_0NT}(\boldsymbol{\alpha})\sum_{s=0}^{K_0-1}\left\|\hat{\beta}_i - \beta_i^0\right\|^s \leq C_{K_0NT}(\boldsymbol{\alpha})\left(1 + 2\left\|\hat{\beta}_i - \beta_i^0\right\|\right),
\end{aligned} \tag{A.4}$$

where a_{ks}'s are finite integers and $C_{K_0NT}(\boldsymbol{\alpha}) = \max_{1\leq i\leq N}\max_{1\leq s\leq k\leq K_0-1}\Pi_{k=1}^{s}a_{ks}\left\|\beta_i^0 - \alpha_k\right\|^{K_0-1-s}$ $= \max_{1\leq l\leq K_0}\max_{1\leq s\leq k\leq K_0-1}\Pi_{k=1}^{s}a_{ks}\left\|\alpha_l^0 - \alpha_k\right\|^{K_0-1-s} = O(1)$ as K_0 is finite. Let $\hat{C}_{K_0} =$

$C_{K_0NT}(\hat{\boldsymbol{\alpha}})$. Combining (A.3)-(A.4) yields $\left\|\hat{b}_i\right\| \leq \frac{\underline{c}_{i,\tilde{x}\tilde{x}}^{-1}}{1-2\hat{C}_{K_0}\lambda_1\underline{c}_{i,\tilde{x}\tilde{x}}^{-1}}\left\{2\left\|\hat{Q}_{i,\tilde{x}\tilde{u}}\right\| + \hat{C}_{K_0}\lambda_1\right\}$. It follows that

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\hat{b}_i\right\|^2 \leq \left(\frac{\underline{c}_{\tilde{x}\tilde{x},NT}^{-1}}{1-2\hat{C}_{K_0}\lambda_1\underline{c}_{\tilde{x}\tilde{x},NT}^{-1}}\right)^2 \frac{1}{N}\sum_{i=1}^{N}\left[2\left\|\hat{Q}_{i,\tilde{x}\tilde{u}}\right\| + \hat{C}_{K_0}\lambda_1\right]^2 = O_P\left(T^{-1}+\lambda_1^2\right)$$

by Assumptions A1(ii)-(iii), where $\underline{c}_{\tilde{x}\tilde{x},NT} = \min_{1\leq i\leq N}\underline{c}_{i,\tilde{x}\tilde{x}}$.

We now demonstrate $\frac{1}{N}\sum_{i=1}^{N}\left\|\hat{b}_i\right\|^2 = O_P\left(T^{-1}\right)$. Let $\boldsymbol{\beta} = \boldsymbol{\beta}^0 + T^{-1/2}\mathbf{v}$ where $\mathbf{v} = (v_1, ..., v_N)$ is a $p \times N$ 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

$$P\left\{\inf_{N^{-1}\sum_{i=1}^{N}\|v_i\|^2=L} Q_{1NT,\lambda_1}^{(K_0)}\left(\boldsymbol{\beta}^0 + T^{-1/2}\mathbf{v}, \hat{\boldsymbol{\alpha}}\right) > Q_{1NT,\lambda_1}^{(K_0)}\left(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right)\right\} \geq 1-\epsilon^*. \tag{A.5}$$

This implies that w.p.a.1 there is a local minimum $\{\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}\}$ such that $N^{-1}\sum_{i=1}^{N}\left\|\hat{b}_i\right\|^2 = O_P\left(T^{-1}\right)$ regardless of the property of $\hat{\boldsymbol{\alpha}}$. By (A.1) and the Cauchy-Schwarz inequality

$$\begin{aligned}
& T\left[Q_{1NT,\lambda_1}^{(K_0)}\left(\boldsymbol{\beta}^0 + T^{-1/2}\mathbf{v}, \hat{\boldsymbol{\alpha}}\right) - Q_{1NT,\lambda_1}^{(K_0)}\left(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right)\right] \\
= & \frac{1}{N}\sum_{i=1}^{N} v_i'\hat{Q}_{i,\tilde{x}\tilde{x}}v_i - \frac{2\sqrt{T}}{N}\sum_{i=1}^{N} v_i'\hat{Q}_{i,\tilde{x}\tilde{u}} + \frac{\lambda_1}{N}\sum_{i=1}^{N}\Pi_{k=1}^{K_0}\left\|\hat{\beta}_i - \hat{\alpha}_k\right\| \\
\geq & \underline{c}_{\tilde{x}\tilde{x},NT}\frac{1}{N}\sum_{i=1}^{N}\|v_i\|^2 - 2\left\{\frac{1}{N}\sum_{i=1}^{N}\|v_i\|^2\right\}^{1/2}\left\{\frac{T}{N}\sum_{i=1}^{N}\left\|\hat{Q}_{i,\tilde{x}\tilde{u}}\right\|^2\right\}^{1/2} \\
\equiv & D_{1NT} - D_{2NT}, \text{ say.}
\end{aligned}$$

By Assumptions A1(ii)-(iii), $\underline{c}_{\tilde{x}\tilde{x},NT}$ is bounded below by $\underline{c}_{\tilde{x}\tilde{x}} > 0$ in large samples and $\frac{T}{N}\sum_{i=1}^{N}\left\|\hat{Q}_{i,\tilde{x}\tilde{u}}\right\|^2 = O_P(1)$. So D_{1NT} dominates D_{2NT} for sufficiently large L. That is $T[Q_{1NT,\lambda_1}^{(K_0)}\left(\boldsymbol{\beta}^0 + T^{-1/2}\mathbf{v}, \hat{\boldsymbol{\alpha}}\right) - Q_{1NT,\lambda_1}^{(K_0)}\left(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right)] > 0$ for sufficiently large L. Consequently, we must have $N^{-1}\sum_{i=1}^{N}\left\|\hat{b}_i\right\|^2 = O_P\left(T^{-1}\right)$.

(iii) Let $P_{NT}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \frac{1}{N}\sum_{i=1}^{N}\Pi_{k=1}^{K_0}\|\beta_i - \alpha_k\|$. By (A.2) and (A.4), as $(N,T)\to\infty$,

$$\begin{aligned}
\left|P_{NT}(\hat{\boldsymbol{\beta}}, \boldsymbol{\alpha}) - P_{NT}\left(\boldsymbol{\beta}^0, \boldsymbol{\alpha}\right)\right| & \leq C_{K_0NT}(\boldsymbol{\alpha})\frac{1}{N}\sum_{i=1}^{N}\left\|\hat{b}_i\right\| + 2C_{K_0NT}(\boldsymbol{\alpha})\frac{1}{N}\sum_{i=1}^{N}\left\|\hat{b}_i\right\|^2 \\
& \leq C_{K_0NT}(\boldsymbol{\alpha})\left\{\frac{1}{N}\sum_{i=1}^{N}\left\|\hat{b}_i\right\|^2\right\}^{1/2} + O_P(T^{-1}) = O_P(T^{-1/2})
\end{aligned} \tag{A.6}$$

By (A.6), and the fact that $P_{NT}\left(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right) = 0$ and that $P_{NT}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}) - P_{NT}(\hat{\boldsymbol{\beta}}, \boldsymbol{\alpha}^0) \leq 0$, we have

$$\begin{aligned}
0 \geq & P_{NT}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}) - P_{NT}(\hat{\boldsymbol{\beta}}, \boldsymbol{\alpha}^0) = P_{NT}\left(\boldsymbol{\beta}^0, \hat{\boldsymbol{\alpha}}\right) - P_{NT}\left(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right) + O_P(T^{-1/2}) \\
= & \frac{1}{N}\sum_{i=1}^{N}\Pi_{k=1}^{K_0}\left\|\beta_i^0 - \hat{\alpha}_k\right\| + O_P(T^{-1/2}) \\
= & \frac{N_1}{N}\Pi_{k=1}^{K_0}\left\|\hat{\alpha}_k - \alpha_1^0\right\| + \frac{N_2}{N}\Pi_{k=1}^{K_0}\left\|\hat{\alpha}_k - \alpha_2^0\right\| + ... + \frac{N_{K_0}}{N}\Pi_{k=1}^{K_0}\left\|\hat{\alpha}_k - \alpha_{K_0}^0\right\| + O_P(T^{-1/2})
\end{aligned} \tag{A.7}$$

By Assumption A1(iv), $N_k/N \to \tau_k \in (0,1)$ for each $k = 1, ..., K_0$. So (A.7) implies that $\Pi_{k=1}^{K_0} \left\|\hat{\alpha}_k - \alpha_l^0\right\| = O_P\left(T^{-1/2}\right)$ for $l = 1, ..., K_0$. It follows that $\left(\hat{\alpha}_{(1)}, ..., \hat{\alpha}_{(K_0)}\right) - (\alpha_1^0, ..., \alpha_{K_0}^0) = O_P\left(T^{-1/2}\right)$. ■

Proof of Theorem 2.2. (i) Fix $k \in \{1, ..., K_0\}$. 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 $l \neq k$. It follows that w.p.a.1 $\left\|\hat{\beta}_i - \hat{\alpha}_l\right\| \neq 0$ for all $i \in G_k^0$ and $l \neq k$. Now, suppose that $\left\|\hat{\beta}_i - \hat{\alpha}_k\right\| \neq 0$ for some $i \in G_k^0$. Then the first order condition (with respect to β_i) for the minimization problem in (2.4) implies that

$$\begin{aligned}
\mathbf{0}_{p\times 1} &= \frac{-2}{\sqrt{T}} \sum_{t=1}^{T} \tilde{x}_{it}\left(\tilde{y}_{it} - \tilde{x}_{it}'\hat{\beta}_i\right) + \sqrt{T}\lambda_1 \sum_{j=1}^{K_0} \frac{\hat{\beta}_i - \hat{\alpha}_j}{\left\|\hat{\beta}_i - \hat{\alpha}_j\right\|} \Pi_{l=1,l\neq j}^{K_0} \left\|\hat{\beta}_i - \hat{\alpha}_l\right\| \\
&= \frac{-2}{\sqrt{T}} \sum_{t=1}^{T} \tilde{x}_{it}\tilde{u}_{it} + \left(2\hat{Q}_{i,\tilde{x}\tilde{x}} + \frac{\lambda_1 \hat{c}_{ki}}{\left\|\hat{\beta}_i - \hat{\alpha}_k\right\|} I_p\right) \sqrt{T}\left(\hat{\beta}_i - \hat{\alpha}_k\right) \\
&\quad + 2\hat{Q}_{i,\tilde{x}\tilde{x}} \sqrt{T}\left(\hat{\alpha}_k - \beta_i^0\right) + \sqrt{T}\lambda_1 \sum_{j=1, j\neq k}^{K_0} \frac{\hat{\beta}_i - \hat{\alpha}_j}{\left\|\hat{\beta}_i - \hat{\alpha}_j\right\|} \Pi_{l=1,l\neq j}^{K_0} \left\|\hat{\beta}_i - \hat{\alpha}_l\right\| \\
&\equiv -\hat{B}_{i1} + \hat{B}_{i2} + \hat{B}_{i3} + \sum_{j=1, j\neq k}^{K_0} \hat{B}_{i4,j}, \text{ say,} \qquad \text{(A.8)}
\end{aligned}$$

where $\hat{c}_{ki} = \Pi_{l=1,l\neq k}^{K_0} \left\|\hat{\beta}_i - \hat{\alpha}_l\right\| \xrightarrow{P} c_k^0 \equiv \Pi_{l=1,l\neq k}^{K_0} \left\|\alpha_k^0 - \alpha_l^0\right\| > 0$ for $i \in G_k^0$ by Theorem 2.1.

Clearly $\hat{B}_{i1} = O_P(1)$ by Assumption A1(i) and $\hat{B}_{i3} = O_P(1)$ by Theorem 2.1(iii) as $i \in G_k^0$. One can also show that $\hat{B}_{i4,j} = \sqrt{T}\lambda_1 O_P(T^{-1/2} + \lambda_1) = O_P(1)$ for each i and j by Theorems 2.1(i) and (iii) and Assumption A2(i). Let $\hat{R}_i = \hat{B}_{i3} + \sum_{j=1,j\neq k}^{K_0} \hat{B}_{i4,j}$. Noting that $\left(\hat{\beta}_i - \hat{\alpha}_k\right)' \hat{B}_{i2} \geq 2\underline{c}_{i,\tilde{x}\tilde{x}} \sqrt{T} \left\|\hat{\beta}_i - \hat{\alpha}_k\right\|^2 + \sqrt{T}\lambda_1 \hat{c}_{ki}$, $\left|\left(\hat{\beta}_i - \hat{\alpha}_k\right)' \hat{R}_i\right| = O_P(\lambda_1)$, we have $\left(\hat{\beta}_i - \hat{\alpha}_k\right)' \hat{B}_{i2}$ $- \left|\left(\hat{\beta}_i - \hat{\alpha}_k\right)' \hat{R}_i\right| \geq \left(\hat{\beta}_i - \hat{\alpha}_k\right)' \hat{B}_{i2}/2$ as $(N,T) \to \infty$. It follows that for all $i \in G_k^0$

$$\begin{aligned}
P(\hat{E}_{kNT,i}) &= P\left(i \notin \hat{G}_k \mid i \in G_k^0\right) = P\left(\hat{B}_{i1} = \hat{B}_{i2} + \hat{R}_i\right) \\
&\leq P\left(\left|\left(\hat{\beta}_i - \hat{\alpha}_k\right)' \hat{B}_{i1}\right| \geq \left|\left(\hat{\beta}_i - \hat{\alpha}_k\right)' \hat{B}_{i2} + \left(\hat{\beta}_i - \hat{\alpha}_k\right)' \hat{R}_i\right|\right) \\
&\leq P\left(\left\|\hat{\beta}_i - \hat{\alpha}_k\right\| \left\|\hat{B}_{i1}\right\| \geq \left(\hat{\beta}_i - \hat{\alpha}_k\right)' \hat{B}_{i2} - \left|\left(\hat{\beta}_i - \hat{\alpha}_k\right)' \hat{R}_i\right|\right) \\
&\leq P\left(\left\|\hat{B}_{i1}\right\| \geq \underline{c}_{i,\tilde{x}\tilde{x}} \sqrt{T} \left\|\hat{\beta}_i - \hat{\alpha}_k\right\| + \frac{\sqrt{T}\lambda_1 \hat{c}_{ki}}{2\left\|\hat{\beta}_i - \hat{\alpha}_k\right\|}\right) \\
&\leq P\left(\left\|\hat{B}_{i1}\right\| \geq \sqrt{2\underline{c}_{i,\tilde{x}\tilde{x}} \hat{c}_{ki} T\lambda_1}\right) \to 0 \text{ as } (N,T) \to \infty,
\end{aligned}$$

where the second and fourth inequalities follow from the Cauchy-Schwarz and triangle inequalities, and Cauchy-Schwarz inequality, respectively, and the last convergence result follows from

Assumptions A1(ii) and A2(i) and the fact that $\hat{c}_{ki} \xrightarrow{P} c_k^0$ for $i \in G_k^0$. Consequently, we may conclude that w.p.a.1 the differences $\hat{\beta}_i - \hat{\alpha}_k$ must reach the point where $\|\beta_i - \alpha_k\|$ is not differentiable with respect to β_i for any $i \in G_k^0$. That is $P\left(\left\|\hat{\beta}_i - \hat{\alpha}_k\right\| = 0 \mid i \in G_k^0\right) \to 1$ as $(N,T) \to \infty$.

For the uniform consistency, observe that $P(\cup_{k=1}^{K_0} \hat{E}_{kNT}) \leq \sum_{k=1}^{K_0} P(\hat{E}_{kNT}) \leq \sum_{k=1}^{K_0} \sum_{i \in G_k^0} P(\hat{E}_{kNT,i})$ and

$$\begin{aligned}
\sum_{k=1}^{K_0} \sum_{i \in G_k^0} P\left(\hat{E}_{kNT,i}\right) &\leq \sum_{k=1}^{K_0} \sum_{i \in G_k^0} P\left(\left\|\hat{B}_{i1}\right\| \geq \sqrt{2\underline{c}_{i,\tilde{x}\tilde{x}} \hat{c}_k T \lambda_1}\right) \\
&\leq N \max_{1 \leq i \leq N} P\left(\left\|\frac{1}{T}\sum_{t=1}^{T} \tilde{x}_{it}\tilde{u}_{it}\right\| \geq \sqrt{\frac{\underline{c}_{i,\tilde{x}\tilde{x}} \hat{c}_{ki} \lambda_1}{2}}\right) \\
&\to 0 \text{ as } (N,T) \to \infty \text{ by Assumption A2(ii).} \qquad \text{(A.9)}
\end{aligned}$$

This completes the proof of (i).

(ii) By pretending each individual's membership is random, we have $P\left(i \in G_k^0\right) = N_k/N \to \tau_k \in (0,1)$ for $k = 1, ..., K_0$ and can interpret previous results as conditional on the group membership assignment. By Bayes theorem,

$$\begin{aligned}
P(\hat{F}_{kNT,i}) &= 1 - P(i \in G_k^0 \mid i \in \hat{G}_k) \\
&= \frac{\sum_{l=1,l\neq k}^{K_0} P(i \in \hat{G}_k | i \in G_l^0) P\left(i \in G_l^0\right)}{P(i \in \hat{G}_k | i \in G_k^0) P\left(i \in G_k^0\right) + \sum_{l=1,l\neq k}^{K_0} P(i \in \hat{G}_k | i \in G_l^0) P\left(i \in G_l^0\right)}. \qquad \text{(A.10)}
\end{aligned}$$

For the numerator, we have by (A.9)

$$\sum_{l=1,l\neq k}^{K_0} \sum_{i \in \hat{G}_k} P\left(i \in \hat{G}_k | i \in G_l^0\right) P\left(i \in G_l^0\right) \leq (K_0 - 1) \sum_{l=1}^{K_0} \sum_{i \in G_l^0} P\left(i \notin \hat{G}_l | i \in G_l^0\right) = o(1).$$

In addition, noting that $P(i \in \hat{G}_k | i \in G_k^0) = 1 - P(i \notin \hat{G}_k | i \in G_k^0) = 1 - o(1)$ uniformly in i and k by (i), we have that $P(i \in \hat{G}_k | i \in G_k^0) P\left(i \in G_k^0\right) + \sum_{l=1,l\neq k}^{K_0} P(i \in \hat{G}_k | i \in G_l^0) P\left(i \in G_l^0\right) \geq P\left(i \in G_k^0\right)/2$ w.p.a.1. It follows that

$$\begin{aligned}
P\left(\cup_{k=1}^{K_0} \hat{F}_{kNT}\right) &\leq \sum_{k=1}^{K_0} P\left(\hat{F}_{kNT}\right) \leq \sum_{k=1}^{K_0} \sum_{i \in \hat{G}_k} P\left(\hat{F}_{kNT,i}\right) \\
&\leq \frac{\sum_{l=1,l\neq k}^{K_0} \sum_{i \in \hat{G}_k} P\left(i \in \hat{G}_k | i \in G_l^0\right) P\left(i \in G_l^0\right)}{\min_{1 \leq i \leq N} \min_{1 \leq k \leq K_0} P\left(i \in G_k^0\right)/2} \\
&= \frac{o(1)}{\min_{1 \leq k \leq K_0} \tau_k/2} = o(1). \blacksquare
\end{aligned}$$

Proof of Corollary 2.3. Noting that $\hat{N}_k = \sum_{i=1}^{N} \mathbf{1}\{i \in \hat{G}_k\}$, $N_k = \sum_{i=1}^{N} \mathbf{1}\{i \in G_k^0\}$, and $\mathbf{1}\{i \in \hat{G}_k\} - \mathbf{1}\{i \in G_k^0\} = \mathbf{1}\{i \in \hat{G}_k \backslash G_k^0\} - \mathbf{1}\{i \in G_k^0 \backslash \hat{G}_k\}$, we have $\hat{N}_k - N_k = \sum_{i=1}^{N} [\mathbf{1}\{i \in \hat{G}_k \backslash G_k^0\} -$

$\mathbf{1}\{i \in G_k^0 \backslash \hat{G}_k\}]$. Then by the implication rule and Markov inequality, for any $\epsilon > 0$,

$$\begin{aligned} P\left(\left|\hat{N}_k - N_k\right| \geq 2\epsilon\right) &\leq P\left(\sum_{i=1}^{N} \mathbf{1}\{i \in \hat{G}_k \backslash G_k^0\} \geq \epsilon\right) + P\left(\sum_{i=1}^{N} \mathbf{1}\{i \in G_k^0 \backslash \hat{G}_k\} \geq \epsilon\right) \\ &= \frac{1}{\epsilon}\sum_{i=1}^{N} P\left(\hat{F}_{kNT,i}\right) + \frac{1}{\epsilon}\sum_{i=1}^{N} P\left(\hat{E}_{kNT,i}\right). \end{aligned}$$

By (A.9), $\sum_{i=1}^{N} P(\hat{E}_{kNT,i}) = \sum_{k=1}^{K_0}\sum_{i\in G_k^0} P(\hat{E}_{kNT,i}) = o(1)$. By the proof of Theorem 2.2(i), $\sum_{i=1}^{N} P(\hat{F}_{kNT,i}) = \sum_{k=1}^{K_0}\sum_{i\in \hat{G}_k} P(\hat{F}_{kNT,i}) = o(1)$. Consequently, $P(|\hat{N}_k - N_k| \geq 2\epsilon) = o(1)$ and the conclusion follows. ■

Proof of Theorem 2.4. To study the oracle property of the Lasso estimator, we utilize conditions from subdifferential calculus (e.g., Bersekas (1995, Appendix B.5)). In particular, necessary and sufficient conditions for $\{\hat{\beta}_i\}$ and $\{\hat{\alpha}_k\}$ to minimize the objective function in (2.4) is that for each $i = 1, ..., N$ (resp. $k = 1, ..., K_0$), $\mathbf{0}_{p\times 1}$ belongs to the subdifferential of $Q_{1NT,\lambda_1}^{(K_0)}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ 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 $k = 1, ..., K_0$, we have

$$\mathbf{0}_{p\times 1} = \frac{-2}{NT}\sum_{t=1}^{T} \tilde{x}_{it}\left(\tilde{y}_{it} - \hat{\beta}_i'\tilde{x}_{it}\right) + \frac{\lambda_1}{N}\sum_{j=1}^{K_0} \hat{e}_{ij}\Pi_{l=1,l\neq j}^{K_0}\left\|\hat{\beta}_i - \hat{\alpha}_l\right\|, \text{ and} \tag{A.11}$$

$$\mathbf{0}_{p\times 1} = \frac{\lambda_1}{N}\sum_{i=1}^{N} \hat{e}_{ik}\Pi_{l=1,l\neq k}^{K_0}\left\|\hat{\beta}_i - \hat{\alpha}_l\right\|, \tag{A.12}$$

where $\hat{e}_{ij} = \frac{\hat{\beta}_i - \hat{\alpha}_j}{\left\|\hat{\beta}_i - \hat{\alpha}_j\right\|}$ if $\left\|\hat{\beta}_i - \hat{\alpha}_j\right\| \neq 0$ and $\|\hat{e}_{ij}\| \leq 1$ if $\left\|\hat{\beta}_i - \hat{\alpha}_j\right\| = 0$. Fix $k \in \{1, ..., K_0\}$. Observe that (a) $\left\|\hat{\beta}_i - \hat{\alpha}_k\right\| = 0$ for any $i \in \hat{G}_k$ by the definition of $\hat{G}_k$, and (b) $\hat{\beta}_i - \hat{\alpha}_l \xrightarrow{P} \alpha_k^0 - \alpha_l^0 \neq 0$ for any $i \in \hat{G}_k$ and $l \neq k$. It follows that $\|\hat{e}_{ik}\| \leq 1$ for any $i \in \hat{G}_k$ and $\hat{e}_{ij} = \frac{\hat{\beta}_i - \hat{\alpha}_j}{\left\|\hat{\beta}_i - \hat{\alpha}_j\right\|} = \frac{\hat{\alpha}_k - \hat{\alpha}_j}{\|\hat{\alpha}_k - \hat{\alpha}_j\|}$ w.p.a.1 for any $i \in \hat{G}_k$ and $j \neq k$. This further implies that w.p.a.1

$$\sum_{i\in\hat{G}_k}\sum_{j=1,j\neq k}^{K_0} \hat{e}_{ij}\Pi_{l=1,l\neq j}^{K_0}\left\|\hat{\beta}_i - \hat{\alpha}_l\right\| = \sum_{i\in\hat{G}_k}\sum_{j=1,j\neq k}^{K_0} \frac{\hat{\alpha}_k - \hat{\alpha}_j}{\|\hat{\alpha}_k - \hat{\alpha}_j\|}\Pi_{l=1,l\neq j}^{K_0}\|\hat{\alpha}_k - \hat{\alpha}_l\| = 0 \text{ and}$$

and

$$\begin{aligned} \sum_{i=1}^{N} \hat{e}_{ik}\Pi_{l=1,l\neq k}^{K_0}\left\|\hat{\beta}_i - \hat{\alpha}_l\right\| &= \sum_{i\in\hat{G}_k} \hat{e}_{ik}\Pi_{l=1,l\neq k}^{K_0}\|\hat{\alpha}_k - \hat{\alpha}_l\| + \sum_{i\in\hat{G}_0} \hat{e}_{ik}\Pi_{l=1,l\neq k}^{K_0}\left\|\hat{\beta}_i - \hat{\alpha}_l\right\| \\ &\quad + \sum_{j=1,j\neq k}^{K_0}\sum_{i\in\hat{G}_j} \hat{e}_{ik}\Pi_{l=1,l\neq k}^{K_0}\|\hat{\alpha}_j - \hat{\alpha}_l\| \\ &= \sum_{i\in\hat{G}_k} \hat{e}_{ik}\Pi_{l=1,l\neq k}^{K_0}\|\hat{\alpha}_k - \hat{\alpha}_l\| + \sum_{i\in\hat{G}_0} \hat{e}_{ik}\Pi_{l=1,l\neq k}^{K_0}\left\|\hat{\beta}_i - \hat{\alpha}_l\right\| = 0. \end{aligned}$$

Then by (A.11) we have $\frac{2}{NT}\sum_{i\in\hat{G}_k}\sum_{t=1}^{T}\tilde{x}_{it}\left(\tilde{y}_{it}-\hat{\alpha}_k'\tilde{x}_{it}\right)+\frac{\lambda_1}{N}\sum_{i\in\hat{G}_0}\hat{e}_{ik}\Pi_{l=1,l\neq k}^{K_0}\left\|\hat{\beta}_i-\hat{\alpha}_l\right\| = \mathbf{0}_{p\times 1}$. It follows that

$$\begin{aligned}\hat{\alpha}_k &= \left(\frac{1}{NT}\sum_{i\in\hat{G}_k}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{x}_{it}'\right)^{-1}\frac{1}{NT}\sum_{i\in\hat{G}_k}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{y}_{it}\\ &\quad+\left(\frac{1}{NT}\sum_{i\in\hat{G}_k}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{x}_{it}'\right)^{-1}\frac{\lambda_1}{2N}\sum_{i\in\hat{G}_0}\hat{e}_{ik}\Pi_{l=1,l\neq k}^{K_0}\left\|\hat{\beta}_i-\hat{\alpha}_l\right\|\equiv\hat{\alpha}_{\hat{G}_k}+\hat{\mathcal{R}}_k,\text{ say.}\end{aligned}$$

In view of the fact that, $\hat{e}_{ik}\Pi_{l=1,l\neq k}^{K_0}\left\|\hat{\beta}_i-\hat{\alpha}_l\right\|\neq 0$ only if $i\in\hat{G}_0$, we have for any $\epsilon>0$

$$P\left(\sqrt{NT}\left\|\hat{\mathcal{R}}_k\right\|\geq\epsilon\right)\leq\sum_{k=1}^{K_0}\sum_{i\in G_k^0}P\left(i\in\hat{G}_0|i\in G_k^0\right)\leq\sum_{k=1}^{K_0}\sum_{i\in G_k^0}P\left(i\notin\hat{G}_k|i\in G_k^0\right)=o\left(1\right)\text{ by (A.9).}$$

So $\left\|\hat{\mathcal{R}}_k\right\|=o_P\left((NT)^{-1/2}\right)$. Then the limit distribution result follows from Theorem 2.5 below. ■

Proof of Theorem 2.5. Noting that $\tilde{y}_{it}=\tilde{x}_{it}'\alpha_k^0+\tilde{x}_{it}'\left(\beta_i^0-\alpha_k^0\right)+\tilde{u}_{it}$, we have

$$\begin{aligned}\sqrt{N_kT}\left(\hat{\alpha}_{\hat{G}_k}-\alpha_k^0\right) &= \left(\frac{1}{N_kT}\sum_{i\in\hat{G}_k}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{x}_{it}'\right)^{-1}\frac{1}{\sqrt{N_kT}}\sum_{i\in\hat{G}_k}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{u}_{it}\\ &\quad+\left(\frac{1}{N_kT}\sum_{i\in\hat{G}_k}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{x}_{it}'\right)^{-1}\frac{1}{\sqrt{N_kT}}\sum_{i\in\hat{G}_k}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{x}_{it}'\left(\beta_i^0-\alpha_k^0\right).\end{aligned}$$

By Assumption A3 and Slutsky theorem, it suffices to prove the theorem by showing that (i) $S_{kNT,1}\equiv\frac{1}{N_kT}\sum_{i\in\hat{G}_k}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{x}_{it}'=\frac{1}{N_kT}\sum_{i\in G_k^0}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{x}_{it}'+o_P\left(1\right)$, (ii) $S_{kNT,2}\equiv\frac{1}{\sqrt{N_kT}}\sum_{i\in\hat{G}_k}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{u}_{it}=\frac{1}{\sqrt{N_kT}}\sum_{i\in G_k^0}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{u}_{it}+o_P\left(1\right)$, (iii) $S_{kNT,3}\equiv\frac{1}{\sqrt{N_kT}}\sum_{i\in\hat{G}_k}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{x}_{it}'\left(\beta_i^0-\alpha_k^0\right)=o_P\left(1\right)$, and (iv) $S_{kNT,4}\equiv(S_{kNT,1}^{-1}-\bar{\Phi}_k^{-1})\mathbb{B}_{KNT}=o_P\left(1\right)$.

Using the fact 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

$$S_{kNT,1}-\frac{1}{N_kT}\sum_{i\in G_k^0}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{x}_{it}'=\frac{1}{N_kT}\sum_{i\in\hat{G}_k\backslash G_k^0}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{x}_{it}'-\frac{1}{N_kT}\sum_{i\in G_k^0\backslash\hat{G}_k}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{x}_{it}'\equiv S_{kNT,11}-S_{kNT,12}.$$

Let $\epsilon>0$. By Theorem 2.2, $P\left(\|S_{kNT,11}\|\geq\epsilon\right)\leq P(\hat{F}_{kNT})\to 0$, and $P\left(\|S_{kNT,12}\|\geq\epsilon\right)\leq P(\hat{E}_{kNT})\to 0$. Then (i) follows. Analogously, writing $S_{kNT,2}-\frac{1}{\sqrt{N_kT}}\sum_{i\in G_k^0}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{u}_{it}=\frac{1}{\sqrt{N_kT}}\sum_{i\in\hat{G}_k\backslash G_k^0}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{u}_{it}-\frac{1}{\sqrt{N_kT}}\sum_{i\in G_k^0\backslash\hat{G}_k}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{u}_{it}\equiv S_{kNT,21}-S_{kNT,22}$, we have $P(\|S_{kNT,21}\|\geq\epsilon)\leq P(\hat{F}_{kNT})\to 0$ and $P\left(\|S_{kNT,22}\|\geq\epsilon\right)\leq P(\hat{E}_{kNT})\to 0$. Then (ii) follows. Noting that $\beta_i^0-\alpha_k^0=0$ if $i\in G_k^0$, $P\left(\|S_{kNT,3}\|\geq\epsilon\right)\leq P(\hat{E}_{kNT})+P(\hat{F}_{kNT})\to 0+0=0$. Lastly, $P\left(\|S_{kNT,4}\|\geq\epsilon\right)\leq P(\hat{E}_{kNT})+P(\hat{F}_{kNT})\to 0+0=0$. ■

Proof of Theorem 2.6. Using Theorems 2.2 and 2.5 and Assumption A5, we can readily show that

$$\begin{aligned} IC_1(K_0,\lambda_1) &= \ln\left[\hat{\sigma}^2_{\hat{G}(K_0,\lambda_1)}\right] + \rho_{1NT} p K_0 \\ &= \ln\left[\frac{1}{NT}\sum_{k=1}^{K_0}\sum_{i\in\hat{G}_k(K_0,\lambda_1)}\sum_{t=1}^{T}\left(\tilde{y}_{it} - \tilde{\alpha}'_{\hat{G}_k(K_0,\lambda_1)}\tilde{x}_{it}\right)^2\right] + o(1) \xrightarrow{P} \ln\left(\sigma_0^2\right). \end{aligned}$$

We consider the cases of under- and over-fitted models separately.

Case 1: Under-fitted model. In this case, we have $K < K_0$. Noting that

$$\begin{aligned} \hat{\sigma}^2_{\hat{G}(K,\lambda_1)} &= \frac{1}{NT}\sum_{k=1}^{K}\sum_{i\in\hat{G}_k(K,\lambda_1)}\sum_{t=1}^{T}\left(\tilde{y}_{it} - \tilde{\alpha}'_{\hat{G}_k(K,\lambda_1)}\tilde{x}_{it}\right)^2 \\ &\geq \min_{1\leq K<K_0}\inf_{G^{(K)}\in\mathcal{G}_K}\frac{1}{NT}\sum_{k=1}^{K}\sum_{i\in G_{K,k}}\sum_{t=1}^{T}\left(\tilde{y}_{it} - \tilde{\alpha}'_{G_{K,k}}\tilde{x}_{it}\right)^2 = \min_{1\leq K<K_0}\inf_{G^{(K)}\in\mathcal{G}_K}\hat{\sigma}^2_{G^{(K)}}, \end{aligned}$$

we have by Assumptions A4-A5 and the Slutsky Lemma

$$\min_{1\leq K<K_0} IC_1(K,\lambda_1) \geq \min_{1\leq K<K_0}\inf_{G^{(K)}\in\mathcal{G}_K}\ln(\hat{\sigma}^2_{G^{(K)}}) + \rho_{1NT} p K \xrightarrow{P} \ln(\underline{\sigma}^2) > \ln(\sigma_0^2).$$

It follows that $P\left(\min_{K\in\Omega_-} IC_1(K,\lambda_1) > IC_1(K_0,\lambda_1)\right) \to 1$.

Case 2: Over-fitted model. Let $K \in \Omega_+$. By Lemma A.1 below and the fact that $\delta^2_{NT}\rho_{1NT} \to \infty$ under Assumption A5, we have

$$\begin{aligned} & P\left(\min_{K\in\Omega_+} IC_1(K,\lambda_1) > IC_1(K_0,\lambda_1)\right) \\ = \; & P\left(\min_{K\in\Omega_+}\left[\delta^2_{NT}\ln\left(\hat{\sigma}^2_{\hat{G}(K,\lambda_1)}/\hat{\sigma}^2_{\hat{G}(K_0,\lambda_1)}\right) + \delta^2_{NT}\rho_{1NT}(K-K_0)\right] > 0\right) \\ = \; & P\left(\min_{K\in\Omega_+}\delta^2_{NT}\left(\hat{\sigma}^2_{\hat{G}(K,\lambda_1)} - \hat{\sigma}^2_{\hat{G}(K_0,\lambda_1)}\right)/\hat{\sigma}^2_{\hat{G}(K_0,\lambda_1)} + \delta^2_{NT}\rho_{1NT}(K-K_0) + o_P(1) > 0\right) \\ \to \; & 1 \text{ as } (N,T)\to\infty. \; \blacksquare \end{aligned}$$

Lemma A.1 *Suppose that the conditions in Theorem 2.6 hold. Let* $\bar{\sigma}^2_{G^0} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{u}^2_{it}$. *Then* $\max_{K_0\leq K\leq K_{\max}}\left|\hat{\sigma}^2_{\hat{G}(K,\lambda_1)} - \bar{\sigma}^2_{G^0}\right| = O_P\left(\delta^{-2}_{NT}\right)$.

Proof. When $K \geq K_0$, following the proof of Theorem 2.1, we can show that

$$\left\|\hat{\beta}_i - \beta_i^0\right\| = O_P(T^{-1/2}+\lambda_1) \text{ for each } i \text{ and } \frac{1}{N}\sum_{i=1}^{N}\Pi_{k=1}^{K}\left\|\beta_i^0 - \hat{\alpha}_k\right\| = O_P(T^{-1/2}).$$

Noting that β_i^0, $i = 1,...,N$, only take K_0 distinct values, the latter implies that the collection $\{\hat{\alpha}_k, k = 1,...,K\}$ contains at least K_0 distinct vectors, say, $\hat{\alpha}_{(1)},...,\hat{\alpha}_{(K_0)}$, such that $\hat{\alpha}_{(k)} -$

$\alpha_k^0 = O_P(T^{-1/2})$ for $k = 1, ..., K_0$. For notational simplicity, we rename the other vectors in the above collection as $\hat{\alpha}_{(K_0+1)}, ..., \hat{\alpha}_{(K)}$. As before, we classify $i \in \hat{G}_k(K, \lambda_1)$ if $\left\|\hat{\beta}_i - \hat{\alpha}_{(k)}\right\| = 0$ for $k = 1, ..., K$, and $i \in \hat{G}_0(K, \lambda_1)$ otherwise. Using arguments like those used in the proof of Theorem 2.2, we can show that

$$\sum_{i \in G_k^0} P\left(\hat{E}_{kNT,i}\right) = o(1) \text{ for } k = 1, ..., K_0 \text{ and } \sum_{i \in \hat{G}_k(K,\lambda_1)} P\left(\hat{F}_{kNT,i}\right) = o(1) \text{ for } k = 1, ..., K_0.$$

The first part implies that $\sum_{i=1}^N P\left(i \in \hat{G}_0(K, \lambda_1) \cup \hat{G}_{K_0+1}(K, \lambda_1) \cup \cdots \cup \hat{G}_K(K, \lambda_1)\right) = o(1)$.

Using the fact 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 $\hat{\sigma}^2_{\hat{G}(K,\lambda_1)} = \frac{1}{NT}\sum_{k=1}^K \sum_{i \in \hat{G}_k(K,\lambda_1)} \sum_{t=1}^T [\hat{u}_{it}(k)]^2 = D_{1NT} + D_{2NT} - D_{3NT} + D_{4NT}$, where $\hat{u}_{it}(k) = \tilde{y}_{it} - \hat{\alpha}'_{\hat{G}_k(K,\lambda_1)} \tilde{x}_{it}$, $D_{1NT} = \frac{1}{NT}\sum_{k=1}^{K_0} \sum_{i \in G_k^0} \sum_{t=1}^T [\hat{u}_{it}(k)]^2$, $D_{2NT} = \frac{1}{NT}\sum_{k=1}^{K_0} \sum_{i \in \hat{G}_k(K,\lambda_1) \backslash G_k^0} \sum_{t=1}^T [\hat{u}_{it}(k)]^2$, $D_{3NT} = \frac{1}{NT}\sum_{k=1}^{K_0} \sum_{i \in G_k^0 \backslash \hat{G}_k(K,\lambda_1)} \sum_{t=1}^T [\hat{u}_{it}(k)]^2$, and $D_{4NT} = \frac{1}{NT}\sum_{k=K_0+1}^{K} \sum_{i \in \hat{G}_k(K,\lambda_1)} \sum_{t=1}^T [\hat{u}_{it}(k)]^2$. Following the proof of Theorem 2.5, we can show that $\hat{\alpha}_{\hat{G}_k(K,\lambda_1)} - \alpha_k^0 = O_P\left(\delta_{NT}^{-1}\right)$ for $k = 1, ..., K_0$. In addition, we can readily show that $D_{1NT} = \bar{\sigma}^2_{G^0} + O_P\left(\delta_{NT}^{-2}\right)$. For D_{2NT}, D_{3NT}, and D_{4NT}, we have that for any $\epsilon > 0$, $P\left(D_{2NT} \geq \delta_{NT}^{-2}\epsilon\right) \leq \sum_{i=1}^{K_0} P(\hat{F}_{kNT}) \to 0$, $P\left(D_{3NT} \geq \delta_{NT}^{-2}\epsilon\right) \leq \sum_{i=1}^{K_0} P(\hat{E}_{kNT}) \to 0$, and $P\left(D_{4NT} \geq \delta_{NT}^{-2}\epsilon\right) \leq \sum_{i=1}^{N} P(i \in \cup_{K_0+1 \leq k \leq K} \hat{G}_k(K, \lambda_1)) \to 0$. It follows that $\hat{\sigma}^2_{\hat{G}(K,\lambda_1)} = \bar{\sigma}^2_{G^0} + O_P\left(\delta_{NT}^{-2}\right)$ for all $K_0 \leq K \leq K_{\max}$. ■

B Proof of the Results in Section 3

We start by proving a useful technical result and then proceed to prove the main results.

Let $V_{iNT}(\beta_i) \equiv [\frac{1}{T}\sum_{t=1}^T \rho(\xi_{it}, \beta_i)]' W_{iNT} [\frac{1}{T}\sum_{t=1}^T \rho(\xi_{it}, \beta_i)]$, and $\bar{V}_i(\beta_i) \equiv \{\frac{1}{T}\sum_{t=1}^T \mathbb{E}[\rho(\xi_{it}, \beta_i)]\}' W_i \frac{1}{T}\sum_{t=1}^T \mathbb{E}[\rho(\xi_{it}, \beta_i)]$. Let $R_{i,T}(\beta_i) = [\frac{1}{T}\sum_{t=1}^T \{\rho(\xi_i, \beta_i) - \mathbb{E}[\rho(\xi_i, \beta_i)]\}]' W_i [\frac{1}{T}\sum_{t=1}^T \{\rho(\xi_i, \beta_i) - \mathbb{E}[\rho(\xi_i, \beta_i)]\}]$.

Lemma B.1 *Suppose Assumption B1(iv) hold. Then* $\underline{c}\left[\frac{1}{2}\bar{V}_i(\beta_i) - R_{i,T}(\beta_i)\right] \leq V_{iNT}(\beta_i) \leq \bar{c}[2\bar{V}_i(\beta_i) + 2R_{i,T}(\beta_i)]$ *for all* $\beta_i \in \mathcal{B}_i$ *w.p.a.1, where* $\underline{c}$ *and* $\bar{c}$ *are some generic positive constants that do not depend on* i *with* $0 < \underline{c} < 1 < \bar{c} < \infty$.

Proof. Noting that $W_{iNT} = W_i + o_P(1)$ uniformly in i under Assumption B1(iv), we have w.p.a.1

$$\underline{c}\left[\frac{1}{T}\sum_{t=1}^T \rho(\xi_{it}, \beta_i)\right]' W_i \left[\frac{1}{T}\sum_{t=1}^T \rho(\xi_{it}, \beta_i)\right] \leq V_{iNT}(\beta_i) \leq \bar{c}\left[\frac{1}{T}\sum_{t=1}^T \rho(\xi_{it}, \beta_i)\right]' W_i \left[\frac{1}{T}\sum_{t=1}^T \rho(\xi_{it}, \beta_i)\right] \tag{B.1}$$

for all $\beta_i \in \mathcal{B}_i$. By the positive definiteness of W_i and the matrix version of the Cauchy-Schwarz inequality, we can readily show that

$$(a-b)' W_i (a-b) \geq \frac{1}{2} a' W_i a - b' W_i b \text{ and } (a-b)' W_i (a-b) \leq 2a' W_i a + 2b' W_i b$$

for any conformable vectors a and b. Taking $a = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\rho\left(\xi_{it},\beta_i\right)\right]$ and $b = \frac{1}{T}\sum_{t=1}^{T}\{\rho\left(\xi_{it},\beta_i\right) - \mathbb{E}\left[\rho\left(\xi_{it},\beta_i\right)\right]\}$, we have

$$\left[\frac{1}{T}\sum_{t=1}^{T}\rho\left(\xi_{it},\beta_i\right)\right]' W_i \left[\frac{1}{T}\sum_{t=1}^{T}\rho\left(\xi_{it},\beta_i\right)\right] \geq \frac{1}{2}\bar{V}_i\left(\beta_i\right) - R_{i,T}\left(\beta_i\right), \text{ and} \tag{B.2}$$

$$\left[\frac{1}{T}\sum_{t=1}^{T}\rho\left(\xi_{it},\beta_i\right)\right]' W_i \left[\frac{1}{T}\sum_{t=1}^{T}\rho\left(\xi_{it},\beta_i\right)\right] \leq 2\bar{V}_i\left(\beta_i\right) + 2R_{i,T}\left(\beta_i\right). \tag{B.3}$$

Combining (B.1)-(B.3) yields the desired results. ■

Proof of Theorem 3.1. (i) Let $Q^{(K_0)}_{2iNT,\lambda_2}\left(\beta_i,\boldsymbol{\alpha}\right) = V_{iNT}\left(\beta_i\right) + \lambda_2\Pi_{k=1}^{K_0}\left\|\beta_i - \alpha_k\right\|$. By the definition of $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\alpha}}$ and the fact that $Q^{(K_0)}_{2NT,\lambda_2}\left(\boldsymbol{\beta},\boldsymbol{\alpha}\right) = \frac{1}{N}\sum_{i=1}^{N}Q^{(K_0)}_{2iNT,\lambda_2}\left(\beta_i,\boldsymbol{\alpha}\right)$, we have

$$\begin{aligned} & Q_{2iNT,\lambda_2}(\tilde{\beta}_i,\tilde{\boldsymbol{\alpha}}) - Q_{2iNT,\lambda_2}\left(\beta_i^0,\tilde{\boldsymbol{\alpha}}\right) \\ = \; & V_{iNT}(\tilde{\beta}_i) - V_{iNT}\left(\beta_i^0\right) + \lambda_2\left\{\Pi_{k=1}^{K}\left\|\tilde{\beta}_i - \tilde{\alpha}_k\right\| - \Pi_{k=1}^{K}\left\|\beta_i^0 - \tilde{\alpha}_k\right\|\right\} \leq 0. \end{aligned} \tag{B.4}$$

By Lemma B.1 and Assumptions B1(i) and (iv), we have that $V_{iNT}(\tilde{\beta}_i) \geq \underline{c}[\frac{1}{2}\bar{V}_i(\tilde{\beta}_i) - \tilde{R}_{i,T}]$ and $V_{iNT}\left(\beta_i^0\right) \leq \bar{c}\left[2\bar{V}_i\left(\beta_i^0\right) + 2R^0_{i,T}\right] = 2\bar{c}R^0_{i,T}$ w.p.a.1, where $\tilde{R}_{i,T} = R_{i,T}(\tilde{\beta}_i)$ and $R^0_{i,T} = R_{i,T}(\beta_i^0)$. It follows that $\underline{c}[\frac{1}{2}\bar{V}_i(\tilde{\beta}_i) - \tilde{R}_{i,T}] - 2\bar{c}R^0_{i,T} + \lambda_2\left\{\Pi_{k=1}^{K}\left\|\tilde{\beta}_i - \tilde{\alpha}_k\right\| - \Pi_{k=1}^{K}\left\|\beta_i^0 - \tilde{\alpha}_k\right\|\right\} \leq 0$, which can be rewritten as

$$\bar{V}_i\left(\tilde{\beta}_i\right) \leq \frac{2}{\underline{c}}\left[2\bar{c}R^0_{i,T} + \underline{c}\tilde{R}_{i,T} - \lambda_2\left(\Pi_{k=1}^{K}\left\|\tilde{\beta}_i - \tilde{\alpha}_k\right\| - \Pi_{k=1}^{K}\left\|\beta_i^0 - \tilde{\alpha}_k\right\|\right)\right]. \tag{B.5}$$

Using arguments like those applied to obtain (A.2) and (A.4), we have

$$\left|\Pi_{k=1}^{K}\left\|\tilde{\beta}_i - \alpha_k\right\| - \Pi_{k=1}^{K}\left\|\beta_i^0 - \alpha_k\right\|\right| \leq C_{K_0NT}\left(\boldsymbol{\alpha}\right)\left(\left\|\tilde{\beta}_i - \beta_i^0\right\| + 2\left\|\tilde{\beta}_i - \beta_i^0\right\|^2\right). \tag{B.6}$$

Noting that $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\rho\left(\xi_i,\beta_i\right)\right] = -\bar{Q}_{i,z\Delta x}\left(\beta_i - \beta_i^0\right)$, we have

$$\max_{1\leq i\leq N}\bar{V}_i(\tilde{\beta}_i) = \max_{1\leq i\leq N}\left(\tilde{\beta}_i - \beta_i^0\right)'\bar{Q}'_{i,z\Delta x}W_i\bar{Q}_{i,z\Delta x}\left(\tilde{\beta}_i - \beta_i^0\right) \geq \underline{c}_{1NT}\max_{1\leq i\leq N}\left\|\tilde{\beta}_i - \beta_i^0\right\|^2 \tag{B.7}$$

where $\underline{c}_{1NT} \equiv \min_{1\leq i\leq N}\mu_{\min}\left(\bar{Q}'_{i,z\Delta x}W_i\bar{Q}_{i,z\Delta x}\right)$ satisfies that $\liminf_{(N,T)\to\infty}\underline{c}_{1NT} \geq \underline{c}_W\underline{c}_{\bar{Q}} > 0$ by Assumptions B1(iii)-(iv). Combining (B.5)-(B.7) yields

$$\underline{c}_{1NT}\left\|\tilde{\beta}_i - \beta_i^0\right\|^2 \leq \frac{2}{\underline{c}}\left[2\bar{c}R^0_{i,T} + \underline{c}\tilde{R}_{i,T} + \lambda_2\tilde{C}_{K_0}\left(\left\|\tilde{\beta}_i - \beta_i^0\right\| + 2\left\|\tilde{\beta}_i - \beta_i^0\right\|^2\right)\right],$$

or equivalently, $(\underline{c}_{1NT} - \frac{4}{\underline{c}}\lambda_2\tilde{C}_{K_0})\left\|\tilde{\beta}_i - \beta_i^0\right\|^2 \leq \frac{2}{\underline{c}}\left[2\bar{c}R^0_{i,T} + \underline{c}\tilde{R}_{i,T} + \lambda_2\tilde{C}_{K_0}\left\|\tilde{\beta}_i - \beta_i^0\right\|\right]$, where $\tilde{C}_{K_0} = C_{K_0NT}\left(\tilde{\boldsymbol{\alpha}}\right)$. It follows that

$$\left\|\tilde{\beta}_i - \beta_i^0\right\| \leq \frac{\frac{2}{\underline{c}}\lambda_2\tilde{C}_{K_0} + \left[\left(\frac{2}{\underline{c}}\lambda_2\tilde{C}_{K_0}\right)^2 + \frac{8}{\underline{c}}(\underline{c}_{1NT} - \frac{4}{\underline{c}}\lambda_2\tilde{C}_{K_0})\left(2\bar{c}R^0_{i,T} + \underline{c}\tilde{R}_{i,T}\right)\right]^{1/2}}{2\left(\underline{c}_{1NT} - \frac{4}{\underline{c}}\lambda_2\tilde{C}_{K_0}\right)} = O_P\left(\eta_{2NT}\right), \tag{B.8}$$

where $\eta_{2NT} \equiv T^{-1/2} + \lambda_2$. Further, noting that $\frac{1}{N}\sum_{i=1}^{N} \tilde{R}_{i,T}^2 = O_P(1)$ and $\frac{1}{N}\sum_{i=1}^{N}(R_{i,T}^0)^2 = O_P(1)$ under Assumptions B1(ii) and (iv), we can readily show that $\frac{1}{N}\sum_{i=1}^{N}\left\|\tilde{\beta}_i - \beta_i^0\right\|^2 = O_P\left(\eta_{2NT}^2\right)$. As in the proof of Theorem 2.1(ii), we can further demonstrate that $\frac{1}{N}\sum_{i=1}^{N}\left\|\tilde{\beta}_i - \beta_i^0\right\|^2 = O_P\left(T^{-1}\right)$.

The proof of (iii) is completely analogous to that of Theorem 2.1(iii), now using the facts that $\left|P_{NT}(\tilde{\boldsymbol{\beta}}, \boldsymbol{\alpha}) - P_{NT}\left(\boldsymbol{\beta}^0, \boldsymbol{\alpha}\right)\right| = O_P\left(T^{-1}\right)$ and that $0 \geq P_{NT}(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}) - P_{NT}(\tilde{\boldsymbol{\beta}}, \boldsymbol{\alpha}^0)$. ■

Proof of Theorem 3.2. (i) Fix $k \in \{1, ..., K_0\}$. By the consistency of $\tilde{\alpha}_k$ and $\tilde{\beta}_i$ with α_k^0 for $i \in G_k^0$, we have $\tilde{\beta}_i - \tilde{\alpha}_l \xrightarrow{P} \alpha_k^0 - \alpha_l^0 \neq 0$ for all $l \neq k$. It follows that w.p.a.1 $\left\|\tilde{\beta}_i - \tilde{\alpha}_l\right\| \neq 0$ for all $i \in G_k^0$ and $l \neq k$. Now, suppose that $\left\|\tilde{\beta}_i - \tilde{\alpha}_k\right\| \neq 0$ for some $i \in G_k^0$. Then the first order condition (with respect to β_i) for the minimization problem in (3.2) implies that

$$
\begin{aligned}
0 &= -2\tilde{Q}_{i,z\Delta x}' W_{iNT} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} z_{it}\left(\Delta y_{it} - \tilde{\beta}_i' \Delta x_{it}\right) + \sqrt{T}\lambda_2 \sum_{j=1}^{K_0} \frac{\tilde{\beta}_i - \tilde{\alpha}_j}{\left\|\tilde{\beta}_i - \tilde{\alpha}_j\right\|} \Pi_{l=1,l\neq j}^{K_0} \left\|\tilde{\beta}_i - \tilde{\alpha}_l\right\| \\
&= -2\tilde{Q}_{i,z\Delta x}' W_{iNT} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} z_{it}\Delta u_{it} + \left(2\tilde{Q}_{i,z\Delta x}' W_{iNT} \tilde{Q}_{i,z\Delta x} + \frac{\lambda_2 \tilde{c}_{ki}}{\left\|\tilde{\beta}_i - \tilde{\alpha}_k\right\|}\right) \sqrt{T}\left(\tilde{\beta}_i - \tilde{\alpha}_k\right) \\
&\quad + 2\tilde{Q}_{i,z\Delta x}' W_{iNT} \tilde{Q}_{i,z\Delta x} \sqrt{T}\left(\tilde{\alpha}_k - \beta_i^0\right) + \sqrt{T}\lambda_2 \sum_{j=1,j\neq k}^{K_0} \frac{\tilde{\beta}_i - \tilde{\alpha}_j}{\left\|\tilde{\beta}_i - \tilde{\alpha}_j\right\|} \Pi_{l=1,l\neq j}^{K_0} \left\|\tilde{\beta}_i - \tilde{\alpha}_l\right\| \\
&\equiv -\tilde{B}_{i1} + \tilde{B}_{i2} + \tilde{B}_{i3} + \sum_{j=1,j\neq k}^{K_0} \tilde{B}_{i4,j}, \text{ say,}
\end{aligned}
$$

where $\tilde{c}_{ki} = \Pi_{l=1,l\neq k}^{K_0}\left\|\tilde{\beta}_i - \tilde{\alpha}_l\right\| \xrightarrow{P} c_k^0 \equiv \Pi_{l=1,l\neq k}^{K_0}\left\|\alpha_k^0 - \alpha_l^0\right\| > 0$ for any $i \in G_k^0$ by Theorem 3.1.

As in the proof of Theorem 2.2, we can readily show that $\tilde{B}_{i1} = O_P(1)$, $\tilde{B}_{i3} = O_P(1)$, and $\tilde{B}_{i4,j} = \sqrt{T}\lambda_2 O_P\left(T^{-1/2} + \lambda_2\right) = O_P(1)$ for each $i \in G_k^0$ and $j = 1, ..., K_0$. Let $\tilde{R}_i = \tilde{B}_{i3} + \sum_{j=1,j\neq k}^{K_0} \tilde{B}_{i4,j}$ and $\tilde{\delta}_{1i} \equiv \mu_{\min}(\tilde{Q}_{i,z\Delta x}' W_{iNT} \tilde{Q}_{i,z\Delta x})$. Noting that $\left(\tilde{\beta}_i - \tilde{\alpha}_k\right)' \tilde{B}_{i2} \geq 2\tilde{\delta}_{1i}\sqrt{T}\left\|\tilde{\beta}_i - \tilde{\alpha}_k\right\|^2 + \sqrt{T}\lambda_2 \tilde{c}_{ki}$ and $\left|\left(\tilde{\beta}_i - \tilde{\alpha}_k\right)' \tilde{R}_i\right| = O_P(\lambda_2)$, we have $\left(\tilde{\beta}_i - \tilde{\alpha}_k\right)' \tilde{B}_{i2} - \left|\left(\tilde{\beta}_i - \tilde{\alpha}_k\right)' \tilde{R}_i\right| \geq \left(\tilde{\beta}_i - \tilde{\alpha}_k\right)' \tilde{B}_{i2}/2$ as $(N, T) \to \infty$. It follows that by Assumption B2(i)

$$
\begin{aligned}
P(\tilde{E}_{kNT,i}) &= P\left(i \notin \tilde{G}_k \mid i \in G_k^0\right) = P\left(\tilde{B}_{i1} = \tilde{B}_{i2} + \tilde{R}_i\right) \\
&\leq P\left(\left|\left(\tilde{\beta}_i - \tilde{\alpha}_k\right)' \tilde{B}_{i1}\right| \geq \left|\left(\tilde{\beta}_i - \tilde{\alpha}_k\right)' \tilde{B}_{i2} + \left(\tilde{\beta}_i - \tilde{\alpha}_k\right)' \tilde{R}_i\right|\right) \\
&\leq P\left(\left\|\tilde{\beta}_i - \tilde{\alpha}_k\right\| \left\|\tilde{B}_{i1}\right\| \geq \left(\tilde{\beta}_i - \tilde{\alpha}_k\right)' \tilde{B}_{i2}/2\right) \\
&\leq P\left(\left\|\tilde{B}_{i1}\right\| \geq \tilde{\delta}_{1i}\sqrt{T}\left\|\tilde{\beta}_i - \tilde{\alpha}_k\right\| + \frac{\sqrt{T}\lambda_2 \tilde{c}_{ki}}{2\left\|\tilde{\beta}_i - \tilde{\alpha}_k\right\|}\right)
\end{aligned}
$$

$$\leq P\left(\left\|\tilde{B}_{i1}\right\| \geq \sqrt{2\tilde{\delta}_{1i}\tilde{c}_{ki}T\lambda_2}\right) \to 0 \text{ as } (N,T) \to \infty,$$

where we use the fact that $\tilde{c}_{ki} \xrightarrow{P} c_k^0$ for $i \in G_k^0$ and $\tilde{\delta}_{1i} \xrightarrow{P} \mu_{\min}(\bar{Q}'_{i,z\Delta x}W_i\bar{Q}_{i,z\Delta x}) \geq \mu_{\min}(\bar{Q}'_{i,z\Delta x}\bar{Q}_{i,z\Delta x})$ $\mu_{\min}(W_i) > 0$ by Assumptions B1(iii)-(iv). It follows that $P\left(\left\|\tilde{\beta}_i - \tilde{\alpha}_k\right\| = 0 \mid i \in G_k^0\right) \to 1$ as $(N,T) \to \infty$. Now, observe that $P(\cup_{k=1}^{K_0}\hat{E}_{kNT}) \leq \sum_{k=1}^{K_0} P(\hat{E}_{kNT}) \leq \sum_{k=1}^{K_0}\sum_{i\in G_k^0} P(\hat{E}_{kNT,i})$ and

$$\begin{aligned}
\sum_{k=1}^{K_0}\sum_{i\in G_k^0} P\left(\tilde{E}_{kNT,i}\right) &\leq \sum_{k=1}^{K_0}\sum_{i\in G_k^0} P\left(\left\|\tilde{B}_{i1}\right\| \geq \sqrt{2\tilde{\delta}_i\tilde{c}_{ki}T\lambda_2}\right) \\
&\leq N \max_{1\leq i\leq N} P\left(\left\|\tilde{Q}'_{i,z\Delta x}W_{iNT}\frac{1}{T}\sum_{t=1}^{T} z_{it}\Delta u_{it}\right\| \geq \sqrt{\tilde{\delta}_{1i}\tilde{c}_{ki}\lambda_2/2}\right) \\
&\leq N \max_{1\leq i\leq N} P\left(\left\|\frac{1}{T}\sum_{t=1}^{T} z_{it}\Delta u_{it}\right\| \geq \sqrt{\frac{\tilde{\delta}_{1i}\tilde{c}_{ki}\lambda_2}{2\tilde{\delta}_{2i}}}\right) \\
&\to 0 \text{ by Assumption B2(ii)},
\end{aligned}$$

where $\tilde{\delta}_{2i} \equiv \left\|\tilde{Q}'_{i,z\Delta x}W_{iNT}\right\|^2 \xrightarrow{P} \text{tr}\left(\bar{Q}'_{i,z\Delta x}W_iW_i\bar{Q}_{i,z\Delta x}\right) \leq [\mu_{\max}(W_i)]^2\left\|\bar{Q}_{i,z\Delta x}\right\|^2 < \infty$ by Assumption B1(iii)-(iv). Consequently, we have shown (i).

(ii) The proof of (i) is almost identical to that of Theorem 2.2(ii) and is omitted. ■

Proof of Theorem 3.4. The proof follows closely from that of Theorem 2.4 and we only sketch it. Based on the subdifferential calculus, the KKT conditions for the minimization of (3.2) are that for each $i = 1, ..., N$ and $k = 1, ..., K_0$,

$$\begin{aligned}
\mathbf{0}_{p\times 1} &= -2\tilde{Q}'_{i,z\Delta x}W_{iNT}\frac{1}{NT}\sum_{t=1}^{T} z_{it}\left(\Delta y_{it} - \tilde{\beta}'_i\Delta x_{it}\right) + \frac{\lambda_2}{N}\sum_{j=1}^{K_0}\tilde{e}_{ij}\Pi_{l=1,l\neq j}^{K_0}\left\|\tilde{\beta}_i - \tilde{\alpha}_l\right\|, \text{ and} \\
\mathbf{0}_{p\times 1} &= \frac{\lambda_1}{N}\sum_{i=1}^{N}\tilde{e}_{ik}\Pi_{l=1,l\neq k}^{K_0}\left\|\tilde{\beta}_i - \tilde{\alpha}_l\right\|,
\end{aligned}$$

where $\tilde{e}_{ij} = \frac{\tilde{\beta}_i - \tilde{\alpha}_j}{\left\|\tilde{\beta}_i - \tilde{\alpha}_j\right\|}$ if $\left\|\tilde{\beta}_i - \tilde{\alpha}_j\right\| \neq 0$ and $\|\tilde{e}_{ij}\| \leq 1$ if $\left\|\tilde{\beta}_i - \tilde{\alpha}_j\right\| = 0$. Fix $k \in \{1, ..., K_0\}$. As in the proof of Theorem 2.4, we can show that $\frac{2}{NT}\sum_{i\in\tilde{G}_k}\tilde{Q}'_{i,z\Delta x}W_{iNT}\sum_{t=1}^{T} z_{it}\left(\Delta y_{it} - \tilde{\alpha}'_k\Delta x_{it}\right) + \frac{\lambda_2}{N}\sum_{i\in\tilde{G}_0}\tilde{e}_{ik}\,\Pi_{l=1,l\neq k}^{K_0}\left\|\tilde{\beta}_i - \tilde{\alpha}_l\right\| = \mathbf{0}_{p\times 1}$ w.p.a.1. It follows that

$$\begin{aligned}
\tilde{\alpha}_k &= \left(\frac{1}{N}\sum_{i\in\tilde{G}_k}\tilde{Q}'_{i,z\Delta x}W_{iNT}\tilde{Q}_{i,z\Delta x}\right)^{-1}\frac{1}{NT}\sum_{i\in\tilde{G}_k}\tilde{Q}'_{i,z\Delta x}W_{iNT}\sum_{t=1}^{T} z_{it}\Delta y_{it} \\
&\quad + \left(\frac{1}{N}\sum_{i\in\tilde{G}_k}\tilde{Q}'_{i,z\Delta x}W_{iNT}\tilde{Q}_{i,z\Delta x}\right)^{-1}\frac{\lambda_2}{2N}\sum_{i\in\tilde{G}_0}\tilde{e}_{ik}\Pi_{l=1,l\neq k}^{K_0}\left\|\tilde{\beta}_i - \tilde{\alpha}_l\right\| \equiv \tilde{\alpha}_{1k} + \tilde{\mathcal{R}}_k, \text{ say.}
\end{aligned}$$

By Theorem 3.2, we can readily show that $P\left(\sqrt{NT}\left\|\tilde{\mathcal{R}}_k\right\| \geq \epsilon\right) = o(1)$ for any $\epsilon > 0$, and

$$\sqrt{N_k T}\left(\tilde{\alpha}_{1k} - \alpha_k^0\right) = \left(\frac{1}{N_k}\sum_{i\in G_k^0} \tilde{Q}_{i,z\Delta x}' W_{iNT} \tilde{Q}_{i,z\Delta x}\right)^{-1} \frac{1}{\sqrt{N_k T}} \sum_{i\in G_k^0} \tilde{Q}_{i,z\Delta x}' W_{iNT} \sum_{t=1}^{T} z_{it}\Delta u_{it} + o_P(1).$$

Under Assumptions B1(iv) and B3(i)-(ii), we have $\frac{1}{N_k}\sum_{i\in G_k^0} \tilde{Q}_{i,z\Delta x}' W_{iNT} \tilde{Q}_{i,z\Delta x} = \frac{1}{N_k}\sum_{i\in G_k^0} \bar{Q}_{i,z\Delta x}' W_i \bar{Q}_{i,z\Delta x} + o_P(1) = A_k + o_P(1)$. Then the result follows from Assumption B3(iii) and Slutsky theorem. ■

Proof of Theorem 3.5. Following the proof of Theorem 2.5, we can readily show that

$$\begin{aligned}
\sqrt{N_k T}\left(\tilde{\alpha}_{\tilde{G}_k} - \alpha_k^0\right) &= \left[\tilde{Q}_{z\Delta x}^{(k)\prime} W_{NT}^{(k)} \tilde{Q}_{z\Delta x}^{(k)}\right]^{-1} \tilde{Q}_{z\Delta x}^{(k)\prime} W_{NT}^{(k)} \sqrt{N_k T}\tilde{Q}_{z\Delta u}^{(k)} + o_P(1) \\
&= \left[Q_{z\Delta x,NT}^{(k)\prime} W_{NT}^{(k)} Q_{z\Delta x,NT}^{(k)}\right]^{-1} Q_{z\Delta x,NT}^{(k)\prime} W_{NT}^{(k)} \sqrt{N_k T} Q_{z\Delta u,NT}^{(k)} + o_P(1),
\end{aligned}$$

where $\tilde{Q}_{z\Delta u}^{(k)} = \frac{1}{N_k T}\sum_{i\in\tilde{G}_k}\sum_{t=1}^{T} z_{it}\Delta u_{it}$ and $Q_{z\Delta u,NT}^{(k)} = \frac{1}{N_k T}\sum_{i\in G_k^0}\sum_{t=1}^{T} z_{it}\Delta u_{it}$. The results then follow from analogous arguments as used in the proof of Theorem 2.5, Assumption B3, and Slutsky theorem. ■

Proof of Theorem 3.6. The proof is analogous to that of Theorem 2.6 and is omitted. ■

REFERENCE

ALVAREZ, J., AND M. ARELLANO (2003): "The Time Series and Cross-Section Asymptotics of Dynamic Panel Data Estimators," *Econometrica* 71, 1121-1159.

ANDO, T., AND J. BAI (2013): "Panel Data Models with Grouped Factor Structure under Unknown Group Membership," *Working Paper*, Dept. of Economics, Columbia University.

BAI, J. (2009): "Panel Data Models with Interactive Fixed Effects," *Econometrica* 77, 1229-1279.

BAI, J. AND T. ANDO (2013): "Multifactor Asset Pricing with a Large Number of Observable Risk Factors and Unobservable Common and Group-specific Factors," *Working Paper*, Dept. of Economics, Columbia University.

BALTAGI, B. H., G. BRESSON, AND A. PIROTTE (2008): "To Pool or Not to Pool?" In L. Mátyás and P. Sevestre (Eds.), *The Econometrics of Panel Data:Fundamentals and Recent Developments in Theory and Practice*, pp. 517-546, 3rd Edition. Springer-Verlag, Berlin.

BALTAGI, B. H., AND J. M. GRIFFIN (1997): "Pooled Estimators v.s. Their Heterogeneous Counterpart in the Context of Dynamic Demand for Gasoline," *Journal of Econometrics* 77, 303-327.

BELLONI, A., V. CHERNOZHUKOV (2013): "Least Squares after Model Selection in High-Dimensional Sparse Models." *Bernoulli* 19, 521-547.

BERTSEKAS, D. (1995): *Nonlinear Programming.* Athena Scientific, Belmont, MA.

BESTER, C. A. AND C. B. HANSEN (2013): "Grouped Effects Estimators in Fixed Effects Models," *Journal of Econometrics*, forthcoming.

BONHOMME, S. AND E. MANRESA (2014): "Grouped Patterns of Heterogeneity in Panel Data," *Working paper*, CEMFI, Madrid.

BOSWORTH, P, S. COLLINS AND C.M. REINHART (1999): "Capital Flows to Developing Economies: Implications for Saving and Investment," *Brookings papers on economic activity* 30, 143-180.

BREITUNG, J. AND M. H. PESARAN (2008): "Unit Roots and Cointegration in Panels," in L. Mátyás and P. Sevestre (eds.), *The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory and Practice*, pp. 279-322. Springer-Verlag, Berlin.

BROWNING, M., AND J. M. CARRO (2007): "Heterogeneity and Microeconometrics Modelling," In R. Blundell, W. K. Newey and T. Persson (Eds.), *Advances in Economics and Econometrics, Theory and Applications: Ninth World Congress of the Econometric Society*, Volume 3, pp. 45-74. Cambridge University Press, New York.

BROWNING, M., AND J. M. CARRO (2010): "Heterogeneity in Dynamic Discrete Choice Models," *Econometrics Journal* 13, 1-39.

BROWNING, M., AND J. M. CARRO (2014): "Dynamic Binary Outcome Models with Maximal Heterogeneity," *Journal of Econometrics* 178, 805-823.

BROWNING, M., M. EJRNÆS, AND J. ALVAREZ (2010): "Modelling Income Processes with Lots of Heterogeneity," *Review of Economic Studies* 77, 1121-1159.

CARROLL, C AND D.N. WEIL (1994): "Saving and Growth: a Reinterpretation," *Carnegie-Rochester Conference Series on Public Policy* 40, 133-192

DEATON, A. (1990): "Saving in Developing Countries: Theory and Review," *Proceedings of the World Bank Annual Conference on Development Economics,* pp. 61-96, The World Bank, Washington, DC.

DEISSENBERG, C., G. FEICHTINGER, W. SEMMLER, AND F. WIRL (2004): "History Dependence and Global Dynamics in Models with Multiple Equilibria," in W. Barnett, C. Deissenberg, and G. Feichtinger (eds.), *Economic Complexity: Non-linear Dynamics, Multi-agents Economies, and Learning*, pp. 91-124. Elsevier, Amsterdam.

DURLAUF, S. N., P. A. JOHNSON, AND J. R. TEMPLE (2005): "Growth Econometric," in P. Aghion and S. Durlauf (eds), *Handbook of Economic Growth*, Vol 1, pp. 555-677. Amsterdam: Elsevier.

EBERHARDT, M. AND F. TEAL (2011): "Econometrics for Grumblers: a New Look at the Literature on Cross-country Growth Empirics.," *Journal of Economic Surveys* 25, 109-155.

EDWARDS, S (1996): "Why Are Latin America's Savings Rates So Low? An International Comparative Analysis" *Journal of Development Economics* 51, 5-44.

FAN, J., AND R. LI (2001): "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association* 96, 1348-1360.

FAZZARI„ S.M., R. G. HUBBARD, and B.C. PETERSEN (1988): "Financial Constraints and Corporate Investment," *Brookings Papers on Economic Activity* 1, 141-206.

FELDSTEIN, M. (1980): "International Differences in Social Security and Saving," *Journal of Public Economics* 14, 225–244.

GOURIEROUX, C., P. C. B. PHILLIPS, and J. YU (2010): "Indirect Inference for Dynamic Panel Models", *Journal of Econometrics* 157, 68-77.

HAHN, J., AND G. KUERSTEINER (2002): "Asymptotically Unbiased Inference for a Dynamic Panel Model with Fixed Effects When Both n and T are Large," *Econometrica* 70, 1639-1657.

HAHN, J., AND G. KUERSTEINER (2011): "Bias Reduction for Dynamic Nonlinear Panel Models with Fixed Effects," *Econometric Theory* 27, 1152-1191.

HAN, C., P. C. B. PHILLIPS AND D. SUL (2014): "X-Differencing and Dynamic Panel Model Estimation", *Econometric Theory* 30, 201-251.

HSIAO, C. (2003): *Analysis of Panel Data.* Cambridge University Press, Cambridge.

HSIAO, C., AND H. PESARAN (2008): "Random Coefficient Panel Data Models," in L. Matyas and P. Sevestre (Eds.), *The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory and Practice*, pp. 187-216, 3rd Edition. Springer-Verlag, Berlin.

HSIAO, C., H. M. PESARAN, AND A. K. TAHMISCIOGLU (1999): "Bayes Estimation of Short-run Coefficients in Dynamic Panel Data Models," in *Analysis of Panels and Limited Dependent Variable Models*, C. Hsiao, K. Lahiri, L.-F. Lee, and M. H. Pesaran, eds., Cambridge University Press, Cambridge, pp. 268-296.

IM, K. S., M. H. PESARAN, AND Y. SHIN, "Testing for Unit Roots in Heterogeneous Panels," *Journal of Econometrics* 115, 53-74.

KASAHARA, H., AND K. SHIMOTSU (2009): "Nonparametric Identification of Finite Mixture Models of Dynamic Discrete Choices," *Econometrica* 77, 135-175.

KIVIET, J. F. (1995): "On Bias, Inconsistency, and Efficiency of Various Estimators in Dynamic Panel Data Models," *Journal of Econometrics* 68, 53-78.

KNEIP, A., R. C. SICKLES, AND W. SONG (2012): "A New Panel Data Treatment for Heterogeneity in Time Trends", *Econometric Theory* 28, 590–628.

KNIGHT, K., AND W. FU (2000): "Asymptotics for Lasso-Type Estimators," *Annals of Statistics 28*, 1356-1378.

LEE, K., M. H. PESARAN, AND R. SMITH (1997): "Growth and Convergence in a Multicountry Empirical Stochastic Growth Model," *Journal of Applied Econometrics* 12, 357-392.

LEE, Y. (2012): "Bias in Dynamic Panel Models under Time Series Misspecification," *Journal of Econometrics* 169, 54-60.

LEVIN, A., C-F. LIN, AND C-S. J. Chu (2002): "Unit Root Tests in Panel Data: Asymptotic and Finite-sample Properties," *Journal of Econometrics* 108, 1-24.

LI, H., J. ZHANG AND J. ZHANG (2007), "Effects of Longevity and Dependency Rates on Saving and Growth " *Journal of Development Economics* 84, 138-154.

LIAO, Z. (2013): "Adaptive GMM Shrinkage Estimation with Consistent Moment Selection," *Econometric Theory* 29, 857-904.

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.

LOAYZA, N., K. SCHMIDT-HEBBEL AND L. SERVÉN (2000): "Saving in Developing Countries: An Overview," *The World Bank Economic Review* 14, 393-414.

LU, X., AND L. SU (2013): "Shrinkage Estimation of Dynamic Panel Data Models with Interactive Fixed Effects," *Working paper*, Singapore Management University.

MADDALA, G. S., R. P. TROST, H. LI, AND F. JOUTZ (1997): "Estimation of Short-run and Long-run Elasticities of Energy Demand from Panel Data Using Shrinkage Estimators," *Journal of Business and Economic Statistics* 15, 90-100.

MOON, H., B. PERRON, P. C. B. PHILLIPS (2014): "Incidental Parameters and Dynamic Panel Modeling," *Oxford Handbook of Panel Data,* forthcoming.

PESARAN, H., Y. SHIN, AND R. SMITH (1999): "Pooled Mean Group Estimation of Dynamic Heterogeneous Panels," *Journal of the American Statistical Association* 94, 621-634.

PESARAN, H., AND R. SMITH (1995): "Estimating Long-run Relationships from Dynamic Heterogeneous Panels," *Journal of Econometrics* 68, 79-113.

PHILLIPS, P. C. B., AND D. SUL (2007a): "Transition Modeling and Econometric Convergence Tests," *Econometrica* 75, 1771-1855.

PHILLIPS, P. C. B., AND D. SUL (2007b): "Bias in Dynamic Panel Estimation with Fixed Effects, Incidental Trends and Cross Section Dependence," *Journal of Econometrics* 137, 162-188.

PRAKASA RAO, B. L. S. (2009): "Conditional Independence, Conditional Mixing and Conditional Association," *Annals of the Institute of Statistical Mathematics* 61, 441-460.

RODRIK, D (2000): "Saving Transitions," *The World Bank Economic Review* 14, 481-507.

SARAFIDIS, V. AND N. WEBER (2011): "A Partially Heterogenous Framework for Analyzing Panel Data," *Working paper*, University of Sydney.

SU, L., AND Q. CHEN (2013): "Testing Homogeneity in Panel Data Models with Interactive Fixed Effects," *Econometric Theory* 29, 1079-1135.

SUN, Y. (2005): "Estimation and Inference in Panel Structure Models," *Working Paper*, Dept. of Economics, UCSD.

TIBSHIRANI, R. J. (1996): "Regression Shrinkage and Selection via the LASSO," *Journal of the Royal Statistical Society, Series B.* 58, 267-288.

TIBSHIRANI, R., M. SAUNDERS, S. ROSSET, J. ZHU and K. KNIGHT (2005): "Sparsity and Smoothness via the Fused Lasso," *Journal of the Royal Statistical Society, Series B* 67, 91-108.

WANG, H., R. LI, AND C-L. TSAI (2007): "Tuning Parameter Selectors for the Smoothly Clipped Absolute Deviation Method," *Biometrika* 94, 553-568.

WHITE, H. (2001): *Asymptotic Theory for Econometricians.* Emerald, UK.

YUAN, M., AND Y. LIN (2006): "Model Selection and Estimation in Regression with Grouped Variables," *Journal of the Royal Statistical Society, Series B.* 68, 49-67.

ZELDES, S. P. (1989): "Consumption and Liquidity Constraints: An Empirical Investigation," *Journal of Political Economy* 97, 305-346.

ZHANG, Y., L. SU, AND P. C. B. PHILLIPS (2012): "Testing for Common Trends in Semiparametric Panel Data Models with Fixed Effects," *Econometrics Journal* 15, 56-100.

ZOU, H. (2006): "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association* 101, 1418-1429.

Supplementary Material On

"**Identifying Latent Structures in Panel Data**"

Liangjun Su[a], Zhentao Shi[b], and Peter C. B. Phillips[c]

[a] *School of Economics, Singapore Management University*

[b] *Department of Economics, Yale University*

[c] *Yale University, University of Auckland, University of Southampton & Singapore Management University*

THIS APPENDIX PROVIDES SOME ADDITIONAL RESULTS FOR THE ABOVE PAPER.

C Some Primitive Assumptions and Technical Lemmas

This appendix presents some primitive assumptions that ensure the high level conditions in Assumptions A1(ii) and A2(ii) hold for non-dynamic panel data models. We then discuss primitive conditions to ensure that they hold for dynamic panels. The verification of Assumption B2(ii) is similar.

ASSUMPTION C1 *(i) For each* $i = 1, ..., N$, $\{(x_{it}, u_{it}) : t = 1, 2, ...\}$ *is strong mixing with mixing coefficients* $\{\alpha_i(\cdot)\}$. $\alpha(\cdot) \equiv \max_{1\le i\le N} \alpha_i(\cdot)$ *satisfies* $\alpha(\tau) \le c_\alpha \rho^\tau$ *for some* $c_\alpha = O(1)$ *and* $\rho \in (0, 1)$. $\mathbb{E}(x_{it}u_{it}) = 0$ *for each* i *and* t.

(ii) There exists a constant $\underline{c}_{\tilde{x}\tilde{x}}$ *such that* $0 < \underline{c}_{\tilde{x}\tilde{x}} \le \min_{1\le i\le N} \mu_{\min}(Q_{i,\tilde{x}\tilde{x}})$.

(iii) Either one of the following two conditions is satisfied: (a) $\sup_{i\ge1}\sup_{t\ge1}\mathbb{E}\|x_{it}\|^{4q} \le C$ *and* $\sup_{i\ge1}\sup_{t\ge1}\mathbb{E}\|x_{it}u_{it}\|^{2q} \le C$ *for some* $q > 1$ *and* $C < \infty$; *(b) There exist three constants* c_{xx}, c_{xu}, *and* c_u *such that* $\sup_{i\ge1}\sup_{t\ge1}\mathbb{E}[\exp(c_{xx}\|x_{it}\|^{2\gamma})] \le C_\gamma$, $\sup_{i\ge1}\sup_{t\ge1}\mathbb{E}[\exp(c_{xx}\|x_{it}u_{it}\|^{\gamma})] \le C_\gamma$, *and* $\sup_{i\ge1}\sup_{t\ge1}\mathbb{E}[\exp(c_u\|u_{it}\|^{\gamma})] \le C_\gamma$ *for some* $C_\gamma < \infty$ *and* $\gamma \in (0, \infty]$.

(iv) T *satisfies one of the following two conditions: (a)* $T/N^\epsilon \to (0, \infty]$ *for* $\epsilon > 1/(2q-1)$ *if C1(iii.a) is satisfied; (b)* $T/(\ln N)^{(1+\gamma)/\gamma} \to \infty$ *if C1(iii.b) is satisfied.*

(v) T *satisfies one of the following two conditions: (a)* $T\lambda_1\{(\ln N)^{-1} + T(NT)^{-1/q}(\ln T)^{-4}(\ln N)^{-2}]\} \to \infty$ *if C1(iii.a) is satisfied; (b)* $T\lambda_1\{(\ln N)^{-1} + T[\ln(NT)]^{-2(1+\gamma)/\gamma}\} \to \infty$ *if C1(iii.b) is satisfied.*

C1(i) requires that each individual time series $\{x_{it} : t = 1, 2, ...\}$ be strong-mixing with geometric mixing rate. If $\{x_{it}\}$ are identically distributed for all individuals within the same group, then the sup $\max_{1\le i\le N}$ is effectively taken with respect to the K_0 groups. C1(ii) requires that the matrices $Q_{i,\tilde{x}\tilde{x}}$ be positive definite uniformly in i and the uniformity is required only over the K_0 groups in the case of group-wise identical distributions. The conditions stated in Assumption C1(iii) pertain to two specific cases related to the moments of $\|x_{it}\|^2$ and $x_{it}u_{it}$: part (a) only requires finite $2q$-th moments whereas part (b) requires the existence of exponential moments. By the Markov inequality, part (b) implies that

$$P\left(\|x_{it}\|^2 \ge v\right) \le \exp\left(1 - \left(\frac{v}{K/c_{xx}}\right)^\gamma\right),$$

where $K = \max(1, \ln C_\gamma)$. That is, the distribution of $\|x_{it}\|^2$ has to decay exponentially fast. The case $\gamma = \infty$ in part (b) corresponds to the case where $\|x_{it}\|$ is uniformly bounded. Similar remarks

hold for $\|x_{it}u_{it}\|$ and $\|u_{it}\|$. When combined with C1(i), the conditions in C2(iii) allow us to apply some exponential inequalities for strong mixing processes; see, e.g., Merlevède, Peilgrad, and Rio (2009, 2011). C1(iv) and (v) are needed to verify Assumption A1(ii) and A2(ii), respectively.

Lemma C.1 *Let $\{\xi_t, t = 1, 2, ...\}$ be a zero-mean strong mixing process, not necessarily stationary, with the mixing coefficients satisfying $\alpha(\tau) \leq c_\alpha \rho^\tau$ for some $c_\alpha > 0$ and $\rho \in (0, 1)$.*

(i) If $\sup_{1\leq t\leq T} |\xi_t| \leq M_T$, then there exists a constant C_0 depending on c_α and ρ such that for any $T \geq 2$ and $\epsilon > 0$,

$$P\left(\left|\sum_{t=1}^{T} \xi_t\right| > \epsilon\right) \leq \exp\left(-\frac{C_0\epsilon^2}{v_0^2 T + M_T^2 + \epsilon M_T (\ln T)^2}\right),$$

where $v_0^2 = \sup_{t\geq 1}\left[Var(\xi_t) + 2\sum_{s=t+1}^{\infty} |Cov(\xi_t, \xi_s)|\right]$.

(ii) If $\sup_{t\geq 1} P(|\xi_t| > v) \leq \exp(1 - (v/b)^\gamma)$ for some $b \in (0, \infty)$ and $\gamma \in (0, \infty]$, then there exist constants C_1 and C_2 depending only on b, c_α, ρ, and γ such that for any $T \geq 4$ and $\epsilon \geq C_0(\ln T)^{\eta_0}$ with η_0, $C_0 > 0$,

$$P_{\mathcal{D}}\left(\left|\sum_{t=1}^{T} \xi_t\right| > \epsilon\right) \leq (T + 1)\exp\left(-\frac{\epsilon^{\frac{\gamma}{1+\gamma}}}{C_1}\right) + \exp\left(-\frac{\epsilon^2}{TC_2}\right).$$

Proof. (i) Merlevède, Peilgrad, and Rio (2009, Theorem 2) prove (i) under the condition $\alpha(\tau) \leq \exp(-2c\tau)$ for some $c > 0$. If $c_\alpha = 1$, we can take $\rho = \exp(-2c)$ and apply the theorem to obtain the claim in (i). Other values of c_α do not alter the conclusion.

(ii) Merlevède, Peilgrad, and Rio (2011, Theorem 1) prove a result that is more general than that in (ii) under the condition $\alpha(\tau) \leq \exp(-c_1\tau^{\gamma_1})$ for some c_1, $\gamma_1 > 0$. If $c_\alpha = 1$ and $\gamma_1 = 1$, we can take $\rho = \exp(-2c_1)$ and apply the theorem to obtain the claim in (ii). Other values of c_α do not alter the conclusion. ■

Lemma C.2 *Let $\hat{Q}_{i,\tilde{x}\tilde{x}} \equiv T^{-1}\sum_{t=1}^{T} \tilde{x}_{it}\tilde{x}_{it}'$. Suppose that Assumptions C1(i)-(iii) hold.*

(i) If C1(iv) holds, then $\min_{1\leq i\leq N} \mu_{\min}(\hat{Q}_{i,\tilde{x}\tilde{x}}) \geq \min_{1\leq i\leq N} \mu_{\min}(Q_{i,\tilde{x}\tilde{x}}) - o_P(1)$;

(ii) If C1(v) holds, then $N \max_{1\leq i\leq N} P\left(\left\|\frac{1}{T}\sum_{t=1}^{T} \tilde{x}_{it}\tilde{u}_{it}\right\| \geq c\sqrt{\lambda_1}\right) \to 0$ as $(N, T) \to \infty$.

Proof. (i) By the Weyl inequality and the fact that $|\mu_{\max}(A)| \leq \|A\|$ for any symmetric matrix A, we have

$$\mu_{\min}(\hat{Q}_{i,\tilde{x}\tilde{x}}) \geq \mu_{\min}(Q_{i,\tilde{x}\tilde{x}}) - \left\|\hat{Q}_{i,\tilde{x}\tilde{x}} - Q_{i,\tilde{x}\tilde{x}}\right\|.$$

We are left to show that $\max_{1\leq i\leq N} \left\|\hat{Q}_{i,\tilde{x}\tilde{x}} - Q_{i,\tilde{x}\tilde{x}}\right\| = o_P(1)$. Noting that $\hat{Q}_{i,\tilde{x}\tilde{x}} = T^{-1}\sum_{t=1}^{T} \tilde{x}_{it}\tilde{x}_{it}' = T^{-1}\sum_{t=1}^{T} x_{it}x_{it}' - \bar{x}_{i\cdot}\bar{x}_{i}'$, it suffices to show that (i1) $\max_{1\leq i\leq N} \left\|T^{-1}\sum_{t=1}^{T}[x_{it}x_{it}' - \mathbb{E}(x_{it}x_{it}')]\right\| = o_P(1)$ and (i2) $\max_{1\leq i\leq N} \left\|T^{-1}\sum_{t=1}^{T}[x_{it} - \mathbb{E}(x_{it})]\right\| = o_P(1)$. We only prove (i1) as the proof of (i2) is analogous.

We first consider the case where Assumption C1(iii.a) hold. Let $\eta_{NT} = (NT)^{1/(2q)}$. Let ι_{sp} be an arbitrary $p \times 1$ vector with $\|\iota_{sp}\| = 1$ for $s = 1, 2$. Let $\zeta_{it} \equiv \iota_{1p}'[x_{it}x_{it}' - \mathbb{E}(x_{it}x_{it}')]\iota_{2p}$, $\zeta_{1it} \equiv \iota_{1p}'[x_{it}x_{it}'\mathbf{1}_{it} - \mathbb{E}(x_{it}x_{it}'\mathbf{1}_{it})]\iota_{2p}$ and $\zeta_{2it} \equiv \iota_{1p}'[x_{it}x_{it}'\bar{\mathbf{1}}_{it} - \mathbb{E}(x_{it}x_{it}'\bar{\mathbf{1}}_{it})]\iota_{2p}$, where $\mathbf{1}_{it} \equiv$

$\mathbf{1}\{\|x_{it}\|^2 \leq \eta_{NT}\}$ and $\bar{\mathbf{1}}_{it} = 1 - \mathbf{1}_{it}$. Note that $\zeta_{it} = \zeta_{1it} + \zeta_{2it}$. Let $v_i^2 = \sup_{t\geq 1}[\mathrm{Var}(\zeta_{1it}) + 2\sum_{s=t+1}^{\infty}\mathrm{Cov}(\zeta_{1it}, \zeta_{1is})]$ and $\bar{v}^2 = \sup_{N\geq 1}\max_{1\leq i\leq N} v_i^2$. The moment conditions in C1(iii.a) and Davydov inequality ensure that $\bar{v}^2 = O(1)$. By the Boole inequality and Lemma C.1(i), for any $\epsilon > 0$,

$$\begin{aligned} P\left(\max_{1\leq i\leq N}\left|T^{-1}\sum_{t=1}^{T}\zeta_{1it}\right| \geq \epsilon\right) &\leq N\max_{1\leq i\leq N} P\left(\left|\sum_{t=1}^{T}\zeta_{1it}\right| \geq T\epsilon\right) \\ &\leq N\max_{1\leq i\leq N}\exp\left(-\frac{C_0T^2\epsilon^2}{v_i^2T + 4\eta_{NT}^2 + 2T\epsilon\eta_{NT}(\ln T)^2}\right) \\ &\leq \exp\left(-\frac{C_0T^2\epsilon^2}{\bar{v}^2T + 4\left(NT\right)^{1/q} + 2T\epsilon\left(NT\right)^{1/(2q)}(\ln T)^2} + \ln N\right) \\ &\to 0 \text{ as } T\to\infty. \end{aligned}$$

By Assumption C1(iii.a), the Boole and Markov inequalities, and the dominated convergence theorem,

$$\begin{aligned} P\left(\max_{1\leq i\leq N}\left|\frac{1}{T}\sum_{t=1}^{T}\zeta_{2it}\right| \geq \epsilon\right) &\leq P\left(\max_{1\leq i\leq N}\max_{1\leq t\leq T}\|x_{it}\|^2 \geq \eta_{NT}\right) \leq NT\max_{1\leq i\leq N}\max_{1\leq t\leq T} P\left(\|x_{it}\|^2 \geq \eta_{NT}\right) \\ &\leq \frac{NT}{\eta_{NT}^{2q}}\max_{1\leq i\leq N}\max_{1\leq t\leq T}\mathbb{E}\left[\|x_{it}\|^{4q}\mathbf{1}\left\{\|x_{it}\|^2 \geq \eta_{NT}\right\}\right] \to 0 \text{ as } T\to\infty. \end{aligned}$$

Noting that ι_{1p} and ι_{2p} are arbitrary unit vectors, we infer that $\max_{1\leq i\leq N}\left\|\frac{1}{T}\sum_{t=1}^{T}\left[x_{it}x_{it}' - \mathbb{E}\left(x_{it}x_{it}'\right)\right]\right\| = o_P(1)$. Then (i) follows.

Next consider the case where Assumption C1(iii.b) holds. By the Boole inequality and Lemma C.1(ii), for any $\epsilon > 0$,

$$\begin{aligned} P\left(\max_{1\leq i\leq N}\left|T^{-1}\sum_{t=1}^{T}\zeta_{it}\right| \geq \epsilon\right) &\leq N\max_{1\leq i\leq N} P\left(\left|\sum_{t=1}^{T}\zeta_{it}\right| \geq T\epsilon\right) \\ &\leq N\max_{1\leq i\leq N}\left[(T+1)\exp\left(-\frac{(T\epsilon)^{\gamma/(1+\gamma)}}{C_1}\right) + \exp\left(-\frac{(T\epsilon)^2}{TC_2}\right)\right] \\ &\leq \exp\left(-\frac{(T\epsilon)^{\gamma/(1+\gamma)}}{C_1} + \ln(T+1) + \ln N\right) + \exp\left(-\frac{T\epsilon^2}{C_2} + \ln N\right) \\ &\to 0 \text{ as } T\to\infty, \end{aligned}$$

provided $T \gg (\ln N)^{(1+\gamma)/\gamma}$. It follows that $\max_{1\leq i\leq N}\left\|\frac{1}{T}\sum_{t=1}^{T}\left[x_{it}x_{it}' - \mathbb{E}\left(x_{it}x_{it}'\right)\right]\right\| = o_P(1)$.

(ii) Noting that $T^{-1}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{u}_{it} = T^{-1}\sum_{t=1}^{T}x_{it}u_{it} - \bar{x}_{i\cdot}\bar{u}_{i\cdot}$, we prove (ii) by showing that (ii1) $N\max_{1\leq i\leq N}P(||T^{-1}\sum_{t=1}^{T}x_{it}u_{it}|| \geq c\sqrt{\lambda_1}) \to 0$, (ii2) $N\max_{1\leq i\leq N}P(||T^{-1}\sum_{t=1}^{T}[x_{it} - \mathbb{E}(x_{it})]|| \geq c\sqrt{\lambda_1}) \to 0$, and (ii3) $N\max_{1\leq i\leq N}P(||T^{-1}\sum_{t=1}^{T}u_{it}|| \geq c\sqrt{\lambda_1}) \to 0$. We only outline the proof of (ii1) as the other two claims can be proved analogously. If Assumption C1(iii.a) holds, by letting $\varsigma_{it} \equiv \iota_{1p}'[x_{it}u_{it} - \mathbb{E}(x_{it}u_{it})]$, $\varsigma_{1it} \equiv \iota_{1p}'[x_{it}u_{it}\mathbf{1}_{it} - \mathbb{E}(x_{it}u_{it}\mathbf{1}_{it})]$ and $\varsigma_{2it} \equiv \iota_{1p}'[x_{it}u_{it}\bar{\mathbf{1}}_{it} - \mathbb{E}(x_{it}u_{it}\bar{\mathbf{1}}_{it})]$

where now $\mathbf{1}_{it} \equiv \mathbf{1}\{\|x_{it}u_{it}\| \leq \eta_{NT}\}$ and $\bar{\mathbf{1}}_{it} = 1 - \mathbf{1}_{it}$, we have

$$\begin{aligned} P\left(\max_{1\leq i\leq N}\left|\frac{1}{T}\sum_{t=1}^{T}\varsigma_{1it}\right| \geq c\sqrt{\lambda_1}\right) &\leq N\max_{1\leq i\leq N} P\left(\left|\sum_{t=1}^{T}\varsigma_{1it}\right| \geq cT\sqrt{\lambda_1}\right) \\ &\leq \exp\left(-\frac{c^2C_0T^2\lambda_1}{v^2T + 4(NT)^{1/q} + 2cT\sqrt{\lambda_1}(NT)^{1/(2q)}(\ln T)^2} + \ln N\right) \\ &\to 0 \text{ as } T\to\infty, \end{aligned}$$

and $P\left(\max_{1\leq i\leq N}\left|\frac{1}{T}\sum_{t=1}^{T}\varsigma_{2it}\right| \geq c\sqrt{\lambda_1}\right) \leq NT\max_{1\leq i\leq N}\max_{1\leq t\leq T} P(\|x_{it}u_{it}\| \geq \eta_{NT}) \to 0$ as $T\to\infty$. Here $v^2 = \sup_{N\geq 1}\max_{1\leq i\leq N}\sup_{t\geq 1}[\mathrm{Var}(\varsigma_{1it}) + 2\sum_{s=t+1}^{\infty}\mathrm{Cov}(\varsigma_{1it},\varsigma_{1is})] = O(1)$ under Assumption C1(iii.a). Similarly, if Assumption C1(iii.b) holds, then

$$\begin{aligned} & P\left(\max_{1\leq i\leq N}\left|\frac{1}{T}\sum_{t=1}^{T}\varsigma_{it}\right| \geq c\sqrt{\lambda_1}\right) \\ \leq\ & N\max_{1\leq i\leq N} P\left(\left|\sum_{t=1}^{T}\varsigma_{it}\right| \geq cT\sqrt{\lambda_1}\right) \\ \leq\ & N\max_{1\leq i\leq N}\left[(T+1)\exp\left(-\frac{\left(cT\sqrt{\lambda_1}\right)^{\gamma/(1+\gamma)}}{C_1}\right) + \exp\left(-\frac{\left(cT\sqrt{\lambda_1}\right)^2}{TC_2}\right)\right] \\ \leq\ & \exp\left(-\frac{\left(cT\sqrt{\lambda_1}\right)^{\gamma/(1+\gamma)}}{C_1} + \ln(T+1) + \ln N\right) + \exp\left(-\frac{c^2T\lambda_1}{C_2} + \ln N\right) \\ \to\ & 0 \text{ as } T\to\infty, \end{aligned}$$

provided $T\lambda_1\{(\ln N)^{-1} + T[\ln(NT)]^{-2(1+\gamma)/\gamma}\} \to \infty$. ■

Evidently Lemma C.2(i) ensures the second part of Assumption A1(ii) and Lemma C.2(ii) ensures Assumption A2(ii). These results rely on the use of Bernstein-type inequalities for strong mixing processes that are not necessarily stationary.

To verify Assumptions A1(ii) and A2(ii) for dynamic panel data models, we need to distinguish two cases based on whether we treat the fixed effects μ_i in (2.1) as random or not. If we follow Hahn and Kuersteiner (2011) and assume that the individual fixed effects are nonrandom and uniformly bounded, then we can assume that $\{(z_{it}\Delta u_{it}), t \geq 1\}$ is strong mixing for each i and verify the assumptions as above. On the other hand, if we assume that μ_i's are random fixed effects, then the notion of strong mixing is generally no longer appropriate for dynamic models. To appreciate the point, take the simple panel AR(1) model as an example:

$$y_{it} = \rho_0 y_{i,t-1} + \mu_i + u_{it},\ i = 1, ..., N,\ t = 1, \ldots, T, \tag{C.1}$$

where $x_{it} = y_{i,t-1}$ and an example of the IV for $\Delta y_{i,t-1}$ would be $z_{it} = y_{i,t-2}$. Even if $\{u_{it}, t \geq 1\}$ is a strong mixing process, $\{y_{it}, t \geq 1\}$ is generally not so if μ_i is stochastic as the dependence between y_{it} and y_{is} is not asymptotically vanishing as $|t-s|$ passes to infinity. In this case, as Hahn and Kuersteiner (2011) suggest, it is natural to adopt the concept of conditional strong mixing (see, e.g., Prakasa Rao, 2009) where the mixing coefficient is defined by conditioning on the fixed effects. Su and Chen (2013) adopt the latter approach in their study of panel data

models with interactive fixed effects and show that the well known Davydov and Bernstein-type inequalities that hold for strong mixing processes also hold for conditional strong mixing processes. A conditional version of the results in Lemma C.1 are also satisfied where all probabilities are defined by conditioning on the σ-field generated by $(\mu_1, ..., \mu_N)$. Then one can verify Assumptions A1(ii) and A2(ii) by following analogous arguments as used in the proof of Lemma C.2(ii).

D Bias Correction

D.1 Bias Correction for the PLS C-Lasso Estimator

Recall from Theorems 2.4 and 2.5 that the bias takes the form

$$b_{kNT} = \bar{\Phi}_k^{-1} \mathbb{B}_{kNT},$$

where $\bar{\Phi}_k \equiv \frac{1}{N_k T} \sum_{i \in G_k^0} \sum_{t=1}^T \tilde{x}_{it} \tilde{x}_{it}'$, and $\mathbb{B}_{kNT} = \frac{1}{N_k^{1/2} T^{1/2}} \sum_{i \in G_k^0} \sum_{s=1}^T \mathbb{E}(x_{is} \tilde{u}_{is}) = -\frac{1}{N_k^{1/2} T^{3/2}} \sum_{i \in G_k^0} \sum_{s=1}^T \sum_{t=1}^T \mathbb{E}(x_{is} u_{it})$ as $\mathbb{E}(x_{it} u_{it}) = 0$. Let $\hat{u}_{is} = y_{it} - x_{it}' \hat{\alpha}_{\hat{G}_k} - \hat{\mu}_i$ and $\hat{\mu}_i = \frac{1}{T} \sum_{t=1}^T (y_{it} - x_{it}' \hat{\alpha}_{\hat{G}_k})$ for all $i \in \hat{G}_k$.[12] We propose to estimate b_{kNT} by

$$\hat{b}_{kNT} = \hat{\Phi}_k^{-1} \hat{\mathbb{B}}_{kNT}$$

where $\hat{\Phi}_k = \frac{1}{\hat{N}_k T} \sum_{i \in \hat{G}_k} \sum_{t=1}^T \tilde{x}_{it} \tilde{x}_{it}'$ and $\hat{\mathbb{B}}_{kNT} = -\frac{1}{N_k^{1/2} T^{3/2}} \sum_{i \in \hat{G}_k} \sum_{s=1}^T \sum_{t=1}^T k_{M_T}(t,s) x_{is} \hat{u}_{it}$. Here $k_{M_T}(t,s) = k_{M_T}^0(|t-s|)$ and $k_{M_T}^0(u)$ denotes the Bartlett kernel:

$$k_{M_T}^0(u) = (1 - |u| / M_T) \mathbf{1}\{|u| \leq M_T\}.$$

Note that we allow dynamic misspecification here. If one is sure that the model is dynamically correctly specified in the sense that $\mathbb{E}(u_{it} | \mathcal{F}_{i,t-1}) = 0$ where $\mathcal{F}_{i,t-1} = \sigma(u_{i,t-1}, u_{i,t-2}, ...; x_{it}, x_{it-1}, ...)$, one can use the one-sided kernel: $k_{M_T}(t,s) = k_{M_T}^1(s-t)$, where

$$k_{M_T}^1(u) = (1 - u/M_T) \mathbf{1}\{0 \leq u \leq M_T\}.$$

Other choices of kernels are possible. So the bias-corrected PLS C-Lasso estimator is given by

$$\hat{\alpha}_k^{(c)} = \hat{\alpha}_k - \frac{1}{\sqrt{\hat{N}_k T}} \hat{\Phi}_k^{-1} \hat{\mathbb{B}}_{kNT}.$$

Similarly, we can obtain the bias-corrected estimator for the post-Lasso estimator $\hat{\alpha}_{\hat{G}_k}$.

Let $\|A\|_a = \{\mathbb{E}\|A\|^a\}^{1/a}$ for any $a \geq 1$. Let C denote a generic positive constant that does not depend on N and T. We add the following assumption.

ASSUMPTION D1. *(i) For each $i = 1, ..., N$, $\{(x_{it}, u_{it}) : t = 1, 2, ...\}$ is strong mixing with mixing coefficients $\{\alpha_i(\cdot)\}$ such that $\alpha_i(\tau) \leq c_{\alpha,i} \rho^\tau$ for some $c_{\alpha,i} < \infty$ and $\rho \in (0,1)$. $\frac{1}{N_k} \sum_{i \in G_k^0} c_{\alpha,i}^{(2q-1)/(2q)} = O(1)$.*

[12] Observing that $\hat{\alpha}_k - \alpha_k^0 = O_P\left((N_k T)^{-1/2} + T^{-1}\right)$ and $\hat{\alpha}_{\hat{G}_k} - \alpha_k^0 = O_P\left((N_k T)^{-1/2} + T^{-1}\right)$, one can use either estimator in the definition of the residuals. We recommend using the post-Lasso estimator $\hat{\alpha}_{\hat{G}_k}$ because of its better finite sample performance.

(ii) Let $x_i \equiv (x_{i1}, ..., x_{iT})'$ *and* $u_i \equiv (u_{i1}, ..., u_{iT})'$. (x_i, u_i) *are independent across* $i \in G_k^0$ *where* $k = 1, ..., K_0$.

(iii) $\max_{i,t} \mathbb{E} \|x_{it}\|^{4q} < C < \infty$ *and* $\max_{i,t} \mathbb{E} \|u_{it}\|^{4q} < C < \infty$ *for some* $q \geq 1$.

(iv) As $(N, T) \to \infty$, $M_T \to \infty$, $M_T^2/T \to 0$, $M_T^2 N_k/T^3 \to 0$, *and* $N_k^{-1/2} T^{1/2} \sum_{i \in G_k^0} \alpha_i (M_T)^{\frac{2q-1}{2q}} \to 0$ *for each* $k = 1, ..., K_0$.

Assumption D1(i) assumes the usual mixing condition. D1(ii) assumes cross sectional independence to simplify the proof which can be relaxed at the cost of lengthy arguments. D1(iii) assumes moment conditions. The last condition in D1(iv) can be easily ensured under D1(i) because for any $M_T \gg -\frac{2q}{(2q-1)\ln q} \ln(N^{1/2}T^{1/2})$ (e.g., $M_T = \left(\ln(N^{1/2}T^{1/2})\right)^{1+\epsilon}$ for some $\epsilon > 0$), we have

$$\begin{aligned} N_k^{-1/2} T^{1/2} \sum_{i \in G_k^0} \alpha_i (M_T)^{(2q-1)/(2q)} &\leq \left(N_k^{-1} \sum_{i \in G_k^0} c_{\alpha,i}^{(2q-1)/(2q)} \right) N_k^{1/2} T^{1/2} \rho^{M_T(2q-1)/(2q)} \\ &= O(1) \exp\left(\ln\left(N_k^{1/2} T^{1/2} \right) + \frac{(2q-1) M_T}{2q} \ln \rho \right) \to 0. \end{aligned}$$

The first three requirements in D1(iv) can be easily satisfied too. For example, if $N_k \propto T^a$ for some $a < 3$, it suffices to set $M_T \propto T^{1/b}$ for some $b > \max\{2, 2/(3-a)\}$.

Proposition D.1 *Suppose that the conditions in Theorem 2.4 hold. Suppose Assumption D1 holds. Then* $\hat{\Phi}_k^{-1} \hat{\mathbb{B}}_{kNT} - \bar{\Phi}_k^{-1} \mathbb{B}_{kNT} = o_P(1)$.

Proof. Noting that $\hat{\Phi}_k^{-1} \hat{\mathbb{B}}_{kNT} - \bar{\Phi}_k^{-1} \mathbb{B}_{kNT} = (\hat{\Phi}_k^{-1} - \bar{\Phi}_k^{-1}) \mathbb{B}_{kNT} + (\hat{\Phi}_k^{-1} - \bar{\Phi}_k^{-1})(\hat{\mathbb{B}}_{kNT} - \mathbb{B}_{kNT})$ $+ \bar{\Phi}_k^{-1} (\hat{\mathbb{B}}_{kNT} - \mathbb{B}_{kNT})$, $\bar{\Phi}_k^{-1} = O(1)$, and $\mathbb{B}_{kNT} = O\left(\sqrt{N_k/T}\right)$, it suffices to show that (i) $\hat{\Phi}_k - \bar{\Phi}_k = o_P(\nu_{NT})$ and (ii) $\hat{\mathbb{B}}_{kNT} - \mathbb{B}_{kNT} = o_P(1)$, where $\nu_{NT} = \min(1, \sqrt{T/N_k})$.

We first prove (i). Note that

$$\begin{aligned} \hat{\Phi}_k - \bar{\Phi}_k &= \frac{1}{\hat{N}_k T} \sum_{i \in \hat{G}_k} \sum_{t=1}^{T} \tilde{x}_{it} \tilde{x}_{it}' - \frac{1}{N_k T} \sum_{i \in G_k^0} \sum_{t=1}^{T} \tilde{x}_{it} \tilde{x}_{it}' \\ &= \frac{1}{\hat{N}_k T} \left(\sum_{i \in \hat{G}_k} - \sum_{i \in G_k^0} \right) \sum_{t=1}^{T} \tilde{x}_{it} \tilde{x}_{it}' + \frac{N_k - \hat{N}_k}{\hat{N}_k N_k T} \sum_{i \in G_k^0} \sum_{t=1}^{T} \tilde{x}_{it} \tilde{x}_{it}' \\ &\equiv \Phi_{k,1} + \Phi_{k,2}, \text{ say.} \end{aligned}$$

By Corollary 2.3, we can readily show that $\Phi_{k,2} = O_P(N_k^{-1}) = o_P(\nu_{NT})$. For any $\epsilon > 0$, we have by the proof of Theorem 2.2, $P(\|\Phi_{k,1}\| \geq \nu_{NT}\epsilon) \leq P(\hat{F}_{kNT}) + P(\hat{E}_{kNT}) = o(1)$. It follows that $\hat{\Phi}_k - \bar{\Phi}_k = o_P(\nu_{NT})$.

We now prove (ii). We first make the following decomposition:

$$\begin{aligned}
\mathbb{B}_{kNT}-\hat{\mathbb{B}}_{kNT} &= \frac{1}{\hat{N}_k^{1/2}T^{3/2}}\sum_{i\in\hat{G}_k}\sum_{s=1}^{T}\sum_{t=1}^{T}k_{M_T}(t,s)\,x_{is}\hat{u}_{it}-\frac{1}{N_k^{1/2}T^{3/2}}\sum_{i\in G_k^0}\sum_{s=1}^{T}\sum_{t=1}^{T}\mathbb{E}\left(x_{is}u_{it}\right)\\
&= \frac{1}{\hat{N}_k^{1/2}T^{3/2}}\sum_{i\in G_k^0}\sum_{s=1}^{T}\sum_{t=1}^{T}k_{M_T}(t,s)\,x_{is}\hat{u}_{it}-\frac{1}{N_k^{1/2}T^{3/2}}\sum_{i\in G_k^0}\sum_{s=1}^{T}\sum_{t=1}^{T}\mathbb{E}\left(x_{is}u_{it}\right)\\
&\quad +o_P(1)\\
&= \frac{1}{\hat{N}_k^{1/2}T^{3/2}}\sum_{i\in G_k^0}\sum_{s=1}^{T}\sum_{t=1}^{T}k_{M_T}(t,s)\,x_{is}\left(\hat{u}_{it}-u_{it}\right)\\
&\quad +\frac{1}{\hat{N}_k^{1/2}T^{3/2}}\sum_{i\in G_k^0}\sum_{s=1}^{T}\sum_{t=1}^{T}k_{M_T}(t,s)\left[x_{is}u_{it}-\mathbb{E}\left(x_{is}u_{it}\right)\right]\\
&\quad +\frac{N_k^{-1/2}-\hat{N}_k^{-1/2}}{T^{3/2}}\sum_{i\in G_k^0}\sum_{s=1}^{T}\sum_{t=1}^{T}k_{M_T}(t,s)\,\mathbb{E}\left(x_{is}u_{it}\right)\\
&\quad +\frac{1}{N_k^{1/2}T^{3/2}}\sum_{i\in G_k^0}\sum_{s=1}^{T}\sum_{t=1}^{T}\left[1-k_{M_T}(t,s)\right]\mathbb{E}\left(x_{is}u_{it}\right)+o_P(1)\\
&\equiv \hat{B}_{kNT,1}+\hat{B}_{kNT,2}+\hat{B}_{kNT,3}+\hat{B}_{kNT,4}+o_P(1),\ \text{say,}
\end{aligned}$$

where the $o_P(1)$ term arises due to the replacement of $\hat{G}_k$ by G_k^0 and this can be easily justified by using the uniform classification consistency result and arguments as used in the proof of Theorem 2.5. We prove (ii) by demonstrating that $\hat{B}_{kNT,s}=o_P(1)$ for $s=1$, 2, 3, and 4.

We first study $\hat{B}_{kNT,1}$. Noting that $\hat{u}_{it}=y_{it}-x_{it}'\hat{\alpha}_{\hat{G}_k}-\hat{\mu}_i=y_{it}-x_{it}'\hat{\alpha}_{\hat{G}_k}-\frac{1}{T}\sum_{t=1}^{T}(y_{it}-x_{it}'\hat{\alpha}_{\hat{G}_k})$ and $y_{it}=x_{it}'\alpha_k^0+\mu_i+u_{it}$ for $i\in G_k^0$, we have that for $i\in G_k^0$

$$\hat{u}_{it}-u_{it}=y_{it}-x_{it}'\hat{\alpha}_{\hat{G}_k}-\frac{1}{T}\sum_{t=1}^{T}(y_{it}-x_{it}'\hat{\alpha}_{\hat{G}_k})-u_{it}=\tilde{x}_{it}'(\alpha_k^0-\hat{\alpha}_{\hat{G}_k})-\bar{u}_i,$$

where $\bar{u}_i=\frac{1}{T}\sum_{t=1}^{T}u_{it}$. Then

$$\begin{aligned}
\hat{B}_{kNT,1} &= \frac{1}{\hat{N}_k^{1/2}T^{3/2}}\sum_{i\in G_k^0}\sum_{s=1}^{T}\sum_{t=1}^{T}k_{M_T}(t,s)\,x_{is}\tilde{x}_{it}'(\alpha_k^0-\hat{\alpha}_{\hat{G}_k})\\
&\quad -\frac{1}{\hat{N}_k^{1/2}T^{3/2}}\sum_{i\in G_k^0}\sum_{s=1}^{T}\sum_{t=1}^{T}k_{M_T}(t,s)\,x_{is}\bar{u}_i\\
&\equiv B_{kNT,1}(1)-B_{kNT,1}(2),\ \text{say.}
\end{aligned}$$

In view of the fact that $\hat{\alpha}_{\hat{G}_k} - \alpha_k^0 = O_P\left((N_k T)^{-1/2} + T^{-1}\right)$ and $\hat{N}_k = N_k\left(1 + o_P(1)\right)$, we have

$$\begin{aligned}
\|B_{kNT,1}(1)\| &= \frac{1}{\hat{N}_k^{1/2} T^{3/2}} \left\| \sum_{i \in G_k^0} \sum_{s=1}^{T} \sum_{t=1}^{T} k_{M_T}(t,s)\, x_{is} \tilde{x}_{it}' (\alpha_k^0 - \hat{\alpha}_{\hat{G}_k}) \right\| \\
&\leq \frac{N_k T^{1/2}}{\hat{N}_k^{1/2}} \left\| \alpha_k^0 - \hat{\alpha}_{\hat{G}_k} \right\| \frac{1}{N_k T^2} \sum_{i \in G_k^0} \sum_{|s-t| \leq M_T} \left\| x_{is} \tilde{x}_{it}' \right\| \\
&= N_k^{1/2} T^{1/2} O_P\left((N_k T)^{-1/2} + T^{-1}\right) O_P(M_T/T) \\
&= O_P\left(1 + N_k^{1/2} T^{-1/2}\right) O_P(M_T/T) = o_P(1)
\end{aligned}$$

where we use the fact that $\frac{1}{N_k T^2} \sum_{i \in G_k^0} \sum_{|s-t| \leq M_T} \|x_{is} \tilde{x}_{it}'\| = O_P(M_T/T)$ by moment calculation and Markov inequality. Let $\bar{B}_{kNT,1}(2) \equiv \frac{1}{N_k^{1/2} T^{3/2}} \sum_{i \in G_k^0} \sum_{s=1}^{T} \sum_{t=1}^{T} k_{M_T}(t,s)\, \omega' x_{is} \bar{u}_i$ where ω is any $p \times 1$ nonrandom vector such that $\|\omega\| = 1$. Then by Assumptions D1(i), (iii) and (iv),

$$\begin{aligned}
\left|\mathbb{E}\left[\bar{B}_{kNT,1}(2)\right]\right| &\leq \frac{1}{N_k^{1/2} T^{5/2}} \sum_{i \in G_k^0} \sum_{s=1}^{T} \sum_{t=1}^{T} \sum_{r=1}^{T} k_{M_T}(t,s) \left|\mathbb{E}\left(\omega' x_{is} u_{ir}\right)\right| \\
&\leq \frac{8}{N_k^{1/2} T^{5/2}} \sum_{i \in G_k^0} \sum_{s=1}^{T} \sum_{t=1}^{T} \sum_{r=1}^{T} k_{M_T}(t,s) \left\|\omega' x_{is}\right\|_{4q} \|u_{ir}\|_{4q}\, \alpha_i(|r-s|)^{(2q-1)/(2q)} \\
&\leq \frac{C N_k^{1/2}}{T^{3/2}} \left\{ \frac{1}{N_k} \sum_{i \in G_k^0} c_{\alpha,i}^{(2q-1)/(2q)} \right\} \left\{ \frac{1}{T} \sum_{t,s,r:\ |s-t| \leq M_T} \rho^{|r-s|(2q-1)/(2q)} \right\} \\
&= N_k^{1/2} T^{-3/2} O(1)\, O(M_T) = O\left(M_T N_k^{1/2} T^{-3/2}\right) = o(1).
\end{aligned}$$

Similarly, by Assumptions D1(i)-(iv),

$$\begin{aligned}
\mathrm{Var}\left(\bar{B}_{kNT,1}(2)\right) &= \frac{1}{N_k T^5} \sum_{i \in G_k^0} \mathrm{Var}\left(\sum_{s=1}^{T} \sum_{t=1}^{T} \sum_{r=1}^{T} k_{M_T}(t,s)\, \omega' x_{is} u_{ir} \right) \\
&\leq \frac{1}{N_k T^5} \sum_{i \in G_k^0} \mathbb{E}\left[\left(\sum_{s=1}^{T} \sum_{t=1}^{T} \sum_{r=1}^{T} k_{M_T}(t,s)\, \omega' x_{is} u_{ir} \right)^2 \right] \\
&= \frac{1}{N_k T^5} \sum_{i \in G_k^0} \sum_{1 \leq t_1, t_2, \ldots, t_6 \leq T} k_{M_T}(t_1,t_2)\, k_{M_T}(t_4,t_5)\, \mathbb{E}\left(\omega' x_{it_2} u_{it_3}\ \omega' x_{it_5} u_{it_6}\right) \\
&\leq \frac{1}{N_k T^5} \sum_{i \in G_k^0} \sum_{\substack{1 \leq t_1, t_2, \ldots, t_6 \leq T \\ |t_1 - t_2| \leq M_T, |t_4 - t_5| \leq M_T}} \left|\mathbb{E}\left(\omega' x_{it_2} u_{it_3}\ \omega' x_{it_5} u_{it_6}\right)\right| \\
&= O\left(M_T^2 / T\right) = o(1).
\end{aligned}$$

Consequently, $\bar{B}_{kNT,1}(2) = o_P(1)$. This, in conjunction with Corollary 2.3, implies that $B_{kNT,1}(2) = o_P(1)$ as ω is arbitrary. Thus we have shown that $\hat{B}_{kNT,1} = o_P(1)$.

For $\hat{B}_{kNT,2}$, note that $\hat{B}_{kNT,2} = \bar{B}_{kNT,2} N_k^{1/2} / \tilde{N}_k^{1/2} = \bar{B}_{kNT,2} \left(1 + o_P(1)\right)$, where $\bar{B}_{kNT,2} = \frac{1}{N_k^{1/2} T^{3/2}} \sum_{i \in G_k^0} \sum_{s=1}^{T} \sum_{t=1}^{T} k_{M_T}(t,s) \left[x_{is} u_{it} - \mathbb{E}\left(x_{is} u_{it}\right)\right]$. By construction $\mathbb{E}(\bar{B}_{kNT,2}) = 0$. By Assumptions D1(ii)-(iii) and Jensen inequality,

$$\begin{aligned}
\mathrm{Var}\left(\omega' \bar{B}_{kNT,2}\right) &= \frac{1}{N_k T^3} \sum_{i \in G_k^0} \mathrm{Var}\left[\sum_{s=1}^{T} \sum_{t=1}^{T} k_{M_T}(t,s)\, \omega' \left[x_{is} u_{it} - \mathbb{E}\left(x_{is} \Delta u_{it}\right)\right]\right] \\
&\leq \frac{1}{N_k T^3} \sum_{i \in G_k^0} \sum_{s=1}^{T} \sum_{t=1}^{T} \sum_{r=1}^{T} \sum_{l=1}^{T} k_{M_T}(t,s)\, k_{M_T}(r,l)\, \mathbb{E}\left(\omega' x_{is} u_{it} u_{ir} x_{il} \omega\right) \\
&\leq \frac{1}{N_k T^3} \sum_{i \in G_k^0} \sum_{|s-t| \leq M_T} \sum_{|r-l| \leq M_T} \left|\mathbb{E}\left(\omega' x_{is} u_{it} u_{ir} x_{il} \omega\right)\right| = O\left(M_T^2 / T\right) = o(1),
\end{aligned}$$

where the last equality follows from the fact that $\left\|\mathbb{E}\left(\omega' x_{is} u_{it} u_{ir} x_{il} \omega\right)\right\| \leq \max_{i,s,t} \left\|x_{is} u_{it}\right\|_2^2 \leq \max_{i,t} \left\|x_{it}\right\|_4^2 \times \max_{i,t} \left\|u_{it}\right\|_4^2 < C < \infty$ by Assumption D1(iii). Then $\bar{B}_{kNT,2} = o_P(1)$ by Chebyshev inequality and thus $\hat{B}_{kNT,2} = o_P(1)$.

By Corollary 2.3 and Davydov inequality,

$$\begin{aligned}
\left\|\hat{B}_{kNT,3}\right\| &= \frac{\left|N_k^{-1} - \hat{N}_k^{-1}\right|}{T^{3/2}(N_k^{-1/2} + \hat{N}_k^{-1/2})} \left\|\sum_{i \in G_k^0} \sum_{s=1}^{T} \sum_{t=1}^{T} k_{M_T}(t,s)\, \mathbb{E}\left(x_{is} u_{it}\right)\right\| \\
&\leq \frac{\left|\hat{N}_k - N_k\right|}{T^{1/2} \hat{N}_k (N_k^{-1/2} + \hat{N}_k^{-1/2})} \left\{\frac{1}{N_k T} \sum_{i \in G_k^0} \sum_{|s-t| \leq M_T} \left\|\mathbb{E}\left(x_{is} u_{it}\right)\right\|\right\} \\
&= o_P(N_k^{-1/2} T^{-1/2}) O(1) = o_P(1).
\end{aligned}$$

By Assumptions D1(i)-(iv) and the Davydov inequality,

$$\begin{aligned}
\left\|\hat{B}_{kNT,4}\right\| &= \frac{1}{N_k^{1/2} T^{3/2}} \sum_{i \in G_k^0} \sum_{s=1}^{T} \sum_{t=1}^{T} \left[1 - k_{M_T}(t,s)\right] \mathbb{E}\left(x_{is} u_{it}\right) \\
&= \left\|\frac{1}{N_k^{1/2} T^{3/2}} \sum_{i \in G_k^0} \sum_{s=1}^{T} \sum_{t=1}^{T} \left[1 - k_{M_T}(t,s)\right] \mathbb{E}\left(x_{is} u_{it}\right)\right\| \\
&\leq \frac{8}{N_k^{1/2} T^{3/2}} \sum_{i \in G_k^0} \sum_{|s-t| > M_T} \alpha_i \left(|s-t|\right)^{(2q-1)/(2q)} \left\|x_{is}\right\|_{4q} \left\|u_{it}\right\|_{4q} \\
&\leq C N_k^{-1/2} T^{1/2} \sum_{i \in G_k^0} \alpha_i \left(M_T\right)^{(2q-1)/(2q)} = o(1).
\end{aligned}$$

This completes the proof of the proposition. ■

With the above result in hand, we can readily show that

$$\begin{aligned}\sqrt{N_kT}\left(\hat{\alpha}_k^{(c)}-\alpha_k^0\right) &= \left[\sqrt{N_kT}\left(\hat{\alpha}_k-\alpha_k^0\right)-\bar{\Phi}_k^{-1}\mathbb{B}_{kNT}\right]+\left(N_k/\hat{N}_k\right)^{1/2}\left[\bar{\Phi}_k^{-1}\mathbb{B}_{kNT}-\hat{\Phi}_k^{-1}\hat{\mathbb{B}}_{kNT}\right]\\ &\quad+\left[1-\left(N_k/\hat{N}_k\right)^{1/2}\right]\bar{\Phi}_k^{-1}\mathbb{B}_{kNT}\\ &= \left[\sqrt{N_kT}\left(\hat{\alpha}_k-\alpha_k^0\right)-\bar{\Phi}_k^{-1}\mathbb{B}_{kNT}\right]+o_P(1)+o_P\left(N_k^{-1}\right)O\left(\left(N_K/T\right)^{1/2}\right)\\ &= \left[\sqrt{N_kT}\left(\hat{\alpha}_k-\alpha_k^0\right)-\bar{\Phi}_k^{-1}\mathbb{B}_{kNT}\right]+o_P(1).\end{aligned}$$

That is, $\sqrt{N_kT}(\hat{\alpha}_k^{(c)}-\alpha_k^0)$ has the desired limiting distribution centered on the origin.

D.2 Bias Correction for the PGMM C-Lasso Estimator

Bias correction for the PGMM C-Lasso estimator in dynamic panel data models can be done analogously. For simplicity we focus on the case where $W_{iNT}=I_d$ for all i. Recall from Theorem 3.4 and the remark regarding Assumption B3(iii) (see (3.6) in particular) that

$$\sqrt{N_kT}\left(\tilde{\alpha}_k-\alpha_k^0\right)-\bar{A}_k^{-1}B_{kNT}\xrightarrow{D}N(0,A_k^{-1}C_kA_k^{-1}) \text{ for } k=1,...,K_0$$

where $\bar{A}_k\equiv\frac{1}{N_k}\sum_{i\in G_k^0}\bar{Q}_{i,z\Delta x}'\bar{Q}_{i,z\Delta x}$ and $B_{kNT}=\frac{1}{N_k^{1/2}T^{3/2}}\sum_{i\in G_k^0}\sum_{s=1}^T\sum_{t=1}^T\mathbb{E}\left(\Delta x_{is}z_{is}'z_{it}\Delta u_{it}\right)$. Based on (3.6), in order to verify Assumption B3(iii) we also need to show

$$V_{kNT} = \frac{1}{N_k^{1/2}T^{1/2}}\sum_{i\in G_k^0}\sum_{t=1}^T\bar{Q}_{i,z\Delta x}'z_{it}\Delta u_{it}\xrightarrow{D}N\left(0,C_k\right), \text{ and} \tag{D.1}$$

$$\begin{aligned}R_{kNT} &= \frac{1}{N_k^{1/2}T^{3/2}}\sum_{i\in G_k^0}\sum_{s=1}^T\sum_{t=1}^T\left\{\left[\Delta x_{is}z_{is}'-\mathbb{E}\left(\Delta x_{is}z_{is}'\right)\right]z_{it}\Delta u_{it}-\mathbb{E}\left(\Delta x_{is}z_{is}'z_{it}\Delta u_{it}\right)\right\}\\ &= o_P(1).\end{aligned} \tag{D.2}$$

The first part is assured by a version of the CLT. Below we first propose an estimate of the bias $\bar{A}_k^{-1}B_{kNT}$ and then demonstrate (D.2).

To correct the bias, we propose to obtain consistent estimates of $\bar{A}_k$ and B_{kNT} respectively by

$$\tilde{A}_k=\frac{1}{\tilde{N}_k}\sum_{i\in\tilde{G}_k}\tilde{Q}_{i,z\Delta x}'\tilde{Q}_{i,z\Delta x} \text{ and } \tilde{B}_{kNT}=\frac{1}{\tilde{N}_k^{1/2}T^{3/2}}\sum_{i\in\tilde{G}_k}\sum_{s=1}^T\sum_{t=1}^T k_{M_T}(t,s)\,\Delta x_{is}z_{is}'z_{it}\Delta\tilde{u}_{it},$$

where $\Delta\tilde{u}_{it}=\Delta y_{it}-\tilde{\alpha}_{\tilde{G}_k}'\Delta x_{it}$ for all $i\in\tilde{G}_k$,[13] $k_{M_T}(t,s)$ is as defined above: $k_{M_T}(t,s)=k_{M_T}^0(|t-s|)$ and $k_{M_T}^0(u)$ denotes the Bartlett kernel: $k_{M_T}^0(u)=(1-|u|/M_T)\,\mathbf{1}\{|u|\le M_T\}$. Note that we also allow dynamic misspecification here. If one is sure that the model is dynamically correctly specified in the sense that $\mathbb{E}(\Delta u_{it}|\mathcal{F}_{i,t-1})=0$ where $\mathcal{F}_{i,t-1}=\sigma(\Delta u_{i,t-1},$

[13] Observe that $\tilde{\alpha}_k-\alpha_k^0=O_P\left(\left(N_kT\right)^{-1/2}+T^{-1}\right)$ and $\tilde{\alpha}_{\tilde{G}_k}-\alpha_k^0=O_P\left(\left(N_kT\right)^{-1/2}\right)$. We recommend using the post-Lasso estimator $\tilde{\alpha}_{\tilde{G}_k}$.

$\Delta x_{it-1}, z_{it}; \Delta u_{i,t-2}, \Delta x_{it-2}, z_{i,t-1}; ...)$, one can use the one-sided kernel: $k_{M_T}(t,s) = k^1_{M_T}(s-t)$, where $k^1_{M_T}(u) = (1-u/M_T)\mathbf{1}\{0 \le u \le M_T\}$. The bias-corrected C-Lasso estimator of α^0_k would be

$$\tilde{\alpha}^{(c)}_k = \tilde{\alpha}_k - \frac{1}{\sqrt{\tilde{N}_k T}} \tilde{A}^{-1}_k \tilde{B}_{kNT}.$$

Note that Theorem 3.4 indicates that there is no need to consider bias correction for the post Lasso estimator $\tilde{\alpha}_{\tilde{G}_k}$.

We add the following assumption.

ASSUMPTION D2. *(i) For each $i = 1, ..., N$, $\{(\Delta x_{it}, z_{it}, \Delta u_{it}) : t = 1, 2, ...\}$ is strong mixing with mixing coefficients $\{\alpha_i(\cdot)\}$. In addition, $\alpha_i(\tau) \le c_{\alpha,i}\rho^\tau$ for some $c_{\alpha,i} < \infty$ and $\rho \in (0,1)$ where $\frac{1}{N_k}\sum_{i\in G^0_k} c^{(2q-1)/(2q)}_{\alpha,i} = O(1)$ and $\frac{1}{N_k}\sum_{i\in G^0_k} c^{(q-1)/q}_{\alpha,i} = O(1)$.*

(ii) Let $x_i \equiv (x_{i1}, ..., x_{iT})'$ and $u_i \equiv (u_{i1}, ..., u_{iT})'$. (x_i, u_i) are independent across $i \in G^0_k$ where $k = 1, ..., K_0$.

(iii) $\max_{i,t} \mathbb{E}\|\Delta x_{it} z'_{it}\|^{4q} < C < \infty$ and $\max_{i,t} \mathbb{E}\|z_{it}\Delta u_{it}\|^{4q} < C < \infty$ for some $q > 1$.

(iv) As $(N,T) \to \infty$, $M_T \to \infty$, $M^2_T/T \to 0$ and $N^{-1/2}_k T^{1/2} \sum_{i\in G^0_k} \alpha_i (M_T)^{(2q-1)/(2q)} \to 0$ for each $k = 1, ..., K_0$.

Assumptions D2(i)-(iv) parallel D1(i)-(iv). The major difference is that we do not need $M^2_T N_k/T^3 \to 0$ in D2(iv) but require $q > 1$ in D2(iii).

Proposition D.2 *Suppose that the conditions of Theorem 3.4 hold. Suppose Assumption D2 holds. Then $\tilde{A}^{-1}_k \tilde{B}_{kNT} - \bar{A}^{-1}_k B_{kNT} = o_P(1)$.*

Proof. Noting that $\tilde{A}^{-1}_k \tilde{B}_{kNT} - \bar{A}^{-1}_k B_{kNT} = (\tilde{A}^{-1}_k - \bar{A}^{-1}_k)B_{kNT} + (\tilde{A}^{-1}_k - \bar{A}^{-1}_k)(\tilde{B}_{kNT} - B_{kNT}) + \bar{A}^{-1}_k(\tilde{B}_{kNT} - B_{kNT})$, $\bar{A}^{-1}_k = O(1)$, and $B_{kNT} = O(\sqrt{N_k/T})$, it suffices to show that (i) $\tilde{A}_k - \bar{A}_k = o_P(\nu_{NT})$ and (ii) $\tilde{B}_{kNT} - B_{kNT} = o_P(1)$, where $\nu_{NT} = \min(1, \sqrt{T/N_k})$.

We first prove (i). Note that

$$\begin{aligned}
\tilde{A}_k - \bar{A}_k &= \frac{1}{\tilde{N}_k}\sum_{i\in\tilde{G}_k} \tilde{Q}'_{i,z\Delta x}\tilde{Q}_{i,z\Delta x} - \frac{1}{N_k}\sum_{i\in G^0_k} \tilde{Q}'_{i,z\Delta x}\tilde{Q}_{i,z\Delta x} \\
&= \frac{1}{\tilde{N}_k}\left(\sum_{i\in\tilde{G}_k} - \sum_{i\in G^0_k}\right)\tilde{Q}'_{i,z\Delta x}\tilde{Q}_{i,z\Delta x} + \frac{N_k - \tilde{N}_k}{\tilde{N}_k N_k}\sum_{i\in G^0_k}\tilde{Q}'_{i,z\Delta x}\tilde{Q}_{i,z\Delta x} \\
&\equiv A_{k,1} + A_{k,2}, \text{ say.}
\end{aligned}$$

By Corollary 3.3, $A_{k,2} = O_P(N^{-1}_k) = o_P(\nu_{NT}))$. For any $\epsilon > 0$, we have by the proof of Theorem 3.2, $P(\|A_{k,1}\| \ge \nu_{NT}\epsilon) \le P(\tilde{F}_{kNT}) + P(\tilde{E}_{kNT}) = o(1)$. It follows that $\tilde{A}_k - \bar{A}_k = o_P(\nu_{NT})$.

Now we prove (ii). We make the following decomposition:

$$
\begin{aligned}
& \tilde{B}_{kNT} - B_{kNT} \\
= \; & \frac{1}{\tilde{N}_k^{1/2} T^{3/2}} \sum_{i \in \tilde{G}_k} \sum_{s=1}^{T} \sum_{t=1}^{T} k_{M_T}(t,s)\, \Delta x_{is} z_{is}' z_{it} \Delta \tilde{u}_{it} - \frac{1}{N_k^{1/2} T^{3/2}} \sum_{i \in G_k^0} \sum_{s=1}^{T} \sum_{t=1}^{T} \mathbb{E}\left(\Delta x_{is} z_{is}' z_{it} \Delta u_{it}\right) \\
= \; & \frac{1}{\tilde{N}_k^{1/2} T^{3/2}} \sum_{i \in G_k^0} \sum_{s=1}^{T} \sum_{t=1}^{T} k_{M_T}(t,s)\, \Delta x_{is} z_{is}' z_{it} \Delta \tilde{u}_{it} - \frac{1}{N_k^{1/2} T^{3/2}} \sum_{i \in G_k^0} \sum_{s=1}^{T} \sum_{t=1}^{T} \mathbb{E}\left(\Delta x_{is} z_{is}' z_{it} \Delta u_{it}\right) \\
& + o_P(1) \\
= \; & \frac{1}{\tilde{N}_k^{1/2} T^{3/2}} \sum_{i \in G_k^0} \sum_{s=1}^{T} \sum_{t=1}^{T} k_{M_T}(t,s)\, \Delta x_{is} z_{is}' z_{it} \left(\Delta \tilde{u}_{it} - \Delta u_{it}\right) \\
& + \frac{1}{\tilde{N}_k^{1/2} T^{3/2}} \sum_{i \in G_k^0} \sum_{s=1}^{T} \sum_{t=1}^{T} k_{M_T}(t,s) \left[\Delta x_{is} z_{is}' z_{it} \Delta u_{it} - \mathbb{E}\left(\Delta x_{is} z_{is}' z_{it} \Delta u_{it}\right)\right] \\
& + \frac{N_k^{-1/2} - \tilde{N}_k^{-1/2}}{T^{3/2}} \sum_{i \in G_k^0} \sum_{s=1}^{T} \sum_{t=1}^{T} k_{M_T}(t,s)\, \mathbb{E}\left(\Delta x_{is} z_{is}' z_{it} \Delta u_{it}\right) \\
& + \frac{1}{N_k^{1/2} T^{3/2}} \sum_{i \in G_k^0} \sum_{s=1}^{T} \sum_{t=1}^{T} \left[1 - k_{M_T}(t,s)\right] \mathbb{E}\left(\Delta x_{is} z_{is}' z_{it} \Delta u_{it}\right) + o_P(1) \\
\equiv \; & B_{kNT,1} + B_{kNT,2} + B_{kNT,3} + B_{kNT,4} + o_P(1), \text{ say,}
\end{aligned}
$$

where the $o_P(1)$ term arises due to the replacement of $\tilde{G}_k$ by G_k^0 and this can be easily justified by using the uniform classification consistency result and arguments as used in the proof of Theorems 2.5. We prove (ii) by demonstrating that $B_{kNT,s} = o_P(1)$ for $s = 1, 2, 3, 4$.

First, noting that $\Delta \tilde{u}_{it} - \Delta u_{it} = (\alpha_k^0 - \tilde{\alpha}_{\tilde{G}_k})' \Delta x_{it}$, $\tilde{\alpha}_{\tilde{G}_k} - \alpha_k^0 = O_P\left((N_k T)^{-1/2}\right)$, and that $N_k / \tilde{N}_k = 1 + o_P(1)$ by Corollary 3.3, we have

$$
\begin{aligned}
\|B_{kNT,1}\| &= \frac{1}{\tilde{N}_k^{1/2} T^{3/2}} \left\| \sum_{i \in G_k^0} \sum_{s=1}^{T} \sum_{t=1}^{T} k_{M_T}(t,s)\, \Delta x_{is} z_{is}' z_{it} (\Delta x_{it})' (\alpha_k^0 - \tilde{\alpha}_{\tilde{G}_k}) \right\| \\
&\leq (\tilde{N}_k T)^{1/2} \left\| \alpha_k^0 - \tilde{\alpha}_{\tilde{G}_k} \right\| \frac{N_k}{\tilde{N}_k} \frac{1}{N_k T^2} \sum_{i \in G_k^0} \sum_{|s-t| \leq M_T} \left\| \Delta x_{is} z_{is}' z_{it} (\Delta x_{it})' \right\| \\
&= O_P(1)\, b_{kNT,1}
\end{aligned}
$$

where $b_{kNT,1} = \frac{1}{N_k T^2} \sum_{i \in G_k^0} \sum_{|s-t| \leq M_T}^{T} \left\| \Delta x_{is} z_{is}' z_{it} (\Delta x_{it})' \right\|$. By Markov inequality, $b_{kNT,1} = O_P(M_T / T)$. It follows that $\|B_{kNT,1}\| = O_P(M_T/T) = o_P(1)$ under Assumption D2(iv).

For $B_{kNT,2}$, note that $B_{kNT,2} = b_{kNT,2} N_k^{1/2} / \tilde{N}_k^{1/2} = b_{kNT,2}\left(1 + o_P(1)\right)$, where

$$
b_{kNT,2} = \frac{1}{N_k^{1/2} T^{3/2}} \sum_{i \in G_k^0} \sum_{s=1}^{T} \sum_{t=1}^{T} k_{M_T}(t,s) \left[\Delta x_{is} z_{is}' z_{it} \Delta u_{it} - \mathbb{E}\left(\Delta x_{is} z_{is}' z_{it} \Delta u_{it}\right)\right]
$$

Let ω be any $p\times 1$ nonrandom vector such that $\|\omega\|=1$. Then $\mathbb{E}\left(\omega' b_{kNT,2}\right)=0$. By Assumptions D2(ii)-(iv) and Jensen inequality,

$$\begin{aligned}
&\operatorname{Var}\left(\omega' b_{kNT,2}\right)\\
=\;& \frac{1}{N_k T^3}\sum_{i\in G_k^0}\operatorname{Var}\left[\sum_{s=1}^{T}\sum_{t=1}^{T} k_{M_T}(t,s)\,\omega'\left\{\Delta x_{is} z_{is}' z_{it}\Delta u_{it}-\mathbb{E}\left(\Delta x_{is} z_{is}' z_{it}\Delta u_{it}\right)\right\}\right]\\
\leq\;& \frac{1}{N_k T^3}\sum_{i\in G_k^0}\sum_{s=1}^{T}\sum_{t=1}^{T}\sum_{r=1}^{T}\sum_{l=1}^{T} k_{M_T}(t,s)\,k_{M_T}(r,l)\,\omega'\mathbb{E}\left[\Delta x_{is} z_{is}' z_{it}\Delta u_{it}\Delta x_{il} z_{il}' z_{ir}\Delta u_{ir}\right]\omega\\
\leq\;& \frac{1}{N_k T^3}\sum_{i\in G_k^0}\sum_{|s-t|\leq M_T}\sum_{|r-l|\leq M_T}\left\|\mathbb{E}\left[\omega'\Delta x_{is} z_{is}' z_{it}\Delta u_{it}\Delta x_{il} z_{il}' z_{ir}\Delta u_{ir}\omega\right]\right\|\\
=\;& O\left(M_T^2/T\right)=o(1),
\end{aligned}$$

where the last equality follows from the fact that $\|\mathbb{E}\left[\omega'\Delta x_{is} z_{is}' z_{it}\Delta u_{it}\Delta x_{il} z_{il}' z_{ir}\Delta u_{ir}\omega\right]\|\leq \max_{i,s}\left\{\mathbb{E}\left\|\Delta x_{is} z_{is}'\right\|^4\right\}^{1/2}\times\max_{i,t}\left\{\mathbb{E}\left\|z_{it}\Delta u_{it}\right\|^4\right\}^{1/2}<C<\infty$ by Assumption D2(iii). It follows that $B_{kNT,2}=o_P(1)$.

By Corollary 3.3 and Davydov inequality,

$$\begin{aligned}
\|B_{kNT,3}\| &= \frac{\left|N_k^{-1}-\tilde{N}_k^{-1}\right|}{T^{3/2}(N_k^{-1/2}+\tilde{N}_k^{-1/2})}\left\|\sum_{i\in G_k^0}\sum_{s=1}^{T}\sum_{t=1}^{T} k_{M_T}(t,s)\,\mathbb{E}\left(\Delta x_{is} z_{is}' z_{it}\Delta u_{it}\right)\right\|\\
&\leq \frac{\left|\tilde{N}_k-N_k\right|}{T^{1/2}\tilde{N}_k(N_k^{-1/2}+\tilde{N}_k^{-1/2})}\left\{\frac{1}{N_k T}\sum_{i\in G_k^0}\sum_{|s-t|\leq M_T}\left\|\mathbb{E}\left(\Delta x_{is} z_{is}' z_{it}\Delta u_{it}\right)\right\|\right\}\\
&= o_P(N_k^{-1/2}T^{-1/2})O(1)=o_P(1).
\end{aligned}$$

By Assumptions D2(i)-(iii) and Davydov inequality,

$$\begin{aligned}
\|B_{kNT,4}\| &= \left\|\frac{1}{N_k^{1/2}T^{3/2}}\sum_{i\in G_k^0}\sum_{s=1}^{T}\sum_{t=1}^{T}\left[1-k_{M_T}(t,s)\right]\mathbb{E}\left(\Delta x_{is} z_{is}' z_{it}\Delta u_{it}\right)\right\|\\
&\leq \frac{8}{N_k^{1/2}T^{3/2}}\sum_{i\in G_k^0}\sum_{|s-t|>M_T}\alpha_i(|s-t|)^{(2q-1)/(2q)}\left\|\Delta x_{is} z_{is}'\right\|_{4q}\left\|z_{it}\Delta u_{it}\right\|_{4q}\\
&\leq CN_k^{-1/2}T^{1/2}\sum_{i\in G_k^0}\alpha_i(M_T)^{(2q-1)/(2q)}=o(1).
\end{aligned}$$

This completes the proof of the proposition. ■

With the above result in hand, we can readily show that

$$\begin{aligned}\sqrt{N_kT}\left(\tilde{\alpha}_k^{(c)}-\alpha_k^0\right) &= \left[\sqrt{N_kT}\left(\tilde{\alpha}_k-\alpha_k^0\right)-\bar{A}_k^{-1}B_{kNT}\right]+\left(N_k/\tilde{N}_k\right)^{1/2}\left[\bar{A}_k^{-1}B_{kNT}-\tilde{A}_k^{-1}\tilde{B}_{kNT}\right]\\ &\quad+\left[1-\left(N_k/\tilde{N}_k\right)^{1/2}\right]\bar{A}_k^{-1}B_{kNT}\\ &= \left[\sqrt{N_kT}\left(\tilde{\alpha}_k-\alpha_k^0\right)-\bar{A}_k^{-1}B_{kNT}\right]+o_P(1)+o_P\left(N_k^{-1}\right)O\left(\left(N_K/T\right)^{1/2}\right)\\ &= \left[\sqrt{N_kT}\left(\tilde{\alpha}_k-\alpha_k^0\right)-\bar{A}_k^{-1}B_{kNT}\right]+o_P(1).\end{aligned}$$

That is, $\sqrt{N_kT}(\tilde{\alpha}_k^{(c)}-\alpha_k^0)$ has the desired limiting distribution centered on the origin.

Now, we demonstrate (D.2). Let $\xi_{is}=\Delta x_{is}z_{is}'-\mathbb{E}\left(\Delta x_{is}z_{is}'\right)$ and $\eta_{it}=z_{it}\Delta u_{it}$. Noting that $E\left(\xi_{is}\right)=0$ and $E\left(\eta_{it}\right)=0$, we have

$$\begin{aligned}R_{kNT} &= \frac{1}{N_k^{1/2}T^{3/2}}\sum_{i\in G_k^0}\sum_{s=1}^{T}\sum_{t=1}^{T}\left[\xi_{is}\eta_{it}-\mathbb{E}\left(\xi_{is}\eta_{it}\right)\right]\\ &= \frac{1}{N_k^{1/2}T^{3/2}}\sum_{i\in G_k^0}\sum_{t=1}^{T}\left[\xi_{it}\eta_{it}-\mathbb{E}\left(\xi_{it}\eta_{it}\right)\right]+\frac{1}{N_k^{1/2}T^{3/2}}\sum_{i\in G_k^0}\sum_{1\le s<t\le T}\left[\xi_{is}\eta_{it}-\mathbb{E}\left(\xi_{is}\eta_{it}\right)\right]\\ &\quad+\frac{1}{N_k^{1/2}T^{3/2}}\sum_{i\in G_k^0}\sum_{1\le t<s\le T}\left[\xi_{is}\eta_{it}-\mathbb{E}\left(\xi_{is}\eta_{it}\right)\right]\\ &\equiv R_{kNT,1}+R_{kNT,2}+R_{kNT,3},\ \text{say.}\end{aligned}$$

It is trivial to show that $R_{kNT,1}=O_P\left(T^{-1}\right)$ by Chebyshev and Davydov inequalities. For $R_{kNT,2}$, we have $\mathbb{E}\left(R_{kNT,2}\right)=0$ by construction, and by Assumption D2(ii) and Jensen inequality

$$\begin{aligned}\mathbb{E}\left(R_{kNT,2}^2\right) &= \frac{1}{N_kT^3}\sum_{i\in G_k^0}\mathrm{Var}\left(\sum_{1\le t_1<t_2\le T}\left[\xi_{it_1}\eta_{it_2}-\mathbb{E}\left(\xi_{it_1}\eta_{it_2}\right)\right]\right)\\ &\le \frac{1}{N_kT^3}\sum_{i\in G_k^0}\sum_{1\le t_1<t_2\le T}\sum_{1\le t_3<t_4\le T}\mathbb{E}\left(\xi_{it_1}\eta_{it_2}\xi_{it_3}\eta_{it_4}\right)\equiv S_{kNT},\ \text{say.}\end{aligned}$$

To bound S_{kNT}, we can consider three subcases: (a) $\#\{t_1,t_2,t_3,t_4\}=4$, (b) $\#\{t_1,t_2,t_3,t_4\}=3$, and (c) $\#\{t_1,t_2,t_3,t_4\}=2$, and use $S_{kNT,a}$, $S_{kNT,b}$, and $S_{kNT,c}$ to denote the last summation when the time indices are restricted to these three cases in order. Apparently, $S_{kNT,c}=O\left(1/T\right)$ under Assumption D2(iii). In case (a), without loss of generality (wlog) assume that $1\le t_1<t_2<t_3<t_4\le T$ and denote $S_{kNT,a}^{(1)}$ as $S_{kNT,a}$ when the time indices are restricted to this subcase. [Note that the other subcases can be analyzed analogously.] Let d_c be the c-th largest difference among $t_{j+1}-t_j$ for $j=1,2,3$. Then

$$\begin{aligned}S_{kNT,a}^{(1)} &= \frac{1}{N_kT^3}\sum_{i\in G_k^0}\left\{\sum_{1\le t_1<t_2<t_3<t_4\le T,t_2-t_1=d_1}+\sum_{1\le t_1<t_2<t_3<t_4\le T,t_3-t_2=d_1}+\sum_{1\le t_1<t_2<t_3<t_4\le T,t_4-t_3=d_1}\right\}\\ &\quad\times\mathbb{E}\left(\xi_{it_1}\eta_{it_2}\xi_{it_3}\eta_{it_4}\right)\\ &\equiv S_{kNT,a1}^{(1)}+S_{kNT,a2}^{(1)}+S_{kNT,a3}^{(1)},\ \text{say.}\end{aligned}$$

By the Davydov inequality and Assumptions D2(i) and (iii),

$$\begin{aligned} S_{NT,a1}^{(1)} &\leq \frac{1}{N_k T^3} \sum_{i \in G_k^0} \sum_{t_1=1}^{T-3} \sum_{t_2=t_1+\max_{j\geq 3}\{t_j - t_{j-1}\}}^{T-2} \sum_{t_3=t_2+1}^{T-1} \sum_{t_4=t_3+1}^{T} \left\|\xi_{it_1}\right\|_{4q} \left\|\eta_{it_2}\xi_{it_3}\eta_{it_4}\right\|_{4q/3} \alpha_i \left(t_2 - t_1\right)^{(q-1)/q} \\ &\leq \frac{C}{N_k T^3} \sum_{i=1}^{N} \sum_{t_1=1}^{T-3} \sum_{t_2=t_1+1}^{T-2} \left(t_2 - t_1\right)^2 \alpha_i \left(t_2 - t_1\right)^{(q-1)/q} \\ &\leq \frac{1}{N_k T} \sum_{i=1}^{N} \sum_{\tau=1}^{\infty} \tau \alpha_i \left(\tau\right)^{(q-1)/q} = O(T^{-1}). \end{aligned}$$

Similarly, we can show that $S_{kNT,as}^{(1)} = O\left(1/T\right)$ for $s = 2, 3$. It follows that $S_{kNT,a}^{(1)} = O\left(1/T\right)$ and $S_{kNT,a}^{(1)} = O\left(1/T\right) = o(1)$. In case (b), wlog assume that $t_4 = t_2$ and $1 \leq t_1 < t_2 < t_3 \leq T$ and we use $S_{kNT,b}^{(1)}$ to $S_{kNT,b}$ when the time indices are restricted to this subcase. Then by the Davydov inequality and Assumptions D2(i) and (iii)

$$\begin{aligned} \left|S_{kNT,b}^{(1)}\right| &= \frac{1}{N_k T^3} \sum_{i \in G_k^0} \sum_{1 \leq t_1 < t_2 < t_3 \leq T} \left|\mathbb{E}\left(\xi_{it_1}\eta_{it_2}^2\xi_{it_3}\right)\right| \\ &\leq \frac{8}{N_k T^3} \sum_{i=1}^{N} \sum_{1 \leq t_1 < t_2 < t_3 \leq T} \left\|\xi_{it_1}\eta_{it_2}^2\right\|_{4q/3} \left\|\xi_{it_3}\right\|_{4q} \alpha_i \left(t_3 - t_2\right)^{(q-1)/q} \\ &\leq \frac{8C}{N_k T} \sum_{i=1}^{N} \sum_{\tau=1}^{\infty} \alpha_i \left(\tau\right)^{(q-1)/q} = O(T^{-1}). \end{aligned}$$

So $S_{kNT,b} = O(T^{-1})$. Consequently, $S_{kNT} = O(T^{-1})$ and $R_{kNT,2} = O_P(T^{-1/2})$ by Chebyshev inequality. By the same token, $R_{kNT,3} = O_P(T^{-1/2})$. Thus we have shown that $R_{kNT} = O_P(T^{-1/2}) = o_P\left(1\right)$.

E Additional Simulation Results

Figures 3- 6 graph the first 50 replications of the information criteria curves, showing how the IC value reacts to changing group number. Each figure provides six panels of (N, T) combinations with the vertical axis giving the IC value and the horizontal axis the trial group number K. As described in Section 4.3, we use $\rho_{jNT} = \frac{2}{3}(NT)^{-1/2}$ and $C_{\lambda_j} = 1$ for $(j = 1, 2)$. The true group number is 3 for each DGP. Examination of the figures shows that in all panel combinations, the IC value falls rapidly as K increases from 1 to 3. When $K > 3$, the IC value typically rises, due to the impact of the penalty, although in many cases the rise in value is slight. Similar phenomena tend to occur in other uses of information criteria, such as lag order determination in time series regression. When $T = 10$, the U-shape in the graphics is clear with the valley lying close to $K = 3$. When $T = 40$, almost all the IC curves have minima at $K = 3$. These outcomes echo the frequency values reported in Table 1.

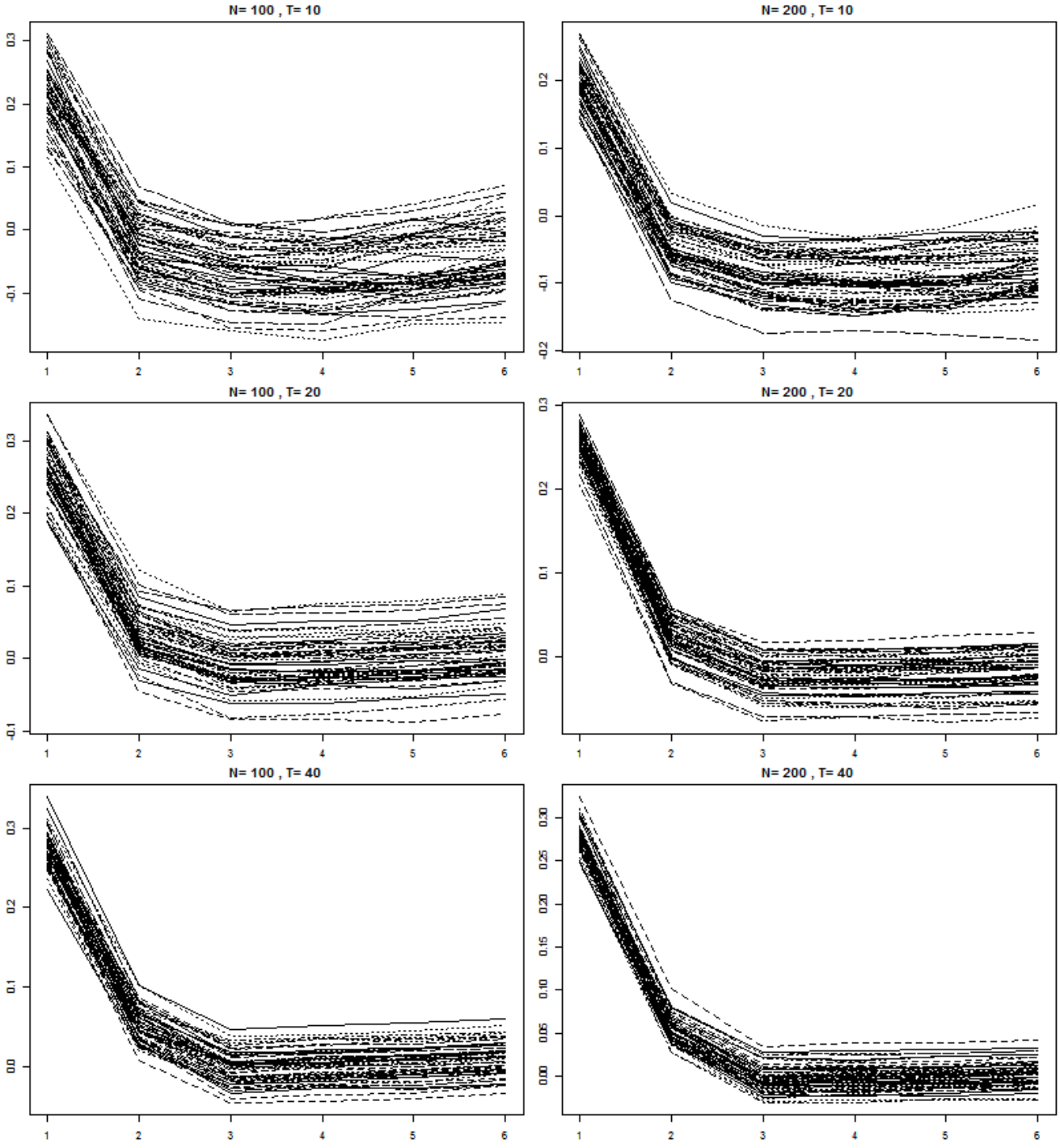

Figure 3: Information criterion of DGP 1 under PLS

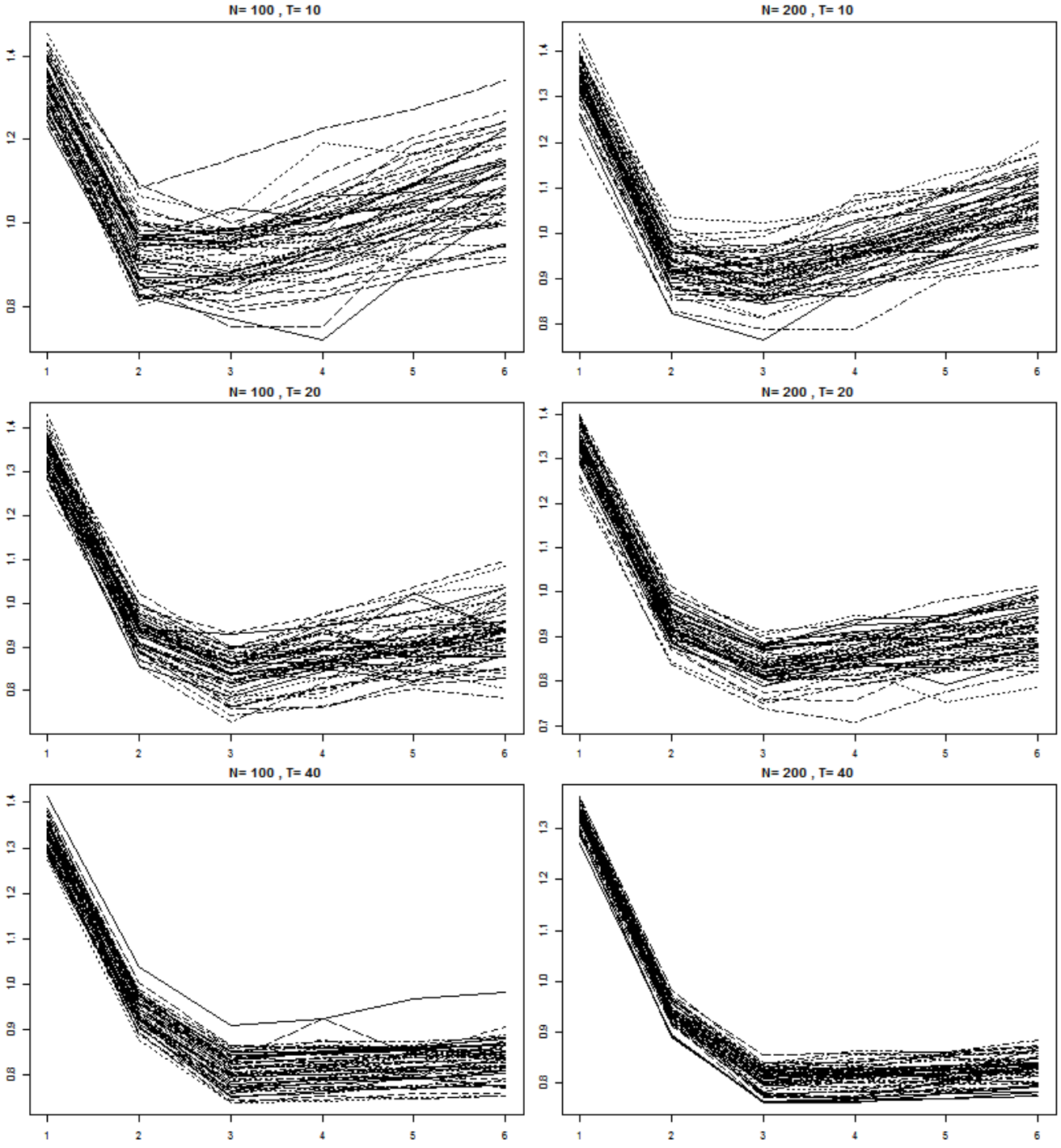

Figure 4: Information criterion of DGP 2 under PGMM

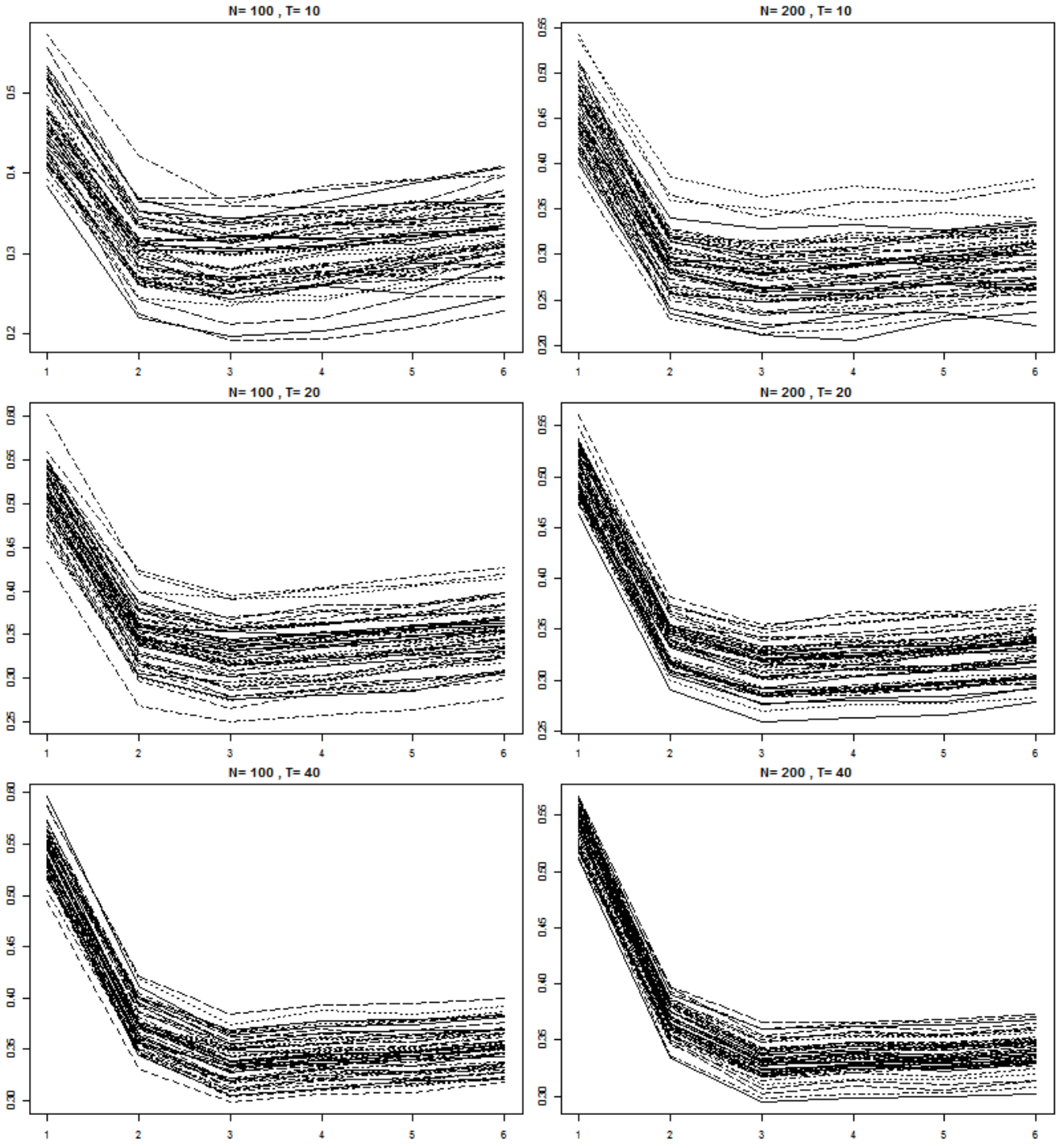

Figure 5: Information criterion of DGP 3 under PLS

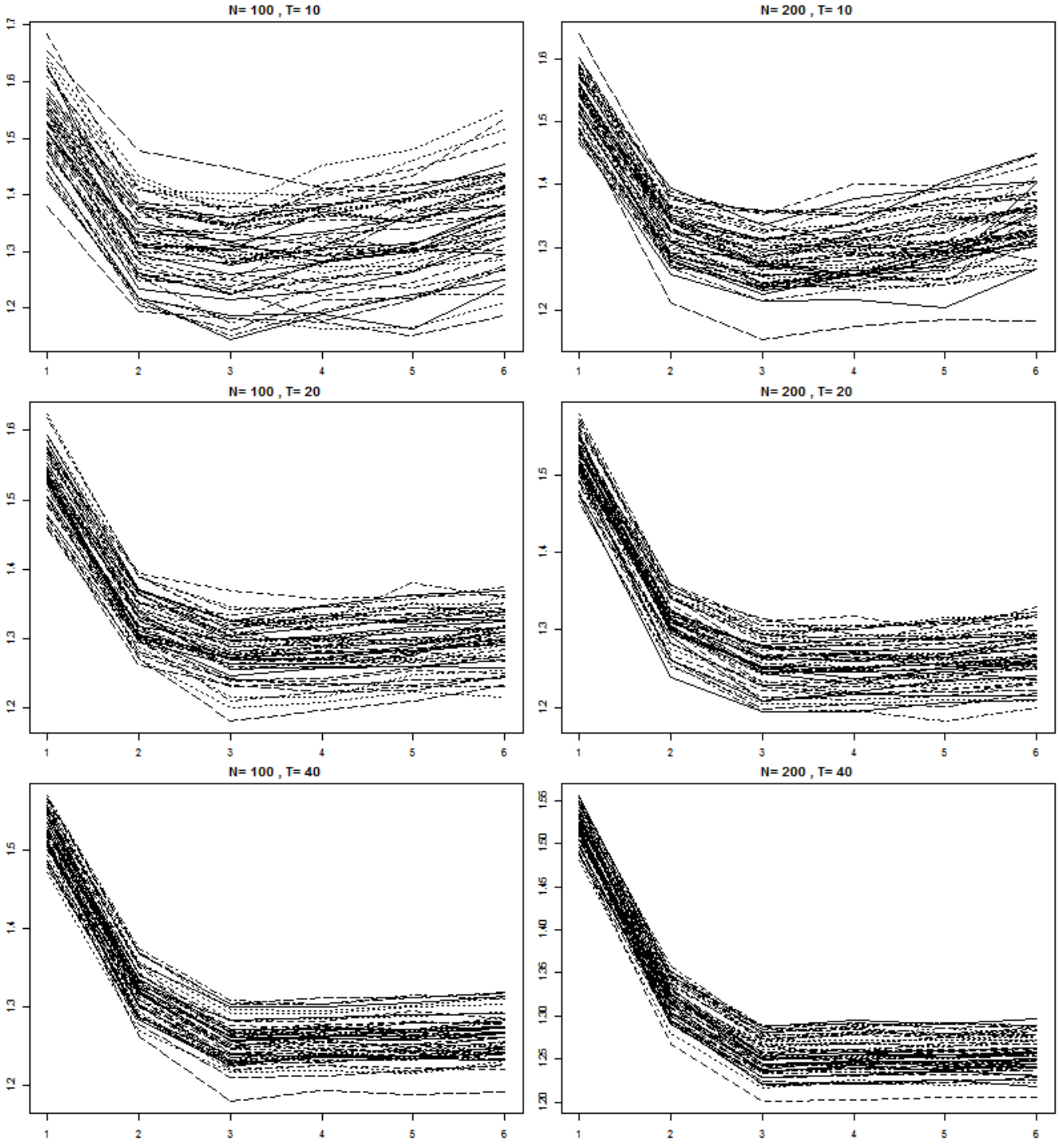

Figure 6: Information criterion of DGP 3 under PGMM

REFERENCE

HAHN, J., AND G. KUERSTEINER (2011): "Bias Reduction for Dynamic Nonlinear Panel Models with Fixed Effects," *Econometric Theory* 27, 1152-1191.

MERLEVÈDE, F., M. PEILGRAD, M., AND E. RIO (2009): "Bernstein Inequality and Moderate Deviations under Strong Mixing Conditions," *IMS collections: High Dimensional Probability V.*, 273-292.

MERLEVÈDE, F., M. PEILGRAD, M., AND E. RIO (2011): "A Bernstein Type Inequality and Moderate Deviations for Weakly Dependent Sequences," *Probability Theory and Related Fields* 151, 435-474.

PRAKASA RAO, B. L. S. (2009): "Conditional Independence, Conditional Mixing and Conditional Association," *Annals of the Institute of Statistical Mathematics* 61, 441-460.

SU, L., AND Q. CHEN (2013): "Testing Homogeneity in Panel Data Models with Interactive Fixed Effects," *Econometric Theory* 29, 1079-1135.