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, Chinese University of Hong Kong*

[c] *Yale University, University of Auckland*

University of Southampton & Singapore Management University

December 29, 2015

Abstract

This paper provides a novel mechanism for identifying and estimating latent group structures in panel data using penalized techniques. We consider both linear and nonlinear models where the regression coefficients are heterogeneous across groups but homogeneous within a group and the group membership is unknown. Two approaches are considered – penalized profile likelihood (PPL) estimation for the general nonlinear models without endogenous regressors, and penalized GMM (PGMM) estimation for linear 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 PPL 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. Empirical applications to both linear and nonlinear models are presented.

JEL Classification: C33, C36, C38, C51

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

*The authors thank the Co-editor Elie Tamer and three anonymous referees for many constructive comments on the previous version of the paper. They also thank Stéphane Bonhomme, Xiaohong Chen, Cheng Hsiao, Joon Park, and Yixiao Sun for discussions on the subject matter and comments on the paper. Su acknowledges support from the Singapore Ministry of Education for Academic Research Fund (AcRF) under the Tier-2 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.

1 Introduction

Panel data are widely used in empirical analysis in many disciplines across the social and medical sciences. 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 2014, ch. 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 typically enter the model additively. This approach is an exemplar of a convenient assumption that facilitates estimation and inference. Nevertheless, this assumption has been frequently questioned and rejected in empirical studies; see Hsiao and Tahmiscioglu (1997), Lee, Pesaran, and Smith (1997), Durlauf, Kourtellos, and Minkin (2001), Phillips and Sul (2007a), Browning and Carro (2007, 2010, 2014), and Su and Chen (2013), among others.

Despite general agreement that slope heterogeneity is endemic in empirical work with panels, few methods are available to allow for heterogeneity in the slopes when the extent of the heterogeneity is unknown. Some researchers assume complete slope heterogeneity where regression coefficients are completely different for different individuals; see the survey by Baltagi, Bresson, and Pirotte (2008) and Hsiao and Pesaran (2008). Others consider panel structure models where individuals belong to a number of homogeneous groups within a broadly heterogeneous population, and 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 membership of each individual. 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 (2016) consider a panel structure model where individuals are grouped according to some external classification, geographic location, or observable explanatory variables; Ando and Bai (2014) consider a multifactor asset-pricing model with group-specific pervasive factors where the group membership is known. Here the group structure is assumed to be *completely known* to the researcher, an approach that is common in practical work because of its convenience. 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 applies finite mixture models. For example, Sun (2005) considers a *parametric* finite mixture linear panel data model, and 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 (2015) consider linear panel data models where the slope coefficients have latent group structure. They modify the K-means algorithm to estimate the models but do not provide any inference theory. Bonhomme and Manresa (2015, BM hereafter) consider a linear panel data model where the additive fixed effects have group structure and apply the K-means algorithm to estimate the model and study its asymptotic properties. Ando and Bai (2015) extend BM's approach to allow for group structure among the interactive fixed effects. In addition, Phillips and Sul (2007a) develop an algorithm for determining group clusters that relies on the estimation of evaporating trend functions to determine convergence clusters. Hahn and Moon (2010) argue that the group structure has sound foundations in game theory or macroeconomic models where multiplicity of Nash equilibria is expected and they consider nonlinear panel data models where the parameter of interest is common to individuals whereas the fixed effects have finite support.

The present paper proposes a new method for econometric estimation and inference in panel models when the regression parameters are heterogenous across groups, individual group membership is unknown, and classification is to be determined empirically. It is an automated data-determined procedure and does not require the specification of any modeling mechanism for the unknown group structure. 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 a key advantage of Lasso technology in coping with parameter sparsity. This advantage is particularly useful when the set of unknown parameters is potentially large but may also embody certain *sparse* features. In a typical panel structure model, the *effective* number of unknown regression 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 parameters are homogeneous. Hence, in many empirical applications the set of unknown 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. It has been used to detect multiple structural changes in the time series setting; see, e.g., Harchaoui and Lévy-Leduc (2010), Chan, Yau, and Zhang (2014), and Qian and Su (2015). 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 used they 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 regression parameter vectors to a particular *unknown* group-level parameter vector, thereby producing a joint shrinkage process to unknown quantities. This process is distinct from the prototypical Lasso method that shrinks an individual parameter to the *known* value zero and the group Lasso method that shrinks a parameter vector to a *known* 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 double asymptotic limit theory for the C-Lasso that demonstrates its capacity to achieve simultaneous classification and estimation in a single step. As mentioned in the Abstract, the paper develops two classes of estimators for panel structure models – penalized profile likelihood (PPL) and penalized GMM (PGMM). The former is applicable to both linear and nonlinear panel models without endogeneity and with or without dynamic structures, while the latter is applicable to linear panel models with endogeneity or dynamic structures. Both broaden the scope of applicability of our method as early literature only considers linear panels without endogeneity. 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. Such a uniform result allows us to establish an *oracle* property of the PPL estimator that, like the BM K-means estimator, is asymptotically equivalent to the corresponding infeasible estimator of the group-specific parameter that is obtained by knowing all individual group identities. Unfortunately, our PGMM estimator generally does not have the oracle property. But the uniform classification consistency property 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 and these estimators are asymptotically as efficient as the oracle ones in both the PPL and PGMM contexts.

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 for both PPL and PGMM estimation and it is shown that this procedure selects the correct number of groups consistently.

The rest of the paper is organized as follows. We study C-Lasso PPL 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. Section 5 contains two empirical applications. Section 6 concludes. Proofs of the main results in the paper are given in Appendices A and B. Additional materials may be found in the Supplemental Material.

For any real matrix A, we write the transpose A', the Frobenius norm $\|A\|$, 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, and $\mathbf{1}\{\cdot\}$ is the indicator function. 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 Profile Likelihood 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

Given a panel data set $\{(y_{it}, x_{it})\}$ for $i = 1, ..., N$ and $t = 1, ..., T$, it is proposed to use fixed effects quasi maximum likelihood to estimate the unknown parameters by solving the minimization problem

$$\min_{\{\beta_i, \mu_i\}} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \psi\left(w_{it}; \beta_i, \mu_i\right). \tag{2.1}$$

Here $-\psi\left(w_{it}; \beta_i, \mu_i\right)$ denotes the logarithm of the pseudo-true conditional density function of y_{it} given x_{it}, the history of (y_{it}, x_{it}), and (μ_i, β_i), where μ_i are scalar individual effects and β_i are $p \times 1$ vectors of parameters of interest. Traditionally, econometric work has assumed that the β_i are common for all cross sectional units, leading to a homogeneous panel with individual heterogeneity modeled through μ_i alone. At the other extreme, the β_i are assumed to differ across individuals and each is estimated at a slow rate without pooling across section. The present paper allows the true values of β_i, denoted β_i^0, to follow a group pattern of the general form

$$\beta_i^0 = \sum_{k=1}^{K_0} \alpha_k^0 \mathbf{1}\left\{i \in G_k^0\right\}. \tag{2.2}$$

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. In the economic growth literature (e.g., Phillips and Sul, 2007a), K_0 corresponds to the number of convergence clubs and countries (indexed by i) within the same k^{th} club share the same (slope) parameter vector α_k^0. In the market entry-exit example (e.g., Hahn and Moon, 2010), K_0 denotes the number of pure Nash equilibria and markets (indexed by i) selecting the same equilibrium over time exhibit the same parameter vector.

For now, we assume that the number of groups, K_0, is known and fixed but that each individual's group membership is unknown. In addition, following Sun (2005), Lin and Ng (2012), and BM, we implicitly assume that individual group membership does not vary over time. Let $\boldsymbol{\alpha} \equiv (\alpha_1, ..., \alpha_{K_0})$ and $\boldsymbol{\beta} \equiv (\beta_1, ..., \beta_N)$. We denote the true values of μ_i, α_k, β_i, $\boldsymbol{\alpha}$, and $\boldsymbol{\beta}$ as μ_i^0, α_k^0, β_i^0, $\boldsymbol{\alpha}^0$, and $\boldsymbol{\beta}^0$, respectively. The econometric task is to infer each individual's group identity and to estimate the group-specific parameters α_k^0. Some examples of models that fall within this framework and the scope of our methodology are as follows.

EXAMPLE 1 (Linear panel) The linear panel structure model is generated according to

$$y_{it} = \beta_i^{0\prime} x_{it} + \mu_i^0 + \varepsilon_{it}, \tag{2.3}$$

where x_{it} is a $p \times 1$ vector of exogenous or predetermined variables, μ_i is an individual fixed effect, β_i is a $p \times 1$ vector of slope parameters, and ε_{it} is the idiosyncratic error term with mean zero. Gaussian quasi-maximum likelihood estimation (QMLE) of β_i and μ_i is achieved by minimizing (2.1) with $\psi\left(w_{it}; \beta_i, \mu_i\right) = \frac{1}{2}\left(y_{it} - \beta_i' x_{it} - \mu_i\right)^2$ and $w_{it} = (y_{it}, x_{it}')'$.

EXAMPLE 2 (Linear panel with quantile restrictions) Consider the model in (2.3) with the quantile restriction: $P\left(\varepsilon_{it} \leq 0|x_{it}, \beta_i^0, \mu_i^0\right) = \tau$; see, e.g., Kato, Galvo, and Montes-Rojas (2012). We can estimate β_i and μ_i by minimizing (2.1) with $\psi\left(w_{it}; \beta_i, \mu_i\right) = \rho_\tau\left(y_{it} - \beta_i' x_{it} - \mu_i\right)$ where $\rho_\tau\left(u\right) = \{\tau - K\left(-u/h\right)\}u$ is a smoothed version of the usual check function with K being a CDF-type kernel function and h a bandwidth parameter.

EXAMPLE 3 (Binary choice panel) The dynamic binary choice panel data model is characterized by $y_{it} = \mathbf{1}\left\{\beta_i^{0\prime} x_{it} + \mu_i^0 - \varepsilon_{it} \geq 0\right\}$, where x_{it}, μ_i, and ε_{it} are as defined in Example 1. In this case, $-\psi\left(w_{it}; \beta_i, \mu_i\right) = y_{it} \ln F\left(y_{it} - \beta_i' x_{it} - \mu_i\right) + \left(1 - y_{it}\right) \ln\left[1 - F\left(y_{it} - \beta_i' x_{it} - \mu_i\right)\right]$, where $w_{it} = \left(y_{it}, x_{it}'\right)'$, and $F\left(\cdot\right)$ denotes the conditional CDF (standard logistic or normal) of ε_{it} given x_{it} and the history of $\left(x_{it}, y_{it}\right)$.

EXAMPLE 4 (Tobit panel) The Tobit panel is characterized by $y_{it} = \max\left(0, \beta_i^{0\prime} x_{it} + \mu_i^0 + \varepsilon_{it}\right)$, where x_{it}, μ_i, and ε_{it} are defined as in the above examples. For clarity, assume that ε_{it}'s are independent and identically distributed (IID) $N\left(0, \sigma_\varepsilon^2\right)$ given x_{it} and the history of $\left(x_{it}, y_{it}\right)$. In this case, $-\psi\left(w_{it}; \beta_i, \mu_i, \sigma_\varepsilon^2\right) = \mathbf{1}\left\{y_{it} = 0\right\} \ln F\left(\left(y_{it} - \beta_i' x_{it} - \mu_i\right)/\sigma_\varepsilon^2\right) + \mathbf{1}\left\{y_{it} > 0\right\} \ln[f\left(\left(y_{it} - \beta_i' x_{it} - \mu_i\right)/\sigma_\varepsilon^2\right)/\sigma_\varepsilon]$, where $w_{it} = \left(y_{it}, x_{it}'\right)'$, f and F denotes the standard normal PDF and CDF, respectively. The presence of the common parameter σ_ε^2 can be addressed by extending the asymptotic analysis below.

2.2 Penalized Profile Likelihood Estimation of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$

Following Hahn and Newey (2004) and Hahn and Kuersteiner (2011), the profile log-likelihood function is

$$Q_{1,NT}\left(\boldsymbol{\beta}\right) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \psi\left(w_{it}; \beta_i, \hat{\mu}_i\left(\beta_i\right)\right), \tag{2.4}$$

where $\hat{\mu}_i\left(\beta_i\right) = \arg\min_{\mu_i} \frac{1}{T} \sum_{t=1}^{T} \psi\left(w_{it}; \beta_i, \mu_i\right)$. Motivated by the literature on group Lasso (e.g., Yuan and Lin 2006), we propose to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ by minimizing the following PPL criterion function

$$Q_{1NT,\lambda_1}^{(K_0)}\left(\boldsymbol{\beta}, \boldsymbol{\alpha}\right) = Q_{1,NT}\left(\boldsymbol{\beta}\right) + \frac{\lambda_1}{N} \sum_{i=1}^{N} \Pi_{k=1}^{K_0} \left\|\beta_i - \alpha_k\right\|, \tag{2.5}$$

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., $\hat{\boldsymbol{\alpha}} \equiv \left(\hat{\alpha}_1, ..., \hat{\alpha}_K\right)$ and $\hat{\boldsymbol{\beta}} \equiv (\hat{\beta}_1, ..., \hat{\beta}_N)$.

The penalty term in (2.5) takes a novel mixed *additive-multiplicative* form that does not appear in the literature. Traditional Lasso includes additive penalty terms to an objective function by differentiating zeros from non-zero-valued parameters to select relevant regressors. In contrast, the C-Lasso has N additive terms, each of which takes a multiplicative form as the product of K_0 separate penalties. The multiplicative component is needed because for each i, β_i^0 can take any one of the K_0 *unknown* values, $\alpha_1^0, ..., \alpha_{K_0}^0$. We do not know *a priori* to which point β_i should shrink, and all K_0 possibilities must be allowed. Each of the K_0 penalty terms in the multiplicative expression permits β_i to shrink to a particular *unknown* group-level parameter vector α_k. The summation component is needed because we need to pull information from all N cross sectional units in order to identify $\left\{\beta_i^0\right\}$ and $\left\{\alpha_k^0\right\}$ jointly. Our 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 zero vector. The main purpose in the latter papers is to select relevant variables while C-Lasso is designed to determine group membership for each individual. As emphasized in the Introduction, both problems enjoy the same motivation of parameter sparsity despite their different nature. C-Lasso has the additional motivation of classification of unknown parameters into *a priori* unknown groups each with their own *unknown* parameters.

Note that the objective function in (2.5) is not convex in $\boldsymbol{\beta}$ even though it is (conditionally) convex in α_k when one fixes α_j for $j \neq k$. The supplement provides an iterative algorithm to obtain the estimates $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$.

2.3 Assumptions

Let $\mu_i(\beta_i) \equiv \arg\min_{\mu_i} \Psi_i(\beta_i, \mu_i)$ where $\Psi_i(\beta_i, \mu_i) \equiv \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\psi(w_{it}; \beta_i, \mu_i)\right]$. Note that $\mu_i^0 = \mu_i\left(\beta_i^0\right)$. Let $U_i(w_{it}; \beta_i, \mu_i) \equiv \partial\psi(w_{it}; \beta_i, \mu_i)/\partial\beta_i$ and $V_i(w_{it}; \beta_i, \mu_i) \equiv \partial\psi(w_{it}; \beta_i, \mu_i)/\partial\mu_i$. Let $U_i^{\mu_i}$ and $U_i^{\mu_i\mu_i}$ denote the first and second derivatives of U_i with respect to μ_i. Define $V_i^{\mu_i}$, $V_i^{\mu_i\mu_i}$, $U_i^{\beta_i}$, and $V_i^{\beta_i}$ similarly. For notational simplicity, denote $U_{it} \equiv U_i\left(w_{it}; \beta_i^0, \mu_i^0\right)$, and similarly for $U_{it}^{\mu_i}$, $U_{it}^{\mu_i\mu_i}$, V_{it}, $V_{it}^{\mu_i}$ and $U_{it}^{\mu_i\mu_i}$. Define

$$
\begin{aligned}
m_{iU} &\equiv \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left(U_{it}^{\mu_i}\right), \; m_{iV} \equiv \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left(V_{it}^{\mu_i}\right), \; m_{iU2} \equiv \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left(U_{it}^{\mu_i\mu_i}\right), \; m_{iV2} \equiv \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left(V_{it}^{\mu_i\mu_i}\right), \\
\mathbb{U}_{it} &\equiv U_{it} - \frac{m_{iU}}{m_{iV}}V_{it}, \;\; \mathbb{U}_{it}^{\beta_i} \equiv U_{it}^{\beta_i} - \frac{m_{iU}}{m_{iV}}V_{it}^{\beta_i\prime}, \text{ and } \mathbb{U}_{it}^{\mu_i} \equiv U_{it}^{\mu_i} - \frac{m_{iU}}{m_{iV}}V_{it}^{\mu_i}.
\end{aligned}
$$

Let $\Omega_{iT} \equiv \frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathbb{E}\left(\mathbb{U}_{it}\mathbb{U}_{is}'\right)$, $\mathbb{H}_{iT} \equiv \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\mathbb{U}_{it}^{\beta_i}]$, and $\mathbb{H}_{kNT} \equiv \frac{1}{N_k}\sum_{i\in G_k^0}\mathbb{H}_{iT}$. Define the two expected Hessian matrices for cross sectional unit i:

$$
H_{i\mu\mu}(\beta_i) \equiv \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[V_i^{\mu_i}(w_{it}; \beta_i, \mu_i(\beta_i))\right] \text{ and } H_{i\beta\beta}(\beta_i) \equiv \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[U_{it}^{\beta_i}(\beta_i) + U_{it}^{\mu_i}(\beta_i)\frac{\partial\mu_i(\beta_i)}{\partial\beta_i'}\right],
$$

where $U_{it}^{\beta_i}(\beta_i) = U_i^{\beta_i}(w_{it}; \beta_i, \mu_i(\beta_i))$, and similarly for $U_{it}^{\mu_i}(\beta_i)$. Let $\min_i$ denote $\min_{1\leq i\leq N}$, and similarly for $\max_i$. We make the following assumptions

ASSUMPTION A1. (i) *For each* i, $\{w_{it} : t = 1, 2, ...\}$ *is stationary strong mixing with mixing coefficients* $\alpha_i(\cdot)$. $\alpha(\cdot) \equiv \max_i \alpha_i(\cdot)$ *satisfies* $\alpha(\tau) \leq c_\alpha\rho^\tau$ *for some* $c_\alpha > 0$ *and* $\rho \in (0, 1)$. $\{w_{it} : t = 1, 2, ...\}$ *are independent across* i.

(ii) *For each* $\eta > 0$, $\min_i[\inf_{(\beta_i,\mu_i):\left\|(\beta_i,\mu_i)-(\beta_i^0,\mu_i^0)\right\|>\eta} \Psi_i(\beta_i, \mu_i) - \Psi_i\left(\beta_i^0, \mu_i^0\right)] > 0$.

(iii) *Let* Θ *denote the parameter space for* $\theta_i = (\beta_i', \mu_i)'$. Θ *is a compact and convex subset of* $\mathbb{R}^{p+1}$ *such that* $\theta_i^0 = (\beta_i^{0\prime}, \mu_i^0)'$ *lies in the interior of* Θ *for each* i.

(iv) *Let* $|v| \equiv \sum_{j=1}^{p+1} v_j$ *and* $D^v\psi(w_{it}; \theta) \equiv \partial^{|v|}\psi(w_{it}; \theta)/(\partial\theta_{(1)}\cdots\partial\theta_{(p+1)})$ *where* $v = (v_1, \cdots, v_{p+1})$ *is a vector of nonnegative integers and* $\theta_{(j)}$ *denotes the* j*th element of* θ. *There exists a function* $M(\cdot)$ *such that* $\sup_{\theta\in\Theta}\|D^v\psi(w_{it}; \theta)\| \leq M(w_{it})$, $\left\|D^v\psi(w_{it}; \theta) - D^v\psi\left(w_{it}; \bar{\theta}\right)\right\| \leq M(w_{it})\left\|\theta - \bar{\theta}\right\|$ *for any* $\theta, \bar{\theta} \in \Theta$ *and* $|v| \leq 3$, *and* $\max_i \mathbb{E}\left|M(w_{it})\right|^q < c_M$ *for some* $c_M < \infty$ *and* $q \geq 6$.

(v) *There exists a constant* $c_H > 0$ *such that* $\min_i \inf_{\beta\in\mathcal{B}} H_{i\mu\mu}(\beta) \geq c_H$ *and* $\min_i \mu_{\min}\left(H_{i\beta\beta}\left(\beta_i^0\right)\right) \geq c_H$.

(vi) *There exists a constant* $c_\alpha > 0$ *such that* $\min_{1\leq k<l\leq K_0}\left\|\alpha_k^0 - \alpha_l^0\right\| \geq c_\alpha$.

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

ASSUMPTION A2. (i) $T\lambda_1^2/(\ln T)^{6+2\nu} \to \infty$ *and* $\lambda_1(\ln T)^\nu \to 0$ *for some* $\nu > 0$ *as* $(N, T) \to \infty$.

(ii) $N^{1/2}T^{-1}(\ln T)^9 \to 0$ *and* $N^2T^{1-q/2} \to c \in [0, \infty)$ *as* $(N, T) \to \infty$.

ASSUMPTION A3. (i) *For each* $k = 1, ..., K_0$, $\Omega_k \equiv \lim_{(N_k,T)\to\infty} \frac{1}{N_k} \sum_{i \in G_k^0} \Omega_{iT}$ *exists and* $\Omega_k > 0$.

(ii) *For each* $k = 1, ..., K_0$, $\mathbb{H}_k \equiv \lim_{(N_k,T)\to\infty} \mathbb{H}_{kNT}$ *exists and* $\mathbb{H}_k > 0$.

Assumption A1(i) imposes conditions on $\{w_{it}\}$, which are commonly assumed for dynamic nonlinear panel data model; see, e.g., Hahn and Kuersteiner (2011) and Lee and Phillips (2015). With more complicated notation, we can relax the stationarity assumption along the time dimension. A1(ii) imposes an identification condition for the joint identification of (β_i, μ_i) for each i. A1(iii) restricts the parameter space and it is possible to allow Θ to be i-dependent. A1(iv) specifies the smoothness and moment conditions on ψ or objects associated with it. A1(v), in conjunction with A1(ii) and (iv), implies that $\min_i[\inf_{\mu_i:|\mu_i-\mu_i(\beta_i)|>\eta} \Psi_i(\beta_i,\mu_i) - \Psi_i(\beta_i,\mu_i(\beta_i))] > 0$ and $\min_i[\inf_{\beta_i:\|\beta_i-\beta_i^0\|>\eta} \Psi_i(\beta_i,\mu_i(\beta_i)) - \Psi_i(\beta_i^0,\mu_i(\beta_i^0))] > 0$. A1(vi) specifies that the group-specific parameters are separated from each other, similar to the separation requirement in Hahn and Moon (2010). A1(vii) implies that each group has an asymptotically non-negligible membership number of individuals as $N \to \infty$. This assumption can also be relaxed at the cost of more lengthy arguments. Assumption A2(i) imposes conditions on λ_1, all of which hold if

$$\lambda_1 \propto T^{-a} \text{ for any } a \in (0, 1/2). \tag{2.6}$$

A2(ii) is needed to ensure some higher order terms are asymptotically negligible. A3 is used to derive the asymptotic bias and variance of the C-Lasso estimator. The theory developed below under these conditions does not require correct specification of the likelihood function and the C-Lasso asymptotics apply under the general QMLE setup.

2.4 Asymptotic Properties of the PPL C-Lasso Estimators

2.4.1 Preliminary Rates of Convergence for Coefficient Estimates

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

Theorem 2.1 *Suppose that Assumption A1 holds and* $\lambda_1 = o(1)$. *Then (i)* $\hat{\beta}_i - \beta_i^0 = O_P\left(T^{-1/2} + \lambda_1\right)$ *for* $i = 1, 2, ..., N$, *(ii)* $\frac{1}{N}\sum_{i=1}^N \left\|\hat{\beta}_i - \beta_i^0\right\|^2 = O_P\left(T^{-1}\right)$, *and (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. Theorem 2.1(i)-(ii) establish the pointwise and mean-square convergence of $\hat{\beta}_i$. Theorem 2.1(iii) indicates that the group-specific parameters $\alpha_1^0, ..., \alpha_K^0$ can 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 $\hat{\alpha}_k$. For notational simplicity, hereafter we simply write $\hat{\alpha}_k$ for $\hat{\alpha}_{(k)}$ as the consistent estimator of α_k^0, and define

$$\hat{G}_k = \left\{i \in \{1, 2, ..., N\} : \hat{\beta}_i = \hat{\alpha}_k\right\} \text{ for } k = 1, ..., K_0. \tag{2.7}$$

2.4.2 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

$$\hat{E}_{kNT,i} \equiv \left\{i \notin \hat{G}_k \mid i \in G_k^0\right\} \text{ and } \hat{F}_{kNT,i} \equiv \left\{i \notin G_k^0 \mid i \in \hat{G}_k\right\}, \tag{2.8}$$

where $i = 1, ..., N$ and $k = 1, ..., K_0$. Let $\hat{E}_{kNT} = \cup_{i \in G_k^0} \hat{E}_{kNT,i}$ and $\hat{F}_{kNT} = \cup_{i \in \hat{G}_k} \hat{F}_{kNT,i}$. $\hat{E}_{kNT}$ 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 $\hat{G}_k$. Both types of errors must be controlled. We use the following definition.

Definition 1. (Consistent classification) The classification is *individually consistent* if $P(\hat{E}_{kNT,i}) \to 0$ as $(N, T) \to \infty \; \forall \; i \in G_k^0$ and $k \in \{1, ..., K_0\}$, and $P(\hat{F}_{kNT,i}) \to 0$ as $(N, T) \to \infty \; \forall \; i \in \hat{G}_k$ and $k \in \{1, ..., K_0\}$. It is *uniformly consistent* if $P(\cup_{k=1}^{K_0} \hat{E}_{kNT}) \to 0$ and $P(\cup_{k=1}^{K_0} \hat{F}_{kNT}) \to 0$ as $(N, T) \to \infty$.

The following theorem establishes uniform consistency for the PPL classifier.

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

REMARK 2. Theorem 2.2 implies that all individuals within a 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 $\hat{G}_0 \equiv \{1, 2, ..., N\} \backslash (\cup_{k=1}^{K_0} \hat{G}_k)$ and $\hat{H}_{iNT} \equiv \{i \in \hat{G}_0\}$. 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, a small percentage of individuals could be left unclassified if we stick with the classification rule in (2.7). To ensure that all individuals are classified into one of the K_0 groups in finite samples, we can modify the classifier. 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 $||\hat{\beta}_i - \hat{\alpha}_l|| = \min\{||\hat{\beta}_i - \hat{\alpha}_1||, ..., ||\hat{\beta}_i - \hat{\alpha}_{K_0}||\}$ and $\sum_{k=1}^{K_0} \mathbf{1}\{\hat{\beta}_i = \hat{\alpha}_k\} = 0$. Since the event $\sum_{k=1}^{K_0} \mathbf{1}\{\hat{\beta}_i = \hat{\alpha}_k\} = 0$ occurs w.p.a.1 uniformly in i, we can ignore it in large samples in subsequent theoretical analysis and restrict our attention to the classification rule in (2.7) to avoid confusion.

Let $\hat{N}_k \equiv \sum_{i=1}^{N} \mathbf{1}\{i \in \hat{G}_k\}$. The following corollary studies the consistency of $\hat{N}_k$.

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

2.4.3 The Oracle Property and Asymptotic Properties of Post-Lasso Estimators

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

Theorem 2.4 *Suppose Assumptions A1-A3 hold. Then* $\sqrt{N_k T}\left(\hat{\alpha}_k - \alpha_k^0\right) - \mathbb{H}_{kNT}^{-1} \mathbb{B}_{kNT} \xrightarrow{D} N\left(0, \mathbb{H}_k^{-1} \Omega_k (\mathbb{H}_k^{-1})'\right)$, *where* $\mathbb{B}_{kNT} = \mathbb{B}_{1kNT} - \mathbb{B}_{2kNT}$, $\mathbb{B}_{1kNT} = \frac{1}{\sqrt{N_k T^3}} \sum_{i \in G_k^0} m_{iV}^{-1} \sum_{s=1}^{T} \sum_{t=1}^{T} V_{is} \mathbb{U}_{it}^{\mu_i}$, *and* $\mathbb{B}_{2kNT} = \frac{1}{2\sqrt{N_k T}} \sum_{i \in G_k^0}$ $m_{iV}^{-2} (m_{iU2} - \frac{m_{iV2}}{m_{iV}} m_{iU}) (\frac{1}{\sqrt{T}} \sum_{t=1}^{T} V_{it})^2$ *for* $k = 1, ..., K_0$.

REMARK 3. $\mathbb{B}_{kNT}$ is written as the difference between two terms that are derived from the first and second order Taylor expansions of the PPL estimating equation, respectively. Comparing the above result with HK, we find that the quantities Ω_k, $\mathbb{H}_k$, and $\mathbb{B}_k$ coincide with the corresponding terms in HK; see the remark after Lemma S1.12 for details. Then we can use the formula in HK to estimate the asymptotic bias and variance with obvious modifications. Alternatively, we can use the jackknife to correct bias; see Hahn and Newey (2004) and Dhaene and Jachmans (2015) for static and dynamic models, respectively.

If group membership is known, the *oracle* estimator of α_k is given by $\hat{\alpha}_{G_k^0} \equiv \arg\min_{\alpha_k} \frac{1}{N_k T} \sum_{i \in G_k^0} \sum_{t=1}^{T} \psi(w_{it}; \alpha_k, \hat{\mu}_i(\alpha_k))$. Then following our asymptotic analysis or that of HK, we can readily show that $\sqrt{N_k T}(\hat{\alpha}_{G_k^0} - \alpha_k^0) - \mathbb{H}_{kNT}^{-1}\mathbb{B}_{kNT} \xrightarrow{D} N\left(0, \mathbb{H}_k^{-1}\Omega_k(\mathbb{H}_k^{-1})'\right)$ under Assumptions A1 and A3. Theorem 2.4 indicates that the PPL estimator $\hat{\alpha}_k$ achieves the same limit distribution as this oracle estimator. In this sense, we say that the PPL estimators $\{\hat{\alpha}_k\}$ enjoy the asymptotic oracle property. In addition, given the estimated groups $\hat{G}_k$, we can obtain the post-Lasso estimator of α_k by $\hat{\alpha}_{\hat{G}_k} \equiv \arg\min_{\alpha_k} \frac{1}{N_k T} \sum_{i \in \hat{G}_k} \sum_{t=1}^{T} \psi(w_{it}; \alpha_k, \hat{\mu}_i(\alpha_k))$. The following theorem reports the asymptotic distribution of $\hat{\alpha}_{\hat{G}_k}$

Theorem 2.5 *Suppose Assumptions A1-A3 hold. Then* $\sqrt{N_k T}(\hat{\alpha}_{\hat{G}_k} - \alpha_k^0) - \mathbb{H}_{kNT}^{-1}\mathbb{B}_{kNT} \xrightarrow{D} N\left(0, \mathbb{H}_k^{-1}\Omega_k(\mathbb{H}_k^{-1})'\right)$ *for* $k = 1, ..., K_0$, *where* $\mathbb{B}_{kNT}$ *is as defined in Theorem 2.4.*

REMARK 4. Theorems 2.4 and 2.5 indicate that $\hat{\alpha}_k$ and $\hat{\alpha}_{\hat{G}_k}$ are asymptotically equivalent. 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 smaller bias. Correspondingly, it would be interesting to compare the higher-order asymptotic properties of $\hat{\alpha}_k$ and $\hat{\alpha}_{\hat{G}_k}$ in future work.

REMARK 5. Note that our asymptotic results are "pointwise" in the sense that the unknown parameters are treated as fixed. The implication is that in finite samples, the distributions of our estimators can be quite different from normal, as discussed in Leeb and Pötscher (2008, 2009). This is a well-known challenge for shrinkage estimators. Despite its importance, developing a thorough theory on uniform inference in this context is beyond the scope of the present work.

2.5 Determination of the Number of Groups

In practice, the exact number of groups is typically unknown. We assume that K_0 is bounded from above by a finite integer $K_{\max}$ and study the determination of the number of groups via some information criterion (IC). By minimizing (2.5) with K_0 replaced by K, 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 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., $\hat{G}_k(K, \lambda_1) \equiv \{i \in \{1, 2, ..., N\} : \hat{\beta}_i(K, \lambda_1) = \hat{\alpha}_k(K, \lambda_1)\}$ for $k = 1, ..., K$. Let $\hat{G}(K, \lambda_1) \equiv \{\hat{G}_1(K, \lambda_1), ..., \hat{G}_K(K, \lambda_1)\}$. The post-Lasso estimator of α_k^0 is denoted as $\hat{\alpha}_{\hat{G}_k(K, \lambda_1)}$. We propose to select K to minimize

$$IC_1(K, \lambda_1) \equiv \frac{2}{NT} \sum_{k=1}^{K} \sum_{i \in \hat{G}_k(K, \lambda_1)} \sum_{t=1}^{T} \psi\left(w_{it}; \hat{\alpha}_{\hat{G}_k(K,\lambda_1)}, \hat{\mu}_i(\hat{\alpha}_{\hat{G}_k(K,\lambda_1)})\right) + \rho_{1NT} pK, \tag{2.9}$$

where ρ_{1NT} is a tuning parameter. Let $\hat{K}(\lambda_1) \equiv \arg\min_{1 \le K \le K_{\max}} IC_1(K, \lambda_1)$. See Wang, Li, and Tsai (2007), Liao (2013), and Lu and Su (2016) for the use of a similar IC in various contexts.

Let $G^{(K)} \equiv (G_{K,1}, ..., G_{K,K})$ be any K-partition of $\{1, 2, ..., N\}$ and $\mathcal{G}_K$ a collection of all such partitions. Let $\hat{\sigma}^2_{G^{(K)}} \equiv \frac{2}{NT} \sum_{k=1}^{K} \sum_{i \in G_{K,k}} \sum_{t=1}^{T} \psi\left(w_{it}; \hat{\alpha}_{G_{K,k}}, \hat{\mu}_i(\hat{\alpha}_{G_{K,k}})\right)$, where $\hat{\alpha}_{G_{K,k}} \equiv \arg\min_{\alpha_k} \frac{1}{N_k T} \sum_{i \in G_{K,k}} \sum_{t=1}^{T} \psi(w_{it}; \alpha_k, \hat{\mu}_i(\alpha_k))$. We add the following two assumptions.

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* $\sigma_0^2 \equiv lim_{(N,T)\to\infty} \frac{2}{NT} \sum_{k=1}^{K_0} \sum_{i \in G_k^0} \sum_{t=1}^{T} \mathbb{E}[\psi(w_{it}; \alpha_k^0, \mu_i^0)]$.

ASSUMPTION A5. *As* $(N,T)\to\infty$, $\rho_{1NT}\to 0$ *and* $\rho_{1NT}T\to\infty$.

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: ρ_{1NT} cannot shrink to zero either too fast or too slowly.

The following theorem justifies the use of (2.9) as a selector criterion for K.

Theorem 2.6 *Suppose Assumptions A1-A5 hold. Then* $P(\hat{K}(\lambda_1)=K_0)\to 1$ *as* $(N,T)\to\infty$.

REMARK 6. As Theorem 2.6 indicates, as long as λ_1 satisfies Assumption A2(i), we can ensure that the correct number of groups is chosen w.p.a.1. In practice, we can fine-tune this parameter over a finite set, e.g., $\Lambda_1\equiv\{\lambda_1=c_jT^{-1/3}, c_j=c_0\gamma^j \text{ for } j=1,...,J\}$ for some $c_0>0$ and $\gamma>1$. That is, we pick up $\lambda_1\in\Lambda_1$ such that $IC_1(\hat{K}(\lambda_1),\lambda_1)$ is minimized. We can show that with such a choice of λ_1, Theorem 2.6 continues to hold. Alternatively, we can consider a data-driven cross-validation procedure.

2.6 The Special Case of Linear Models

For the linear model in (2.3) with $\mathbb{E}\left(\varepsilon_{it}|x_{it},\mu_i^0\right)=0$, we have $\psi\left(w_{it};\beta_i,\mu_i\right)=\frac{1}{2}\left(y_{it}-\beta_i'x_{it}-\mu_i\right)^2$, $\hat{\mu}_i\left(\beta_i\right)=\bar{y}_{i\cdot}-\beta_i'\bar{x}_{i\cdot}$, and $Q_{1,NT}\left(\boldsymbol{\beta}\right)=\frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T\left(\tilde{y}_{it}-\beta_i'\tilde{x}_{it}\right)^2$, where $\bar{y}_{i\cdot}=\frac{1}{T}\sum_{t=1}^T y_{it}$, $\tilde{y}_{it}=y_{it}-\bar{y}_{i\cdot}$, and $\bar{x}_{i\cdot}$ and $\tilde{x}_{it}$ are analogously defined. So the PPL problem becomes the penalized least squares (PLS) problem considered in Su, Shi, and Phillips (2014, SSP hereafter). In addition, we can verify that $\mu_i\left(\beta_i\right)=\mathbb{E}(\bar{y}_{i\cdot})-\beta_i'\mathbb{E}(\bar{x}_{i\cdot})$, $U_i\left(w_{it};\beta_i,\mu_i\right)=-\left(y_{it}-\beta_i'x_{it}-\mu_i\right)x_{it}$, $V_i\left(w_{it};\beta_i,\mu_i\right)=-\left(y_{it}-\beta_i'x_{it}-\mu_i\right)$, $U_i^{\mu_i}\left(w_{it};\beta_i,\mu_i\right)=x_{it}=V_i^{\beta_i}\left(w_{it};\beta_i,\mu_i\right)$, $V_i^{\mu_i}\left(w_{it};\beta_i,\mu_i\right)=1$, $U_i^{\beta_i}\left(w_{it};\beta_i,\mu_i\right)=x_{it}x_{it}'$, $\mathbb{U}_{it}=-\varepsilon_{it}\left[x_{it}-\mathbb{E}\left(\bar{x}_{i\cdot}\right)\right]$, $\mathbb{U}_{it}^{\beta_i}=\left[x_{it}-\mathbb{E}\left(\bar{x}_{i\cdot}\right)\right]x_{it}'$, $\mathbb{U}_{it}^{\mu_i}=x_{it}-\mathbb{E}\left(\bar{x}_{i\cdot}\right)$, $H_{i\mu\mu}\left(\beta_i\right)=1$, $H_{i\beta\beta}\left(\beta_i\right)=\frac{1}{T}\sum_{t=1}^T\mathbb{E}\{[(x_{it}-\mathbb{E}(\bar{x}_{i\cdot})]\left[x_{it}-\mathbb{E}(\bar{x}_{i\cdot})\right]'\}$,

$$\begin{aligned}\Omega_{iT} &= \frac{1}{T}\sum_{t=1}^T\sum_{s=1}^T\mathbb{E}\{\varepsilon_{it}\varepsilon_{is}\left[x_{it}-\mathbb{E}\left(\bar{x}_{i\cdot}\right)\right]\left[x_{is}-\mathbb{E}\left(\bar{x}_{i\cdot}\right)\right]'\}, \text{ and}\\ \mathbb{H}_{iT} &= \frac{1}{T}\sum_{t=1}^T\mathbb{E}\{\left[x_{it}-\mathbb{E}\left(\bar{x}_{i\cdot}\right)\right]\left[x_{it}-\mathbb{E}\left(\bar{x}_{i\cdot}\right)\right]'\}.\end{aligned}$$

With the above calculations, we can readily verify that Assumptions A1(ii), (iv)-(v) and A3 hold under weak conditions. In addition, we can show that

$$\mathbb{B}_{1kNT}=\frac{-1}{\sqrt{N_kT^3}}\sum_{i\in G_k^0}\sum_{t=1}^T\sum_{s=1}^T\varepsilon_{it}\left[x_{is}-\mathbb{E}\left(\bar{x}_{i\cdot}\right)\right]=\mathbb{B}_{1k}+o_P\left(1\right) \text{ and } \mathbb{B}_{2kNT}=0,$$

where $\mathbb{B}_{1k}=\frac{-1}{\sqrt{N_kT^3}}\sum_{i\in G_k^0}\sum_{t=1}^T\sum_{s=1}^T\mathbb{E}\left(\varepsilon_{it}x_{is}\right)$. When x_{it} is strictly exogenous so that $\mathbb{E}(\varepsilon_{it}x_{is})=0$ for all t,s, and i, $\mathbb{B}_{1kNT}=o_P\left(1\right)$ and there is no need to make bias correction. When x_{it} is predetermined, various bias correction formulae have been proposed; see Kiviet (1995), Hahn and Kuersteiner (2002), Phillips and Sul (2007b), and Lee (2012), among others. Jackknife methods can also be applied to correct for bias.

The post-Lasso and oracle estimators of α_k^0 become $\hat{\alpha}_{\hat{G}_k}=(\sum_{i\in\hat{G}_k}\sum_{t=1}^T\tilde{x}_{it}\tilde{x}_{it}')^{-1}\sum_{i\in\hat{G}_k}\sum_{t=1}^T\tilde{x}_{it}\tilde{y}_{it}$ and $\hat{\alpha}_{G_k^0}=(\sum_{i\in G_k^0}\sum_{t=1}^T\tilde{x}_{it}\tilde{x}_{it}')^{-1}\sum_{i\in G_k^0}\sum_{t=1}^T\tilde{x}_{it}\tilde{y}_{it}$. The IC formula in (2.9) now reduces to $IC_1\left(K,\lambda_1\right)=\hat{\sigma}^2_{\hat{G}(K,\lambda_1)}+\rho_{1NT}pK$, where $\hat{\sigma}^2_{\hat{G}(K,\lambda_1)}=\frac{1}{NT}\sum_{k=1}^K\sum_{i\in\hat{G}_k(K,\lambda_1)}\sum_{t=1}^T(\tilde{y}_{it}-\hat{\alpha}'_{\hat{G}_k(K,\lambda_1)}\tilde{x}_{it})^2$ with $\hat{\alpha}_{\hat{G}_k(K,\lambda_1)}$ being analogously defined as $\hat{\alpha}_{\hat{G}_k}$. In practice, $\hat{\sigma}^2_{\hat{G}(K,\lambda_1)}$ is frequently replaced by its natural logarithm as in standard BIC to obtain

$$IC_1(K,\lambda_1) = \ln\left[\hat{\sigma}^2_{\hat{G}(K,\lambda_1)}\right] + \rho_{1NT}pK, \tag{2.10}$$

which will be used in our simulations and applications. But because the fixed effects are eliminated in the within-group transformed model, the $\sqrt{T}$-convergence rates of their estimates won't play a role to ensure the selection consistency of IC_1. SSP show that the requirement on ρ_{1NT} can be relaxed with Assumption A5 replaced by:

ASSUMPTION A5*. *As* $(N,T) \to \infty$, $\rho_{1NT} \to 0$ *and* $\rho_{1NT}\delta^2_{NT} \to \infty$ *where* $\delta_{NT} = N^{1/2}T^{1/2}$ *if* x_{it} *is strictly exogenous and* $\min(N^{1/2}T^{1/2}, T)$ *otherwise.*

2.7 Extension to the Mixed Panel Structure Models

In some applications, certain parameters of interest may be common across all individuals whereas others are group-specific. For instance, Pesaran, Shin, and Smith (1999) constrain the long-run coefficients to be identical across individuals while assuming the short-run coefficients to be heterogenous, or in our case, group-specific. Example 4 above is another instance. To keep up with the early notation, we write the negative log-likelihood function as $\psi(w_{it};\beta_i,\gamma,\mu_i)$ where γ is the common parameter and the β_i have a group structure as before. The negative profile log-likelihood function now becomes $Q_{1,NT}(\boldsymbol{\beta},\gamma) = \frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T\psi(w_{it};\beta_i,\gamma,\hat{\mu}_i(\beta_i,\gamma))$, where $\hat{\mu}_i(\beta_i,\gamma) = \arg\min_{\mu_i}\frac{1}{T}\sum_{t=1}^T\psi(w_{it};\beta_i,\gamma,\mu_i)$. Then we can estimate $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ by minimizing the following PPL criterion function

$$Q^{(K_0)}_{1NT,\lambda_1}(\boldsymbol{\beta},\boldsymbol{\alpha},\gamma) = Q_{1,NT}(\boldsymbol{\beta},\gamma) + \frac{\lambda_1}{N}\sum_{i=1}^N \Pi_{k=1}^{K_0}\|\beta_i - \alpha_k\|. \tag{2.11}$$

Our previous analysis can be followed to establish uniform consistency for the classifier and the oracle property for the resulting estimators of the group-specific parameters α_k and the common parameter γ.

When we have time effects $\{\gamma_t\}$, we generally cannot eliminate them through transformation even in a linear panel structure model because of the slope heterogeneity. In this case, we need to estimate $\gamma = (\gamma_1,...,\gamma_T)'$ jointly with $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ in (2.11). A formal asymptotic analysis of this case is left for future work.

3 Penalized GMM Estimation of Panel Structure Models

This section considers penalized GMM estimation of linear panel structure models when some regressors are lagged dependent variables or endogenous.

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

To stay focused, we restrict attention to the linear panel structure model in (2.3).[1] We consider the first differenced system

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

[1] Extension to general nonlinear panel data models with endogeneity and nonadditive fixed effects (e.g., Fernández-Val and Lee 2013) is possible but rigorous analysis raises additional statistical challenges and is left for future research.

where, e.g., $\Delta y_{it} = y_{it} - y_{i,t-1}$ for $t = 1, ..., T$ and $i = 1, ..., N$, and we assume that y_{i0} and x_{i0} are observed. Let z_{it} be a $d \times 1$ vector of instruments for Δx_{it} with $d \geq p$. Define $\Delta y_i = (\Delta y_{i1}, ..., \Delta y_{iT})'$, with similar definitions for Δx_i and $\Delta \varepsilon_i$. We propose to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ by minimizing the following penalized GMM (PGMM) criterion function[2]

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

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 $d \times d$ symmetric matrix that is asymptotically nonsingular and $\lambda_2 = \lambda_{2NT}$ is a tuning parameter. Minimizing (3.2) produces the PGMM estimates $\tilde{\boldsymbol{\alpha}}$ and $\tilde{\boldsymbol{\beta}}$ where $\tilde{\boldsymbol{\alpha}} \equiv (\tilde{\alpha}_1, ..., \tilde{\alpha}_{K_0})$ and $\tilde{\boldsymbol{\beta}} \equiv (\tilde{\beta}_1, ..., \tilde{\beta}_N)$.

3.2 Basic Assumptions

Let $\tilde{Q}_{i,z\Delta x} \equiv \frac{1}{T} \sum_{t=1}^{T} z_{it} (\Delta x_{it})'$ and $\bar{Q}_{i,z\Delta x} \equiv \mathbb{E}[\tilde{Q}_{i,z\Delta x}]$. Let $\xi_{it} \equiv (\Delta y_{it}, (\Delta x_{it})', z_{it}')'$, $\rho(\xi_{it}, \beta) \equiv z_{it}(\Delta y_{it} - \beta' \Delta x_{it})$ and $\bar{\rho}_{i,T}(\beta) \equiv \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \{\rho(\xi_{it}, \beta) - \mathbb{E}[\rho(\xi_{it}, \beta)]\}$. Let $\mathcal{B}_i$ denote the parameter space for β_i. We make the following assumptions.

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) $\sup_{\beta \in \mathcal{B}_i} \|\bar{\rho}_{i,T}(\beta)\| = O_P(1)$, $\frac{1}{N} \sum_{i=1}^{N} \|\bar{\rho}_{i,T}(\beta_i)\|^2 = O_P(1)$ *where* $\beta_i \in \mathcal{B}_i$, *and* $P(\max_i \|\bar{\rho}_{i,T}(\beta_i)\| \geq C(\ln T)^{3+\nu}) = o\left(N^{-1}\right)$ *for any* $C > 0$ *and* $\nu > 0$.

(iii) $P\left(\max_i \left\|\tilde{Q}_{i,z\Delta x} - \bar{Q}_{i,z\Delta x}\right\| \geq \eta\right) = o\left(N^{-1}\right)$ *for any* $\eta > 0$ *and* $\liminf_{(N,T)\to\infty} \min_i \mu_{\min}(\bar{Q}_{i,z\Delta x}' \bar{Q}_{i,z\Delta x}) = \underline{c}_{\bar{Q}}^2 > 0$.

(iv) *There exist nonrandom matrices* W_i *such that* $P(\max_i \|W_{iNT} - W_i\| \geq \eta) = o\left(N^{-1}\right)$ *for any* $\eta > 0$ *and* $\liminf_{N\to\infty} \min_i \mu_{\min}(W_i) = \underline{c}_W > 0$.

(v) *There exists a constant* $c_\alpha > 0$ *such that* $\min_{1 \leq k < l \leq K_0} \left\|\alpha_k^0 - \alpha_l^0\right\| \geq c_\alpha$.

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

ASSUMPTION B2. (i) $T\lambda_2^2/(\ln T)^{6+2\nu} \to \infty$ *and* $\lambda_2(\ln T)^\nu \to 0$ *for some* $\nu > 0$ *as* $(N, T) \to \infty$.

(ii) *For any given* $c > 0$, $N \max_i P\left(\left\|T^{-1} \sum_{t=1}^{T} z_{it} \Delta \varepsilon_{it}\right\| \geq c\lambda_2\right) \to 0$ *as* $(N, T) \to \infty$.

ASSUMPTION B3. (i) *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* $(N, T) \to \infty$.

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

Assumption B1(i) specifies moment conditions to identify β_i^0. B1(ii) is a high level condition. Its first part can be verified by applying Donsker's theorem. 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 $\Sigma_i > 0$ and any nonrandom $\omega \in \mathbb{R}^d$ with $\|\omega\| = 1$, then $\bar{\rho}_{i,T}(\beta_i) \xrightarrow{D} N(0, \Sigma_i)$ and the first part of B1(ii) follows. The second and third parts of B1(ii) can be verified by the Markov inequality and the application of Lemma S1.2(iii) in the supplement under strong

[2] 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 $\tilde{Q}_{2,NT}(\boldsymbol{\beta}) = [\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} z_{it} (\Delta y_{it} - \beta_i' \Delta x_{it})]' W_{NT} [\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} z_{it} (\Delta y_{it} - \beta_i' \Delta x_{it})]$, where W_{NT} is asymptotically nonsingular. We also found that Arellano and Bond (1991) GMM estimation is not applicable to handle unobserved slope heterogeneity. Noticing this, Fernández-Val and Lee (2013) used a criterion similar to $Q_{2,NT}(\boldsymbol{\beta})$ in the nonlinear panel setup. As we shall see, the use of $Q_{2,NT}(\boldsymbol{\beta})$ means that the PGMM estimator generally does not have the oracle property.

mixing conditions. B1(iii) provides a rank condition to identify β_i^0. B1(iv) is automatically satisfied for $W_{iNT} = I_d$, the $d \times d$ identity matrix. B1(v)-(vi) and B2(i) parallel A1(vi)-(vii) and A2(i). B2(ii) holds true by Lemma S1.2 in the supplement if $\{(z_{it}, \Delta\varepsilon_{it}), t \geq 1\}$ is strong mixing with geometric decay rate and $z_{it}\Delta\varepsilon_{it}$ has six plus moments.

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

$$
\begin{aligned}
& \frac{1}{\sqrt{N_k T}} \sum_{i\in G_k^0} \tilde{Q}_{i,z\Delta x}' \sum_{t=1}^{T} z_{it}\Delta\varepsilon_{it} \\
= \; & \frac{1}{N_k^{1/2}T^{3/2}} \sum_{i\in G_k^0}\sum_{s=1}^{T}\sum_{t=1}^{T} \mathbb{E}\left(\Delta x_{is} z_{is}' z_{it}\Delta\varepsilon_{it}\right) + \frac{1}{N_k^{1/2}T^{3/2}} \sum_{i\in G_k^0}\sum_{s=1}^{T}\sum_{t=1}^{T} \mathbb{E}\left(\Delta x_{is} z_{is}'\right) z_{it}\Delta\varepsilon_{it} \\
& + \frac{1}{N_k^{1/2}T^{3/2}} \sum_{i\in G_k^0}\sum_{s=1}^{T}\sum_{t=1}^{T} \left\{\left[\Delta x_{is} z_{is}' - \mathbb{E}\left(\Delta x_{is} z_{is}'\right)\right] z_{it}\Delta\varepsilon_{it} - \mathbb{E}\left(\Delta x_{is} z_{is}' z_{it}\Delta\varepsilon_{it}\right)\right\} \\
\equiv \; & B_{kNT} + V_{kNT} + R_{kNT}, \text{ say,}
\end{aligned}
\tag{3.3}
$$

where B_{kNT} and V_{kNT} contribute to the asymptotic bias and variance, respectively, and R_{kNT} is a term that is asymptotically negligible under suitable conditions. Then B3(ii) is satisfied with $W_{iNT} = I_d$ if $V_{kNT} = \frac{1}{N_k^{1/2}T^{1/2}}\sum_{i\in G_k^0}\sum_{t=1}^{T} \bar{Q}_{i,z\Delta x}' z_{it}\Delta\varepsilon_{it} \xrightarrow{D} N(0, C_k)$ and $R_{kNT} = o_P(1)$, both of which can be verified by strengthening the conditions given in (a)-(c) above. Note that $\bar{A}_k^{-1} B_{kNT}$ signifies the asymptotic bias of $\tilde{\alpha}_k$, which may not vanish asymptotically but can be corrected; see Section S2.2 in the supplement.[3]

3.3 Asymptotic Properties of the PGMM Estimators

3.3.1 Preliminary Rates of Convergence

We first establish the preliminary consistency rate of $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}})$.

Theorem 3.1 *Suppose Assumption B1 holds and* $\lambda_2 = o(1)$. *Then (i)* $\tilde{\beta}_i - \beta_i^0 = O_P\left(T^{-1/2} + \lambda_2\right)$ *for* $i = 1, ..., N$, *(ii)* $\frac{1}{N}\sum_{i=1}^{N}\left\|\tilde{\beta}_i - \beta_i^0\right\|^2 = O_P\left(T^{-1}\right)$, *and (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. Remark 1 applies here with obvious modifications. As before, hereafter we simply write $\tilde{\alpha}_k$ for $\tilde{\alpha}_{(k)}$ as the consistent estimator of α_k^0, and define $\tilde{G}_k \equiv \{i \in \{1, 2, ..., N\} : \tilde{\beta}_i = \tilde{\alpha}_k\}$ for $k = 1, ..., K_0$.

[3] If Conditions (a)-(b) are satisfied and $E\left\|z_{it}\Delta\varepsilon_{it}\right\|^{2+\sigma} > 0$, by the Davydov inequality, we have $\|B_{kNT}\| \leq \frac{1}{T\sqrt{N_k T}}\sum_{i\in G_k^0}\sum_{t=1}^{T}\sum_{s=1}^{T}\left\|E\left[\Delta x_{is} z_{is}' z_{it}\Delta\varepsilon_{it}\right]\right\| = O\left((N/T)^{1/2}\right)$, which is $o(1)$ if $T \gg N$ and usually asymptotically non-negligible otherwise.

3.3.2 Classification Consistency

Let $\tilde{E}_{kNT,i} \equiv \{i \notin \tilde{G}_k \mid i \in G_k^0\}$ and $\tilde{F}_{kNT,i} \equiv \{i \notin G_k^0 \mid i \in \tilde{G}_k\}$ for $i = 1, ..., N$ and $k = 1, ..., K_0$. Let $\tilde{E}_{kNT} \equiv \cup_{i \in G_k^0} \tilde{E}_{kNT,i}$ and $\tilde{F}_{kNT} \equiv \cup_{i \in \tilde{G}_k} \tilde{F}_{kNT,i}$. We establish uniform classification consistency in the next theorem.

Theorem 3.2 *Suppose that Assumptions B1-B2 hold. Then (i) $P(\cup_{k=1}^{K_0} \tilde{E}_{kNT}) \leq \sum_{k=1}^{K_0} P(\tilde{E}_{kNT}) \to 0$ as $(N,T) \to \infty$, and (ii) $P(\cup_{k=1}^{K_0} \tilde{F}_{kNT}) \leq \sum_{k=1}^{K_0} P(\tilde{F}_{kNT}) \to 0$ as $(N,T) \to \infty$.*

REMARK 8. Remark 2 also holds for the above theorem with obvious modifications. Let $\tilde{G}_0 \equiv \{1, 2, ..., N\} \backslash (\cup_{k=1}^{K_0} \tilde{G}_k)$ and $\tilde{H}_{iNT} = \{i \in \tilde{G}_0\}$. Theorem 3.2(i) implies that $P(\cup_{1 \leq i \leq N} \tilde{H}_{iNT}) \leq \sum_{k=1}^{K_0} P(\tilde{E}_{kNT})$ $\to 0$, meaning that all individuals are classified into one of the K_0 groups w.p.a.1.

Let $\tilde{N}_k \equiv \sum_{i=1}^{N} \mathbf{1}\{i \in \tilde{G}_k\}$. The following corollary parallels Corollary 2.3.

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

3.3.3 Improved Convergence and Asymptotic Properties of Post-Lasso

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

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

REMARK 9. In contrast to the PPL 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 $\tilde{\alpha}_{G_k^0}$ of α_k^0 by minimizing the following objective function

$$\tilde{Q}_{NT}(\alpha_k) = \left[\frac{1}{N_k T} \sum_{i \in G_k^0} \sum_{t=1}^{T} z_{it} \left(\Delta y_{it} - \alpha_k' \Delta x_{it}\right)\right]' W_{NT}^{(k)} \left[\frac{1}{N_k T} \sum_{i \in G_k^0} \sum_{t=1}^{T} z_{it} \left(\Delta y_{it} - \alpha_k' \Delta x_{it}\right)\right], \quad (3.4)$$

where $W_{NT}^{(k)}$ is a $d \times d$ symmetric positive definite matrix. Let $Q_{z\Delta x,NT}^{(k)} = \frac{1}{N_k T} \sum_{i \in G_k^0} \sum_{t=1}^{T} z_{it} \left(\Delta x_{it}\right)'$ 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 $\tilde{\alpha}_{G_k^0} = [Q_{z\Delta x,NT}^{(k)\prime} W_{NT}^{(k)} Q_{z\Delta x,NT}^{(k)}]^{-1}\ Q_{z\Delta x,NT}^{(k)\prime} W_{NT}^{(k)} Q_{z\Delta y,NT}^{(k)}$. We can readily show that the asymptotic distribution of $\tilde{\alpha}_{G_k^0}$ is typically different from that of $\tilde{\alpha}_k$. 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 (3.4) and obtain the post-Lasso GMM estimator of α_k^0 given by $\tilde{\alpha}_{\tilde{G}_k} = [\tilde{Q}_{z\Delta x}^{(k)\prime} W_{NT}^{(k)} \tilde{Q}_{z\Delta x}^{(k)}]^{-1} \tilde{Q}_{z\Delta x}^{(k)\prime} W_{NT}^{(k)} \tilde{Q}_{z\Delta y}^{(k)}$, where $\tilde{Q}_{z\Delta x}^{(k)} = \frac{1}{N_k T} \sum_{i \in \tilde{G}_k} \sum_{t=1}^{T} z_{it} \left(\Delta x_{it}\right)'$ 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) $\frac{1}{\sqrt{N_k T}} \sum_{i \in G_k^0} \sum_{t=1}^{T} z_{it} \Delta \varepsilon_{it} \xrightarrow{D} N(0, V_k)$.

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

Theorem 3.5 *Suppose 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. To prove the above theorem, we first apply Theorem 3.2 and show that $\sqrt{N_k T}\left(\tilde{\alpha}_{\tilde{G}_k} - \alpha_k^0\right) = \sqrt{N_k T}(\tilde{\alpha}_{G_k^0} - \alpha_k^0) + o_P(1)$. That is, the post-Lasso GMM estimator $\tilde{\alpha}_{\tilde{G}_k}$ is asymptotically equivalent to the oracle estimator $\tilde{\alpha}_{G_k^0}$. 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}. Alternatively, we can consider Arellano and Bond (1991) GMM estimation based on the estimated groups. The procedure is standard and details are omitted.

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 Assumption B3(i) (which is unrealistic before knowing the group identity), and $B_{kNT} = 0$ in Assumption B3(ii), 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 $\sqrt{N_k T}\left(\tilde{\alpha}_k - \alpha_k^0\right) \xrightarrow{D} N(0, \Omega_k)$. Thus in this special case the C-Lasso estimator $\tilde{\alpha}_k$ has the oracle property. But $B_{kNT} = 0$ typically requires $T \gg N$, a condition that we do not usually want to impose. For this reason we recommend the post-Lasso estimator $\tilde{\alpha}_{\tilde{G}_k}$ in practice.

3.4 Determination of the Number of Groups

When K_0 is unknown, we minimize the PGMM criterion function in (3.2) with K_0 replaced by K to obtain the C-Lasso estimates $\{\tilde{\beta}_i(K, \lambda_2), \tilde{\alpha}_k(K, \lambda_2)\}$ of $\{\beta_i, \alpha_k\}$. As above, we classify individual i into group $\tilde{G}_k(K, \lambda_2)$ if and only if $\tilde{\beta}_i(K, \lambda_2) = \tilde{\alpha}_k(K, \lambda_2)$. Let $\tilde{G}(K, \lambda_2) \equiv \{\tilde{G}_1(K, \lambda_2), ..., \tilde{G}_K(K, \lambda_2)\}$. The post-Lasso GMM estimate of α_k^0 is given by $\tilde{\alpha}_{\tilde{G}_k(K,\lambda_2)} \equiv [\tilde{Q}_{z\Delta x}^{(K,k)\prime} W_{NT}^{(k)} \tilde{Q}_{z\Delta x}^{(K,k)}]^{+} \tilde{Q}_{z\Delta x}^{(K,k)\prime} W_{NT}^{(k)} \tilde{Q}_{z\Delta y}^{(K,k)}$, where $\tilde{Q}_{z\Delta x}^{(K,k)} = \frac{1}{N_k T}\sum_{i\in\tilde{G}_k(K,\lambda_2)}\sum_{t=1}^{T} z_{it}(\Delta x_{it})'$, $\tilde{Q}_{z\Delta y}^{(K,k)} = \frac{1}{N_k T}\sum_{i\in\tilde{G}_k(K,\lambda_2)}\sum_{t=1}^{T} z_{it}\Delta y_{it}$, 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 IC:

$$IC_2(K, \lambda_2) = \ln\left[\tilde{\sigma}^2_{\tilde{G}(K,\lambda_2)}\right] + \rho_{2NT} pK, \tag{3.5}$$

where ρ_{2NT} is a tuning parameter. Let $\tilde{K}(\lambda_2) \equiv \arg\min_{1\le K\le K_{\max}} IC_2(K, \lambda_2)$. 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 $\tilde{G}_k(K, \lambda_2)$ being replaced by $G_{K,k}$.

To proceed, we add the following two assumptions, which parallel earlier Assumptions A4-A5.

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

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

The following theorem proves consistency of $\tilde{K}(\lambda_2)$.

Theorem 3.6 *Suppose Assumptions B1-B2 and B4-B6 hold. Then* $P(\tilde{K}(\lambda_2) = K_0) \to 1$ *as* $(N, T) \to \infty$.

4 Simulations

To evaluate the finite-sample performance of the classification and estimation procedure, we consider three data generating processes (DGPs) that cover linear and non-linear panels of static and dynamic models. 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 use sample sizes $N = 100, 200$ and time spans $T = 15, 25, 50$. Throughout these DGPs, the fixed effect μ_i^0 and the idiosyncratic error ε_{it} are standard normal, independent across i and t, and mutually independent. ε_{it} is also independent of all regressors.

DGP 1 (Linear static panel) The observations (y_{it}, x_{it}) are generated from the linear panel structure model as in Example 1. The exogenous regressor $x_{it} = (0.2\mu_i^0 + e_{it1}, 0.2\mu_i^0 + e_{it2})'$, where $e_{it1}, e_{it2} \sim$ IID $N(0, 1)$, are mutually independent, and independent of μ_i^0. The true coefficients are $(0.4, 1.6)$, $(1, 1)$, $(1.6, 0.4)$ for the three groups, respectively. PLS will be applied in this DGP.

DGP 2 (Linear panel AR(1)) The dependent variable is determined by its lag term and two exogenous regressors x_{it2} and x_{it3} in $y_{it} = \beta_{i1}^0 y_{i,t-1} + \beta_{i2}^0 x_{it2} + \beta_{i3}^0 x_{it3} + \mu_i^0(1 - \beta_{i1}^0) + \varepsilon_{it},$. The exogenous variables are standard normal and mutually independent. For each i, the initial value is specified to guarantee that the time series $(y_{i0}, \ldots, y_{iT})$ is strictly stationary. The true coefficients are $(0.4, 1.6, 1.6)$, $(0.6, 1, 1)$, and $(0.8, 0.4, 0.4)$. The AR(1) coefficients represent weak, moderate, and strong persistence, respectively. PGMM will be used to estimate the first-differenced model with the instruments $(y_{i,t-2}, y_{i,t-3}, \Delta x_{it2}, \Delta x_{it3})$. In the Supplementary Material, the same DGP is also estimated by PLS.

DGP 3 (Probit panel AR(1)) As in Example 3, the binary dependent variable $y_{it} = \mathbf{1}\{\beta_{i1}^0 y_{i,t-1} + \beta_{i2}^0 x_{it} + \beta_{i3}^0 + \mu_i^0 - \varepsilon_{it} > 0\}$. The exogenous regressor $x_{it} = 0.1\mu_i^0 + e_{it}$, where $e_{it} \sim$ IID $N(0, 1)$ and is independent of all other variables. The true coefficients are $(1, -1, 0.5)$, $(0.5, 0, -0.25)$, and $(0, 1, 0)$. β_{i1}^0 and β_{i2}^0 are identifiable in this model, whereas β_{i3}^0 is unidentifiable as it is absorbed into the individual heterogeneity. PPL will be implemented in this DGP.

Since both classification consistency and the oracle property hinge on the correct number of groups, our first simulation exercise is designed to assess how well the proposed IC selects the number of groups. Asymptotically, all sequences ρ_{1NT} work if they satisfy Assumption A5 or 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. We experimented with many alternatives, and found that $\rho_{jNT} = \frac{2}{3}(NT)^{-1/2}$, $j = 1, 2$, work fairly well in the linear models and so does $\rho_{1NT} = \frac{1}{4}(\ln \ln T)/T$ in the Probit model. They are used throughout the simulations as well as the empirical applications.

Regarding the C-Lasso tuning parameter, we specify $\lambda_j = c_{\lambda_j} s_Y^2 T^{-1/3}$ for $j = 1, 2$ in the linear models, where s_Y^2 is the sample variance of $\tilde{y}_{it}$ for PLS or the sample variance of Δy_{it} for PGMM, and $c_{\lambda_j} \in \{0.125, 0.25, 0.5, 1, 2\}$ $(j = 1,\ 2)$, a geometrically increasing sequence. In the Probit model, we also specify $\lambda_1 = c_{\lambda_1} s_Y^2 T^{-1/3}$, but the constant $c_{\lambda_1} \in \{0.0125, 0.025, 0.05, 0.1, 0.2\}$, as this criterion function is at a different magnitude from the linear models. Following Remark 6, we pick up from the set of candidate values the λ_1 that minimizes $IC_1(\hat{K}(\lambda_1), \lambda_1)$ and similarly the λ_2 that minimizes $IC_2(\tilde{K}(\lambda_2), \lambda_2)$. We run 500 replications for each DGP. Table 1 displays the empirical probability that a particular group size from 1 to 5 is selected according to the IC when the true number of groups is 3. In the linear models, the IC achieves almost perfect selection of the true group number when $T = 25$. In the Probit model, the correct determination rate is also close to 100% when $T = 50$, although the rate is lower than that of the linear models when $T = 25$, due to the presence of incidental parameters. These statistics demonstrate the usefulness of the IC.

Table 1: Frequency of selecting $K = 1, \ldots, 5$ groups when $K_0 = 3$

N	T	DGP 1					DGP 2					DGP 3				
		1	2	**3**	4	5	1	2	**3**	4	5	1	2	**3**	4	5
100	15	0	0	**0.994**	0.004	0.002	0	0.232	**0.762**	0.004	0.002					
100	25	0	0	**1**	0	0	0	0.016	**0.984**	0	0	0	0.096	**0.646**	0.242	0.016
100	50	0	0	**1**	0	0	0	0	**1**	0	0	0	0	**0.986**	0.014	0
200	15	0	0	**0.890**	0.106	0.004	0	0.022	**0.970**	0.008	0					
200	25	0	0	**1**	0	0	0	0	**1**	0	0	0	0.106	**0.668**	0.226	0
200	50	0	0	**1**	0	0	0	0	**1**	0	0	0	0	1	0	0

Next, given the true number of groups, we focus on the classification of individual units and the point estimation of post-Lasso.[4] Due to space limitation, all tabulated results are produced under $c_{\lambda_j} = 0.5$, $j = 1, 2$, for the linear models, and $c_{\lambda_1} = 0.05$ for the Probit model. The outcomes are found robust over the specified range of constants. Column 4 of Tables 2 shows the percentage of correct classification of the N units, calculated as $\frac{1}{N} \sum_{k=1}^{K_0} \sum_{i \in \hat{G}_k} \mathbf{1}\{\beta_i^0 = \alpha_k^0\}$, averaged over the Monte Carlo replications. Columns 5–7 summarize the post-Lasso estimator's root-mean-squared error (RMSE), bias, and the coverage probability of the two-sided nominal 95% confidence interval. To save space, we report the results for the first coefficient $\boldsymbol{\alpha}_1 = (\alpha_{k1})_{k=1}^{K_0}$ in each model. As $\boldsymbol{\alpha}_1$ is a $K_0 \times 1$ vector, we averaged over the statistics by their weight N_k/N, $k = 1, \ldots, K_0$. For example, in DGP 1 and 3 the coverage probability is computed as $\sum_{k=1}^{K_0} \frac{N_k}{N} \mathbf{1} \left\{ \hat{\alpha}_{\hat{G}_k,1} - 1.96\hat{\sigma}_{k1} \leq \alpha_{k1}^0 \leq \hat{\alpha}_{\hat{G}_k,1} + 1.96\hat{\sigma}_{k1} \right\}$, where $\hat{\sigma}_{k1}$ is the estimated standard deviation of $\hat{\alpha}_{\hat{G}_k,1}$; in DGP 2, $\tilde{\alpha}_{\tilde{G}_k,1}$ and $\tilde{\sigma}_{k1}$ replace their counterparts $\hat{\alpha}_{\hat{G}_k,1}$ and $\hat{\sigma}_{k1}$, respectively. For comparison purpose, Columns 8–10 show the corresponding statistics of the *oracle* estimator $\hat{\alpha}_{G_k^0,1}$ or $\tilde{\alpha}_{G_k^0,1}$. The only difference between the oracle estimator and the post-Lasso estimator is that the former utilizes the true group identity G_k^0, which is infeasible in practice, while the latter is based on the data-determined group $\hat{G}_k$ or $\tilde{G}_k$.

As expected, the correct classification percentage approaches 100% as T increases, and the oracle estimator's RMSE and bias are typically smaller than those of post-Lasso. When $T = 50$, post-Lasso and the oracle perform almost identically in DGP 1. In DGP 2, the PGMM confidence interval covers the true parameter slightly more often than the oracle, since the estimated standard deviation is inflated by a few misclassified units, which hide as outliers against the majority of the group members. The same reason explains the mild discrepancy of RMSE in DGPs 2 and 3. However, the confidence interval in DGP 3 under-covers due to the extra complexity of the bias caused by incidental parameters. Here the bias in post-Lasso is effectively reduced by the half-panel jackknife (Dhaene and Jochmans, 2015), but it cannot be completely eliminated in finite samples.

5 Empirical Applications

In this section we illustrate the use of C-Lasso in two cross-country studies. We explore the determinants of savings rates via a linear dynamic panel model and the relationship between civil war incidence and poverty

[4] Here we report the results for post-Lasso under K_0. In Table S3 of the Supplementary Material, we also compare the RMSE and bias of post-Lasso and C-Lasso under K_0 and the IC-determined group number $\hat{K}$ or $\tilde{K}$.

Table 2: Classification and Point Estimation of $\boldsymbol{\alpha}_1$

	N	T	% of correct classification	Post-Lasso RMSE	Post-Lasso Bias	Post-Lasso Coverage	Oracle RMSE	Oracle Bias	Oracle Coverage
DGP 1	100	15	0.8935	0.0594	0.0105	0.8758	0.0463	0.0012	0.9336
	100	25	0.9674	0.0384	0.0018	0.9344	0.0353	0.0001	0.9362
	100	50	0.9964	0.0249	0.0000	0.9528	0.0245	-0.0002	0.9348
	200	15	0.8987	0.0432	0.0077	0.8650	0.0324	-0.0013	0.9410
	200	25	0.9661	0.0272	0.0015	0.9228	0.0250	-0.0006	0.9394
	200	50	0.9966	0.0174	-0.0001	0.9496	0.0171	-0.0002	0.9424
DGP 2	100	15	0.8063	0.0711	-0.0123	0.9562	0.0502	-0.0037	0.9090
	100	25	0.8974	0.0461	-0.0060	0.9760	0.0351	0.0011	0.9336
	100	50	0.9689	0.0278	-0.0011	0.9860	0.0242	-0.0010	0.9320
	200	15	0.8151	0.0557	-0.0159	0.9436	0.0352	-0.0017	0.9308
	200	25	0.9037	0.0328	-0.0047	0.9664	0.0252	-0.0006	0.9442
	200	50	0.9711	0.0193	-0.0014	0.9842	0.0164	0.0000	0.9304
DGP 3	100	25	0.7941	0.1701	0.0805	0.7856	0.1077	0.0114	0.9376
	100	50	0.9456	0.0859	0.0231	0.8970	0.0752	0.0090	0.9504
	200	25	0.8277	0.1325	0.0777	0.7214	0.0821	0.0116	0.9104
	200	50	0.9527	0.0635	0.0223	0.8818	0.0573	0.0121	0.9280

via a dynamic Probit model. Due to space limitation, we only report the estimated coefficients in the main text. Summary statistics, group membership, and additional details of implementation can be found in the Supplementary Material.

5.1 Savings Rate Dynamic Panel Modeling and Classification

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

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

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

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 ε_{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). The first-order lagged savings rate 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. 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 56 countries. Substantial heterogeneity across countries was observed in all these major macroeconomic indicators. Evidence of within group homogeneity is therefore particularly important in supporting panel data pooling techniques.

This dynamic panel model can be estimated by either PLS or PGMM. We first try PLS, which has higher correct classification ratio in our simulation when $T = 15$. Following the simulation, ρ_{1NT} is set as $\frac{2}{3}(NT)^{-1/2}$, and the IC picks two groups and the tuning parameter constant $c_{\lambda_1} = 1.55$ over all combinations of $K = 1, \ldots, 5$ and c_{λ_1} in a geometrically increasing sequence of 10 points in $(0.2, \ldots, 2)$. Based on this choice of tuning parameter, the data determine the group identities. Interestingly, some geographic features remain salient in the classification. For example, we observe a strong collection of Asian countries in Group 1. In particular, except for South Korea and the city state Singapore, Group 1 includes all Eastern Asian and Southeastern Asian countries in our sample, namely, China, Japan, Indonesia, Malaysia, Philippines, and Thailand.

Table 3: PLS and PGMM estimation results

Variables		PLS			PGMM	
	Pooled FE	Group1	Group2	Pooled GMM	Group1	Group2
Lagged savings	0.7609***	0.6952***	0.6939***	0.5854	0.4026	0.6373**
	(0.0322)	(0.0433)	(0.0449)	(0.4588)	(0.3095)	(0.3197)
Inflation	−0.0145	−0.1601***	0.1967***	0.0350	−0.1647**	0.4128***
	(0.0324)	(0.0388)	(0.0435)	(0.0621)	(0.0733)	(0.0758)
Interest rate	−0.0346	−0.1490***	0.1226***	−0.0333	−0.1580**	0.1395*
	(0.0313)	(0.0397)	(0.0408)	(0.0598)	(0.0729)	(0.0775)
GDP growth	0.2027***	0.2892***	0.1127**	0.2081***	0.1853***	0.2061**
	(0.0353)	(0.0413)	(0.0517)	(0.0541)	(0.0627)	(0.0908)

Note: *** 1% significant, ** 5% significant, * 10% significant

Columns 3–4 in Table 3 report the results for the PLS-based post-Lasso estimation, in comparison with those for the pooled FE estimation in Column 2. The estimates are bias-corrected by the half-panel jackknife (Dhaene and Jochmans, 2015), and the standard errors (in parentheses) are clustered at the country level. 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 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 at the 5% level, which suggests that conventional wisdom is universally relevant and applies both within and across groups.

To check the robustness of PLS, we compare it with PGMM on the same data. The IC with $\rho_{2NT} = \frac{2}{3}(NT)^{-1/2}$ is minimized at $K = 2$ and $c_{\lambda_2} = 0.72$ among the same combinations of K and c_{λ_2} for PLS. The estimated group identities reveal 84% overlap with the PLS classified membership, and the coefficients in columns 6–7 of Table 3 are comparable to those from PLS.

5.2 Dynamic Probit Panel Modeling of Civil War Conflict

According to a conservative estimate, direct casualties from civil conflicts were at least 16.2 million in the second half of the twentieth century, a figure five times as large as the inter-state toll (Fearon and Laitin, 2003). Civil war damage to national development has attracted interest among economists and political scientists, looking at both causes and consequences, and leading to an explosion of research output (Miguel, Satyanath, and Sergenti, 2004; Besley and Persson, 2010; Nunn and Qian, 2014). A comprehensive overview was given in Blattman and Miguel (2010).

This section revisits the connection between civil wars and poverty, a topic of enduring research interest.[5] Cross-country empirical work mostly follows Fearon and Laitin (2003) and Collier and Hoeffler (2004) in regressing war onset or incidence against posited causes of civil conflict. Country-specific heterogeneity is handled either by control variables or fixed effects. In view of the measurement error in many macro variables and the difficulty in exhausting all relevant factors, Djankov and Reynal-Querol (2010) explored the fixed effect approach in linear regressions. Group-specific heterogeneity was also investigated after identifying groups using observed information relating to former colonial families (Djankov and Reynal-Querol, 2010) or continental regions (Esteban, Mayoral, and Ray, 2012). Without such information, PPL can deal with unobservable country- and group-specific heterogeneity simultaneously in a nonlinear model.

We use the replication data in Fearon and Laitin (2003). Since most of the variables in the dataset are time-invariant country characteristics, we collect *civil war incidence*, *GDP per capita* and *log population* to generate a balanced panel of 38 countries and 39 years spanning from 1960 to 1998.[6] We specify a panel AR(1) Probit model as in DGP 3 to capture the high persistence of civil war incidence, and we transform *GDP per capita* and *log population* into growth rates to avoid non-stationarity.

The IC with $\rho_{1NT} = \frac{1}{4}(\ln\ln T)/T$ selects two groups and the tuning parameter constant $c_{\lambda_1} = 0.046$ from all the combinations of $K = 1, \ldots, 5$, and c_{λ_1} from 10 points in the geometrically increasing sequence $(0.01, \ldots, 0.1)$. C-Lasso classifies 23 countries into a "high-occurrence" group (with mean civil war incidence 0.4302), and the other 15 countries into a "low-occurrence" group (with mean incidence 0.2263). In terms of geographic features, Iran and Jordan are separated from all the other 12 Asian countries, most of which are plagued by civil wars; the four included European countries (Cyprus, Russia, UK, Yugoslavia) all fall into the low-occurrence group.

Table 4 displays the estimated PPL coefficients along with those for standard Probit and FE Probit regressions. Again, the estimates are bias-corrected by the half-panel jackknife, and the standard errors are clustered at the country level. Obviously, civil war incidence is highly persistent, and its association with GDP per capita growth remains robust in Probit and FE Probit regressions. However, the effects are distinguished in the two groups: the negative coefficient is statistically significant in the low-occurrence

[5] It was taken as a stylized fact in pooled regressions that "civil wars are more likely to occur in countries that are poor" (Blattman and Miguel, 2010), but Djankov and Reynal-Querol (2010) found "the statistical association between poverty and civil wars disappears once we include country fixed effects [into a linear panel model]."

[6] Between 1960–1998, there are 102 countries with data on all the three variables available. However, 61 of them had no civil war in the sample period and 3 had on-going civil wars throughout the time. These countries are dropped from the data, since they are associated with infinite ($-\infty$ or ∞) fixed effects in a FE Probit model.

group, but no such relationship is found in the high-occurrence group.

Table 4: Probit, FE Probit and PPL estimation results

Variables	Probit		FE Probit		Post-Lasso PPL			
					high-occurrence		low-occurrence	
	coef.	s.e.	coef.	s.e.	coef.	s.e.	coef.	s.e.
Lagged civil war	3.1955^{***}	0.1156	3.2649^{***}	0.1140	3.3012^{***}	0.1363	2.9630^{***}	0.2707
GDP per capita growth	-0.4359^{***}	0.1155	-0.3854^{***}	0.1389	0.1591	0.1193	-1.2072^{***}	0.2220
population growth	-0.0125	0.1107	0.0162	0.1284	-0.0448	0.1429	0.2811	0.1736

Note: *** 1% significant, ** 5% significant, * 10% significant

6 Conclusion

We propose a novel and systematic approach to identify and estimate latent group structures in panel data, developing panel penalized profile likelihood (PPL) and panel GMM (PGMM) methods for classification and estimation, and providing asymptotic properties for use in inference. The PPL method enjoys the oracle property but PGMM 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 general approach to classifying and estimating panel models with unknown homogeneous groups, heterogeneity across groups, and an unknown number of groups. Simulations show that the approach has good finite sample performance and can be readily implemented in practical work. Two applications reveal the advantages of data-determined identification of latent group structures in empirical panel modeling.

The present work raises interesting issues for further research. First, it may be appealing to consider a more general framework that allows the number (K_0) of groups to grow with the sample size. Close examination of the theory provided in this paper suggests that it is possible to permit K_0 to increases with N but at a very slow rate. Second, both the linear and nonlinear models may be extended to include time effects or interactive fixed effects (IFE). In linear models with IFE but without endogeneity, we remark that the present approach can be used in conjunction with principal component analysis to address cross sectional dependence modeled through IFE. Extension to nonlinear models or to models with endogeneity will raise new statistical and computational challenges. Third, our method can be extended to nonstationary panels where panel unit and cointegrating relationships may possess latent group structures. Some of these topics will be explored in future work.

APPENDIX

A Proofs of the Results in Section 2

Proof of Theorem 2.1. (i) Let $Q_{1NT,i}(\beta_i) = \frac{1}{T}\sum_{t=1}^{T}\psi(w_{it};\beta_i,\hat{\mu}_i(\beta_i))$ 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 $\hat{b}_i=\hat{\beta}_i-\beta_i^0$. Since $\hat{\mu}_i(\beta_i)=\arg\min_{\mu_i}\frac{1}{T}\sum_{t=1}^{T}\psi(w_{it};\beta_i,\mu_i)$, we have $\frac{1}{T}\sum_{t=1}^{T}V_i(w_{it};\beta_i,\hat{\mu}_i(\beta_i))=0\ \forall\ \beta_i$. Then by second order Taylor expansion and the envelope theorem, we have

$$\begin{aligned} Q_{1NT,i}(\hat{\beta}_i)-Q_{1NT,i}\left(\beta_i^0\right) &= \frac{1}{T}\sum_{t=1}^{T}\psi\left(w_{it};\hat{\beta}_i,\hat{\mu}_i(\hat{\beta}_i)\right)-\frac{1}{T}\sum_{t=1}^{T}\psi\left(w_{it};\beta_i^0,\hat{\mu}_i\left(\beta_i^0\right)\right)\\ &= \hat{b}_i'\hat{S}_i+\frac{1}{2}\hat{b}_i'\hat{H}_{i\beta\beta}(\breve{\beta}_i)\hat{b}_i, \end{aligned} \tag{A.1}$$

where $\breve{\beta}_i$ lies between $\hat{\beta}_i$ and β_i^0 elementwise, $\hat{S}_i=\frac{1}{T}\sum_{t=1}^{T}U_i\left(w_{it};\beta_i^0,\hat{\mu}_i\left(\beta_i^0\right)\right)$, and

$$\hat{H}_{i\beta\beta}(\beta_i)=\frac{1}{T}\sum_{t=1}^{T}\left[U_i^{\beta_i}(w_{it};\beta_i,\hat{\mu}_i(\beta_i))+U_i^{\mu_i}(w_{it};\beta_i,\hat{\mu}_i(\beta_i))\frac{\partial\hat{\mu}_i(\beta_i)}{\partial\beta_i'}\right]. \tag{A.2}$$

By Lemmas S1.6 and S1.10 in the Supplement, $\hat{S}_i=O_P\left(T^{-1/2}\right)$, $\frac{1}{N}\sum_{i=1}^{N}||\hat{S}_i||^2=O_P\left(T^{-1}\right)$, and $c_{\hat{H}}\equiv \min_{1\le i\le N}\mu_{\min}(\hat{H}_{i\beta\beta}(\breve{\beta}_i))\ge c_H-o_P(1)$. By the triangle and reverse triangle inequalities,

$$\begin{aligned} \left|\Pi_{k=1}^{K_0}\left\|\hat{\beta}_i-\alpha_k\right\|-\Pi_{k=1}^{K_0}\left\|\beta_i^0-\alpha_k\right\|\right| &\le \left|\Pi_{k=1}^{K_0-1}\left\|\hat{\beta}_i-\alpha_k\right\|\left\{\left\|\hat{\beta}_i-\alpha_{K_0}\right\|-\left\|\beta_i^0-\alpha_{K_0}\right\|\right\}\right|\\ &\quad+\left|\Pi_{k=1}^{K_0-2}\left\|\hat{\beta}_i-\alpha_k\right\|\left\|\beta_i^0-\alpha_{K_0}\right\|\left\{\left\|\hat{\beta}_i-\alpha_{K_0-1}\right\|-\left\|\beta_i^0-\alpha_{K_0-1}\right\|\right\}\right|\\ &\quad+\cdots\\ &\quad+\left|\Pi_{k=2}^{K_0}\left\|\beta_i^0-\alpha_k\right\|\left\{\left\|\hat{\beta}_i-\alpha_1\right\|-\left\|\beta_i^0-\alpha_1\right\|\right\}\right|\\ &\le \hat{c}_{iNT}(\boldsymbol{\alpha})\left\|\hat{\beta}_i-\beta_i^0\right\|, \end{aligned} \tag{A.3}$$

where $\hat{c}_{iNT}(\boldsymbol{\alpha})=\Pi_{k=1}^{K_0-1}||\hat{\beta}_i-\alpha_k||+\Pi_{k=1}^{K_0-2}||\hat{\beta}_i-\alpha_k||\left\|\beta_i^0-\alpha_{K_0}\right\|+...+\Pi_{k=2}^{K_0}\left\|\beta_i^0-\alpha_k\right\|=O_P(1)$. By (A.1), (A.3) and the fact that $Q_{1iNT,\lambda_1}^{(K_0)}(\hat{\beta}_i,\hat{\boldsymbol{\alpha}})-Q_{1iNT,\lambda_1}^{(K_0)}\left(\beta_i^0,\hat{\boldsymbol{\alpha}}\right)\le 0$, we have $c_{\hat{H}}||\hat{b}_i||^2\le 2(||\hat{S}_i||+\hat{c}_{iNT}(\hat{\boldsymbol{\alpha}})\lambda_1)||\hat{b}_i||$. Then, by Assumptions A1(v)

$$\left\|\hat{b}_i\right\|\le 2c_{\hat{H}}^{-1}\left(\left\|\hat{S}_i\right\|+\hat{c}_{iNT}(\hat{\boldsymbol{\alpha}})\lambda_1\right)=O_P\left(T^{-1/2}+\lambda_1\right). \tag{A.4}$$

(ii) 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

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

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

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

Noting that $c_{\hat{H}}=c_H-o_P(1)$ and $\frac{T}{N}\sum_{i=1}^N||\hat{S}_i||^2=O_P(1)$, 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 $N^{-1}\sum_{i=1}^N||\hat{b}_i||^2=O_P\left(T^{-1}\right)$.

(iii) Let $P_{NT}\left(\boldsymbol{\beta},\boldsymbol{\alpha}\right)=\frac{1}{N}\sum_{i=1}^N\Pi_{k=1}^{K_0}\left\|\beta_i-\alpha_k\right\|$. By the Minkowski inequality and the result in (i), as $(N,T)\to\infty$,

$$\begin{aligned}\hat{c}_{iNT}\left(\boldsymbol{\alpha}\right) \leq\; & \Pi_{k=1}^{K_0-1}\left\{\left\|\hat{\beta}_i-\beta_i^0\right\|+\left\|\beta_i^0-\alpha_k\right\|\right\}+\Pi_{k=1}^{K_0-2}\left\{\left\|\hat{\beta}_i-\beta_i^0\right\|+\left\|\beta_i^0-\alpha_k\right\|\right\}\left\|\beta_i^0-\alpha_{K_0}\right\| \\ & +\cdots+\Pi_{k=2}^{K_0}\left\|\beta_i^0-\alpha_k\right\| \\ =\; & \sum_{s=0}^{K_0-1}\left\|\hat{\beta}_i-\beta_i^0\right\|^s\Pi_{k=1}^s a_{ks}\left\|\beta_i^0-\alpha_k\right\|^{K_0-1-s} \\ \leq\; & C_{K_0}\left(\boldsymbol{\alpha}\right)\sum_{s=0}^{K_0-1}\left\|\hat{\beta}_i-\beta_i^0\right\|^s\leq C_{K_0}\left(\boldsymbol{\alpha}\right)\left(1+2\left\|\hat{\beta}_i-\beta_i^0\right\|\right), \qquad \text{(A.6)}\end{aligned}$$

where the a_{ks} are finite integers and $C_{K_0}\left(\boldsymbol{\alpha}\right)\equiv\max_{1\leq l\leq K_0}\max_{1\leq s\leq K_0-1}\Pi_{k=1}^s a_{ks}\left\|\alpha_l^0-\alpha_k\right\|^{K_0-1-s}=O\left(1\right)$ as K_0 is finite. By (A.3) and (A.6), as $(N,T)\to\infty$,

$$\begin{aligned}\left|P_{NT}(\hat{\boldsymbol{\beta}},\boldsymbol{\alpha})-P_{NT}\left(\boldsymbol{\beta}^0,\boldsymbol{\alpha}\right)\right| &\leq C_{K_0}\left(\boldsymbol{\alpha}\right)\frac{1}{N}\sum_{i=1}^N\left\|\hat{b}_i\right\|+2C_{K_0}\left(\boldsymbol{\alpha}\right)\frac{1}{N}\sum_{i=1}^N\left\|\hat{b}_i\right\|^2 \\ &\leq C_{K_0}\left(\boldsymbol{\alpha}\right)\left\{\frac{1}{N}\sum_{i=1}^N\left\|\hat{b}_i\right\|^2\right\}^{1/2}+O_P(T^{-1})=O_P(T^{-1/2}). \qquad \text{(A.7)}\end{aligned}$$

By (A.7), 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

$$\begin{aligned}0 \geq\; & P_{NT}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\alpha}})-P_{NT}(\hat{\boldsymbol{\beta}},\boldsymbol{\alpha}^0)=P_{NT}\left(\boldsymbol{\beta}^0,\hat{\boldsymbol{\alpha}}\right)-P_{NT}\left(\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)+O_P(T^{-1/2}) \\ =\; & \frac{1}{N}\sum_{i=1}^N\Pi_{k=1}^{K_0}\left\|\beta_i^0-\hat{\alpha}_k\right\|+O_P(T^{-1/2}) \\ =\; & \frac{N_1}{N}\Pi_{k=1}^{K_0}\left\|\hat{\alpha}_k-\alpha_1^0\right\|+\cdots+\frac{N_{K_0}}{N}\Pi_{k=1}^{K_0}\left\|\hat{\alpha}_k-\alpha_{K_0}^0\right\|+O_P(T^{-1/2}). \qquad \text{(A.8)}\end{aligned}$$

By Assumption A1(vii), $N_k/N\to\tau_k\in(0,1)$ for each $k=1,...,K_0$. So (A.8) 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) First, we fix $k\in\{1,...,K_0\}$. By the consistency of $\hat{\alpha}_k$ and $\hat{\beta}_i$ in Theorem 2.1 and Assumptions A1(vi)-(vii), $\hat{\beta}_i-\hat{\alpha}_l\overset{P}{\to}\alpha_k^0-\alpha_l^0\neq 0$ for all $i\in G_k^0$ and $l\neq k$ and $\hat{c}_{ki}\equiv\Pi_{l=1,l\neq k}^{K_0}||\hat{\beta}_i-\hat{\alpha}_l||\overset{P}{\to}c_k^0\equiv\Pi_{l=1,l\neq k}^{K_0}\left\|\alpha_k^0-\alpha_l^0\right\|\geq c_\alpha^{K_0-1}>0$ for $i\in G_k^0$. Now, suppose that $||\hat{\beta}_i-\hat{\alpha}_k||\neq 0$ for some $i\in G_k^0$. By the envelope theorem, the first order condition (with respect to β_i) for the minimization problem in (2.5) yields that

$$\begin{aligned}
\mathbf{0}_{p\times 1} &= \frac{1}{\sqrt{T}}\sum_{t=1}^{T} U_i(w_{it};\hat{\beta}_i,\hat{\mu}_i(\hat{\beta}_i)) + \sqrt{T}\lambda_1 \sum_{j=1}^{K_0} \hat{e}_{ij}\Pi_{l=1,l\neq j}^{K_0}\left\|\hat{\beta}_i-\hat{\alpha}_l\right\| \\
&= \frac{1}{\sqrt{T}}\sum_{t=1}^{T} U_{it} + \left(\frac{\lambda_1\hat{c}_{ki}}{\left\|\hat{\beta}_i-\hat{\alpha}_k\right\|}I_p + \bar{H}_{i\beta\beta}\right)\sqrt{T}\left(\hat{\beta}_i-\hat{\alpha}_k\right) + \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left[U_i\left(w_{it};\beta_i^0,\hat{\mu}_i\left(\beta_i^0\right)\right)-U_{it}\right] \\
&\quad + \bar{H}_{i\beta\beta}\sqrt{T}\left(\hat{\alpha}_k-\alpha_k^0\right) + \sqrt{T}\lambda_1 \sum_{j=1,j\neq k}^{K_0} \hat{e}_{ij}\Pi_{l=1,l\neq j}^{K_0}\left\|\hat{\beta}_i-\hat{\alpha}_l\right\| \\
&\equiv \hat{B}_{i1}+\hat{B}_{i2}+\hat{B}_{i3}+\hat{B}_{i4}+\hat{B}_{i5}, \qquad\qquad (A.9)
\end{aligned}$$

where $\hat{e}_{ij}=\frac{\hat{\beta}_i-\hat{\alpha}_j}{\left\|\hat{\beta}_i-\hat{\alpha}_j\right\|}$ if $||\hat{\beta}_i-\hat{\alpha}_j||\neq 0$ and $\|\hat{e}_{ij}\|\leq 1$ otherwise, the second equality follows from the first order Taylor expansion and rearrangement of terms, $\bar{H}_{i\beta\beta}\equiv\hat{H}_{i\beta\beta}\left(\bar{\beta}_i\right)$, $\hat{H}_{i\beta\beta}\left(\cdot\right)$ is defined in (A.2), $\bar{\beta}_i$ lies between $\hat{\beta}_i$ and β_i^0 elementwise.

Let $\varkappa_{1NT}=(T^{-1/2}\left(\ln T\right)^3+\lambda_1)\left(\ln T\right)^{\nu}$. Let C denote a generic constant that may vary across lines. By (A.4) and Lemmas S1.6-S1.7 in the Supplement, we can readily show that

$$P\left(\max_i\left\|\hat{\beta}_i-\beta_i^0\right\|\geq C\varkappa_{1NT}\right)=o\left(N^{-1}\right) \text{ for some } C>0, \qquad (A.10)$$

which in conjunction with the proof of Theorem 2.1(iii), implies that

$$P\left(\sqrt{T}\left\|\hat{\alpha}_k-\alpha_k^0\right\|\geq C\left(\ln T\right)^{\nu}\right)=o\left(N^{-1}\right) \text{ and } P\left(\max_{i\in G_k^0}\left|\hat{c}_{ki}-c_k^0\right|\geq c_k^0/2\right)=o\left(N^{-1}\right). \qquad (A.11)$$

By (A.10)-(A.11), $P\left(\max_{i\in G_k^0}\left\|\hat{B}_{i5}\right\|\geq C\sqrt{T}\lambda_1\varkappa_{1NT}\right)=o\left(N^{-1}\right)$. Combining these results with those in Lemmas S1.6(v) and S1.11(i), we have $P\left(\Xi_{kNT}\right)=1-o\left(N^{-1}\right)$, where

$$\begin{aligned}
\Xi_{kNT} \equiv\ & \left\{\max_{i\in G_k^0}\left|\hat{c}_{ki}-c_k^0\right|\leq c_k^0/2\right\} \cap \left\{\max_{i\in G_k^0}\left\|\bar{H}_{i\beta\beta}-H_{i\beta\beta}\left(\beta_i^0\right)\right\|\leq c_H/2\right\} \cap \left\{\max_{i\in G_k^0}\left\|\hat{B}_{i3}\right\|\leq C\left(\ln T\right)^{3+\nu}\right\} \\
& \cap\left\{\max_{i\in G_k^0}\left\|\hat{B}_{i4}\right\|\leq C\left(\ln T\right)^{\nu}\right\}\cap\left\{\max_{i\in G_k^0}\left\|\hat{B}_{i5}\right\|\leq C\sqrt{T}\lambda_1\varkappa_{1NT}\right\}.
\end{aligned}$$

Then conditional on Ξ_{kNT}, we have uniformly in $i\in G_k^0$

$$\begin{aligned}
\left\|\left(\hat{\beta}_i-\hat{\alpha}_k\right)'\left(\hat{B}_{i2}+\hat{B}_{i3}+\hat{B}_{i4}+\hat{B}_{i5}\right)\right\| &\geq \left\|\left(\hat{\beta}_i-\hat{\alpha}_k\right)'\hat{B}_{i2}\right\|-\left\|\left(\hat{\beta}_i-\hat{\alpha}_k\right)'\left(\hat{B}_{i3}+\hat{B}_{i4}+\hat{B}_{i5}\right)\right\| \\
&\geq \sqrt{T}\lambda_1\hat{c}_{ki}\left\|\hat{\beta}_i-\hat{\alpha}_k\right\|-C\left\|\hat{\beta}_i-\hat{\alpha}_k\right\|\left[2\left(\ln T\right)^{3+\nu}+\sqrt{T}\lambda_1\varkappa_{1NT}\right] \\
&\geq \sqrt{T}\lambda_1 c_k^0\left\|\hat{\beta}_i-\hat{\alpha}_k\right\|/4 \text{ for sufficiently large } (N,T),
\end{aligned}$$

where the last inequality follows because $\sqrt{T}\lambda_1\gg 2\left(\ln T\right)^{3+\nu}+\sqrt{T}\lambda_1\varkappa_{1NT}$ by Assumption A2(i). Then for all $i\in G_k^0$ we have

$$\begin{aligned}
P(\hat{E}_{kNT,i}) &= P\left(i\notin\hat{G}_k \mid i\in G_k^0\right)=P\left(-\hat{B}_{i1}=\hat{B}_{i2}+\hat{B}_{i3}+\hat{B}_{i4}+\hat{B}_{i5}\right) \\
&\leq P\left(\left|\left(\hat{\beta}_i-\hat{\alpha}_k\right)'\hat{B}_{i1}\right|\geq\left|\left(\hat{\beta}_i-\hat{\alpha}_k\right)'\left(\hat{B}_{i2}+\hat{B}_{i3}+\hat{B}_{i4}+\hat{B}_{i5}\right)\right|\right) \\
&\leq P\left(\left\|\hat{B}_{i1}\right\|\geq\sqrt{T}\lambda_1 c_k^0/4,\ \Xi_{kNT}\right)+P\left(\Xi_{kNT}^c\right)\to 0 \text{ as } (N,T)\to\infty,
\end{aligned}$$

where Ξ_{kNT}^c denotes the complement of Ξ_{kNT} and the convergence follows by Lemma S1.6(iv) and Assumption A2 . Consequently, we conclude that with probability $1-o\left(N^{-1}\right)$ the difference $\hat{\beta}_i-\hat{\alpha}_k$ 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 $P\left(\left\|\hat{\beta}_i-\hat{\alpha}_k\right\|=0 \mid i\in G_k^0\right)=1-o\left(N^{-1}\right)$.

For uniform consistency, we have: $P(\cup_{k=1}^{K_0}\hat{E}_{kNT})\leq\sum_{k=1}^{K_0}P(\hat{E}_{kNT})\leq\sum_{k=1}^{K_0}\sum_{i\in G_k^0}P(\hat{E}_{kNT,i})$ and by Lemma S1.6(iv)

$$\begin{aligned}\sum_{k=1}^{K_0}\sum_{i\in G_k^0}P\left(\hat{E}_{kNT,i}\right) &\leq \sum_{k=1}^{K_0}\sum_{i\in G_k^0}\left[P\left(\left\|\hat{B}_{i1}\right\|\geq\sqrt{T}\lambda_1 c_k^0/4,\ \Xi_{kNT}\right)+P\left(\Xi_{kNT}^c\right)\right]\\ &\leq N\max_{1\leq i\leq N}P\left(\left\|\frac{1}{T}\sum_{t=1}^{T}U_{it}\right\|\geq\lambda_1 c_\alpha^{K_0-1}/4\right)+o\left(1\right)=o\left(1\right). \quad\quad (A.12)\end{aligned}$$

This completes the proof of (i).

(ii) 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,

$$\begin{aligned}P(\hat{F}_{kNT,i}) &= 1-P(i\in G_k^0\mid i\in\hat{G}_k)\\ &= \frac{\sum_{l=1,l\neq k}^{K_0}P(i\in\hat{G}_k|i\in G_l^0)P\left(i\in G_l^0\right)}{P(i\in\hat{G}_k|i\in G_k^0)P\left(i\in G_k^0\right)+\sum_{l=1,l\neq k}^{K_0}P(i\in\hat{G}_k|i\in G_l^0)P\left(i\in G_l^0\right)}. \quad\quad (A.13)\end{aligned}$$

For the numerator, we have by (A.12)

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

In addition, noting that $P(i\in\hat{G}_k|i\in G_k^0)=1-P(i\notin\hat{G}_k|i\in G_k^0)=1-o\left(1\right)$ 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

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

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 the 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.12), $\sum_{i=1}^{N} P(\hat{E}_{kNT,i}) = \sum_{k=1}^{K_0} \sum_{i\in G_k^0} P(\hat{E}_{kNT,i}) = o(1)$. 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 Karush–Kuhn–Tucker (KKT) conditions for $\{\hat{\beta}_i\}$ and $\{\hat{\alpha}_k\}$ to minimize the objective function in (2.5) are 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 by the envelope theorem,

$$\mathbf{0}_{p\times 1} = \frac{1}{T}\sum_{t=1}^{T} U_i\left(w_{it}; \hat{\beta}_i, \hat{\mu}_i(\hat{\beta}_i)\right) + \lambda_1 \sum_{j=1}^{K_0} \hat{e}_{ij} \Pi_{l=1,l\neq j}^{K_0} \left\|\hat{\beta}_i - \hat{\alpha}_l\right\|, \tag{A.14}$$

$$\mathbf{0}_{p\times 1} = \frac{\lambda_1}{N}\sum_{i=1}^{N} \hat{e}_{ik} \Pi_{l=1,l\neq k}^{K_0} \left\|\hat{\beta}_i - \hat{\alpha}_l\right\|, \tag{A.15}$$

where $\hat{e}_{ij}$ is defined after (A.9). 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 $\hat{e}_{ij} = \frac{\hat{\beta}_i - \hat{\alpha}_j}{\|\hat{\beta}_i - \hat{\alpha}_j\|} = \frac{\hat{\alpha}_k - \hat{\alpha}_j}{\|\hat{\alpha}_k - \hat{\alpha}_j\|}$ for any $i \in \hat{G}_k$ and $j \neq k$. Then by the fact that $\hat{\beta}_i = \hat{\alpha}_k$ $\forall i \in \hat{G}_k$ and (A.15),

$$\sum_{i\in\hat{G}_k} \sum_{j=1,j\neq k}^{K_0} \hat{e}_{ij} \Pi_{l=1,l\neq j}^{K_0} \left\|\hat{\beta}_i - \hat{\alpha}_l\right\| = \sum_{i\in\hat{G}_k} \sum_{j=1,j\neq k}^{K_0} \frac{\hat{\alpha}_k - \hat{\alpha}_j}{\|\hat{\alpha}_k - \hat{\alpha}_j\|} \Pi_{l=1,l\neq j}^{K_0} \|\hat{\alpha}_k - \hat{\alpha}_l\| = \mathbf{0}_{p\times 1} \tag{A.16}$$

and

$$\begin{aligned}
\mathbf{0}_{p\times 1} &= \sum_{i=1}^{N} \hat{e}_{ik} \Pi_{l=1,l\neq k}^{K_0} \left\|\hat{\beta}_i - \hat{\alpha}_l\right\| \\
&= \sum_{i\in\hat{G}_k} \hat{e}_{ik} \Pi_{l=1,l\neq k}^{K_0} \|\hat{\alpha}_k - \hat{\alpha}_l\| + \sum_{i\in\hat{G}_0} \hat{e}_{ik} \Pi_{l=1,l\neq k}^{K_0} \left\|\hat{\beta}_i - \hat{\alpha}_l\right\| + \sum_{j=1,j\neq k}^{K_0} \sum_{i\in\hat{G}_j} \hat{e}_{ik} \Pi_{l=1,l\neq k}^{K_0} \|\hat{\alpha}_j - \hat{\alpha}_l\| \\
&= \sum_{i\in\hat{G}_k} \hat{e}_{ik} \Pi_{l=1,l\neq k}^{K_0} \|\hat{\alpha}_k - \hat{\alpha}_l\| + \sum_{i\in\hat{G}_0} \hat{e}_{ik} \Pi_{l=1,l\neq k}^{K_0} \left\|\hat{\beta}_i - \hat{\alpha}_l\right\|,
\end{aligned} \tag{A.17}$$

where the last equality follows from the fact that $\sum_{j=1,j\neq k}^{K_0} \sum_{i\in\hat{G}_j} \hat{e}_{ik} \Pi_{l=1,l\neq k}^{K_0} \|\hat{\alpha}_j - \hat{\alpha}_l\| = \sum_{j=1,j\neq k}^{K_0} \sum_{i\in\hat{G}_j} \frac{\hat{\alpha}_j - \hat{\alpha}_k}{\|\hat{\alpha}_j - \hat{\alpha}_k\|} \Pi_{l=1,l\neq k}^{K_0} \|\hat{\alpha}_j - \hat{\alpha}_l\| = 0$. Then averaging both sides of (A.14) over $i \in \hat{G}_k$ and using (A.16)-(A.17), we have

$$\mathbf{0}_{p\times 1} = \frac{1}{N_k T} \sum_{i\in\hat{G}_k} \sum_{t=1}^{T} U_i\left(w_{it}; \hat{\alpha}_k, \hat{\mu}_i(\hat{\alpha}_k)\right) + \frac{\lambda_1}{N_k} \sum_{i\in\hat{G}_0} \hat{e}_{ik} \Pi_{l=1,l\neq k}^{K_0} \left\|\hat{\beta}_i - \hat{\alpha}_l\right\|. \tag{A.18}$$

Using $U_i\left(w_{it}; \hat{\alpha}_k, \hat{\mu}_i(\hat{\alpha}_k)\right) = U_i\left(w_{it}; \alpha_k^0, \hat{\mu}_i(\alpha_k^0)\right) + \hat{H}_{(k)}(\hat{\alpha}_k - \alpha_k^0)$ with $\hat{H}_{(k)} \equiv \frac{1}{N_k T} \sum_{i\in\hat{G}_k} \sum_{t=1}^{T} [U_i^{\beta_i}(w_{it}; \breve{\alpha}_k,$ $\hat{\mu}_i(\breve{\alpha}_k)) + U_i^{\mu_i}(w_{it}; \breve{\alpha}_k^0, \hat{\mu}_i(\breve{\alpha}_k)) \frac{\partial \hat{\mu}_i(\breve{\alpha}_k)}{\partial \alpha_k'}]$ and $\breve{\alpha}_k$ lying between $\hat{\alpha}_k$ and α_k^0 elementwise, and re-arranging terms in (A.18) yield

$$\hat{\alpha}_k - \alpha_k^0 = -\hat{H}_{(k)}^{-1} \frac{1}{N_k T} \sum_{i\in\hat{G}_k} \sum_{t=1}^{T} U_i\left(w_{it}; \alpha_k^0, \hat{\mu}_i(\alpha_k^0)\right) + \hat{\mathcal{R}}_k,$$

where $\hat{\mathcal{R}}_k = \hat{H}_{(k)}^{-1} \frac{\lambda_1}{N} \sum_{i \in \hat{G}_0} \hat{e}_{ik} \Pi_{l=1, l \neq k}^{K_0} \left\| \hat{\beta}_i - \hat{\alpha}_l \right\|$. 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 $\epsilon > 0$

$$P\left(\sqrt{NT} \left\| \hat{\mathcal{R}}_k \right\| \geq \epsilon\right) \leq \sum_{k=1}^{K_0} \sum_{i \in G_k^0} P\left(i \in \hat{G}_0 | i \in G_k^0\right) \leq \sum_{k=1}^{K_0} \sum_{i \in G_k^0} P\left(i \notin \hat{G}_k | i \in G_k^0\right) = o(1) \text{ by (A.6).}$$

So $\left\| \hat{\mathcal{R}}_k \right\| = o_P((NT)^{-1/2})$. By Lemmas S1.11-S1.12, $\frac{1}{\sqrt{N_k T}} \sum_{i \in \hat{G}_k} \sum_{t=1}^{T} U_i \left(w_{it}; \alpha_k^0, \hat{\mu}_i(\alpha_k^0)\right) + \mathbb{B}_{kNT} \xrightarrow{D} N\left(0, \Omega_k\right)$ and $\hat{H}_{(k)} = \mathbb{H}_{kNT} + o_P\left(\nu_{NT}\right)$ where $\nu_{NT} = \min(1, \sqrt{T/N_k})$. These results in conjunction with the fact that $\mathbb{B}_{kNT} = O_P(\sqrt{N_k/T})$ and Assumption A3(ii) imply that $\sqrt{N_k T}\left(\hat{\alpha}_k - \alpha_k^0\right) - \mathbb{H}_{kNT}^{-1} \mathbb{B}_{kNT} \xrightarrow{D} N\left(0, \mathbb{H}_k^{-1} \Omega_k (\mathbb{H}_k^{-1})'\right)$. ■

Proof of Theorem 2.5. For the post-Lasso estimator, we have the following first order conditions: $\frac{1}{N_k T} \sum_{i \in \hat{G}_k} \sum_{t=1}^{T} U_i(w_{it}; \hat{\alpha}_{\hat{G}_k}, \hat{\mu}_i(\hat{\alpha}_{\hat{G}_k})) = \mathbf{0}_{p \times 1}$. Following the analyses of $\hat{\mu}_i(\beta_i)$, $\hat{\beta}_i$ and $\partial \hat{\mu}_i(\beta_i)/\partial \beta_i$ in Lemmas S1.5, S1.7, and S1.9, we can readily establish the consistency of $\hat{\mu}_i(\alpha_k)$, $\hat{\alpha}_{\hat{G}_k}$, and $\partial \hat{\mu}_i(\beta_i)/\partial \beta_i$ in the absence of the Lasso penalty term. By Taylor expansion, we have

$$\hat{\alpha}_{\hat{G}_k} - \alpha_k^0 = -\hat{H}_{\hat{G}_k}^{-1} \frac{1}{N_k T} \sum_{i \in \hat{G}_k} \sum_{t=1}^{T} U_i \left(w_{it}; \alpha_k^0, \hat{\mu}_i(\alpha_k^0)\right),$$

where $\hat{H}_{\hat{G}_k} \equiv \frac{1}{N_k T} \sum_{i \in \hat{G}_k} \sum_{t=1}^{T} [U_i^{\alpha_k}(w_{it}; \check{\alpha}_{\hat{G}_k}, \hat{\mu}_i(\check{\alpha}_{\hat{G}_k})) + U_i^{\mu_i}(w_{it}; \check{\alpha}_k^0, \hat{\mu}_i(\check{\alpha}_{\hat{G}_k})) \partial \hat{\mu}_i(\check{\alpha}_{\hat{G}_k})/\partial \alpha_k']$ and $\check{\alpha}_{\hat{G}_k}$ lies between $\hat{\alpha}_{\hat{G}_k}$ and α_k^0 elementwise. Following the analysis of $\hat{H}_{(k)}$ in Lemma S1.13, we can also show that $\hat{H}_{\hat{G}_k} = \mathbb{H}_k + o_P(1)$. This result, in conjunction with Lemma S1.12, implies that $\sqrt{N_k T}(\hat{\alpha}_{\hat{G}_k} - \alpha_k^0) - \mathbb{H}_k^{-1} \mathbb{B}_{kNT} \xrightarrow{D} N\left(0, \mathbb{H}_k^{-1} \Omega_k (\mathbb{H}_k^{-1})'\right)$. ■

Proof of Theorem 2.6. Let $\mathcal{K} = \{1, 2, ..., K_{\max}\}$. We divide $\mathcal{K}$ into three subsets: $\mathcal{K}_0 \equiv \{K_0\}$, $\mathcal{K}_- \equiv \{K \in \mathcal{K} : K < K_0\}$, and $\mathcal{K}_+ \equiv \{K \in \mathcal{K} : K > K_0\}$, in which the true, under-, and over-fitted models are produced, respectively. Let $\hat{\sigma}_{\hat{G}(K,\lambda_1)}^2 = \frac{2}{NT} \sum_{k=1}^{K} \sum_{i \in \hat{G}_k(K,\lambda_1)} \sum_{t=1}^{T} \psi(w_{it}; \hat{\alpha}_{\hat{G}_k(K,\lambda_1)}, \hat{\mu}_i(\hat{\alpha}_{\hat{G}_k(K,\lambda_1)}))$. Using Theorems 2.2 and 2.5 and Assumption A5, we can readily show that $IC_1(K_0, \lambda_1) = \hat{\sigma}_{\hat{G}(K_0,\lambda_1)}^2 + \rho_{1NT} p K_0 = \frac{2}{NT} \sum_{k=1}^{K_0} \sum_{i \in \hat{G}_k(K_0,\lambda_1)} \sum_{t=1}^{T} \psi\left(w_{it}; \alpha_k^0, \mu_i^0\right) + o_P(1) \xrightarrow{P} \ln\left(\sigma_0^2\right)$. 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

$$\hat{\sigma}_{\hat{G}(K,\lambda_1)}^2 \geq \min_{1 \leq K < K_0} \inf_{G^{(K)} \in \mathcal{G}_K} \frac{2}{NT} \sum_{k=1}^{K} \sum_{i \in G_{K,k}} \sum_{t=1}^{T} \psi\left(w_{it}; \hat{\alpha}_{G_{K,k}}, \hat{\mu}_i(\hat{\alpha}_{G_{K,k}})\right) = \min_{1 \leq K < K_0} \inf_{G^{(K)} \in \mathcal{G}_K} \hat{\sigma}_{G^{(K)}}^2,$$

we have by Assumptions A4-A5 that $\min_{1 \leq K < K_0} IC_1(K, \lambda_1) \geq \min_{1 \leq K < K_0} \inf_{G^{(K)} \in \mathcal{G}_K} \hat{\sigma}_{G^{(K)}}^2 + \rho_{1NT} p K \xrightarrow{P} \ln(\underline{\sigma}^2) > \ln(\sigma_0^2)$. 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 S1.14 in the supplemental appendix and the fact that $T\rho_{1NT} \to \infty$ under Assumption A5, we have

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

It follows that $P(\hat{K}(\lambda_1) = K_0) \to 1$ as $(N,T) \to \infty$. ■

B Proofs 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* $P(\underline{c}[\frac{1}{2}\bar{V}_i(\beta_i) - R_{i,T}(\beta_i)] \leq V_{iNT}(\beta_i) \leq \bar{c}[2\bar{V}_i(\beta_i) + 2R_{i,T}(\beta_i)]) = 1 - o(N^{-1})$ *for all* $\beta_i \in \mathcal{B}_i$*, 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. Let $\Lambda_{NT} \equiv \{\max_i \|W_{iNT} - W_i\| \leq C\underline{c}_W\}$ for some $C \in (0,1)$. Then $P(\Lambda_{NT}) = 1 - o(N^{-1})$ by Assumption B1(iv). On the set Λ_{NT}, we have

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

for all $\beta_i \in \mathcal{B}_i$. By positive definiteness of W_i and simple manipulations, 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 $a = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\rho(\xi_{it},\beta_i)]$ and $b = \frac{1}{T}\sum_{t=1}^{T}\{\rho(\xi_{it},\beta_i) - \mathbb{E}[\rho(\xi_{it},\beta_i)]\}$, we have

$$\begin{aligned}\left[\frac{1}{T}\sum_{t=1}^{T}\rho(\xi_{it},\beta_i)\right]' W_i\left[\frac{1}{T}\sum_{t=1}^{T}\rho(\xi_{it},\beta_i)\right] &\geq \frac{1}{2}\bar{V}_i(\beta_i) - R_{i,T}(\beta_i), \text{ and} &\text{(B.2)}\\ \left[\frac{1}{T}\sum_{t=1}^{T}\rho(\xi_{it},\beta_i)\right]' W_i\left[\frac{1}{T}\sum_{t=1}^{T}\rho(\xi_{it},\beta_i)\right] &\leq 2\bar{V}_i(\beta_i) + 2R_{i,T}(\beta_i). &\text{(B.3)}\end{aligned}$$

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

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

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

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 $V_{iNT}(\beta_i^0) \leq \bar{c}[2\bar{V}_i(\beta_i^0) + 2R_{i,T}^0] = 2\bar{c}R_{i,T}^0$ on the set Λ_{NT}, where $\tilde{R}_{i,T} = R_{i,T}(\tilde{\beta}_i)$ and $R_{i,T}^0 = R_{i,T}(\beta_i^0)$. It follows that $\underline{c}[\frac{1}{2}\bar{V}_i(\tilde{\beta}_i) - \tilde{R}_{i,T}] - 2\bar{c}R_{i,T}^0 + \lambda_2\left\{\Pi_{k=1}^{K}\left\|\tilde{\beta}_i - \tilde{\alpha}_k\right\| - \Pi_{k=1}^{K}\left\|\beta_i^0 - \tilde{\alpha}_k\right\|\right\} \leq 0$, which can be rewritten as

$$\bar{V}_i\left(\tilde{\beta}_i\right) \leq \frac{2}{\underline{c}}\left[2\bar{c}R_{i,T}^0 + \underline{c}\tilde{R}_{i,T} - \lambda_2\left(\Pi_{k=1}^{K}\left\|\tilde{\beta}_i - \tilde{\alpha}_k\right\| - \Pi_{k=1}^{K}\left\|\beta_i^0 - \tilde{\alpha}_k\right\|\right)\right]. \quad \text{(B.4)}$$

By the arguments used to obtain (A.3) and (A.6), we have

$$\left|\Pi_{k=1}^{K_0}\left\|\tilde{\beta}_i - \alpha_k\right\| - \Pi_{k=1}^{K_0}\left\|\beta_i^0 - \alpha_k\right\|\right| \leq C_{K_0}(\boldsymbol{\alpha})\left(\left\|\tilde{\beta}_i - \beta_i^0\right\| + 2\left\|\tilde{\beta}_i - \beta_i^0\right\|^2\right). \quad \text{(B.5)}$$

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

$$\max_{1\leq i\leq N}\bar{V}_i(\tilde{\beta}_i) = \max_{1\leq i\leq N}\left(\tilde{\beta}_i - \beta_i^0\right)' \bar{Q}_{i,z\Delta x}' W_i\bar{Q}_{i,z\Delta x}\left(\tilde{\beta}_i - \beta_i^0\right) \geq \underline{c}_{1NT}\max_{1\leq i\leq N}\left\|\tilde{\beta}_i - \beta_i^0\right\|^2, \quad \text{(B.6)}$$

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 $\liminf_{(N,T)\to\infty} \underline{c}_{1NT} \geq \underline{c}_W \underline{c}_{\bar{Q}}^2 > 0$ by Assumptions B1(iii)-(iv). Combining (B.4)-(B.6) yields

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

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

$$\begin{aligned}\left\|\tilde{\beta}_i - \beta_i^0\right\| &\leq \frac{\frac{2}{\underline{c}}\lambda_2\tilde{C}_{K_0} + \left[\left(\frac{2}{\underline{c}}\lambda_2\tilde{C}_{K_0}\right)^2 + \frac{8}{\underline{c}}(\underline{c}_{1NT} - \frac{4}{\underline{c}}\lambda_2\tilde{C}_{K_0})\left(2\bar{c}R_{i,T}^0 + \underline{c}\tilde{R}_{i,T}\right)\right]^{1/2}}{2\left(\underline{c}_{1NT} - \frac{4}{\underline{c}}\lambda_2\tilde{C}_{K_0}\right)} \\ &= O_P\left(T^{-1/2} + \lambda_2\right). \qquad \text{(B.7)}\end{aligned}$$

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 $|P_{NT}(\tilde{\boldsymbol{\beta}}, \boldsymbol{\alpha}) - P_{NT}\left(\boldsymbol{\beta}^0, \boldsymbol{\alpha}\right)| = O_P\left(T^{-1/2}\right)$ 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) First, we fix $k \in \{1, ..., K_0\}$. By the consistency of $\tilde{\alpha}_k$ and $\tilde{\beta}_i$ in Theorem 3.1 and Assumptions B1(v)-(vi), we have $\tilde{\beta}_i - \tilde{\alpha}_l \xrightarrow{P} \alpha_k^0 - \alpha_l^0 \neq 0$ for all $i \in G_k^0$ and $l \neq k$ and $\tilde{c}_{ki} \equiv \Pi_{l=1,l\neq k}^{K_0}||\tilde{\beta}_i - \tilde{\alpha}_l|| \xrightarrow{P} c_k^0 \equiv \Pi_{l=1,l\neq k}^{K_0}||\alpha_k^0 - \alpha_l^0|| \geq c_\alpha^{K_0-1} > 0$ for any $i \in G_k^0$. 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

$$\begin{aligned}\mathbf{0}_{p\times 1} &= -2\tilde{Q}'_{i,z\Delta x} W_{iNT}\frac{1}{\sqrt{T}}\sum_{t=1}^T z_{it}\left(\Delta y_{it} - \tilde{\beta}_i'\Delta x_{it}\right) + \sqrt{T}\lambda_2\sum_{j=1}^{K_0}\tilde{e}_{ij}\Pi_{l=1,l\neq j}^{K_0}\left\|\tilde{\beta}_i - \tilde{\alpha}_l\right\| \\ &= -2\tilde{Q}'_{i,z\Delta x} W_{iNT}\frac{1}{\sqrt{T}}\sum_{t=1}^T z_{it}\Delta\varepsilon_{it} + \left\{\frac{\lambda_2\tilde{c}_{ki}}{\left\|\tilde{\beta}_i - \tilde{\alpha}_k\right\|}I_p + 2\tilde{Q}'_{i,z\Delta x}W_{iNT}\tilde{Q}_{i,z\Delta x}\right\}\sqrt{T}\left(\tilde{\beta}_i - \tilde{\alpha}_k\right) \\ &\quad + 2\tilde{Q}'_{i,z\Delta x}W_{iNT}\tilde{Q}_{i,z\Delta x}\sqrt{T}\left(\tilde{\alpha}_k - \alpha_k^0\right) + \sqrt{T}\lambda_2\sum_{j=1,j\neq k}^{K_0}\tilde{e}_{ij}\Pi_{l=1,l\neq j}^{K_0}\left\|\tilde{\beta}_i - \tilde{\alpha}_l\right\| \\ &\equiv -\tilde{B}_{i1} + \tilde{B}_{i2} + \tilde{B}_{i3} + \tilde{B}_{i4}, \text{ say,} \qquad \text{(B.8)}\end{aligned}$$

where $\tilde{e}_{ij} = \frac{\tilde{\beta}_i - \tilde{\alpha}_j}{||\tilde{\beta}_i - \tilde{\alpha}_j||}$ if $||\tilde{\beta}_i - \tilde{\alpha}_j|| \neq 0$ and $\|\tilde{e}_{ij}\| \leq 1$ if $||\tilde{\beta}_i - \tilde{\alpha}_j|| = 0$. Following the proof of Lemma S1.7, we can show that $P\left(\max_i ||\tilde{\beta}_i - \beta_i^0|| \geq \eta\right) = o\left(N^{-1}\right)$ for any given $\eta > 0$. With this, by (B.7) and Assumptions B2(ii)-(iv), we can readily show that

$$P\left(\max_i ||\tilde{\beta}_i - \beta_i^0|| \geq C\varkappa_{2NT}\right) = o\left(N^{-1}\right) \text{ for some } C > 0, \qquad \text{(B.9)}$$

where $\varkappa_{2NT} = (T^{-1/2}(\ln T)^3 + \lambda_2)(\ln T)^\nu$. This, in conjunction with the proof of Theorem 3.1(iii), implies that

$$P\left(\sqrt{T}||\tilde{\alpha}_k - \alpha_k^0|| \geq C(\ln T)^\nu\right) = o\left(N^{-1}\right) \text{ and } P\left(\max_{i\in G_k^0}\left|\tilde{c}_{ki} - c_k^0\right| \geq c_k^0/2\right) = o\left(N^{-1}\right). \qquad \text{(B.10)}$$

By (B.9)-(B.10), $P(\max_{i\in G_k^0}||\tilde{B}_{i4}|| \geq C\sqrt{T}\lambda_2\varkappa_{2NT}) = o\left(N^{-1}\right)$. By Assumptions B1(iii)-(iv), we have $P(\max_{i\in G_k^0}||\tilde{Q}_{i,z\Delta x}'W_{iNT}\tilde{Q}_{i,z\Delta x} - \bar{Q}_{i,z\Delta x}'W_i\bar{Q}_{i,z\Delta x}|| \geq \eta) = o\left(N^{-1}\right)$ for any $\eta > 0$. This result, in conjunction with (B.10), implies that $P(\max_{i\in G_k^0}||\tilde{B}_{i3}|| \geq C(\ln T)^{\nu}) = o\left(N^{-1}\right)$ for some $C > 0$. It follows that $P\left(\Gamma_{kNT}\right) = 1 - o\left(N^{-1}\right)$, where

$$\begin{aligned}\Gamma_{kNT} \quad \equiv \quad & \left\{\max_{i\in G_k^0}\left|\tilde{c}_{ki} - c_k^0\right| \leq c_k^0/2\right\} \cap \left\{\max_{i\in G_k^0}\left\|W_{iNT} - W_i\right\| \leq \underline{c}_W/2\right\} \cap \left\{\max_{i\in G_k^0}\left\|\tilde{Q}_{i,z\Delta x} - \bar{Q}_{i,z\Delta x}\right\| \leq \underline{c}_{\bar{Q}}/2\right\} \\ & \cap\left\{\max_{i\in G_k^0}\left\|\tilde{B}_{i3}\right\| \leq C\left(\ln T\right)^{\nu}\right\} \cap \left\{\max_{i\in G_k^0}\left\|\tilde{B}_{i4}\right\| \leq C\sqrt{T}\lambda_2\varkappa_{2NT}\right\}.\end{aligned}$$

Then conditional on Γ_{kNT}, we have uniformly in $i \in G_k^0$

$$\begin{aligned}\left(\tilde{\beta}_i - \tilde{\alpha}_k\right)'\left(\tilde{B}_{i2} + \tilde{B}_{i3} + \tilde{B}_{i4}\right) \quad \geq \quad & \left\|\left(\tilde{\beta}_i - \tilde{\alpha}_k\right)'\tilde{B}_{i2}\right\| - \left\|\left(\tilde{\beta}_i - \tilde{\alpha}_k\right)'\left(\tilde{B}_{i3} + \tilde{B}_{i4}\right)\right\| \\ \geq \quad & \sqrt{T}\lambda_2\tilde{c}_{ki}\left\|\tilde{\beta}_i - \tilde{\alpha}_k\right\| - C\left\|\tilde{\beta}_i - \tilde{\alpha}_k\right\|\left[\left(\ln T\right)^{\nu} + \sqrt{T}\lambda_2\varkappa_{2NT}\right] \\ \geq \quad & \sqrt{T}\lambda_2 c_k^0\left\|\tilde{\beta}_i - \tilde{\alpha}_k\right\|/4 \text{ for sufficiently large } (N, T),\end{aligned}$$

because $\sqrt{T}\lambda_2 \gg (\ln T)^{\nu} + \sqrt{T}\lambda_2\varkappa_{2NT}$ by Assumption B2(i). Then by Assumptions B2(i)-(ii)

$$\begin{aligned}P(\tilde{E}_{kNT,i}) \quad = \quad & P\left(i \notin \tilde{G}_k \mid i \in G_k^0\right) = P\left(\tilde{B}_{i1} = \tilde{B}_{i2} + \tilde{B}_{i3} + \tilde{B}_{i4}\right) \\ \leq \quad & P\left(\left|\left(\tilde{\beta}_i - \tilde{\alpha}_k\right)'\tilde{B}_{i1}\right| \geq \left|\left(\tilde{\beta}_i - \tilde{\alpha}_k\right)'\left(\tilde{B}_{i2} + \tilde{B}_{i3} + \tilde{B}_{i4}\right)\right|\right) \\ \leq \quad & P\left(\left\|\tilde{B}_{i1}\right\| \geq \sqrt{T}\lambda_2 c_k^0/4,\ \Gamma_{kNT}\right) + P\left(\Gamma_{kNT}^c\right) \to 0 \text{ as } (N, T) \to \infty.\end{aligned}$$

It follows that $P\left(||\tilde{\beta}_i - \tilde{\alpha}_k|| = 0 \mid i \in G_k^0\right) \to 1$ 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 by Assumption B2(ii)

$$\begin{aligned}\sum_{k=1}^{K_0}\sum_{i\in G_k^0}P\left(\tilde{E}_{kNT,i}\right) \quad \leq \quad & \sum_{k=1}^{K_0}\sum_{i\in G_k^0}\left[P\left(\left\|\tilde{B}_{i1}\right\| \geq \sqrt{T}\lambda_2 c_k^0/4,\ \Gamma_{kNT}\right) + P\left(\Gamma_{kNT}^c\right)\right] \\ \leq \quad & N\max_{1\leq i\leq N}P\left(\left\|\tilde{Q}_{i,z\Delta x}'W_{iNT}\frac{1}{T}\sum_{t=1}^{T}z_{it}\Delta\varepsilon_{it}\right\| \geq \lambda_2 c_k^0/4, \Gamma_{kNT}\right) + o(1) \\ \leq \quad & N\max_{1\leq i\leq N}P\left(\left\|\frac{1}{T}\sum_{t=1}^{T}z_{it}\Delta\varepsilon_{it}\right\| \geq \lambda_2 c_{\alpha}^{K_0-1}/(16\underline{c}_{\bar{Q}}\underline{c}_W)\right) + o(1) = o(1),\end{aligned}$$

where we use the fact that $\left\|\tilde{Q}_{i,z\Delta x}\right\|\left\|W_{iNT}\right\| \geq \left(\left\|\bar{Q}_{i,z\Delta x}\right\| - \left\|\tilde{Q}_{i,z\Delta x} - \bar{Q}_{i,z\Delta x}\right\|\right)\left(\left\|W_i\right\| - \left\|W_{iNT} - W_i\right\|\right) \geq \underline{c}_{\bar{Q}}\underline{c}_W/4$ on the set Γ_{kNT}. 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. Based on the subdifferential calculus, the KKT conditions for the minimization of (3.2) are that for each $i = 1, ..., N$ and $k = 1, ..., K_0$,

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

where $\tilde{e}_{ij}$ is defined after (B.8). Fix $k \in \{1, ..., K_0\}$. As in the proof of Theorem 2.4, we can show that $\frac{2}{NT}\sum_{i\in\tilde{G}_k}\tilde{Q}'_{i,z\Delta x}W_{iNT}\sum_{t=1}^{T}z_{it}\left(\Delta y_{it}-\tilde{\alpha}'_k\Delta x_{it}\right)+\frac{\lambda_2}{N}\sum_{i\in\tilde{G}_0}\tilde{e}_{ik}\Pi_{l=1,l\neq k}^{K_0}||\tilde{\beta}_i-\tilde{\alpha}_l||=\mathbf{0}_{p\times 1}$ w.p.a.1. It follows that $\tilde{\alpha}_k=\tilde{\alpha}_{1k}+\tilde{\mathcal{R}}_k$, where $\tilde{\alpha}_{1k}=(\frac{1}{N}\sum_{i\in\tilde{G}_k}\tilde{Q}'_{i,z\Delta x}W_{iNT}\tilde{Q}_{i,z\Delta x})^{-1}\times\frac{1}{NT}\sum_{i\in\tilde{G}_k}\tilde{Q}'_{i,z\Delta x}W_{iNT}\sum_{t=1}^{T}z_{it}\Delta y_{it}$ and $\tilde{\mathcal{R}}_k=(\frac{1}{N}\sum_{i\in\tilde{G}_k}\tilde{Q}'_{i,z\Delta x}W_{iNT}\tilde{Q}_{i,z\Delta x})^{-1}\frac{\lambda_2}{2N}\sum_{i\in\tilde{G}_0}\tilde{e}_{ik}\Pi_{l=1,l\neq k}^{K_0}||\tilde{\beta}_i-\tilde{\alpha}_l||$. By Theorem 3.2, we can readily show that $P(\sqrt{NT}\left\|\tilde{\mathcal{R}}_k\right\|\geq\epsilon)=o(1)$ for any $\epsilon>0$, and

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

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. By Theorem 3.2, we can readily show that

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

where $\tilde{Q}^{(k)}_{z\Delta\varepsilon}=\frac{1}{N_kT}\sum_{i\in\tilde{G}_k}\sum_{t=1}^{T}z_{it}\Delta\varepsilon_{it}$ and $Q^{(k)}_{z\Delta\varepsilon,NT}=\frac{1}{N_kT}\sum_{i\in G_k^0}\sum_{t=1}^{T}z_{it}\Delta\varepsilon_{it}$. The results then follow by Assumption B3 and the Slutsky theorem. ■

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

REFERENCE

ANDO, T., AND J. BAI (2014): "Asset Pricing with a General Multifactor Structure," *Journal of Financial Econometrics*, forthcoming.

ANDO, T., AND J. BAI (2015): "Panel Data Models with Grouped Factor Structure under Unknown Group Membership," *Journal of Applied Econometrics*, forthcoming.

ARELLANO, M., and S. BOND (1991): "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations," *Review of Economic Studies* 58, 277-297.

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.

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.

BESLEY, T. AND T. PERSSON (2010): "State Capacity, Conflict, and Development," *Econometrica* 78, 1-34.

BESTER, C. A., AND C. B. HANSEN (2016): "Grouped Effects Estimators in Fixed Effects Models," *Journal of Econometrics* 190, 197-208.

BLATTMAN C. AND E. MIGUEL (2010): "Civil Wars," *Journal of Economic Literature* 48, 3-57.

BONHOMME, S., AND E. MANRESA (2015): "Grouped Patterns of Heterogeneity in Panel Data," *Econometrica* 83, 1147-1184.

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.

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.

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

CHAN, N.H., C.Y. YAU, AND R-M. ZHANG (2014): "Group Lasso for Structural Break Time Series," *Journal of the American Statistical Association* 109, 590-599.

COLLIER, P. AND A. HOEFFLER (2004): "Greed and Grievance in Civil Wars," *Oxford Economic Papers* 56, 563-595.

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

DHAENE, G., AND K. JOCHMANS (2015): "Split-panel Jackknife Estimation of Fixed-effect Models," *Review of Economic Studies* 82, 991-1030.

DJANKOV, S. AND M. REYNAL-QUEROL (2010): "Poverty and Civil War: Revisiting the Evidence," *Review of Economics and Statistics* 92, 1035-1041.

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.

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

ESTEBAN, J., L. MAYORAL AND D. RAY (2012): "Ethnicity and Conflict: An Empirical Study," *American Economic Review* 102, 1310-1342.

FEARON, J.D. AND D. D. LAITIN (2003): "Ethnicity, Insurgency, and Civil War," *The American Political Science Review* 97, 75-90.

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

FERNÁNDEZ-VAL, I., AND L. LEE (2013): "Panel Data Models with Nonadditive Unobservable Heterogeneity: Estimation and Inference," *Quantitative Economics* 4, 453–481.

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.

HAHN, J., AND H.R. MOON (2010): "Panel Data Models with Finite Number of Multiple Equilibria," *Econometric Theory* 26, 863-881.

HAHN, J., AND W. NEWEY (2004): "Jackknife and Analytical Bias Reduction for Nonlinear Panel Models," *Econometrica* 72, 1295-1319.

HARCHAOUI, Z., AND C. LÉVY-LEDUC (2010): "Multiple Change-point Estimation with a Total Variation Penalty," *Journal of the American Statistical Association* 105, 1481-1493.

HSIAO, C. (2014): *Analysis of Panel Data*, 3rd edition. Cambridge University Press, New York.

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., AND A. K. TAHMISCIOGLU (1997): "A Panel Analysis of Liquidity Constraints and Firm Investment," *Journal of the American Statistical Association 92*, 455-465.

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

KATO, K., A.F. GAVAO, AND G.V. MONTES-ROJAS (2012): "Asymptotics for Panel Quantile Regression Models with Individual Effects," *Journal of Econometrics* 170, 76-91.

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

LEE, K., M. H. PESARAN, AND R. SMITH (1997): "Growth and Convergence in a Multi-country 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.

LEE, Y., AND P.C.B. PHILLIPS (2015): "Model Selection in the Presence of Incidental Parameters," *Journal of Econometrics* 188, 474-489.

LEEB, H., AND P.M. PÖTSCHER (2008): "Sparse Estimators and the Oracle Property, or the Return of Hodges' Estimator," *Journal of Econometrics* 142, 201–211.

LEEB, H., AND P.M. PÖTSCHER (2009): " On the Distribution of Penalized Maximum Likelihood Estimators: the LASSO, SCAD, and Thresholding," *Journal of Multivariate Analysis* 100, 2065–2082.

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 (2016): "Shrinkage Estimation of Dynamic Panel Data Models with Interactive Fixed Effects," *Journal of Econometrics* 190, 148-175.

MIGUEL, E., S. SATYANATH AND E. SERGENTI (2004): "Economic Shocks and Civil Conflict: an Instrumental Variables Approach," *Journal of Political Economy* 112, 725-753.

NUNN, N. AND N. QIAN (2014): "US Food Aid and Civil Conflict," *American Economic Review* 104, 1630-1666.

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.

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.

QIAN, J., AND L. SU (2015): "Shrinkage Estimation of Regression Models with Multiple Structural Changes," *Econometric Theory*, forthcoming.

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

SARAFIDIS, V., AND N. WEBER (2015): "A Partially Heterogenous Framework for Analyzing Panel Data," *Oxford Bulletin of Economics and Statistics* 77, 274-296.

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

SU, L., Z. SHI, AND P.C.B. PHILLIPS (2014): "Identifying Latent Structures in Panel Data," *Cowles Foundation Discussion Papers* 1965, Yale University.

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.

Online Supplement to "Identifying Latent Structures in Panel Data"[1]

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

[a]*School of Economics, Singapore Management University*
[b]*Department of Economics, Chinese University of Hong Kong*
[c] *Yale University, University of Auckland,*
University of Southampton & Singapore Management University

This supplement is composed of four parts. Section S1 contains the proofs of some technical lemmas for the proofs of the main results in Section 2. Section S2 gives bias correction formulae in linear panel data models for both PPL and PGMM estimation. Sections S3 and S4 contain some additional simulation and applications results, respectively.

S1 Some Technical Lemmas for the Proofs of the Main Results in Section 2 of the Paper

In this appendix, we state and prove some technical lemmas that are used in the proofs of the main results in Section 2. We first state an exponential inequality for strong mixing processes.

Lemma S1.1 *Let $\{\zeta_t, t = 1, 2, ...\}$ 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 $\rho \in (0,1)$. If $\sup_{1\leq t\leq T} |\zeta_t| \leq M_T$, then there exists a constant C_0 depending on c_α and ρ such that for any $T \geq 2$ and $\epsilon > 0$,*

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

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

Proof. 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. ■

The above lemma is used in the proof of the following lemma.

Lemma S1.2 *(i) Let $\xi(w_{it};\phi)$ be a $\mathbb{R}^{d_\xi}$-valued function indexed by the parameter $\phi \in \Phi$, where Φ is a convex compact set in $\mathbb{R}^{p+1}$ and $\mathbb{E}[\xi(w_{it};\phi)] = 0$ for all i, t, and $\phi \in \Phi$. Assume that there exists a function $M(w_{it})$ such that $\left\|\xi(w_{it};\phi) - \xi(w_{it};\bar{\phi})\right\| \leq M(w_{it})\left\|\phi - \bar{\phi}\right\|$ for all $\phi, \bar{\phi} \in \Phi$ and $sup_\phi \|\xi(w_{it};\phi)\| \leq M(w_{it})$. Assume that $\mathbb{E}|M(w_{it})|^q < \infty$ for some $q \geq 6$ such that $N = O(T^{q/2-1})$. Let $\{\phi_i\}$ be a nonstochastic sequence in Φ. Then*

(i) $\max_{1\leq i\leq N}\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\xi(w_{it};\phi_i)\right\| = O_P((\ln T)^3)$,

(ii) $\max_{1\leq i\leq N} P\left(\left\|\frac{1}{T}\sum_{t=1}^{T}\xi(w_{it};\phi_i)\right\| \geq c\lambda\right) = o\left(N^{-1}\right)$ *for any given $c > 0$,*

(iii) $P\left(\max_{1\leq i\leq N}\left\|\frac{1}{T}\sum_{t=1}^{T}\xi(w_{it};\phi_i)\right\| \geq c\lambda\right) = o\left(N^{-1}\right)$ *for any given $c > 0$ if $N^2 = O(T^{q/2-1})$,*

where $\lambda = \lambda_{NT}$ satisfies $(\ln T)^3 = o\left(T^{1/2}\lambda\right)$.

[1]Acknowledgements made in the leading footnote of the main paper apply also to this Online Supplement. In particular, Su acknowledges support from the Singapore Ministry of Education for Academic Research Fund (AcRF) under the Tier-2 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.

Proof. (i) Let $\eta_{NT} = T^{1/2}$. Let ι_ξ be an arbitrary $d_\xi \times 1$ nonrandom vector with $\|\iota_\xi\| = 1$. Let $\mathbf{1}_{it} = \mathbf{1}\{\|\xi(w_{it};\phi_i)\| \le \eta_{NT}\}$ and $\bar{\mathbf{1}}_{it} = 1 - \mathbf{1}_{it}$. Define

$$\begin{aligned}
\xi_1(w_{it};\phi_i) &= \iota_\xi' \{\xi(w_{it};\phi_i)\mathbf{1}_{it} - \mathbb{E}[\xi(w_{it};\phi_i)\mathbf{1}_{it}]\}, \\
\xi_2(w_{it};\phi_i) &= \iota_\xi' \xi(w_{it};\phi_i)\bar{\mathbf{1}}_{it}, \text{ and } \xi_{3it} = -\iota_\xi' \mathbb{E}[\xi(w_{it};\phi_i)\bar{\mathbf{1}}_{it}].
\end{aligned}$$

Apparently $\xi_1(w_{it};\phi_i)+\xi_2(w_{it};\phi_i)+\xi_{3it} = \iota_\xi'\xi(w_{it};\phi_i)$ as $\mathbb{E}[\xi(w_{it};\phi_i)] = 0$. We prove the lemma by showing that (i1) $\max_{1\le i\le N}\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^T \xi_1(w_{it};\phi_i)\right\| = O_P((\ln T)^3)$, (i2) $P\left[\max_{1\le i\le N}\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^T \xi_2(w_{it};\phi_i)\right\| \ge c(\ln T)^3\right]$ $= o(1)$ for any given $c > 0$, and (i3) $\max_{1\le i\le N}\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^T \xi_{3it}\right\| = O((\ln T)^3)$.

First, we prove (i3). By the Hölder and Markov inequalities

$$\begin{aligned}
\max_{1\le i\le N}\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^T \xi_{3it}\right\| &\le T^{1/2}\max_{1\le i\le N}\max_{1\le t\le T}\|\mathbb{E}[\xi(w_{it};\phi_i)\bar{\mathbf{1}}_{it}]\| \\
&\le T^{1/2}\max_{1\le i\le N}\max_{1\le t\le T}\left\{\mathbb{E}\|\xi(w_{it};\phi_i)\|^{q/2}\right\}^{2/q}\left\{P\left(\|\xi(w_{it};\phi_i)\| > T^{1/2}\right)\right\}^{(q-2)/q} \\
&\le T^{1/2}c_{1q}\max_{1\le i\le N}\max_{1\le t\le T}\left\{P\left(\|\xi(w_{it};\phi_i)\| > T^{1/2}\right)\right\}^{(q-2)/q} \\
&\le T^{1/2}c_{1q}\max_{1\le i\le N}\max_{1\le t\le T}\left\{T^{-q/2}\mathbb{E}(\|\xi(w_{it};\phi_i)\|^q)\right\}^{(q-2)/q} \\
&= c_{1q}c_{2q}T^{(3-q)/2} = o((\ln T)^3) \text{ for any } q \ge 3,
\end{aligned}$$

where $c_{1q} \equiv \max_{1\le i\le N}\max_{1\le t\le T}\left\{\mathbb{E}\|\xi(w_{it};\phi_i)\|^{q/2}\right\}^{2/q}$ and $c_{2q} \equiv \max_{1\le i\le N}\max_{1\le t\le T}\{\mathbb{E}(\|\xi(w_{it};\phi_i)\|^q)\}^{(q-2)/q}$.

Next, we prove (i2). Noting that $\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^T \xi_2(w_{it};\phi_i)\right\| \ge (\ln T)^3$ implies that $\max_{1\le t\le T}\|\xi(w_{it};\phi_i)\| > \eta_{NT}$, by the Boole and Markov inequalities, the dominated convergence theorem, and the stated conditions, we have

$$\begin{aligned}
P\left[\max_{1\le i\le N}\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^T \xi_2(w_{it};\phi_i)\right\| \ge c(\ln T)^3\right] &\le P\left[\max_{1\le i\le N}\max_{1\le t\le T}\|\xi(w_{it};\phi_i)\| > \eta_{NT}\right] \\
&\le NT\max_{1\le i\le N}\max_{1\le t\le T}P(M(w_{it}) > \eta_{NT}) \\
&\le \frac{NT}{T^{q/2}}\max_{1\le i\le N}\max_{1\le t\le T}\mathbb{E}\left[|M(w_{it})|^q\mathbf{1}\left\{M(w_{it}) > T^{1/2}\right\}\right] \\
&= o\left(NT^{1-q/2}\right) = o(1).
\end{aligned}$$

Now, we prove (i1). We observe that for any $C > 0$,

$$P\left[\max_{1\le i\le N}\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^T \xi_1(w_{it};\phi_i)\right\| \ge C(\ln T)^3\right] \le \sum_{i=1}^N P\left[\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^T \xi_1(w_{it};\phi_i)\right\| \ge C(\ln T)^3\right].$$

We choose $\varepsilon > 0$ and divide Φ into subsets Φ_j, $j = 1,...,n_\varepsilon$ such that $\|\phi - \bar{\phi}\| < \varepsilon/\sqrt{T}$ for all $\phi, \bar{\phi} \in \Phi_j$, where $n_\varepsilon = O\left(T^{(p+1)/2}\right)$. Then

$$\begin{aligned}
\sum_{i=1}^N P\left[\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^T \xi_1(w_{it};\phi_i)\right\| \ge C(\ln T)^3\right] &\le \sum_{i=1}^N P\left[\sup_{\phi\in\Phi}\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^T \xi_1(w_{it};\phi)\right\| \ge C(\ln T)^3\right] \\
&\le \sum_{i=1}^N\sum_{j=1}^{n_\varepsilon} P\left[\sup_{\phi\in\Phi_j}\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^T \xi_1(w_{it};\phi)\right\| \ge C(\ln T)^3\right].
\end{aligned}$$

Let $\phi_j \in \Phi_j$. Then for any $\phi \in \Phi_j$ we have

$$\begin{aligned}\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\xi_1\left(w_{it};\phi\right)\right\| &\leq \left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\xi_1\left(w_{it};\phi_j\right)\right\| + \left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left[\xi_1\left(w_{it};\phi_j\right)-\xi_1\left(w_{it};\phi\right)\right]\right\| \\ &\leq \left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\xi_1\left(w_{it};\phi_j\right)\right\| + \frac{2}{\sqrt{T}}\sum_{t=1}^{T}M\left(w_{it}\right)\left\|\phi-\phi_j\right\| \\ &\leq \left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\xi_1\left(w_{it};\phi_j\right)\right\| + \left|\frac{2\varepsilon}{T}\sum_{t=1}^{T}\left\{M\left(w_{it}\right)-\mathbb{E}\left[M\left(w_{it}\right)\right]\right\}\right| + \frac{2\varepsilon}{T}\sum_{t=1}^{T}\mathbb{E}\left[M\left(w_{it}\right)\right].\end{aligned}$$

It follows that

$$\begin{aligned}P\left[\sup_{\phi\in\Phi_j}\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\xi_1\left(w_{it};\phi\right)\right\| \geq C\left(\ln T\right)^3\right] &\leq P\left[\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\xi_1\left(w_{it};\phi_j\right)\right\| \geq C\left(\ln T\right)^3/3\right] \\ &\quad + P\left[\left|\frac{2\varepsilon}{T}\sum_{t=1}^{T}M\left(w_{it}\right)-\mathbb{E}\left[M\left(w_{it}\right)\right]\right| \geq C\left(\ln T\right)^3/3\right]\end{aligned}$$

as $P\left[\frac{\varepsilon}{T}\sum_{t=1}^{T}\mathbb{E}\left[M\left(w_{it}\right)\right] \geq C\left(\ln T\right)^3/3\right]=0$. Then

$$\begin{aligned}P\left[\max_{1\leq i\leq N}\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\xi\left(w_{it};\phi_i\right)\right\| \geq C\left(\ln T\right)^3\right] &\leq \sum_{i=1}^{N}\sum_{j=1}^{n_\varepsilon}P\left[\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\xi_1\left(w_{it};\phi_j\right)\right\| \geq C\left(\ln T\right)^3/3\right] \\ &\quad + \sum_{i=1}^{N}\sum_{j=1}^{n_\varepsilon}P\left[\left|\frac{2\varepsilon}{T}\sum_{t=1}^{T}M\left(w_{it}\right)-\mathbb{E}\left[M\left(w_{it}\right)\right]\right| \geq C\left(\ln T\right)^3/3\right].\end{aligned}$$

For the first term, we have by Lemma S1.1

$$\begin{aligned}\sum_{i=1}^{N}\sum_{j=1}^{n_\varepsilon}P\left[\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\xi_1\left(w_{it};\phi_j\right)\right\| \geq C\left(\ln T\right)^3/3\right] &\leq CNn_\varepsilon\exp\left(-\frac{C^2C_0T\left(\ln T\right)^6/9}{\bar{v}^2T+4\eta_{NT}^2+\frac{2C}{3}\eta_{NT}T^{1/2}\left(\ln T\right)^{3+2}}\right) \\ &= C\exp\left(-\frac{C^2C_0T\left(\ln T\right)^6/9}{\bar{v}^2T+4T+\frac{2C}{3}T\left(\ln T\right)^5}+\ln N+\ln n_\varepsilon\right) \\ &\to 0 \text{ for sufficiently large } C.\end{aligned}$$

Similarly, we can show that $\sum_{i=1}^{N}\sum_{j=1}^{n_\varepsilon}P\left[\left\|\frac{2\varepsilon}{T}\sum_{t=1}^{T}M\left(w_{it}\right)-\mathbb{E}\left[M\left(w_{it}\right)\right]\right\| \geq C\left(\ln T\right)^3/3\right]=o\left(1\right)$. Then (i1) follows. This completes the proof of (i).

(ii) Let ξ_1, ξ_2, and ξ_{3it} be as defined in (i). Noting that ξ_{3it} is nonrandom, it suffices to show that for any given $c>0$, we have (ii1) $N\max_{1\leq i\leq N}P(||\frac{1}{T}\sum_{t=1}^{T}\xi_1\left(w_{it};\phi_i\right)|| \geq c\lambda)=o\left(1\right)$, (ii2) $N\max_{1\leq i\leq N}P(||\frac{1}{T}\sum_{t=1}^{T}\xi_2\left(w_{it};\phi_i\right)|| \geq c\lambda)=o\left(1\right)$, and (ii3) $\max_{1\leq i\leq N}\left\|\frac{1}{T}\sum_{t=1}^{T}\xi_{3it}\right\|=o\left(\lambda\right)$. Following the analysis of $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\xi_{3it}$ in (i), we have

$$\max_{1\leq i\leq N}\left\|\frac{1}{T}\sum_{t=1}^{T}\xi_{3it}\right\| \leq c_{1q}c_{2q}T^{(2-q)/2}=o(\lambda)$$

where we use the fact that $\lambda \gg T^{-1/2}(\ln T)^3$ and $q \geq 3$ by the stated conditions. Thus, (ii3) follows. Following the analysis of $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\xi_3(w_{it};\phi_i)$ in (i2), we have

$$
\begin{aligned}
N\max_{1\leq i\leq N}P\left(\left\|\frac{1}{T}\sum_{t=1}^{T}\xi_2(w_{it};\phi_i)\right\|\geq c\lambda\right) &\leq N\max_{1\leq i\leq N}P\left(\max_{1\leq t\leq T}\|\xi(w_{it};\phi_i)\|>\eta_{NT}\right)\\
&\leq NT\max_{1\leq i\leq N}\max_{1\leq t\leq T}P(M(w_{it})>\eta_{NT})\\
&= o\left(NT^{1-q/2}\right)=o(1).
\end{aligned}
$$

That is, (ii2) follows. For (ii1), the analysis is similar to that of $\max_{1\leq i\leq N}\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\xi_1(w_{it};\phi_i)\right\|$ in (i1) with $(\ln T)^3$ replaced by $T^{1/2}\lambda$. We now require $T^{1/2}\lambda/(\ln T)^3\to\infty$ as $(N,T)\to\infty$. This completes the proof of (ii).

(iii) Let ξ_1, ξ_2, and ξ_{3it} be as defined in (i). Noting that ξ_{3it} is nonrandom, it suffices to show that for any given $c>0$, we have (iii1) $N\cdot P(\max_{1\leq i\leq N}||\frac{1}{T}\sum_{t=1}^{T}\xi_1(w_{it};\phi_i)||\geq c\lambda)=o(1)$, (iii2) $N\cdot P(\max_{1\leq i\leq N}||\frac{1}{T}\sum_{t=1}^{T}\xi_2(w_{it};\phi_i)||\geq c\lambda)=o(1)$, and (iii3) $\max_{1\leq i\leq N}\left\|\frac{1}{T}\sum_{t=1}^{T}\xi_{3it}\right\|=o(\lambda)$. Following the analysis of $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\xi_{3it}$ in (i), we have

$$
\max_{1\leq i\leq N}\left\|\frac{1}{T}\sum_{t=1}^{T}\xi_{3it}\right\|\leq c_{1q}c_{2q}T^{(2-q)/2}=o(\lambda),
$$

where we use the fact that $\lambda \gg T^{-1/2}(\ln T)^3$ and $q\geq 6$ by the stated conditions. Thus, (iii3) follows. Following the analysis of $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\xi_3(w_{it};\phi_i)$ in (i2), we have

$$
\begin{aligned}
N\cdot P\left(\max_{1\leq i\leq N}\left\|\frac{1}{T}\sum_{t=1}^{T}\xi_2(w_{it};\phi_i)\right\|\geq c\lambda\right) &\leq N\cdot P\left(\max_{1\leq i\leq N}\max_{1\leq t\leq T}\|\xi(w_{it};\phi_i)\|>\eta_{NT}\right)\\
&\leq N^2T\max_{1\leq i\leq N}\max_{1\leq t\leq T}P(M(w_{it})>\eta_{NT})\\
&= o\left(N^2T^{1-q/2}\right)=o(1).
\end{aligned}
$$

That is, (iii2) follows. For (iii1), the analysis is similar to that of $\max_{1\leq i\leq N}\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\xi_1(w_{it};\phi_i)\right\|$ in (i1) with $(\ln T)^3$ replaced by $T^{1/2}\lambda$. We now require $T^{1/2}\lambda/(\ln T)^3\to\infty$ as $(N,T)\to\infty$. This completes the proof of (ii). ■

Recall that $\hat{\Psi}_i(\beta,\mu)=\frac{1}{T}\sum_{t=1}^{T}\psi(w_{it};\beta,\mu)$ and $\Psi_i(\beta,\mu)=\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\psi(w_{it};\beta,\mu)]$. Recall that $\hat{\mu}_i(\beta_i)=\arg\min_{\mu_i}\frac{1}{T}\sum_{t=1}^{T}\psi(w_{it};\beta_i,\mu_i)$. The following three lemmas study the properties of $\hat{\Psi}_i(\beta,\mu)$ and $\hat{\mu}_i(\beta_i)$.

Lemma S1.3 *For any $\eta>0$, we have* $P\left[\max_{1\leq i\leq N}\sup_{(\beta,\mu)}\left|\hat{\Psi}_i(\beta,\mu)-\Psi_i(\beta,\mu)\right|\geq\eta\right]=o\left(N^{-1}\right)$.

Proof. The proof is analogous to that of Lemma S1.2(iii). ■

Lemma S1.4 *For any $\eta>0$, we have* $P\left[\max_{1\leq i\leq N}|\hat{\mu}_i(\beta_i)-\mu_i(\beta_i)|\geq\eta\right]=o\left(N^{-1}\right)$.

Proof. Let $\varepsilon=\min_i\left[\inf_{\mu_i:|\mu_i-\mu_i(\beta_i)|>\eta}\Psi_i(\beta_i,\mu_i)-\Psi_i(\beta_i,\mu_i(\beta_i))\right]$. Then $\varepsilon>0$ by Assumptions A1(ii) and (v). Then conditional on the event $A\equiv\left\{\max_{1\leq i\leq N}\sup_{(\beta,\mu)}\left|\hat{\Psi}_i(\beta,\mu)-\Psi_i(\beta,\mu)\right|\leq\frac{1}{3}\varepsilon\right\}$, we have

$$
\begin{aligned}
\inf_{|\mu_i-\mu_i(\beta_i)|>\eta}\hat{\Psi}_i(\beta_i,\mu_i) &\geq \inf_{|\mu_i-\mu_i(\beta_i)|>\eta}\Psi_i(\beta_i,\mu_i)-\frac{1}{3}\varepsilon\\
&\geq \Psi_i(\beta_i,\mu_i(\beta_i))+\frac{2}{3}\varepsilon\\
&\geq \hat{\Psi}_i(\beta_i,\mu_i(\beta_i))+\frac{1}{3}\varepsilon.
\end{aligned}
$$

On the other hand, $\hat{\Psi}_i(\beta_i, \hat{\mu}_i(\beta_i)) \leq \hat{\Psi}_i(\beta_i, \mu_i(\beta_i))$. It follows that $P(\max_{1\leq i\leq N} |\hat{\mu}_i(\beta_i) - \mu_i(\beta_i)| \leq \eta) \leq P(A) = o(N^{-1})$ by Lemma S1.3. ■

Lemma S1.5 *(i)* $\hat{\mu}_i(\beta_i) - \mu_i(\beta_i) = O_P(T^{-1/2})$ *for each* i,

(ii) $\max_{1\leq i\leq N} |\hat{\mu}_i(\beta_i) - \mu_i(\beta_i)| = O_P\left(T^{-1/2}(\ln T)^3\right)$,

(iii) $\max_{1\leq i\leq N} |\Psi_i(\beta_i, \hat{\mu}_i(\beta_i)) - \Psi_i(\beta_i, \mu_i(\beta_i))| = O_P\left(T^{-1/2}(\ln T)^3\right)$,

(iv) $P\left(\max_{1\leq i\leq N} |\hat{\mu}_i(\beta_i) - \mu_i(\beta_i)| \geq CT^{-1/2}(\ln T)^{3+\nu}\right) = o(N^{-1})$ *for any* $\nu > 0$ *and* $C > 0$,

(v) $P\left(\max_{1\leq i\leq N} |\Psi_i(\beta_i, \hat{\mu}_i(\beta_i)) - \Psi_i(\beta_i, \mu_i(\beta_i))| \geq CT^{-1/2}(\ln T)^{3+\nu}\right) = o(N^{-1})$ *for any* $\nu > 0$ *and* $C > 0$.

Proof. (i)-(ii) Noting that $\hat{\mu}_i(\beta_i) = \arg\min_{\mu_i} \frac{1}{T}\sum_{t=1}^T \psi(w_{it}; \beta_i, \mu_i)$, we have

$$\begin{aligned} 0 &= \frac{1}{T}\sum_{t=1}^T V_i(w_{it}; \beta_i, \hat{\mu}_i(\beta_i)) \\ &= \frac{1}{T}\sum_{t=1}^T V_i(w_{it}; \beta_i, \mu_i(\beta_i)) + \frac{1}{T}\sum_{t=1}^T V_i^{\mu_i}(w_{it}; \beta_i, \breve{\mu}_i(\beta_i))[\hat{\mu}_i(\beta_i) - \mu_i(\beta_i)], \end{aligned}$$

where $\breve{\mu}_i(\beta_i)$ lies between $\hat{\mu}_i(\beta_i)$ and $\mu_i(\beta_i)$ for each i. It follows that

$$\hat{\mu}_i(\beta_i) - \mu_i(\beta_i) = -\left[\frac{1}{T}\sum_{t=1}^T V_i^{\mu_i}(w_{it}; \beta_i, \breve{\mu}_i(\beta_i))\right]^{-1} \frac{1}{T}\sum_{t=1}^T V_i(w_{it}; \beta_i, \mu_i(\beta_i)) \tag{S1}$$

provided $\frac{1}{T}\sum_{t=1}^T V_i^{\mu_i}(w_{it}; \beta_i, \breve{\mu}_i(\beta_i))$ is asymptotically nonvanishing. Let $V_{it}(\beta_i) = V_i(w_{it}; \beta_i, \mu_i(\beta_i))$. Noting that $\mathbb{E}[V_{it}(\beta_i)] = 0$ and

$$\begin{aligned} \operatorname{Var}\left(\frac{1}{T}\sum_{t=1}^T V_{it}(\beta_i)\right) &= \frac{1}{T^2}\sum_{t=1}^T\sum_{s=1}^T \operatorname{Cov}(V_{it}(\beta_i), V_{is}(\beta_i)) \\ &\leq 8\max_{i,t}\{\mathbb{E}|V_{it}(\beta_i)|^q\}^{2/q} \frac{1}{T^2}\sum_{t=1}^T\sum_{s=1}^T \alpha(|t-s|)^{1-2/q} \\ &\leq 8\max_{i,t}\{\mathbb{E}|V_{it}(\beta_i)|^q\}^{2/q} \frac{1}{T}\sum_{\tau=1}^\infty \alpha(\tau)^{1-2/q} = O\left(\frac{1}{T}\right) \end{aligned}$$

by the Davydov inequality (e.g., Corollary A.2 in Hall and Heyde (1980)), we have $\frac{1}{T}\sum_{t=1}^T V_{it}(\beta_i) = O_P(T^{-1/2})$ by the Chebyshev inequality. In addition, by a simple application of Lemma S1.2(i), we can show that $\max_{1\leq i\leq N}\left|\frac{1}{T}\sum_{t=1}^T V_{it}(\beta_i)\right| = O_P(T^{-1/2}(\ln T)^3)$.

For $\frac{1}{T}\sum_{t=1}^T V_i^{\mu_i}(w_{it}; \beta_i, \breve{\mu}_i(\beta_i))$ we make the following decomposition:

$$\begin{aligned} \frac{1}{T}\sum_{t=1}^T V_i^{\mu_i}(w_{it}; \beta_i, \breve{\mu}_i(\beta_i)) &= \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[V_i^{\mu_i}(w_{it}; \beta_i, \mu_i(\beta_i))\right] \\ &\quad + \frac{1}{T}\sum_{t=1}^T \left\{V_i^{\mu_i}(w_{it}; \beta_i, \mu_i(\beta_i)) - \mathbb{E}\left[V_i^{\mu_i}(w_{it}; \beta_i, \mu_i(\beta_i))\right]\right\} \\ &\quad + \frac{1}{T}\sum_{t=1}^T \left\{V_i^{\mu_i}(w_{it}; \beta_i, \breve{\mu}_i(\beta_i)) - V_i^{\mu_i}(w_{it}; \beta_i, \mu_i(\beta_i))\right\}. \end{aligned} \tag{S2}$$

By Assumption A1(v), $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[V_i^{\mu_i}\left(w_{it};\beta_i,\mu_i\left(\beta_i\right)\right)\right] = H_{i\mu\mu}\left(\beta_i\right) \geq c_H > 0$ uniformly in i. By a simple application of Lemma S1.2(i), we have

$$\max_{1\leq i\leq N}\left|\frac{1}{T}\sum_{t=1}^{T}\left\{V_i^{\mu_i}\left(w_{it};\beta_i,\mu_i\left(\beta_i\right)\right)-\mathbb{E}\left[V_i^{\mu_i}\left(w_{it};\beta_i,\mu_i\left(\beta_i\right)\right)\right]\right\}\right| = o_P\left(1\right).$$

Next, by Assumption A1, and Lemmas S1.2(i) and S1.4, we have

$$\begin{aligned}
&\max_{1\leq i\leq N}\left|\frac{1}{T}\sum_{t=1}^{T}\left[V_i^{\mu_i}\left(w_{it};\beta_i,\breve{\mu}_i\left(\beta_i\right)\right)-V_i^{\mu_i}\left(w_{it};\beta_i,\mu_i\left(\beta_i\right)\right)\right]\right| \\
\leq\ & \max_{1\leq i\leq N}\frac{1}{T}\sum_{t=1}^{T}M\left(w_{it}\right)\left|\breve{\mu}_i\left(\beta_i\right)-\mu_i\left(\beta_i\right)\right| \\
\leq\ & \left\{\max_{1\leq i\leq N}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[M\left(w_{it}\right)\right]+\max_{1\leq i\leq N}\left|\frac{1}{T}\sum_{t=1}^{T}\left\{M\left(w_{it}\right)-\mathbb{E}\left[M\left(w_{it}\right)\right]\right\}\right|\right\}\max_{1\leq i\leq N}\left|\hat{\mu}_i\left(\beta_i\right)-\mu_i\left(\beta_i\right)\right| \\
\leq\ & \left[c_M^{1/q}+o_P\left(1\right)\right]o_P\left(1\right)=o_P\left(1\right).
\end{aligned} \tag{S3}$$

It follows that $\frac{1}{T}\sum_{t=1}^{T}V_i^{\mu_i}\left(w_{it};\beta_i,\breve{\mu}_i\left(\beta_i\right)\right) = H_{i\mu\mu}(\beta_i)+o_P\left(1\right)$ uniformly in i, $\hat{\mu}_i\left(\beta_i\right)-\mu_i\left(\beta_i\right) = O_P(T^{-1/2})$ for each i, and $\max_{1\leq i\leq N}\left|\hat{\mu}_i\left(\beta_i\right)-\mu_i\left(\beta_i\right)\right| = O_P(T^{-1/2}\left(\ln T\right)^3)$.

(iii) In view of the definition that $\Psi_i\left(\beta_i,\mu_i\right) = \mathbb{E}\left[\psi\left(w_{it};\beta_i,\mu_i\right)\right]$, we have $\max_{1\leq i\leq N}\left|\Psi_i\left(\beta_i,\hat{\mu}_i\left(\beta_i\right)\right)-\Psi_i(\beta_i,\,\mu_i\left(\beta_i\right))\right| = \max_{i,t}\mathbb{E}\left|M\left(w_{it}\right)\right|\left|\hat{\mu}_i\left(\beta_i\right)-\mu_i\left(\beta_i\right)\right| = O_P(T^{-1/2}\left(\ln T\right)^3)$.

(iv) We define the following events:

$$\begin{aligned}
A_1 &\equiv \left\{\max_{1\leq i\leq N}\left|\hat{\mu}_i\left(\beta_i\right)-\mu_i\left(\beta_i\right)\right|\leq c_H/(6c_M^{1/q})\right\}, \\
A_2 &\equiv \left\{\max_{1\leq i\leq N}\left|\frac{1}{T}\sum_{t=1}^{T}\left\{M\left(w_{it}\right)-\mathbb{E}\left[M\left(w_{it}\right)\right]\right\}\right|\leq c_M^{1/q}/2\right\}, \\
A_3 &\equiv \left\{\max_{1\leq i\leq N}\left|\frac{1}{T}\sum_{t=1}^{T}\left[V_i^{\mu_i}\left(w_{it};\beta_i,\breve{\mu}_i\left(\beta_i\right)\right)-V_i^{\mu_i}\left(w_{it};\beta_i,\mu_i\left(\beta_i\right)\right)\right]\right|\leq c_H/4\right\}, \\
A_4 &\equiv \left\{\min_{1\leq i\leq N}\left|\frac{1}{T}\sum_{t=1}^{T}V_i^{\mu_i}\left(w_{it};\beta_i,\mu_i\left(\beta_i\right)\right)\right|\geq c_H/2\right\}, \\
A_5 &\equiv \left\{\min_{1\leq i\leq N}\left|\frac{1}{T}\sum_{t=1}^{T}V_i^{\mu_i}\left(w_{it};\beta_i,\breve{\mu}_i\left(\beta_i\right)\right)\right|\geq c_H/4\right\}.
\end{aligned}$$

Let A_j^c denote the complement of A_j for $j = 1,2,3,4,5$. Let $\delta_i = \hat{\mu}_i\left(\beta_i\right)-\mu_i\left(\beta_i\right)$. By Lemmas S1.4 and S1.2(iii), $P\left(A_1^c\right) = o\left(N^{-1}\right)$ and $P\left(A_2^c\right) = o\left(N^{-1}\right)$. Then by (S3)

$$\begin{aligned}
&P\left(A_3^c\right) \\
\leq\ & P\left(\left\{\max_{1\leq i\leq N}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[M\left(w_{it}\right)\right]+\max_{1\leq i\leq N}\left|\frac{1}{T}\sum_{t=1}^{T}\left\{M\left(w_{it}\right)-\mathbb{E}\left[M\left(w_{it}\right)\right]\right\}\right|\right\}\max_{1\leq i\leq N}\left|\delta_i\right|\geq c_H/4\right) \\
\leq\ & P\left(\left\{\max_{1\leq i\leq N}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[M\left(w_{it}\right)\right]+\max_{1\leq i\leq N}\left|\frac{1}{T}\sum_{t=1}^{T}\left\{M\left(w_{it}\right)-\mathbb{E}\left[M\left(w_{it}\right)\right]\right\}\right|\right\}\max_{1\leq i\leq N}\left|\delta_i\right|\geq c_H/4,\ A_2\right) \\
&+P\left(A_2^c\right) \\
\leq\ & P\left(3c_M^{1/q}\max_{1\leq i\leq N}\left|\delta_i\right|\geq c_H/2\right)+P\left(A_2^c\right) \\
\leq\ & P\left(A_1^c\right)+P\left(A_2^c\right)=o\left(N^{-1}\right).
\end{aligned}$$

Let $V_{it}^{\mu_i}(\beta_i) \equiv V_i^{\mu_i}(w_{it};\beta_i,\mu_i(\beta_i))$. Noting that $\frac{1}{T}\sum_{t=1}^T V_{it}^{\mu_i}(\beta_i) = \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[V_{it}^{\mu_i}(\beta_i)\right] + \frac{1}{T}\sum_{t=1}^T \{V_{it}^{\mu_i}(\beta_i)$ $-\mathbb{E}\left[V_{it}^{\mu_i}(\beta_i)\right]\}$, $\min_{1\le i\le N}\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[V_{it}^{\mu_i}(\beta_i)\right] \ge c_H$ by Assumption A1(v) and

$$P\left(\max_{1\le i\le N}\left|\frac{1}{T}\sum_{t=1}^T \{V_{it}^{\mu_i}(\beta_i) - \mathbb{E}\left[V_{it}^{\mu_i}(\beta_i)\right]\right| \ge c_H/2\right) = o\left(N^{-1}\right) \text{ by Lemma S1.2(iii),}$$

we have $P(A_4) = P\left(\min_{1\le i\le N}\left|\frac{1}{T}\sum_{t=1}^T V_i^{\mu_i}(w_{it};\beta_i,\mu_i(\beta_i))\right| \ge c_H/2\right) = 1 - o\left(N^{-1}\right)$. It follows that

$$P(A_5) \ge P(A_3\cap A_4) \ge 1 - P(A_3^c) - P(A_4^c) = 1 - o\left(N^{-1}\right).$$

Consequently, we have that by Lemma S1.2(iii),

$$\begin{aligned}
& N\cdot P\left(\max_{1\le i\le N}|\hat{\mu}_i(\beta_i) - \mu_i(\beta_i)| \ge T^{-1/2}(\ln T)^{3+\nu}\right) \\
= \; & N\cdot P\left(\left(\frac{c_H}{4}\right)^{-1}\max_{1\le i\le N}\left|\frac{1}{T}\sum_{t=1}^T V_i(w_{it};\beta_i,\mu_i(\beta_i))\right| \ge T^{-1/2}(\ln T)^{3+\nu},\; A_5\right) + N\cdot P(A_5^c) \\
\le \; & N\cdot P\left(\max_{1\le i\le N}\left|\frac{1}{T}\sum_{t=1}^T V_i(w_{it};\beta_i,\mu_i(\beta_i))\right| \ge c_H T^{-1/2}(\ln T)^{3+\nu}/4\right) + N\cdot P(A_5^c) \\
= \; & o(1) + o(1) = o(1).
\end{aligned}$$

(v) Noting that $|\Psi_i(\beta_i,\hat{\mu}_i(\beta_i)) - \Psi_i(\beta_i,\mu_i(\beta_i))| \le \mathbb{E}[M(w_{it})]\,|\hat{\mu}_i(\beta_i) - \mu_i(\beta_i)| \le c_M^{1/q}|\hat{\mu}_i(\beta_i) - \mu_i(\beta_i)|$ by Assumption A1(iv), the result follows directly from (iv). ■

Recall that $\hat{S}_i \equiv \frac{1}{T}\sum_{t=1}^T U_i\left(w_{it};\beta_i^0,\hat{\mu}_i\left(\beta_i^0\right)\right)$. Let $S_i \equiv \frac{1}{T}\sum_{t=1}^T U_i\left(w_{it};\beta_i^0,\mu_i^0\right)$. The next lemma studies the asymptotic properties of $\hat{S}_i$, S_i, and their difference.

Lemma S1.6 *(i)* $S_i = O_P\left(T^{-1/2}\right)$ *and* $\hat{S}_i - S_i = O_P\left(T^{-1/2}\right)$ *for each* i,

(ii) $\max_{1\le i\le N}\|S_i\| = O_P(T^{-1/2}(\ln T)^3)$ *and* $\max_{1\le i\le N}\left\|\hat{S}_i - S_i\right\| = O_P(T^{-1/2}(\ln T)^3)$,

(iii) $\frac{1}{N}\sum_{i=1}^N\left\|\hat{S}_i\right\|^2 = O_P\left(T^{-1}\right)$,

(iv) $\max_{1\le i\le N}P(\|S_i\| \ge \eta\lambda_1) = o\left(N^{-1}\right)$ *for any given constant* $\eta > 0$,

(v) $P\left(\max_{1\le i\le N}\left\|\hat{S}_i - S_i\right\| \ge CT^{-1/2}(\ln T)^{3+\nu}\right) = o\left(N^{-1}\right)$ *for any* $\nu > 0$ *and* $C > 0$.

Proof. (i) Let ι_p be an arbitrary $p\times 1$ nonrandom vector with $\|\iota_p\| = 1$. Recall that $U_{it} = U_i\left(w_{it};\beta_i^0,\mu_i^0\right)$. Note that $\mathbb{E}(U_{it}) = 0$ and

$$\begin{aligned}
\mathrm{Var}\left(\iota_p' S_i\right) &= \frac{1}{T^2}\sum_{t=1}^T\sum_{s=1}^T \mathrm{Cov}\left(\iota_p' U_{it}, U_{is}'\iota_p\right) \le 8\max_{i,t}\left\{\mathbb{E}\left|\iota_p' U_{it}\right|^q\right\}^{2/q}\frac{1}{T^2}\sum_{t=1}^T\sum_{s=1}^T \alpha(|t-s|)^{1-2/q} \\
&\le 8\max_{i,t}\left\{\mathbb{E}\left|\iota_p' U_{it}\right|^q\right\}^{2/q}\frac{1}{T}\sum_{\tau=1}^\infty \alpha(\tau)^{1-2/q} = O\left(\frac{1}{T}\right)
\end{aligned}$$

by the Davydov inequality (e.g., Corollary A.2 in Hall and Heyde (1980)). Then $S_i = \frac{1}{T}\sum_{t=1}^T U_i\left(w_{it};\beta_i^0,\mu_i^0\right) = O_P\left(T^{-1/2}\right)$ by the Chebyshev inequality. By second order Taylor expansion

$$\begin{aligned}
\hat{S}_i - S_i &= \frac{1}{T}\sum_{t=1}^T\left[U_i\left(w_{it};\beta_i^0,\hat{\mu}_i\left(\beta_i^0\right)\right) - U_i\left(w_{it};\beta_i^0,\mu_i^0\right)\right] \\
&= \frac{1}{T}\sum_{t=1}^T U_i^{\mu_i}\left(w_{it};\beta_i^0,\mu_i\left(\beta_i^0\right)\right)\left[\hat{\mu}_i\left(\beta_i^0\right) - \mu_i\left(\beta_i^0\right)\right] \\
&\quad + \frac{1}{2T}\sum_{t=1}^T U_i^{\mu_i\mu_i}\left(w_{it};\beta_i^0,\breve{\mu}_i\left(\beta_i^0\right)\right)\left[\hat{\mu}_i\left(\beta_i^0\right) - \mu_i\left(\beta_i^0\right)\right]^2,
\end{aligned}$$

where $\breve{\mu}_i\left(\beta_i^0\right)$ lies between $\hat{\mu}_i\left(\beta_i^0\right)$ and $\mu_i\left(\beta_i^0\right)$. By Assumptions A1, Lemma S1.5, and the Markov inequality, one can readily show that the first term is $O_P\left(T^{-1/2}\right)$ and the second is $O_P\left(T^{-1}\right)$. It follows that $\hat{S}_i - S_i = O_P\left(T^{-1/2}\right)$.

(ii) By a simple application of Lemma S1.2(i), $\max_{1\leq i\leq N}\|S_i\| = O_P(T^{-1/2}(\ln T)^3)$. Next,

$$\begin{aligned}
\max_{1\leq i\leq N}\left\|\hat{S}_i - S_i\right\| &\leq \max_{1\leq i\leq N}\frac{1}{T}\sum_{t=1}^{T}M\left(w_{it}\right)\left|\hat{\mu}_i\left(\beta_i^0\right)-\mu_i\left(\beta_i^0\right)\right| \\
&\leq \left\{\max_{1\leq i\leq N}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[M\left(w_{it}\right)\right]+\max_{1\leq i\leq N}\left|\frac{1}{T}\sum_{t=1}^{T}\left\{M\left(w_{it}\right)-\mathbb{E}\left[M\left(w_{it}\right)\right]\right\}\right|\right\} \\
&\quad \times \max_{1\leq i\leq N}\left|\hat{\mu}_i\left(\beta_i^0\right)-\mu_i\left(\beta_i^0\right)\right| \\
&= \left\{O\left(1\right)+o_P\left(1\right)\right\}O_P(T^{-1/2}\left(\ln T\right)^3) = O_P(T^{-1/2}\left(\ln T\right)^3).
\end{aligned}$$

(iii) By the Cauchy-Schwarz inequality, $\frac{1}{N}\sum_{i=1}^{N}\left\|\hat{S}_i\right\|^2 \leq \frac{2}{N}\sum_{i=1}^{N}\|S_i\|^2 + \frac{2}{N}\sum_{i=1}^{N}\left\|\hat{S}_i - S_i\right\|^2$. The first term in $O_P\left(T^{-1}\right)$ by the Markov inequality and the calculation in (i). Using the decomposition of $\hat{S}_i - S_i$ in (i), we can readily show that the second term is $O_P\left(T^{-1}\right)$. Then $\frac{1}{N}\sum_{i=1}^{N}\left\|\hat{S}_i\right\|^2 = O_P\left(T^{-1}\right)$.

(iv) The result follows by a simple application of Lemma S1.2(ii) and Assumption A2.

(v) The proof is similar to that of (ii) but we now apply Lemmas S1.2(iii) and S1.5(iv). ■

The next lemma establishes the uniform consistency of $\hat{\beta}_i$.

Lemma S1.7 *For any $\eta > 0$, we have $P\left(\max_{1\leq i\leq N}\left\|\hat{\beta}_i - \beta_i^0\right\|\right) > \eta) = o\left(N^{-1}\right)$.*

Proof. Recall that $Q_{1NT,\lambda_1}^{(K_0)}\left(\boldsymbol{\beta},\boldsymbol{\alpha}\right) = Q_{1,NT}\left(\boldsymbol{\beta}\right) + \frac{\lambda_1}{N}\sum_{i=1}^{N}\Pi_{k=1}^{K_0}\left\|\beta_i - \alpha_k\right\|$ where $Q_{1,NT}\left(\boldsymbol{\beta}\right) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\psi\left(w_{it};\beta_i,\hat{\mu}_i\left(\beta_i\right)\right) = \frac{1}{N}\sum_{i=1}^{N}\hat{\Psi}_i\left(\beta_i,\hat{\mu}_i\left(\beta_i\right)\right)$. Noting that $(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\alpha}}) = \arg\min_{(\boldsymbol{\beta},\boldsymbol{\alpha})} Q_{1NT,\lambda_1}^{(K_0)}\left(\boldsymbol{\beta},\boldsymbol{\alpha}\right)$, we have $Q_{1NT,\lambda_1}^{(K_0)}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\alpha}}) \leq Q_{1NT,\lambda_1}^{(K_0)}\left(\boldsymbol{\beta}^0,\hat{\boldsymbol{\alpha}}\right)$ and

$$\hat{\Psi}_i\left(\hat{\beta}_i,\hat{\mu}_i(\hat{\beta}_i)\right) + \lambda_1\Pi_{k=1}^{K_0}\left\|\hat{\beta}_i - \hat{\alpha}_k\right\| \leq \hat{\Psi}_i\left(\beta_i^0,\hat{\mu}_i\left(\beta_i^0\right)\right) + \lambda_1\Pi_{k=1}^{K_0}\left\|\beta_i^0 - \hat{\alpha}_k\right\| \text{ for } i = 1,...,N.$$

Let $\varepsilon = \min_i\left[\inf_{\beta_i:\left\|\beta_i-\beta_i^0\right\|>\eta}\Psi_i\left(\beta_i,\mu_i\left(\beta_i\right)\right) - \Psi_i\left(\beta_i^0,\mu_i\left(\beta_i^0\right)\right)\right] > 0$. Define three events $A_1 \equiv \{\max_{1\leq i\leq N}\sup_{(\beta,\mu)}|\hat{\Psi}_i\left(\beta,\mu\right) - \Psi_i\left(\beta,\mu\right)| \leq \frac{1}{6}\varepsilon\}$ and $A_2 \equiv \{\max_{1\leq i\leq N}\sup_{\beta_i}|\Psi_i\left(\beta_i,\hat{\mu}_i\left(\beta_i\right)\right) - \Psi_i\left(\beta_i,\mu_i\left(\beta_i\right)\right)| \leq \frac{1}{6}\varepsilon\}$ and $A_3 \equiv \{\lambda_1\max_{\beta_i:\beta_i\in\mathcal{B}}\Pi_{k=1}^{K_0}\left\|\beta_i - \hat{\alpha}_k\right\| \leq \frac{1}{6}\varepsilon\}$. By Lemmas S1.3, S1.5(v) and Assumption A2(i), $P\left(A_1\cap A_2\cap A_3\right) \geq 1 - P\left(A_1^c\right) - P\left(A_2^c\right) - P\left(A_3^c\right) = 1 - o\left(N^{-1}\right)$. Then conditional on $A_1\cap A_2\cap A_3$, we have uniformly in i

$$\begin{aligned}
&\inf_{\beta_i:\left\|\beta_i-\beta_i^0\right\|>\eta}\hat{\Psi}_i\left(\beta_i,\hat{\mu}_i\left(\beta_i\right)\right) + \lambda_1\Pi_{k=1}^{K_0}\left\|\beta_i - \hat{\alpha}_k\right\| \\
\geq\ &\inf_{\beta_i:\left\|\beta_i-\beta_i^0\right\|>\eta}\Psi_i\left(\beta_i,\hat{\mu}_i\left(\beta_i\right)\right) - \frac{1}{6}\varepsilon + 0 \geq \inf_{\beta_i:\left\|\beta_i-\beta_i^0\right\|>\eta}\Psi_i\left(\beta_i,\mu_i\left(\beta_i\right)\right) - \frac{1}{6}\varepsilon - \frac{1}{6}\varepsilon \\
\geq\ &\Psi_i\left(\beta_i^0,\mu_i\left(\beta_i^0\right)\right) + \varepsilon - \frac{1}{6}\varepsilon - \frac{1}{6}\varepsilon \\
\geq\ &\Psi_i\left(\beta_i^0,\hat{\mu}_i\left(\beta_i^0\right)\right) - \frac{1}{6}\varepsilon + \varepsilon - \frac{1}{6}\varepsilon - \frac{1}{6}\varepsilon \\
\geq\ &\hat{\Psi}_i\left(\beta_i^0,\hat{\mu}_i\left(\beta_i^0\right)\right) - \frac{1}{6}\varepsilon - \frac{1}{6}\varepsilon + \varepsilon - \frac{1}{6}\varepsilon - \frac{1}{6}\varepsilon \\
=\ &\hat{\Psi}_i\left(\beta_i^0,\hat{\mu}_i\left(\beta_i^0\right)\right) + \frac{1}{3}\varepsilon \\
\geq\ &\hat{\Psi}_i\left(\beta_i^0,\hat{\mu}_i\left(\beta_i^0\right)\right) + \lambda_1\Pi_{k=1}^{K_0}\left\|\beta_i^0 - \hat{\alpha}_k\right\| + \frac{1}{6}\varepsilon.
\end{aligned}$$

On the other hand, $\hat{\Psi}_i\left(\hat{\beta}_i, \hat{\mu}_i(\hat{\beta}_i)\right) + \lambda_1 \Pi_{k=1}^{K_0} \left\|\hat{\beta}_i - \hat{\alpha}_k\right\| \leq \hat{\Psi}_i\left(\beta_i^0, \hat{\mu}_i\left(\beta_i^0\right)\right) + \lambda_1 \Pi_{k=1}^{K_0} \left\|\beta_i^0 - \hat{\alpha}_k\right\|$. It follows that $P\left(\max_{1\leq i\leq N} \left\|\hat{\beta}_i - \beta_i^0\right\|\right) > \eta) = o\left(N^{-1}\right)$. ■

To state and prove the next lemma, we follow Hahn and Newey (2004) and introduce some notation. Let F_i and $\hat{F}_i$ denote the cumulative and empirical distribution functions of w_{it}, respectively. Let $F_i(\epsilon) \equiv F_i + \epsilon\sqrt{T}(\hat{F}_i - F_i)$ for $\epsilon \in \left[0, T^{-1/2}\right]$. For fixed β_i and ϵ, let $\mu_i(\beta_i, F_i(\epsilon)) \equiv \arg\min_{\mu_i} \int \psi(\cdot; \beta_i, \mu_i)\, dF_i(\epsilon)$, which is the solution to the estimating equation

$$0 = \int V_i\left(\cdot; \beta_i, \mu_i\left(\beta_i, F_i(\epsilon)\right)\right) dF_i(\epsilon). \tag{S4}$$

Define $\mu_i^{\beta_i}(\epsilon) = \partial\mu_i(\beta_i, F_i(\epsilon))/\partial\beta_i$. Apparently, $F_i(0) = F_i$, $F_i\left(T^{-1/2}\right) = \hat{F}_i$,

$$\begin{aligned}
\mu_i(\beta_i) &= \mu_i(\beta_i, F_i(0)), \\
\hat{\mu}_i(\beta_i) &= \mu_i\left(\beta_i, F_i\left(T^{-1/2}\right)\right), \\
\frac{\partial\mu_i(\beta_i)}{\partial\beta_i} &= \frac{\partial\mu_i(\beta_i, F_i(0))}{\partial\beta_i} = \mu_i^{\beta_i}(0), \text{ and} \\
\frac{\partial\hat{\mu}_i(\beta_i)}{\partial\beta_i} &= \frac{\partial\mu_i\left(\beta_i, F_i\left(T^{-1/2}\right)\right)}{\partial\beta_i} = \mu_i^{\beta_i}\left(T^{-1/2}\right).
\end{aligned}$$

We study the properties of $\mu_i(\beta_i, F_i(\epsilon))$ and $\mu_i^{\beta_i}(\epsilon)$ in the next two lemmas.

Lemma S1.8 *(i)* $P\left(\max_{1\leq i\leq N} \max_{0\leq\epsilon\leq T^{-1/2}} |\mu_i(\beta_i, F_i(\epsilon)) - \mu_i(\beta_i)| \geq \eta\right) = o_P\left(N^{-1}\right)$ *for any* $\eta > 0$,

(ii) $\max_{1\leq i\leq N, \max\|\beta_i - \beta_i^0\| = o(1)} \left|\mu_i(\beta_i) - \mu_i\left(\beta_i^0\right)\right| = o(1)$,

(iii) $P\left(\max_{1\leq i\leq N, \max\|\beta_i - \beta_i^0\| = o(1)} \left|\hat{\mu}_i(\beta_i) - \hat{\mu}_i\left(\beta_i^0\right)\right| \geq \eta\right) = o\left(N^{-1}\right)$ *for any* $\eta > 0$.

Proof. (i) Let $\varepsilon = \min_i\left[\inf_{\mu_i:|\mu_i - \mu_i(\beta_i)|>\eta} \Psi_i(\beta_i, \mu_i) - \Psi_i(\beta_i, \mu_i(\beta_i))\right] > 0$. Noting that

$$\int \psi(\cdot; \beta_i, \mu_i)\, dF_i(\epsilon) = \left(1 - \epsilon\sqrt{T}\right)\Psi_i(\beta_i, \mu_i) + \epsilon\sqrt{T}\hat{\Psi}_i(\beta_i, \mu_i),$$

we have

$$\begin{aligned}
\left|\int \psi(\cdot; \beta_i, \mu_i)\, dF_i(\epsilon) - \Psi_i(\beta_i, \mu_i)\right| &\leq \epsilon\sqrt{T}\left|\hat{\Psi}_i(\beta_i, \mu_i) - \Psi_i(\beta_i, \mu_i)\right| \\
&\leq \left|\hat{\Psi}_i(\beta_i, \mu_i) - \Psi_i(\beta_i, \mu_i)\right|.
\end{aligned}$$

By Lemma S1.3, we have $P[A] = o\left(N^{-1}\right)$ where

$$A \equiv \left\{\max_{0\leq\epsilon\leq T^{-1/2}} \max_{1\leq i\leq N} \left|\int \psi(\cdot; \beta_i, \mu_i)\, dF_i(\epsilon) - \Psi_i(\beta_i, \mu_i)\right| \geq \varepsilon/3\right\}.$$

Therefore for every $\epsilon \in \left[0, T^{-1/2}\right]$ and conditional on the event A, we have

$$\begin{aligned}
\inf_{\mu_i:|\mu_i - \mu_i(\beta_i)|>\eta} \int \psi(\cdot; \beta_i, \mu_i)\, dF_i(\epsilon) &\geq \inf_{\mu_i:|\mu_i - \mu_i(\beta_i)|>\eta} \Psi_i(\beta_i, \mu_i) - \frac{1}{3}\varepsilon \\
&\geq \Psi_i(\beta_i, \mu_i(\beta_i)) + \frac{2}{3}\varepsilon \\
&\geq \int \Psi_i(\beta_i, \mu_i(\beta_i))\, dF_i(\epsilon) + \frac{1}{3}\varepsilon.
\end{aligned}$$

On the other hand, we have $\int \psi\left(\cdot;\beta_i,\mu_i\left(\beta_i,F_i\left(\epsilon\right)\right)\right)dF_i\left(\epsilon\right) \leq \int \psi\left(\cdot;\beta_i,\mu_i\left(\beta_i\right)\right)dF_i\left(\epsilon\right)$ by the definition of $\mu_i\left(\beta_i,F_i\left(\epsilon\right)\right)$. It follows that $P\left(\max_{1\leq i\leq N}\max_{0\leq\epsilon\leq T^{-1/2}}\left|\mu_i\left(\beta_i,F_i\left(\epsilon\right)\right)-\mu_i\left(\beta_i\right)\right|\geq\eta\right)=o\left(N^{-1}\right)$.

(ii) Let $\eta>0$ be given. Let $\varepsilon=\min_i\left[\inf_{\mu_i:\left|\mu_i-\mu_i\left(\beta_i^0\right)\right|>\eta}\Psi_i\left(\beta_i,\mu_i\right)-\Psi_i\left(\beta_i^0,\mu_i\left(\beta_i^0\right)\right)\right]>0$ and $\bar{M}=\max_{i,t}\mathbb{E}\left[M\left(w_{it}\right)\right]$. Note that $\max_i\left|\int\left[\psi\left(\cdot;\beta_i,\mu_i\left(\beta_i^0\right)\right)-\psi\left(\cdot;\beta_i^0,\mu_i\left(\beta_i^0\right)\right)\right]dF_i\right|\leq\bar{M}\max_i\left\|\beta_i-\beta_i^0\right\|=o\left(1\right)$, implying that $\left|\int\left[\psi\left(\cdot;\beta_i,\mu_i\left(\beta_i^0\right)\right)-\psi\left(\cdot;\beta_i^0,\mu_i\left(\beta_i^0\right)\right)\right]dF_i\right|\leq\varepsilon/3$ when $\max_i\left\|\beta_i-\beta_i^0\right\|\leq\varepsilon/(3\bar{M})$. Then for all β_i with $\max\left\|\beta_i-\beta_i^0\right\|\leq\varepsilon/(3\bar{M})$, we have

$$\begin{aligned}\inf_{\mu_i:\left|\mu_i-\mu_i\left(\beta_i^0\right)\right|>\eta}\int\psi\left(\cdot;\beta_i,\mu_i\right)dF_i &\geq \Psi_i\left(\beta_i^0,\mu_i\left(\beta_i^0\right)\right)+\varepsilon\\ &\geq \Psi_i\left(\beta_i,\mu_i\left(\beta_i^0\right)\right)+\frac{2}{3}\varepsilon\\ &= \int\psi_i\left(\cdot;\beta_i,\mu_i\left(\beta_i^0\right)\right)dF_i+\frac{2}{3}\varepsilon.\end{aligned}$$

On the other hand, we have $\int\psi\left(\cdot;\beta_i,\mu_i\left(\beta_i\right)\right)dF_i\leq\int\psi\left(\cdot;\beta_i,\mu_i\left(\beta_i^0\right)\right)dF_i$ by the definition of $\mu_i\left(\beta_i\right)$. It follows that $\max_{1\leq i\leq N,\max\left\|\beta_i-\beta_i^0\right\|=o(1)}\left|\mu_i\left(\beta_i\right)-\mu_i\left(\beta_i^0\right)\right|=o\left(1\right)$.

(iii) By the triangle inequality,

$$\max_{1\leq i\leq N}\left|\hat{\mu}_i\left(\beta_i\right)-\hat{\mu}_i\left(\beta_i^0\right)\right|\leq\max_{1\leq i\leq N}\left|\hat{\mu}_i\left(\beta_i\right)-\mu_i\left(\beta_i\right)\right|+\max_{1\leq i\leq N}\left|\hat{\mu}_i\left(\beta_i^0\right)-\mu_i\left(\beta_i^0\right)\right|+\max_{1\leq i\leq N}\left|\mu_i\left(\beta_i\right)-\mu_i\left(\beta_i^0\right)\right|.$$

By Lemma S1.5(iv), $P\left(\max_{1\leq i\leq N}\left|\hat{\mu}_i\left(\beta_i\right)-\mu_i\left(\beta_i\right)\right|\geq\eta/3\right)=o\left(N^{-1}\right)$. The last term in the above displayed equation is $o\left(1\right)$ uniformly in the set $\max_i\left\|\beta_i-\beta_i^0\right\|=o\left(1\right)$ by (ii). It follows that $P(\max_{1\leq i\leq N,\max\left\|\beta_i-\beta_i^0\right\|=o(1)}$ $\left|\hat{\mu}_i\left(\beta_i\right)-\hat{\mu}_i\left(\beta_i^0\right)\right|\geq\eta)=o\left(N^{-1}\right)$ for any $\eta>0$. ■

Lemma S1.9 *(i)* $P\left(\max_{1\leq i\leq N}\max_{0\leq\epsilon\leq T^{-1/2}}\left\|\frac{\partial\mu_i\left(\beta_i,F_i\left(\epsilon\right)\right)}{\partial\beta_i}-\frac{\partial\mu_i\left(\beta_i\right)}{\partial\beta_i}\right\|\geq\eta\right)=o\left(N^{-1}\right)$ *for any* $\eta>0$,

(ii) $\max_{1\leq i\leq N,\max\left\|\beta_i-\beta_i^0\right\|=o(1)}\left\|\frac{\partial\mu_i\left(\beta_i\right)}{\partial\beta_i}-\frac{\partial\mu_i\left(\beta_i^0\right)}{\partial\beta_i}\right\|=o\left(1\right)$,

(iii) $P\left(\max_{1\leq i\leq N,\max\left\|\beta_i-\beta_i^0\right\|=o(1)}\left\|\frac{\partial\hat{\mu}_i\left(\beta_i\right)}{\partial\beta_i}-\frac{\partial\hat{\mu}_i\left(\beta_i^0\right)}{\partial\beta_i}\right\|\right)=o\left(N^{-1}\right)$ *for any* $\eta>0$.

Proof. (i) Differentiating both sides of (S4) with respect to β_i yields

$$0=\int V_i^{\beta_i}\left(\cdot;\beta_i,\mu_i\left(\beta_i,F_i\left(\epsilon\right)\right)\right)dF_i\left(\epsilon\right)+\int V_i^{\mu_i}\left(\cdot;\beta_i,\mu_i\left(\beta_i,F_i\left(\epsilon\right)\right)\right)dF_i\left(\epsilon\right)\frac{\partial\mu_i\left(\beta_i,F_i\left(\epsilon\right)\right)}{\partial\beta_i}.$$

It follows that

$$\mu_i^{\beta_i}\left(\epsilon\right)\equiv\frac{\partial\mu_i\left(\beta_i,F_i\left(\epsilon\right)\right)}{\partial\beta_i}=-\frac{\int V_i^{\beta_i}\left(\cdot;\beta_i,\mu_i\left(\beta_i,F_i\left(\epsilon\right)\right)\right)dF_i\left(\epsilon\right)}{\int V_i^{\mu_i}\left(\cdot;\beta_i,\mu_i\left(\beta_i,F_i\left(\epsilon\right)\right)\right)dF_i\left(\epsilon\right)}. \tag{S5}$$

Noting that $\int V_i^{\mu_i}\left(\cdot;\beta_i,\mu_i\left(\beta_i\right)\right)dF_i=H_{i\mu\mu}\left(\beta_i\right)>c_H>0$ uniformly in i by Assumption A1(v), it suffices to show that

$$P\left(\max_{1\leq i\leq N}\max_{0\leq\epsilon\leq T^{-1/2}}\left\|\int V_i^{\beta_i}\left(\cdot;\beta_i,\mu_i\left(\beta_i,F_i\left(\epsilon\right)\right)\right)dF_i\left(\epsilon\right)-\int V_i^{\beta_i}\left(\cdot;\beta_i,\mu_i\left(\beta_i\right)\right)dF_i\right\|\geq\eta/2\right)=o\left(N^{-1}\right), \tag{S6}$$

and

$$P\left(\max_{1\leq i\leq N}\max_{0\leq\epsilon\leq T^{-1/2}}\left|\int V_i^{\mu_i}\left(\cdot;\beta_i,\mu_i\left(\beta_i,F_i\left(\epsilon\right)\right)\right)dF_i\left(\epsilon\right)-\int V_i^{\mu_i}\left(\cdot;\beta_i,\mu_i\left(\beta_i\right)\right)dF_i\right|\geq\eta/2\right)=o\left(N^{-1}\right). \tag{S7}$$

By the triangle inequality,

$$\begin{aligned}
&\left\| \int V_i^{\beta_i}\left(\cdot;\beta_i,\mu_i\left(\beta_i,F_i\left(\epsilon\right)\right)\right)dF_i\left(\epsilon\right) - \int V_i^{\beta_i}\left(\cdot;\beta_i,\mu_i\left(\beta_i\right)\right)dF_i \right\| \\
\leq\ & \left\| \int \left[V_i^{\beta_i}\left(\cdot;\beta_i,\mu_i\left(\beta_i,F_i\left(\epsilon\right)\right)\right) - V_i^{\beta_i}\left(\cdot;\beta_i,\mu_i\left(\beta_i\right)\right)\right]dF_i\left(\epsilon\right) \right\| + \left\| \int V_i^{\beta_i}\left(\cdot;\beta_i,\mu_i\left(\beta_i\right)\right)d\left[F_i\left(\epsilon\right)-F_i\right] \right\| \\
=\ & \left\| \int \left[V_i^{\beta_i}\left(\cdot;\beta_i,\mu_i\left(\beta_i,F_i\left(\epsilon\right)\right)\right) - V_i^{\beta_i}\left(\cdot;\beta_i,\mu_i\left(\beta_i\right)\right)\right]dF_i\left(\epsilon\right) \right\| + \epsilon\sqrt{T} \left\| \int V_i^{\beta_i}\left(\cdot;\beta_i,\mu_i\left(\beta_i\right)\right)d\left[\hat{F}_i-F_i\right] \right\|.
\end{aligned}$$

Using Lemma S1.2(iii), we have

$$P\left(\max_{1\leq i\leq N} \max_{0\leq\epsilon\leq T^{-1/2}} \epsilon\sqrt{T} \left\| \int V_i^{\beta_i}\left(\cdot;\beta_i,\mu_i\left(\beta_i\right)\right)d\left[\hat{F}_i-F_i\right] \right\| \geq \eta/4 \right) = o\left(N^{-1}\right).$$

In addition, by Lemma S1.8(i),

$$\begin{aligned}
& P\left(\max_{1\leq i\leq N} \max_{0\leq\epsilon\leq T^{-1/2}} \left\| \int \left[V_i^{\beta_i}\left(\cdot;\beta_i,\mu_i\left(\beta_i,F_i\left(\epsilon\right)\right)\right) - V_i^{\beta_i}\left(\cdot;\beta_i,\mu_i\left(\beta_i\right)\right)\right]dF_i\left(\epsilon\right) \right\| \geq \eta/4 \right) \\
\leq\ & P\left(\max_{1\leq i\leq N} \int M\left(\cdot\right)dF_i\left(\epsilon\right) \max_{1\leq i\leq N} \max_{0\leq\epsilon\leq T^{-1/2}} \left|\mu_i\left(\beta_i,F_i\left(\epsilon\right)\right)-\mu_i\left(\beta_i\right)\right| \geq \eta/4 \right) = o\left(N^{-1}\right).
\end{aligned}$$

Then (S6) follows. Analogously we can prove (S7).

(ii) Recall that

$$\frac{\partial\mu_i\left(\beta_i\right)}{\partial\beta_i} = -\frac{\int V_i^{\beta_i}\left(\cdot;\beta_i,\mu_i\left(\beta_i\right)\right)dF_i}{\int V_i^{\mu_i}\left(\cdot;\beta_i,\mu_i\left(\beta_i\right)\right)dF_i}. \tag{S8}$$

To prove (ii), it suffices to show that

$$\max_{1\leq i\leq N,\max\left\|\beta_i-\beta_i^0\right\|=o(1)} \left\| \int V_i^{\beta_i}\left(\cdot;\beta_i,\mu_i\left(\beta_i\right)\right)dF_i - \int V_i^{\beta_i}\left(\cdot;\beta_i^0,\mu_i\left(\beta_i^0\right)\right)dF_i \right\| = o\left(1\right),$$

and

$$\max_{1\leq i\leq N,\max\left\|\beta_i-\beta_i^0\right\|=o(1)} \left\| \int V_i^{\mu_i}\left(\cdot;\beta_i,\mu_i\left(\beta_i\right)\right)dF_i - \int V_i^{\mu_i}\left(\cdot;\beta_i^0,\mu_i\left(\beta_i^0\right)\right)dF_i \right\| = o\left(1\right).$$

We only show the first result as the proof of the second one is similar. By Assumption A1(iv) and Lemma S1.8(ii),

$$\begin{aligned}
& \max_{1\leq i\leq N,\max\left\|\beta_i-\beta_i^0\right\|=o(1)} \left\| \int V_i^{\beta_i}\left(\cdot;\beta_i,\mu_i\left(\beta_i\right)\right)dF_i - \int V_i^{\beta_i}\left(\cdot;\beta_i^0,\mu_i\left(\beta_i^0\right)\right)dF_i \right\| \\
\leq\ & \max_{i,t}\mathbb{E}\left[M\left(w_{it}\right)\right] \max_{1\leq i\leq N,\max\left\|\beta_i-\beta_i^0\right\|=o(1)} \left\{ \left\|\beta_i-\beta_i^0\right\| + \left|\mu_i\left(\beta_i\right)-\mu_i\left(\beta_i^0\right)\right| \right\} = o\left(1\right).
\end{aligned}$$

(iii) By the triangle inequality,

$$\begin{aligned}
\max_{1\leq i\leq N} \left\| \frac{\partial\hat{\mu}_i\left(\beta_i\right)}{\partial\beta_i} - \frac{\partial\hat{\mu}_i\left(\beta_i^0\right)}{\partial\beta_i} \right\| \leq\ & \max_{1\leq i\leq N} \left\| \frac{\partial\hat{\mu}_i\left(\beta_i\right)}{\partial\beta_i} - \frac{\partial\mu_i\left(\beta_i\right)}{\partial\beta_i} \right\| + \max_{1\leq i\leq N} \left\| \frac{\partial\hat{\mu}_i\left(\beta_i^0\right)}{\partial\beta_i} - \frac{\partial\mu_i\left(\beta_i^0\right)}{\partial\beta_i} \right\| \\
& + \max_{1\leq i\leq N} \left\| \frac{\partial\mu_i\left(\beta_i\right)}{\partial\beta_i} - \frac{\partial\mu_i\left(\beta_i^0\right)}{\partial\beta_i} \right\|.
\end{aligned}$$

Noting that $P\left(\max_{1\leq i\leq N}\left\|\frac{\partial\hat{\mu}_i(\beta_i)}{\partial\beta_i} - \frac{\partial\mu_i(\beta_i)}{\partial\beta_i}\right\| \geq \eta/3\right) = o\left(N^{-1}\right)$ by (i) and the last term in the above displayed equation is $o\left(1\right)$ uniformly in the set $\max_i\left\|\beta_i-\beta_i^0\right\| = o_P\left(1\right)$ by (ii), we have $P\Big(\max_{1\leq i\leq N,\max\left\|\beta_i-\beta_i^0\right\|=o(1)}$ $\left\|\frac{\partial\hat{\mu}_i(\beta_i)}{\partial\beta_i} - \frac{\partial\hat{\mu}_i\left(\beta_i^0\right)}{\partial\beta_i}\right\| \geq \eta\Big) = o\left(N^{-1}\right)$for any $\eta > 0$. ■

Recall from (A.2) that

$$\hat{H}_{i\beta\beta}\left(\beta_i\right)=\frac{1}{T}\sum_{t=1}^{T}\left[U_i^{\beta_i}\left(w_{it};\beta_i,\hat{\mu}_i\left(\beta_i\right)\right)+U_i^{\mu_i}\left(w_{it};\beta_i,\hat{\mu}_i\left(\beta_i\right)\right)\frac{\partial\hat{\mu}_i\left(\beta_i\right)}{\partial\beta_i'}\right].$$

Let $\check{H}_{i\beta\beta}\left(\beta_i\right)=\frac{1}{T}\sum_{t=1}^{T}[U_i^{\beta_i}\left(w_{it};\beta_i,\mu_i\left(\beta_i\right)\right)+U_i^{\mu_i}\left(w_{it};\beta_i,\mu_i\left(\beta_i\right)\right)\frac{\partial\mu_i(\beta_i)}{\partial\beta_i'}]$. Note that $H_{i\beta\beta}\left(\beta_i\right)=\mathbb{E}[\check{H}_{i\beta\beta}\left(\beta_i\right)]$, where $H_{i\beta\beta}\left(\cdot\right)$ is defined in Section 2.3. The next lemma study the asymptotics of $\hat{H}_{i\beta\beta}\left(\beta_i\right)$.

Lemma S1.10 *(i)* $P\left(\max_{1\le i\le N}\left\|\hat{H}_{i\beta\beta}\left(\breve{\beta}_i\right)-H_{i\beta\beta}\left(\beta_i^0\right)\right\|\ge\eta\right)=o\left(N^{-1}\right).$
(ii) $c_{\hat{H}}\equiv\min_{1\le i\le N}\mu_{\min}\left(\hat{H}_{i\beta\beta}(\breve{\beta}_i)\right)\ge c_H-o_P\left(1\right).$

Proof. (i) By the triangle inequality,

$$\begin{aligned}&\max_{1\le i\le N}\left\|\hat{H}_{i\beta\beta}\left(\breve{\beta}_i\right)-H_{i\beta\beta}\left(\beta_i^0\right)\right\|\\ \le\;&\max_{1\le i\le N}\left\|\hat{H}_{i\beta\beta}\left(\breve{\beta}_i\right)-\hat{H}_{i\beta\beta}\left(\beta_i^0\right)\right\|+\max_{1\le i\le N}\left\|\hat{H}_{i\beta\beta}\left(\beta_i^0\right)-\check{H}_{i\beta\beta}\left(\beta_i^0\right)\right\|+\max_{1\le i\le N}\left\|\check{H}_{i\beta\beta}\left(\beta_i^0\right)-H_{i\beta\beta}\left(\beta_i^0\right)\right\|.\end{aligned}$$

We prove (i) by showing that (i1) $P(\max_{1\le i\le N}||\hat{H}_{i\beta\beta}(\breve{\beta}_i)-\hat{H}_{i\beta\beta}\left(\beta_i^0\right)||\ge\eta/3)=o\left(N^{-1}\right)$, (i2) $P(\max_{1\le i\le N}||\hat{H}_{i\beta\beta}\left(\beta_i^0\right)-\check{H}_{i\beta\beta}\left(\beta_i^0\right)||\ge\eta/3)=o\left(N^{-1}\right)$, and (i3) $P(\max_{1\le i\le N}||\check{H}_{i\beta\beta}\left(\beta_i^0\right)-H_{i\beta\beta}\left(\beta_i^0\right)||\ge\eta/3)=o\left(N^{-1}\right)$. For (i1), we make the following decomposition:

$$\begin{aligned}\hat{H}_{i\beta\beta}\left(\breve{\beta}_i\right)-\hat{H}_{i\beta\beta}\left(\beta_i^0\right)&=\frac{1}{T}\sum_{t=1}^{T}\left[U_i^{\beta_i}\left(w_{it};\breve{\beta}_i,\hat{\mu}_i\left(\breve{\beta}_i\right)\right)-U_i^{\beta_i}\left(w_{it};\beta_i^0,\hat{\mu}_i\left(\beta_i^0\right)\right)\right]\\&\quad+\frac{1}{T}\sum_{t=1}^{T}\left[U_i^{\mu_i}\left(w_{it};\breve{\beta}_i,\hat{\mu}_i\left(\breve{\beta}_i\right)\right)\frac{\partial\hat{\mu}_i\left(\breve{\beta}_i\right)}{\partial\beta_i'}-U_i^{\mu_i}\left(w_{it};\beta_i^0,\hat{\mu}_i\left(\beta_i^0\right)\right)\frac{\partial\hat{\mu}_i\left(\beta_i^0\right)}{\partial\beta_i'}\right]\\&\equiv H_{11i}+H_{12i},\text{ say.}\end{aligned}$$

For H_{11i}, we have

$$\max_{1\le i\le N}\left\|H_{11i}\right\|\le\max_{1\le i\le N}\frac{1}{T}\sum_{t=1}^{T}M\left(w_{it}\right)\left\{\left\|\breve{\beta}_i-\beta_i^0\right\|+\left\|\hat{\mu}_i\left(\breve{\beta}_i\right)-\hat{\mu}_i\left(\beta_i^0\right)\right\|\right\}.$$

Using the arguments as used in the proof of Lemma S1.5(iv), we can show that

$$P\left(\max_{1\le i\le N}\frac{1}{T}\sum_{t=1}^{T}M\left(w_{it}\right)\le 2c_M^{1/q}\right)=1-o\left(N^{-1}\right).$$

Then by Lemmas S1.7 and S1.8(iii), we can readily show that $P\left(\max_{1\le i\le N}\left\|H_{11i}\right\|\ge\eta/6\right)=o\left(N^{-1}\right)$. For H_{12i}, we make the following decomposition:

$$\begin{aligned}H_{12i}&=\frac{1}{T}\sum_{t=1}^{T}\left[U_i^{\mu_i}\left(w_{it};\breve{\beta}_i,\hat{\mu}_i\left(\breve{\beta}_i\right)\right)-U_i^{\mu_i}\left(w_{it};\beta_i^0,\hat{\mu}_i\left(\beta_i^0\right)\right)\right]\frac{\partial\hat{\mu}_i\left(\breve{\beta}_i\right)}{\partial\beta_i'}\\&\quad+\frac{1}{T}\sum_{t=1}^{T}U_i^{\mu_i}\left(w_{it};\beta_i^0,\hat{\mu}_i\left(\beta_i^0\right)\right)\left[\frac{\partial\hat{\mu}_i\left(\breve{\beta}_i\right)}{\partial\beta_i'}-\frac{\partial\hat{\mu}_i\left(\beta_i^0\right)}{\partial\beta_i'}\right]\\&\equiv H_{12i,1}+H_{12i,2},\text{ say.}\end{aligned}$$

Following the analysis of H_{11i} and applying Lemmas S1.8(i) and (iii) and Lemmas S1.9(i) and (iii), we can readily show that $P\left(\max_{1\leq i\leq N}\|H_{12i,s}\|\geq\eta/12\right)=o\left(N^{-1}\right)$ for $s=1,2$. Then $P\left(\max_{1\leq i\leq N}\|H_{12i}\|\geq\eta/6\right)=o\left(N^{-1}\right)$. Consequently, we have $P\left(\max_{1\leq i\leq N}||\hat{H}_{i\beta\beta}\left(\breve{\beta}_i\right)-\hat{H}_{i\beta\beta}\left(\beta_i^0\right)||\geq\eta/3\right)=o\left(N^{-1}\right)$.

To prove (i2), we make the following decomposition:

$$\begin{aligned}\hat{H}_{i\beta\beta}\left(\beta_i^0\right)-\check{H}_{i\beta\beta}\left(\beta_i^0\right) &= \frac{1}{T}\sum_{t=1}^{T}\left[U_i^{\beta_i}\left(w_{it};\beta_i^0,\hat{\mu}_i\left(\beta_i^0\right)\right)-U_i^{\beta_i}\left(w_{it};\beta_i^0,\mu_i\left(\beta_i^0\right)\right)\right]\\ &\quad+\frac{1}{T}\sum_{t=1}^{T}\left[U_i^{\mu_i}\left(w_{it};\beta_i^0,\hat{\mu}_i\left(\beta_i^0\right)\right)\frac{\partial\hat{\mu}_i\left(\beta_i^0\right)}{\partial\beta_i'}-U_i^{\mu_i}\left(w_{it};\beta_i^0,\mu_i\left(\beta_i^0\right)\right)\frac{\partial\mu_i\left(\beta_i^0\right)}{\partial\beta_i'}\right]\\ &\equiv H_{21i}+H_{22i}.\end{aligned}$$

Following the analysis of $\max_{1\leq i\leq N}||\hat{H}_{i\beta\beta}(\breve{\beta}_i)-\hat{H}_{i\beta\beta}\left(\beta_i^0\right)||$ and using Lemmas S1.2, S1.7, S1.8, and S1.9 and Assumption A1, we can show $P\left(\max_{1\leq i\leq N}\|H_{2si}\|\geq\eta/6\right)=o\left(N^{-1}\right)$ for $s=1,2$. Then (i2) holds.

Next,

$$\begin{aligned}\check{H}_{i\beta\beta}\left(\beta_i^0\right)-H_{i\beta\beta}\left(\beta_i^0\right) &= \frac{1}{T}\sum_{t=1}^{T}\left\{U_i^{\beta_i}\left(w_{it};\beta_i^0,\mu_i\left(\beta_i^0\right)\right)-\mathbb{E}\left[U_i^{\beta_i}\left(w_{it};\beta_i^0,\mu_i\left(\beta_i^0\right)\right)\right]\right\}\\ &\quad+\frac{1}{T}\sum_{t=1}^{T}\left\{U_i^{\mu_i}\left(w_{it};\beta_i^0,\mu_i\left(\beta_i^0\right)\right)-\mathbb{E}\left[U_i^{\mu_i}\left(w_{it};\beta_i^0,\mu_i\left(\beta_i^0\right)\right)\right]\right\}\frac{\partial\mu_i\left(\beta_i^0\right)}{\partial\beta_i'}\\ &\equiv H_{31i}+H_{32i},\ \text{say.}\end{aligned}$$

Using Lemma S1.2, we can show $P\left(\max_{1\leq i\leq N}\|H_{3si}\|\geq\eta/6\right)=o\left(N^{-1}\right)$ for $s=1,2$. Then (i3) holds. This completes the proof of (i).

(ii) By the Weyl inequality and the fact that $|\mu_{\max}(A)|\leq\|A\|$ for any symmetric matrix A, we have $\mu_{\min}(\hat{H}_{i\beta\beta}(\breve{\beta}_i))\geq\mu_{\min}(H_{i0}\left(\beta_i^0\right))-||\hat{H}_{i\beta\beta}(\breve{\beta}_i)-H_{i\beta\beta}\left(\beta_i^0\right)||$. Then by (i) and Assumption A1(v), $c_{\hat{H}}\equiv\min_{1\leq i\leq N}\mu_{\min}(\hat{H}_{i\beta\beta}(\breve{\beta}_i))\geq\min_{1\leq i\leq N}\mu_{\min}(H_{i\beta\beta}\left(\beta_i^0\right))-\max_{1\leq i\leq N}||\hat{H}_{i\beta\beta}(\breve{\beta}_i)-H_{i\beta\beta}\left(\beta_i^0\right)||\geq c_H-o_P(1)$. ■

Lemma S1.11 *Recall that* $\bar{H}_{i\beta\beta}\equiv\hat{H}_{i\beta\beta}\left(\bar{\beta}_i\right)$ *where* $\bar{\beta}_i$ *lies between* $\hat{\beta}_i$ *and* β_i^0 *elementwise. Then*
(i) $P\left(\max_{1\leq i\leq N}\left\|\bar{H}_{i\beta\beta}-H_{i\beta\beta}\left(\beta_i^0\right)\right\|\geq\eta\right)=o\left(N^{-1}\right)$ *for any* $\eta>0$,
(ii) $\max_{1\leq i\leq N}\left\|\bar{H}_{i\beta\beta}\right\|=O_P(1)$.

Proof. (i) The proof is identical to that of Lemma S1.10(i) with $\breve{\beta}_i$ replaced by $\bar{\beta}_i$.
(ii) By (i) and the triangle inequality, $\max_{1\leq i\leq N}\left\|\bar{H}_{i\beta\beta}\right\|\leq\max_{1\leq i\leq N}\left\|H_{i\beta\beta}\left(\beta_i^0\right)\right\|+\max_{1\leq i\leq N}||\bar{H}_{i\beta\beta}-H_{i\beta\beta}\left(\beta_i^0\right)||=O(1)+o_P(1)=O_P(1)$. ■

Recall that $U_{it}=U_i\left(w_{it};\beta_i^0,\mu_i^0\right)$, $U_{it}^{\mu_i}=U_i^{\mu_i}\left(w_{it};\beta_i^0,\mu_i^0\right)$, $U_{it}^{\mu_i\mu_i}=U_i^{\mu_i\mu_i}\left(w_{it};\beta_i^0,\mu_i^0\right)$, and similarly for V_{it} and $V_{it}^{\mu_i}$. Recall that $m_{iU}=\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left(U_{it}^{\mu_i}\right)$, $m_{iV}=\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left(V_{it}^{\mu_i}\right)$, $m_{iU2}=\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left(U_{it}^{\mu_i\mu_i}\right)$, $m_{iV2}=\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left(V_{it}^{\mu_i\mu_i}\right)$, and $\mathbb{U}_{it}=U_{it}-\frac{m_{iU}}{m_{iV}}V_{it}$. The next two lemmas are essential to establish the asymptotic distribution of the C-Lasso and post-Lasso estimators.

Lemma S1.12 *Let* $\hat{S}_{\hat{G}_k}\equiv\frac{1}{\sqrt{N_kT}}\sum_{i\in\hat{G}_k}\sum_{t=1}^{T}U_i\left(w_{it};\alpha_k^0,\hat{\mu}_i(\alpha_k^0)\right)$. *Then* $\hat{S}_{\hat{G}_k}+\mathbb{B}_{kNT}\xrightarrow{D}N\left(0,\Omega_k\right)$.

Proof. Let $\hat{S}_{G_k^0}\equiv\frac{1}{\sqrt{N_kT}}\sum_{i\in G_k^0}\sum_{t=1}^{T}U_i\left(w_{it};\alpha_k^0,\hat{\mu}_i(\alpha_k^0)\right)$. 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

$$\begin{aligned}\hat{S}_{\hat{G}_k}-\hat{S}_{G_k^0} &= \frac{1}{\sqrt{N_kT}}\sum_{i\in\hat{G}_k\backslash G_k^0}\sum_{t=1}^{T}U_i\left(w_{it};\alpha_k^0,\hat{\mu}_i(\alpha_k^0)\right)-\frac{1}{\sqrt{N_kT}}\sum_{i\in G_k^0\backslash\hat{G}_k}\sum_{t=1}^{T}U_i\left(w_{it};\alpha_k^0,\hat{\mu}_i(\alpha_k^0)\right)\\ &\equiv \hat{S}_{k,1}-\hat{S}_{k,2},\ \text{say.}\end{aligned}$$

Let $\epsilon > 0$ be an arbitrary constant. By Theorem 2.2, $P\left(\left\|\hat{S}_{k,1}\right\| \geq \epsilon\right) \leq P(\hat{F}_{kNT}) \to 0$, and $P\left(\left\|\hat{S}_{k,2}\right\| \geq \epsilon\right) \leq P(\hat{E}_{kNT}) \to 0$. Thus $\hat{S}_{\hat{G}_k} = \hat{S}_{G_k^0} + o_P(1)$ and it suffices to prove the lemma by showing that (i) $\hat{S}_{G_k^0} + \mathbb{B}_{kNT} = \frac{1}{\sqrt{N_k T}} \sum_{i \in G_k^0} \sum_{t=1}^T \mathbb{U}_{it} + o_P(1)$, and (ii) $\frac{1}{\sqrt{N_k T}} \sum_{i \in G_k^0} \sum_{t=1}^T \mathbb{U}_{it} \xrightarrow{D} N(0, \Omega_k)$.

Part (i): We prove $\hat{S}_{G_k^0} + \mathbb{B}_{kNT} = \frac{1}{\sqrt{N_k T}} \sum_{i \in G_k^0} \sum_{t=1}^T \mathbb{U}_{it} + o_P(1)$. By second order Taylor expansion,

$$\begin{aligned} \hat{S}_{G_k^0} &= \frac{1}{\sqrt{N_k T}} \sum_{i \in G_k^0} \sum_{t=1}^T U_{it} + \frac{1}{\sqrt{N_k T}} \sum_{i \in G_k^0} \sum_{t=1}^T U_{it}^{\mu_i} \left[\hat{\mu}_i(\alpha_k^0) - \mu_i^0\right] \\ &\quad + \frac{1}{2\sqrt{N_k T}} \sum_{i \in G_k^0} \sum_{t=1}^T U_i^{\mu_i \mu_i}\left(w_{it}; \alpha_k^0, \mu_i^*\right) \left[\hat{\mu}_i(\alpha_k^0) - \mu_i^0\right]^2 \\ &\equiv S_{k,1} + S_{k,2} + S_{k,3}, \text{ say,} \end{aligned} \tag{S9}$$

where μ_i^* lies between $\hat{\mu}_i(\alpha_k^0)$ and μ_i^0. We will show that $S_{k,1}$ contributes to the asymptotic variance of $\hat{S}_{G_k^0}$, $S_{k,3}$ contributes to the asymptotic bias, and $S_{k,2}$ contributes to both. We analyze $S_{k,3}$ first. Let $S_{k,3}^0 = \frac{1}{2\sqrt{N_k T}} \sum_{i \in G_k^0} \sum_{t=1}^T U_{it}^{\mu_i \mu_i} \left[\hat{\mu}_i(\alpha_k^0) - \mu_i^0\right]^2$. By Assumption A1, the Markov inequality, and Lemma S1.5(ii), we have

$$\begin{aligned} \left\|S_{k,3} - S_{k,3}^0\right\| &= \frac{1}{2\sqrt{N_k T}} \sum_{i \in G_k^0} \sum_{t=1}^T \left\|U_i^{\mu_i \mu_i}\left(w_{it}; \alpha_k^0, \mu_i^*\right) - U_i^{\mu_i \mu_i}\left(w_{it}; \alpha_k^0, \mu_i^0\right)\right\| \left[\hat{\mu}_i(\alpha_k^0) - \mu_i^0\right]^2 \\ &\leq \left\{\frac{1}{2N_k T} \sum_{i \in G_k^0} \sum_{t=1}^T M(w_{it})\right\} \sqrt{N_k T} \left|\hat{\mu}_i(\alpha_k^0) - \mu_i^0\right|^3 \\ &= O_P(1) \sqrt{N_k T} O_P\left(T^{-3/2} (\ln T)^9\right) = O_P\left(N_k^{1/2} T^{-1} (\ln T)^9\right) = o_P(1). \end{aligned}$$

By (S1) in the proof of Lemma S1.5,

$$S_{k,3}^0 = \frac{1}{2\sqrt{N_k T}} \sum_{i \in G_k^0} \sum_{t=1}^T U_{it}^{\mu_i \mu_i} \left(\frac{\frac{1}{T}\sum_{t=1}^T V_{it}}{\frac{1}{T}\sum_{t=1}^T V_i^{\mu_i}\left(w_{it}; \alpha_k^0, \breve{\mu}_i(\alpha_k^0)\right)}\right)^2.$$

As in the analysis of $S_{k,3} - S_{k,3}^0$, by Lemmas S1.5(ii) and S1.2(i) and the fact that $\max_{1 \leq i \leq N} \left|\frac{1}{T}\sum_{t=1}^T V_{it}\right| = O_P(T^{-1/2} (\ln T)^3)$, we can readily show that $S_{k,3}^0 = S_{k,3}^{00} + O_P(N_k^{1/2} T^{-1} (\ln T)^9) = S_{k,3}^{00} + o_P(1)$, where

$$\begin{aligned} S_{k,3}^{00} &\equiv \frac{1}{2\sqrt{N_k T}} \sum_{i \in G_k^0} \sum_{t=1}^T U_{it}^{\mu_i \mu_i} \left(\frac{\frac{1}{T}\sum_{t=1}^T V_{it}}{\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[V_{it}^{\mu_i}\right]}\right)^2 \\ &= \frac{1}{2\sqrt{N_k T}} \sum_{i \in G_k^0} m_{iU2} m_{iV}^{-2} \left(\frac{1}{\sqrt{T}} \sum_{t=1}^T V_{it}\right)^2 + O_P(N_k^{1/2} T^{-1} (\ln T)^9), \end{aligned}$$

where we also use the fact that $\max_{1 \leq i \leq N} \left\|\frac{1}{T}\sum_{t=1}^T U_{it}^{\mu_i \mu_i} - m_{iU2}\right\| = O_P(T^{-1/2} (\ln T)^3$ by Lemma S1.2(i). Thus, we have

$$S_{k,3} = \frac{1}{2\sqrt{N_k T}} \sum_{i \in G_k^0} m_{iU2} m_{iV}^{-2} \left(\frac{1}{\sqrt{T}} \sum_{t=1}^T V_{it}\right)^2 + o_P(1). \tag{S10}$$

Now, we study $S_{k,2}$. By Lemma S1.5(ii), (S1) in its proof, and the fact that $\max_{1\le i\le N}\left|\frac{1}{T}\sum_{t=1}^{T}V_{it}\right| = O_P\left(T^{-1/2}(\ln T)^3\right)$ and $\max_{1\le i\le N}\left|\frac{1}{T}\sum_{t=1}^{T}V_{it}^{\mu_i}-m_{iV}\right| = O_P\left(T^{-1/2}(\ln T)^3\right)$, we have

$$\begin{aligned}\hat{\mu}_i(\alpha_k^0)-\mu_i^0 &= -\frac{\frac{1}{T}\sum_{t=1}^{T}V_{it}}{\frac{1}{T}\sum_{t=1}^{T}V_i^{\mu_i}(w_{it};\alpha_k^0,\check{\mu}_i(\alpha_k^0))} = -\frac{\frac{1}{T}\sum_{t=1}^{T}V_{it}}{\frac{1}{T}\sum_{t=1}^{T}V_{it}^{\mu_i}}+O_P\left(T^{-1}(\ln T)^6\right)\\ &= -m_{iV}^{-1}\frac{1}{T}\sum_{t=1}^{T}V_{it}+O_P\left(T^{-1}(\ln T)^6\right) \text{ uniformly in } i\in G_k^0.\end{aligned}$$

But the above expansion is not sufficient to study $S_{k,2}$ and we need to get better control on the remainder term. Noting that $\hat{\mu}_i(\beta_i^0)=\arg\min_{\mu_i}\frac{1}{T}\sum_{t=1}^{T}\psi(w_{it};\beta_i^0,\mu_i)$, we have

$$\begin{aligned}0 &= \frac{1}{T}\sum_{t=1}^{T}V_i\left(w_{it};\beta_i^0,\hat{\mu}_i(\beta_i^0)\right)\\ &= \frac{1}{T}\sum_{t=1}^{T}V_{it}+\frac{1}{T}\sum_{t=1}^{T}V_{it}^{\mu_i}\left[\hat{\mu}_i(\beta_i^0)-\mu_i(\beta_i^0)\right]+\frac{1}{2T}\sum_{t=1}^{T}V_i^{\mu_i\mu_i}\left(w_{it};\beta_i^0,\check{\mu}_i(\beta_i^0)\right)\left[\hat{\mu}_i(\beta_i^0)-\mu_i(\beta_i^0)\right]^2,\end{aligned}$$

where $\check{\mu}_i(\beta_i^0)$ lies between $\hat{\mu}_i(\beta_i^0)$ and $\mu_i(\beta_i^0)$ for each i. It follows that

$$\begin{aligned}&\hat{\mu}_i(\beta_i^0)-\mu_{i0}(\beta_i^0)\\ &= -\left[\frac{1}{T}\sum_{t=1}^{T}V_{it}^{\mu_i}\right]^{-1}\left\{\frac{1}{T}\sum_{t=1}^{T}V_{it}+\frac{1}{2T}\sum_{t=1}^{T}V_i^{\mu_i\mu_i}\left(w_{it};\beta_i^0,\check{\mu}_i(\beta_i^0)\right)\left[\hat{\mu}_i(\beta_i^0)-\mu_i(\beta_i^0)\right]^2\right\}\\ &= -\left[\frac{1}{T}\sum_{t=1}^{T}V_{it}^{\mu_i}\right]^{-1}\left\{\frac{1}{T}\sum_{t=1}^{T}V_{it}+\frac{1}{2T}\sum_{t=1}^{T}V_{it}^{\mu_i\mu_i}\left[\hat{\mu}_i(\beta_i^0)-\mu_i(\beta_i^0)\right]^2\right\}+O_P\left(T^{-3}(\ln T)^9\right)\\ &= -\left[\frac{1}{T}\sum_{t=1}^{T}V_{it}^{\mu_i}\right]^{-1}\left\{\frac{1}{T}\sum_{t=1}^{T}V_{it}+\frac{1}{2}m_{iV}^{-2}\frac{1}{T}\sum_{t=1}^{T}V_{it}^{\mu_i\mu_i}\left(\frac{1}{T}\sum_{t=1}^{T}V_{it}\right)^2\right\}+O_P\left(T^{-3}(\ln T)^9\right)\\ &= -\left[\frac{1}{T}\sum_{t=1}^{T}V_{it}^{\mu_i}\right]^{-1}\left\{\frac{1}{T}\sum_{t=1}^{T}V_{it}+\frac{1}{2}m_{iV}^{-2}m_{iV2}\left(\frac{1}{T}\sum_{t=1}^{T}V_{it}\right)^2\right\}+O_P\left(T^{-3}(\ln T)^9\right),\end{aligned}\tag{S11}$$

where we use the fact $\max_{1\le i\le N}\left|\frac{1}{T}\sum_{t=1}^{T}\left[V_i^{\mu_i\mu_i}\left(w_{it};\beta_i,\check{\mu}_i(\beta_i^0)\right)-V_{it}^{\mu_i\mu_i}\right]\right| \le \max_{1\le i\le N}\frac{1}{T}\sum_{t=1}^{T}M(w_{it})$ $\times\max_{1\le i\le N}\left|\check{\mu}_i(\beta_i^0)-\mu_i(\beta_i^0)\right| = O_P(T^{-1/2}(\ln T)^3)$ and $\max_{1\le i\le N}\left|\frac{1}{T}\sum_{t=1}^{T}V_{it}^{\mu_i\mu_i}-m_{iV2}\right| = O_P(T^{-1/2}(\ln T)^3)$ by Lemma S1.2(i). It follows that

$$\begin{aligned}S_{k,2} &= \frac{-1}{\sqrt{N_kT}}\sum_{i\in G_k^0}\sum_{t=1}^{T}U_{it}^{\mu_i}\left\{\left[\frac{1}{T}\sum_{t=1}^{T}V_{it}^{\mu_i}\right]^{-1}\left\{\frac{1}{T}\sum_{t=1}^{T}V_{it}+\frac{1}{2}m_{iV}^{-2}m_{iV2}\left(\frac{1}{T}\sum_{t=1}^{T}V_{it}\right)^2\right\}+O_P(T^{-3}(\ln T)^9)\right\}\\ &= -\frac{1}{\sqrt{N_kT}}\sum_{i\in G_k^0}\sum_{t=1}^{T}V_{it}\frac{\frac{1}{T}\sum_{t=1}^{T}U_{it}^{\mu_i}}{\frac{1}{T}\sum_{t=1}^{T}V_{it}^{\mu_i}}\\ &\quad -\frac{1}{2\sqrt{N_kT}}\sum_{i\in G_k^0}\sum_{t=1}^{T}U_{it}^{\mu_i}\frac{m_{iV}^{-2}m_{iV2}\left(\frac{1}{T}\sum_{t=1}^{T}V_{it}\right)^2}{\frac{1}{T}\sum_{t=1}^{T}V_{it}^{\mu_i}}+o_P(1)\\ &\equiv -S_{k,21}-S_{k,22}+o_P(1).\end{aligned}\tag{S12}$$

For $S_{k,21}$, we make the following decomposition:

$$\begin{aligned}
S_{k,21} &= \frac{1}{\sqrt{N_kT}}\sum_{i\in G_k^0}\sum_{t=1}^{T}V_{it}\frac{m_{iU}}{m_{iV}}+\frac{1}{\sqrt{N_kT}}\sum_{i\in G_k^0}\sum_{t=1}^{T}V_{it}\frac{\frac{1}{T}\sum_{t=1}^{T}\left(U_{it}^{\mu_i}-m_{iU}\right)}{m_{iV}}\\
&\quad+\frac{1}{\sqrt{N_kT}}\sum_{i\in G_k^0}\sum_{t=1}^{T}V_{it}m_{iU}\left\{\frac{1}{\frac{1}{T}\sum_{t=1}^{T}V_{it}^{\mu_i}}-\frac{1}{m_{iV}}\right\}\\
&\quad+\frac{1}{\sqrt{N_kT}}\sum_{i\in G_k^0}\sum_{t=1}^{T}V_{it}\frac{1}{T}\sum_{t=1}^{T}\left(U_{it}^{\mu_i}-m_{iU}\right)\left\{\frac{1}{\frac{1}{T}\sum_{t=1}^{T}V_{it}^{\mu_i}}-\frac{1}{m_{iV}}\right\}\\
&\equiv S_{k,21a}+S_{k,21b}+S_{k,21c}+S_{k,21d}. \qquad\text{(S13)}
\end{aligned}$$

Apparently, $S_{k,21b}=\frac{1}{\sqrt{N_kT^3}}\sum_{i\in G_k^0}m_{iV}^{-1}\sum_{s=1}^{T}\sum_{t=1}^{T}V_{is}\left[U_{it}^{\mu_i}-\mathbb{E}(U_{it}^{\mu_i})\right]$. For $S_{k,21c}$, we can use the fact that $\max_{1\le i\le N}\left|\frac{1}{T}\sum_{t=1}^{T}V_{it}\right|=O_P\left(T^{-1/2}\left(\ln T\right)^3\right)$ and $\max_{1\le i\le N}\left|\frac{1}{T}\sum_{t=1}^{T}V_{it}^{\mu_i}-m_{iV}\right|=O_P\left(T^{-1/2}\left(\ln T\right)^3\right)$ to show that

$$\begin{aligned}
S_{k,21c} &= \frac{-1}{\sqrt{N_kT}}\sum_{i\in G_k^0}\sum_{t=1}^{T}V_{it}m_{iU}\frac{\frac{1}{T}\sum_{t=1}^{T}\left(V_{it}^{\mu_i}-m_{iV}\right)}{m_{iV}\frac{1}{T}\sum_{t=1}^{T}V_{it}^{\mu_i}}\\
&= \frac{-1}{\sqrt{N_kT^3}}\sum_{i\in G_k^0}m_{iU}m_{iV}^{-2}\sum_{s=1}^{T}\sum_{t=1}^{T}V_{is}\left(V_{it}^{\mu_i}-m_{iV}\right)+o_P\left(1\right).
\end{aligned}$$

It follows that

$$S_{k,21b}+S_{k,21c}=\frac{1}{\sqrt{N_kT^3}}\sum_{i\in G_k^0}m_{iV}^{-1}\sum_{t=1}^{T}\sum_{s=1}^{T}V_{is}\mathbb{U}_{it}^{\mu_i}+o_P\left(1\right), \qquad\text{(S14)}$$

where we use the definition of $\mathbb{U}_{it}^{\mu_i}(\equiv U_{it}^{\mu_i}-\frac{m_{iU}}{m_{iV}}V_{it}^{\mu_i})$ and the fact that $\mathbb{E}\left[\mathbb{U}_{it}^{\mu_i}\right]=0$. For $S_{k,21d}$, we can bound it directly:

$$\begin{aligned}
\left|S_{k,21d}\right| &\le \sqrt{N_kT}\max_{1\le i\le N}\left|\frac{1}{T}\sum_{t=1}^{T}V_{it}\right|\max_{1\le i\le N}\left\|\frac{1}{T}\sum_{t=1}^{T}\left(U_{it}^{\mu_i}-m_{iU}\right)\right\|\max_{1\le i\le N}\left|\frac{1}{\frac{1}{T}\sum_{t=1}^{T}V_{it}^{\mu_i}}-\frac{1}{m_{iV}}\right|\\
&= \sqrt{N_kT}O_P\left(T^{-3/2}\left(\ln T\right)^9\right)=o_P\left(1\right). \qquad\text{(S15)}
\end{aligned}$$

Combining (S13)-(S15) yields

$$S_{k,21}=\frac{1}{\sqrt{N_kT}}\sum_{i\in G_k^0}\sum_{t=1}^{T}V_{it}\frac{m_{iU}}{m_{iV}}+\frac{1}{\sqrt{N_kT^3}}\sum_{i\in G_k^0}m_{iV}^{-1}\sum_{t=1}^{T}\sum_{s=1}^{T}V_{is}\mathbb{U}_{it}^{\mu_i}+o_P\left(1\right). \qquad\text{(S16)}$$

In addition, for $S_{k,22}$ we have

$$\begin{aligned}
S_{k,22} &= \frac{1}{2\sqrt{N_kT}}\sum_{i\in G_k^0}\left(\frac{1}{T}\sum_{t=1}^{T}U_{it}^{\mu_i}\right)m_{iV}^{-3}m_{iV2}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}V_{it}\right)^2\\
&= \frac{1}{2\sqrt{N_kT}}\sum_{i\in G_k^0}m_{iU}m_{iV}^{-3}m_{iV2}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}V_{it}\right)^2+o_P\left(1\right). \qquad\text{(S17)}
\end{aligned}$$

Combining (S12)-(S17) yields

$$S_{k,2} = -\frac{1}{\sqrt{N_kT}}\sum_{i\in G_k^0}\sum_{t=1}^{T}V_{it}\frac{m_{iU}}{m_{iV}}$$
$$-\left\{\frac{1}{\sqrt{N_kT^3}}\sum_{i\in G_k^0}m_{iV}^{-1}\sum_{t=1}^{T}\sum_{s=1}^{T}V_{is}\mathbb{U}_{it}^{\mu_i}+\frac{1}{2\sqrt{N_kT}}\sum_{i\in G_k^0}m_{iU}m_{iV}^{-3}m_{iV2}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}V_{it}\right)^2\right\}+o_P(1).\text{(S18)}$$

Then by (S9), (S10), and (S18)

$$\hat{S}_{G_k^0} = \frac{1}{\sqrt{N_kT}}\sum_{i\in G_k^0}\sum_{t=1}^{T}\left(U_{it}-\frac{m_{iU}}{m_{iV}}V_{it}\right)$$
$$-\left\{\frac{1}{\sqrt{N_kT^3}}\sum_{i\in G_k^0}m_{iV}^{-1}\sum_{t=1}^{T}\sum_{s=1}^{T}V_{is}\mathbb{U}_{it}^{\mu_i}-\frac{1}{2\sqrt{N_kT}}\sum_{i\in G_k^0}m_{iV}^{-2}\left[m_{iU2}-\frac{m_{iV2}}{m_{iV}}m_{iU}\right]\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}V_{it}\right)^2\right\}$$
$$+o_P(1)$$
$$= \frac{1}{\sqrt{N_kT}}\sum_{i\in G_k^0}\sum_{t=1}^{T}\mathbb{U}_{it}-\mathbb{B}_{kNT}+o_P(1).$$

This completes the proof of (i).

Part (ii): We prove $\mathbb{Z}_{NT}\equiv\frac{1}{\sqrt{N_kT}}\sum_{i\in G_k^0}\sum_{t=1}^{T}\mathbb{U}_{it}\xrightarrow{D}N(0,\Omega_k)$. Let $Z_{iT}\equiv\frac{1}{\sqrt{N_kT}}\sum_{t=1}^{T}\iota_p'\mathbb{U}_{it}$ where ι_p is an arbitrary $p\times1$ nonrandom vector with $\|\iota_p\|=1$. Then $\iota_p'\mathbb{Z}_{NT}=\sum_{i\in G_k^0}Z_{iT}$. Noting that $\mathbb{E}(\mathbb{U}_{it})=0$, we have by Assumption A3(i)

$$\mathrm{Var}(\mathbb{Z}_{NT})=\frac{1}{N_kT}\sum_{i\in G_k^0}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathbb{E}(\mathbb{U}_{it}\mathbb{U}_{is}')=\frac{1}{N_k}\sum_{i\in G_k^0}\Omega_{iT}\to\Omega_k>0,$$

where $\Omega_{iT}=\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathbb{E}(\mathbb{U}_{it}\mathbb{U}_{is}')$. By the Lindeberg-Feller central limit theorem (e.g., Theorem 5.6 in White (2001)), it suffices to verify the Lindeberg condition:

$$S_{NT}\equiv\sum_{i\in G_k^0}\mathbb{E}\left[Z_{iT}^2\mathbf{1}\{|Z_{iT}|>\varepsilon\}\right]\to0\text{ for any given }\varepsilon>0.$$

By the Cauchy-Schwarz and Markov inequalities,

$$\sum_{i\in G_k^0}\mathbb{E}\left[Z_{iT}^2\mathbf{1}\{|Z_{iT}|>\varepsilon\}\right]\le\sum_{i\in G_k^0}\left\{\mathbb{E}\left(Z_{iT}^4\right)\right\}^{1/2}\{P(|Z_{iT}|>\varepsilon)\}^{1/2}\le\frac{1}{\varepsilon^2}\sum_{i\in G_k^0}\mathbb{E}\left(Z_{iT}^4\right).$$

By straightforward moment conditions and properties of strong mixing processes, we can readily show that

$$\mathbb{E}\left(Z_{iT}^4\right)=\frac{1}{N_k^2T^2}\left(\sum_{t=1}^{T}\iota_p'\mathbb{U}_{it}\right)^4=O\left(N_k^{-2}\right)\text{ uniformly in }i.$$

It follows that $S_{NT}\to0$ for any $\varepsilon>0$ and $\mathbb{Z}_{NT}\equiv\frac{1}{\sqrt{N_kT}}\sum_{i\in G_k^0}\sum_{t=1}^{T}\mathbb{U}_{it}\xrightarrow{D}N(0,\Omega_k)$. ■

REMARK. Note that $\mathbb{U}_i(w_{it};\beta_i,\mu_i)\equiv U_i(w_{it};\beta_i,\mu_i)-\frac{m_{iU}}{m_{iV}}V_i(w_{it};\beta_i,\mu_i)$ and $\mathbb{U}_{it}$ correspond to $U_i(x_{it};\theta,\gamma_i)$ and U_{it} in Hahn and Kuersteiner (2011, HK hereafter), respectively. Let $\mathbb{U}_i^{\mu_i}$ and $\mathbb{U}_i^{\mu_i\mu_i}$ denote the first and second derivatives of $\mathbb{U}_i$ with respect to μ_i. Let $\mathbb{U}_{it}^{\mu_i} = \mathbb{U}_i^{\mu_i}\left(w_{it};\beta_i^0,\mu_i^0\right)$ and $\mathbb{U}_{it}^{\mu_i\mu_i} = \mathbb{U}_i^{\mu_i\mu_i}(w_{it};\beta_i^0,\mu_i^0)$. Following HK, the asymptotic bias term of $\hat{S}_{\hat{G}_k}$ takes the form:

$$\mathbb{B}_{kNT}^{HK} = \frac{1}{\sqrt{N_k T}} \sum_{i\in G_k^0} \left[m_{iV}^{-1}\frac{1}{\sqrt{T}}\sum_{t=1}^T V_{it}\right]\left[\frac{1}{\sqrt{T}}\sum_{t=1}^T\left(\mathbb{U}_{it}^{\mu_i} - \frac{m_{i\mathbb{U}2}}{2m_{iV}}V_{it}\right)\right],$$

where $m_{i\mathbb{U}2} \equiv \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left(\mathbb{U}_{it}^{\mu_i\mu_i}\right)$. Note that

$$\begin{aligned}\mathbb{B}_{kNT}^{HK} &= \frac{1}{\sqrt{N_k T^3}}\sum_{i\in G_k^0} m_{iV}^{-1}\sum_{s=1}^T\sum_{t=1}^T V_{is}\mathbb{U}_{it}^{\mu_i} - \frac{1}{2\sqrt{N_k T}}\sum_{i\in G_k^0} m_{iV}^{-2} m_{i\mathbb{U}2}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^T V_{it}\right)^2 \\ &\equiv \mathbb{B}_{1kNT}^{HK} - \mathbb{B}_{2kNT}^{HK}, \text{ say.}\end{aligned}$$

Let $\mathbb{B}_{1kNT}$ and $\mathbb{B}_{2kNT}$ be as defined in Theorem 2.4. Apparently, $\mathbb{B}_{1kNT}^{HK} = \mathbb{B}_{1kNT}$. Noting that $m_{i\mathbb{U}2} = \frac{1}{T}\sum_{t=1}^T \mathbb{E}(U_{it}^{\mu_i\mu_i} - \frac{m_{iU}}{m_{iV}}V_i^{\mu_i\mu_i}) = m_{iU2} - \frac{m_{iU}}{m_{iV}}m_{iV2}$ with $m_{iU2} = \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left(U_{it}^{\mu_i\mu_i}\right)$, we have

$$\mathbb{B}_{2kNT}^{HK} = \frac{1}{2\sqrt{N_k T}}\sum_{i\in G_k^0} m_{iV}^{-2}\left(m_{iU2} - \frac{m_{iU}}{m_{iV}}m_{iV2}\right)\left(\frac{1}{\sqrt{T}}\sum_{t=1}^T V_{it}\right)^2 = \mathbb{B}_{2kNT}.$$

It follows that $\mathbb{B}_{kNT}^{HK} = \mathbb{B}_{kNT}$.

Lemma S1.13 *Let* $\hat{H}_{(k)} \equiv \frac{1}{N_k T}\sum_{i\in\hat{G}_k}\sum_{t=1}^T [U_i^{\beta_i}(w_{it};\breve{\alpha}_k,\hat{\mu}_i(\breve{\alpha}_k)) + U_i^{\mu_i}\left(w_{it};\breve{\alpha}_k^0,\hat{\mu}_i(\breve{\alpha}_k)\right)\frac{\partial\hat{\mu}_i(\breve{\alpha}_k)}{\partial\alpha_k'}]$ *and* $\breve{\alpha}_k$ *lying between* $\hat{\alpha}_k$ *and* α_k^0 *elementwise. Then* $\hat{H}_{(k)} = \mathbb{H}_{kNT} + o_P(\nu_{NT})$ *where* $\nu_{NT} = \min(1, \sqrt{T/N_k})$.

Proof. As in the proof of Lemma S1.12, we can readily show that $\hat{H}_{(k)} = \hat{H}_{G_k^0} + o_P(1)$ where $\hat{H}_{G_k^0} = \frac{1}{N_k T}\sum_{i\in G_k^0}\sum_{t=1}^T [U_i^{\beta_i}(w_{it};\breve{\alpha}_k,\hat{\mu}_i(\breve{\alpha}_k)) + U_i^{\mu_i}\left(w_{it};\breve{\alpha}_k^0,\hat{\mu}_i(\breve{\alpha}_k)\right)\frac{\partial\hat{\mu}_i(\breve{\alpha}_k)}{\partial\alpha_k'}]$. For $\hat{H}_{G_k^0}$ We make the following decomposition

$$\begin{aligned}\hat{H}_{G_k^0} &\equiv \frac{1}{N_k T}\sum_{i\in G_k^0}\sum_{t=1}^T\left[U_i^{\beta_i}\left(w_{it};\alpha_k^0,\mu_i(\alpha_k^0)\right) + U_i^{\mu_i}\left(w_{it};\alpha_k^0,\mu_i(\alpha_k^0\right)\frac{\partial\mu_i(\alpha_k^0)}{\partial\alpha_k'}\right], \\ &\quad + \frac{1}{N_k T}\sum_{i\in G_k^0}\sum_{t=1}^T\left[U_i^{\beta_i}\left(w_{it};\breve{\alpha}_k,\hat{\mu}_i(\breve{\alpha}_k)\right) - U_i^{\alpha_k}\left(w_{it};\alpha_k^0,\mu_i(\alpha_k^0)\right)\right] \\ &\quad + \frac{1}{NT}\sum_{i\in\hat{G}_k}\sum_{t=1}^T\left[U_i^{\mu_i}\left(w_{it};\alpha_k^0,\mu_i(\alpha_k^0\right)\frac{\partial\mu_i(\alpha_k^0)}{\partial\alpha_k'} - U_i^{\mu_i}\left(w_{it};\breve{\alpha}_k^0,\hat{\mu}_i(\breve{\alpha}_k)\right)\frac{\partial\hat{\mu}_i(\breve{\alpha}_k)}{\partial\alpha_k'}\right] \\ &\equiv H_{G_k^0,1} + H_{G_k^0,2} + H_{G_k^0,3}.\end{aligned}$$

Using the arguments in the proof of Lemma S1.10, we can readily show that $H_{G_k^0,s} = o_P(\nu_{NT})$ for $s = 2, 3$. In addition, by the Chebyshev inequality we can show that $H_{G_k^0,1} = \bar{H}_{G_k^0,1} + o_P(\nu_{NT})$, where $\bar{H}_{G_k^0,1} = \mathbb{E}[\bar{H}_{G_k^0,1}]$. Then by (S5) with $\epsilon = 0$,

$$\begin{aligned}\bar{H}_{G_k^0,1} &= \frac{1}{N_k T}\sum_{i\in G_k^0}\sum_{t=1}^T \mathbb{E}\left[U_i^{\beta_i}\left(w_{it};\alpha_k^0,\mu_i(\alpha_k^0)\right) + U_i^{\mu_i}\left(w_{it};\alpha_k^0,\mu_i(\alpha_k^0\right)\frac{\partial\mu_i(\alpha_k^0)}{\partial\alpha_k'}\right] \\ &= \frac{1}{N_k T}\sum_{i\in G_k^0}\sum_{t=1}^T \mathbb{E}\left[U_{it}^{\beta_i} - U_{it}^{\mu_i}\frac{\frac{1}{T}\sum_{t=1}^T\mathbb{E}\left(V_{it}^{\beta_i}\right)'}{m_{iV}}\right] \\ &= \frac{1}{N_k T}\sum_{i\in G_k^0}\sum_{t=1}^T \mathbb{E}\left[U_{it}^{\beta_i} - \frac{m_{iU}}{m_{iV}}V_{it}^{\beta_i\prime}\right] = \frac{1}{N_k T}\sum_{i\in G_k^0}\sum_{t=1}^T\mathbb{E}\left[\mathbb{U}_{it}^{\beta_i}\right] = \mathbb{H}_{kNT}.\end{aligned}$$

It follows that $\hat{H}_{(k)} = \mathbb{H}_{kNT} + o_P(\nu_{NT})$. ■

REMARK. When $\{\mathbb{U}_{it}, t \geq 1\}$ are serially uncorrelated, we have

$$\Omega_k = \lim_{(N_k,T)\to\infty} \frac{1}{N_k} \sum_{i \in G_k^0} \Omega_{iT} = \lim_{(N_k,T)\to\infty} \frac{1}{N_k} \sum_{i \in G_k^0} \frac{1}{T} \sum_{t=1}^T \mathbb{E}\left(\mathbb{U}_{it}\mathbb{U}_{it}'\right).$$

When the likelihood function is correctly specified, we can apply the second Bartlett identity (i.e., the information matrix equality) to obtain

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[U_{it}U_{it}'\right] &= -\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[U_{it}^{\beta_i}\right], \\
\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[U_{it}V_{it}\right] &= -\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[U_{it}^{\mu_i}\right] = -m_{iU} = -\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[V_{it}^{\beta_i}\right], \\
\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[V_{it}^2\right] &= -\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[V_{it}^{\mu_i}\right] = -m_{iV},
\end{aligned}$$

when $i \in G_k^0$. Then

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left(\mathbb{U}_{it}\mathbb{U}_{it}'\right) &= \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[\left(U_{it} - \frac{m_{iU}}{m_{iV}}V_{it}\right)\left(U_{it} - \frac{m_{iU}}{m_{iV}}V_{it}\right)'\right] \\
&= \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[U_{it}U_{it}' + \frac{m_{iU}m_{iU}'}{m_{iV}^2}V_{it}^2 - U_{it}V_{it}\frac{m_{iU}'}{m_{iV}} - \frac{m_{iU}}{m_{iV}}U_{it}'V_{it}\right] \\
&= \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[-U_{it}^{\beta_i} - \frac{m_{iU}m_{iU}'}{m_{iV}} + \frac{m_{iU}m_{iU}'}{m_{iV}} + \frac{m_{iU}m_{iU}'}{m_{iV}}\right] \\
&= -\frac{1}{N_kT}\sum_{i \in G_k^0}\sum_{t=1}^T \mathbb{E}\left[U_{it}^{\beta_i} - \frac{m_{iU}}{m_{iV}}V_{it}^{\beta_i \prime}\right].
\end{aligned}$$

It follows that

$$\Omega_k = -\lim_{(N_k,T)\to\infty} \frac{1}{N_k}\sum_{i \in G_k^0} \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left(\mathbb{U}_{it}\mathbb{U}_{it}'\right) = -\mathbb{H}_k.$$

Lemma S1.14 *Suppose the conditions in Theorem 2.6 hold. Recall that* $\hat{\sigma}^2_{\hat{G}(K,\lambda_1)} = \frac{2}{NT}\sum_{k=1}^K \sum_{i \in \hat{G}_k(K,\lambda_1)} \sum_{t=1}^T \psi\left(w_{it}; \hat{\alpha}_{\hat{G}_k(K,\lambda_1)}, \hat{\mu}_i(\hat{\alpha}_{\hat{G}_k(K,\lambda_1)})\right)$. *Let* $\bar{\sigma}^2_{G^0} = \frac{2}{NT}\sum_{i=1}^N \sum_{t=1}^T \psi\left(w_{it}; \beta_i^0, \mu_i^0\right)$. *Then* $\max_{K_0 \leq K \leq K_{\max}} |\hat{\sigma}^2_{\hat{G}(K,\lambda_1)} - \bar{\sigma}^2_{G^0}| = O_P\left(T^{-1}\right)$.

Proof. When $K \geq K_0$, following the proof of Theorem 2.1, we can show that $\left\|\hat{\beta}_i - \beta_i^0\right\| = O_P(T^{-1/2} + \lambda_1)$ for each i, and

$$\frac{1}{N}\sum_{i=1}^N \Pi_{k=1}^K \left\|\beta_i^0 - \hat{\alpha}_k\right\| = \frac{N_1}{N}\Pi_{k=1}^K \left\|\hat{\alpha}_k - \alpha_1^0\right\| + \cdots + \frac{N_{K_0}}{N}\Pi_{k=1}^K \left\|\hat{\alpha}_k - \alpha_{K_0}^0\right\| = O_P(T^{-1/2}).$$

Then by Assumption A1(vii), $\Pi_{k=1}^K \left\|\hat{\alpha}_k - \alpha_l^0\right\| = O_P(T^{-1/2})$ for $l = 1, ..., K_0$. It follows that the collection $\mathcal{C} \equiv \{\hat{\alpha}_k, k = 1, ..., K\}$ contains at least K_0 distinct vectors, say, $\hat{\alpha}_1, ..., \hat{\alpha}_{K_0}$, possibly after relabeling the vectors, such that

$$\left\|\hat{\alpha}_k - \alpha_k^0\right\| \Pi_{l=K_0+1}^K \left\|\hat{\alpha}_l - \alpha_k^0\right\| = O_P(T^{-1/2}) \text{ for } k = 1, ..., K_0.$$

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. Suppose that $i \in \hat{G}_k(K,\lambda_1)$ for $k \in \{K_0+1, ..., K\}$. Then by the pointwise consistency of $\hat{\beta}_i$, we know that the probability limit of $\hat{\alpha}_k$ must be given by one of the columns in $\boldsymbol{\alpha}^0 = \left(\alpha_1^0, ..., \alpha_{K_0}^0\right)$ and it converges in probability to the true value at the rate $T^{-1/2} + \lambda_1$. Apparently, if $\mathcal{C}$ contains n_k elements with probability limit given by α_k^0, we can derive that $\left\|\hat{\alpha}_k - \alpha_k^0\right\| = O_P(\min(T^{-1/(2n_k)}, T^{-1/2}+\lambda_1))$ for $k = 1, ..., K_0$. Without loss of generality, assume that if $n_k > 1$ for $k \in \{1, ..., K_0\}$, $\hat{G}_k(K,\lambda_1)$ contains the maximum number of elements among the subsets $\hat{G}_l(K,\lambda_1)$ with $\mathrm{plim}_{(N,T)\to\infty}\hat{\alpha}_l = \alpha_k^0$.

Using arguments like those in the proof of Theorem 2.2, we can show that

$$\sum_{i\in G_k^0} P\left(\hat{E}_{kNT,i}\right) = o(1) \text{ for } k = 1, ..., K_0 \text{ and } \sum_{i\in \hat{G}_k(K,\lambda_1)} P\left(\hat{F}_{kNT,i}\right) = o(1) \text{ for } k = 1, ..., K_0. \tag{S19}$$

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

Let $\hat{\psi}_{it}(k) = 2\psi\left(w_{it}; \hat{\alpha}_{\hat{G}_k(K,\lambda_1)}(K,\lambda_1), \hat{\mu}_i(\hat{\alpha}_{\hat{G}_k(K,\lambda_1)})\right)$. 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

$$\hat{\sigma}^2_{\hat{G}(K,\lambda_1)} = \frac{1}{NT}\sum_{k=1}^K \sum_{i\in \hat{G}_k(K,\lambda_1)} \sum_{t=1}^T \hat{\psi}_{it}(k) = D_{1NT} + D_{2NT} - D_{3NT} + D_{4NT},$$

where

$$\begin{aligned} D_{1NT} &= \frac{1}{NT}\sum_{k=1}^{K_0}\sum_{i\in G_k^0}\sum_{t=1}^T \hat{\psi}_{it}(k), \quad D_{2NT} = \frac{1}{NT}\sum_{k=1}^{K_0}\sum_{i\in \hat{G}_k(K,\lambda_1)\backslash G_k^0}\sum_{t=1}^T \hat{\psi}_{it}(k), \\ D_{3NT} &= \frac{1}{NT}\sum_{k=1}^{K_0}\sum_{i\in G_k^0\backslash \hat{G}_k(K,\lambda_1)}\sum_{t=1}^T \hat{\psi}_{it}(k), \text{ and } D_{4NT} = \frac{1}{NT}\sum_{k=K_0+1}^{K}\sum_{i\in \hat{G}_k(K,\lambda_1)}\sum_{t=1}^T \hat{\psi}_{it}(k). \end{aligned}$$

Let $\delta_{NT} = \min(\sqrt{NT}, T)$. By (S19), 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 $\hat{\sigma}^2_{\hat{G}(K,\lambda_1)} = D_{1NT} + o_P\left(\delta_{NT}^{-2}\right)$ for all $K_0 \leq K \leq K_{\max}$.

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 $k = 1, ..., K_0$. Then by Taylor expansion, we can readily show that

$$\begin{aligned} D_{1NT} &= \frac{2}{NT}\sum_{k=1}^{K_0}\sum_{i\in G_k^0}\sum_{t=1}^T \psi\left(w_{it}; \hat{\alpha}_{\hat{G}_k(K,\lambda_1)}(K,\lambda_1), \hat{\mu}_i(\hat{\alpha}_{\hat{G}_k(K,\lambda_1)})\right) \\ &= \frac{2}{NT}\sum_{k=1}^{K_0}\sum_{i\in G_k^0}\sum_{t=1}^T \psi\left(w_{it}; \alpha_k^0, \hat{\mu}_i\left(\alpha_k^0\right)\right) \\ &\quad + \frac{2}{NT}\sum_{k=1}^{K_0}\sum_{i\in G_k^0}\sum_{t=1}^T U_i\left(w_{it}; \alpha_k^0, \hat{\mu}_i\left(\alpha_k^0\right)\right)\left[\hat{\alpha}_{\hat{G}_k(K,\lambda_1)} - \alpha_k^0\right] \\ &\quad + \frac{2}{NT}\sum_{k=1}^{K_0}\sum_{i\in G_k^0}\sum_{t=1}^T V_i\left(w_{it}; \alpha_k^0, \hat{\mu}_i\left(\alpha_k^0\right)\right)\frac{\partial\hat{\mu}_i\left(\alpha_k^0\right)}{\partial\alpha_k}\left[\hat{\alpha}_{\hat{G}_k(K,\lambda_1)} - \alpha_k^0\right] + O_P\left(\delta_{NT}^{-2}\right). \\ &\equiv D_{1NT,1} + D_{1NT,2} + D_{1NT,3} + O_P\left(\delta_{NT}^{-2}\right). \end{aligned}$$

By Lemma S1.12 and the fact that $\mathbb{B}_k = O_P\left((N/T)^{1/2}\right)$, we can show that $D_{1NT,2} = O_P\left(\delta_{NT}^{-2}\right)$. Let $\bar{D}_{1NT,3} \equiv \frac{1}{NT}\sum_{k=1}^{K_0}\sum_{i\in G_k^0}\sum_{t=1}^{T} V_i\left(w_{it};\alpha_k^0,\hat{\mu}_i\left(\alpha_k^0\right)\right)\frac{\partial\hat{\mu}_i(\alpha_k^0)}{\partial\alpha_k}$. Then

$$\begin{aligned}
\bar{D}_{1NT,3} &= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} V_{it}\frac{\partial\mu_i\left(\beta_i^0\right)}{\partial\beta_i} + \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} V_{it}\left(\frac{\partial\hat{\mu}_i\left(\beta_i^0\right)}{\partial\beta_i} - \frac{\partial\mu_i\left(\beta_i^0\right)}{\partial\beta_i}\right) \\
&\quad + \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} V_{it}^{\mu_i}\left(\hat{\mu}_i\left(\beta_i^0\right)-\mu_i^0\right)\frac{\partial\hat{\mu}_i\left(\alpha_k^0\right)}{\partial\alpha_k} + O_P\left(T^{-1}\right) \\
&\equiv \bar{D}_{1NT,31} + \bar{D}_{1NT,32} + \bar{D}_{1NT,33} + O_P\left(T^{-1}\right).
\end{aligned}$$

By the Chebyshev and Davydov inequalities, we can readily show that $\bar{D}_{1NT,31} = O_P\left((NT)^{-1/2}\right)$. By (S5),

$$\begin{aligned}
\frac{\partial\hat{\mu}_i\left(\beta_i^0\right)}{\partial\beta_i} - \frac{\partial\mu_i\left(\beta_i^0\right)}{\partial\beta_i} &= \frac{\partial\mu_i\left(\beta_i^0, F_i\left(T^{-1/2}\right)\right)}{\partial\beta_i} - \frac{\partial\mu_i\left(\beta_i^0, F_i\left(0\right)\right)}{\partial\beta_i} \\
&= \frac{\int V_i^{\beta_i}\left(\cdot;\beta_i^0,\mu_i^0\right)dF_i}{\int V_i^{\mu_i}\left(\cdot;\beta_i^0,\mu_i^0\right)dF_i} - \frac{\int V_i^{\beta_i}\left(\cdot;\beta_i^0,\hat{\mu}_i\left(\beta_i^0\right)\right)d\hat{F}_i}{\int V_i^{\mu_i}\left(\cdot;\beta_i^0,\hat{\mu}_i\left(\beta_i^0\right)\right)d\hat{F}_i} \\
&= \frac{n_{iV}}{m_{iV}} - \frac{\hat{n}_{iV}}{\hat{m}_{iV}} = \frac{n_{iV}\hat{m}_{iV} - \hat{n}_{iV}n_{iV}}{m_{iV}\hat{m}_{iV}} \\
&= \frac{n_{iV}\left(\hat{m}_{iV}-m_{iV}\right) + \left(n_{iV}-\hat{n}_{iV}\right)m_{iV}}{m_{iV}\hat{m}_{iV}} \qquad \text{(S20)}
\end{aligned}$$

where $n_{iV} \equiv \int V_i^{\mu_i}\left(\cdot;\beta_i^0,\mu_i^0\right)dF_i$, $\hat{n}_{iV} \equiv \int V_i^{\beta_i}\left(\cdot;\beta_i^0,\hat{\mu}_i\left(\beta_i^0\right)\right)d\hat{F}_i$, $\hat{m}_{iV} \equiv \int V_i^{\mu_i}\left(\cdot;\beta_i^0,\hat{\mu}_i\left(\beta_i^0\right)\right)d\hat{F}_i$, and recall $m_{iV} \equiv \int V_i^{\mu_i}\left(\cdot;\beta_i^0,\mu_i^0\right)dF_i$. Then by (S11) and Lemma S1.2(i), we can show that

$$\begin{aligned}
\bar{D}_{1NT,32} &= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} V_{it}m_{iV}^{-2}n_{iV}\left(\hat{m}_{iV}-m_{iV}\right) - \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} V_{it}m_{iV}^{-1}\left(\hat{n}_{iV}-n_{iV}\right) + O_P\left(\delta_{NT}^{-1}\right), \\
\frac{1}{N}\sum_{i=1}^{N}\left(\hat{m}_{iV}-m_{iV}\right)^2 &= \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{1}{T}\sum_{T=1}^{T}\left[V_i^{\mu_i}\left(w_{it};\beta_i^0,\hat{\mu}_i\left(\beta_i^0\right)\right)-m_{iV}\right]\right\}^2 \\
&\leq \frac{2}{N}\sum_{i=1}^{N}\left\{\frac{1}{T}\sum_{T=1}^{T}\left[V_i^{\mu_i}\left(w_{it};\beta_i^0,\hat{\mu}_i\left(\beta_i^0\right)\right)-V_{it}^{\mu_i}\right]\right\}^2 + \frac{2}{N}\sum_{i=1}^{N}\left\{\frac{1}{T}\sum_{T=1}^{T}\left(V_{it}^{\mu_i}-m_{iV}\right)\right\}^2 \\
&= O_P\left(T^{-1}\right) + O_P\left(T^{-1}\right) = O_P\left(T^{-1}\right),
\end{aligned}$$

and similarly $\frac{1}{N}\sum_{i=1}^{N}\left(\hat{n}_{iV}-n_{iV}\right)^2 = O_P\left(T^{-1}\right)$. Then

$$\begin{aligned}
\left\|\bar{D}_{1NT,32}\right\| &\leq \left\{\frac{1}{N}\sum_{i=1}^{N}\left\|m_{iV}^{-2}n_{iV}\right\|\left(\frac{1}{T}\sum_{t=1}^{T}V_{it}\right)^2\right\}^{1/2}\left\{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{m}_{iV}-m_{iV}\right)^2\right\}^{1/2} \\
&\quad + \left\{\frac{1}{N}\sum_{i=1}^{N}\left|m_{iV}^{-1}\right|\left(\frac{1}{T}\sum_{t=1}^{T}V_{it}\right)^2\right\}^{1/2}\left\{\frac{1}{N}\sum_{i=1}^{N}\left\|\hat{n}_{iV}-n_{iV}\right\|^2\right\}^{1/2} + O_P\left(\delta_{NT}^{-1}\right) \\
&= O_P\left(T^{-1}\right) + O_P\left(T^{-1}\right) + O_P\left(\delta_{NT}^{-1}\right) = O_P\left(\delta_{NT}^{-1}\right).
\end{aligned}$$

For $\bar{D}_{1NT,33}$, using (S11), (S20), and Lemma S1.2(i), we can readily show that

$$\begin{aligned}
\bar{D}_{1NT,33} &= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} V_{it}^{\mu_i}\left(\hat{\mu}_i\left(\beta_i^0\right)-\mu_i^0\right)\frac{\partial \mu_i\left(\alpha_k^0\right)}{\partial \alpha_k} + O_P\left(\delta_{NT}^{-1}\right) \\
&= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} V_{it}\frac{\partial \mu_i\left(\alpha_k^0\right)}{\partial \alpha_k} + O_P\left(\delta_{NT}^{-1}\right) \\
&= O_P\left((NT)^{-1/2}\right) + O_P\left(\delta_{NT}^{-1}\right) = O_P\left(\delta_{NT}^{-1}\right).
\end{aligned}$$

Then $\bar{D}_{1NT,3} = O_P\left(\delta_{NT}^{-1}\right)$ and $D_{1NT,3} = O_P\left(\delta_{NT}^{-2}\right)$.

By Taylor expansion,

$$\begin{aligned}
D_{1NT,1} - \bar{\sigma}_{G^0}^2 &= \frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\psi\left(w_{it};\beta_i^0,\hat{\mu}_i\left(\beta_i^0\right)\right) - \bar{\sigma}_{G^0}^2 \\
&= \frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} V_i\left(w_{it};\beta_i^0,\mu_i^0\right)\left[\hat{\mu}_i\left(\beta_i^0\right)-\mu_i^0\right] + \frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} V_i^{\mu_i}\left(w_{it};\beta_i^0,\breve{\mu}_i\left(\beta_i^0\right)\right)\left[\hat{\mu}_i\left(\beta_i^0\right)-\mu_i^0\right]^2 \\
&\equiv D_{1NT,11} + D_{1NT,12}
\end{aligned}$$

Using (S11), we can readily show that $D_{1NT,11} = O_P\left(T^{-1}\right)$ and $D_{1NT,12} = O_P\left(T^{-1}\right)$. Then $D_{1NT} = \bar{\sigma}_{G^0}^2 + O_P\left(T^{-1}\right)$. It follows that $\hat{\sigma}_{\hat{G}(K,\lambda_1)}^2 - \bar{\sigma}_{G^0}^2 = O_P\left(T^{-1}\right)$ for each $K_0 \leq K \leq K_{\max}$. ■

S2 Bias Correction in Linear Panel Data Models

S2.1 Bias Correction for the PPL C-Lasso Estimator

For the linear models considered in Section 2.6, the bias of the Lasso and post-Lasso estimator takes the form

$$b_{kNT} = \mathbb{H}_{kNT}^{-1}\mathbb{B}_{kNT} = \mathbb{H}_{kNT}^{-1}\mathbb{B}_{1kNT},$$

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

$$\hat{b}_{kNT} = \widehat{\mathbb{H}}_{kNT}^{-1}\hat{\mathbb{B}}_{1kNT}$$

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

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

Dynamic misspecification is permitted here. If the model is dynamically correctly specified in the sense that $\mathbb{E}\left(\varepsilon_{it}|\mathcal{F}_{i,t-1}\right) = 0$ where $\mathcal{F}_{i,t-1} = \sigma(u_{i,t-1},\, u_{i,t-2}, ...;\, x_{it},\, x_{it-1}, ...)$, a one-sided kernel can be used with $k_{M_T}\left(t,s\right) = k_{M_T}^1\left(s-t\right)$, where

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

[2] 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.

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

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

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

Let $x_i \equiv (x_{i1}, ..., x_{iT})'$ and $\varepsilon_i \equiv (\varepsilon_{i1}, ..., \varepsilon_{iT})'$. 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* $i = 1, ..., N$, $\{(x_{it}, \varepsilon_{it}) : t = 1, 2, ...\}$ *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* $\rho \in (0,1)$. $\frac{1}{N_k}\sum_{i \in G_k^0} c_{\alpha,i}^{(2q-1)/q} = O(1)$.

(ii) (x_i, ε_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* $\max_{i,t} \mathbb{E}\|\varepsilon_{it}\|^{4q} < C < \infty$ *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 S2.1 *Suppose Assumption D1 holds. Then* $\widehat{\mathbb{H}}_{kNT}^{-1}\hat{\mathbb{B}}_{1kNT} - \mathbb{H}_{kNT}^{-1}\mathbb{B}_{1kNT} = o_P(1)$.

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

We first prove (i). Let $\bar{\mathbb{H}}_{kNT} \equiv \frac{1}{N_k T}\sum_{i \in G_k^0}\sum_{t=1}^T \tilde{x}_{it}\tilde{x}_{it}'$. It suffices to show that (i1) $\widehat{\mathbb{H}}_{kNT} - \bar{\mathbb{H}}_{kNT} = o_P(\nu_{NT})$ and (i2) $\bar{\mathbb{H}}_{kNT} - \mathbb{H}_{kNT} = o_P(\nu_{NT})$. Note that

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

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

$o_P(\nu_{NT})$. Now,

$$\begin{aligned}
\bar{\mathbb{H}}_{kNT}-\mathbb{H}_{kNT} &= \frac{1}{N_kT}\sum_{i\in G_k^0}\sum_{t=1}^{T}\left\{\tilde{x}_{it}\tilde{x}_{it}'-\mathbb{E}\{\left[x_{it}-\mathbb{E}\left(\bar{x}_{i\cdot}\right)\right]\left[x_{it}-\mathbb{E}\left(\bar{x}_{i\cdot}\right)\right]'\}\right\} \\
&= \frac{1}{N_kT}\sum_{i\in G_k^0}\sum_{t=1}^{T}\left\{\left[x_{it}-\mathbb{E}\left(\bar{x}_{i\cdot}\right)\right]\left[x_{it}-\mathbb{E}\left(\bar{x}_{i\cdot}\right)\right]-\mathbb{E}\{\left[x_{it}-\mathbb{E}\left(\bar{x}_{i\cdot}\right)\right]\left[x_{it}-\mathbb{E}\left(\bar{x}_{i\cdot}\right)\right]'\right\} \\
&\quad +\frac{1}{N_kT}\sum_{i\in G_k^0}\sum_{t=1}^{T}\left\{\tilde{x}_{it}\tilde{x}_{it}'-\left[x_{it}-\mathbb{E}\left(\bar{x}_{i\cdot}\right)\right]\left[x_{it}-\mathbb{E}\left(\bar{x}_{i\cdot}\right)\right]\right\}\} \\
&\equiv H_{k,3}+H_{k,4}.
\end{aligned}$$

Let ω_1 and ω_2 be arbitrary nonrandom $p\times 1$ vectors such that $\|\omega_1\|=\|\omega_2\|=1$.By Assumptions D1(i)-(ii) and the Davydov inequality, we can readily show that

$$\mathbb{E}\left[\left(\omega_1'H_{k,3}\omega_2\right)^2\right]=\mathrm{Var}\left(\omega_1'H_{k,3}\omega_2\right)=\frac{1}{(N_kT)^2}\sum_{i\in G_k^0}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathbb{E}\left(\zeta_{1,it}\zeta_{1,is}\right)=O\left((N_kT)^{-1}\right).$$

where $\zeta_{1,it}=\omega_1'\left[x_{it}-\mathbb{E}\left(\bar{x}_{i\cdot}\right)\right]\left[x_{it}-\mathbb{E}\left(\bar{x}_{i\cdot}\right)\right]-\mathbb{E}\{\left[x_{it}-\mathbb{E}\left(\bar{x}_{i\cdot}\right)\right]\left[x_{it}-\mathbb{E}\left(\bar{x}_{i\cdot}\right)\right]'\}\omega_2$. It follows that $H_{k,3}=O_P\left((N_kT)^{-1/2}\right)$. Note that

$$\omega_1'H_{k,4}\omega_2=\frac{1}{N_kT}\sum_{i\in G_k^0}\sum_{t=1}^{T}\omega_1'x_{it}\left[\mathbb{E}\left(\bar{x}_{i\cdot}\right)-\bar{x}_{i\cdot}\right]'\omega_2=\frac{-1}{N_kT^2}\sum_{i\in G_k^0}\sum_{t=1}^{T}\sum_{t=1}^{T}\zeta_{2,its}$$

where $\zeta_{2,its}=\omega_1'[x_{it}-\mathbb{E}\left(x_{it}\right)]\left[x_{is}-\mathbb{E}\left(x_{is}\right)\right]'\omega_2$. By Assumptions D1(i)-(ii) and Lemma A.2(ii) in Gao (2007), we have

$$\mathbb{E}\left[\left(\omega_1'H_{k,4}\omega_2\right)^2\right]=\mathrm{Var}\left(\omega_1'H_{k,4}\omega_2\right)=\frac{1}{N_k^2T^4}\sum_{i\in G_k^0}\mathbb{E}\left[\left(\sum_{t=1}^{T}\sum_{s=1}^{T}\zeta_{2,its}\right)^2\right]=O\left(N_k^{-1}T^{-2}\right).$$

It follows that $H_{k,3}=O_P(N_k^{-1/2}T^{-1})$. Consequently, $\bar{\mathbb{H}}_{kNT}-\mathbb{H}_{kNT}=O_P\left((N_kT)^{-1/2}\right)=o_P\left(\nu_{NT}\right)$. This completes the proof of (i).

We now prove (ii). Let $\bar{\mathbb{B}}_{1kNT}=\mathbb{E}\left(\mathbb{B}_{1kNT}\right)=\frac{-1}{N_k^{1/2}T^{3/2}}\sum_{i\in G_k^0}\sum_{s=1}^{T}\sum_{t=1}^{T}\mathbb{E}\left(\varepsilon_{it}x_{is}\right)$. We prove (ii) by showing that (ii1) $\mathbb{B}_{1kNT}-\bar{\mathbb{B}}_{1kNT}=o_P(1)$, and (ii2) $\hat{\mathbb{B}}_{1kNT}-\bar{\mathbb{B}}_{1kNT}=o_P(1)$. For (ii1), we have

$$\begin{aligned}
\omega_1'\left(\mathbb{B}_{1kNT}-\bar{\mathbb{B}}_{1kNT}\right) &= \frac{-1}{N_k^{1/2}T^{3/2}}\sum_{i\in G_k^0}\sum_{s=1}^{T}\sum_{t=1}^{T}\omega_1'\left\{\varepsilon_{it}\left[x_{is}-\mathbb{E}\left(\bar{x}_{i\cdot}\right)\right]-\mathbb{E}\left(\varepsilon_{it}x_{is}\right)\right\} \\
&= \frac{-1}{N_k^{1/2}T^{3/2}}\sum_{i\in G_k^0}\sum_{s=1}^{T}\sum_{t=1}^{T}\zeta_{3,its},
\end{aligned}$$

where $\zeta_{3,its}=\omega_1'\left\{\varepsilon_{it}\left[x_{is}-\mathbb{E}\left(\bar{x}_{i\cdot}\right)\right]-\mathbb{E}\left(\varepsilon_{it}x_{is}\right)\right\}$. Then by Assumptions D1(i)-(ii) and Cauchy-Schwarz inequality,

$$\begin{aligned}
\mathbb{E}\left[\omega_1'\left(\mathbb{B}_{1kNT}-\bar{\mathbb{B}}_{1kNT}\right)\right]^2 &= \mathrm{Var}\left(\omega_1'\left(\mathbb{B}_{1kNT}-\bar{\mathbb{B}}_{1kNT}\right)\right)=\frac{1}{N_kT^3}\sum_{i\in G_k^0}\mathbb{E}\left(\sum_{s=1}^{T}\sum_{t=1}^{T}\zeta_{3,its}\right)^2 \\
&\leq \frac{2}{N_kT^3}\sum_{i\in G_k^0}\mathbb{E}\left(\sum_{s=1}^{T}\sum_{t=1}^{T}\omega_1'\varepsilon_{it}\left[x_{is}-\mathbb{E}\left(\bar{x}_{i\cdot}\right)\right]\right)^2+\frac{2}{N_kT^3}\sum_{i\in G_k^0}\left(\sum_{s=1}^{T}\sum_{t=1}^{T}\omega_1'\mathbb{E}\left(\varepsilon_{it}x_{is}\right)\right)^2.
\end{aligned}$$

For the first term, we can apply Lemma A.2(ii) in Gao (2007) and show that it is $O\left(T^{-1}\right)$. For the second term, we can apply the Davydov inequality directly to show that it is bounded from above by

$$\frac{2}{N_k T^3}\sum_{i\in G_k^0}\left(8T\left\|\varepsilon_{it}\right\|_{4q}\left\|\omega_1' x_{is}\right\|_{4q}\sum_{s=1}^{T}\alpha_i\left(s\right)^{(2q-1)/(2q)}\right)^2 = O\left(T^{-1}\right).$$

It follows that $\mathbb{B}_{1kNT}-\bar{\mathbb{B}}_{1kNT}=O\left(T^{-1/2}\right)=o_P\left(1\right).$

We now show (ii2), we first make the following decomposition:

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

where the $o_P\left(1\right)$ 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 $\hat{B}_{kNT,s}=o_P\left(1\right)$ for $s=1$, 2, 3, and 4.

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

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

where $\bar{\varepsilon}_{i\cdot}=\frac{1}{T}\sum_{t=1}^{T}\varepsilon_{it}$. Then

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

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

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

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 the Markov inequality. Let $\bar{B}_{kNT,1}(2) \equiv \frac{1}{N_k^{1/2} T^{3/2}} \sum_{i \in G_k^0} \sum_{s=1}^{T} \sum_{t=1}^{T} k_{M_T}(t,s)\, \omega' x_{is} \bar{\varepsilon}_{i\cdot}$, where ω is any $p \times 1$ nonrandom vector such that $\|\omega\| = 1$. Then by Assumptions D1(i), (iii) and (iv),

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

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

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

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 $\bar{B}_{kNT,2} = \frac{1}{N_k^{1/2} T^{3/2}} \sum_{i \in G_k^0} \sum_{s=1}^{T} \sum_{t=1}^{T} k_{M_T}(t,s)\left[x_{is} \varepsilon_{it} - \mathbb{E}\left(x_{is} \varepsilon_{it}\right)\right]$. By construction $\mathbb{E}(\bar{B}_{kNT,2}) = 0$. By Assumptions D1(ii)-(iii) and

Jensen inequality,

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

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

By Corollary 2.3 and the Davydov inequality,

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

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

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

This completes the proof of the proposition. ■

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

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

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

S2.2 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 (3.3) in particular) that

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

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 $B_{kNT} = \frac{1}{N_k^{1/2} T^{3/2}} \sum_{i \in G_k^0} \sum_{s=1}^T \sum_{t=1}^T \mathbb{E}\left(\Delta x_{is} z'_{is} z_{it} \Delta\varepsilon_{it}\right)$. Based on (3.3), in order to verify Assumption B3(iii) we also need to show

$$V_{kNT} = \frac{1}{N_k^{1/2} T^{1/2}} \sum_{i \in G_k^0} \sum_{t=1}^T \bar{Q}'_{i,z\Delta x} z_{it} \Delta\varepsilon_{it} \xrightarrow{D} N\left(0, C_k\right), \text{ and} \tag{S1}$$

$$R_{kNT} = \frac{1}{N_k^{1/2} T^{3/2}} \sum_{i \in G_k^0} \sum_{s=1}^T \sum_{t=1}^T \left\{\left[\Delta x_{is} z'_{is} - \mathbb{E}\left(\Delta x_{is} z'_{is}\right)\right] z_{it} \Delta\varepsilon_{it} - \mathbb{E}\left(\Delta x_{is} z'_{is} z_{it} \Delta\varepsilon_{it}\right)\right\} = o_P(1). \tag{S2}$$

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 (S2).

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

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

where $\Delta\tilde{\varepsilon}_{it} = \Delta y_{it} - \tilde{\alpha}'_{\tilde{G}_k} \Delta x_{it}$ for all $i \in \tilde{G}_k$,[3] $k_{M_T}(t,s)$ is as defined above: $k_{M_T}(t,s) = k^0_{M_T}(|t-s|)$ and $k^0_{M_T}(u)$ denotes the Bartlett kernel: $k^0_{M_T}(u) = (1 - |u|/M_T)\, \mathbf{1}\{|u| \le M_T\}$. Note that we also allow dynamic misspecification here. If one is sure that the model is dynamically correctly specified in the sense that $\mathbb{E}(\Delta\varepsilon_{it} | \mathcal{F}_{i,t-1}) = 0$ where $\mathcal{F}_{i,t-1} = \sigma(\Delta\varepsilon_{i,t-1}, \Delta x_{it-1}, z_{it}; \Delta\varepsilon_{i,t-2}, \Delta x_{it-2}, z_{i,t-1}; ...)$, 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_k^0 would be

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

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

Let $x_i \equiv (x_{i1}, ..., x_{iT})'$ and $\varepsilon_i \equiv (\varepsilon_{i1}, ..., \varepsilon_{iT})'$. We add the following assumption.

ASSUMPTION D2. *(i) For each $i = 1, ..., N$, $\{(\Delta x_{it}, z_{it}, \Delta\varepsilon_{it}) : t = 1, 2, ...\}$ 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 $\frac{1}{N_k} \sum_{i \in G_k^0} c_{\alpha,i}^{(2q-1)/(2q)} = O(1)$ and $\frac{1}{N_k} \sum_{i \in G_k^0} c_{\alpha,i}^{(q-1)/q} = O(1)$.*

(ii) (x_i, ε_i) are independent across $i \in G_k^0$ where $k = 1, ..., K_0$.

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

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

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

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

Proposition S2.2 *Suppose that the conditions of Theorem 3.4 hold. Suppose Assumption D2 holds. Then* $\tilde{A}_k^{-1}\tilde{B}_{kNT} - \bar{A}_k^{-1}B_{kNT} = o_P(1)$.

Proof. Noting that $\tilde{A}_k^{-1}\tilde{B}_{kNT} - \bar{A}_k^{-1}B_{kNT} = (\tilde{A}_k^{-1} - \bar{A}_k^{-1})B_{kNT} + (\tilde{A}_k^{-1} - \bar{A}_k^{-1})(\tilde{B}_{kNT} - B_{kNT}) + \bar{A}_k^{-1}(\tilde{B}_{kNT} - B_{kNT})$, $\bar{A}_k^{-1} = O(1)$, 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

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

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

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

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

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 Theorem 2.5. We prove (ii) by demonstrating that $B_{kNT,s} = o_P(1)$ for $s = 1, 2, 3, 4$.

First, noting that $\Delta\tilde{\varepsilon}_{it} - \Delta\varepsilon_{it} = (\alpha_k^0 - \tilde{\alpha}_{\tilde{G}_k})'\Delta x_{it}$, $\tilde{\alpha}_{\tilde{G}_k} - \alpha_k^0 = O_P\left((N_kT)^{-1/2}\right)$, and that $N_k/\tilde{N}_k =$

$1+o_P(1)$ by Corollary 3.3, we have

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

where $b_{kNT,1}=\frac{1}{N_kT^2}\sum_{i\in G_k^0}\sum_{|s-t|\leq M_T}^{T}\|\Delta x_{is}z_{is}'z_{it}(\Delta x_{it})'\|$. By the 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 $B_{kNT,2}=b_{kNT,2}N_k^{1/2}/\tilde{N}_k^{1/2}=b_{kNT,2}(1+o_P(1))$, where

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

Let ω be any $p\times1$ nonrandom vector such that $\|\omega\|=1$. Then $\mathbb{E}(\omega'b_{kNT,2})=0$. By Assumptions D2(ii)-(iv) and Jensen inequality,

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

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

By Corollary 3.3 and the Davydov inequality,

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

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

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

This completes the proof of the proposition. ■

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

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

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

Now, we demonstrate (S2). Let $\xi_{is} = \Delta x_{is} z_{is}' - \mathbb{E}\left(\Delta x_{is} z_{is}'\right)$ and $\eta_{it} = z_{it}\Delta\varepsilon_{it}$. Noting that $\mathbb{E}\left(\xi_{is}\right) = 0$ and $\mathbb{E}\left(\eta_{it}\right) = 0$, we have

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

It is trivial to show that $R_{kNT,1} = O_P\left(T^{-1}\right)$ by the 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

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

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, $S_{kNT,c} = O(1/T)$ under Assumption D2(iii). In case (a), without loss of generality (wlog) assume that $1 \leq t_1 < t_2 < t_3 < t_4 \leq T$ 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),

$$\begin{aligned} S_{NT,a1}^{(1)} &\le \frac{1}{N_kT^3}\sum_{i\in G_k^0}\sum_{t_1=1}^{T-3}\sum_{t_2=t_1+\max_{j\ge3}\{t_j-t_{j-1}\}}^{T-2}\sum_{t_3=t_2+1}^{T-1}\sum_{t_4=t_3+1}^{T}\left\|\xi_{it_1}\right\|_{4q}\left\|\eta_{it_2}\xi_{it_3}\eta_{it_4}\right\|_{4q/3}\alpha_i\left(t_2-t_1\right)^{(q-1)/q} \\ &\le \frac{C}{N_kT^3}\sum_{i=1}^{N}\sum_{t_1=1}^{T-3}\sum_{t_2=t_1+1}^{T-2}\left(t_2-t_1\right)^2\alpha_i\left(t_2-t_1\right)^{(q-1)/q} \\ &\le \frac{1}{N_kT}\sum_{i=1}^{N}\sum_{\tau=1}^{\infty}\tau\alpha_i\left(\tau\right)^{(q-1)/q}=O(T^{-1}). \end{aligned}$$

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 $1\le t_1<t_2<t_3\le T$ 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)

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

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

S3 Numerical Algorithm and Additional Simulation Results

S3.1 Numerical Algorithm

In this section, we propose an iterative algorithm to obtain the PPL estimates $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ in Section 2. A similar algorithm applies for PGMM estimation. Documented computer code is available online.

Step 1 Start with an initial value $\hat{\boldsymbol{\alpha}}^{(0)}=(\hat{\alpha}_1^{(0)},...,\hat{\alpha}_K^{(0)})$ and $\hat{\boldsymbol{\beta}}^{(0)}=(\hat{\beta}_1^{(0)},...,\hat{\beta}_N^{(0)})$ such that $\sum_{i=1}^N\|\hat{\beta}_i^{(0)}-\hat{\alpha}_k^{(0)}\|\neq0$ for each $k=2,...,K$. Set the iteration index $r=1$.

Step 2 Given $\hat{\boldsymbol{\alpha}}^{(r-1)}=(\hat{\alpha}_1^{(r-1)},...,\hat{\alpha}_K^{(r-1)})$ and $\hat{\boldsymbol{\beta}}^{(r-1)}=(\hat{\beta}_1^{(r-1)},\ ...,\hat{\beta}_N^{(r-1)})$, we first choose $(\boldsymbol{\beta},\alpha_1)$ to minimize

$$Q_{1NT,\lambda_1}^{(r,1,K)}\left(\boldsymbol{\beta},\alpha_1\right)=Q_{1,NT}\left(\boldsymbol{\beta}\right)+\frac{\lambda_1}{N}\sum_{i=1}^{N}\left\|\beta_i-\alpha_1\right\|\Pi_{k\neq1}^{K}\left\|\hat{\beta}_i^{(r-1)}-\hat{\alpha}_k^{(r-1)}\right\|,$$

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

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

and obtain the updated estimate $(\hat{\boldsymbol{\beta}}^{(r,2)}, \hat{\alpha}_2^{(r)})$ of $(\boldsymbol{\beta}, \alpha_2)$. Repeat this procedure until we choose $(\boldsymbol{\beta}, \alpha_K)$ to minimize

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

and obtain the updated estimate $(\hat{\boldsymbol{\beta}}^{(r,K)}, \hat{\alpha}_K^{(r)})$ of $(\boldsymbol{\beta}, \alpha_K)$. Let $\hat{\boldsymbol{\alpha}}^{(r)} = (\hat{\alpha}_1^{(r)}, ..., \hat{\alpha}_K^{(r)})$ and $\hat{\mathcal{Q}}_{1NT}^{(r,K)} = \sum_{k=1}^{K} Q_{1NT,\lambda_1}^{(r,k,K)}(\hat{\boldsymbol{\beta}}^{(r,k)}, \hat{\boldsymbol{\alpha}}_k^{(r)})$. Update the iteration index from r to $r+1$.

Step 3 Repeat Step 2 until a convergence criterion is achieved, e.g., when

$$\hat{\mathcal{Q}}_{1NT}^{(r-1,K)} - \hat{\mathcal{Q}}_{1NT}^{(r,K)} < \epsilon_{\text{tol}} \quad \text{and} \quad \frac{\sum_{k=1}^{K} \left\|\hat{\alpha}_k^{(r)} - \hat{\alpha}_k^{(r-1)}\right\|^2}{\sum_{k=1}^{K} \left\|\hat{\alpha}_k^{(r-1)}\right\|^2 + 0.0001} < \epsilon_{\text{tol}},$$

where ϵ_{tol} is some prescribed tolerance level (e.g., 0.0001). Define the final iterative estimate of $\boldsymbol{\alpha}$ as $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1^{(R)}, ..., \hat{\alpha}_K^{(R)})$ for a sufficiently large R such that the convergence criterion is met. Intuitively, individual i is classified to group $\hat{G}_k$ if $\hat{\beta}_i^{(R,k)} = \hat{\alpha}_k$; otherwise, $\hat{\beta}_i$ is assigned to be the $\alpha_k^{(R)}$ that is closest to some $\hat{\beta}_i^{(R,l)}$, $l = 1, \ldots, K$. In either case, we can write the individual estimate as $\hat{\beta}_i = \hat{\alpha}_{k^*}^{(R)}$, where $k^* = \operatorname{argmin}_{k\in\{1,\ldots,K\}} \|\hat{\beta}_i^{(R,l^*(k))} - \hat{\alpha}_k^{(R)}\|$ and $l^*(k) = \operatorname{argmin}_{l\in\{1,\ldots,K\}} \|\hat{\beta}_i^{(R,l)} - \hat{\alpha}_k^{(R)}\|$.

S3.2 Convexity, Choice of Initial Value, and Convergence of the Algorithm

The optimization of $Q_{1NT,\lambda_1}^{(r,k,K)}(\boldsymbol{\beta}, \alpha_k)$ is conducted on the $(Np + p)$-dimensional parameter space for $(\boldsymbol{\beta}, \alpha_k)$. When N is non-trivial, this is a high-dimensional optimization problem. Obviously, in the penalty term $\beta_1, \cdots, \beta_N$ and α_k are jointly convex, given $\Pi_{l=k+1}^{K} \left\|\hat{\beta}_i^{(r-1)} - \hat{\alpha}_l^{(r-1)}\right\|$ and $\Pi_{l=1}^{k-1} \left\|\hat{\beta}_i^{(r)} - \hat{\alpha}_l^{(r)}\right\|$ for each $i = 1, \ldots, N$. If $Q_{1,NT}(\boldsymbol{\beta})$ is convex in $\boldsymbol{\beta}$, then $Q_{1NT,\lambda_1}^{(r,k,K)}(\boldsymbol{\beta}, \alpha_k)$, as the summation of $Q_{1,NT}(\boldsymbol{\beta})$ and the penalty, is also convex in $(\boldsymbol{\beta}, \alpha_k)$. Convexity can substantially reduce the computational burden of high-dimensional optimization.

A convex $Q_{1,NT}(\boldsymbol{\beta})$ is common in panel data models. Convexity apparently holds in the linear models in Examples 1 and 2. It also holds in the nonlinear models in Example 3 with $F(\cdot)$ as the standard logistic or normal CDF, and in Example 4 after re-parameterizing the original parameter $(\beta_i, \mu_i, \sigma_\varepsilon^2)$ into $(\theta_{1i} = \beta_i/\sigma_\varepsilon^2$, $\theta_{2i} = \mu_i/\sigma_\varepsilon^2$, $\theta_3 = 1/\sigma_\varepsilon^2)$. We utilize the convexity throughout our numerical works.

Given the convexity in each substep (r, k), the proposed algorithm consists of a sequence of convex problems implemented in an iterative manner. In particular, the only difference between the standard Lasso and a single substep of PPL is that Lasso shrinks the coefficients to a known center (zero), while the center of PPL is determined in the convex programming. Thus a PPL iteration has the same computational complexity as Lasso, which is $O(N^3T)$ in our context of panel linear regression (Efron, Hastie and Johnstone, 2004, p.443). The computational cost of a single iteration is minimal.

Since the additive-multiplicative penalty is not jointly convex in all the parameters $(\boldsymbol{\beta}, \boldsymbol{\alpha})$, we can take advantage of convexity in each substep for $(\boldsymbol{\beta}, \alpha_k)$ but not simultaneously for $(\boldsymbol{\beta}, \boldsymbol{\alpha})$. As a consequence of the non-convexity, the sequence $\hat{\mathcal{Q}}_{1NT}^{(r,K)}, r = 1, \ldots, R$, depends on the initial value $(\hat{\boldsymbol{\alpha}}^{(0)}, \hat{\boldsymbol{\beta}}^{(0)})$, and $\hat{\mathcal{Q}}_{1NT}^{(R,K)}$ might terminate at a local minimum but not a global minimum.

A natural initial value is to set $(\alpha_k^{(0)} = 0)_{k=1}^K$ and each $\beta_i^{(0)}$ as the QMLE from the i-th individual time series $(w_{i1}, \ldots, w_{iT})$. Denote this particular choice $(\boldsymbol{\beta}^{\text{init}}, \boldsymbol{\alpha}^{\text{init}})$, and we use it in all the simulations as well as applications, if not explicitly stated otherwise. As we compare it with other possible choices, for example $(\alpha_k^{(0)} = 0)_{k=1}^K$ and $(\beta_i^{(0)} = 1)_{i=1}^N$, starting at $(\boldsymbol{\beta}^{\text{init}}, \boldsymbol{\alpha}^{\text{init}})$ often makes the algorithm converge in fewer iterations.

Although a formal investigation of the algorithm's computational complexity to attain the global optimum is beyond the scope of the paper, we explore its numerical convergence and sensitivity to initial values through a numerical example. We use the real data of savings rate in Section 5.1, and apply PLS and PGMM given the number of groups and the tuning parameters selected according to the IC. The left subgraph in Figure S1 shows the path of $\hat{\mathcal{Q}}_{1NT}^{(r,K)}, r = 1, \ldots, R$, and the right subgraph displays its PGMM counterpart. Each of the ten paths is associated with a different starting point. First, the bold black curve is the path that starts at $(\boldsymbol{\beta}^{\text{init}}, \boldsymbol{\alpha}^{\text{init}})$. Next, we perturb the initial value to be $((\beta_i^{\text{init}} + e_i)_{i=1}^N, \boldsymbol{\alpha}^{\text{init}})$, where e_i is a vector of p elements, each of which is randomly drawn from Uniform$(-1, 1)$ and is independent across i. This is a substantial perturbation, in view of the magnitude of the estimates in Table 3. We use the perturbed initial values to generate the other nine curves.

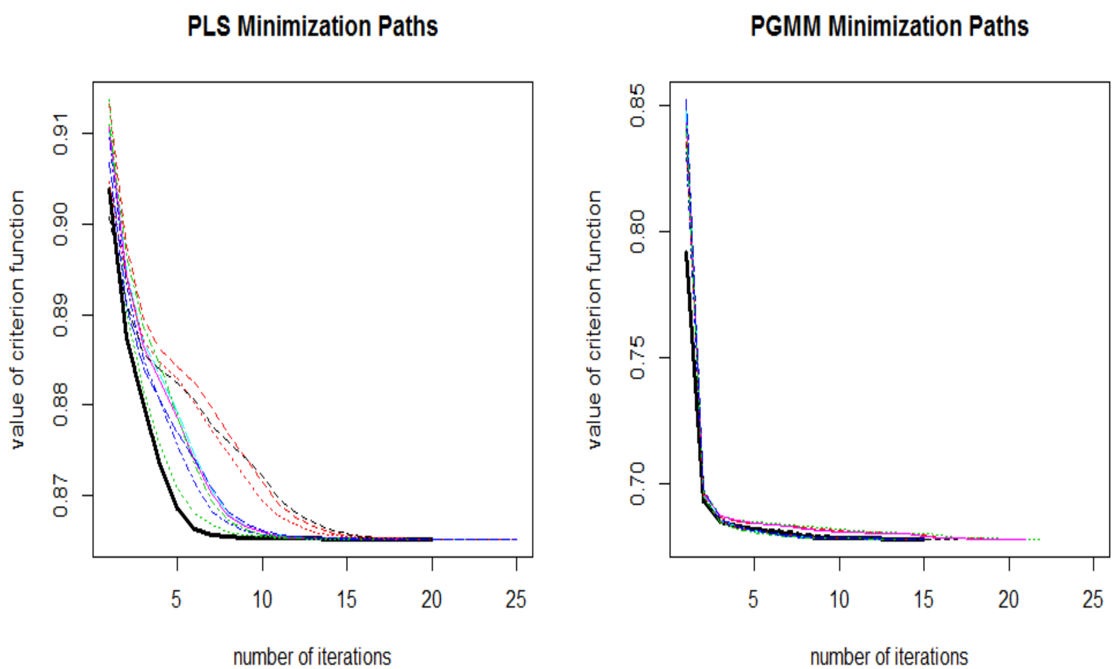

Figure S1: PLS and PGMM paths starting at different initial values

Figure S1 illustrates the robustness of C-Lasso to initial values. We observe in each subgraph that the criterion functions descend fast in the first few iterations, and then the paths turn almost flat until the tolerance level is reached. All paths converge to the same value of the criterion function in this experiment.

S3.3 Additional Simulation Results

In this section we carry out two more simulation exercises, one using PGMM to estimate a static panel model with endogenous regressors as in DGP 4 below, and the other using PLS to estimate the linear panel AR(1) in DGP2.

DGP 4 (Linear static panel with endogeneity.) We maintain the linear panel structure model with two explanatory variables as in DGP 1, but the first regressors is endogenous as it is generated from the following underlying reduced-form equation: $x_{it1} = 0.2\mu_i^0 + 0.5z_{it1} + 0.5z_{it2} + 0.5e_{it}$, where z_{it1} and z_{it2} are two excluded instrumental variables, and the reduced-form error e_{it} and the structural-equation idiosyncratic shock ε_{it} follow a bivariate normal distribution:

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

The second regressor x_{it2} is exogenous, and $(x_{it2}, z_{it1}, z_{it2}) \sim$ IID $N(0, I_3)$ is independent of $(e_{it}, \varepsilon_{it})$. All variables are independent across i and t. The econometrician observes $(y_{it}, x_{it1}, x_{it2}, z_{it1}, z_{it2})$. The true coefficients of the three groups are $(0.2, 1.8)$, $(1, 1)$ and $(1.8, 0.2)$, respectively.

We report the statistics in Tables S1 and S2, which correspond to Tables 1 and 2, respectively, in the main text. The choice of tuning parameters are exactly the same as described in Section 4. When we compare PLS estimation with PGMM in DGP 2, we find that the PLS works better in determining the correct number of groups and in classifying the individual units. The 95% coverage probabilities are comparable to those of PGMM when $T = 50$, but are lower than PGMM when T is small. Similar to PPL in DGP 3, the lower coverage probabilities is caused by the bias. The analytical bias correction removes the bias asymptotically, but the effect is limited when T is small, as is shown in the oracle. The post-Lasso has larger coverage probability than the oracle, as the estimated standard deviation is inflated by a few misclassified units.

Table S1: Frequency of selecting $K = 1, \ldots, 5$ groups when $K_0 = 3$

N	T	DGP 4					DGP 2 (PLS)				
		1	2	**3**	4	5	1	2	**3**	4	5
100	15	0	0.022	**0.902**	0.076	0	0	0.106	**0.894**	0	0
100	25	0	0	**0.966**	0.028	0.006	0	0	**1**	0	0
100	50	0	0	**0.996**	0.004	0	0	0	**1**	0	0
200	15	0	0	**0.940**	0.058	0.002	0	0	**1**	0	0
200	25	0	0	**0.950**	0.046	0.004	0	0	**1**	0	0
200	50	0	0	**0.994**	0.006	0	0	0	**1**	0	0

Table S3 reports the RMSE and bias of $\boldsymbol{\alpha}_1$ from post-Lasso and C-Lasso under the true K_0 and the IC-determined $\hat{K}$ (or $\tilde{K}$ for PGMM). These estimates are bias corrected whenever necessary in the DGPs. For example, the RMSE of PPL under K_0 is calculated as $\left(\frac{1}{S}\sum_{s=1}^{S}\sum_{k=1}^{K_0}\frac{\hat{N}_k^{(s)}}{N}\left(\hat{\alpha}_{k,1}^{(s)}(K_0,\lambda_1) - \alpha_{k,1}^0\right)^2\right)^{1/2}$, where s and S are the index and the total number of simulation replications, respectively, and $\hat{N}_k^{(s)} = \sum_{i=1}^{N}\mathbf{1}\{i \in \hat{G}_k^{(s)}(K,\lambda_1)\}$ is the estimated number of units in the k-th group. This quantity differs from its counterpart in Table 2 as each group-specific estimate is weighted by $\hat{N}_k^{(s)}/N$, instead of N_k/N, to take into account the uncertainty in classification. The bias is computed similarly. The post-Lasso's RMSE and bias under the known K_0 are close to the oracle. The performance of C-Lasso is in general comparable to that

Table S2: Classification and point estimation of $\boldsymbol{\alpha}_1$ in additional simulations

	N	T	% of correct classification	Post-Lasso RMSE	Post-Lasso Bias	Post-Lasso Coverage	Oracle RMSE	Oracle Bias	Oracle Coverage
DGP 4	100	15	0.8287	0.1583	0.0462	0.7850	0.0806	0.0018	0.9344
	100	25	0.9281	0.0883	0.0195	0.8880	0.0617	0.0009	0.9380
	100	50	0.9885	0.0517	0.0075	0.9406	0.0437	-0.0012	0.9422
	200	15	0.8378	0.1155	0.0484	0.7860	0.0577	-0.0016	0.9454
	200	25	0.9320	0.0643	0.0199	0.8742	0.0436	0.0001	0.9506
	200	50	0.9881	0.0364	0.0074	0.9356	0.0311	-0.0005	0.9450
DGP 2	100	15	0.8907	0.0413	0.0061	0.9148	0.0352	0.0041	0.8524
(PLS)	100	25	0.9511	0.0261	0.0041	0.9710	0.0241	0.0028	0.9076
	100	50	0.9908	0.0160	0.0015	0.9908	0.0156	0.0013	0.9334
	200	15	0.8949	0.0294	0.0064	0.9154	0.0253	0.0052	0.8576
	200	25	0.9520	0.0188	0.0037	0.9714	0.0178	0.0036	0.8808
	200	50	0.9912	0.0113	0.0017	0.9934	0.0111	0.0015	0.9282

of post-Lasso, although C-Lasso appears to have larger RMSE in the PGMM estimation of DGP 2, where it does not enjoy the oracle property.

When $K \neq K_0$, we generalize the definition of the set of true group-specific parameters. For $K < K_0$, we shrink $\boldsymbol{\alpha}_1^0 = (\alpha_{1,1}^0, \ldots, \alpha_{K_0,1}^0)$ into a K-element subset $\boldsymbol{\alpha}_1^0(K)$. For $K > K_0$, we augment $\boldsymbol{\alpha}_1^0$ by adding $K - K_0$ elements choosing from $\alpha_{k,1}^0, \ldots, \alpha_{K_0,1}^0$ so that the resulting $\boldsymbol{\alpha}_1^0(K)$ contains $\boldsymbol{\alpha}_1^0$. Elements are eliminated or concatenated in each replication to fit $\hat{\boldsymbol{\alpha}}(\hat{K}^{(s)}, \lambda_1)$. In this scenario, the RMSE is calculated as $\left(\frac{1}{S}\sum_{s=1}^{S}\sum_{k=1}^{\hat{K}^{(s)}}\frac{\hat{N}_k^{(s)}}{N}\left(\hat{\alpha}_{k,1}(\hat{K}^{(s)}, \lambda_1) - \alpha_{k,1}^0(\hat{K}^{(s)})\right)^2\right)^{1/2}$. According to the simulation, the effect of not knowing K_0 is noticeable when $T = 15$ in the linear models and $T = 25$ in the nonlinear model, but it does not necessarily enlarges the RMSE, for the estimator under K_0 is also noisy when T is small. The discrepancy of the RMSE and bias between K_0 and $\hat{K}$ (or $\tilde{K}$) quickly vanishes when T grows.

S4 Additional Application Results

S4.1 More on Savings Rate Modeling and Classification

All data are downloaded from the World Bank.[4] We extract all countries with all the variables in (5.1) available. Using the time span 1995–2010, we were able to construct a balanced panel of 57 countries. We remove one outlier Bulgaria, whose 1997 economic collapse produced hyperinflation in the CPI that significantly distorted the overall mean and the standard deviation. In total we collect 56 countries. The summary statistics are shown in Table S4.

In the implementation, we scale-normalize all the variables for each individual unit to guarantee that the coefficients are comparable. Moreover, in PGMM we use Δy_{t-2} and a constant as two excluded IVs. Although the constant is uncorrelated with the endogenous variable, adding it here stabilizes the post-Lasso estimation in finite samples.

Table S5 displays the group membership. The country names in bold are the 47 coincidences of PLS and PGMM classification out of the total 56 countries.

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

Table S3: Estimation of $\boldsymbol{\alpha}_1$ by post-lasso and C-Lasso under K_0 and $\hat{K}$ or $\tilde{K}$

			Post-Lasso				C-Lasso				Oracle	
			$K = K_0$		$K = \hat{K}$		$K = K_0$		$K = \hat{K}$			
	N	T	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias
DGP 1	100	15	0.0596	0.0108	0.0829	0.0092	0.0619	0.0133	0.0839	0.0120	0.0463	0.0012
	100	25	0.0385	0.0019	0.0385	0.0019	0.0396	0.0040	0.0396	0.0040	0.0353	0.0001
	100	50	0.0249	0.0000	0.0249	0.0000	0.0255	0.0011	0.0255	0.0011	0.0245	-0.0002
	200	15	0.0434	0.0079	0.1373	0.0081	0.0457	0.0107	0.1353	0.0114	0.0324	-0.0013
	200	25	0.0273	0.0015	0.0273	0.0015	0.0280	0.0040	0.0280	0.0040	0.0250	-0.0006
	200	50	0.0174	-0.0001	0.0174	-0.0001	0.0181	0.0011	0.0181	0.0011	0.0171	-0.0002
DGP 2	100	15	0.0848	-0.0090	0.0787	-0.0016	0.1311	-0.0372	0.1188	-0.0250	0.0502	-0.0037
(PGMM)	100	25	0.0556	-0.0055	0.0561	-0.0051	0.1042	-0.0267	0.1045	-0.0255	0.0351	0.0011
	100	50	0.0278	-0.0012	0.0278	-0.0012	0.0418	-0.0130	0.0418	-0.0130	0.0242	-0.0010
	200	15	0.0712	-0.0141	0.0743	-0.0145	0.1491	-0.0399	0.1483	-0.0383	0.0352	-0.0017
	200	25	0.0333	-0.0051	0.0333	-0.0051	0.0932	-0.0284	0.0932	-0.0284	0.0252	-0.0006
	200	50	0.0193	-0.0014	0.0193	-0.0014	0.0277	-0.0134	0.0277	-0.0134	0.0164	0.0000
DGP 3	100	25	0.1722	0.0587	0.1516	0.0727	0.2154	0.0615	0.1641	0.0688	0.1077	0.0114
	100	50	0.0853	0.0379	0.0878	0.0383	0.1178	0.0487	0.1191	0.0489	0.0752	0.0090
	200	25	0.1342	0.0483	0.1401	0.0649	0.1826	0.0487	0.1441	0.0573	0.0821	0.0116
	200	50	0.0632	0.0264	0.0632	0.0264	0.0948	0.0372	0.0948	0.0372	0.0573	0.0121
DGP 4	100	15	0.1691	0.0487	0.1803	0.0376	0.2148	0.1087	0.2102	0.0941	0.0806	0.0018
	100	25	0.0724	0.0189	0.1217	0.0207	0.0882	0.0523	0.1323	0.0539	0.0617	0.0009
	100	50	0.0450	0.0031	0.0645	0.0042	0.0532	0.0204	0.0707	0.0215	0.0437	-0.0012
	200	15	0.1271	0.0512	0.1348	0.0466	0.1777	0.1128	0.1793	0.1074	0.0577	-0.0016
	200	25	0.0513	0.0153	0.1392	0.0235	0.0720	0.0498	0.1485	0.0577	0.0436	0.0001
	200	50	0.0314	0.0036	0.0549	0.0049	0.0399	0.0221	0.0602	0.0234	0.0311	-0.0005
DGP 2	100	15	0.0482	0.0081	0.0487	0.0065	0.0747	0.0297	0.0715	0.0254	0.0352	0.0041
(PLS)	100	25	0.0263	0.0043	0.0263	0.0043	0.0418	0.0189	0.0418	0.0189	0.0241	0.0028
	100	50	0.0160	0.0016	0.0160	0.0016	0.0218	0.0085	0.0218	0.0085	0.0156	0.0013
	200	15	0.0295	0.0064	0.0295	0.0064	0.0567	0.0293	0.0567	0.0293	0.0253	0.0052
	200	25	0.0188	0.0037	0.0188	0.0037	0.0307	0.0174	0.0307	0.0174	0.0178	0.0036
	200	50	0.0113	0.0017	0.0113	0.0017	0.0171	0.0084	0.0171	0.0084	0.0111	0.0015

Table S4: Summary statistics for the savings dataset

	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

Table S5: Estimated group membership

PLS	PGMM
Group 1: (31 countries) **Armenia**, **Australia**, **Bahamas**, **Belarus**, **Bolivia**, **Botswana**, **Cape Verde**, **China**, **Czech**, Guatemala, **Honduras**, **Hungary**, **Indonesia**, **Israel**, **Italy**, **Japan**, **Jordan**, **Latvia**, **Malawi**, **Malaysia**, **Mauritius**, **Mexico**, Mongolia, **Panama**, **Paraguay**, **Philippines**, **Romania**, **South Africa**, **Sri Lanka**, **Thailand**, **Ukraine**	Group 1: (36 countries) **Armenia**, **Australia**, **Bahamas**, **Belarus**, **Bolivia**, **Botswana**, **Cape Verde**, **China**, **Czech**, Egypt, **Honduras**, **Hungary**, India, **Indonesia**, **Israel**, **Italy**, **Japan**, **Jordan**, Kenya, **Latvia**, **Malawi**, **Malaysia**, Malta, **Mauritius**, **Mexico**, **Panama**, **Paraguay**, **Philippines**, **Romania**, **South Africa**, **Sri Lanka**, Swaziland, Switzerland, **Thailand**, **Ukraine**, United Kingdom
Group 2: (25 countries) **Bangladesh**, **Canada**, **Costa Rica**, **Dominican**, Egypt, **Guyana**, **Iceland**, India, Kenya, **Korea (Rep.)**, **Lithuania**, Malta, **Netherlands**, **Papua New Guinea**, **Peru**, **Russian**, **Singapore**, Swaziland, Switzerland, **Syrian**, **Tanzania**, **Uganda**, United Kingdom, **United States**, **Uruguay**	Group 2: (20 countries) **Bangladesh**, **Canada**, **Costa Rica**, **Dominican**, Guatemala, **Guyana**, **Iceland**, **Korea (Rep.)**, **Lithuania**, Mongolia, **Netherlands**, **Papua New Guinea**, **Peru**, **Russian**, **Singapore**, **Syrian**, **Tanzania**, **Uganda**, **United States**, **Uruguay**

S4.2 More on the Civil War Application

The replication data of Fearon and Laitin (2003) can be downloaded from Fearon's personal web page.[5] The data span from 1945–1998,[6] but the panel is highly unbalanced. Following Collier and Hoeffler (2004), Djankov and Reynal-Querol (2010) and Blattman and Miguel (2010), we choose 1960 as the staring year to generate a balanced panel of $N = 38$, as many countries' civil war incidence are always 0 or 1 between 1960 and 1998.

In the regression, the dependent variable is the civil war incidence, and the explanatory variables are the lagged civil war incidence, the one-period difference of log GDP per capita and the one-period difference of log population. Moreover, in views of the natural scaling of the binary variable, we keep the original dependent variable and the lagged dependent variable. For the other two continuously distributed variables, we follow the practice as in the savings rate application to scale-normalize each time series by the individual sample standard deviation. To ensure that the estimated coefficients are comparable, we further multiply these two scale-normalized variables by the overall standard deviation of the lagged dependent variable so that all the explanatory regressors are of the same scale. Furthermore, the Probit regressions for the individual time series are unstable in those countries with only 1 or 2 incidences. Therefore the C-Lasso initial values are set as the pooled FE Probit coefficient estimates.

The summary statistics are displayed in Table S6. Membership is reported under "high-occurrence" and

[5] `https://www.stanford.edu/group/fearon-research/cgi-bin/wordpress/wp-content/uploads/2013/10/apsr03repdata.zip`

[6] The original data end at 1999, but no population information is provided for any country in the last year.

Table S6: Summary statistics for the civil war dataset

	mean	median	s.e.	min	max
Civil war incidence	0.352	0	0.478	0	1
GDP per capita growth	0.020	0.024	0.040	-0.811	0.306
Population growth	0.012	0.015	0.076	-0.507	0.661

"low-occurrence" groups with results as follows.

High-occurrence group (23 countries): Guatemala, Peru, Argentina, Mali, Senegal, Chad, Congo (Dem.), Congo (Rep.), Somalia, Morocco, Sudan, Turkey, Iraq, Lebanon, Afghanistan, China, Pakistan, Sri Lanka, Nepal, Cambodia, Laos, Philippines, Indonesia

Low-occurrence group (15 countries): Haiti, Dominican, El Salvador, Nicaragua, UK, Yugoslavia, Cyprus, Russia, Liberia, Nigeria, Central African Republic, Ethiopia, South Africa, Iran, Jordan

S4.3 Linear Dynamic Modeling of Democracy

In this section, we use the data provided by Bonhomme and Manresa (2015) to revisit the link between income growth and democracy across countries. Following BM's Equation (22), we specify a linear dynamic model, where the dependent variable is a country's democracy index (measured by Freedom House indicator between 0 (the lowest) and 1 (the highest)), and the explanatory variables are the first-order lagged democracy index and the income (measured by the logarithm of GDP per capita).

The dataset contains a balanced panel of 84 countries and 8 periods at a five year interval over 1965–2000. We use PLS to estimate the model in this short panel. Many developed countries, such as the United States or United Kingdom, kept their democracy index at the highest level throughout the time. Due to the lack of within-group variation in these countries, we scale normalize each variable by its pooled standard deviation. This standardization makes sure that the parameter y_{it-1} can still be interpreted as the autoregressive coefficient, and the magnitude is comparable with the income coefficient.

Table S7: Summary statistics for the democracy dataset

	mean	median	s.e.	min	max
Democracy index	0.5760	0.6667	0.3712	0	1
GDP per capita (in logarithm)	8.2981	8.3039	1.0685	6.0937	10.4450

Following practice in the simulation, the IC with $\rho_{1NT} = \frac{2}{3}(NT)^{-1/2}$ picks out $K = 3$ and $c_{\lambda_1} = 1.20$ in all combinations of $K = 1, \ldots, 5$ and c_{λ_1} in a geometrically increasing sequence of 10 points in $(0.2, \ldots, 2)$. Under $K = 3$ and $c_{\lambda_1} = 1.20$, C-Lasso categorizes the 84 countries into the following groups:

Group 1 (30 countries): Belgium, Bolivia, Brazil, Canada, Dominican, Ecuador, El Salvador, Finland, Guatemala, Guinea, Iceland, Indonesia, Italy, Japan, Jordan, Luxembourg, Mali, Morocco, Nepal, Panama, Peru, Philippines, Portugal, Romania, South Africa, Thailand, Turkey, United Kingdom, Uruguay, Venezuela

Group 2 (36 countries): Algeria, Argentina, Australia, Austria, Barbados, Burkina Faso, Burundi, Cameroon, Chile, China, Colombia, Costa Rica, Cote d'Ivoire, Denmark, Egypt, France, Gabon, Ghana, Greece, India, Iran, Israel, Jamaica, Kenya, Malawi, Malaysia, Mexico, Nigeria, Norway, Paraguay, Rwanda, Spain, Sweden, Togo, Trinidad and Tobago, United States

Group 3 (18 countries): Benin, Chad, Congo (Rep.), Honduras, Ireland, Korea (Rep.), Madagascar, Netherlands, New Zealand, Nicaragua, Niger, Sri Lanka, Switzerland, Syrian, Tanzania, Tunisia, Uganda, Zambia

The post-Lasso and pooled FE estimates are shown in Table S8. We focus on the coefficient for income. The common FE coefficient is positive and significant. The positive effect is echoed by Groups 1 and 2, but contrasts with Group 3, which consists mainly of low-income and low-democracy nations combined with a few selected OECD countries. OECD countries such as Ireland, Netherlands, New Zealand and Switzerland maintained their democracy index at 1 throughout the sample period. The lack of variation in the dependent variable makes them uninformative about the income coefficient.

Table S8: PLS estimation results

	Pooled FE		PLS					
			Group 1		Group 2		Group 3	
	coef.	s.e.	coef.	s.e.	coef.	s.e.	coef.	s.e.
Lagged democracy	0.4993^{***}	0.0491	0.5141^{***}	0.0643	0.0954	0.0733	-0.0543	0.0521
Income	0.2552^{***}	0.0489	0.6545^{***}	0.0930	0.1550^{***}	0.0448	-0.5542^{***}	0.0860

Note: ***1% significant, ** 5% significant, * 10% significant

REFERENCE

BLATTMAN C. AND E. MIGUEL (2010): "Civil Wars," *Journal of Economic Literature* 48, 3-57.

BONHOMME, S., AND E. MANRESA (2015): "Grouped Patterns of Heterogeneity in Panel Data," *Econometrica* 83, 1147-1184.

COLLIER, P. AND A. HOEFFLER (2004): "Greed and Grievance in Civil Wars," *Oxford Economic Papers* 56, 563-595.

DJANKOV, S. AND M. REYNAL-QUEROL (2010): "Poverty and Civil War: Revisiting the Evidence," *Review of Economics and Statistics* 92, 1035-1041.

EFRON, B., T. HASTIE, I. JOHNSTONE and R. TIBSHIRANI (2004), "Least Angle Regression," *Annals of Statistics* 32, 407-499.

FEARON, J.D. AND D. D. LAITIN (2003): "Ethnicity, Insurgency, and Civil War," *The American Political Science Review* 97, 75-90.

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

HAHN, J., AND W. NEWEY (2004): "Jackknife and Analytical Bias Reduction for Nonlinear Panel Models," *Econometrica* 72, 1295-1319.

HALL, P. AND C.C. HEYDE (1980): *Martingale Limit Theory and Its Applications.* Academic Press, New York.

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.

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