Econometrica, Vol. 84, No. 6 (November, 2016), 2215–2264

IDENTIFYING LATENT STRUCTURES IN PANEL DATA

BY LIANGJUN SU, ZHENTAO SHI, AND PETER C. B. PHILLIPS[1]

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 in both classification and estimation. Empirical applications to both linear and nonlinear models are presented.

KEYWORDS: 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.

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, Chapter 6)). It is therefore widely acknowledged that an important feature of good empirical modeling is to control for heterogeneity in the data as well as for potential heterogeneity in the response mechanisms that figure within the model. Since heterogeneity is a latent feature of the data and its extent is unknown a priori, respecting the potential influence of heterogeneity on model specification is a

[1]The authors thank a co-editor and three anonymous referees for many constructive comments on the previous version of the paper. They also thank Stéphane Bonhomme, Xiaohong Chen, Han Hong, Cheng Hsiao, Arthur Lewbel, 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 MOE2012-T2-2-021 and the funding support provided by the Lee Kong Chian Fund for Excellence. Phillips acknowledges NSF support under Grants SES-0956687 and SES-1285258. Address correspondence to: Liangjun Su.

 DOI: 10.3982/ECTA12560

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) considered a panel structure model where individuals are grouped according to some external classification, geographic location, or observable explanatory variables; Ando and Bai (2015) considered 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) considered a *parametric* finite mixture linear panel data model, and Kasahara and Shimotsu (2009) and Browning and Carro (2014) studied 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) considered linear panel data models where the slope coefficients have latent group structure. They modified the K-means algorithm to estimate the models but did not provide any inference theory. Bonhomme and Manresa (2015, BM hereafter) considered a linear panel data model where the additive fixed effects have group structure and applied the K-means algorithm to estimate the model and study its asymptotic properties. Ando and Bai (2016) extended BM's approach to allow for group structure among the interactive fixed effects. In addition, Phillips and Sul (2007a) developed an algorithm for determining group clusters that relies on the estimation of evaporating trend functions to determine convergence clusters. Hahn and Moon (2010) argued that the group structure has sound foundations in game theory or macroeconomic models where multiplicity of Nash equilibria is expected and they considered 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 heterogeneous 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 $\{\beta_i, i = 1, \ldots, N\}$ 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, for example, 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 1 (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 (Su, Shi, and Phillips (2016)).

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 $i = 1, \ldots, N$ and $t = 1, \ldots, T$, it is proposed to use fixed effects quasi-maximum likelihood to estimate the unknown parameters by solving the minimization problem

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

Here $-\psi(w_{it};\beta_i,\mu_i)$ 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

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

Here $\alpha_j^0\neq\alpha_k^0$ for any $j\neq k$, $\bigcup_{k=1}^{K_0}G_k^0=\{1,2,\ldots,N\}$, and $G_k^0\cap G_j^0=\emptyset$ 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 kth that 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 that 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 $\boldsymbol{\alpha}\equiv(\alpha_1,\ldots,\alpha_{K_0})$ and $\boldsymbol{\beta}\equiv(\beta_1,\ldots,\beta_N)$. 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 $\psi(w_{it};\beta_i,\mu_i)=\frac{1}{2}(y_{it}-\beta_i'x_{it}-\mu_i)^2$ and $w_{it}=(y_{it},x_{it}')'$.

EXAMPLE 2—Linear Panel With Quantile Restrictions: Consider the model in (2.3) with the quantile restriction: $P(\varepsilon_{it}\leq 0|x_{it},\beta_i^0,\mu_i^0)=\tau$; see, for example, Kato, Gavao, and Montes-Rojas (2012). We can estimate β_i and μ_i

by minimizing (2.1) with $\psi(w_{it};\beta_i,\mu_i)=\rho_\tau(y_{it}-\beta_i'x_{it}-\mu_i)$, where $\rho_\tau(u)=\{\tau-K(-u/h)\}u$ 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 $y_{it}=\mathbf{1}\{\beta_i^{0\prime}x_{it}+\mu_i^0-\varepsilon_{it}\geq 0\}$, where x_{it}, μ_i, and ε_{it} are as defined in Example 1. In this case, $-\psi(w_{it};\beta_i,\mu_i)=y_{it}\ln F(y_{it}-\beta_i'x_{it}-\mu_i)+(1-y_{it})\ln[1-F(y_{it}-\beta_i'x_{it}-\mu_i)]$, where $w_{it}=(y_{it},x_{it}')'$, and $F(\cdot)$ denotes the conditional CDF (standard logistic or normal) of ε_{it} given x_{it} and the history of (x_{it},y_{it}).

EXAMPLE 4—Tobit Panel: The Tobit panel is characterized by $y_{it}=\max(0,\beta_i^{0\prime}x_{it}+\mu_i^0+\varepsilon_{it})$, 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 (i.i.d.) $N(0,\sigma_\varepsilon^2)$ given x_{it} and the history of (x_{it},y_{it}). In this case, $-\psi(w_{it};\beta_i,\mu_i,\sigma_\varepsilon^2)=\mathbf{1}\{y_{it}=0\}\ln F((y_{it}-\beta_i'x_{it}-\mu_i)/\sigma_\varepsilon^2)+\mathbf{1}\{y_{it}>0\}\ln[f((y_{it}-\beta_i'x_{it}-\mu_i)/\sigma_\varepsilon^2)/\sigma_\varepsilon]$, where $w_{it}=(y_{it},x_{it}')'$, and f and F denote 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

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

where $\hat{\mu}_i(\beta_i)=\arg\min_{\mu_i}\frac{1}{T}\sum_{t=1}^{T}\psi(w_{it};\beta_i,\mu_i)$. 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:

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

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 ith and kth columns of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$, respectively, that is, $\hat{\boldsymbol{\alpha}}\equiv(\hat{\alpha}_1,\ldots,\hat{\alpha}_K)$ and $\hat{\boldsymbol{\beta}}\equiv(\hat{\beta}_1,\ldots,\hat{\beta}_N)$.

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, $\alpha_1^0, \ldots, \alpha_{K_0}^0$. 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 $\{\beta_i^0\}$ and $\{\alpha_k^0\}$ 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 Supplemental Material 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 $\Psi_i(\beta_i, \mu_i) \equiv \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\psi(w_{it}; \beta_i, \mu_i)]$. Note that $\mu_i^0 = \mu_i(\beta_i^0)$. 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 $U_{it} \equiv U_i(w_{it}; \beta_i^0, \mu_i^0)$, 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

$$
\begin{aligned}
&m_{iU} \equiv \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\big(U_{it}^{\mu_i}\big), \quad m_{iV} \equiv \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\big(V_{it}^{\mu_i}\big),\\
&m_{iU2} \equiv \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\big(U_{it}^{\mu_i\mu_i}\big), \quad m_{iV2} \equiv \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\big(V_{it}^{\mu_i\mu_i}\big),\\
&\mathbb{U}_{it} \equiv U_{it} - \frac{m_{iU}}{m_{iV}} V_{it}, \quad \mathbb{U}_{it}^{\beta_i} \equiv U_{it}^{\beta_i} - \frac{m_{iU}}{m_{iV}} V_{it}^{\beta_{i'}}, \quad \text{and}\\
&\mathbb{U}_{it}^{\mu_i} \equiv U_{it}^{\mu_i} - \frac{m_{iU}}{m_{iV}} V_{it}^{\mu_i}.
\end{aligned}
$$

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

unit i:

$$H_{i\mu\mu}(\beta_i) \equiv \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[V_i^{\mu_i}\big(w_{it};\beta_i,\mu_i(\beta_i)\big)\big] \quad \text{and}$$

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

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 $\min_{1\le i\le N}$, and similarly for $\max_i$. We make the following assumptions:

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

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

(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* $v = (v_1, \ldots, v_{p+1})$ *is a vector of nonnegative integers and* $\theta_{(j)}$ *denotes the* j*th element of* θ. *There exists a function* $M(\cdot)$ *such that* $\sup_{\theta\in\Theta}\|D^v\psi(w_{it};\theta)\| \le M(w_{it})$, $\|D^v\psi(w_{it};\theta) - D^v\psi(w_{it};\overline{\theta})\| \le M(w_{it})\|\theta - \overline{\theta}\|$ *for any* $\theta, \overline{\theta} \in \Theta$ *and* $|v| \le 3$, *and* $\max_i \mathbb{E}|M(w_{it})|^q < c_M$ *for some* $c_M < \infty$ *and* $q \ge 6$.

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

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

(vii) K_0 *is fixed and* $N_k/N \to \tau_k \in (0,1)$ *for each* $k = 1, \ldots, K_0$ *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* $k = 1, \ldots, K_0$, $\Omega_k \equiv \lim_{(N_k,T)\to\infty} \frac{1}{N_k}\sum_{i\in G_k^0}\Omega_{iT}$ *exists and* $\Omega_k > 0$.

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

Assumption A1(i) imposes conditions on $\{w_{it}\}$, which are commonly assumed for dynamic nonlinear panel data models; see, for example, Hahn and Kuersteiner (2011) and Lee and Phillips (2015). With more complicated notation, we can relax the stationarity assumption along the time dimension. Assumption A1(ii) imposes an identification condition for the joint identification of (β_i, μ_i) for each i. Assumption A1(iii) restricts the parameter space and it is possible to allow Θ to be i-dependent. Assumption

A1(iv) specifies the smoothness and moment conditions on ψ or objects associated with it. Assumption A1(v), in conjunction with Assumptions 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 $\min_i[\inf_{\beta_i:\|\beta_i-\beta_i^0\|>\eta} \Psi_i(\beta_i,\mu_i(\beta_i)) - \Psi_i(\beta_i^0,\mu_i(\beta_i^0))] > 0$. Assumption A1(vi) specifies that the group-specific parameters are separated from each other, similarly to the separation requirement in Hahn and Moon (2010). Assumption A1(vii) implies that each group has an asymptotically nonnegligible 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

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

Assumption A2(ii) is needed to ensure some higher-order terms are asymptotically negligible. Assumption 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) $\hat\beta_i - \beta_i^0 = O_P(T^{-1/2} + \lambda_1)$ *for* $i = 1, 2, \ldots, N$,
(ii) $\frac{1}{N}\sum_{i=1}^N \|\hat\beta_i - \beta_i^0\|^2 = O_P(T^{-1})$, *and*
(iii) $(\hat\alpha_{(1)}, \ldots, \hat\alpha_{(K_0)}) - (\alpha_1^0, \ldots, \alpha_{K_0}^0) = O_P(T^{-1/2})$, *where* $(\hat\alpha_{(1)}, \ldots, \hat\alpha_{(K_0)})$ *is a suitable permutation of* $(\hat\alpha_1, \ldots, \hat\alpha_{K_0})$.

REMARK 1: Theorem 2.1(i)–(ii) establishes the pointwise and mean square convergence of $\hat\beta_i$. Theorem 2.1(iii) indicates that the group-specific parameters $\alpha_1^0, \ldots, \alpha_K^0$ can be estimated consistently by $\hat\alpha_1, \ldots, \hat\alpha_{K_0}$ 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

$$\hat G_k = \left\{ i \in \{1, 2, \ldots, N\} : \hat\beta_i = \hat\alpha_k \right\} \quad \text{for} \quad k = 1, \ldots, K_0. \tag{2.7}$$

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

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

where $i = 1, \ldots, N$ and $k = 1, \ldots, K_0$. Let $\hat{E}_{kNT} = \bigcup_{i \in G_k^0} \hat{E}_{kNT,i}$ and $\hat{F}_{kNT} = \bigcup_{i \in \hat{G}_k} \hat{F}_{kNT,i}$. $\hat{E}_{kNT}$ 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 $k \in \{1, \ldots, K_0\}$, and $P(\hat{F}_{kNT,i}) \to 0$ as $(N, T) \to \infty$ $\forall i \in \hat{G}_k$ and $k \in \{1, \ldots, K_0\}$. It is *uniformly consistent* if $P(\bigcup_{k=1}^{K_0} \hat{E}_{kNT}) \to 0$ and $P(\bigcup_{k=1}^{K_0} \hat{F}_{kNT}) \to 0$ 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) $P(\bigcup_{k=1}^{K_0} \hat{E}_{kNT}) \leq \sum_{k=1}^{K_0} P(\hat{E}_{kNT}) \to 0$ *as* $(N, T) \to \infty$, *and*
(ii) $P(\bigcup_{k=1}^{K_0} \hat{F}_{kNT}) \leq \sum_{k=1}^{K_0} P(\hat{F}_{kNT}) \to 0$ *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 $\hat{G}_0 \equiv \{1, 2, \ldots, N\} \setminus (\bigcup_{k=1}^{K_0} \hat{G}_k)$ and $\hat{H}_{iNT} \equiv \{i \in \hat{G}_0\}$. Theorem 2.2(i) implies that $P(\bigcup_{1 \leq i \leq N} \hat{H}_{iNT}) \leq \sum_{k=1}^{K_0} P(\hat{E}_{kNT}) \to 0$. 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 $k = 1, \ldots, K_0$, and $i \in \hat{G}_l$ for some $l = 1, \ldots, K_0$ if $\|\hat{\beta}_i - \hat{\alpha}_l\| = \min\{\|\hat{\beta}_i - \hat{\alpha}_1\|, \ldots, \|\hat{\beta}_i - \hat{\alpha}_{K_0}\|\}$ 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* $k = 1, \ldots, K_0$.

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* $\sqrt{N_k T}(\hat{\alpha}_k - \alpha_k^0) - \mathbb{H}_{kNT}^{-1}\mathbb{B}_{kNT} \xrightarrow{D} N(0, \mathbb{H}_k^{-1}\Omega_k(\mathbb{H}_k^{-1})')$, *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* $k = 1, \ldots, K_0$.

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 $\sqrt{N_k T}(\hat{\alpha}_{G_k^0} - \alpha_k^0) - \mathbb{H}_{kNT}^{-1}\mathbb{B}_{kNT} \xrightarrow{D} N(0, \mathbb{H}_k^{-1}\Omega_k(\mathbb{H}_k^{-1})')$ 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 $\hat{\alpha}_{\hat{G}_k} \equiv \arg\min_{\alpha_k} \frac{1}{N_k T} \times \sum_{i \in \hat{G}_k} \sum_{t=1}^{T} \psi(w_{it}; \alpha_k, \hat{\mu}_i(\alpha_k))$. The following theorem reports the asymptotic distribution of $\hat{\alpha}_{\hat{G}_k}$.

THEOREM 2.5: *Suppose Assumptions A1–A3 hold. Then* $\sqrt{N_k T}(\hat{\alpha}_{\hat{G}_k} - \alpha_k^0) - \mathbb{H}_{kNT}^{-1}\mathbb{B}_{kNT} \xrightarrow{D} N(0, \mathbb{H}_k^{-1}\Omega_k(\mathbb{H}_k^{-1})')$ *for* $k = 1, \ldots, K_0$, *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) studied post-Lasso estimators which apply OLS to the model selected by first-step penalized estimators and showed 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 $\hat{\beta}_i(K,\lambda_1) = \hat{\alpha}_k(K,\lambda_1)$, that is, $\hat{G}_k(K,\lambda_1) \equiv \{i \in \{1,2,\ldots,N\} : \hat{\beta}_i(K,\lambda_1) = \hat{\alpha}_k(K,\lambda_1)\}$ for $k = 1,\ldots,K$. Let $\hat{G}(K,\lambda_1) \equiv \{\hat{G}_1(K,\lambda_1),\ldots,\hat{G}_K(K,\lambda_1)\}$. 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

$$\mathrm{IC}_1(K,\lambda_1) \equiv \frac{2}{NT}\sum_{k=1}^{K}\sum_{i\in\hat{G}_k(K,\lambda_1)}\sum_{t=1}^{T}\psi\big(w_{it};\hat{\alpha}_{\hat{G}_k(K,\lambda_1)},\hat{\mu}_i(\hat{\alpha}_{\hat{G}_k(K,\lambda_1)})\big) + \rho_{1NT}\,pK, \tag{2.9}$$

where ρ_{1NT} is a tuning parameter. Let $\hat{K}(\lambda_1) \equiv \arg\min_{1\le K\le K_{\max}} \mathrm{IC}_1(K,\lambda_1)$. See Wang, Li, and Tsai (2007), Liao (2013), and Lu and Su (2016) for the use of a similar IC in various contexts.

Let $G^{(K)} \equiv (G_{K,1},\ldots,G_{K,K})$ be any K-partition of $\{1,2,\ldots,N\}$ and $\mathcal{G}_K$ a collection of all such partitions. Let $\hat{\sigma}^2_{G^{(K)}} \equiv \frac{2}{NT}\sum_{k=1}^{K}\sum_{i\in G_{K,k}}\sum_{t=1}^{T}\psi(w_{it};\hat{\alpha}_{G_{K,k}}, \hat{\mu}_i(\hat{\alpha}_{G_{K,k}}))$, where $\hat{\alpha}_{G_{K,k}} \equiv \arg\min_{\alpha_k}\frac{1}{N_kT}\sum_{i\in G_{K,k}}\sum_{t=1}^{T}\psi(w_{it};\alpha_k,\hat{\mu}_i(\alpha_k))$. 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 underfitted models yield asymptotic mean square errors that are larger than σ_0^2, which is delivered by the true model. Assumption 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, for example, $\Lambda_1\equiv\{\lambda_1=c_jT^{-1/3},\ c_j=c_0\gamma^j \text{ for } j=1,\dots,J\}$ for some $c_0>0$ and $\gamma>1$. That is, we pick up $\lambda_1\in\Lambda_1$ such that $\mathrm{IC}_1(\hat{K}(\lambda_1),\lambda_1)$ 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 $\mathbb{E}(\varepsilon_{it}|x_{it},\mu_i^0)=0$, we have $\psi(w_{it};\beta_i,\mu_i)=\frac{1}{2}(y_{it}-\beta_i'x_{it}-\mu_i)^2$, $\hat{\mu}_i(\beta_i)=\overline{y}_{i\cdot}-\beta_i'\overline{x}_{i\cdot}$, and $Q_{1,NT}(\boldsymbol{\beta})=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\tilde{y}_{it}-\beta_i'\tilde{x}_{it})^2$, where $\overline{y}_{i\cdot}=\frac{1}{T}\sum_{t=1}^{T}y_{it}$, $\tilde{y}_{it}=y_{it}-\overline{y}_{i\cdot}$, and $\overline{x}_{i\cdot}$ 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 $\mu_i(\beta_i)=\mathbb{E}(\overline{y}_{i\cdot})-\beta_i'\mathbb{E}(\overline{x}_{i\cdot})$, $U_i(w_{it};\beta_i,\mu_i)=-(y_{it}-\beta_i'x_{it}-\mu_i)x_{it}$, $V_i(w_{it};\beta_i,\mu_i)=-(y_{it}-\beta_i'x_{it}-\mu_i)$, $U_i^{\mu_i}(w_{it};\beta_i,\mu_i)=x_{it}=V_i^{\beta_i}(w_{it};\beta_i,\mu_i)$, $V_i^{\mu_i}(w_{it};\beta_i,\mu_i)=1$, $U_i^{\beta_i}(w_{it};\beta_i,\mu_i)=x_{it}x_{it}'$, $\mathbb{U}_{it}=-\varepsilon_{it}[x_{it}-\mathbb{E}(\overline{x}_{i\cdot})]$, $\mathbb{U}_{it}^{\beta_i}=[x_{it}-\mathbb{E}(\overline{x}_{i\cdot})]x_{it}'$, $\mathbb{U}_{it}^{\mu_i}=x_{it}-\mathbb{E}(\overline{x}_{i\cdot})$, $H_{i\mu\mu}(\beta_i)=1$, $H_{i\beta\beta}(\beta_i)=\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\{[(x_{it}-\mathbb{E}(\overline{x}_{i\cdot})][x_{it}-\mathbb{E}(\overline{x}_{i\cdot})]'\}$,

$$\Omega_{iT}=\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathbb{E}\big\{\varepsilon_{it}\varepsilon_{is}\big[x_{it}-\mathbb{E}(\overline{x}_{i\cdot})\big]\big[x_{is}-\mathbb{E}(\overline{x}_{i\cdot})\big]'\big\},\quad\text{and}$$

$$\mathbb{H}_{iT}=\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big\{\big[x_{it}-\mathbb{E}(\overline{x}_{i\cdot})\big]\big[x_{it}-\mathbb{E}(\overline{x}_{i\cdot})\big]'\big\}.$$

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

$$\mathbb{B}_{1kNT} = \frac{-1}{\sqrt{N_k T^3}} \sum_{i \in G_k^0} \sum_{t=1}^{T} \sum_{s=1}^{T} \varepsilon_{it} \big[x_{is} - \mathbb{E}(\overline{x}_{i\cdot}) \big] = \mathbb{B}_{1k} + o_P(1) \quad \text{and}$$

$$\mathbb{B}_{2kNT} = 0,$$

where $\mathbb{B}_{1k} = \frac{-1}{\sqrt{N_k T^3}} \sum_{i \in G_k^0} \sum_{t=1}^{T} \sum_{s=1}^{T} \mathbb{E}(\varepsilon_{it} x_{is})$. When x_{it} is strictly exogenous so that $\mathbb{E}(\varepsilon_{it} x_{is}) = 0$ for all t, s, and i, $\mathbb{B}_{1kNT} = o_P(1)$ 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 $\hat{\alpha}_{G_k^0} = (\sum_{i \in G_k^0} \sum_{t=1}^{T} \tilde{x}_{it} \tilde{x}_{it}')^{-1} \times \sum_{i \in G_k^0} \sum_{t=1}^{T} \tilde{x}_{it} \tilde{y}_{it}$. The IC formula in (2.9) now reduces to $\mathrm{IC}_1(K, \lambda_1) = \hat{\sigma}^2_{\hat{G}(K,\lambda_1)} + \rho_{1NT} pK$, 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

$$\mathrm{IC}_1(K, \lambda_1) = \ln \big[\hat{\sigma}^2_{\hat{G}(K,\lambda_1)} \big] + \rho_{1NT} pK, \tag{2.10}$$

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 will not play a role to ensure the selection consistency of IC_1. SSP showed that the requirement on ρ_{1NT} can be relaxed with Assumption A5 replaced by the following:

ASSUMPTION A5*: *As* $(N, T) \to \infty$, $\rho_{1NT} \to 0$ *and* $\rho_{1NT} \delta_{NT}^2 \to \infty$ *where* $\delta_{NT} = N^{1/2} T^{1/2}$ *if* 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) constrained the long-run coefficients to be identical across individuals while assuming the short-run coefficients to be heterogeneous, 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:

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

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 $\gamma=(\gamma_1,\ldots,\gamma_T)'$ 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).[2] We consider the first-differenced system

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

where, for example, $\Delta y_{it}=y_{it}-y_{i,t-1}$ for $t=1,\ldots,T$ and $i=1,\ldots,N$, 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 $\Delta y_i=(\Delta y_{i1},\ldots,\Delta y_{iT})'$, with similar definitions for Δx_i and $\Delta\varepsilon_i$. We propose to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ by minimizing the following

[2]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.

penalized GMM (PGMM) criterion function:[3]

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

where $Q_{2,NT}(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N}[\frac{1}{T}\sum_{t=1}^{T} z_{it}(\Delta y_{it} - \beta_i'\Delta x_{it})]'W_{iNT}[\frac{1}{T}\sum_{t=1}^{T} z_{it}(\Delta y_{it} - \beta_i'\Delta x_{it})]$, 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 $\tilde{\boldsymbol{\alpha}} \equiv (\tilde{\alpha}_1,\ldots,\tilde{\alpha}_{K_0})$ and $\tilde{\boldsymbol{\beta}} \equiv (\tilde{\beta}_1,\ldots,\tilde{\beta}_N)$.

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 $\overline{Q}_{i,z\Delta x} \equiv \mathbb{E}[\tilde{Q}_{i,z\Delta x}]$. 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 $\overline{\rho}_{i,T}(\beta) \equiv \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\{\rho(\xi_{it},\beta) - \mathbb{E}[\rho(\xi_{it},\beta)]\}$. Let $\mathcal{B}_i$ denote the parameter space for β_i. We make the following assumptions.

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

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

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

(iv) *There exist nonrandom matrices* W_i *such that* $P(\max_i\|W_{iNT} - W_i\| \geq \eta) = o(N^{-1})$ *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* $\min_{1\leq k<l\leq K_0}\|\alpha_k^0 - \alpha_l^0\| \geq c_\alpha$.

(vi) K_0 *is fixed and* $N_k/N \to \tau_k \in (0,1)$ *for each* $k = 1,\ldots,K_0$ *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* $c > 0$, $N\max_i P(\|T^{-1}\sum_{t=1}^{T} z_{it}\Delta\varepsilon_{it}\| \geq c\lambda_2) \to 0$ *as* $(N,T)\to\infty$.

ASSUMPTION B3: (i) *For each* $k = 1,\ldots,K_0$, $\overline{A}_k \equiv \frac{1}{N_k}\sum_{i\in G_k^0}\overline{Q}_{i,z\Delta x}'W_i\overline{Q}_{i,z\Delta x} \to A_k > 0$ *as* $(N,T)\to\infty$.

[3]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's (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.

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

Assumption B1(i) specifies moment conditions to identify β_i^0. Assumption 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 $\mathrm{Var}(\omega' \overline{\rho}_{i,T}(\beta_i)) \to \omega' \Sigma_i \omega \in (0, \infty)$ as $T \to \infty$ for some $\Sigma_i > 0$ and any nonrandom $\omega \in \mathbb{R}^d$ with $\|\omega\| = 1$, then $\overline{\rho}_{i,T}(\beta_i) \xrightarrow{D} N(0, \Sigma_i)$ and the first part of Assumption B1(ii) follows. The second and third parts of Assumption B1(ii) can be verified by the Markov inequality and the application of Lemma S1.2(iii) in the Supplemental Material under strong mixing conditions. Assumption B1(iii) provides a rank condition to identify β_i^0. Assumption B1(iv) is automatically satisfied for $W_{iNT} = I_d$, the $d \times d$ identity matrix. Assumptions B1(v)–(vi) and B2(i) parallel Assumptions A1(vi)–(vii) and A2(i). Assumption B2(ii) holds true by Lemma S1.2 in the Supplemental Material 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.

Assumption B3(i)–(ii) can be verified under various primitive conditions. For example, if (a) $\mathbb{E}\|z_{it}(\Delta x_{it})'\|^{2+\sigma} > 0$ 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 i.i.d. along the individual dimension for all $i \in G_k^0$, and (d) $W_i = W \ \forall i \in G_k^0$, then Assumption 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 Assumption B3(ii), for simplicity we assume that $W_{iNT} = I_d$ and make the following decomposition:

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

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 Assumption B3(ii) is satisfied with $W_{iNT} = I_d$ if $V_{kNT} = \frac{1}{N_k^{1/2}T^{1/2}} \sum_{i\in G_k^0} \sum_{t=1}^T \overline{Q}'_{i,z\Delta x} z_{it}\Delta\varepsilon_{it} \xrightarrow{D} N(0, C_k)$ and $R_{kNT} = o_P(1)$, both of which can be verified by strengthening the conditions given in (a)–(c) above. Note that $\overline{A}_k^{-1} B_{kNT}$ signifies the asymptotic bias of $\tilde{\alpha}_k$, which may not vanish asymptotically but can be corrected; see Section S2.2 in the Supplemental Material.[4]

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) $\tilde{\beta}_i - \beta_i^0 = O_P(T^{-1/2} + \lambda_2)$ *for* $i = 1, \ldots, N$,
(ii) $\frac{1}{N}\sum_{i=1}^N \|\tilde{\beta}_i - \beta_i^0\|^2 = O_P(T^{-1})$, *and*
(iii) $(\tilde{\alpha}_{(1)}, \ldots, \tilde{\alpha}_{(K_0)}) - (\alpha_1^0, \ldots, \alpha_{K_0}^0) = O_P(T^{-1/2})$, *where* $(\tilde{\alpha}_{(1)}, \ldots, \tilde{\alpha}_{(K_0)})$ *is a suitable permutation of* $(\tilde{\alpha}_1, \ldots, \tilde{\alpha}_{K_0})$.

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 $\tilde{G}_k \equiv \{i \in \{1, 2, \ldots, N\} : \tilde{\beta}_i = \tilde{\alpha}_k\}$ for $k = 1, \ldots, K_0$.

3.3.2. *Classification Consistency*

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

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

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

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

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

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:

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

where $W_{NT}^{(k)}$ is a $d \times d$ symmetric positive definite matrix. Let $Q_{z\Delta x,NT}^{(k)} = \frac{1}{N_k T}\sum_{i\in G_k^0}\sum_{t=1}^{T} z_{it}(\Delta x_{it})'$ 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 $\tilde{Q}_{z\Delta x}^{(k)} = \frac{1}{N_k T}\sum_{i\in\tilde{G}_k}\sum_{t=1}^{T} z_{it}(\Delta x_{it})'$ 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* $k = 1, \ldots, K_0$, $W_{NT}^{(k)} \xrightarrow{P} W^{(k)} > 0$ *as* $(N, T) \to \infty$.

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

(iii) $\frac{1}{\sqrt{N_k T}} \sum_{i\in G^0_k} \sum_{t=1}^{T} z_{it}\Delta\varepsilon_{it} \xrightarrow{D} N(0, V_k)$.

Assumption B4 is standard in GMM estimation and it can be verified under various primitive conditions that allow for both conditional heteroscedasticity 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* $\sqrt{N_k T}(\tilde{\alpha}_{\tilde{G}_k} - \alpha^0_k) \xrightarrow{D} N(0, \Omega_k)$, *where* $\Omega_k = [Q^{(k)\prime}_{z\Delta x} W^{(k)} Q^{(k)}_{z\Delta x}]^{-1} Q^{(k)\prime}_{z\Delta x} W^{(k)} V_k W^{(k)} Q^{(k)}_{z\Delta x} [Q^{(k)\prime}_{z\Delta x} W^{(k)} Q^{(k)}_{z\Delta x}]^{-1}$ *and* $k = 1, \ldots, K_0$.

REMARK 10: To prove the above theorem, we first apply Theorem 3.2 and show that $\sqrt{N_k T}(\tilde{\alpha}_{\tilde{G}_k} - \alpha^0_k) = \sqrt{N_k T}(\tilde{\alpha}_{G^0_k} - \alpha^0_k) + o_P(1)$. That is, the post-Lasso GMM estimator $\tilde{\alpha}_{\tilde{G}_k}$ is asymptotically equivalent to the oracle estimator $\tilde{\alpha}_{G^0_k}$. 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 $W^{(k)}_{NT}$ to be a consistent estimator of V_k^{-1}. Alternatively, we can consider Arellano and Bond's (1991) GMM estimation based on the estimated groups. The procedure is standard and details are omitted.

REMARK 11: If $W_{iNT} = W^{(k)}_{NT}$, $\overline{Q}_{i,z\Delta x} = Q^{(k)}_{z\Delta x}$ for each $i \in G^0_k$ in Assumption B3(i) (which is unrealistic before knowing the group identity), and $B_{kNT} = 0$ in Assumption B3(ii), then $A_k = Q^{(k)\prime}_{z\Delta x} W^{(k)} Q^{(k)}_{z\Delta x}$, $C_k = Q^{(k)\prime}_{z\Delta x} W^{(k)} \Omega_k W^{(k)} Q^{(k)}_{z\Delta x}$, and $\sqrt{N_k T}(\tilde{\alpha}_k - \alpha^0_k) \xrightarrow{D} N(0, \Omega_k)$. 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 $\tilde{G}(K, \lambda_2) \equiv \{\tilde{G}_1(K, \lambda_2), \ldots, \tilde{G}_K(K, \lambda_2)\}$. The post-Lasso GMM estimate of α^0_k is given by

$$\tilde{\alpha}_{\tilde{G}_k(K,\lambda_2)} \equiv \big[\tilde{Q}^{(K,k)\prime}_{z\Delta x} W^{(k)}_{NT} \tilde{Q}^{(K,k)}_{z\Delta x}\big]^{+} \tilde{Q}^{(K,k)\prime}_{z\Delta x} W^{(k)}_{NT} \tilde{Q}^{(K,k)}_{z\Delta y},$$

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 $k = 1, 2, \ldots, K$. Let

$$\tilde{\sigma}^2_{\tilde{G}(K,\lambda_2)} = \frac{1}{NT} \sum_{k=1}^{K} \sum_{i \in \tilde{G}_k(K,\lambda_2)} \sum_{t=1}^{T} \left[\Delta y_{it} - \tilde{\alpha}'_{\tilde{G}_k(K,\lambda_1)} \Delta x_{it}\right]^2.$$

We propose to select K to minimize the following IC:

$$\text{IC}_2(K, \lambda_2) = \ln\left[\tilde{\sigma}^2_{\tilde{G}(K,\lambda_2)}\right] + \rho_{2NT}\, pK, \tag{3.5}$$

where ρ_{2NT} is a tuning parameter. Let $\tilde{K}(\lambda_2) \equiv \arg\min_{1 \le K \le K_{\max}} \text{IC}_2(K, \lambda_2)$. As before, for any $G^{(K)} = (G_{K,1}, \ldots, G_{K,K}) \in \mathcal{G}_K$, define

$$\tilde{\sigma}^2_{G^{(K)}} = \frac{1}{NT} \sum_{k=1}^{K} \sum_{i \in G_{K,k}} \sum_{t=1}^{T} \left[\Delta y_{it} - \tilde{\alpha}'_{G_{K,k}} \Delta x_{it}\right]^2,$$

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* $\sigma^2_{\Delta\varepsilon} = \text{plim}_{(N,T)\to\infty} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (\Delta\varepsilon_{it})^2$.

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 nonlinear 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 $e_{it1}, e_{it2} \sim$ i.i.d. $N(0, 1)$, 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 $e_{it} \sim$ i.i.d. $N(0, 1)$ 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 $\mathrm{IC}_1(\hat{K}(\lambda_1), \lambda_1)$ and similarly the λ_2 that minimizes $\mathrm{IC}_2(\tilde{K}(\lambda_2), \lambda_2)$. We run 500 replications for each DGP. Table I 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.

Next, given the true number of groups, we focus on the classification of individual units and the point estimation of post-Lasso.[5] 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 Table II 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 DGPs 1 and 3, the coverage probability is computed as $\sum_{k=1}^{K_0}\frac{N_k}{N}\mathbf{1}\{\hat{\alpha}_{\hat{G}_k,1} - 1.96\hat{\sigma}_{k1} \le \alpha_{k1}^0 \le \hat{\alpha}_{\hat{G}_k,1} + 1.96\hat{\sigma}_{k1}\}$, 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 purposes, 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

[5]Here we report the results for post-Lasso under K_0. In Table SIII of the Supplemental 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 I

FREQUENCY OF SELECTING $K = 1, \ldots, 5$ GROUPS WHEN $K_0 = 3$

		DGP 1					DGP 2					DGP 3				
N	T	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

TABLE II

CLASSIFICATION AND POINT ESTIMATION OF $\boldsymbol{\alpha}_1$

				Post-Lasso			Oracle		
	N	T	% of Correct Classification	RMSE	Bias	Coverage	RMSE	Bias	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

undercovers 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 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 Supplemental Material.

5.1. *Savings Rate Dynamic Panel Modeling and Classification*

Understanding the disparate savings behavior across countries is a long-standing 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 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 data set 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 the 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.

Columns 3–4 in Table III 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 insignif-

TABLE III
PLS AND PGMM ESTIMATION RESULTS[a]

	PLS			PGMM		
Variables	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)

[a]Note: *** 1% significant, ** 5% significant, * 10% significant.

icance of inflation and interest rates in the determination of savings rate. This result also lends support to the *conventional wisdom* that, across countries, higher savings 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 III 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.[6] 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 data set 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.[7] 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 nonstationarity.

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 IV 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

[6]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]."

[7]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 ongoing 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.

TABLE IV

PROBIT, FE PROBIT, AND PPL ESTIMATION RESULTS[a]

					Post-Lasso PPL			
	Probit		FE Probit		High-Occurrence		Low-Occurrence	
Variables	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

[a]Note: *** 1% significant, ** 5% significant, * 10% significant.

two groups: the negative coefficient is statistically significant in the low-occurrence group, but no such relationship is found in the high-occurrence group.

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 increase 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 $Q^{(K_0)}_{1iNT,\lambda_1}(\beta_i, \boldsymbol{\alpha}) = Q_{1NT,i}(\beta_i) + \lambda_1 \prod_{k=1}^{K_0} \|\beta_i - \alpha_k\|$. 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 the-

orem, we have

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

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

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

By Lemmas S1.6 and S1.10 in the Supplemental Material, $\hat{S}_i = O_P(T^{-1/2})$, $\frac{1}{N}\sum_{i=1}^{N}\|\hat{S}_i\|^2 = O_P(T^{-1})$, 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,

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

where $\hat{c}_{iNT}(\boldsymbol{\alpha}) = \prod_{k=1}^{K_0-1}\|\hat{\beta}_i - \alpha_k\| + \prod_{k=1}^{K_0-2}\|\hat{\beta}_i - \alpha_k\|\|\beta_i^0 - \alpha_{K_0}\| + \cdots + \prod_{k=2}^{K_0}\|\beta_i^0 - \alpha_k\| = O_P(1)$. By (A.1), (A.3), and the fact that $Q_{1iNT,\lambda_1}^{(K_0)}(\hat{\beta}_i, \hat{\boldsymbol{\alpha}}) - Q_{1iNT,\lambda_1}^{(K_0)}(\beta_i^0, \hat{\boldsymbol{\alpha}}) \le 0$, 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 As-

sumption A1(v),

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

(ii) Let $\boldsymbol{\beta} = \boldsymbol{\beta}^0 + T^{-1/2}\mathbf{v}$, where $\mathbf{v} = (v_1, \ldots, v_N)$ 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

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

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

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

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, $T[Q_{1NT,\lambda_1}^{(K_0)}(\boldsymbol{\beta}^0 + T^{-1/2}\mathbf{v}, \hat{\boldsymbol{\alpha}}) - Q_{1NT,\lambda_1}^{(K_0)}(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0)] > 0$ for sufficiently large L. Consequently, we must have $N^{-1}\sum_{i=1}^N \|\hat{b}_i\|^2 = O_P(T^{-1})$.

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

$$\begin{aligned}
\text{(A.6)} \qquad \hat{c}_{iNT}(\boldsymbol{\alpha}) \le{}& \prod_{k=1}^{K_0-1}\big\{\big\|\hat{\beta}_i - \beta_i^0\big\| + \big\|\beta_i^0 - \alpha_k\big\|\big\} \\
&+ \prod_{k=1}^{K_0-2}\big\{\big\|\hat{\beta}_i - \beta_i^0\big\| + \big\|\beta_i^0 - \alpha_k\big\|\big\}\big\|\beta_i^0 - \alpha_{K_0}\big\| \\
&+ \cdots + \prod_{k=2}^{K_0}\big\|\beta_i^0 - \alpha_k\big\|
\end{aligned}$$

$$= \sum_{s=0}^{K_0-1} \|\hat{\beta}_i - \beta_i^0\|^s \prod_{k=1}^{s} a_{ks} \|\beta_i^0 - \alpha_k\|^{K_0-1-s}$$

$$\leq C_{K_0}(\boldsymbol{\alpha}) \sum_{s=0}^{K_0-1} \|\hat{\beta}_i - \beta_i^0\|^s \leq C_{K_0}(\boldsymbol{\alpha})\big(1 + 2\|\hat{\beta}_i - \beta_i^0\|\big),$$

where the a_{ks} are finite integers and

$$C_{K_0}(\boldsymbol{\alpha}) \equiv \max_{1\leq l\leq K_0} \max_{1\leq s\leq K_0-1} \prod_{k=1}^{s} a_{ks} \|\alpha_l^0 - \alpha_k\|^{K_0-1-s} = O(1)$$

as K_0 is finite. By (A.3) and (A.6), as $(N, T) \to \infty$,

$$\text{(A.7)} \quad \big|P_{NT}(\hat{\boldsymbol{\beta}}, \boldsymbol{\alpha}) - P_{NT}\big(\boldsymbol{\beta}^0, \boldsymbol{\alpha}\big)\big|$$

$$\leq C_{K_0}(\boldsymbol{\alpha}) \frac{1}{N} \sum_{i=1}^{N} \|\hat{b}_i\| + 2C_{K_0}(\boldsymbol{\alpha}) \frac{1}{N} \sum_{i=1}^{N} \|\hat{b}_i\|^2$$

$$\leq C_{K_0}(\boldsymbol{\alpha}) \left\{ \frac{1}{N} \sum_{i=1}^{N} \|\hat{b}_i\|^2 \right\}^{1/2} + O_P\big(T^{-1}\big) = O_P\big(T^{-1/2}\big).$$

By (A.7), and the fact that $P_{NT}(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0) = 0$ and that $P_{NT}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}) - P_{NT}(\hat{\boldsymbol{\beta}}, \boldsymbol{\alpha}^0) \leq 0$, we have

$$\text{(A.8)} \quad 0 \geq P_{NT}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}) - P_{NT}\big(\hat{\boldsymbol{\beta}}, \boldsymbol{\alpha}^0\big)$$

$$= P_{NT}\big(\boldsymbol{\beta}^0, \hat{\boldsymbol{\alpha}}\big) - P_{NT}\big(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\big) + O_P\big(T^{-1/2}\big)$$

$$= \frac{1}{N} \sum_{i=1}^{N} \prod_{k=1}^{K_0} \|\beta_i^0 - \hat{\alpha}_k\| + O_P\big(T^{-1/2}\big)$$

$$= \frac{N_1}{N} \prod_{k=1}^{K_0} \|\hat{\alpha}_k - \alpha_1^0\| + \cdots + \frac{N_{K_0}}{N} \prod_{k=1}^{K_0} \|\hat{\alpha}_k - \alpha_{K_0}^0\| + O_P\big(T^{-1/2}\big).$$

By Assumption A1(vii), $N_k/N \to \tau_k \in (0, 1)$ for each $k = 1, \ldots, K_0$. So (A.8) implies that $\prod_{k=1}^{K_0} \|\hat{\alpha}_k - \alpha_l^0\| = O_P(T^{-1/2})$ for $l = 1, \ldots, K_0$. It follows that $(\hat{\alpha}_{(1)}, \ldots, \hat{\alpha}_{(K_0)}) - (\alpha_1^0, \ldots, \alpha_{K_0}^0) = O_P(T^{-1/2})$. *Q.E.D.*

PROOF OF THEOREM 2.2: (i) First, we fix $k \in \{1, \ldots, K_0\}$. By the consistency of $\hat{\alpha}_k$ and $\hat{\beta}_i$ in Theorem 2.1 and Assumption A1(vi)–(vii), $\hat{\beta}_i - \hat{\alpha}_l \xrightarrow{P} \alpha_k^0 - \alpha_l^0 \neq$

0 for all $i \in G_k^0$ and $l \neq k$ and $\hat{c}_{ki} \equiv \prod_{l=1,l\neq k}^{K_0} \|\hat{\beta}_i - \hat{\alpha}_l\| \xrightarrow{P} c_k^0 \equiv \prod_{l=1,l\neq k}^{K_0} \|\alpha_k^0 - \alpha_l^0\| \geq c_\alpha^{K_0-1} > 0$ 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

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

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

Let $\varkappa_{1NT} = (T^{-1/2}(\ln T)^3 + \lambda_1)(\ln T)^\nu$. Let C denote a generic constant that may vary across lines. By (A.4) and Lemmas S1.6–S1.10 in the Supplemental Material, we can readily show that

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

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

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

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

$1-o(N^{-1})$, where

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

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

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

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

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

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 $1-o(N^{-1})$, 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, $P(\|\hat{\beta}_i-\hat{\alpha}_k\|=0 | i \in G_k^0) = 1-o(N^{-1})$.

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

$$\begin{aligned}&\text{(A.12)} \quad \sum_{k=1}^{K_0}\sum_{i\in G_k^0}P(\hat{E}_{kNT,i}) \\ &\qquad \le \sum_{k=1}^{K_0}\sum_{i\in G_k^0}\big[P\big(\|\hat{B}_{i1}\| \ge \sqrt{T}\lambda_1 c_k^0/4, \Xi_{kNT}\big) + P\big(\Xi_{kNT}^c\big)\big]\end{aligned}$$

$$\leq N \max_{1\leq i\leq N} P\left(\left\|\frac{1}{T}\sum_{t=1}^{T} U_{it}\right\| \geq \lambda_1 c_\alpha^{K_0-1}/4\right) + o(1) = o(1).$$

This completes the proof of (i).

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

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

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

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

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

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

Q.E.D.

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 $\mathbf{1}\{i \in \hat{G}_k\} - \mathbf{1}\{i \in G_k^0\} = \mathbf{1}\{i \in \hat{G}_k \setminus G_k^0\} - \mathbf{1}\{i \in G_k^0 \setminus \hat{G}_k\}$, we have $\hat{N}_k - N_k = \sum_{i=1}^N [\mathbf{1}\{i \in \hat{G}_k \setminus G_k^0\} - \mathbf{1}\{i \in G_k^0 \setminus \hat{G}_k\}]$. Then by the implication rule and the Markov inequality, for any $\epsilon > 0$,

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

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. *Q.E.D.*

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 $i = 1, \ldots, N$ (resp. $k = 1, \ldots, K_0$), $\mathbf{0}_{p\times 1}$ 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 $i = 1, \ldots, N$ and $k = 1, \ldots, K_0$, we have by the envelope theorem,

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

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

where $\hat{e}_{ij}$ is defined after (A.9). Fix $k \in \{1, \ldots, K_0\}$. Observe that (a) $\|\hat{\beta}_i - \hat{\alpha}_k\| = 0$ 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),

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

and

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

where the last equality follows from the fact that

$$\begin{aligned}&\sum_{j=1,j\neq k}^{K_0}\sum_{i\in\hat{G}_j}\hat{e}_{ik}\prod_{l=1,l\neq k}^{K_0}\|\hat{\alpha}_j-\hat{\alpha}_l\| \\ &= \sum_{j=1,j\neq k}^{K_0}\sum_{i\in\hat{G}_j}\frac{\hat{\alpha}_j-\hat{\alpha}_k}{\|\hat{\alpha}_j-\hat{\alpha}_k\|}\prod_{l=1,l\neq k}^{K_0}\|\hat{\alpha}_j-\hat{\alpha}_l\| = 0.\end{aligned}$$

Then, averaging both sides of (A.14) over $i\in\hat{G}_k$ and using (A.16)–(A.17), we have

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

Using $U_i(w_{it};\hat{\alpha}_k,\hat{\mu}_i(\hat{\alpha}_k)) = U_i(w_{it};\alpha_k^0,\hat{\mu}_i(\alpha_k^0)) + \hat{H}_{(k)}(\hat{\alpha}_k-\alpha_k^0)$ with $\hat{H}_{(k)}\equiv \frac{1}{N_kT}\sum_{i\in\hat{G}_k}\sum_{t=1}^{T}[U_i^{\beta_i}(w_{it};\breve{\alpha}_k,\hat{\mu}_i(\breve{\alpha}_k)) + U_i^{\mu_i}(w_{it};\breve{\alpha}_k^0,\hat{\mu}_i(\breve{\alpha}_k))\frac{\partial\hat{\mu}_i(\breve{\alpha}_k)}{\partial\alpha_k'}]$ and $\breve{\alpha}_k$ lying

between $\hat{\alpha}_k$ and α_k^0 elementwise, and rearranging terms in (A.18) yield

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

where $\hat{\mathcal{R}}_k = \hat{H}_{(k)}^{-1} \frac{\lambda_1}{N} \sum_{i \in \hat{G}_0} \hat{e}_{ik} \prod_{l=1, l \neq k}^{K_0} \|\hat{\beta}_i - \hat{\alpha}_l\|$. In view of the fact that $\hat{e}_{ik} \prod_{l=1, l \neq k}^{K_0} \|\hat{\beta}_i - \hat{\alpha}_l\| \neq 0$ only if $i \in \hat{G}_0$, we have, for any $\epsilon > 0$,

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

So $\|\hat{\mathcal{R}}_k\| = o_P((NT)^{-1/2})$. By Lemmas S1.11–S1.12, $\frac{1}{\sqrt{N_k T}} \sum_{i \in \hat{G}_k} \sum_{t=1}^{T} U_i(w_{it}; \alpha_k^0, \hat{\mu}_i(\alpha_k^0)) + \mathbb{B}_{kNT} \xrightarrow{D} N(0, \Omega_k)$ and $\hat{H}_{(k)} = \mathbb{H}_{kNT} + o_P(\nu_{NT})$, 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 $\sqrt{N_k T}(\hat{\alpha}_k - \alpha_k^0) - \mathbb{H}_{kNT}^{-1} \mathbb{B}_{kNT} \xrightarrow{D} N(0, \mathbb{H}_k^{-1} \Omega_k (\mathbb{H}_k^{-1})')$. *Q.E.D.*

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

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

where $\hat{H}_{\hat{G}_k} \equiv \frac{1}{N_k T} \sum_{i \in \hat{G}_k} \sum_{t=1}^{T} [U_i^{\alpha_k}(w_{it}; \breve{\alpha}_{\hat{G}_k}, \hat{\mu}_i(\breve{\alpha}_{\hat{G}_k})) + U_i^{\mu_i}(w_{it}; \breve{\alpha}_k^0, \hat{\mu}_i(\breve{\alpha}_{\hat{G}_k})) \times \partial \hat{\mu}_i(\breve{\alpha}_{\hat{G}_k})/\partial \alpha_k']$ and $\breve{\alpha}_{\hat{G}_k}$ 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 $\sqrt{N_k T}(\hat{\alpha}_{\hat{G}_k} - \alpha_k^0) - \mathbb{H}_k^{-1} \mathbb{B}_{kNT} \xrightarrow{D} N(0, \mathbb{H}_k^{-1} \Omega_k (\mathbb{H}_k^{-1})')$. *Q.E.D.*

PROOF OF THEOREM 2.6: Let $\mathcal{K} = \{1, 2, \ldots, K_{\max}\}$. 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 $\hat{\sigma}^2_{\hat{G}(K,\lambda_1)} = \frac{2}{NT}\sum_{k=1}^{K}\sum_{i\in\hat{G}_k(K,\lambda_1)}\sum_{t=1}^{T}\psi(w_{it};\hat{\alpha}_{\hat{G}_k(K,\lambda_1)},\hat{\mu}_i(\hat{\alpha}_{\hat{G}_k(K,\lambda_1)}))$. Using Theorems 2.2 and 2.5 and Assumption A5, we can readily show that $\mathrm{IC}_1(K_0,\lambda_1) = \hat{\sigma}^2_{\hat{G}(K_0,\lambda_1)} + \rho_{1NT}pK_0 = \frac{2}{NT}\sum_{k=1}^{K_0}\sum_{i\in\hat{G}_k(K_0,\lambda_1)}\sum_{t=1}^{T}\psi(w_{it};\alpha^0_k,\mu^0_i) + o_P(1) \xrightarrow{P} \ln(\sigma_0^2)$. We consider the cases of under- and over-fitted models separately.

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

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

we have by Assumptions A4–A5 that

$$\min_{1\leq K<K_0}\mathrm{IC}_1(K,\lambda_1) \geq \min_{1\leq K<K_0}\inf_{G^{(K)}\in\mathcal{G}_K}\hat{\sigma}^2_{G^{(K)}} + \rho_{1NT}pK \xrightarrow{P} \ln(\underline{\sigma}^2) > \ln(\sigma_0^2).$$

It follows that $P(\min_{K\in\Omega_-}\mathrm{IC}_1(K,\lambda_1) > \mathrm{IC}_1(K_0,\lambda_1)) \to 1$.

Case 2: Over-fitted model. Let $K \in \Omega_+$. By Lemma S1.14 in the Supplemental Material and the fact that $T\rho_{1NT} \to \infty$ under Assumption A5, we have

$$\begin{aligned}&P\Big(\min_{K\in\Omega_+}\mathrm{IC}_1(K,\lambda_1) > \mathrm{IC}_1(K_0,\lambda_1)\Big)\\ &= P\Big(\min_{K\in\Omega_+}\big[T\big(\hat{\sigma}^2_{\hat{G}(K,\lambda_1)} - \hat{\sigma}^2_{\hat{G}(K_0,\lambda_1)}\big) + T\rho_{1NT}(K-K_0)\big] > 0\Big)\\ &\to 1 \quad \text{as} \quad (N,T)\to\infty.\end{aligned}$$

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

APPENDIX 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 $\overline{V}_i(\beta_i) \equiv \{\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\rho(\xi_{it},\beta_i)]\}'W_i\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\rho(\xi_{it},\beta_i)]$. 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) *holds. Then* $P(\underline{c}[\frac{1}{2}\overline{V}_i(\beta_i) - R_{i,T}(\beta_i)] \leq V_{iNT}(\beta_i) \leq \overline{c}[2\overline{V}_i(\beta_i) + 2R_{i,T}(\beta_i)]) = 1 - o(N^{-1})$ *for all* $\beta_i \in \mathcal{B}_i$, *where* $\underline{c}$ *and* $\overline{c}$ *are some generic positive constants that do not depend on* i *with* $0 < \underline{c} < 1 < \overline{c} < \infty$.

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

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

for all $\beta_i \in \mathcal{B}_i$. By positive definiteness of W_i and simple manipulations, we can readily show that

$$\begin{aligned}&(a-b)'W_i(a-b) \geq \frac{1}{2}a'W_ia - b'W_ib \quad \text{and} \\ &(a-b)'W_i(a-b) \leq 2a'W_ia + 2b'W_ib\end{aligned}$$

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

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

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

Combining (B.1)–(B.3) yields the desired results. *Q.E.D.*

PROOF OF THEOREM 3.1: (i) Let $Q^{(K_0)}_{2iNT,\lambda_2}(\beta_i,\boldsymbol{\alpha}) = V_{iNT}(\beta_i) + \lambda_2\prod_{k=1}^{K_0}\|\beta_i - \alpha_k\|$. Then $Q^{(K_0)}_{2NT,\lambda_2}(\boldsymbol{\beta},\boldsymbol{\alpha}) = \frac{1}{N}\sum_{i=1}^{N}Q^{(K_0)}_{2iNT,\lambda_2}(\beta_i,\boldsymbol{\alpha})$. By the definition of $(\tilde{\boldsymbol{\beta}},\tilde{\boldsymbol{\alpha}})$, we have

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

By Lemma B.1 and Assumption B1(i) and (iv), we have that $V_{iNT}(\tilde{\beta}_i) \geq \underline{c}[\frac{1}{2}\overline{V}_i(\tilde{\beta}_i) - \tilde{R}_{i,T}]$ and $V_{iNT}(\beta_i^0) \leq \overline{c}[2\overline{V}_i(\beta_i^0) + 2R^0_{i,T}] = 2\overline{c}R^0_{i,T}$ on the set Λ_{NT}, where $\tilde{R}_{i,T} = R_{i,T}(\tilde{\beta}_i)$ and $R^0_{i,T} = R_{i,T}(\beta_i^0)$. It follows that $\underline{c}[\frac{1}{2}\overline{V}_i(\tilde{\beta}_i) - \tilde{R}_{i,T}] -$

$2\overline{c}R_{i,T}^0 + \lambda_2\{\prod_{k=1}^K \|\tilde{\beta}_i - \tilde{\alpha}_k\| - \prod_{k=1}^K \|\beta_i^0 - \tilde{\alpha}_k\|\} \le 0$, which can be rewritten as

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

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

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

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

$$\begin{aligned}\max_{1\le i\le N} \overline{V}_i(\tilde{\beta}_i) &= \max_{1\le i\le N}\left(\tilde{\beta}_i - \beta_i^0\right)'\overline{Q}_{i,z\Delta x}'W_i\overline{Q}_{i,z\Delta x}\left(\tilde{\beta}_i - \beta_i^0\right) \\ &\ge \underline{c}_{1NT}\max_{1\le i\le N}\|\tilde{\beta}_i - \beta_i^0\|^2,\end{aligned} \tag{B.6}$$

where $\underline{c}_{1NT} \equiv \min_{1\le i\le N}\mu_{\min}(\overline{Q}_{i,z\Delta x}'W_i\overline{Q}_{i,z\Delta x})$ satisfies that $\liminf_{(N,T)\to\infty}\underline{c}_{1NT} \ge \underline{c}_W\underline{c}_Q^2 > 0$ by Assumption B1(iii)–(iv). Combining Assumptions (B.4)–(B.6) yields

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

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

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

As in the proof of Theorem 2.1(ii), we can further demonstrate that $\frac{1}{N}\sum_{i=1}^N \|\tilde{\beta}_i - \beta_i^0\|^2 = O_P(T^{-1})$.

The proof of (iii) is completely analogous to that of Theorem 2.1(iii), now using the facts that $|P_{NT}(\tilde{\boldsymbol{\beta}}, \boldsymbol{\alpha}) - P_{NT}(\boldsymbol{\beta}^0, \boldsymbol{\alpha})| = O_P(T^{-1/2})$ and that $0 \ge P_{NT}(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}) - P_{NT}(\tilde{\boldsymbol{\beta}}, \boldsymbol{\alpha}^0)$. *Q.E.D.*

PROOF OF THEOREM 3.2: (i) First, we fix $k \in \{1, \ldots, K_0\}$. By the consistency of $\tilde{\alpha}_k$ and $\tilde{\beta}_i$ in Theorem 3.1 and Assumption 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 $\tilde{c}_{ki} \equiv \prod_{l=1,l\neq k}^{K_0} \|\tilde{\beta}_i - \tilde{\alpha}_l\| \xrightarrow{P} c_k^0 \equiv \prod_{l=1,l\neq k}^{K_0} \|\alpha_k^0 - \alpha_l^0\| \geq c_\alpha^{K_0-1} > 0$ for any $i \in G_k^0$. Now, suppose that $\|\tilde{\beta}_i - \tilde{\alpha}_k\| \neq 0$ 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

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

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 $P(\max_i \|\tilde{\beta}_i - \beta_i^0\| \geq \eta) = o(N^{-1})$ for any given $\eta > 0$. With this, by (B.7) and Assumption B2(ii)–(iv), we can readily show that

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

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

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

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

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

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

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

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

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

It follows that $P(\|\tilde{\beta}_i - \tilde{\alpha}_k\| = 0 | i \in G_k^0) \to 1$ as $(N,T) \to \infty$. Now, observe that $P(\bigcup_{k=1}^{K_0}\hat{E}_{kNT}) \leq \sum_{k=1}^{K_0}P(\hat{E}_{kNT}) \leq \sum_{k=1}^{K_0}\sum_{i\in G_k^0}P(\hat{E}_{kNT,i})$ and by Assumption B2(ii),

$$\begin{aligned}&\sum_{k=1}^{K_0}\sum_{i\in G_k^0}P(\tilde{E}_{kNT,i}) \\ &\quad\leq \sum_{k=1}^{K_0}\sum_{i\in G_k^0}\left[P\left(\|\tilde{B}_{i1}\| \geq \sqrt{T}\lambda_2 c_k^0/4, \Gamma_{kNT}\right) + P\left(\Gamma_{kNT}^c\right)\right]\end{aligned}$$

$$\leq N \max_{1\leq i\leq N} P\left(\left\|\tilde{Q}'_{i,z\Delta x} W_{iNT} \frac{1}{T}\sum_{t=1}^{T} z_{it}\Delta\varepsilon_{it}\right\| \geq \lambda_2 c_k^0/4, \Gamma_{kNT}\right) + o(1)$$

$$\leq N \max_{1\leq i\leq N} P\left(\left\|\frac{1}{T}\sum_{t=1}^{T} z_{it}\Delta\varepsilon_{it}\right\| \geq \lambda_2 c_\alpha^{K_0-1}/(16\underline{c}_{\overline{Q}}\underline{c}_W)\right) + o(1) = o(1),$$

where we use the fact that $\|\tilde{Q}_{i,z\Delta x}\|\|W_{iNT}\| \geq (\|\overline{Q}_{i,z\Delta x}\| - \|\tilde{Q}_{i,z\Delta x} - \overline{Q}_{i,z\Delta x}\|) \times (\|W_i\| - \|W_{iNT} - W_i\|) \geq \underline{c}_{\overline{Q}}\underline{c}_W/4$ 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. *Q.E.D.*

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 $i = 1, \ldots, N$ and $k = 1, \ldots, K_0$,

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

$$\mathbf{0}_{p\times 1} = \frac{\lambda_1}{N}\sum_{i=1}^{N} \tilde{e}_{ik} \prod_{l=1,l\neq k}^{K_0} \|\tilde{\beta}_i - \tilde{\alpha}_l\|,$$

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

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

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

PROOF OF THEOREM 3.5: By Theorem 3.2, we can readily show that

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

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

PROOF OF THEOREM 3.6: The proof is analogous to that of Theorem 2.6 and is omitted. *Q.E.D.*

REFERENCES

ANDO, T., AND J. BAI (2015): "Asset Pricing With a General Multifactor Structure," *Journal of Financial Econometrics*, 13, 556–604. [2216]

——— (2016): "Panel Data Models With Grouped Factor Structure Under Unknown Group Membership," *Journal of Applied Econometrics*, 31, 163–191. [2217]

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. [2231,2235]

BALTAGI, B. H., G. BRESSON, AND A. PIROTTE (2008): "To Pool or Not to Pool?" in *The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory and Practice* (Third Ed.), ed. by L. Mátyás and P. Sevestre. Berlin: Springer-Verlag, 517–546. [2216]

BELLONI, A., AND V. CHERNOZHUKOV (2013): "Least Squares After Model Selection in High-Dimensional Sparse Models," *Bernoulli*, 19, 521–547. [2227]

BERTSEKAS, D. (1995): *Nonlinear Programming*. Belmont, MA: Athena Scientific. [2252]

BESLEY, T., AND T. PERSSON (2010): "State Capacity, Conflict, and Development," *Econometrica*, 78, 1–34. [2242]

BESTER, C. A., AND C. B. HANSEN (2016): "Grouped Effects Estimators in Fixed Effects Models," *Journal of Econometrics*, 190, 197–208. [2216]

BLATTMAN, C., AND E. MIGUEL (2010): "Civil Wars," *Journal of Economic Literature*, 48, 3–57. [2242,2243]

BONHOMME, S., AND E. MANRESA (2015): "Grouped Patterns of Heterogeneity in Panel Data," *Econometrica*, 83, 1147–1184. [2217]

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. [2240]

BROWNING, M., AND J. M. CARRO (2007): "Heterogeneity and Microeconometrics Modelling," in *Advances in Economics and Econometrics, Theory and Applications: Ninth World Congress of the Econometric Society*, Vol. 3, ed. by R. Blundell, W. K. Newey, and T. Persson. New York: Cambridge University Press, 45–74. [2216]

——— (2010): "Heterogeneity in Dynamic Discrete Choice Models," *Econometrics Journal*, 13, 1–39. [2216]

——— (2014): "Dynamic Binary Outcome Models With Maximal Heterogeneity," *Journal of Econometrics*, 178, 805–823. [2216,2217]

CARROLL, C., AND D. N. WEIL (1994): "Saving and Growth: A Reinterpretation," *Carnegie-Rochester Conference Series on Public Policy*, 40, 133–192. [2241]

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. [2218]

COLLIER, P., AND A. HOEFFLER (2004): "Greed and Grievance in Civil Wars," *Oxford Economic Papers*, 56, 563–595. [2243]

DEATON, A. (1990): "Saving in Developing Countries: Theory and Review," in *Proceedings of the World Bank Annual Conference on Development Economics*. Washington, D.C.: The World Bank, 61–96. [2240]

DHAENE, G., AND K. JOCHMANS (2015): "Split-Panel Jackknife Estimation of Fixed-Effect Models," *Review of Economic Studies*, 82, 991–1030. [2240,2241]

DJANKOV, S., AND M. REYNAL-QUEROL (2010): "Poverty and Civil War: Revisiting the Evidence," *Review of Economics and Statistics*, 92, 1035–1041. [2243]

DUALAUF, S. N., A. KOURTELLOS, AND A. MINKIN (2001): "The Local Solow Growth Model," *European Economic Review*, 45, 928–940. [2216]

EDWARDS, S. (1996): "Why Are Latin America's Savings Rates So Low? An International Comparative Analysis," *Journal of Development Economics*, 51, 5–44. [2240,2241]

ESTEBAN, J., L. MAYORAL, AND D. RAY (2012): "Ethnicity and Conflict: An Empirical Study," *American Economic Review*, 102, 1310–1342. [2243]

FEARON, J. D., AND D. D. LAITIN (2003): "Ethnicity, Insurgency, and Civil War," *The American Political Science Review*, 97, 75–90. [2242,2243]

FELDSTEIN, M. (1980): "International Differences in Social Security and Saving," *Journal of Public Economics*, 14, 225–244. [2240]

FERNÁNDEZ-VAL, I., AND L. LEE (2013): "Panel Data Models With Nonadditive Unobservable Heterogeneity: Estimation and Inference," *Quantitative Economics*, 4, 453–481. [2230,2231]

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. [2229]

——— (2011): "Bias Reduction for Dynamic Nonlinear Panel Models With Fixed Effects," *Econometric Theory*, 27, 1152–1191. [2221,2223]

HAHN, J., AND H. R. MOON (2010): "Panel Data Models With Finite Number of Multiple Equilibria," *Econometric Theory*, 26, 863–881. [2217,2220,2224]

HAHN, J., AND W. NEWEY (2004): "Jackknife and Analytical Bias Reduction for Nonlinear Panel Models," *Econometrica*, 72, 1295–1319. [2221,2226]

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. [2218]

HSIAO, C. (2014): *Analysis of Panel Data* (Third Ed.). New York: Cambridge University Press. [2215]

HSIAO, C., AND H. PESARAN (2008): "Random Coefficient Panel Data Models," in *The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory and Practice* (Third Ed.), ed. by L. Mátyás and P. Sevestre. Berlin: Springer-Verlag, 187–216. [2216]

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. [2216]

KASAHARA, H., AND K. SHIMOTSU (2009): "Nonparametric Identification of Finite Mixture Models of Dynamic Discrete Choices," *Econometrica*, 77, 135–175. [2217]

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. [2220]

KIVIET, J. F. (1995): "On Bias, Inconsistency, and Efficiency of Various Estimators in Dynamic Panel Data Models," *Journal of Econometrics*, 68, 53–78. [2229]

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. [2216]

LEE, Y. (2012): "Bias in Dynamic Panel Models Under Time Series Misspecification," *Journal of Econometrics*, 169, 54–60. [2229]

LEE, Y., AND P. C. B. PHILLIPS (2015): "Model Selection in the Presence of Incidental Parameters," *Journal of Econometrics*, 188, 474–489. [2223]

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. [2227]

——— (2009): "On the Distribution of Penalized Maximum Likelihood Estimators: The LASSO, SCAD, and Thresholding," *Journal of Multivariate Analysis*, 100, 2065–2082. [2227]

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. [2240]

LIAO, Z. (2013): "Adaptive GMM Shrinkage Estimation With Consistent Moment Selection," *Econometric Theory*, 29, 857–904. [2227]

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. [2217,2220]

LOAYZA, N., K. SCHMIDT-HEBBEL, AND L. SERVÉN (2000): "Saving in Developing Countries: An Overview," *The World Bank Economic Review*, 14, 393–414. [2241,2242]

LU, X., AND L. SU (2016): "Shrinkage Estimation of Dynamic Panel Data Models With Interactive Fixed Effects," *Journal of Econometrics*, 190, 148–175. [2227]

MIGUEL, E., S. SATYANATH, AND E. SERGENTI (2004): "Economic Shocks and Civil Conflict: An Instrumental Variables Approach," *Journal of Political Economy*, 112, 725–753. [2242]

NUNN, N., AND N. QIAN (2014): "US Food Aid and Civil Conflict," *American Economic Review*, 104, 1630–1666. [2242]

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. [2229]

PHILLIPS, P. C. B., AND D. SUL (2007a): "Transition Modeling and Econometric Convergence Tests," *Econometrica*, 75, 1771–1855. [2216,2217,2220]

——— (2007b): "Bias in Dynamic Panel Estimation With Fixed Effects, Incidental Trends and Cross Section Dependence," *Journal of Econometrics*, 137, 162–188. [2229]

QIAN, J., AND L. SU (2015): "Shrinkage Estimation of Regression Models With Multiple Structural Changes," *Econometric Theory*, first published online 23 June 2015, 1–58, DOI:10.1017/S0266466615000237. [2218]

RODRIK, D. (2000): "Saving Transitions," *The World Bank Economic Review*, 14, 481–507. [2240]

SARAFIDIS, V., AND N. WEBER (2015): "A Partially Heterogenous Framework for Analyzing Panel Data," *Oxford Bulletin of Economics and Statistics*, 77, 274–296. [2217]

SU, L., AND Q. CHEN (2013): "Testing Homogeneity in Panel Data Models With Interactive Fixed Effects," *Econometric Theory*, 29, 1079–1135. [2216]

SU, L., Z. SHI, AND P. C. B. PHILLIPS (2014): "Identifying Latent Structures in Panel Data," Cowles Foundation Discussion Papers 1965, Yale University. [2228]

——— (2016): "Supplement to 'Identifying Latent Structures in Panel Data'," *Econometrica Supplemental Material*, 84, http://dx.doi.org/10.3982/ECTA12560. [2219]

SUN, Y. (2005): "Estimation and Inference in Panel Structure Models," Working Paper, Dept. of Economics, UCSD. [2217,2220]

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. [2218]

TIBSHIRANI, R. J. (1996): "Regression Shrinkage and Selection via the LASSO," *Journal of the Royal Statistical Society, Series B*, 58, 267–288. [2222]

WANG, H., R. LI, AND C.-L. TSAI (2007): "Tuning Parameter Selectors for the Smoothly Clipped Absolute Deviation Method," *Biometrika*, 94, 553–568. [2227]

WHITE, H. (2001): *Asymptotic Theory for Econometricians*. London: Emerald. [2232]

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. [2218,2221,2222]

School of Economics, Singapore Management University, 90 Stamford Road, Singapore 178903, Singapore; ljsu@smu.edu.sg,

Dept. of Economics, The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China; zhentao.shi@cuhk.edu.hk,

and

Cowles Foundation for Research in Economics, Yale University, P.O. Box 208281, New Haven, CT 06520, U.S.A., University of Auckland, University of Southampton, and Singapore Management University; peter.phillips@yale.edu.

Co-editor Elie Tamer handled this manuscript.

Manuscript received June, 2014; final revision received December, 2015.