Taylor & Francis
Taylor & Francis Group

Journal of Business & Economic Statistics

ISSN: 0735-0015 (Print) 1537-2707 (Online) Journal homepage: https://www.tandfonline.com/loi/ubes20

# Sieve Estimation of Time-Varying Panel Data Models With Latent Structures

Liangjun Su, Xia Wang & Sainan Jin

# Sieve Estimation of Time-Varying Panel Data Models With Latent Structures

**Liangjun Su**
School of Economics, Singapore Management University, Singapore (ljsu@smu.edu.sg)

**Xia Wang**
Lingnan (University) College, Sun Yat-Sen University, Guangzhou, Guangdong, China, and Fujian Key Laboratory of Statistical Sciences, Xiamen University, Xiamen, Fujian, China (wangxia25@mail.sysu.edu.cn)

**Sainan Jin**
School of Economics, Singapore Management University, Singapore (snjin@smu.edu.sg)

We propose a heterogeneous time-varying panel data model with a latent group structure that allows the coefficients to vary over both individuals and time. We assume that the coefficients change smoothly over time and form different unobserved groups. When treated as smooth functions of time, the individual functional coefficients are heterogeneous across groups but homogeneous within a group. We propose a penalized-sieve-estimation-based classifier-Lasso (C-Lasso) procedure to identify the individuals' membership and to estimate the group-specific functional coefficients in a single step. The classification exhibits the desirable property of uniform consistency. The C-Lasso estimators and their post-Lasso versions achieve the oracle property so that the group-specific functional coefficients can be estimated as well as if the individuals' membership were known. Several extensions are discussed. Simulations demonstrate excellent finite sample performance of the approach in both classification and estimation. We apply our method to study the heterogeneous trending behavior of GDP per capita across 91 countries for the period 1960–2012 and find four latent groups.

KEY WORDS: Classifier-Lasso; Functional coefficient; Heterogeneity; Latent structure; Panel data; Penalized sieve estimation; Polynomial splines; Time-varying coefficients.

## 1. INTRODUCTION

Longitudinal or panel datasets have become widely available nowadays. Analysis of panel datasets has various advantages over that of pure cross-sectional or time series datasets, among which the most important one is perhaps that the panel data provide researchers a flexible way to model both heterogeneity among cross-sectional units and possible structural changes over time. For example, influenced by preference changes, technological progress, institutional transformation, and economic transition, the functional relationships between economic variables may change over time. For this reason, numerous studies have been devoted to test for structural changes in panel data models; see Han and Park (1989), Bai and Lluís Carrion-I-Silvestre (2009), Bai (2010), Kim (2011), Chen and Huang (2014), Li, Qian, and Su (2016), and Qian and Su (2016), among others. On the other hand, panel 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. Despite the fact that traditional panel data models frequently assume homogeneous slopes for the ease of estimation and inference, such an assumption has been frequently rejected in empirical studies (e.g., Lee, Pesaran, and Smith 1997; Durlauf, Kourtellos, and Minkin 2001; Juárez and Steel 2010; Su and Chen 2013) and there has been increasing interest in modeling slope heterogeneity in panel data models.

Although individual heterogeneity and structural changes are likely to coexist, existing panel data models only address at most one of these two important features. First, the studies on the panel data models with structural changes can be grouped into two categories, one is to consider abrupt changes and the other is to model smooth changes. For the former approach, see, for example, Bai (2010), Kim (2011), and Qian and Su (2016). The latter approach is mainly motivated from the time-varying (functional) coefficient model or nonparametric regression model in the time series framework. For example, Li, Chen, and Gao (2011) generalize Cai, Fan, and Yao's (2000) and Cai's (2007) time-varying coefficient model to the panel data framework, and develop a local linear dummy variable approach to estimate the functional coefficients; Robinson (2012) introduced a nonparametric trending model with cross-sectional dependence


Journal of Business & Economic Statistics
April 2019, Vol. 37, No. 2
DOI: 10.1080/07350015.2017.1340299
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/jbes.

and estimates the trend by the kernel method; Chen, Gao, and Li (2012) extend Robinson's (2012) nonparametric trending model to a semiparametric partially linear panel data model. Nevertheless, all parameters of interest, of finite or infinite dimension, in these models are assumed to be common across all cross-sectional units. Second, econometricians and statisticians have tried to address the potential slope heterogeneity in panel data models for a long time, say, through the random coefficient models in econometrics (e.g., Hsiao 2014, chap. 6; Hsiao and Pesaran 2008) and the random effects model in statistics (e.g., Diggle et al. 2003, chap. 9). More recently, Su, Shi, and Phillips (2016, SSP hereafter) propose a novel variant of Lasso to estimate heterogeneous linear panel data models, where the slope parameters are heterogeneous across groups but homogeneous within a group and the group membership is unknown. But they do not allow the coefficients to change over time.

In this article, we propose a heterogeneous time-varying panel data model with latent group structures to capture individual heterogeneity and smooth structural changes over time simultaneously. To the best of our knowledge, this is the first model to capture these two important features together. As individual heterogeneity and smooth structural changes are likely to coexist, our model appears more realistic than existing models and is expected to have much broader empirical applications. Following Cai (2007), we model the time-varying coefficients as smooth functions of time which can be estimated by nonparametric sieve or kernel methods. We could allow each individual unit to have distinct functional coefficients and estimate them individually but only with a slow convergence rate. Here, we adopt the latent group structure advocated by SSP and assume that the individuals belong to $K$ different groups, and the individual functional coefficients are heterogeneous across groups but homogeneous within a group. The major difficulty lies in the fact that the individuals' group membership is unknown. Our interest is to infer the individuals' group membership and estimate the group-specific functional coefficients at the same time.

In terms of statistical methodology, we propose a penalized-sieve-estimation-based classifier-Lasso (C-Lasso) procedure to identify the individuals' membership and to estimate the group-specific functional coefficients simultaneously. Since our estimation procedure is an iterative procedure and computationally involved, we prefer the sieve method to the kernel method to approximate the unknown functional coefficients. In particular, we propose to use polynomial B-splines given their good approximation properties and stable numerical properties; see, for example, Huang, Wu, and Zhou (2004), Huang and Shen (2004), and Xue and Yang (2006). The penalty term in our penalized sieve estimation (PSE) is constructed in the spirit of SSP's C-Lasso procedure which aims to shrink each individual coefficient to one of the $K$ unknown groups. Our procedure achieves classification and estimation in a single step. The classification exhibits the desirable property of uniform consistency. The PSE-based C-Lasso estimators and their post-Lasso versions achieve the oracle property of Fan and Li (2001) so that the group-specific functional coefficients can be estimated as well as if the individuals' membership were known. We also propose a data-driven method to determine the number of groups. Simulations demonstrate excellent finite-sample performance of our approach in both classification and estimation. We apply our method to study the heterogeneous trending behavior of GDP per capita across 91 countries for the period 1960–2012 and find four latent groups.

It is worth mentioning that recently grouping or homogeneity pursuit has generated a lot of interest in statistics. The fused Lasso of Tibshirani et al. (2005) can be regarded as an effort of exploring slope homogeneity. Bondell and Reich (2008) proposed a method called OSCAR to simultaneously select variables while grouping them into predictive clusters. Shen and Huang (2010) developed an algorithm called grouping pursuit by using the truncated $L_1$ penalty to penalize differences for all pairs of coordinates. Such an algorithm is further extended by Zhu, Shen, and Pan (2013) to allow for simultaneous grouping pursuit and feature selection. To explore homogeneity of coefficients, Ke, Fan, and Wu (2015) proposed a new method called clustering algorithm in regression via data-driven segmentation (CARDS), which is extended to the panel setup by Wang, Phillips, and Su (2017). Nevertheless, almost all of these papers consider linear data models in the cross-sectional framework.

The rest of the article is organized as follows. In Section 2, we introduce our time-varying panel data model with latent group structures. In Section 3, we consider the PSE for this model. We examine the asymptotic properties of the estimators in Section 4 and discuss several possible extensions in Section 5. Section 6 provides Monte Carlo study and empirical illustration. Section 7 concludes. All proofs of the main results are relegated to online Appendix A. Further technical details and the numerical algorithm are contained on the online supplementary Appendix.

*Notation.* For an $m \times n$ real matrix $A$, we denote its transpose as $A'$, its Frobenius norm as $\|A\|(\equiv [\mathrm{tr}(AA')]^{1/2})$ and its Moore–Penrose generalized inverse as $A^+$. When $A$ is symmetric, we use $\mu_{\max}(A)$ and $\mu_{\min}(A)$ to denote its largest and smallest eigenvalues, respectively. Let $\|A\|_{\mathrm{sp}}(\equiv [\mu_{\max}(AA')]^{1/2})$ denote the spectral norm of $A$. $\mathbb{I}_a$ and $\mathbf{0}_{a\times b}$ denote the $a \times a$ identity matrix and $a \times b$ matrix of zeros. $\mathbf{1}\{\cdot\}$ denotes the indicator function. We use "p.s.d." to abbreviate "positive semidefinite." 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 $T$ tend to infinity jointly. For a vector-valued function $\alpha(\cdot)$ defined on [0, 1], we use $\|\alpha\|_2$ to denote its $L_2$-norm: $\|\alpha\|_2 \equiv \{\int_0^1 \|\alpha(v)\|^2 dv\}^{1/2}$. Given sequences of positive numbers $a_{NT}$ and $b_{NT}$, $a_{NT} \lesssim b_{NT}$ and $b_{NT} \gtrsim a_{NT}$ mean $a_{NT}/b_{NT}$ is bounded, and $a_{NT} \asymp b_{NT}$ means that both $a_{NT} \lesssim b_{NT}$ and $a_{NT} \gtrsim b_{NT}$ hold. When $a_{NT}$ and $b_{NT}$ are random, $a_{NT} \lesssim b_{NT}$ and $b_{NT} \gtrsim a_{NT}$ mean $a_{NT}/b_{NT}$ is stochastically bounded and $a_{NT} \asymp b_{NT}$ means that both $a_{NT}/b_{NT}$ and $b_{NT}/a_{NT}$ are stochastically bounded.

## 2. TIME-VARYING PANEL STRUCTURE MODEL

In this section, we introduce the time-varying panel structure model. The dependent variable $Y_{it}$ is generated according to the following time-varying panel structure model:

$$Y_{it} = \gamma_i + \beta_{it}' X_{it} + u_{it}, \quad u_{it} = \sigma_i(X_{it})\varepsilon_{it}, \tag{2.1}$$

where $i = 1, 2, \ldots, N$, $t = 1, 2, \ldots, T$, $X_{it}$ is a $p \times 1$ vector of regressors, $\gamma_i$'s are unobserved individual fixed effects that may be correlated with some components of $X_{it}$ and are assumed to be different for different individuals, $\varepsilon_{it}$ has mean zero and variance one and is independent of the process $\{X_{it}\}$ so that $u_{it}$ is the idiosyncratic error term with conditional variance $\sigma_i^2(X_{it})$ given $X_{it}$, and $\beta_{it} = \beta_i(t/T)$ is a $p \times 1$ vector of time-varying slope coefficients exhibiting the following latent group structure:

$$\beta_{it} = \sum_{k=1}^{K} \alpha_k(t/T) \cdot \mathbf{1}\{i \in G_k\}. \tag{2.2}$$

We assume that $\|\alpha_j - \alpha_k\|_2 \neq 0$ for any $j \neq k$, $\cup_{k=1}^{K} G_k = \{1, 2, \ldots, N\}$, and $G_j \cap G_k = \varnothing$ for any $j \neq k$. Let $N_k = \#G_k$ denote the cardinality of the set $G_k$. For the moment we assume that the number of groups, $K$, is known and fixed, but each individual's group membership is unknown. We will propose an information criterion to determine $K$ in Section 4.4.

Interestingly, our model in (2.1) and (2.2) does not appear as restrictive as the time-invariant panel data models considered in Lin and Ng (2012), Bonhomme and Manresa (2015), and SSP. All the latter authors assume that an individual cannot change its group identity during the whole sampling period. As a matter of fact, this restrictive assumption also serves as an important motivation for our article. To see this point, we can go back to the SSP's framework. When the regression coefficients do not change over time, the model is given by

$$Y_{it} = \gamma_i + \beta_i' X_{it} + u_{it}, \quad i = 1, \ldots, N, \quad t = 1, \ldots, T,$$

where $\beta_i$'s have some grouped patterns. For simplicity, suppose that there are only two groups with the first and second half of individuals belonging to Groups 1 and 2, respectively. In this case, the number of groups (2 here) and the group identity for each individual remain fixed during the whole time period. To allow for the change of group membership for some individuals, it is natural to model $\beta_i$ as $\beta_i(t/T)$. In this case, we say that individuals $i$ and $j$ belong to the same group (say Group 1) if only if $\beta_i(t/T) = \beta_j(t/T)$ for $t = 1, \ldots, T$. It is possible that

$$\beta_i(t/T) = \beta_j(t/T) \quad \text{for all } t = 1, \ldots, T_0 \quad \text{and}$$
$$\beta_i(t/T) \neq \beta_j(t/T) \quad \text{for some } t = T_0 + 1, \ldots, T,$$

in which case $i$ and $j$ belong to the same group (say Group 1) until time $T_0$ and different groups after that. In this case, the total number of groups is generally not 2 but 3 at least, and our PSE method introduced below can identify the emergence of new groups asymptotically. That is, by enlarging the number of groups ($K$), we effectively allow the change of group membership for some individuals over the whole time period. In essence, the number of groups should not be regarded as given at the beginning of the sampling period. Instead, it is determined throughout the whole sampling period, which makes our model very attractive in comparison with existing panel structure models. In short, the change of group membership has been built into our model through the use of time-varying functional coefficients.

Our interest is to estimate the time-varying group-specific functional coefficients $\alpha_k(\cdot)$, $k = 1, 2, \ldots, K$, and to infer each individual's group identity. Following the literature on smooth time-varying regression models (e.g., Cai 2007; Robinson 2012; Chen, Gao, and Li 2012, Zhang, Su, and Phillips 2011), we assume that $\beta_i(\cdot)$'s and $\alpha_k(\cdot)$'s are smooth functions of $t/T$. See also Robinson (1989, 1991) for the discussion on the use of $t/T$ rather than $t$ as an argument of the functions.

Our model (2.1) is fairly general, and it includes a variety of panel data models as special cases.

1. If $X_{it} = 1$ and $\beta_i(\cdot) = \beta(\cdot)$ for some function $\beta(\cdot)$ and for each $i = 1, \ldots, N$, then the model in (2.1) becomes the nonparametric trending panel data model studied by Robinson (2012):
$$Y_{it} = \gamma_i + \beta(t/T) + u_{it}. \tag{2.3}$$
2. If $\beta_i(\cdot) = \beta(\cdot)$ for some function $\beta(\cdot)$ and for each $i = 1, \ldots, N$, then ( 2.1) becomes the time-varying functional coefficient panel data model studied by Li, Chen, and Gao (2011):
$$Y_{it} = \gamma_i + \beta(t/T)' X_{it} + u_{it}. \tag{2.4}$$
3. If $\beta_i(v) = \beta_i$ and $\alpha_k(v) = \alpha_k$ for any $v \in (0, 1]$, $i = 1, \ldots, N$, and $k = 1, \ldots, K$, then model (2.1) becomes the linear time-invariant panel structure model considered by SSP.
4. If $X_{it} = 1$, then model (2.1) becomes the nonparametric trending panel structure model:
$$Y_{it} = \gamma_i + \beta_i(t/T) + u_{it}, \tag{2.5}$$
where $\beta_{it} = \beta_i(t/T)$ satisfies the latent group structure in (2.2). Obviously, this model generalizes that of Robinson (2012) to allow for heterogeneous trending behavior.

In sum, our model in (2.1) can be regarded as an extension of that of SSP or that of Li, Chen, and Gao (2011). It extends the time-invariant model of SSP to allow time-varying coefficients and the homogeneous functional coefficient model of Li, Chen, and Gao (2011) to allow heterogeneous time-varying functional coefficients. It captures the smooth structural changes over time and the individuals' heterogeneity across groups simultaneously, and is thus expected to have much broader empirical applications than existing models in the literature. For example, as our empirical application demonstrates, the logarithm of the gross domestic product (GDP) per capita across countries exhibit heterogenous grouped patterns over time. For another example, the beneficial effects of foreign direct investment (FDI) on economic growth in host countries may exhibit both smooth structural changes and cross-country heterogeneity (see Cai, Chen, and Fang 2014). In either case, one has to apply the methodology developed in this article.

Hereafter, we use the superscript 0 to denote the true values or functions. In particular, we use $\beta_i^0(\cdot)$ and $\alpha_k^0(\cdot)$ to denote the true functional coefficients and $G_k^0$ the true value of $G_k$.

## 3. PENALIZED SIEVE ESTIMATION

In this section, we introduce the PSE method.

### 3.1 Sieve Approximation of Time-Varying Coefficients

We propose to estimate $\beta_i(v)$ and $\alpha_k(v)$ by polynomial splines of order $d$. Let $J_0 = J_0(N, T)$ be a prescribed integer that depends on $(N, T)$. Divide $[0, 1]$ into $(J_0 + 1)$ subintervals $I_j =$

$[\upsilon_j, \upsilon_{j+1})$ for $j = 0, 1, \ldots, J_0 - 1$ and $I_{J_0} = [\upsilon_{J_0}, 1]$, where $\mathcal{V} \equiv \{\upsilon_j\}_{j=1}^{J_0}$ is a sequence of equally spaced points (interior knots),

$$\upsilon_{-(d-1)} = \cdots = \upsilon_{-1} = \upsilon_0 = 0 < \upsilon_1 < \upsilon_2 < \cdots < \upsilon_{J_0} < 1 = \upsilon_{J_0+1} = \cdots = \upsilon_{J_0+d},$$

$\upsilon_j = jh$ for $j = 1, \ldots, J_0$, and $h = 1/(J_0 + 1)$ denotes the distance between two neighboring points. Let $\mathbb{G} = \mathbb{G}_{d,\mathcal{V}}$ denote the space of polynomial splines of order $d$ based on $\mathcal{V}$. It consists of functions $g$ satisfying: (i) $g$ is a polynomial of degree $d - 1$ on each of the subintervals $\{I_j\}_{j=0}^{J_0}$, (ii) for $d \geq 2$, $g$ is $d - 2$ times continuously differentiable on [0, 1]. Let $J = J_0 + d$. We use $B(\upsilon) = (B_{-d+1}(\upsilon), B_{-d+2}(\upsilon), \ldots, B_{J_0}(\upsilon))'$ to denote a basis system of the space $\mathbb{G}$. In this article, we focus on B-splines of order $d$ (or degree $d - 1$) because of the good approximation properties of splines and the stable numerical properties of B-splines. In particular, we will use cubic B-splines in our simulations and application, corresponding to $d = 4$. For more discussions on splines or B-splines, we refer the readers directly to Schumaker (1981), DeVore and Lorentz (1993), de Boor (2001), or the survey article by Chen (2007). See the online Appendix A for some basic properties of B-splines that are used in our analysis.

Given the spline basis system $B(\upsilon)$, we can approximate the square-integrable functions $\beta_i(\upsilon)$ and $\alpha_k(\upsilon)$ by $\pi_i'B(\upsilon)$ and $\omega_k'B(\upsilon)$ for some $J \times p$ matrices $\pi_i = (\pi_{i,1}, \ldots, \pi_{i,p})$ and $\omega_k = (\omega_{k,1}, \ldots, \omega_{k,p})$. Note that for notational simplicity we choose the same basis functions with the same interior knots and polynomial order to approximate different functions of interest. Then, we can rewrite the model in (2.1) as

$$Y_{it} = \gamma_i + [X_{it} \otimes B(t/T)]'\text{vec}(\pi_i) + e_{it}, \tag{3.1}$$

where $e_{it} = u_{it} + \beta_{it}'X_{it} - [X_{it} \otimes B(t/T)]'\text{vec}(\pi_i)$, and $\pi_i = \omega_k$ if $i \in G_k$ for $i = 1, \ldots, N$ and $k = 1, \ldots, K$.

## 3.2 Penalized Sieve Estimation of $\pi_i$ and $\omega_k$

Given the representation of the model in (3.1), we could estimate $\pi_i$ by minimizing the following least-squares objective function:

$$Q_{0,NT}(\pi, \gamma) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\{Y_{it} - \gamma_i - Z_{it}'\text{vec}(\pi_i)\}^2,$$

where $\pi = (\text{vec}(\pi_1)', \ldots, \text{vec}(\pi_N)')'$, $\gamma = (\gamma_1, \ldots, \gamma_N)'$, and $Z_{it} \equiv X_{it} \otimes B(t/T)$. Since the individual effects $\gamma_i$'s are not of primary interest, we concentrate them out and obtain the following concentrated objective function:

$$Q_{1,NT}(\pi) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}[\tilde{Y}_{it} - \tilde{Z}_{it}'\text{vec}(\pi_i)]^2, \tag{3.2}$$

where $\tilde{Z}_{it} = Z_{it} - \frac{1}{T}\sum_{t=1}^{T} Z_{it}$ and $\tilde{Y}_{it} = Y_{it} - \frac{1}{T}\sum_{t=1}^{T} Y_{it}$. By minimizing the objective function in (3.2), we obtain the least-squares estimator of $\pi$ by $\tilde{\pi} = (\text{vec}(\tilde{\pi}_1)', \ldots, \text{vec}\ (\tilde{\pi}_N)')'$, where

$$\text{vec}(\tilde{\pi}_i) = \left(\frac{1}{T}\sum_{t=1}^{T}\tilde{Z}_{it}\tilde{Z}_{it}'\right)^{+}\left(\frac{1}{T}\sum_{t=1}^{T}\tilde{Z}_{it}\tilde{Y}_{it}\right) \quad \text{for } i = 1, \ldots, N. \tag{3.3}$$

Let $\omega = (\text{vec}(\omega_1)', \ldots, \text{vec}(\omega_K)')'$ and $\tilde{Z}_i = (\tilde{Z}_{i1}, \ldots, \tilde{Z}_{iT})'$. To estimate $\pi$ and $\omega$ together, we consider the following penalized least-squares objective function:

$$Q_{NT,\lambda}^{(K)}(\pi, \omega) = Q_{1,NT}(\pi) + \frac{\lambda}{N}\sum_{i=1}^{N}\tilde{\sigma}_i^{2-K}\prod_{k=1}^{K}\|\tilde{V}_i\text{vec}(\pi_i - \omega_k)\|, \tag{3.4}$$

where $\lambda = \lambda(N, T)$ is a tuning parameter, $\tilde{V}_i = \{\text{diag}(\frac{J}{T}\tilde{Z}_i'\tilde{Z}_i)\}^{1/2}$, and $\tilde{\sigma}_i = \{\frac{1}{T}\sum_{t=1}^{T}[\tilde{Y}_{it} - \tilde{Z}_{it}'\text{vec}(\tilde{\pi}_i)]^2\}^{1/2}$ is an estimator of the sample standard deviation of $\{u_{it}\}_{t=1}^{T}$. Minimizing objective function in (3.4) yields the PSE-based *classifier-Lasso* (C-Lasso hereafter) estimators $\hat{\pi} = (\text{vec}(\hat{\pi}_1)', \ldots, \text{vec}(\hat{\pi}_N)')'$ and $\hat{\omega} = (\text{vec}(\hat{\omega}_1)', \ldots, \text{vec}(\hat{\omega}_K)')'$ of $\pi$ and $\omega$, respectively.

The objective function in (3.4) is in the same spirit as that in SSP if we replace $\tilde{\sigma}_i$ and $\tilde{V}_i$ by one and an identity matrix, respectively. We apply $\tilde{\sigma}_i$ and $\tilde{V}_i$ to ensure the scale-invariant property of our objective function: $Q_{NT}(\pi, \omega)$ remains unchanged when one changes the scales of either $\tilde{Z}_{it}$ or $\tilde{Y}_{it}$ by changing those of $X_{it}$ and $Y_{it}$. Note that the objective function in (3.4) is not convex in $\pi$ or $\omega$. In the supplementary Appendix C we provide an iterative algorithm to obtain the estimators $\hat{\pi}$ and $\hat{\omega}$. Given these estimators, we can obtain the estimators of $\beta_i(\upsilon)$'s and $\alpha_k(\upsilon)$'s as follows:

$$\hat{\beta}_i(\upsilon) = \hat{\pi}_i'B(\upsilon), \quad \text{and} \quad \hat{\alpha}_k(\upsilon) = \hat{\omega}_k'B(\upsilon) \quad \text{for } i = 1, \ldots, N, \quad \text{and} \quad k = 1, \ldots, K. \tag{3.5}$$

We will study the asymptotic properties of these estimators in the next section.

*Remark 1.* Alternatively, one can extend the $K$-means algorithm to our framework. The latter approach was adopted by Lin and Ng (2012) in linear panel data models with additive fixed effects, by Bonhomme and Manresa (2015) for linear models with grouped additive effects, and by Ando and Bai (2016) in linear panel data models with grouped interactive fixed effects. There are three major differences between the C-Lasso and $K$-means methods. First, the C-Lasso estimation needs to specify the number of groups ($K$) and the tuning parameter ($\lambda$), while the $K$-means estimation requires the specification of $K$ only. Despite this, it is hard to tell which method should be preferred as the additional parameter $\lambda$ may offer some degree of freedom in finite samples. Second, the $K$-means algorithm forces all individuals to be classified into one of the $K$ groups while the C-Lasso procedure may leave some individuals unclassified for small values of $\lambda$. For large values of $\lambda$, the C-Lasso can also classify all individuals to one of the $K$ groups and produce similar results as the $K$-means algorithm. But it is hard to tell whether we should force all individuals to be classified. In fact, when $T$ is not large, forcing all individuals to be classified via either the $K$-means algorithm or the use of a large value of $\lambda$ for the C-Lasso tends to yield a large proportion of misclassification. In contrast, when $\lambda$ is not large enough, the C-Lasso allows for some individuals to be left unclassified, which could yield better finite sample performance for the estimators of the group-specific functional coefficients especially when $T$ is not large. For large $T$, the choice of $\lambda$ does not matter very much and the two methods generally produce highly consistent classification results. Third, computationally the C-Lasso is much less demanding than the $K$-means algorithm. This is because the $K$-means estimation is NP-hard and the C-Lasso problem, despite its nonconvexity, can be transformed into a sequence of convex problems (see the supplementary Appendix C).

We will show that C-Lasso estimators of the group-specific functional coefficients and their post-Lasso versions are oracally efficient—they are asymptotically equivalent to the corresponding infeasible estimators of the group-specific functional coefficients that are obtained by knowing all individual group identities. Following the theoretical studies in Bonhomme and Manresa (2015) and Ando and Bai (2016), we conjecture that the $K$-means estimators also exhibit the oracle property. If this is the case, the two types of estimators for the group-specific functional coefficients are asymptotically equivalent.

## 4. ASYMPTOTIC THEORY

In this section, we first establish the preliminary convergence rates for $\hat{\beta}_i(v)$ and $\hat{\alpha}_k(v)$, and then study the consistency of the classification. We also establish the asymptotic distributions of $\hat{\alpha}_k(v)$'s and their post-Lasso versions and study the determination of $K$.

### 4.1 Preliminary Rates of Convergence for Coefficient Estimates

Let $\min_{i,t}$ and $\max_{i,t}$ denote $\min_{1\le i\le N}\min_{1\le t\le T}$ and $\max_{1\le i\le N}\max_{1\le t\le T}$, respectively. Let $C^{(\gamma)}[0,1]$ denote the space of functions that are $\gamma$th-order continuously differentiable on $[0,1]$, where $\gamma \ge 1$. Let $X_{it}^{(2)} = X_{it}$ if $X_{it}$ does not contain 1 and $X_{it} = (1, X_{it}^{(2)\prime})'$ otherwise.

To study the consistency of $\hat{\beta}_i(v)$ and $\hat{\alpha}_k(v)$, we make the following assumptions.

*Assumption A1.*

(i) Let $X_i^{(2)} = (X_{i1}^{(2)}, \ldots, X_{iT}^{(2)})'$ and $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iT})'$. $\{(X_i^{(2)}, \varepsilon_i)\}$ are independently distributed over $i$.
(ii) For each $i = 1, \ldots, N$, the process $\{(X_{it}^{(2)}, \varepsilon_{it}),\ t = 1, 2, \ldots\}$ is strong mixing with mixing coefficient $\alpha(j)$ satisfying $\alpha(j) \le c_\alpha \rho^j$ for some $c_\alpha < \infty$ and $\rho \in [0, 1)$.
(iii) $\max_{i,t} E\|X_{it}\|^q \le \bar{c}_x < \infty$ and $\max_{i,t} E|u_{it}|^q \le \bar{c}_u < \infty$ for some $q > 6$.
(iv) There exist positive constants $\underline{c}_{xx}$ and $\bar{c}_{xx}$ such that $\underline{c}_{xx} \le \min_{i,t} \mu_{\min}(\mathrm{Var}\ (X_{it}^{(2)})) \le \max_{i,t}\ \mu_{\max}(E(X_{it}X_{it}')) \le \bar{c}_{xx}$ whenever $X_{it} \ne 1$. There exists $\underline{c}_\sigma > 0$ such that $\lim_{T\to\infty} \min_i \bar{\sigma}_{i,T}^2 \ge \underline{c}_\sigma$, where $\bar{\sigma}_{i,T}^2 \equiv T^{-1}\sum_{t=1}^T E(u_{it}^2)$.
(v) For $k = 1, 2, \ldots, K$, $\alpha_k^0 \in C^{(\gamma)}[0,1]$ for some $2 \le \gamma \le d+1$. There exists $\underline{c}_\alpha > 0$ such that $\min_{1\le j\ne k\le K} \|\alpha_j^0 - \alpha_k^0\|_2 \ge \underline{c}_\alpha$.
(vi) $N_k/N \to \tau_k \in (0,1)$ for each $k = 1, \ldots, K$ as $N \to \infty$.

*Assumption A2.*

(i) As $(N,T) \to \infty$, $J \to \infty$, $J^2/T \to 0$, $\lambda J^{(K+1)/2} \to 0$, and $N^2 T^{1-q/2}(\ln N)^{q\epsilon_0/2} \to 0$ for some $\epsilon_0 > 1$.
(ii) As $(N,T) \to \infty$, $J/\ln T \to \infty$, $\lambda J^{\gamma+(K-1)/2} \to \infty$ and $\lambda\sqrt{T} J^{(K-1)/2}/(\ln T)^3 \to \infty$, and $\lambda(\ln T)^{\upsilon} \to 0$ for some $\upsilon > 0$.

Assumptions A1(i)–(ii) require that $\{X_{it}^{(2)}, \varepsilon_{it}\}$ be independently distributed over individuals and weakly dependent over time. We assume that $\{(X_{it}^{(2)}, \varepsilon_{it}),\ t = 1, 2, \ldots\}$ is a strong mixing process with a geometric decay rate, which can be satisfied by many well-known linear processes such as ARMA processes and a variety of nonlinear processes. Note that we allow serial correlation in $\{u_{it}\}$ and lagged dependent variables in $X_{it}^{(2)}$. When $X_{it}^{(2)}$ contains lagged dependent variables (e.g., $Y_{i,t-1}$), the strong mixing condition imposes some restrictions on the fixed effects $\lambda_i$ and the error terms $u_{it}$. In this case, we can assume that $\lambda_i$'s are nonrandom and $u_{it}$'s have Lebesgue-integrable characteristic functions (Andrews 1984). If $\lambda_i$'s are stochastic, we can follow Hahn and Kuersteiner (2011) and Su and Chen (2013) and adopt the concept of conditional strong mixing, where the mixing coefficients are defined by conditioning on the fixed effects. A1(iii) imposes moment conditions for $X_{it}$ and $u_{it}$. A1(iv) imposes the identification condition that ensures the large dimensional matrix $\frac{J}{T}\sum_{t=1}^T \tilde{Z}_{it}\tilde{Z}_{it}'$ (see (3.3)) is asymptotically nonsingular and the preliminary estimator $\tilde{\sigma}_i^2$ of $\bar{\sigma}_{i,T}^2$ is uniformly bounded away from zero with probability approaching one (w.p.a.1); see Lemmas A.3 and A.5 in the online Appendix A. Note that $X_{it}$ may contain 1 or not and we allow $X_{it} = 1$. When $X_{it} = 1$, the first part of Assumption A1(iv) is not relevant. Assumption A1(vi) is also assumed in SSP and it implies that each group has an asymptotically nonnegligible number of members as $N \to \infty$.

The first part of Assumption A1(v) imposes smooth conditions on the group-specific functional coefficients $\alpha_k^0$ (and thus the individual functional coefficients $\beta_i^0$). By Theorem 12.6 in de Boor (2001, p. 149), there exists $\omega_k^0 \in \mathbb{R}^J$ such that

$$\sup_{v\in[0,1]} \left\|\alpha_k^0(v) - \omega_k^{0\prime}B(v)\right\| = O(h^\gamma) = O(J^{-\gamma}) \quad \text{for} \quad k = 1, \ldots, K. \tag{4.1}$$

Similarly, there exists $\pi_i^0 \in \mathbb{R}^J$ such that

$$\sup_{v\in[0,1]} \left\|\beta_i^0(v) - \pi_i^{0\prime}B(v)\right\| = O(h^\gamma) = O(J^{-\gamma}) \text{ for } i = 1, \ldots, N, \tag{4.2}$$

and $\pi_i^0 = \omega_k^0$ if $i \in G_k^0$. The second part of A1(v) implies conditions for the identification of the group-specific functional coefficients. By the triangle inequality, (A.1) in online Appendix A, and (4.1), we have

$$\begin{aligned}\underline{c}_\alpha &\le \left\|\alpha_j^0 - \alpha_k^0\right\|_2 \le \left\|(\omega_j^0 - \omega_k^0)'B\right\|_2 + \left\|\alpha_j^0 - \omega_j^{0\prime}B\right\|_2 \\ &\quad + \left\|\alpha_k^0 - \omega_k^{0\prime}B\right\|_2 \\ &= \left\{\mathrm{tr}\left(\left(\omega_j^0 - \omega_k^0\right)'\int B(v)B(v)'dv\left(\omega_j^0 - \omega_k^0\right)\right)\right\}^{1/2} + O(J^{-\gamma}) \\ &\asymp J^{-1/2}\left\|\omega_j^0 - \omega_k^0\right\| \text{ for any } j \ne k.\end{aligned}$$

That is,

$$\left\|\omega_j^0 - \omega_k^0\right\| \asymp J^{1/2} \quad \text{for any } j \ne k, \tag{4.3}$$

which will be used in the proof of Theorem 4.1.

Assumptions A2 imposes conditions on $N$, $T$, $J$, and $\lambda$. It requires that $\lambda$ shrinks to zero at a suitable rate such that the penalty term can effectively distinguish individuals in one group from those in the other groups asymptotically. The range in which $\lambda$ converges to zero mainly depends on $T$ and $J$ but not $N$. The intuition is clear: $J$ controls the bias from the sieve approximation and the effective number of parameters in the sieve estimation; $T$, in conjunction with $J$, controls the speed at which one can estimate the individual functional coefficients $\pi_i(\cdot)$'s and the group-specific functional coefficients $\omega_k(\cdot)$'s. Clearly, A2 allows the choice of a wide range of values of $\lambda$ and $J$ provided the corresponding functions are sufficiently smooth and $q$ is large enough.

The following theorem studies the preliminary convergence rates of the estimators of $\pi_i^0$ and $\omega_k^0$.

*Theorem 4.1.* Suppose Assumptions A1 and A2(i) hold. Then

(i) $\|\hat{\pi}_i - \pi_i^0\| = O_P(J^{-\gamma+1/2} + JT^{-1/2} + \lambda J^{(K+1)/2})$ for $i = 1, 2, \ldots, N$,
(ii) $N^{-1}\sum_{i=1}^{N}\|\hat{\pi}_i - \pi_i^0\|^2 = O_P(J^{-2\gamma+1} + J^2T^{-1})$,
(iii) $\|\hat{\omega}_{(k)} - \omega_k^0\| = O_P(J^{-\gamma+1/2} + JT^{-1/2})$ for $k = 1, 2, \ldots, K$, where $(\hat{\omega}_{(1)}, \ldots, \hat{\omega}_{(K)})$ is a suitable permutation of $(\hat{\omega}_1, \ldots, \hat{\omega}_K)$.

Theorems 4.1(i) and (ii) establish the pointwise and mean-square convergence of $\hat{\pi}_i$, respectively. The first two terms, namely, $J^{-\gamma+1/2}$ and $JT^{-1/2}$ in part (i) reflect the contributions of the usual asymptotic bias and variance terms of sieve estimation, respectively, and the last term $\lambda J^{(K+1)/2}$ signifies the effect of the penalty term in the C-Lasso procedure. For small enough $\lambda$, that is, if $\lambda \lesssim \max(T^{-1/2}J^{(1-K)/2}, J^{-\gamma-K/2})$, we obtain the usual convergence rate for the coefficient estimates when B-splines are used. Interestingly, the mean-square convergence of $\hat{\pi}_i$ and the pointwise convergence of $\hat{\omega}_{(k)}$ do not depend on $\lambda$, which is analogous to the results of SSP in the parametric setting. See the proof in online Appendix A for details. In particular, we show in the proof of Theorem 4.1(iii) that the convergence rate of $\hat{\omega}_{(k)}$ depends on the mean-square but not the pointwise convergence rate of $\hat{\pi}_i$. Note that Assumption A2(i) ensures that $\|\hat{\pi}_i - \pi_i^0\| = o_P(1)$ and $\|\hat{\omega}_{(k)} - \omega_k^0\| = o_P(1)$.

For notational simplicity, hereafter we will write $\hat{\omega}_k$ for $\hat{\omega}_{(k)}$ and $\hat{\alpha}_k(\cdot)$ for $\hat{\alpha}_{(k)}(\cdot)$ where $\hat{\alpha}_{(k)}(\cdot) = \hat{\omega}_{(k)}' B(\cdot)$. Then, we can define the estimated groups:

$$\hat{G}_k = \{i \in \{1, 2, \ldots, N\} : \hat{\pi}_i = \hat{\omega}_k\} \text{ for } k = 1, \ldots, K. \quad (4.4)$$

The following corollary establishes the pointwise and $L_2$ convergence rates of $\hat{\beta}_i(\cdot)$ and $\hat{\alpha}_k(\cdot)$.

*Corollary 4.1.* Suppose Assumptions A1 and A2(i) hold. Then

(i) $\sup_{v\in[0,1]}\|\hat{\beta}_i(v) - \beta_i^0(v)\| = O_P(J^{-\gamma+1/2} + JT^{-1/2} + \lambda J^{(K+1)/2})$ and $\int_0^1 \|\hat{\beta}_i(v) - \beta_i^0(v)\|^2 dv = O_P(J^{-2\gamma} + JT^{-1} + \lambda^2 J^K)$ for $i = 1, 2, \ldots, N$;
(ii) $\sup_{v\in[0,1]}\|\hat{\alpha}_k(v) - \alpha_k^0(v)\| = O_P(J^{-\gamma+1/2} + JT^{-1/2})$ and $\int_0^1 \|\hat{\alpha}_k(v) - \alpha_k^0(v)\|^2 dv = O_P(J^{-2\gamma} + JT^{-1})$ for $k = 1, 2, \ldots, K$.

Similar results hold when we replace the integration by the sample mean. That is, $\frac{1}{T}\sum_{t=1}^{T}||\hat{\beta}_i(t/T) - \beta_i^0(t/T)||^2 = O_P(J^{-2\gamma} + JT^{-1} + \lambda^2 J^K)$ for $i = 1, 2, \ldots, N$, and $\frac{1}{T}\sum_{t=1}^{T}||\hat{\alpha}_k(t/T) - \alpha_k^0(t/T)||^2 = O_P(J^{-2\gamma} + JT^{-1})$ for $k = 1, 2, \ldots, K$.

## 4.2 Classification Consistency

We define the following sequences of events:

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

where $i = 1, 2, \ldots, N$ and $k = 1, 2, \ldots, K$. Let $\hat{E}_{k,NT} = \cup_{i\in G_k^0}\hat{E}_{k,NT,i}$ and $\hat{F}_{k,NT} = \cup_{i\in\hat{G}_k}\hat{F}_{k,NT,i}$. The events $\hat{E}_{k,NT}$ and $\hat{F}_{k,NT}$ mimic Type I and Type II errors in statistical tests: $\hat{E}_{k,NT}$ denotes the error event of not classifying an individual in the $k$th group into the $k$th group; $\hat{F}_{k,NT}$ denotes the error event of classifying an individual that does not belong to the $k$th group into the $k$th group. Following SSP's definition, we say that the classification is *uniformly consistent* if $P(\cup_{k=1}^K \hat{E}_{k,NT}) \to 0$ and $P(\cup_{k=1}^K \hat{F}_{k,NT}) \to 0$ as $(N, T) \to 0$, that is, the probability of committing either type of errors shrinks to zero asymptotically.

The following theorem establishes the classification consistency of our method.

*Theorem 4.2.* Suppose Assumptions A1 and A2 hold. Then

(i) $P(\cup_{k=1}^K \hat{E}_{k,NT}) \le \sum_{k=1}^K P(\hat{E}_{k,NT}) \to 0$ as $(N, T) \to \infty$;
(ii) $P(\cup_{k=1}^K \hat{F}_{k,NT}) \le \sum_{k=1}^K P(\hat{F}_{k,NT}) \to 0$ as $(N, T) \to \infty$.

Theorem 4.2 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.

*Remark 2.* Let $\hat{G}_0$ denote the group of individuals in $\{1, 2, \ldots, N\}$ that are not classified into any of the $K$ groups, that is, $\hat{G}_0 = \{1, 2, \ldots, N\}\backslash(\cup_{k=1}^K \hat{G}_k)$. Define the events $\hat{H}_{iNT} = \{i \in \hat{G}_0\}$. Theorem 4.2(i) implies that $P(\cup_{1\le i\le N}\hat{H}_{iNT}) \le \sum_{k=1}^K P(\hat{E}_{kNT}) \to 0$. That is, all individuals can be classified into one of the $K$ groups w.p.a.1. Nevertheless, when $T$ is not large, it is possible for a small number of individuals to be left unclassified if we stick with the classification rule in (4.4). To ensure that all individuals are classified into one of the $K$ groups, say, if one is sure that there are no isolated individuals as in our simulations, one can modify the classifier a little bit. In particular, for any $i \in \hat{G}_0$ we can classify it to $\hat{G}_l$ for some $l = 1, \ldots, K$ if

$$\|\hat{\pi}_i - \hat{\omega}_l\| = \min\{\|\hat{\pi}_i - \hat{\omega}_1\|, \ldots, \|\hat{\pi}_i - \hat{\omega}_K\|\}.$$

Since the event $\cup_{1\le i\le N}\hat{H}_{iNT}$ occurs with probability approaching zero, we can follow SSP to ignore it in subsequent asymptotic analysis and restrict our attention to the classification rule in (4.4) to avoid confusion. That is, $\hat{G}_k = \{i \in \{1, \ldots, N\} : \hat{\pi}_i = \hat{\omega}_k\}$ for $k = 1, \ldots, K$.

## 4.3 Post-Lasso Estimator and Oracle Property

Given the estimated groups $\{\hat{G}_k, k = 1, \ldots, K\}$ defined in (4.4), we can readily pool the observations within each estimated group to obtain the post-Lasso sieve estimator of the corresponding group-specific functional coefficients by

$$\hat{\alpha}_{\hat{G}_k}(v) = \hat{\omega}'_{\hat{G}_k} B(v), \tag{4.5}$$

where for $k = 1, \ldots, K$,

$$\mathrm{vec}(\hat{\omega}_{\hat{G}_k}) = \left( \sum_{i \in \hat{G}_k} \sum_{t=1}^{T} \tilde{Z}_{it} \tilde{Z}'_{it} \right)^{+} \sum_{i \in \hat{G}_k} \sum_{t=1}^{T} \tilde{Z}_{it} \tilde{Y}_{it}. \tag{4.6}$$

When the group identity for each individual is known, we obtain the oracle estimators: $\hat{\alpha}_{G_k^0}(v) = \hat{\omega}'_{G_k^0} B(v)$, where $\mathrm{vec}(\hat{\omega}_{G_k^0}) = (\sum_{i \in G_k^0} \sum_{t=1}^{T} \tilde{Z}_{it} \tilde{Z}'_{it})^{+} \sum_{i \in G_k^0} \sum_{t=1}^{T} \tilde{Z}_{it} \tilde{Y}_{it}$.

Let $u_i = (u_{i1}, \ldots, u_{iT})'$. Then, $\mathrm{var}(u_i|X_i) = \Sigma_i^{1/2} V_i \Sigma_i^{1/2}$, where $\Sigma_i = \mathrm{diag}(\sigma_i^2(X_{i1}), \ldots, \sigma_i^2(X_{iT}))$ and $V_i = E(\varepsilon_i \varepsilon_i')$. Let $X_{it}^{(\sigma)} = \sigma_i(X_{it}) X_{it}$. We add the following assumption:

*Assumption A3.*

(i) For $k = 1, \ldots, K$, there exist two positive constants $\underline{c}_V$ and $\bar{c}_V$ such that $0 < \underline{c}_V \le \lim_{(N,T)\to\infty} \min_{i \in G_k^0} \mu_{\min}(V_i) \le \lim_{(N,T)\to\infty} \max_{i \in G_k^0} \mu_{\max}(V_i) \le \bar{c}_V \delta_{NT}$ for some nondecreasing sequence $\delta_{NT}$ satisfying $\delta_{NT} N^{-1} \to 0$.

(ii) $\max_{i,t} E\|X_{it}^{(\sigma)}\|^q \le \bar{c}_x^{(\sigma)}$ for some constant $\bar{c}_x^{(\sigma)} < \infty$ and $q > 6$.

(iii) There exist two positive constants $\underline{c}_{xx}^{(\sigma)}$ and $\bar{c}_{xx}^{(\sigma)}$ such that $\underline{c}_{xx}^{(\sigma)} \le \min_{i,t} \mu_{\min}(\mathrm{Var}(X_{it}^{(2,\sigma)})) \le \max_{i,t} \mu_{\max}(E(X_{it}^{(\sigma)} X_{it}^{(\sigma)\prime})) \le \bar{c}_{xx}^{(\sigma)}$, where $X_{it}^{(2,\sigma)} = X_{it}^{(\sigma)}$ if $X_{it}^{(\sigma)}$ does not contain nonrandom element and $X_{it}^{(2,\sigma)}$ is a collection of the random elements of $X_{it}^{(\sigma)}$ otherwise.

(iv) As $(N, T) \to \infty$, $NTJ^{-2\gamma} \to 0$.

Assumption A3(i) imposes conditions on the variance–covariance matrix of $\varepsilon_i$ to verify the Lindeberg condition for a central limit theorem to hold. For its minimum eigenvalue, we only need it to be bounded away from zero by a positive constant. Such a condition can be easily satisfied. For example, if we follow Huang, Wu, and Zhou (2004) and assume that $\varepsilon_{it}$ can be decomposed into two components: $\varepsilon_{it} = \varepsilon_{it}^{(1)} + \varepsilon_{it}^{(2)}$, where $\varepsilon_{it}^{(1)}$ is an arbitrary mean zero process and $\varepsilon_{it}^{(2)}$ is an independent process of "measurement errors" that are independent over time and have mean zero and constant positive variance $\sigma^2$, then the lower bound for the minimum eigenvalue is given by $\sigma^2$. For the maximum eigenvalue, we allow it to be constant or divergent as $(N, T) \to \infty$. If there is no serial correlation, then $V_i$ is diagonal and the condition requires that the maximum value of the diagonal elements of $V_i$ to be of order $O(\delta_{NT})$, where $\delta_{NT} = o(N)$. For any $m \times n$ matrix $A = \{a_{ij}\}$, note that $\|A\|_{\mathrm{sp}}^2 \le \|A\|_1 \|A\|_\infty$, where $\|A\|_1 = \max_{1 \le j \le n} \sum_{i=1}^{m} |a_{ij}|$ and $\|A\|_\infty = \max_{1 \le i \le m} \sum_{j=1}^{n} |a_{ij}|$. Since $V_i$ is symmetric and p.s.d., we have $\|V_i\|_1 = \|V_i\|_\infty$ and $\mu_{\max}(V_i) = \|V_i\|_{\mathrm{sp}} \le \|V_i\|_1$. So our maximum eigenvalue condition will be satisfied if the column sums of $V_i$ are bounded by the order $\delta_{NT} = o(N)$. This condition is automatically satisfied under our strong mixing and moment conditions on $\{\varepsilon_{it}\}$ if $\varepsilon_{it}$ has the same second moment across $i$. Assumption A3(i) says that the central limit theorem can hold without the strong mixing conditions or identical moments across individuals. Assumptions A3(ii) and (iii) parallel the first part of A1(iii) and A1(iv), respectively. If $\sigma_i(X_{it}) = \underline{\sigma}_0 > 0$ a.s., A3(ii) and (iii) would become redundant. A3(iv) ensures that the asymptotic biases of the estimators $\hat{\alpha}_k(v)$ and $\hat{\alpha}_{\hat{G}_k}(v)$ do not contribute to their limit distributions.

The following theorem gives the oracle property of the PSE-based C-Lasso estimators and their post-Lasso versions.

*Theorem 4.3.* Suppose Assumptions A1–A3 hold. Then, for $k = 1, 2, \ldots, K$, we have

(i) $\sqrt{N_k T/J} \mathbb{S}_k^{-1/2} [\hat{\alpha}_k(v) - \alpha_k(v)] \xrightarrow{D} N(0, \mathbb{I}_p)$,

(ii) $\sqrt{N_k T/J} \mathbb{S}_k^{-1/2} [\hat{\alpha}_{\hat{G}_k}(v) - \alpha_k(v)] \xrightarrow{D} N(0, \mathbb{I}_p)$,

where $\mathbb{S}_k^{-1/2}$ is the symmetric square root of $\mathbb{S}_k^{-1}$, $\mathbb{S}_k = (\mathbb{I}_p \otimes B(v))' (J \bar{\mathbb{Q}}_{k,\tilde{z}\tilde{z}})^{-1} \{\frac{J}{N_k T} \sum_{i \in G_k^0} \tilde{Z}_i' \Sigma_i^{1/2} V_i \Sigma_i^{1/2} \tilde{Z}_i\} (J \bar{\mathbb{Q}}_{k,\tilde{z}\tilde{z}})^{-1} (\mathbb{I}_p \otimes B(v))$, and $\bar{\mathbb{Q}}_{k,\tilde{z}\tilde{z}} = \frac{1}{N_k T} \sum_{i \in G_k^0} \sum_{t=1}^{T} \tilde{Z}_{it} \tilde{Z}'_{it}$.

Theorem 4.3 indicates that both the C-Lasso estimator $\hat{\alpha}_k(v)$ and the post-Lasso version $\hat{\alpha}_{\hat{G}_k}(v)$ are asymptotically equivalent to the infeasible estimator $\hat{\alpha}_{G_k^0}(v)$. The latter can be obtained only if one knows all individuals' group identity. In this sense, our C-Lasso estimators and their post-Lasso versions have the oracle efficiency. Despite the asymptotic equivalence between the C-Lasso and post-Lasso estimators, it is well known that the post-Lasso estimators typically outperform the C-Lasso estimators and are thus recommended for practical use.

*Remark 3.* As a referee points out, it does not appear transparent to see the relative rates on $N$ and $T$ to obtain all the asymptotic properties so far because they are related to Assumptions A2(i)–(ii) and A3(iv). To gain some insight, we focus on the case where all functions of interest are sufficiently smooth so that the approximation biases are asymptotically negligible and all terms associated with $\gamma$ in Assumptions A2(ii) and A3(iv) do not matter. In this case, the single most important requirement on $(N, T)$ is given by the last part of Assumption A2(i) because other conditions are essentially imposed on $J$ and $\lambda$. This part of the assumption holds if $q$ or $T$ or both are sufficiently large. If $\{(X_{it}^{(2)}, u_{it})\}$ is sub-Gaussian as in Bonhomme and Manresa (2015), $q = \infty$ and $N$ can grow at any polynomial rate faster than $T$. The first two conditions in Assumption A2(i) require that $J$ diverge to infinity at a rate slower than $\sqrt{T}$ (i.e., $1 \ll J \ll T^{1/2}$) and all the other conditions in Assumption A2 would be satisfied if

$$\max\left( J^{-\gamma - \frac{K-1}{2}}, T^{-1/2} J^{-\frac{K-1}{2}} (\ln T)^3 \right) \ll \lambda \ll \min\left( J^{-\frac{K+1}{2}}, J^{-1}/(\ln T)^{\upsilon} \right).$$

It is easy to see that such $\lambda$ exists under that condition that $1 \ll J \ll T^{1/2}$ and $\gamma \ge 2$. When $N$ and $T$ pass to infinity at the same rate (as is commonly assumed in large dimensional macro panels), our choice of $J_0$ (or $J \equiv J_0 + d$) and $\lambda$ in the following simulation and application sections would meet the above conditions and Assumption A3(iv) provided $q > 6$ and $\gamma > 3$.

*Remark 4.* For statistical inference, one needs to estimate $\mathbb{S}_k$ consistently. Suppose that one can estimate $\Theta_k \equiv \frac{J}{N_k T} \sum_{i \in \hat{G}_k} \tilde{Z}_i' \Sigma_i^{1/2} V_i \Sigma_i^{1/2} \tilde{Z}_i$ by $\tilde{\Theta}_k$ such that $||\tilde{\Theta}_k - \Theta_k||_{\mathrm{sp}} = o_P(1)$. Define

$$\tilde{\mathbb{S}}_k = (\mathbb{I}_p \otimes B(v))' (J \tilde{\mathbb{Q}}_{k,\tilde{z}\tilde{z}})^{-1} \tilde{\Theta}_k (J \tilde{\mathbb{Q}}_{k,\tilde{z}\tilde{z}})^{-1} (\mathbb{I}_p \otimes B(v)),$$

where $\tilde{\mathbb{Q}}_{k,\tilde{z}\tilde{z}} = \frac{1}{\hat{N}_k T}\sum_{i\in\hat{G}_k}\sum_{t=1}^{T}\tilde{Z}_{it}\tilde{Z}_{it}'$ and $\hat{N}_k = \#\hat{G}_k$. Following SSP, we can readily show that $\hat{N}_k/N_k \xrightarrow{P} 1$, $||J\tilde{\mathbb{Q}}_{k,\tilde{z}\tilde{z}} - J\bar{\mathbb{Q}}_{k,\tilde{z}\tilde{z}}||_{\text{sp}} = o_P(1)$, and $\tilde{\mathbb{S}}_k - \mathbb{S}_k = o_P(1)$. Under various primitive conditions, one can propose the corresponding consistent estimator $\tilde{\Theta}_k$; see, for example, Su and Jin (2012). The procedure is standard and thus omitted.

### 4.4 Determination of the Number of Groups

In practice, the exact number of groups, $K$, is typically unknown. Here, we assume that $K$ is bounded from above by a finite integer $K_{\max}$ and denote the true value of $K$ as $K_0$. We propose a BIC-type information criterion (IC) to determine the data-driven choice of $K$.

By minimizing the objective function in (3.4), we obtain the C-Lasso estimates $\{\hat{\pi}_i(K,\lambda)\}_{i=1}^N$ and $\{\hat{\omega}_k(K,\lambda)\}_{k=1}^K$ of $\{\pi_i\}_{i=1}^N$ and $\{\omega_k\}_{k=1}^K$ for $K = 1,\ldots,K_{\max}$, where we make the dependence of the estimators on $(K,\lambda)$ explicit. When $K = 1$, one effectively works on the nonpenalized sieve estimation. As before, we can classify individual $i$ into group $\hat{G}_k(K,\lambda)$ if and only if $\hat{\pi}_i(K,\lambda) = \hat{\omega}_k(K,\lambda)$, that is,

$$\hat{G}_k(K,\lambda) = \{i \in \{1,2,\ldots,N\} : \hat{\pi}_i(K,\lambda) = \hat{\omega}_k(K,\lambda)\} \quad \text{for } k = 1,\ldots,K. \tag{4.7}$$

Denote $\hat{G}(K,\lambda) = \{\hat{G}_1(K,\lambda),\ldots,\hat{G}_K(K,\lambda)\}$. Conditional on the classification, we could define the post-Lasso estimate of $\omega_k$ as follows:

$$\text{vec}(\hat{\omega}_{\hat{G}_k(K,\lambda)}) = \left(\sum_{i\in\hat{G}_k(K,\lambda)}\sum_{t=1}^{T}\tilde{Z}_{it}\tilde{Z}_{it}'\right)^{+}\sum_{i\in\hat{G}_k(K,\lambda)}\sum_{t=1}^{T}\tilde{Z}_{it}\tilde{Y}_{it}. \tag{4.8}$$

In addition, define $\hat{\sigma}^2_{\hat{G}(K,\lambda)} = \frac{1}{NT}\sum_{k=1}^{K}\sum_{i\in\hat{G}_k(K,\lambda)}\sum_{t=1}^{T}[\tilde{Y}_{it} - \tilde{Z}_{it}'\text{vec}(\hat{\omega}_{\hat{G}_k(K,\lambda)})]^2$. Then, we choose $\hat{K}$ to minimize the following information criterion:

$$\text{IC}(K,\lambda) = \ln\left[\hat{\sigma}^2_{\hat{G}(K,\lambda)}\right] + \rho_{NT}JpK, \tag{4.9}$$

where $\rho_{NT}$ is a tuning parameter.

Let $G^{(K)} = (G_{K,1},\ldots,G_{K,K})$ be any $K$ -partition of the set of individual indices $\{1,2,\ldots,N\}$. Let $\mathcal{G}_K$ denote the collection of such partitions. Let $\hat{\sigma}^2_{G^{(K)}} = \frac{1}{NT}\sum_{k=1}^{K}\sum_{i\in G_{K,k}}\sum_{t=1}^{T}[\tilde{Y}_{it} - \tilde{Z}_{it}'\text{vec}(\hat{\omega}_{G_{K,k}})]^2$, where $\text{vec}(\hat{\omega}_{G_{K,k}}) = (\sum_{i\in G_{K,k}}\sum_{t=1}^{T}\tilde{Z}_{it}\tilde{Z}_{it}')^{+}\sum_{i\in G_{K,k}}\sum_{t=1}^{T}\tilde{Z}_{it}\tilde{Y}_{it}$. The following assumptions are useful in the asymptotic development.

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

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

Assumption A4 requires that all under-fitted models yield asymptotic mean square errors that are larger than $\sigma_0^2$, which is delivered by the true model. A5 imposes usual conditions for the consistency of model selection, namely, the penalty coefficient $\rho_{NT}$ cannot shrink to zero either too fast or too slowly.

The following theorem justifies the use of (4.9) as a criterion to select $K$.

*Theorem 4.4.* Suppose that Assumptions A1–A5 hold. Then $P(\hat{K} = K_0)\to 1$ as $(N,T)\to\infty$.

Theorem 4.4 implies that the IC could determine the number of groups w.p.a.1. Of course, to implement it, one still needs to choose the tuning parameter $\rho_{NT}$. Motivated by BIC, we will set $\rho_{NT} = J_0\ln(NT)/(NT)$ in our simulations and application.

## 5. EXTENSIONS

In this section, we discuss several possible extensions of the time-varying panel structure model studied above.

### 5.1 Mixed Time-Varying Panel Structure Models

Consider the time-varying panel data models where some of the functional coefficients in $\beta_{it}$'s are common across all individuals, whereas the others are group-specific. Write $\beta_{it} = (\beta_{it}^{(1)\prime},\beta_t^{(2)\prime})'$ where $\beta_{it}^{(1)} = \beta_i^{(1)}(t/T)$ is a $p_1\times 1$ vector of heterogenous functional coefficients that exhibits the following latent group structure

$$\beta_{it}^{(1)} = \sum_{k=1}^{K}\alpha_k(t/T)\cdot\mathbf{1}\{i\in G_k\},$$

and $\beta_t^{(2)}$ is $(p-p_1)\times 1$ vector of homogenous functional coefficients. Partition $X_{it}$ conformably as $X_{it} = (X_{it}^{(1)\prime},X_{it}^{(2)\prime})'$. The model becomes

$$Y_{it} = \gamma_i + \beta_{it}^{(1)\prime}X_{it}^{(1)} + \beta_t^{(2)\prime}X_{it}^{(2)} + u_{it}, \tag{5.1}$$

where $u_{it}$ and $\gamma_i$ are defined as before. Our PSE method can be extended to this model straightforwardly. Given the spline basis system $B(v)$, we can approximate $\beta_i^{(1)}(v)$, $\beta^{(2)}(v)$, and $\alpha_k(v)$ by $\pi_i^{(1)\prime}B(v)$, $\pi^{(2)\prime}B(v)$, and $\omega_k'B(v)$, respectively. Let $\pi^{(1)} = (\text{vec}(\pi_1^{(1)})',\ldots,\text{vec}(\pi_N^{(1)})')'$, $\pi^{(2)} = \text{vec}(\pi^{(2)})$, and $\omega = (\text{vec}(\omega_1)',\ldots,\text{vec}(\omega_K)')'$. Now, we can estimate $\pi^{(1)}$, $\pi^{(2)}$ and $\omega$ by minimizing the following objective function:

$$Q_{NT,\lambda}^{(K)}(\pi^{(1)},\pi^{(\mathbf{2})},\omega) = Q_{1,NT}\left(\pi^{(1)},\pi^{(2)}\right) + \frac{\lambda}{N}\sum_{i=1}^{N}\tilde{\sigma}_i^{2-K}\prod_{k=1}^{K}\left\|\tilde{V}_i\text{vec}\left(\pi_i^{(1)}-\omega_k\right)\right\|, \tag{5.2}$$

where

$$Q_{1,NT}(\pi^{(1)},\pi^{(2)}) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{\tilde{Y}_{it} - \tilde{Z}_{it}^{(1)\prime}\text{vec}\left(\pi_i^{(1)}\right) - \tilde{Z}_{it}^{(2)\prime}\text{vec}(\pi^{(2)})\right\}^2,$$

$\tilde{Y}_{it} = Y_{it} - \frac{1}{T}\sum_{t=1}^{T}Y_{it}$, $\tilde{Z}_{it}^{(\ell)} = Z_{it}^{(\ell)} - \frac{1}{T}\sum_{t=1}^{T}Z_{it}^{(\ell)}$, $Z_{it}^{(\ell)} = X_{it}^{(\ell)}\otimes B(t/T)$ for $\ell = 1,2$, $\tilde{V}_i = \{\text{diag}(\frac{J}{T}\tilde{Z}_i'\tilde{Z}_i)\}^{1/2}$, $\tilde{Z}_i = (\tilde{Z}_{i1},\ldots,\tilde{Z}_{iT})'$, $\tilde{Z}_{it} = (\tilde{Z}_{it}^{(1)\prime},\tilde{Z}_{it}^{(2)\prime})'$, and $\tilde{\sigma}_i = \{\frac{1}{T}\sum_{t=1}^{T}[\tilde{Y}_{it} - \tilde{Z}_{it}'\text{vec}(\tilde{\pi}_i)]^2\}^{1/2}$, and $\tilde{\pi}_i = (\tilde{\pi}_i^{(1)},\tilde{\pi}^{(2)})$ is a preliminary estimate of $\pi_i = (\pi_i^{(1)},\pi^{(2)})$ obtained as in Section 3.2.

Let $\hat{\pi}^{(1)}$=(vec($\hat{\pi}_1^{(1)}$)$'$, ..., vec($\hat{\pi}_N^{(1)}$)$'$)$'$, $\hat{\pi}^{(2)}$=vec($\hat{\pi}^{(2)}$), and $\hat{\omega} = (\mathrm{vec}(\hat{\omega}_1)', \ldots, \mathrm{vec}(\hat{\omega}_K)')'$ be the estimators of $\pi^{(1)}$, $\pi^{(2)}$, and $\omega$, respectively. We obtain the estimators of $\beta_i^{(1)}(v)$'s, $\beta^{(2)}(v)$, and $\alpha_k(v)$'s as follows:

$$\hat{\beta}_i^{(1)}(v) = \hat{\pi}_i^{(1)\prime}B(v),\ \hat{\beta}^{(2)}(v) = \hat{\pi}^{(2)\prime}B(v),\ \text{and}\ \hat{\alpha}_k(v) = \hat{\omega}_k'B(v), \tag{5.3}$$

where $i = 1, \ldots, N$, and $k = 1, \ldots, K$. Following the analyses in Sections 4.1– 4.3, we can establish the asymptotic properties of the above estimators. In particular, we can establish the uniform consistency of the classification based on the PSE method and the oracle properties of $\hat{\alpha}_k(v)$ and $\hat{\beta}^{(2)}(v)$ and their post-Lasso versions. We omit the details due to the space constraint.

## 5.2 Unbalanced Panels

To broaden the applications of our model, we now consider an extension to unbalanced panels. Like Atak, Linton, and Xiao (2011) and for notational simplicity, we consider an unbalanced panel in which consecutive observations on individual units are available, but the number of time periods available may vary from unit to unit. The model becomes

$$Y_{it} = \gamma_i + \beta_{it}'X_{it} + u_{it},\ i = 1, \ldots, N,\ t = t_i, \ldots, T_i, \tag{5.4}$$

where $\beta_{it}$'s have the latent group pattern in (2.2), and the other notations are defined as in Section 2. Let $\tau_i = T_i - t_i + 1$ and $n = \sum_{i=1}^N \tau_i$. Note that $\tau_i$ and $n$ denote the total number of observations for individual $i$ and the whole sample, respectively. Now, we can estimate $\pi$ and $\omega$ by minimizing the following objective function:

$$Q_{n,\lambda}^{(K)}(\pi, \omega) = Q_{1,n}(\pi) + \frac{\lambda}{N}\sum_{i=1}^N \tilde{\sigma}_i^{2-K}\prod_{k=1}^K \|\tilde{V}_i\mathrm{vec}(\pi_i - \omega_k)\|, \tag{5.5}$$

where

$$Q_{1,n}(\pi) = \frac{1}{N}\sum_{i=1}^N \frac{1}{\tau_i}\sum_{t=t_i}^{T_i}\{\tilde{Y}_{it} - \tilde{Z}_{it}'\ \mathrm{vec}(\pi_i)\}^2,$$

$\tilde{Y}_{it} = Y_{it} - \frac{1}{\tau_i}\sum_{t=t_i}^{T_i} Y_{it}$, $\tilde{Z}_{it} = Z_{it} - \frac{1}{\tau_i}\sum_{t=t_i}^{T_i} Z_{it}$, $Z_{it} = X_{it} \otimes B(t/T)$, $\tilde{V}_i = \{\mathrm{diag}(\frac{J}{\tau_i}\tilde{Z}_i'\tilde{Z}_i)\}^{1/2}$, $\tilde{Z}_i = (\tilde{Z}_{it_i}, \ldots, \tilde{Z}_{iT_i})'$, $\tilde{\sigma}_i = \{\frac{1}{\tau_i}\sum_{t=t_i}^{T_i}[\tilde{Y}_{it} - \tilde{Z}_{it}'\mathrm{vec}(\tilde{\pi}_i)]^2\}^{1/2}$, and $\tilde{\pi}_i$ is a preliminary estimate of $\pi_i$ obtained as in Section 3.2. Let $\hat{\pi}$ =(vec($\hat{\pi}_1$)$'$, ..., vec($\hat{\pi}_N$)$'$)$'$ and $\hat{\omega} = (\mathrm{vec}(\hat{\omega}_1)', \ldots, \mathrm{vec}(\hat{\omega}_K)')'$ of $\pi$ and $\omega$, respectively. The formulas for the estimators $\hat{\beta}_i(v)$ and $\hat{\alpha}_k(v)$ of $\beta_i(v)$ and $\alpha_k(v)$ are the same as those given in (3.5). Define the estimated group $\hat{G}_k$ as in Section 4.1. The post-Lasso estimator of $\alpha_k(v)$ becomes

$$\hat{\alpha}_{\hat{G}_k}(v) = \hat{\omega}_{\hat{G}_k}'B(v), \tag{5.6}$$

where for $k = 1, \ldots, K$,

$$\mathrm{vec}(\hat{\omega}_{\hat{G}_k}) = \left(\sum_{i\in\hat{G}_k}\frac{1}{\tau_i}\sum_{t=t_i}^{T_i}\tilde{Z}_{it}\tilde{Z}_{it}'\right)^+ \sum_{i\in\hat{G}_k}\frac{1}{\tau_i}\sum_{t=t_i}^{T_i}\tilde{Z}_{it}\tilde{Y}_{it}. \tag{5.7}$$

Let $\underline{T} \equiv \min_{1\le i\le N}\tau_i$. To study the asymptotic properties of these estimators, we assume that $\underline{T} \to \infty$ and the conditions in Assumptions A1–A3 continue to hold with $T$ replaced by $\underline{T}$. We need $\underline{T}\to\infty$ for the pointwise and mean square convergence results in Theorem 4.1, which are needed for the proofs of the uniform classification consistency and the oracle properties of $\hat{\alpha}_k(v)$ and its post-Lasso version. With some change in notation, the results in Theorem 4.1, Corollary 4.2, and Theorems 4.3 and 4.4 continue to hold. In particular, the results in Theorem 4.4 become:

(i) $\sqrt{N_k\underline{T}/J}\mathbb{S}_k^{-1/2}[\hat{\alpha}_k(v) - \alpha_k(v)] \xrightarrow{D} N(0, \mathbb{I}_p)$,

(ii) $\sqrt{N_k\underline{T}/J}\mathbb{S}_k^{-1/2}[\hat{\alpha}_{\hat{G}_k}(v) - \alpha_k(v)] \xrightarrow{D} N(0, \mathbb{I}_p)$, where $\mathbb{S}_k = (\mathbb{I}_p \otimes B(v))'(J\bar{\mathbb{Q}}_{k,\tilde{z}\tilde{z}})^{-1}\{\frac{J}{N_k}\sum_{i\in G_k^0}\frac{\underline{T}}{\tau_i^2}\tilde{Z}_i'\Sigma_i^{1/2}V_i\Sigma_i^{1/2}\tilde{Z}_i\}(J\bar{\mathbb{Q}}_{k,\tilde{z}\tilde{z}})^{-1}(\mathbb{I}_p \otimes B(v))$, and $\bar{\mathbb{Q}}_{k,\tilde{z}\tilde{z}} = \frac{1}{N_k}\sum_{i\in G_k^0}\frac{1}{\tau_i}\sum_{t=t_i}^{T_i}\tilde{Z}_{it}\tilde{Z}_{it}'$ for $k = 1, \ldots, K$.

## 5.3 Panels with Cross-Sectional Dependence

We can also allow for cross-sectional dependence in our model. A popular way to introduce the cross-sectional dependence is via the use of the interactive fixed effects:

$$Y_{it} = \beta_{it}'X_{it} + \gamma_i'F_t + u_{it}, \tag{5.8}$$

where $\gamma_i$ and $F_t$ denote an $R \times 1$ vector of factor loadings and common factors, respectively, both of which can be correlated with $\{X_{it}\}$, $\beta_{it}$'s have the latent group pattern in (2.2), and the other notations are defined as in Section 2. When $R = 1$, $F_t = 1$, the model in (5.8) becomes the model in (2.1) with additive fixed effects. Let $F = (F_1, \ldots, F_T)'$ and $\Lambda$= $(\gamma_1, \ldots, \gamma_N)'$. Following Bai and Ng (2002), Moon and Weidner (2015), and Su and Ju (2017), we impose the identification restrictions: $T^{-1}F'F/T = \mathbb{I}_R$, $\Gamma'\Gamma$ = diagonal with descending diagonal elements. Let $Y_i \equiv (Y_{i1}, \ldots, Y_{iT})'$ and $Z_i \equiv (Z_{i1}, \ldots, Z_{iT})'$, where recall that $Z_{it} = X_{it} \otimes B(t/T)$. We propose to estimate $\{\pi_i\}$, $\{\omega_k\}$, $F$, and $\Gamma$ by minimizing the following penalized objective function:

$$Q_{0NT,\lambda}^{(K)}(\pi, \omega, F, \Gamma) = Q_{0,NT}(\pi, F, \Gamma) + \frac{\lambda}{N}\sum_{i=1}^N\prod_{k=1}^K \|\mathrm{vec}(\pi_i - \omega_k)\| \tag{5.9}$$

where $Q_{0,NT}(\pi, F, \Gamma) = \frac{1}{NT}\sum_{i=1}^N \|Y_i - Z_i\mathrm{vec}(\pi_i) - F\gamma_i\|^2$, and $F$ and $\Gamma$ satisfy the above identification restrictions. Following Moon and Weidner (2015) and Su and Ju (2017), we can concentrate $\Gamma$ and $F$ out in turn and obtain the profile objective function

$$Q_{NT,\lambda}^{(K)}(\pi, \omega) = Q_{1,NT}(\pi) + \frac{\lambda}{N}\sum_{i=1}^N\prod_{k=1}^K \|\mathrm{vec}(\pi_i - \omega_k)\|, \tag{5.10}$$

where $Q_{1,NT}(\pi) = \frac{1}{T}\sum_{r=R+1}^T \mu_r[\frac{1}{N}\sum_{i=1}^N (Y_i - Z_i\mathrm{vec}(\pi_i))(Y_i - Z_i\mathrm{vec}(\pi_i))']$ and $\mu_r(A)$ denotes the $r$th largest eigenvalue of $A$ by counting multiple eigenvalues multiple times.

Minimizing the criterion function in (5.10) produces the C-Lasso estimators $\hat{\pi}$ =(vec($\hat{\pi}_1$)$'$, ..., vec($\hat{\pi}_N$)$'$)$'$ and $\hat{\omega} = (\mathrm{vec}(\hat{\omega}_1)', \ldots, \mathrm{vec}(\hat{\omega}_K)')'$ of $\pi$ and $\omega$. The estimators $\hat{F}$ and $\hat{\Gamma}$ of $F$ and $\Gamma$ are obtained as the solutions to the following eigenvalue problem:

$$\left[\frac{1}{NT}\sum_{i=1}^{N}(Y_i - Z_i\mathrm{vec}(\hat{\pi}_i))(Y_i - Z_i\mathrm{vec}(\hat{\pi}_i))'\right]\hat{F} = \hat{F}V_{NT} \quad \text{and}$$
$$\hat{\Gamma} = (\hat{\gamma}_1, \ldots, \hat{\gamma}_N)', \tag{5.11}$$

where $V_{NT}$ is a diagonal matrix consisting of the $R$ largest eigenvalues of the above matrix in the square bracket, arranged in the descending order, and $\hat{\gamma}_i = T^{-1}\hat{F}'(Y_i - Z_i\mathrm{vec}(\hat{\pi}_i))$. The formulas for the estimators $\hat{\beta}_i(v)$ and $\hat{\alpha}_k(v)$ of $\beta_i(v)$ and $\alpha_k(v)$ are the same as those given in (3.5). Following the technical analyses in Su and Ju (2017) and those in Sections 4.1–4.3, we can establish the asymptotic properties of the above estimators. In particular, we can establish the uniform consistency of the classification and the oracle properties of $\hat{\alpha}_k(v)$ and its post-Lasso version. For brevity, we omit the details.

## 6. MONTE CARLO STUDY AND EMPIRICAL ILLUSTRATION

### 6.1 Monte Carlo Study

In this section, we evaluate the finite sample performance of the information criterion in determining the number of groups and the C-Lasso and post-Lasso estimates.

*6.1.1 Data-Generating Processes.* We consider three data-generation processes (DGPs). In all DGPs, the fixed effect $\gamma_i$ and the idiosyncratic error $u_{it}$ follow the standard normal distribution and are mutually independent across both $i$ and $t$. The observations in each DGP are drawn from three groups with $N_1 : N_2 : N_3 = 0.3 : 0.3 : 0.4$. We consider four combinations of the sample sizes with $N = 50$, 100 and $T = 40$, 80.

**DGP 1** (Trending panel structure model). $Y_{it}$ is generated via $Y_{it} = \gamma_i + \beta_i^0(t/T) + u_{it}$, where

$$\beta_i^0(v) = \begin{cases} \alpha_1^0(v) = 6F(v, 0.5, 0.1) & \text{if } i \in G_1^0 \\ \alpha_2^0(v) = 6[2v - 6v^2 + 4v^3 \\ \quad + F(v; 0.7, 0.05)] & \text{if } i \in G_2^0 \\ \alpha_3^0(v) = 6[4v - 8v^2 + 4v^3 \\ \quad + F(v; 0.6, 0.05)] & \text{if } i \in G_3^0 \end{cases}, \tag{6.1}$$

$F(\cdot; \mu, s) = \{1 + \exp[-(\cdot - \mu)/s]\}^{-1}$ denotes the cumulative distribution function of the logistic distribution with location and scale parameters given by $\mu$ and $s$, respectively.

**DGP 2** (Time-varying panel structure model with an exogenous regressor). $Y_{it}$ is generated via $Y_{it} = \gamma_i + \beta_{i,1}^0(t/T) + \beta_{i,2}^0(t/T)X_{it} + u_{it}$, where $\{X_{it}\}$ is an IID $N(0, 1)$ sequence, $\beta_{i,1}^0(v) = \frac{1}{2}\beta_i^0(v)$ with $\beta_i^0(v)$ given in (6.1),

$$\beta_{i,2}^0(v) = \begin{cases} \alpha_{1,2}^0(v) = 3[2v - 4v^2 + 2v^3 \\ \quad + F(v; 0.6, 0.1)] & \text{if } i \in G_1^0 \\ \alpha_{2,2}^0(v) = 3[v - 3v^2 + 2v^3 \\ \quad + F(v; 0.7, 0.04)] & \text{if } i \in G_2^0 \\ \alpha_{3,2}^0(v) = 3[0.5v - 0.5v^2 \\ \quad + F(v; 0.4, 0.07)] & \text{if } i \in G_3^0 \end{cases}, \tag{6.2}$$

and $F$ is defined as above. Here, the first element in the group-specific parameter vector $\alpha_k^0(\cdot)$ is given by $\alpha_{k,1}^0(\cdot) = \frac{1}{2}\alpha_k^0(\cdot)$ with $\alpha_k^0(\cdot)$ defined in (6.1). The left and right panels of Figure 1 depict the group-specific time trends $\alpha_{k,1}^0(\cdot)$ and $\alpha_{k,2}^0(\cdot)$ for $k = 1, 2, 3$, respectively.

**DGP 3** (Time-varying dynamic panel structure model). $Y_{it}$ is generated via $Y_{it} = \gamma_i + \beta_{i,3}^0(t/T)Y_{it-1} + u_{it}$, where

$$\beta_{i,3}^0(v) = \begin{cases} \alpha_{1,3}^0(v) = \frac{3}{2}[-0.5 + 2v - 5v^2 + 2v^3 \\ \quad + F(v; 0.6, 0.03)] & \text{if } i \in G_1^0 \\ \alpha_{2,3}^0(v) = \frac{3}{2}[-0.5 + v - 3v^2 + 2v^3 \\ \quad + F(v; 0.2, 0.04)] & \text{if } i \in G_2^0 \\ \alpha_{3,3}^0(v) = \frac{3}{2}[-0.5 + 0.5v - 0.5v^2 \\ \quad + F(v; 0.8, 0.07)] & \text{if } i \in G_3^0 \end{cases}, \tag{6.3}$$

and $F$ is defined as above.

In addition, we also check the performance of our method when the error terms exhibit weak cross-sectional dependence and when the number of groups is large. The simulation results are quite similar to those reported below. Due to the space limit, we do not reports these results in this article.

*6.1.2 Determination of the Number of Groups.* In this subsection, we assess the performance of (4.9) in determining the number of groups. We set $J_0 = \lfloor (NT)^{1/6} \rfloor$, the number of knots in the cubic B-spline approximation, where $\lfloor A \rfloor$ denotes

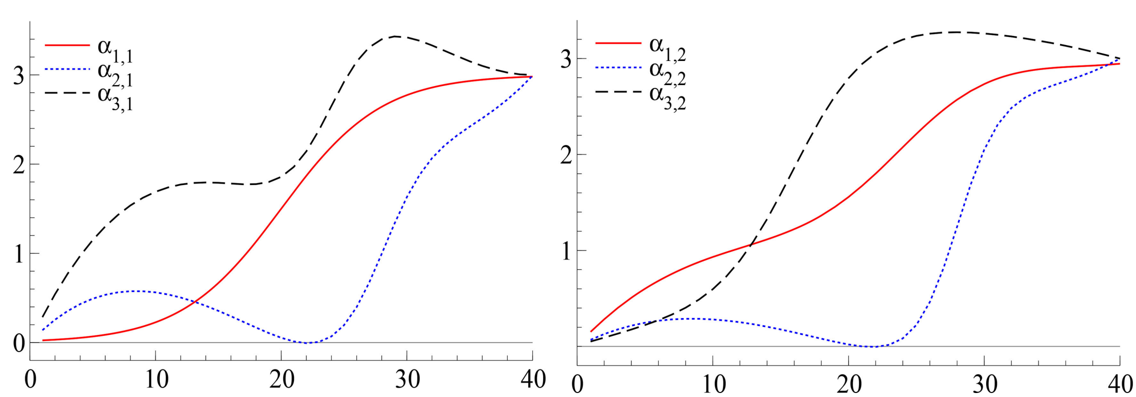


Figure 1. The plots of the group-specific functional coefficients in DGP 2 (left panel: solid, dotted and dashed lines for $\alpha_{1,1}^0(\cdot)$, $\alpha_{2,1}^0(\cdot)$, and $\alpha_{3,1}^0(\cdot)$, respectively in DGP 2; right panel: solid, dotted and dashed lines for $\alpha_{1,2}^0(\cdot)$, $\alpha_{2,2}^0(\cdot)$, and $\alpha_{3,2}^0(\cdot)$, respectively in DGP 2).

Table 1. The performance of information criterion in determining the number of groups

| DGP | $N$ | $T$ | $K=1$ | $K=2$ | $K=3$ | $K=4$ | $K=5$ | $K=6$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 50 | 40 | 0.000 | 0.015 | 0.985 | 0.000 | 0.000 | 0.000 |
| | 50 | 80 | 0.000 | 0.000 | 0.985 | 0.015 | 0.000 | 0.000 |
| | 100 | 40 | 0.000 | 0.000 | 0.995 | 0.005 | 0.000 | 0.000 |
| | 100 | 80 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 |
| 2 | 50 | 40 | 0.000 | 0.075 | 0.925 | 0.000 | 0.000 | 0.000 |
| | 50 | 80 | 0.000 | 0.000 | 0.990 | 0.010 | 0.000 | 0.000 |
| | 100 | 40 | 0.000 | 0.010 | 0.990 | 0.000 | 0.000 | 0.000 |
| | 100 | 80 | 0.000 | 0.000 | 0.980 | 0.020 | 0.000 | 0.000 |
| 3 | 50 | 40 | 0.000 | 0.120 | 0.880 | 0.000 | 0.000 | 0.000 |
| | 50 | 80 | 0.000 | 0.030 | 0.970 | 0.000 | 0.000 | 0.000 |
| | 100 | 40 | 0.000 | 0.020 | 0.980 | 0.000 | 0.000 | 0.000 |
| | 100 | 80 | 0.000 | 0.005 | 0.995 | 0.000 | 0.000 | 0.000 |

Note: The main entries are the empirical probability that a specific number of groups is selected based on 200 replications.

Table 2. Two types of classification error in percentages

| | | | $c_\lambda = 1$ | | $c_\lambda = 2$ | | $c_\lambda = 4$ | |
|---|---|---|---|---|---|---|---|---|
| DGP | $N$ | $T$ | $\bar{P}(\hat{E})$ | $\bar{P}(\hat{F})$ | $\bar{P}(\hat{E})$ | $\bar{P}(\hat{F})$ | $\bar{P}(\hat{E})$ | $\bar{P}(\hat{F})$ |
| 1 | 50 | 40 | 0.460 | 0.434 | 0.380 | 0.357 | 0.600 | 0.570 |
| | 50 | 80 | 0.110 | 0.080 | 0.000 | 0.000 | 0.010 | 0.009 |
| | 100 | 40 | 0.580 | 0.606 | 0.855 | 0.747 | 0.425 | 0.415 |
| | 100 | 80 | 0.015 | 0.015 | 0.005 | 0.005 | 0.010 | 0.010 |
| 2 | 50 | 40 | 2.930 | 2.734 | 2.870 | 2.617 | 2.100 | 1.959 |
| | 50 | 80 | 0.570 | 0.389 | 0.340 | 0.289 | 0.230 | 0.114 |
| | 100 | 40 | 1.645 | 1.547 | 2.665 | 2.460 | 1.765 | 1.700 |
| | 100 | 80 | 0.800 | 0.984 | 0.120 | 0.114 | 0.070 | 0.068 |
| 3 | 50 | 40 | 6.730 | 6.542 | 6.535 | 6.178 | 7.720 | 6.917 |
| | 50 | 80 | 2.175 | 2.052 | 1.995 | 1.873 | 2.230 | 2.208 |
| | 100 | 40 | 5.585 | 5.692 | 5.360 | 5.294 | 5.970 | 5.973 |
| | 100 | 80 | 1.135 | 1.087 | 1.050 | 1.008 | 1.190 | 1.084 |

the integer part of $A$. We set $\lambda = c_\lambda (NT)^{-(2K+3)/24}$ and consider various values of $c_\lambda$ to examine the sensitivity of the IC's performance to the choice of $\lambda$. We consider $c_\lambda = 1, 2, 4$ but only report the results for $c_\lambda = 1$ here to save space. The results for $c_\lambda = 2$ and 4 are quite similar and available upon request from the authors.

For each DGP, we simulate 200 datasets for each of the four combinations of $N$ and $T$. We evaluate the IC for $K = 1, 2, \ldots, K_{\max}$ with $K_{\max} = 6$ and select the optimal number of groups by minimizing the IC in (4.9). Table 1 reports the empirical probability that a specific number of groups is selected based on 200 replications. As shown in the table, our IC works fairly well.

*6.1.3 Classification and Estimation.* As shown in the previous subsection, the IC in Section 4.4 works fairly well in finite samples. In this subsection, we assume that the number of groups is known and focus on the classification and estimation.

We set the tuning parameter $\lambda$ as above. We set the initial values of $\pi_i$'s to be $\tilde{\pi}_i$'s and those of $\omega_k$'s to be zero. We have also tried other initial values and found that the classification and estimation results are quite similar to those reported here, suggesting the robustness of our algorithm to the initial values of parameters.

We run 200 replications for each DGP and classify individual $i$ into group $k$ if $\|\hat{\beta}_i - \hat{\alpha}_k\|$ achieves the minimum. To measure the accuracy of classification, we report two types of classification errors as defined in Section 4.2, that is, $\bar{P}(\hat{E}) = \frac{1}{N}\sum_{i=1}^{N} \hat{P}(\cup_{k=1}^{K} \hat{E}_{k,NT,i})$ and $\bar{P}(\hat{F}) = \frac{1}{N}\sum_{i=1}^{N} \hat{P}(\cup_{k=1}^{K} \hat{F}_{k,NT,i})$, where $\hat{P}$ denotes the empirical average probabilities across 200 replications. Table 2 reports the classification errors. The results with different $c_\lambda$'s are quite similar, indicating the robustness of our algorithm to the choice of tuning parameter. Moreover, the classification errors $\bar{P}(\hat{E})$ and $\bar{P}(\hat{F})$ are all below 3% for each scenario of the first two DGPs. The classification errors are a little bit large for the dynamic panel data models, but are still acceptable. In particular, all of them shrink toward zero quickly as $T$ increases.

For the estimation, Figure 2 depicts the three true group-specific trends and their post-Lasso estimates in DGP 2 for the case $N = 100, T = 40$ based on 200 replications. As shown in Figure 2, the fitted trends approximate the true trends pretty well, indicating the excellent behavior of our estimation procedure. To measure the accuracy of estimation for the group-specific functional coefficients, we define the weighted root-mean-square-error (RMSE) of the estimates $\hat{\alpha}_k(t/T)$'s in DGP 1 for each replication as follows:

$$\mathrm{RMSE}(\hat{\alpha}.) = \frac{1}{N}\sum_{k=1}^{3} N_k \mathrm{RMSE}(\hat{\alpha}_k),$$

where $\mathrm{RMSE}(\hat{\alpha}_k) = \{\frac{1}{T}\sum_{t=1}^{T}[\hat{\alpha}_k(t/T) - \alpha_k^0(t/T)]^2\}^{1/2}$ for $k = 1, 2, 3$. The weighted RMSEs of the estimates of $\alpha_{k,1}^0(t/T)$ and $\alpha_{k,2}^0(t/T)$ in DGP 2 are similarly defined. Table 3 reports the average of these RMSEs across 200 replications for both the C-Lasso and post-Lasso estimators for $c_\lambda = 1, 2, 4$, in comparison with the oracle estimators. As shown in Table 3, the RMSEs are quite similar for different choices of $c_\lambda$ and generally decline as $T$ increases for fixed $N$. The RMSEs of the post-Lasso estimators are less than those of the C-Lasso estimators in all cases and they are close to those of the oracle estimators when $T = 80$. This suggests that the post-Lasso estimators tend to outperform the C-Lasso estimators and would be recommended for practical use.

## 6.2 Empirical Illustration

As a key indicator of a country's standard of living, GDP per capita has been one of the most important variables in economics; see, for example, Solow (1956), Cass (1965), and Barro (1991, 1996). It not only provides a useful statistic for comparison of wealth across countries but also describes the development of a particular country. However, the exact realization of GDP per capita is not very useful in comparison due to the existence of the short term fluctuations. In fact, policy makers often target on long-lasting changes rather than short transitory fluctuations. This prompts us to extract the trend of GDP per capita, which can capture the medium-to long-term changes and have some implications on economic modeling, testing, and forecasting. For example, most of the existing literature assumes a linear

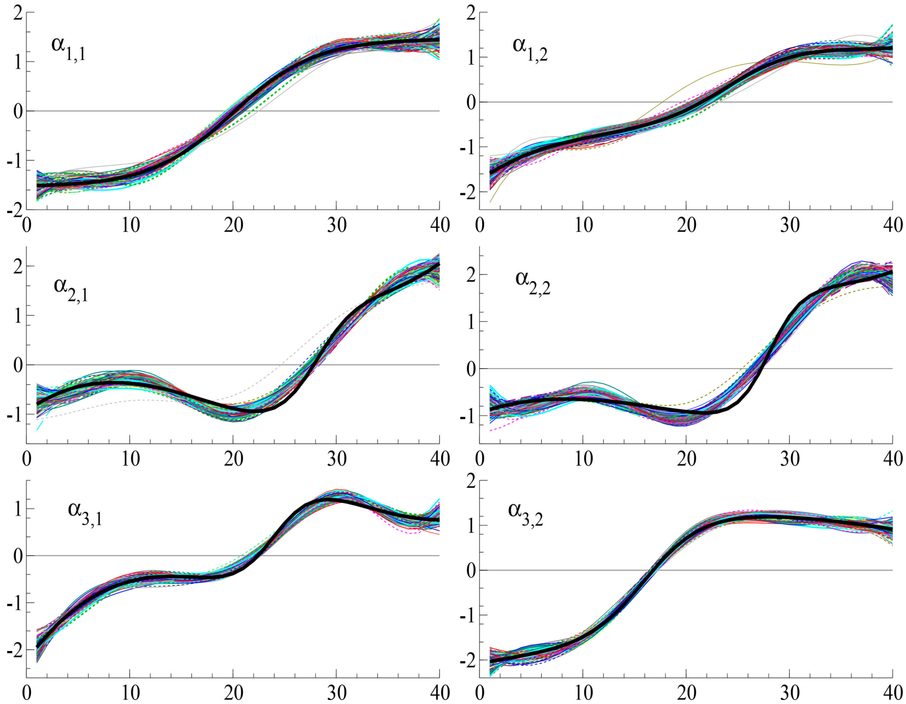


Figure 2. True trends (the heavy black line) and the post-Lasso estimators for DGP2 ($N = 100$, $T = 40$).

trend behavior for the GDP per capita when testing trend stationarity against unit root; see, for example, Fleissig and Strauss (1999) and Lluís et al. (2005). If the underlying trend is nonlinear in fact, then the conclusions can be misleading.

In this section, we use our time-varying panel data model with latent structure to estimate the heterogeneous trending behavior of GDP per capita across countries. In comparison with Robinson's (2012) nonparametric panel trend model, our model allows for unobserved cross-sectional heterogeneity. This is important as it is hard to believe the GDP per capita for all countries exhibit the same trend over time. Although macroeconomists have some consensus that globalization leads to the synchronization of business cycles across countries (Kose, Prasad, and Terrones 2003), it is unrealistic to assume all the countries share the same trend. In fact, a stream of empirical studies confirm the cross-country divergence rather than convergence implied by the neoclassical growth models (Barro 1991), and thus provide ample evidences on the cross-sectional heterogeneity. To account for the cross-sectional heterogeneity, applied researchers usually select a small group of countries

Table 3. Root mean squared errors of the C-Lasso and post-Lasso estimates

| | | | | | $c_\lambda = 1$ | | $c_\lambda = 2$ | | $c_\lambda = 4$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| DGP | coeff | $N$ | $T$ | oracle | C-Lasso | post-Lasso | C-Lasso | post-Lasso | C-Lasso | post-Lasso |
| 1 | $\alpha_k^0$ | 50 | 40 | 0.166 | 0.218 | 0.167 | 0.233 | 0.167 | 0.205 | 0.169 |
| | | 50 | 80 | 0.150 | 0.165 | 0.151 | 0.170 | 0.150 | 0.165 | 0.150 |
| | | 100 | 40 | 0.153 | 0.220 | 0.155 | 0.261 | 0.156 | 0.206 | 0.154 |
| | | 100 | 80 | 0.116 | 0.141 | 0.116 | 0.214 | 0.116 | 0.160 | 0.116 |
| 2 | $\alpha_{k,1}^0$ | 50 | 40 | 0.117 | 0.143 | 0.123 | 0.175 | 0.126 | 0.141 | 0.119 |
| | | 50 | 80 | 0.096 | 0.111 | 0.099 | 0.138 | 0.096 | 0.110 | 0.098 |
| | | 100 | 40 | 0.097 | 0.126 | 0.099 | 0.185 | 0.101 | 0.133 | 0.100 |
| | | 100 | 80 | 0.076 | 0.106 | 0.082 | 0.180 | 0.076 | 0.110 | 0.076 |
| | $\alpha_{k,2}^0$ | 50 | 40 | 0.120 | 0.144 | 0.126 | 0.165 | 0.128 | 0.141 | 0.123 |
| | | 50 | 80 | 0.096 | 0.112 | 0.101 | 0.127 | 0.096 | 0.110 | 0.099 |
| | | 100 | 40 | 0.097 | 0.122 | 0.099 | 0.185 | 0.102 | 0.127 | 0.099 |
| | | 100 | 80 | 0.065 | 0.096 | 0.072 | 0.183 | 0.065 | 0.099 | 0.065 |
| 3 | $\alpha_{k,3}^0$ | 50 | 40 | 0.297 | 0.418 | 0.397 | 0.472 | 0.453 | 0.432 | 0.403 |
| | | 50 | 80 | 0.201 | 0.323 | 0.292 | 0.334 | 0.304 | 0.318 | 0.301 |
| | | 100 | 40 | 0.210 | 0.394 | 0.383 | 0.397 | 0.375 | 0.408 | 0.399 |
| | | 100 | 80 | 0.146 | 0.307 | 0.271 | 0.311 | 0.283 | 0.302 | 0.284 |

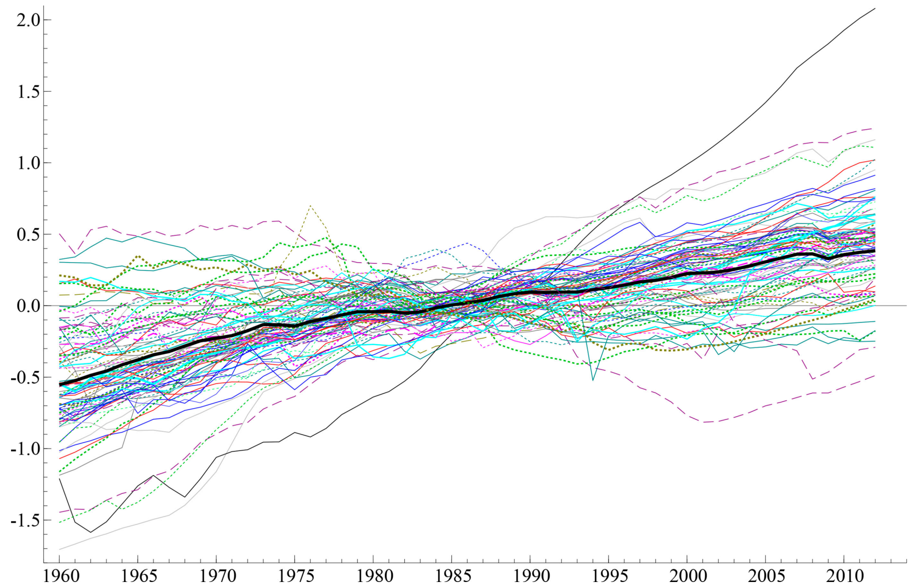


Figure 3. GDP per capita (logarithm and demean) for 91 countries between 1960 and 2012 (logarithm and demeaned, the thick black curve denotes the value for the whole world).

(e.g., OECD countries) that they think would share slope homogeneity, and then conduct statistical analysis for the selected countries. However, such a selection appears arbitrary which may further induce misleading results. As mentioned above, our model provides a data-driven classification before we embark on the estimation and inference procedure for the group-specific trending behavior. Hence, it is useful to extract the group-specific trends of GDP per capita across countries based on our new methodology.

*6.2.1 Data and Setting.* Denote the GDP per capita as $Y_{it}$. Then, we estimate the following trending panel structure model:

$$\log Y_{it} = \gamma_i + \beta_i(t/T) + u_{it}$$

using the annual data from 1960 to 2012 for as many countries as possible. We obtain the GDP per capita data from Federal Reserve Economic Data (FRED), measured in terms of 2005 U.S. dollars. By deleting countries with missing observations, we obtain a balanced panel that contains $N = 91$ countries and $T = 53$ observations for each country. As we are interested in the common trend across countries, we take logarithm for the data. By taking logarithm, the slope of the trend could be roughly interpreted as the growth rate of GDP per capita up to a scaling factor $T$. The data series are depicted in Figure 3. To show the path of the data more clearly, we report the demeaning data. It is obvious that the time path of GDP per capita exhibits noticeable heterogeneity. The whole world realization, which is marked by the thick black curve in the figure, is not a reasonable representative of the economic development.

We estimate the trending panel structure model in (2.5) by using the iterative algorithm introduced in the online Appendix C. To implement the penalized least-squares estimation with cubic B-spline approximation, we set the number of knots ($J_0$) and tuning parameter $\lambda$ as in the simulation section.

*6.2.2 Estimation Results.* To determine the appropriate number of groups, we choose $K$ to minimize the information criterion in Section 4.4. Table 4 reports the ICs for the number of group $K = 1, 2, \ldots, 6$ with different tuning parameters $c_\lambda = 0.5, 1, 2, 4$. The results show that the IC is robust to the tuning parameter and always achieves the minimum when $K = 4$. Figure 4 depicts the estimated trends for the four estimated groups and Figure 5 reports the realization of GDP per capita (logarithm and demeaned) and the trend for each group. To save space, we do not report the detailed estimation results here. A detailed report for the empirical results and discussions can be found in the online Appendix D.

## 6.3 Further Discussion on Potential Applications

The proposed time-varying panel data model with latent structures could capture the smooth structural changes over time and the individuals' heterogeneity across groups simultaneously. The model is flexible and hence is expected to have much broad applications in empirical study. As mentioned before, changes induced by policy switch, preference change, and technology progress can cause structural changes of the functional relationships between economic variables. Besides, individual units sampled from different backgrounds are expected to have

Table 4. The information criterion for different numbers of groups

| | $K$ | | | | | |
|---|---|---|---|---|---|---|
| $c_\lambda$ | 1 | 2 | 3 | 4 | 5 | 6 |
| 0.5 | − 2.584 | − 3.174 | − 3.344 | − **3.605** | − 3.234 | − 2.500 |
| 1 | − 2.584 | − 3.174 | − 3.417 | − **3.559** | − 2.922 | − 2.504 |
| 2 | − 2.584 | − 3.174 | − 3.375 | − **3.407** | − 3.237 | − 2.555 |
| 4 | − 2.584 | − 3.174 | − 3.344 | − **3.697** | − 3.409 | − 2.687 |

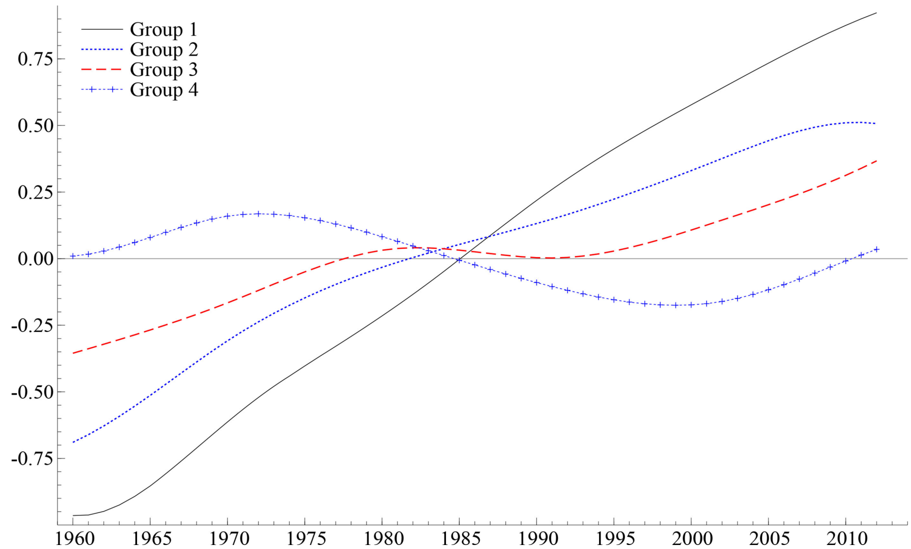


Figure 4. The estimated trends for the four estimated groups of economies.

heterogeneous features. To handle the individual heterogeneity, many empirical studies classify units to different groups based on some external criterion. For example, in macroeconomic studies involving countries, researchers often consider the OECD countries and the emerging economies separately. In microeconomic studies, individuals are usually classified into low income group, middle income group and high income group. By adopting our method, one does not need to classify units into different groups a priori. Our method could identify individuals' membership endogenously. Here, we discuss two potential applications.

The first example is the energy intensity. Energy intensity is a measure of the energy efficiency of a nation's economy. It is calculated as units of energy per unit of GDP. High-energy intensities indicate a high price or cost of converting energy into GDP. The trend of the energy intensity reveals the changes of the economic energy efficiency. Due to the different stages of economic development that different countries attain, the trend of energy intensity varies across countries. Hence, we can consider the following trending panel structure model to estimate the trend of energy intensity for various countries:

$$y_{it} = \gamma_i + \beta_i(t/T) + u_{it},$$

where $\gamma_i$ is the individual effect and $\beta_i(t/T)$ satisfies the latent group structure in our article.

The second example is the beneficial effects of foreign direct investment (FDI) on economic growth in host countries over a long period of time. As mentioned before, the relationships

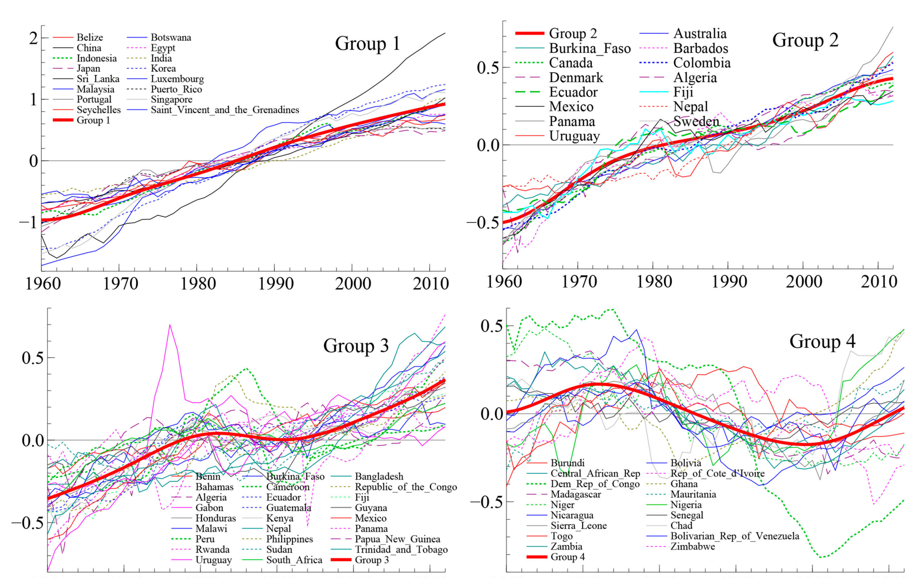


Figure 5. The GDP per capita (logarithm and demean) for countries in each group and the estimated group-specific trend (thick solid curve).

between variables tend to change during a long period. In addition, due to the difference of absorptive capacities in different host countries, the FDI effects tend to be heterogeneous. To capture the time-varying relationship and the cross-country heterogeneous absorptive capacities simultaneously, we consider the following model:

$$y_{it} = \alpha_i + \beta_{it}^{(1)}(\mathrm{FDI}/Y)_{it} + \beta_t^{(2)}\log(\mathrm{DI}/Y)_{it} + \beta_t^{(3)}n_{it} + \beta_t^{(4)}h_{it} + \beta_t^{(5)}((\mathrm{FDI}/Y)_{it} \times h_{it}) + u_{it},$$

where $y_{it}$ denotes the growth rate of GDP per capita in country/region $i$ during the period $t$, $n_{it}$ is the logarithm of population growth rate, $h_{it}$ is the human capital, and $\alpha_i$ is the individual effect used to control the unobserved country-specific heterogeneity. Here, FDI and DI refer to foreign direct investment and domestic investment, respectively; $Y$ represents the total output. Hence, $(\mathrm{FDI}/Y)_{it}$ denotes the average ratio between the FDI and the total output during the period $t$ in country/region $i$, and $(\mathrm{DI}/Y)_{it}$ is defined in the same fashion for the domestic investment. In the model, $\beta_{it}^{(1)}$ exhibits the latent group structure in (2.2), and $\beta_t^{(j)} = \beta^{(j)}(t/T)$, $j = 2, 3, \ldots, 5$, are homogeneous functional coefficients. This model is the mixed time-varying panel structure model that can be estimated by using the technique given in Section 5.1. Alternatively, we can also allow $\beta_t^{(j)}$, $j = 2, 3, \ldots, 5$, to be heterogeneous and have the latent group structure. In either case, the model extends the typical empirical growth equation

$$y_{it} = \alpha_i + \beta^{(1)}(\mathrm{FDI}/Y)_{it} + \beta^{(2)}\log(\mathrm{DI}/Y)_{it} + \beta^{(3)}n_{it} + \beta^{(4)}h_{it} + \beta^{(5)}((\mathrm{FDI}/Y)_{it} \times h_{it}) + u_{it}.$$

See Kottaridi and Stengos (2010) and Cai, Chen and Fang (2014), and references therein.

## 7. CONCLUSION

In this article, we propose a time-varying panel data model with latent group structures to capture individual heterogeneity and smooth structural changes over time simultaneously. We focus on the penalized sieve estimation (PSE) of such a model, where the penalty term is constructed to achieve simultaneous classification and estimation. The PSE achieves the uniform classification consistency and oracle property. We also propose a BIC-type information criterion to determine the unknown number of groups. Simulations are conducted to evaluate the finite sample performance of the proposed information criterion and PSE method. We apply our method to study the heterogeneous trending behavior of GDP per capita across 91 countries for the period 1960–2012 and find four latent groups.

Several extensions are possible. First, one can consider general functional coefficient panel data models with latent group structures where the coefficients are functions of certain random covariates. More generally, one can consider other types of nonparametric or semiparametric panel data models (e.g., the partially linear single-index panel data model of Chen, Gao, and Li 2013) with latent group structures. Second, as discussed in Section 5.3 we can also allow for cross-sectional dependence in our model. But the asymptotic theory is extremely involved and we leave it for future research.

## SUPPLEMENTARY MATERIALS

The online supplementary Appendix presents the proofs of the technical lemmas and numerical algorithm in the article.

## ACKNOWLEDGMENTS

The authors thank a co-editor, an associate editor, and three anonymous referees for many constructive comments on the early version of the article. Su acknowledges financial support from the Singapore Ministry of Education for Academic Research Fund (AcRF) under the Tier 2 Grant MOE2012-T2-2-021 and funding support provided by the Lee Kong Chian Fund for Excellence. Wang acknowledges financial supports from the National Science Foundation of China (No. 71401160, No. 71371160), the Ministry of Education of Humanities and Social Sciences Project of China (No. 14YJC790120), and Fujian Provincial Key Laboratory of Statistics (Xiamen University, No. 2016003).

*[Received April 2015. Revised May 2017.]*

## REFERENCES

Ando, T., and Bai, J. (2016), "Panel Data Models with Grouped Factor Structure under Unknown Group Membership," *Journal of Applied Econometrics*, 31, 163–191. [337,338]

Andrews, D. W. K. (1984), "Non-Strong Mixing Autoregressive Processes," *Journal of Applied Probability*, 21, 930–934. [338]

Atak, A., Linton, O., and Xiao, Z. (2011), "A Semiparametric Panel Model for Unbalanced Data with Application to Climate Change in the United Kingdom," *Journal of Econometrics*, 164, 92–115. [342]

Bai, J. (2010), "Common Breaks in Means and Variances for Panel Data," *Journal of Econometrics*, 157, 78–92. [334]

Bai, J., and Lluís Carrion-I-Silvestre, J. (2009), "Structural Changes, Common Stochastic Trends, and Unit Roots in Panel Data," *Review of Economic Studies*, 76, 471–501. [334]

Bai, J., and Ng, S. (2002), "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70, 191–221. [342]

Barro, R. J. (1991), "Economic Growth in A Cross Section of Countries," *Quarterly Journal of Economics*, 106, 407–443. [344,345]

——— (1996), *Determinants of Economic Growth: A Cross-Country Empirical Study*, Cambridge: The MIT Press. [344]

Bondell, H. D., and Reich, B. J. (2008), "Simultaneous Regression Shrinkage, Variable Selection, and Supervised Clustering of Predictors with OSCAR," *Biometrics*, 64, 115–123. [345]

Bonhomme, S., and Manresa, E. (2015), "Grouped Patterns of Heterogeneity in Panel Data," *Econometrica*, 83, 1147–1184. [336,337,338,340]

Cai, Z. (2007), "Trending Time-Varying Coefficient Time Series Models with Serially Correlated Errors," *Journal of Econometrics*, 136, 163–188. [334,335,336]

Cai, Z., Chen, L., and Fang, Y. (2014), "Quantile Panel Data Models with Partially Varying Coefficients: Estimating the Growth Effect of FDI," Working Paper, Xiamen University. [336,348]

Cai, Z., Fan, J., and Yao, Q. (2000), "Functional-Coefficient Regression Models for Nonlinear Time Series," *Journal of the American Statistical Association*, 95, 941–956. [334]

Cass, D. (1965), "Optimum Growth in An Aggregative Model of Capital Accumulation," *Review of Economic Studies*, 32, 233–240. [344]

Chen, J., Gao, J., and Li, D. (2012), "Semiparametric Trending Panel Data Models with Cross-Sectional Dependence," *Journal of Econometrics*, 171, 71–85. [335,336]

Chen, J., Gao, J., and Li, D. (2013), "Estimation in Partially Linear Single-Index Panel Data Models With Fixed Effects," *Journal of Business & Economic Statistics*, 31, 315–330. [348]

Chen, B., and Huang, L. (2014), "Nonparametric Testing for Smooth Structural Changes in Panel Data Models," Working Paper, Department of Economics, University of Rochester. [334]

Chen, X. (2007), "Large Sample Sieve Estimation of Semi-Nonparametric Models," in *Handbook of Econometrics*, eds. J. Heckman and E. Leamer, (Vol. 6B), New York: Elsevier, pp. 5549–5632. [337]

de Boor, C. (2001), *A Practical Guide to Splines*, New York: Springer-Verlag. [337,338]

DeVore, R. A., and Lorentz, G. G. (1993), *Constructive Approximation*, New York: Springer-Verlag. [337]

Diggle, P. J., Heagerty, P., Liang, K.-Y., and Zeger, S. L. (2003), *Analysis of Longitudinal Data*, Oxford: Oxford University Press. [335]

Durlauf, S., Kourtellos, A., and Minkin, A. (2001), "The Local Solow Growth Model," *European Economic Review*, 45, 928–940. [334]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle properties," *Journal of the American Statistical Association*, 96, 1348–1360. [335]

Fleissig, A. R., and Strauss, J. (1999), "Is OECD Real Per Capita GDP Trend or Difference Stationary? Evidence from Panel Unit Root Tests," *Journal of Macroeconomics*, 21, 673–690. [345]

Hahn, J., and Kuersteiner, G. (2011), "Bias Reduction for Dynamic Nonlinear Panel Models with Fixed Effects," *Econometric Theory*, 27, 1152–1191. [338]

Han, A. K., and Park, D. (1989), "Testing for Structural Change in Panel Data: Application to Study of U.S. Foreign Trade in Manufacturing Goods," *Review of Economics and Statistics*, 71, 135–142. [334]

Hsiao, C. (2014), *Analysis of Panel Data*, Cambridge: Cambridge University Press. [335]

Hsiao, C., and Pesaran, M. (2008), "Random Coefficient Models," in *The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory and Practice*, eds. L. Matyas and P. Sevestre, (3rd ed.), Berlin: Springer-Verlag, pp. 187–216. [335]

Huang, J. Z., and Shen, H. (2004), "Functional Coefficient Regression Models for Non-linear Time Series: A Polynomial Spline Approach," *Scandinavian Journal of Statistics*, 31, 515–534. [335]

Huang, J. Z., Wu, C. O., and Zhou, L. (2004), "Polynomial Spline Estimation and Inference for Varying Coefficient Models with Longitudinal Data," *Statistica Sinica*, 14, 763–788. [335,340]

Juárez, M. A., and Steel, M. F. J. (2010), "Model-Based Clustering of Non-Gaussian Panel Data Based on Skew-*t* Distributions," *Journal of Business & Economic Statistics*, 28, 52–66. [334]

Kottaridi, C., and Stengos, T. (2010), "Foreign Direct Investment, Human Capital and Nonlinearities in Economic Growth," *Journal of Macroeconomics*, 32, 858–871. [348]

Ke, Z., Fan, J., and Wu, Y. (2015), "Homogeneity Pursuit," *Journal of American Statistical Association*, 110, 175–194. [335]

Kim, D. (2011), "Estimating a Common Deterministic Time Trend Break in Large Panels with Cross-Sectional Dependence," *Journal of Econometrics*, 164, 310–330. [334]

Kose, M. A., Prasad, E. S., and Terrones, M. E. (2003), "How Does Globalization Affect the Synchronization of Business Cycles?," *American Economic Review*, 93, 57–62. [345]

Lee, K., Pesaran, M., and Smith, R. (1997), "Growth and Convergence in a Multi-Country Empirical Stochastic Solow Model," *Journal of Applied Econometrics*, 12, 357–392. [334]

Li, D., Chen, J., and Gao, J. (2011), "Nonparametric Time-varying Coefficient Panel Data Models with Fixed Effects," *Econometrics Journal*, 14, 387–408. [334]

Li, D., Qian, J., and Su, L. (2016), "Panel Data Models with Interactive Fixed Effects and Multiple Structural Breaks," *Journal of the American Statistical Association*, 111, 1804–1819. [334]

Lin, C.-C., and Ng, S. (2012), "Estimation of Panel Data Models with Parameter Heterogeneity When Group Membership Is Unknown," *Journal of Econometric Methods*, 1, 42–55. [336,337]

Lluís Carrion-I-Silvestre, J., Barrio-Castro, D., and López-Bazo, E. (2005), "Breaking the Panels: An Application to the GDP Per Capita," *Econometrics Journal*, 8, 159–175. [345]

Moon, H. R., and Weidner, M. (2015), "Linear Regression for Panel with Unknown Number of Factors as Interactive Fixed Effects," *Econometrica*, 83, 1543–1579. [342]

Qian, J., and Su, L. (2016), "Shrinkage Estimation of Common Breaks in Panel Data Models via Adaptive Group Fused Lasso," *Journal of Econometrics*, 191, 86–109. [334]

Robinson, P. M. (1989), "Nonparametric Estimation of Time-Varying Parameters," in *Statistical Analysis and Forecasting of Economic Structural Change*, ed. P. Hackl, Berlin: Springer, pp. 253–264. [336]

——— (1991), "Time-Varying Nonlinear Regression," in *Economic Structure Change Analysis and Forecasting*, eds. P. Hackl and A. H. Westlund, Berlin: Springer, pp. 179–190. [336]

——— (2012), "Nonparametric Trending Regression with Cross-sectional Dependence," *Journal of Econometrics*, 169, 4–14. [334,336,345]

Schumaker, L. (1981),*Spline Functions: Basic Theory*, New York: Wiley. [337]

Solow, R. M. (1956), "A Contribution to the Theory of Economic Growth," *Quarterly Journal of Economics*, 70, 65–94. [344]

Shen, X., and Huang, H-C. (2010), "Group Pursuit Through a Regularization Solution Surface," *Journal of American Statistical Association*, 105, 727–739. [335]

Su, L., and Chen, Q. (2013), "Testing Homogeneity in Panel Data Models with Interactive Fixed Effects," *Econometric Theory*, 29, 1079–1135. [334,338]

Su, L., and Jin, S. (2012), "Sieve Estimation of Panel Data Models with Cross Section Dependence," *Journal of Econometrics*, 169, 34–47. [341]

Su, L., and Ju, G. (2017), "Identifying Latent Grouped Effects in Panel Data Models with Interactive Fixed Effects," *Journal of Econometrics*, forthcoming. [342,343]

Su, L., Shi, Z., and Phillips, P. C. B. (2016), "Identifying Latent Structures in Panel Data," *Econometrica*, 84, 2215–2264. [335]

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005), "Sparsity and Smoothness via the Fused Lasso," *Journal of the Royal Statistical Society*, Series B, 67, 91–108. [335]

Wang, W., Phillips, P. C. B., and Su, L. (2017), "Homogeneity Pursuit in Panel Data Models: Theory and Applications," Cowles Foundation Discussion Paper No. 2063, Yale University. [335]

Xue, L., and Yang, L. (2006), "Additive Coefficient Modeling via Polynomial Spline," *Statistica Sinica*, 16, 1423–1446. [335]

Zhang, Y., Su, L., and Phillips, P. C. B. (2011), "Testing for Common Trends in Semi-parametric Panel Data Models with Fixed Effects," *Econometrics Journal*, 15, 56–100. [336]

Zhu, Y., Shen, X., and Pan, W. (2013), "Simultaneous Grouping Pursuit and Feature Selection over an Undirected Graph," *Journal of American Statistical Association*, 108, 713–725. [335]