Estimation and Identification of Latent Group Structures in Panel Data *

Ali Mehrabani†

College of Business and Analytics, Southern Illinois University, Carbondale, IL, U.S.A.

December 4, 2022

Abstract

This paper provides a framework for joint estimation and identification of latent group structures in panel data models using a pairwise fusion penalized approach. The latent structure of the model allows individuals to be classified into different groups where the number of groups and the group membership are unknown. The individuals within a group have common slope parameters, while parameter heterogeneity is allowed across the groups. A penalized least squares (PLS) approach is introduced for models with exogenous regressors. When the model contains endogenous regressors, a penalized generalized method of moment (PGMM) is introduced. To implement the proposed approach, an alternating direction method of multipliers algorithm has been developed. The proposed method is further illustrated by simulation studies which demonstrate the finite sample performance of the method, and is applied in an empirical analysis.

Key Words: *ADMM algorithm, classification, dynamic panel, high dimensionality, oracle property, pairwise adaptive group fused Lasso, parameter heterogeneity.*

JEL Classification: C33, C36, C38, C51.

*I gratefully thank the Co-editor Serena Ng, an Associate Editor, and two anonymous referees for their constructive comments on the paper. I also thank Gloria Gonzalez-Rivera, Tae-Hwy Lee, Shujie Ma, Esfandiar Maasoumi, Shahnaz Parsaeian, M. Hashem Pesaran, and Aman Ullah for discussions on the subject matter and comments on the paper.
†Address correspondence to: ali.mehrabani@siu.edu.

1 Introduction

Panel data offer great opportunities in empirical research. Nevertheless, in practice, they typically involve aggregate data from various units (such as workers, firms, countries) that are different in some unobservable aspects to researchers. Accordingly, the researchers face a trade-off between using flexible methods to model the unobservable heterogeneity, and using pooled models that avoid the heterogeneity by assuming to some extent homogeneous coefficients for all individual units. To overcome this challenge, recently, latent group structures in panel data literature have received considerable attention. The most important advantage of the latent group structure is that unlike completely heterogenous or fully homogenous models, it allows panel units to be classified into groups, where the individuals within a group share the same slope parameters, while heterogeneity exists across the groups. Inspired by the literature, this paper introduces a simple and fast method to jointly identify and estimate latent group structures in panel data models when the number of groups and the individuals' group identities are both unknown.

A common approach to model heterogeneity in econometric analysis is to assume complete slope heterogeneity. This assumption avoids misspecification, but does not gain from working with panel data, and could result in imprecise estimates even if the time dimension is long (see, Baltagi and Griffin (1997)). Nonetheless, conventional panel data models often avoid the heterogeneity and assume the regression parameters are the same across individuals, and the unobserved heterogeneity is modeled through individual-specific effects (fixed effect and random effect models). This assumption exploits cross-section averaging and causes higher efficiency, but at the cost of estimation bias and inconsistency, which is supported by an increasing number of studies due to a better forecast performance of the associated estimators (see for example, Baltagi et al. (1989), Maddala (1991), Maddala and Hu (1996), Baltagi and Griffin (1997), and Hoogstrate et al. (2000)). In spite of a better forecast performance, it is often difficult to justify the slope homogeneity assumption in the empirical work, as pointed out by Hsiao and Tahmiscioglu (1997), Phillips and Sul (2007), Browning and Carro (2007), and Su and Chen (2013). This discussion motivated much of the recent research on the latent group structures in panel data analysis including Sun (2005), Lin and Ng (2012), Deb and Trivedi (2013), Bonhomme and Manresa (2015), Sarafidis and Weber (2015), Ando and Bai (2016), Bester and Hansen (2016), Su et al. (2016), Lu and Su (2017), Su and Ju (2018), Wang et al. (2018), Su et al. (2019), Gu and Volgushev (2019), Liu et al. (2020), and Wang and Su (2021), among others. Moreover, the group structure has sound foundations

in game theory or macroeconomic models where multiplicity of Nash equilibria is expected (Hahn and Moon (2010)). The latent group structure models partition individuals in different groups and allow the within group individuals share common coefficients, while the groups are assumed to have slope heterogeneity. Since the group membership and the number of groups are unknown in these models, the determination of the true number of groups and each individual's group identity are the key questions. Several approaches have been proposed to address these questions. Sun (2005), Kasahara and Shimotsu (2009), and Browning and Carro (2007) consider finite mixture models. Su et al. (2016) develop a new variant of the Lasso (least absolute shrinkage and selection operator) procedure, called classifier-Lasso (C-Lasso), to achieve classification in panel structure models where the penalty takes an additive-multiplicative form. The C-Lasso method of Su et al. (2016) has been extended to allow for two-way component errors, interactive fixed effects, non-stationary regressors, and semi-parametric specification, respectively, in Lu and Su (2017), Su and Ju (2018), Huang et al. (2020), and Su et al. (2019). Lin and Ng (2012) and Sarafidis and Weber (2015) extend the K-means algorithm to the panel regression framework with latent group structures, but the asymptotic properties of the estimators and the procedures are not provided. Bonhomme and Manresa (2015) and Ando and Bai (2016) modify the K-means algorithm to estimate the time-varying grouped patterns of heterogeneity and unobserved group interactive fixed effects, respectively. Wang et al. (2018) extend the CARDS (clustering algorithm in regression via data-driven segmentation) method of Ke et al. (2015) to panel structure models where the latent group structures exist in vectors of slope parameters. Recently, Liu et al. (2020) extend the modified K-means algorithm of Bonhomme and Manresa (2015) to estimate and identify the latent group structures in panel data. Wang and Su (2021) extend the sequential binary segmentation algorithm (SBSA) of Bai (1997) for break detection from the time series setup to the panel data framework to identify the latent group structures.

These methods make important contributions by empirically estimating the group identities. However, to implement them, one often needs to determine the number of groups first. Consequently, the estimation error often accumulates across the two steps and leads to suboptimal performance. The objective of this paper is to provide a new framework to jointly estimate and identify the latent group structures without a priori knowledge of classification or a natural basis for separating slope coefficients into groups.

Inspired by the adaptive group fused Lasso of Qian and Su (2016), and the pairwise fusion concave penalty of Ma and Huang (2017), we propose a penalized procedure with a pairwise fusion penalty to automatically estimate slope parameters and identify group identities where both the number of groups and the individual group identities are unknown. Our method and mainly our model is different from theirs in several important aspects. Qian and Su (2016) consider estimation and inference of common structural breaks in panel data models using an adaptive group fused Lasso. Their method cannot be used to classify individuals into different groups because there is no natural ordering across individuals, also a different algorithm to locate common individuals is required. Ma and Huang (2017) consider the problem of identifying subgroups among observations, using a concave pairwise fusion penalty. Clearly, their model is different from the model considered here to estimate and identify the latent group structures. Besides, the penalty term in Ma and Huang (2017) is imposed through concave penalties such as the SCAD (smoothly clipped absolute deviations penalty) of Fan and Li (2001) and the MCP (minimax concave penalty) of Zhang (2010), but our penalty is imposed through an adaptive group fused Lasso. The other main differences of our method from theirs lies in three aspects: 1) we impose the penalty on slope vector differences, whereas their method applies the penalty on the intercepts, 2) we consider both penalized least squares and penalized generalized method of moments estimations and show their asymptotic properties, while Ma and Huang (2017) only consider penalized least squares, 3) we assign different weights $\{\dot{w}_{ij}\}$, based on preliminary estimates of the slope parameters to penalize different coefficient differences, however these weights are not feasible in their study. Ma and Huang (2017) use concave penalties because these penalties provide the unbiasedness property. They argue that the Lasso penalty generates large biases. This is due the fact that the Lasso penalty is not adaptive for discriminating large from small differences. As a result, over-penalizing large differences due to shrinking small differences towards zero prevents its consistency property. However, our panel regression allows us to use the adaptive group Lasso penalty that assigns different weights to penalize the pairwise differences and avoids this shortcoming of the Lasso. Since our proposed framework utilizes a pairwise adaptive group fused Lasso penalty, we denote our estimation procedure as PAGFL. To implement our method, we derive an ADMM (alternating direction method of multipliers) algorithm (Boyd et al. (2011)), and show the convergence properties of our ADMM algorithm. It is worth mentioning that, the ADMM has good convergence properties for convex loss functions with the L_p, $p \geq 1$, penalties (see Boyd et al. (2011) and Chi and Lang

(2015)). Thus, not only our adaptive Lasso penalty has similar properties to the concave penalties, it enjoys the convexity and hence computational expedience.

We develop two classes of estimators for panel structure models to estimate the slope parameters: penalized least squares (PLS) and penalized generalized method of moments (PGMM). The PLS can be applied to static or dynamic panel models without endogenous regressors, while the PGMM is suitable for panel models with endogeneity or dynamic structures. We show that the PLS method is an oracle procedure (using the language of Fan and Li (2001)), in the sense that the PLS estimator classifies the right individuals in the right groups (classification consistency), and asymptotically is equivalent to the oracle estimator. The oracle estimator is obtained from least squares regression by assuming that the true group structure is known. Similarly, our PGMM estimator satisfies the classification consistency, but its oracle property does not hold generally. Our asymptotic results hold under $(N, T) \to \infty$ jointly, where T is the time series dimension, and N is the cross-section dimension. The major contribution of our method compared to the existing methods in the literature is that it asymptotically identifies the true structure while estimating the model parameters consistently without relying on correct initial estimates of the number of groups. This implies that our estimation and classification consistency results hold without requiring a priori correct estimation or knowledge of the number of groups. This is of crucial importance as in most empirical research the number of groups is often unknown to practitioners. Furthermore, our proposed approach allows the number of groups and the number of individuals within each group to be either divergent or fixed, which makes our method applicable to a large body of applications.

In comparison with the C-Lasso of Su et al. (2016), the K-means algorithm, the CARDS algorithm of Wang et al. (2018), and the SBSA of Wang and Su (2021), our method has both pros and cons. First, the C-Lasso procedure of Su et al. (2016) is not a convex problem[1], and the K-means algorithm has been shown to be NP-hard, and can get trapped in suboptimal local minima. Unlike the K-means algorithm and the C-Lasso method, our PAGFL approach admits a simple and fast iterative algorithm that is guaranteed to converge to the unique global minimizer. Therefore, the computation burden of our approach is not as much as the K-means algorithm and the C-Lasso. Our penalty term contains $\binom{N}{2}$ pairwise differences of p slope coefficients, which includes several redundant constraints, and can impose computational challenges when $N \times p$ is very large.[2] Second,

[1] However, the numerical solution can be transformed into a sequence of convex problems.

[2] We experimented different simulation studies, and faced the computation challenge in implementing ADMM when $N \times p$ was more than 60,000.

the C-Lasso needs to specify two tuning parameters, one for determining the number of groups, and the other one for the penalty term. The CARDS method needs the choice of three tuning parameters, one is used to control the number of segments, the other two are used for the between-, and within-segment penalties. Unlike the C-Lasso, and CARDS methods, but like the K-means algorithm and the SBSA method, our PAGFL approach relies on the choice of one tuning parameter. We propose and validate a BIC-type information criteria to determine the tuning parameter. The K-means algorithm and the SBSA method require the tuning parameter to determine the number of groups, while we need the tuning parameter in our penalty term. Third, the SBSA of Wang and Su (2021) and the CARDS method of Wang et al. (2018) rely on ordered segmentations to identify the latent group structure and construct the Lasso-type penalties, respectively, which are sensitive to the choice of initial estimators, and often it may be difficult to construct one. The K-means algorithm is also sensitive to the choice of initial estimators, as discussed in Bonhomme and Manresa (2015). Unlike, the SBSA and CARDS methods, but like the C-Lasso and the K-means algorithm our PAGFL method does not rely on order segmentation. However, our method requires initial consistent estimators to produce the adaptive weights in the penalty term. Fourth, the C-Lasso may leave some individuals unclassified. Su et al. (2016) recommend to classify these unclassified individuals to one of the existing groups with closest distance. However, our method is not subject to this issue because it does not require the knowledge of the number of groups.

The reminder of this paper is organized as follows. Section 2 describes our fixed effect panel model, the PLS and PGMM estimation methods depending on whether the regressors are endogenous. Sections 3 and 4 analyze the asymptotic properties of the PLS and PGMM estimators, respectively. Section 5 presents the computation and algorithm. Monte Carlo results are given in section 6. In Section 7, we apply our estimators to a simple model of inter-temporal dynamics of the unemployment rate in the U.S. states. Conclusions and final remarks are given is section 8. All proofs and detailed calculations are provided in the Online Supplemental Appendix.

A brief word on notation: For an $m \times n$ real matrix A, we write the transpose A', the Frobenius norm as $\|A\| = (\text{tr}(AA'))^{1/2}$, and its spectral norm as $\|A\|_{sp}$. When A is symmetric, we use $\mu_{max}(A)$ and $\mu_{min}(A)$ to denote the largest and smallest eigenvalues, respectively. I_p and $\mathbf{0}_{p\times 1}$ denote $p \times p$ identity matrix and $p \times 1$ vector of zeros. $\mathbf{1}(\cdot)$ denotes the indicator function, "p.d." and "p.s.d." abbreviate "positive definite" and "positive semi-definite", respectively. The operators

$\xrightarrow{p}$, $\xrightarrow{D}$, and *plim* denote respectively, convergence in probability, convergence in distribution, and probability limit. We use $(N,T) \to \infty$ to signify that N and T pass jointly to infinity.

2 Model and Penalized Estimation

In this section, we consider a linear panel structure model with an unknown number of groups, and group membership.

2.1 The Model

Consider the following linear panel data model

$$y_{it} = \beta_i^{0\prime} x_{it} + \eta_i + u_{it}, \qquad i = 1, \ldots, N, \quad t = 1, \ldots, T, \tag{2.1}$$

where y_{it} is the dependent variable, x_{it} is a $p \times 1$ vector of regressors explaining y_{it}, η_i is the individual fixed effect that may be correlated with the regressors, u_{it} is the idiosyncratic error term with zero mean, T is the number of observations, and N is the number of individual units. We assume that β_i^0 is a $p \times 1$ vector of slope parameters that admits a possible grouping structure of the form

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

where $\alpha_l^0 \neq \alpha_k^0$ for any $l, k = 1, \ldots, K_0$, with $l \neq k$, $G_l^0 \cap G_k^0 = \emptyset$, and $\mathcal{G}_{K_0}^0 = \{G_1^0, G_2^0, \ldots, G_{K_0}^0\}$ forms a partition of $\{1, 2, \ldots, N\}$. Let N_k be the number of individual units in G_k^0, and the $pK_0 \times 1$ matrix of $\boldsymbol{\alpha}_{K_0}$, and the $pN \times 1$ matrix $\boldsymbol{\beta}$ be defined as

$$\boldsymbol{\alpha}_{K_0} = (\alpha_1', \alpha_2', \ldots, \alpha_{K_0}')' \quad \text{and} \quad \boldsymbol{\beta} = (\beta_1', \beta_2', \ldots, \beta_N')', \tag{2.3}$$

and let $\boldsymbol{\alpha}_{K_0}^0$ and $\boldsymbol{\beta}^0$ denote the true values of $\boldsymbol{\alpha}_{K_0}$ and $\boldsymbol{\beta}$. In practice, the number of groups, K_0, is unknown. However, it is usually reasonable to assume that K_0 is smaller than N. Our goal is

to estimate the regression coefficients $\boldsymbol{\alpha}_{K_0}^0$, the number of groups K_0, and identify the latent group structure.

We consider two cases about the exogeneity or endogeneity of the regressors:

(a) $\mathbb{E}(x_{is}u_{it}) = 0$, for all $1 \leq s \leq t \leq T$;

(b) $\mathbb{E}(x_{it}u_{it}) \neq 0$, for $t = 1, \ldots, T$.

The first case occurs when the regressors are weakly exogenous which allows for lagged values of y_{it} to be included in x_{it}, so that least squares criteria are appropriate. The second case happens when the regressors contain either lagged dependent variables or endogenous regressors that are correlated with the error term. In this case, we assume there exists a $q \times 1$ vector of instruments z_{it} with $q \geq p$.

Since the individual effects, η_i, are not of main interest, in case (a), we concentrate them out and obtain the following equation

$$\tilde{y}_{it} = \beta_i^{0\prime}\tilde{x}_{it} + \tilde{u}_{it}, \qquad i = 1, \ldots, N, \quad t = 1, \ldots, T, \tag{2.4}$$

where, e.g., $\tilde{x}_{it} = x_{it} - T^{-1}\sum_{t=1}^{T} x_{it}$. In case (b), to eliminate the effect of μ_i in the estimation procedure, we consider the first-differenced equation

$$\Delta y_{it} = \beta_i^{0\prime}\Delta x_{it} + \Delta u_{it}, \tag{2.5}$$

where, e.g., $\Delta y_{it} = y_{it} - y_{i,t-1}$ for $i = 1, \ldots, N$, and $t = 1, ..., T$, by assuming that we have observations on y_{i0} and x_{i0}.

2.2 Penalized Least Squares (PLS) Estimation

To estimate the model in (2.4) under case (a), we propose minimizing the following objective function

$$Q_{1,NT}(\boldsymbol{\beta}, \lambda_1) = \frac{1}{T}\sum_{i=1}^{N}\sum_{t=1}^{T}(\tilde{y}_{it} - \beta_i'\tilde{x}_{it})^2 + \frac{\lambda_1}{N}\sum_{1\leq i<j\leq N}\sum \dot{w}_{ij}\|\beta_i - \beta_j\|, \tag{2.6}$$

where $\lambda_1 \geq 0$ is a tuning parameter, and $\dot{w}_{ij}$ is a data-driven weight defined by

$$\dot{w}_{ij} = \|\dot{\beta}_i - \dot{\beta}_j\|^{-\kappa}, \quad \text{for } i, j = 1, \ldots, N, \tag{2.7}$$

where $\dot{\beta}_i$ and $\dot{\beta}_j$ are preliminary consistent estimates of β_i and β_j, respectively, and κ is a user-specified positive constant that usually takes value 2 in the literature of adaptive Lasso.

To obtain the adaptive weights $\{\dot{w}_{ij} : i, j \in \{1, \ldots, N\}\}$, we propose to obtain the preliminary estimates $\dot{\boldsymbol{\beta}} = (\dot{\beta}_1', \ldots, \dot{\beta}_N')'$, by minimizing the first term in equation (2.6) which results in the ordinary least squares. Thus for the i-th element of $\dot{\boldsymbol{\beta}}$, we have

$$\dot{\beta}_i = \Big(\sum_{t=1}^{T} \tilde{x}_{it} \tilde{x}_{it}' \Big)^{-1} \sum_{t=1}^{T} \tilde{x}_{it} \tilde{y}_{it}. \tag{2.8}$$

The objective function in (2.6) is related to the literature on adaptive Lasso (Zou (2006)), group Lasso (Yuan and Lin (2006)), fused Lasso (Tibshirani et al. (2005)) and group fused Lasso (Qian and Su (2015)), however, they are not applicable here. Qian and Su (2015) determine the unknown number of structural breaks in time series regression framework which is different from the purpose of this paper. The other listed papers above aim at determining the nonzero coefficients from the zero ones, and are not applicable here because our aim is to determine the unknown group structure.

It is worth emphasizing that the minimization of (2.6) is a convex optimization problem, thus it does not suffer from multiple local minima issue, and its global minimizer $\widehat{\boldsymbol{\beta}} = \arg\min Q_{1,NT}(\boldsymbol{\beta}, \lambda_1)$, can be efficiently solved. We suppress the dependence of $\widehat{\boldsymbol{\beta}} \equiv \widehat{\boldsymbol{\beta}}(\lambda_1)$ on λ_1 unless necessary, and choose the tuning parameter using a data-driven method proposed in Section 3.4.

The penalty in (2.6) shrinks some of the pairs $\beta_i - \beta_j$ to zero, thus we can partition the slope parameters into groups. In practice, let $\{\hat{\alpha}_1, \ldots, \hat{\alpha}_{\widehat{K}}\}$ be the distinct values of $\widehat{\boldsymbol{\beta}}$. Then, we define $\widehat{\mathcal{G}}_{\widehat{K}} = \{\hat{G}_1, \ldots, \hat{G}_{\widehat{K}}\}$ which forms a partition of $\{1, 2, \ldots, N\}$, with $\hat{G}_k = \{i : \hat{\beta}_i = \hat{\alpha}_k, 1 \leq i \leq N\}$, for any $1 \leq k \leq \widehat{K}$. We denote $\widehat{\boldsymbol{\alpha}}_{\widehat{K}} = (\hat{\alpha}_1', \ldots, \hat{\alpha}_{\widehat{K}}')'$, $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_1', \ldots, \hat{\beta}_N')'$, $\widehat{\mathcal{G}}_{\widehat{K}}$, and $\widehat{K}$, respectively, as the PLS estimates of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\mathcal{G}_{K_0}^0$, and K_0, using the PAGFL procedure.

Post-Lasso Least Squares Estimation

Given the fact that we estimate the group structure, we obtain the post-Lasso least squares estimator of α_k for $k = 1, \ldots, \widehat{K}$ as

$$\hat{\alpha}^p_{\hat{G}_k} = \Big(\sum_{i \in \hat{G}_k} \sum_{t=1}^{T} \tilde{x}_{it} \tilde{x}'_{it} \Big)^{-1} \sum_{i \in \hat{G}_k} \sum_{t=1}^{T} \tilde{x}_{it} \tilde{y}_{it}, \tag{2.9}$$

where $\widehat{K}$, and $\hat{G}_k$ are the estimated number of groups, and the groups identities via the PAGFL procedure. We denote the post-Lasso least squares estimator of $\boldsymbol{\alpha}$ by $\widehat{\boldsymbol{\alpha}}^p_{\widehat{K}} = (\hat{\alpha}^p_{\hat{G}_1}{}', \ldots, \hat{\alpha}^p_{\hat{G}_{\widehat{K}}}{}')'$.

2.3 Penalized GMM (PGMM) Estimation

In case (b), we propose to estimate $\boldsymbol{\beta}$ by minimizing the following objective function

$$\begin{aligned} Q_{2,NT}(\boldsymbol{\beta}, \lambda_2) = &\sum_{i=1}^{N} \Big[\frac{1}{T} \sum_{t=1}^{T} z_{it} (\Delta y_{it} - \beta_i' \Delta x_{it}) \Big]' W_{i,NT} \Big[\frac{1}{T} \sum_{t=1}^{T} z_{it} (\Delta y_{it} - \beta_i' \Delta x_{it}) \Big] \\ &+ \frac{\lambda_2}{N} \sum_{1 \le i < j \le N} \sum \ddot{w}_{ij} \|\beta_i - \beta_j\|, \end{aligned} \tag{2.10}$$

where $\lambda_2 \ge 0$ is a tuning parameter, $W_{i,NT}$ is a $q \times q$ p.d. matrix, and $\ddot{w}_{ij}$ is a data-driven weight defined by

$$\ddot{w}_{ij} = \|\ddot{\beta}_i - \ddot{\beta}_j\|^{-\kappa}, \quad \text{for } i, j = 1, \ldots, N, \tag{2.11}$$

where $\ddot{\beta}_i$ and $\ddot{\beta}_j$ are preliminary consistent estimates of β_i and β_j, respectively, and κ is a user-specified positive constant that usually takes value 2 in the literature.

To obtain the adaptive weights $\{\ddot{w}_{ij} : i, j \in \{1, \ldots, N\}\}$, we propose to obtain the preliminary estimates $\ddot{\boldsymbol{\beta}} = (\ddot{\beta}_1', \ldots, \ddot{\beta}_N')'$ by minimizing the first term in equation (2.10). Thus, for the i-th element of $\ddot{\boldsymbol{\beta}}$, we have

$$\ddot{\beta}_i = \Big[\Big(\frac{1}{T} \sum_{t=1}^{T} \Delta x_{it} z_{it}' \Big) W_{i,NT} \Big(\frac{1}{T} \sum_{t=1}^{T} z_{it} \Delta x_{it}' \Big) \Big]^{-1} \Big(\frac{1}{T} \sum_{t=1}^{T} \Delta x_{it} z_{it}' \Big) W_{i,NT} \Big(\frac{1}{T} \sum_{t=1}^{T} z_{it} \Delta y_{it} \Big). \tag{2.12}$$

The first term in the definition of the objective function in (2.10) is different from the usual GMM objective function in the panel setting where only one weight matrix is needed and the double

summation $\sum_{i=1}^{N}\sum_{t=1}^{T}$ occurs twice, one before the weight and the other after the weight matrix. The reason is that because the true group membership of individual units is unknown, we cannot apply the usual GMM objective function here.

It is worth emphasizing that the minimization of (2.10) is a convex optimization problem, hence it does not suffer from multiple local minima issue, and its global minimizer $\widetilde{\boldsymbol{\beta}} = \arg\min Q_{2,NT}(\boldsymbol{\beta}, \lambda_2)$, can be efficiently solved. We suppress the dependence of $\widehat{\boldsymbol{\beta}} \equiv \widehat{\boldsymbol{\beta}}(\lambda_2)$ on λ_2 unless necessary, and choose the tuning parameter using a data-driven method proposed in Section 4.4.

The penalty in (2.10) shrinks some of the pairs $\beta_i - \beta_j$ to zero, hence we can partition the slope parameters into groups. In practice, let $\{\tilde{\alpha}_1, \ldots, \tilde{\alpha}_{\widetilde{K}}\}$ be the distinct values of $\widetilde{\boldsymbol{\beta}}$. Then, we define $\widetilde{\mathcal{G}}_{\widetilde{K}} = \{\tilde{G}_1, \ldots, \tilde{G}_{\widetilde{K}}\}$ which forms a partition of $\{1, 2, \ldots, N\}$, with $\tilde{G}_k = \{i : \tilde{\beta}_i = \tilde{\alpha}_k, 1 \leq i \leq N\}$, for any $1 \leq k \leq \widetilde{K}$. We denote $\widetilde{\boldsymbol{\alpha}}_{\widetilde{K}} = (\tilde{\alpha}_1', \ldots, \tilde{\alpha}_{\widetilde{K}}')'$, $\widetilde{\boldsymbol{\beta}} = (\tilde{\beta}_1', \ldots, \tilde{\beta}_N')'$, $\widetilde{\mathcal{G}}_{\widetilde{K}}$, and $\widetilde{K}$, respectively, as the penalized generalized method of moments (PGMM) estimates of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\mathcal{G}_{K_0}$, and K_0, using the PAGFL procedure.

Post-Lasso GMM Estimation

Given the fact that we estimate the group structure, we obtain the post-Lasso GMM estimator of α_k for $k = 1, \ldots, \widetilde{K}$ as

$$\tilde{\alpha}^p_{\tilde{G}_k} = \left(\tilde{Q}^{(k)\prime}_{z\Delta x}\tilde{W}^{(k)}_{NT}\tilde{Q}^{(k)}_{z\Delta x}\right)^{-1}\tilde{Q}^{(k)\prime}_{z\Delta x}\tilde{W}^{(k)}_{NT}\tilde{Q}^{(k)}_{z\Delta y}, \tag{2.13}$$

where $\widetilde{K}$, and $\tilde{G}_k$ are the estimated number of groups, and the groups identities via the PAGFL procedure, $\tilde{Q}^{(k)}_{z\Delta x} = T^{-1}\sum_{i\in\tilde{G}_k}\sum_{t=1}^{T} z_{it}(\Delta x_{it})'$, $\tilde{Q}^{(k)}_{z\Delta y} = T^{-1}\sum_{i\in\tilde{G}_k}\sum_{t=1}^{T} z_{it}\Delta y_{it}$, and $\tilde{W}^{(k)}_{NT}$ is a group-specific $q \times q$ p.d. symmetric weight matrix. We denote the post-Lasso GMM estimator of $\boldsymbol{\alpha}$ by $\widetilde{\boldsymbol{\alpha}}^p_{\widetilde{K}} = (\tilde{\alpha}^{p\prime}_{\tilde{G}_1}, \ldots, \tilde{\alpha}^{p\prime}_{\tilde{G}_{\widetilde{K}}})'$.

3 Asymptotic properties of the PLS estimators

In this section, we provide the asymptotic properties of the PLS estimator and the associated post-Lasso estimator.

3.1 Assumptions

Let $\hat{Q}_{i,\tilde{x}\tilde{x}} = T^{-1}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{x}_{it}'$ and $\hat{Q}_{i,\tilde{x}\tilde{u}} = T^{-1}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{u}_{it}$. Define $J_{min} = \min_{1\le l<k\le K_0}\|\alpha_l^0 - \alpha_k^0\|$ which denotes the minimum degree of heterogeneity in the slope coefficients between groups. Define the $pK_0 \times pK_0$ matrix $\Phi_{NT} = diag(\Phi_{NT,1},\ldots,\Phi_{NT,K_0})$, where $\Phi_{NT,k} = \sum_{i\in G_k^0}\hat{Q}_{i,\tilde{x}\tilde{x}}$, and the $pK_0 \times 1$ matrix $\Psi_{NT}^u = diag(\Psi_{NT,1}^u,\ldots,\Psi_{NT,K_0}^u)$, where $\Psi_{NT,k}^u = \sum_{i\in G_k}\hat{Q}_{i,\tilde{x}\tilde{u}}$, for $k = 1,\ldots,K_0$.

To study the asymptotic properties of the PLS estimator and the post-Lasso estimator, we make the following assumptions.

Assumption A.1 *(i)* $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{u}_{it} = O_p(1)$ *for each* $i = 1,\ldots,N$.

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

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

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

Assumption A.2 *(i)* $T^{1/2}J_{min}\to\infty$ *as* $(N,T)\to\infty$.

(ii) $plim_{(N,T)\to\infty}T^{1/2}\lambda_1 J_{min}^{-\kappa} = c \in [0,\infty)$.

(iii) $plim_{(N,T)\to\infty}N_k\,T^{(\kappa+1)/2}\,\lambda_1/N = \infty$, *for each* $k = 1,\ldots,K_0$.

Assumption A.3 *Let* $\mathbb{D}_{K_0} = diag(\sqrt{N_1},\ldots,\sqrt{N_{K_0}})\otimes I_p$, *and* S *denote an arbitrary* $l\times pK_0$ *selection matrix such that* $\|S\|$ *is finite, and* $l\in\{1,2,\ldots,pK_0\}$ *is a fixed integer.*

(i) There exists $\Phi_0 > 0$ *such that* $\|\,\mathbb{D}_{K_0}^{-1}\,\Phi_{NT}\,\mathbb{D}_{K_0}^{-1} - \Phi_0\,\|_{sp} = o_p(1)$.

(ii) $\sqrt{T}\,S\,\Phi_0^{-1}\,\mathbb{D}_{K_0}^{-1}\,\Psi_{NT}^u - S\,\Phi_0^{-1}\,\mathbb{B}_{NT} \xrightarrow{D} N(0, S\Phi_0^{-1}\Psi_0\Phi_0^{-1}S')$ *as* $(N,T)\to\infty$, *where* $\mathbb{B}_{NT} = diag(\mathbb{B}_{NT,1},\ldots,\mathbb{B}_{NT,K_0})$, $\mathbb{B}_{NT,k} = \frac{1}{\sqrt{N_kT}}\sum_{i\in G_k^0}\sum_{t=1}^{T}\mathbb{E}(\tilde{x}_{it}\tilde{u}_{it})$ *is either zero or of order* $O(\sqrt{N_k/T})$ *depending on whether* x_{it} *is strictly exogenous.*

Assumption A.1(i) will be mostly satisfied in large dimensional panel data models with weakly exogenous regressors and can be replaced with sufficient or primitive conditions on the process $\{(x_{it},u_{it}), t\ge 1\}$ that ensure the central limit theory. Note that this assumption allows both conditional heteroscedasticity and serial correlation in $\{u_{it}, t\ge 1\}$. Also, Assumption A.1(iii) can

be easily verified from this assumption. The first part of Assumption A.1(ii) is standard in the literature, and the second one imposes restriction on the moments of x_{it}, the dependence structure on the regressors processes, and the relative rates at which N and T pass to infinity. Su et al. (2016) give details on sufficient and primitive conditions that ensure this assumption. Assumption A.1(iv) implies that as $N \to \infty$, the number of individuals within each group can be either fixed or diverge to infinity.[3] Assumption A.2 mainly specifies conditions on J_{min}, λ_1, N, and T. We use the probability limit in A.2(ii)–(iii) because we allow λ_1 to be data-driven and hence random. Assumption A.2(i) allows the minimum degree of heterogeneity size, J_{min}, to shrink to zero as $T \to \infty$, but at a rate slower than $T^{-1/2}$. Assumptions A.2(ii) and A.2(iii) are used to show the consistency and classification consistency of the PAGFL. In addition, we allow the number of groups K_0 to diverge to infinity at a slow rate. Noting that when the dimension of the PLS or post-Lasso diverge to infinity, we cannot derive the asymptotic normality directly, we make Assumption A.3 to provide conditions to ensure the asymptotic normality for any linear combination of the PLS or post-Lasso estimators, but it can be replaced with various commonly primitive conditions. If K_0 remains fixed as $N \to \infty$, we can replace the selection matrix S by an identity matrix.

3.2 Consistency

The following theorem establishes the consistency of $\hat{\beta}_i$ for $i = 1, \ldots, N$.

Theorem 3.1 *Suppose that Assumptions A.1 and A.2(ii) hold. Then for* $i = 1, \ldots, N$,

(i) $\hat{\beta}_i - \beta_i^0 = O_p(T^{-1/2})$,

(ii) $\frac{1}{N}\sum_{i=1}^{N} \|\hat{\beta}_i - \beta_i^0\|^2 = O_p(T^{-1})$.

Theorem 3.1(i) and (ii), respectively, establish the pointwise and mean square convergence rates of $\hat{\beta}_i$.

The following theorem establishes the classification consistency.

Theorem 3.2 *Suppose that Assumptions A.1 and A.2(ii)–(iii) hold. Then*

$$P\Big(\|\hat{\beta}_i - \hat{\beta}_j\| = 0 \text{ for all } i \,\&\, j \in G_k^0,\ k \in \{1, \ldots, K_0\}\Big) \to 1, \text{ as } (N, T) \to \infty.$$

[3] A main reason that our method allows K_0 to diverge, is because we require T to be sufficiently large enough to produce consistent preliminary estimates of the slope coefficients.

Theorem 3.2 says that with probability approaching one all the zero vectors in $\{\|\beta_i - \beta_j\|, 1 \leq i, j \leq N\}$ must be estimated as exactly zero by the PLS method so that the estimated number of groups cannot be large than K_0 when T is sufficiently large. These results together with the consistency results in Theorem 3.1 imply that the PAGFL has the ability to consistently identify the true group structure with the correct number of individual units within each group when the minimum group size J_{min} does not shrink to zero too fast.

Corollary 3.3 *Suppose that Assumptions A.1 and A.2 hold. Then*

(i) $\lim_{(N,T)\to\infty} P(\hat{K} = K_0) = 1$,

(ii) $\lim_{(N,T)\to\infty} P(\hat{G}_1 = G_1^0, \ldots, \hat{G}_{K_0} = G_{K_0}^0) = 1$.

The above corollary implies that, we can determine the correct number of groups, as long as the minimum degree of heterogeneity, J_{min}, remains fixed or shrinks to zero at a rate slower than $T^{-1/2}$ as $T \to \infty$.

3.3 Limiting Distribution of the PLS and post-Lasso Estimators

In this section, we study the asymptotic distribution of the PLS and post-Lasso estimators. Note that if each individual's group membership is known, the oracle estimator is the within group estimator of α_k^0 which can be formulated as $\bar{\alpha}_k = \left(\sum_{i\in G_k^0}\sum_{t=1}^T \tilde{x}_{it}\tilde{x}_{it}'\right)^{-1}\sum_{i\in G_k^0}\sum_{t=1}^T \tilde{x}_{it}\tilde{y}_{it}$.

The following theorem reports the limiting distribution of the post-Lasso estimator $\widehat{\boldsymbol{\alpha}}^p_{\widehat{K}}$.

Theorem 3.4 *Suppose that Assumptions A.1–A.3 hold. Then, we have*

$$\sqrt{T}SD_{K_0}(\widehat{\boldsymbol{\alpha}}^p_{\widehat{K}} - \boldsymbol{\alpha}^0) - SD_{K_0}\Phi_{NT}^{-1}D_{K_0}\mathbb{B}_{NT} \xrightarrow{D} N(0, S\Phi_0^{-1}\Psi_0\Phi_0^{-1}S'). \tag{3.1}$$

To report the limiting distribution of the PLS estimator, $\widehat{\boldsymbol{\alpha}}_{\widehat{K}}$, we make the following Assumption A.4 which is similar to Assumption A.2(ii), with the difference that we require that $J_{\min}^{-\kappa}$ or λ_1 to tend to zero at a faster rate than the one we need for the post-Lasso estimator. This is because the PLS estimator includes the penalty term which is the two summations over the individuals, and is of order $O_p(N^2)$. Hence, for the penalty term to vanish we need the group slope coefficients to be sufficiently separated.

Assumption A.4 $plim_{(N,T)\to\infty}(NT)^{1/2}\lambda_1 J_{min}^{-\kappa} = 0$.

Theorem 3.5 *Suppose that Assumptions A.1–A.4 hold. Then, we have*

$$\sqrt{T}SD_{K_0}(\widehat{\boldsymbol{\alpha}}_{\widehat{K}} - \boldsymbol{\alpha}^0) - SD_{K_0}\Phi_{NT}^{-1}D_{K_0}\mathbb{B}_{NT} \xrightarrow{D} N(0, S\Phi_0^{-1}\Psi_0\Phi_0^{-1}S'). \tag{3.2}$$

Since the dimensions of the PLS and post-Lasso estimators diverge to infinity when $K_0 \to \infty$, following the literature on inference with a diverging number of parameters, we prove the asymptotic normality for any arbitrary linear combinations of elements of $\widehat{\boldsymbol{\alpha}}_{\widehat{K}}$ or $\widehat{\boldsymbol{\alpha}}^p_{\widehat{K}}$. The asymptotic result in Theorem 3.5 holds under Assumption A.5, which is because the PLS estimator includes the penalty of order $O_p(N^2)$, and for the penalty term to vanish we need $J_{\min}^{-\kappa}$ or λ_1 to offset the linear combinations of $\sqrt{N_k}$ rates of convergency.

Theorem 3.4 and Theorem 3.5 indicate that both the PLS estimator and the post-Lasso estimator achieve the same limiting distribution as the oracle within group estimator. Therefore, we say that the PLS estimator has the asymptotic oracle property. We note that the oracle estimator is the infeasible estimator, because it can be obtained if the group structure is known. In addition, $\mathbb{B}_{NT,k}$ is not equal to zero in case of dynamic panel data models. In fact, it is well known in the literature that the fixed effect estimator has an asymptotic bias of order $O(1/T)$. This suggests that in dynamic panel models $\mathbb{B}_{NT,k} = O(\sqrt{N_k/T})$ and bias correction is required, unless the rate at which T goes to infinity is faster than that of N_k. There are various methods proposed in the literature to estimate the bias term such as Kiviet (1995), Hahn and Kuersteiner (2002), Phillips and Sul (2007), Lee (2012), Gourieroux et al. (2010) and Han et al. (2014), among others, and we refer the readers to these papers.

3.4 Choosing the Tuning Parameter λ_1

Let $\widehat{\boldsymbol{\alpha}}^p_{\widehat{K}_{\lambda_1}} \equiv \widehat{\boldsymbol{\alpha}}^p_{\widehat{K}_{\lambda_1}}(\widehat{\mathcal{G}}_{\widehat{K}_{\lambda_1}}) = \left(\hat{\alpha}_1^p(\widehat{\mathcal{G}}_{\widehat{K}_{\lambda_1}})', \ldots, \hat{\alpha}^p_{\widehat{K}_{\lambda_1}}(\widehat{\mathcal{G}}_{\widehat{K}_{\lambda_1}})'\right)'$ denote the post-Lasso estimates of the regression coefficients based on the group structure in $\widehat{\mathcal{G}}_{\widehat{K}_{\lambda_1}} \equiv \widehat{\mathcal{G}}_{\widehat{K}_{\lambda_1}}(\lambda_1) = \{\hat{G}_1(\lambda_1), \ldots, \hat{G}_{\widehat{K}_{\lambda_1}}(\lambda_1)\}$, where we make the dependence of the estimates on λ_1 explicit. Let $\hat{\sigma}^2_{\widehat{\mathcal{G}}_{\widehat{K}_{\lambda_1}}} = \frac{1}{NT}\sum_{k=1}^{\widehat{K}_{\lambda_1}}\sum_{i\in\hat{G}_k(\lambda_1)}\sum_{t=1}^{T}(\tilde{y}_{it} - \hat{\alpha}_k^p(\hat{\mathcal{G}}_{\widehat{K}_{\lambda_1}})'\tilde{x}_{it})^2$. Following Wang et al. (2007), Zhang et

al. (2010), Qian and Su (2015), and Qian and Su (2016), we propose to select the tuning parameter λ_1 by minimizing the following information criterion (IC):

$$IC_1(\lambda_1) = \hat{\sigma}^2_{\widehat{\mathcal{G}}_{\widehat{K}_{\lambda_1}}} + \rho_{1,NT} p \widehat{K}_{\lambda_1}, \tag{3.3}$$

where $\rho_{1,NT}$ is a tuning parameter.

We proceed to describe the asymptotic properties of (3.3). Let $\Lambda = [0, \lambda_{1,\max}]$ be a bounded interval in $\mathbb{R}^+$. We divide Λ into three subsets Λ_0, Λ_-, and Λ_+ which are defined as follows

$$\Lambda_0 = \{\lambda_1 \in \Lambda : \widehat{K}_{\lambda_1} = K_0\}, \quad \Lambda_- = \{\lambda_1 \in \Lambda : \widehat{K}_{\lambda_1} < K_0\}, \quad \Lambda_+ = \{\lambda_1 \in \Lambda : \widehat{K}_{\lambda_1} > K_0\}.$$

The sets Λ_0, Λ_-, and Λ_+ denote subsets of Λ in which the true, under-, and over-number of groups are produced by our PAGFL procedure, respectively. They are random because $\hat{K}_{\lambda_1}$ has to be determined based on the random sample, but we suppress their dependence on the sample sizes N and T for notational simplicity. Let $\mathcal{G}_{(K)} = \{G_{(K,1)}, \ldots, G_{(K,K)}\}$ be any K-partition of the set of individual indices $\{1, \ldots, N\}$, and let $\mathfrak{G}_K$ denote the collection of such partitions. Let $\hat{\sigma}^2_{\mathcal{G}_{(K)}} = (NT)^{-1} \sum_{k=1}^{K} \sum_{i \in G_{(K,k)}} \sum_{t=1}^{T} (\tilde{y}_{it} - \hat{\alpha}'_{G_{(K,k)}} \tilde{x}_{it})^2$, where $\hat{\alpha}_{G_{(K,k)}} = (\sum_{i \in G_{(K,k)}} \sum_{t=1}^{T} \tilde{x}_{it} \tilde{x}'_{it})^{-1} \sum_{i \in G_{(K,k)}} \sum_{t=1}^{T} \tilde{x}_{it} \tilde{y}_{it}$.

Let $\lambda^0_{1,NT}$ denote an element in Λ^0 that satisfies the conditions on λ_1 is Assumptions A.2(ii)–(iii). We make the following assumptions, to state the next asymptotic result.

Assumption A.5 *As* $(N,T) \to \infty$, $\min_{1 \le K \le K_0} \inf_{\mathcal{G}_{(K)} \in \mathfrak{G}_K} \hat{\sigma}^2_{\mathcal{G}_{(K)}} \xrightarrow{P} \underline{\sigma}^2 > \sigma_0^2$, *where* $\sigma_0^2 = plim_{(N,T)\to\infty} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{u}_{it}^2$.

Assumption A.6 *As* $(N,T) \to \infty$, $K_0\, \rho_{1,NT} \to 0$, *and* $N \rho_{1,NT} \to \infty$.

Assumption A.5 is intuitively clear and applies under primitive conditions in a variety of models. It requires that all under-fitted models yield asymptotic mean squared errors that are larger than that of the true model, σ_0^2. Assumption A.6 reflects the usual conditions for the consistency of model selection, that is, the penalty coefficient $\rho_{1,NT}$ cannot shrink to zero either too fast or too slowly.

Theorem 3.6 *Suppose that Assumptions A.1, A.2(i), A.3, A.5, and A.6 hold. Then,*

$$P\Big(\inf_{\lambda_1 \in \Lambda_- \cup \Lambda_+} IC_1(\lambda_1) > IC_1(\lambda^0_{1,NT})\Big) \to 1 \qquad as \qquad (N,T) \to \infty. \tag{3.4}$$

Theorem 3.6 indicates that the tuning parameters (λ_1's) that yield the over-estimated or under-estimated number of group structures fail to minimize the information criterion with probability approaching one. Consequently, the minimizer of $IC_1(\lambda_1)$ can only be the one that produces K_0 number of groups. We note that the proof of Theorem 3.6 does not require λ_1 to satisfy Assumptions A.2(ii)–(iii).

4 Asymptotic properties of the PGMM estimators

In this section, we provide the asymptotic properties of the PGMM estimator and the associated post-Lasso estimator.

4.1 Assumptions

Let $\tilde{Q}_{i,z\Delta x} = T^{-1}\sum_{t=1}^T z_{it}\Delta x_{it}'$, $\tilde{Q}_{i,z\Delta y} = T^{-1}\sum_{t=1}^T z_{it}\Delta y_{it}$, $\tilde{Q}_{i,z\Delta u} = T^{-1}\sum_{t=1}^T z_{it}\Delta u_{it}$, $\bar{Q}_{i,z\Delta x} = T^{-1}\sum_{t=1}^T \mathbb{E}(z_{it}\Delta x_{it}')$, and $\bar{Q}_{i,z\Delta y} = T^{-1}\sum_{t=1}^T \mathbb{E}(z_{it}\Delta y_{it})$. Let $\xi_{it} = (\Delta y_{it}, (\Delta x_{it})', z_{it}')'$, $\rho(\xi_{it},\beta_i) = z_{it}(\Delta y_{it} - \beta_i'\Delta x_{it})$, and $\bar{\rho}_{i,T}(\beta_i) = T^{-1/2}\sum_{t=1}^T \Big[\rho(\xi_{it},\beta_i) - \mathbb{E}(\rho(\xi_{it},\beta_i))\Big]$. For each group $k = 1,\ldots,K_0$, let $W_{NT}^{(k)}$ be a $q\times q$ p.d. matrix, $Q_{z\Delta x,NT}^{(k)} = T^{-1}\sum_{i\in G_k^0}\sum_{t=1}^T z_{it}(\Delta x_{it})'$, and $Q_{z\Delta u,NT}^{(k)} = T^{-1}\sum_{i\in G_k^0}\sum_{t=1}^T z_{it}\Delta u_{it}$. Define the $pK_0\times pK_0$ matrix $\Upsilon_{NT} \equiv \Upsilon_{NT}(\mathcal{G}^0_{K_0}) = diag\Big(\Upsilon_{NT,1}(\mathcal{G}^0_{K_0}),\ldots,\Upsilon_{NT,K_0}(\mathcal{G}^0_{K_0})\Big)$, and the $pK_0\times 1$ vector $\Xi^u_{NT} \equiv \Xi^u_{NT}(\mathcal{G}^0_{K_0}) = diag\Big(\Xi^u_{NT,1}(\mathcal{G}^0_{K_0}),\ldots,\Xi^u_{NT,K_0}(\mathcal{G}^0_{K_0})\Big)$, where $\Upsilon_{NT,k}(\mathcal{G}^0_{K_0}) = Q_{z\Delta x,NT}^{(k)\prime} W_{NT}^{(k)} Q_{z\Delta x,NT}^{(k)}$, and $\Xi^u_{NT,k}(\mathcal{G}^0_{K_0}) = Q_{z\Delta x,NT}^{(k)\prime} W_{NT}^{(k)} Q_{z\Delta u,NT}^{(k)}$. Similarly, define the $pK_0\times pK_0$ matrix $\bar{\Upsilon}_{NT} \equiv \bar{\Upsilon}_{NT}(\mathcal{G}^0_{K_0}) = diag\Big(\bar{\Upsilon}_{NT,1}(\mathcal{G}^0_{K_0}),\ldots,\bar{\Upsilon}_{NT,K_0}(\mathcal{G}^0_{K_0})\Big)$, where $\bar{\Upsilon}_{NT,k}(\mathcal{G}^0_{K_0}) = \sum_{i\in G_k^0}\tilde{Q}_{i,z\Delta x}' W_{i,NT}\tilde{Q}_{i,z\Delta x}$, and the $pK_0\times 1$ vector $\bar{\Xi}^u_{NT} \equiv \bar{\Xi}^u_{NT}(\mathcal{G}^0_{K_0}) = diag\Big(\bar{\Xi}^u_{NT,1}(\mathcal{G}^0_{K_0}),\ldots,\bar{\Xi}^u_{NT,K_0}(\mathcal{G}^0_{K_0})\Big)$, where $\bar{\Xi}^u_{NT,k}(\mathcal{G}^0_{K_0}) = \sum_{i\in G_k^0}\tilde{Q}_{i,z\Delta x}' W_{i,NT}\tilde{Q}_{i,z\Delta u}$.

To study the asymptotic properties of the PGMM and the post-Lasso estimators, we make the following assumptions.

Assumption B.1 *(i)* $\mathbb{E}(\rho(\xi_{it},\beta_i^0)) = 0$, *for each* $i = 1,\ldots,N$ *and* $t = 1,\ldots,T$.

(ii) $\sup_{\beta_i} \|\bar{\rho}_{i,T}(\beta_i)\| = O_p(1)$, *and* $\frac{1}{N}\sum_{i=1}^{N} \|\bar{\rho}_{i,T}(\beta_i)\|^2 = O_p(1)$, *for any* β_i *and* $i = 1, \ldots, N$.

(iii) $\tilde{Q}_{i,z\Delta x} \xrightarrow{P} \bar{Q}_{i,z\Delta x} > 0$, *for each* $i = 1, \ldots, N$. *There exists a positive constant* $\underline{c}_{\bar{Q}}$ *such that* $\lim_{(N,T)\to\infty} \min_{1\le i\le N} \mu_{min}(\bar{Q}'_{i,z\Delta x}\bar{Q}_{i,z\Delta x}) = \underline{c}_{\bar{Q}}$.

(iv) There exist non-random matrices W_i *such that* $\max_{1\le i\le N} \|W_{i,NT} - W_i\| = o_p(1)$, *and* $\liminf_{(N,T)\to\infty} \min_{1\le i\le N} \mu_{min}(W_i) = \underline{c}_W > 0$.

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

Assumption B.2 *(i)* $T^{1/2} J_{min} \to \infty$ *as* $(N, T) \to \infty$.

(ii) $plim_{(N,T)\to\infty} T^{1/2}\lambda_1 J_{min}^{-\kappa} = c \in [0, \infty)$.

(iii) $plim_{(N,T)\to\infty} N_k T^{(\kappa+1)/2}\lambda_2/N = \infty$, *for each* $k = 1, \ldots, K_0$.

Assumption B.3 *Let* $\mathbb{D}_{K_0} = diag(\sqrt{N_1}, \ldots, \sqrt{N_{K_0}}) \otimes I_p$, *and* S *denote an arbitrary* $l \times pK_0$ *selection matrix such that* $\|S\|$ *is finite, where* $l \in \{1, 2, \ldots, pK_0\}$ *is a fixed integer.*

(i) There exists $\Upsilon_0 > 0$ *such that* $\|D_{K_0}^{-3}\Upsilon_{NT} D_{K_0}^{-1} - \Upsilon_0\|_{sp} = o_p(1)$.

(ii) $\sqrt{T} S\Upsilon_0^{-1} D_{K_0}^{-3}\Xi_{NT}^{u} \xrightarrow{D} N(0, S\Upsilon_0^{-1} V_0 \Upsilon_0^{-1} S')$.

Assumption B.4 *(i) There exists* $\bar{\Upsilon}_0 > 0$ *such that* $\|D_{K_0}^{-1}\bar{\Upsilon}_{NT} D_{K_0}^{-1} - \bar{\Upsilon}_0\|_{sp} = o_p(1)$.

(ii) $\sqrt{T} S\bar{\Upsilon}_0^{-1} D_{K_0}^{-1}\bar{\Xi}_{NT}^{u} - S\bar{\Upsilon}_0^{-1}\mathbb{B}_{NT} \xrightarrow{D} N(0, S\tilde{\Upsilon}_0^{-1}\bar{V}_0\bar{\Upsilon}_0^{-1} S')$.

Assumption B.1(i) specifies moment conditions to identify β_i^0. Assumption B.1(ii) is needed as we do not specify the data structure, where its first part can generally be verified, see Su et al. (2016) for a discussion. Assumption B.1(iii) together with Assumption B.1(i) provide a rank condition for the identification of β_i^0. If one sets $W_{i,NT} = I_q$, then Assumption B.1(iv) is automatically satisfied. Assumption B.1(v) implies that as $N \to \infty$, the number of individuals within each group can be either fixed or diverge to infinity. Assumption B.3 and Assumption B.4 specify conditions for deriving the limiting distributions of the post-Lasso estimator and Lasso estimator, respectively. These assumptions can be verified under various primitive conditions, see Su et al. (2016) who give details on some conditions. Assumption B.2 parallels Assumption A.2, which specifies conditions on J_{min}, λ_2, N, and T.

4.2 Consistency

The following theorem establishes the consistency of the PGMM estimator, $\tilde{\beta}_i$ for $i = 1, \ldots, N$.

Theorem 4.1 *Suppose that Assumptions B.1 and B.2(ii) hold. Then*

(i) $\tilde{\beta}_i - \beta_i^0 = O_p(T^{-1/2})$ *for* $i = 1, \ldots, N$,

(ii) $\frac{1}{N}\sum_{i=1}^{N} \|\tilde{\beta}_i - \beta_i^0\|^2 = O_p(T^{-1})$.

Theorem 4.1(i) and (ii), respectively, establish the pointwise and mean square convergence rates of $\{\tilde{\beta}_i : i = 1, \ldots, N\}$.

The following theorem establishes the classification consistency.

Theorem 4.2 *Suppose Assumptions B.1 and B.2(i)–(ii) hold. Then*

$$P\Big(\|\tilde{\beta}_i - \tilde{\beta}_j\| = 0 \text{ for all } i \,\&\, j \in G_k^0,\ k \in \{1, \ldots, K_0\}\Big) \to 1, \text{ as } (N, T) \to \infty.$$

Theorem 4.2 says that with probability approaching one all the zero vectors in $\{\|\beta_i - \beta_j\|, 1 \leq i, j \leq N\}$ must be estimated as exactly zero by the PGMM method so that the estimated number of groups cannot be different from K_0 when T is sufficiently large. This result together with the consistency result in Theorem 4.1 imply that the PAGFL has the ability to identify the true group structure with the correct number of individual units within each group consistently when the minimum degree of heterogeneity, J_{min}, does not shrink to zero too fast.

Corollary 4.3 *Suppose that Assumptions B.1 and B.2 hold. Then*

(i) $\lim_{(N,T)\to\infty} P(\widetilde{K} = K_0) = 1$,

(ii) $\lim_{(N,T)\to\infty} P(\tilde{G}_1 = G_1^0, \ldots, \tilde{G}_{K_0} = G_{K_0}^0) = 1$.

The above corollary implies that, as long as J_{min} remains fixed or shrinks to zero at a rate slower than $T^{-1/2}$ as $T \to \infty$, we can determine the correct number of groups.

4.3 Limiting Distribution of the PGMM and post-Lasso Estimators

In this section, we study the asymptotic distribution of the PGMM and post-Lasso estimators. Note that if each individual's group membership is known, the oracle estimator is the solution to a usual GMM objective function which can be formulated as below

$$\breve{\alpha}_k = \left[Q_{z\Delta x,NT}^{(k)\prime} W_{NT}^{(k)} Q_{z\Delta x,NT}^{(k)} \right]^{-1} Q_{z\Delta x,NT}^{(k)\prime} W_{NT}^{(k)} Q_{z\Delta y,NT}^{(k)}, \tag{4.1}$$

where $Q_{z\Delta y,NT}^{(k)} = T^{-1} \sum_{i \in G_k^0} \sum_{t=1}^{T} z_{it}(\Delta y_{it})$ for each $k = 1, \ldots, K_0$.

The following theorem reports the limiting distribution of the post-Lasso estimator, $\widetilde{\boldsymbol{\alpha}}_{\widetilde{K}}^{p}$.

Theorem 4.4 *Suppose Assumptions B.1–B.3 hold. Then*

$$\sqrt{T} S D_{K_0} (\widetilde{\boldsymbol{\alpha}}_{\widetilde{K}}^{p} - \boldsymbol{\alpha}^0) \xrightarrow{D} N(0, S \Upsilon_0^{-1} V_0 \Upsilon_0^{-1} S').$$

The above theorem says that the post-Lasso GMM estimator $\widetilde{\boldsymbol{\alpha}}_{\widetilde{K}}^{p}$ asymptotically has the same limiting distribution as the infeasible oracle estimator $\breve{\boldsymbol{\alpha}} = (\breve{\alpha}_1', \ldots, \breve{\alpha}_{K_0}')'$, which further indicates that the post-lasso GMM estimator has the oracle property.

The following theorem reports the limiting distribution of the PGMM estimator, $\widetilde{\boldsymbol{\alpha}}_{\widetilde{K}}$, but we need to make the following assumotion:

Assumption B.5 $plim_{(N,T)\to\infty} (NT)^{1/2} \lambda_2 J_{min}^{-\kappa} = 0$.

Similar to Assumption A.4 in the PLS estimation, we need Assumption B.5 for the penalty term in the Lasso estimator to vanish.

Theorem 4.5 *Suppose Assumptions B.1–B.2, and B.4–B.5 hold. Then*

$$\sqrt{T} S D_{K_0} (\widetilde{\boldsymbol{\alpha}}_{\widetilde{K}} - \boldsymbol{\alpha}^0) - S D_{K_0} \bar{\Upsilon}_{NT}^{-1} D_{K_0} \mathbb{B}_{NT} \xrightarrow{D} N(0, S \bar{\Upsilon}_0^{-1} \bar{V}_0 \bar{\Upsilon}_0^{-1} S').$$

Apparently, the PGMM estimator does not have the same asymptotic distribution as the oracle estimator under general conditions. This is because the true group membership of individual units is unknown, and hence the PGMM estimator is the solution to the objective function in (2.10) which is different from the usual GMM objective function where only one weight matrix is needed

and the double summation occurs twice, one before the weight and the other after the weight matrix. However, since the post-Lasso GMM estimator is the solution to the usual GMM objective function for the individuals within each group, and the procedure consistently estimate the true group memberships, the post-Lasso GMM estimator has the oracle property.

Remark 4.1 *In the special case where for each* $i \in G_k^0$ *and* $k = 1, \ldots, K_0$, *conditions* $(i)\, W_{i,NT} = W_{NT}^{(k)}$, $(ii)\, N_k^{-1}\bar{\Upsilon}_{NT,k}(G_{K_0}^0)$ *shares the same probability limit as* $N_k^{-2}\Upsilon_{NT,k}(G_{K_0}^0)$, $(iii)\, N_k^{-1/2}\bar{\Xi}_{NT,k}^u(G_{K_0}^0)$ *shares the same probability limit as* $N_k^{-3/2}\Upsilon_{NT,k}^u(G_{K_0}^0)$, *and* $(iv)\, \mathbb{B}_{NT,k} = 0$, *hold, then* $\bar{\Upsilon}_0 = \Upsilon_0$, *and* $\bar{V}_0 = V_0$. *Therefore, the PGMM estimator has the oracle property.*

4.4 Choosing the Tuning Parameter λ_2

Let $\widetilde{\boldsymbol{\alpha}}_{\widetilde{K}_{\lambda_2}}^p \equiv \widetilde{\boldsymbol{\alpha}}_{\widetilde{K}_{\lambda_2}}^p(\widetilde{\mathcal{G}}_{\widetilde{K}_{\lambda_2}}) = \left(\tilde{\alpha}_1^p(\widetilde{\mathcal{G}}_{\widetilde{K}_{\lambda_2}})', \ldots, \tilde{\alpha}_{\widetilde{K}_{\lambda_2}}^p(\widetilde{\mathcal{G}}_{\widetilde{K}_{\lambda_2}})'\right)'$ denote the post-Lasso estimates of the regression coefficients based on the group structure in $\widetilde{\mathcal{G}}_{\widetilde{K}_{\lambda_2}} \equiv \widetilde{\mathcal{G}}_{\widetilde{K}_{\lambda_2}}(\lambda_2) = \{\tilde{G}_1(\lambda_2), \ldots, \tilde{G}_{\widetilde{K}_{\lambda_2}}(\lambda_2)\}$, where we make the dependence of the estimates on λ_2 explicit. Let $\tilde{\sigma}^2_{\widetilde{\mathcal{G}}_{\widetilde{K}_{\lambda_2}}} = (NT)^{-1}\sum_{k=1}^{\widetilde{K}_{\lambda_2}}\sum_{i\in\tilde{G}_k(\lambda_2)}\sum_{t=1}^{T}(\Delta y_{it} - \tilde{\alpha}_k^p(\widetilde{\mathcal{G}}_{\widetilde{K}_{\lambda_2}})'\Delta x_{it})^2$. We propose to select the tuning parameter λ_2 by minimizing the following IC:

$$IC_2(\lambda_2) = \tilde{\sigma}^2_{\widetilde{\mathcal{G}}_{\widetilde{K}_{\lambda_2}}} + \rho_{2,NT}\, p \widetilde{K}_{\lambda_2}, \tag{4.2}$$

where $\rho_{2,NT}$ is a tuning parameter.

We proceed to describe the asymptotic properties of (4.2). Let $\bar{\Lambda} = [0, \lambda_{2,\max}]$ be a bounded interval in $\mathbb{R}^+$. We divide Λ into three subsets $\bar{\Lambda}_0$, $\bar{\Lambda}_-$, and $\bar{\Lambda}_+$ which are defined as follows

$$\bar{\Lambda}_0 = \{\lambda_2 \in \bar{\Lambda} : \widetilde{K}_{\lambda_2} = K_0\}, \quad \bar{\Lambda}_- = \{\lambda_2 \in \bar{\Lambda} : \widetilde{K}_{\lambda_2} < K_0\}, \quad \bar{\Lambda}_+ = \{\lambda_2 \in \bar{\Lambda} : \widetilde{K}_{\lambda_2} > K_0\}.$$

The sets $\bar{\Lambda}_0$, $\bar{\Lambda}_-$, and $\bar{\Lambda}_+$ denote subsets of Λ in which the true, under-, and over-number of groups are produced by our PAGFL procedure, respectively. We suppress their dependence on the sample sizes N and T for notational simplicity. Let $\mathcal{G}_{(K)} = \{G_{(K,1)}, \ldots, G_{(K,K)}\}$ be any K-partition of the set of individual indices $\{1, \ldots, N\}$, and let $\mathfrak{G}_K$ denote the collection of such partitions. Let $\tilde{\sigma}^2_{\mathcal{G}_{(K)}} = (NT)^{-1}\sum_{k=1}^{K}\sum_{i\in G_{(K,k)}}\sum_{t=1}^{T}(\Delta y_{it} - \tilde{\alpha}'_{G_{(K,k)}}\Delta x_{it})^2$, where $\tilde{\alpha}_{G_{(K,k)}} = \left(\tilde{Q}_{z\Delta x}^{(K,k)\prime}\tilde{W}_{NT}^{(K,k)}\tilde{Q}_{z\Delta x}^{(K,k)}\right)^{-1}\tilde{Q}_{z\Delta x}^{(K,k)\prime}\tilde{W}_{NT}^{(K,k)}\tilde{Q}_{z\Delta y}^{(K,k)}$, $\tilde{Q}_{z\Delta x}^{(K,k)} = T^{-1}\sum_{i\in G_{(K,k)}}\sum_{t=1}^{T} z_{it}\Delta x'_{it}$, $\tilde{Q}_{z\Delta y}^{(K,k)} = T^{-1}\sum_{i\in G_{(K,k)}}\sum_{t=1}^{T} z_{it}\Delta y_{it}$, and $\tilde{W}_{NT}^{(K,k)}$ is defined as before but with $k = 1, 2, \ldots, K$.

Let $\lambda_{2,NT}^0$ denote an element in $\bar{\Lambda}_0$ that satisfies the conditions on λ_2 in Assumptions B.2(ii)–(iii). We make the following assumptions, to state the next asymptotic result.

Assumption B.6 *As* $(N,T) \to \infty$, $\min_{1\le K\le K_0} \inf_{\mathcal{G}_{(K)}\in\mathfrak{G}_K} \tilde{\sigma}^2_{\mathcal{G}_{(K)}} \xrightarrow{P} \underline{\sigma}^2 > \sigma_0^2$, *where* $\sigma_0^2 = plim_{(N,T)\to\infty} \frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \Delta u_{it}^2$.

Assumption B.7 *As* $(N,T)\to\infty$, $K_0\,\rho_{2,NT}\to 0$, *and* $N\rho_{2,NT}\to\infty$.

Assumptions B.6 and B.7 parallel earlier Assumptions A.5–A.6. The following theorem implies that the minimizer of $IC_2(\lambda_2)$ can only be the one that produces the correct number of estimated group structure.

Theorem 4.6 *Suppose that Assumptions B.1, B.2(i), B.3, B.6, and B.7 hold. Then,*

$$P\Big(\inf_{\lambda_2\in\bar{\Lambda}_-\cup\bar{\Lambda}_+} IC_2(\lambda_2) > IC_2(\lambda_{2,NT}^0)\Big)\to 1 \qquad as \qquad (N,T)\to\infty. \tag{4.3}$$

5 Computation and Algorithm

The objective functions in (2.6) and (2.10) are not separable in β_i, which makes it difficult to compute the estimates directly. Thus, we define a new set of parameters $\delta_{ij} = \beta_i - \beta_j$ and reparameterize the criterion functions separately for PLS and PGMM and describe the implementation below.

5.1 PLS Computation

Reparametrizing the objective function in (2.6), is equivalent to the constraint optimization problem below

$$\min S_1(\boldsymbol{\beta},\boldsymbol{\delta}) = \frac{1}{2}\sum_{i=1}^N\sum_{t=1}^T(\tilde{y}_{it}-\beta_i'\tilde{x}_{it})^2 + \lambda_1^*\sum_{1\le i<j\le N}\sum \dot{w}_{ij}\|\delta_{ij}\|,$$

$$\text{subject to } \beta_i-\beta_j-\delta_{ij}=0,$$

where $\boldsymbol{\delta} = \{\delta_{ij}, i < j\}'$ and $\lambda_1^* = T\lambda_1/2N$. By the augmented Lagrangian method, the estimates of the parameters can be obtained by minimizing

$$L_1(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\nu}) = S_1(\boldsymbol{\beta}, \boldsymbol{\delta}) + \sum_{1\le i<j\le N}\sum \nu'_{ij}(\beta_i - \beta_j - \delta_{ij}) + \frac{\vartheta}{2}\sum_{1\le i<j\le N}\sum \|\beta_i - \beta_j - \delta_{ij}\|^2,$$

where $\boldsymbol{\nu} = \{\nu'_{ij}, i < j\}'$ are lagrange multipliers and ϑ is the penalty parameter. We can obtain the estimates of $(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\nu})$ through iterations by the ADMM, as we describe in the rest of this section.

The iteration process consists of updating $\boldsymbol{\beta}, \boldsymbol{\delta}$ and $\boldsymbol{\nu}$ iteratively. For a given $(\boldsymbol{\delta}, \boldsymbol{\nu})$, we obtain the updates of $\boldsymbol{\beta}$ by setting the derivative $\partial L_1(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\nu})/\partial\boldsymbol{\beta}$ to zero, where

$$L_1(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\nu}) = \frac{1}{2}\|\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}\|^2 + \frac{\vartheta}{2}\|\Lambda\boldsymbol{\beta} - \boldsymbol{\delta} + \vartheta^{-1}\boldsymbol{\nu}\|^2 + C,$$

and C is a constant independent of $\boldsymbol{\beta}$, $\tilde{\boldsymbol{y}} = (\tilde{y}'_1, \ldots, \tilde{y}'_N)'$, $\tilde{y}_i = (\tilde{y}_{i1}, \ldots, \tilde{y}_{iT})'$ for each $i = 1, \ldots, N$, $\tilde{\boldsymbol{X}} = diag(\tilde{X}_1, \ldots, \tilde{X}_N)$, $\tilde{X}_i = (\tilde{x}_{i1}, \ldots, \tilde{x}_{iT})'$ for each $i = 1, \ldots, N$. Besides, $\Lambda = \nabla \otimes I_p$, where $\nabla = \{(e_i - e_j), 1 \le i < j \le N\}'$ and e_i is an $N \times 1$ vector whose ith element is one and the remaining ones are zero. Further, we note that the minimizer of $L_1(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\nu})$ with respect to δ_{ij}, for given $(\boldsymbol{\beta}, \boldsymbol{\nu})$, has a closed form solution and is unique. In practice, for given $(\boldsymbol{\beta}, \boldsymbol{\nu})$, the minimization problem with respect to δ_{ij} is equivalent to the following minimization

$$\min \frac{\vartheta}{2}\sum_{1\le i<j\le N}\sum \|\zeta_{ij} - \delta_{ij}\|^2 + \lambda_1^*\sum_{1\le i<j\le N}\sum \dot{w}_{ij}\|\delta_{ij}\|,$$

where $\zeta_{ij} = \beta_i - \beta_j + \vartheta^{-1}\nu_{ij}$. Thus, the closed form solution is

$$\hat{\delta}_{ij} = ST(\zeta_{ij}, \lambda_1^*/\vartheta), \tag{5.1}$$

where $ST(\boldsymbol{a}, b) = (1 - b/\|\boldsymbol{a}\|)_+ \boldsymbol{a}$ is the groupwise soft thresholds rule, and $(c)_+ = 1(c > 0)c$.

We track the progress of the ADMM based on the primal residual at step m, $r^{(m)} = \Lambda\boldsymbol{\beta}^{(m)} - \boldsymbol{\delta}^{(m)}$, and stop the algorithm when $\|r^{(m)}\| < \epsilon$. The algorithm can be summarized as following:

PLS Algorithm:

1. **Initialization:** Find initial estimates of $\beta_i^{(0)}$ by minimizing the first term of (2.6) for all $i = 1, \ldots, N$. Let the initial values of $\boldsymbol{\nu}^{(0)} = 0$, and $\delta_{ij}^{(0)} = \beta_i^{(0)} - \beta_j^{(0)}$.

2. **Iterations:** At iteration $m \geq 1$, for given $\boldsymbol{\delta}^{(m-1)}$ and $\boldsymbol{\nu}^{(m-1)}$,

   (a) update $\boldsymbol{\beta}^{(m)}$ which is the minimizer of $L_1(\boldsymbol{\beta}, \boldsymbol{\delta}^{(m)}, \boldsymbol{\nu}^{(m)})$ and takes the form below

   $$\boldsymbol{\beta}^{(m)} = \Big[\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}} + \vartheta\Lambda'\Lambda\Big]^{-1}\Big[\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{y}} + \vartheta\Lambda'\Big(\boldsymbol{\delta}^{(m-1)} - \vartheta^{-1}\boldsymbol{\nu}^{(m-1)}\Big)\Big];$$

   (b) update the value of δ_{ij} at the (m)th iteration by (5.1), after replacing

   $$\zeta_{ij} = \beta_i^{(m)} - \beta_j^{(m)} + \vartheta^{-1}\nu_{ij}^{(m-1)};$$

   (c) update the value ν_{ij} by

   $$\nu_{ij}^{(m)} = \nu_{ij}^{(m-1)} + \vartheta(\beta_i^{(m)} - \beta_j^{(m)} - \delta_{ij}^{(m)});$$

   (d) terminate the algorithm if the stopping rule $\|r^{(m)}\| < \epsilon$ is met at step m. Then, $(\boldsymbol{\beta}^{(m)}, \boldsymbol{\delta}^{(m)}, \boldsymbol{\nu}^{(m)})$ are the PAGFL estimates, denoted by $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\nu}})$.

Proposition 1 *The primal residual* $r^{(m)} = \Lambda\boldsymbol{\beta}^{(m)} - \boldsymbol{\delta}^{(m)}$ *and the dual residual* $s^{(m)} = \vartheta\Lambda(\boldsymbol{\beta}^{(m)} - \boldsymbol{\beta}^{(m-1)})$ *of the ADMM satisfy the following conditions:*

(i) $\lim_{m\to\infty} \|r^{(m)}\|^2 = 0$,

(ii) $\lim_{m\to\infty} \|s^{(m)}\|^2 = 0$.

Proposition 1 shows that both the primal and dual feasibility are achieved by the algorithm. Further, as the objective function in (2.6) is convex, the algorithm converges to an optimal point.

5.2 PGMM Computation

Similarly, by reparametrizing the objective function in (2.10), the minimization is equivalent to the constraint optimization problem below

$$\min S_2(\boldsymbol{\beta}, \boldsymbol{\delta}) = \frac{1}{2}\sum_{i=1}^{N}\Big[\sum_{t=1}^{T} z_{it}(\Delta y_{it} - \beta_i'\Delta x_{it})\Big]' W_{i,NT}\Big[\sum_{t=1}^{T} z_{it}(\Delta y_{it} - \beta_i'\Delta x_{it})\Big] + \lambda_2^* \sum_{1\leq i<j\leq N}\sum \ddot{w}_{ij}\|\delta_{ij}\|,$$

subject to $\beta_i - \beta_j - \delta_{ij} = 0$,

where $\boldsymbol{\delta} = \{\delta_{ij}, i < j\}'$, and $\lambda_2^* = T^2\lambda_2/2N$. By the augmented Lagrangian method, the estimates of the parameters can be obtained by minimizing

$$L_2(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\nu}) = S_2(\boldsymbol{\beta}, \boldsymbol{\delta}) + \sum_{1\le i<j\le N}\sum \nu'_{ij}(\beta_i - \beta_j - \delta_{ij}) + \frac{\vartheta}{2}\sum_{1\le i<j\le N}\sum \|\beta_i - \beta_j - \delta_{ij}\|^2,$$

where $\boldsymbol{\nu} = \{\nu'_{ij}, i < j\}'$ are lagrange multipliers and ϑ is the penalty parameter. We can obtain the estimates of $(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\nu})$ through iterations by the ADMM.

The iteration process consists of updating $\boldsymbol{\beta}, \boldsymbol{\delta}$ and $\boldsymbol{\nu}$ iteratively. For a given $(\boldsymbol{\delta}, \boldsymbol{\nu})$, we obtain the updates of $\boldsymbol{\beta}$ by setting the derivative $\partial L_2(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\nu})/\partial\boldsymbol{\beta}$ to zero, where

$$L_2(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\nu}) = \frac{1}{2}(\Delta\boldsymbol{y} - \Delta\boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{Z}(\Delta\boldsymbol{y} - \Delta\boldsymbol{X}\boldsymbol{\beta}) + \frac{\vartheta}{2}\|\Lambda\boldsymbol{\beta} - \boldsymbol{\delta} + \vartheta^{-1}\boldsymbol{\nu}\|^2 + C,$$

where C is a constant independent of $\boldsymbol{\beta}$, $\Delta\boldsymbol{y} = (\Delta y'_1, \ldots, \Delta y'_N)'$, $\Delta y_i = (\Delta y_{i1}, \ldots, \Delta y_{iT})'$, $\boldsymbol{Z} = diag(Z_1, \ldots, Z_N)$, $Z_i = (z_{i1}, \ldots, z_{iT})'$, $\Delta\boldsymbol{X} = diag(\Delta X_1, \ldots, \Delta X_N)$, $\Delta X_i = (\Delta x_{i1}, \ldots, \Delta x_{iT})'$ for each $i = 1, \ldots, N$, and $\boldsymbol{W} = diag(W_{1,NT}, \ldots, W_{N,NT})$. Moreover, we note that the minimizer of $L_2(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\nu})$ with respect to δ_{ij}, for given $(\boldsymbol{\beta}, \boldsymbol{\nu})$, has a closed form solution and is unique. In practice, for given $(\boldsymbol{\beta}, \boldsymbol{\nu})$, the minimizer problem with respect to δ_{ij} is equivalent to the following minimization

$$\frac{\vartheta}{2}\sum_{1\le i<j\le N}\sum \|\zeta_{ij} - \delta_{ij}\|^2 + \lambda_2^*\sum_{1\le i<j\le N}\sum \ddot{w}_{ij}\|\delta_{ij}\|,$$

where $\zeta_{ij} = \beta_i - \beta_j + \vartheta^{-1}\nu_{ij}$. Thus, the closed form solution is

$$\tilde{\delta}_{ij} = ST(\zeta_{ij}, \lambda_2^*/\vartheta). \tag{5.2}$$

Similarly, we track the progress of the ADMM based on the primal residual at step m, $r^{(m)} = \Lambda\boldsymbol{\beta}^{(m)} - \boldsymbol{\delta}^{(m)}$, and stop the algorithm when $\|r^{(m)}\| < \epsilon$. The algorithm can be summarized in below:

PGMM Algorithm:

1. **Initialization:** Find initial estimates of $\beta_i^{(0)}$ by minimizing the first term of (2.10) for all $i = 1, \ldots, N$. Let the initial values of $\boldsymbol{\nu}^{(0)} = 0$, and $\delta_{ij}^{(0)} = \beta_i^{(0)} - \beta_j^{(0)}$.

2. **Iterations:** At iteration $m \ge 1$, for given $\boldsymbol{\delta}^{(m-1)}$ and $\boldsymbol{\nu}^{(m-1)}$,

(a) update $\boldsymbol{\beta}^{(m)}$, which is the minimizer of $L_2(\boldsymbol{\beta}, \boldsymbol{\delta}^{(m)}, \boldsymbol{\nu}^{(m)})$, takes the form below

$$\boldsymbol{\beta}^{(m)} = \Big[\Delta \boldsymbol{X}' \boldsymbol{Z}' \boldsymbol{W} \boldsymbol{Z} \Delta \boldsymbol{X} + \vartheta \Lambda' \Lambda\Big]^{-1} \Big[\Delta \boldsymbol{X}' \boldsymbol{Z}' \boldsymbol{W} \boldsymbol{Z} \Delta \boldsymbol{y} + \vartheta \Lambda' \Big(\boldsymbol{\delta}^{(m-1)} - \vartheta^{-1} \boldsymbol{\nu}^{(m-1)}\Big)\Big];$$

(b) update the value of δ_{ij} at the (m)th iteration by (5.2), after replacing

$$\zeta_{ij} = \beta_i^{(m)} - \beta_j^{(m)} + \vartheta^{-1} \nu_{ij}^{(m-1)};$$

(c) update the value ν_{ij} by

$$\nu_{ij}^{(m)} = \nu_{ij}^{(m-1)} + \vartheta(\beta_i^{(m)} - \beta_j^{(m)} - \delta_{ij}^{(m)});$$

(d) terminate the algorithm if the stopping rule $\|r^{(m)}\| < \epsilon$ is met at step m. Then, $(\boldsymbol{\beta}^{(m)}, \boldsymbol{\delta}^{(m)}, \boldsymbol{\nu}^{(m)})$ are the PAGFL estimates, denoted by $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\delta}}, \tilde{\boldsymbol{\nu}})$.

Proposition 2 *The primal residual* $r^{(m)} = \Lambda \boldsymbol{\beta}^{(m)} - \boldsymbol{\delta}^{(m)}$ *and the dual residual* $s^{(m)} = \vartheta \Lambda (\boldsymbol{\beta}^{(m+1)} - \boldsymbol{\beta}^{(m)})$ *of the ADMM satisfy the following conditions:*

(i) $\lim_{m \to \infty} \|r^{(m)}\|^2 = 0$,

(ii) $\lim_{m \to \infty} \|s^{(m)}\|^2 = 0$.

Proposition 2 shows that both the primal and dual feasibility are achieved by the algorithm. Further, as the objective function in (2.10) is convex, the algorithm converges to an optimal point.

6 Monte Carlo Simulation

In this section, we investigate the finite sample performance of our PAGFL method. We consider six Monte Carlo experiments which are similar to those considered in Su et al. (2016) and Wang et al. (2018). The first three experiments consider data generating processes (DGPs) of static panel data models and deal with the PLS estimation. The fourth experiment is concerned with the PLS and PGMM estimation of dynamic panel data models. In this experiment, we focus on DGPs with a lagged dependent variable and multiple exogenous regressors. Finally, in the last two experiments, we consider both static and dynamic panel data DGPs where the number of group

structures is relatively large. For each experiment, we evaluate the performance of our PAGFL method by considering the following three criteria:

(i) Estimation Consistency: We report the Root Mean Squared Errors (RMSE), and the bias of the estimated regression coefficients, which are measured by

$$\text{RMSE } (\hat{\boldsymbol{\beta}}) = \sqrt{\frac{1}{Np}\sum_{i=1}^{N}\|\hat{\beta}_i - \beta_i^0\|^2}, \tag{6.1}$$

$$\text{Bias} = \sqrt{\frac{1}{Np}\sum_{i=1}^{N}\sum_{l=1}^{p}(\hat{\beta}_{i,l} - \beta_{i,l}^0)}. \tag{6.2}$$

(ii) Consistency of $\hat{K}$: We report the empirical percentage of selecting the true number of groups. That is, in our simulation designs we measure the percentage of the number of times $\hat{K} = K_0$.

(iii) Classification Consistency: To measure the similarity between the estimated grouping structure, $\hat{\mathcal{G}}$, and the true grouping structure, $\mathcal{G}^0$, similar to Ke et al. (2015) and Wang et al. (2018), we report the normalized mutual information (NMI) measure. The NMI measure for two classification $\mathcal{A} = \{A_1, A_2, \dots\}$, and $\mathcal{B} = \{B_1, B_2, \dots\}$, on the same set $\{1, \dots, N\}$, is defined as

$$\text{NMI}(\mathcal{A}, \mathcal{B}) = \frac{I(\mathcal{A}, \mathcal{B})}{\sqrt{H(\mathcal{A})H(\mathcal{B})}},$$

where

$$I(\mathcal{A}, \mathcal{B}) = \sum_{i,j}(|A_i \cap B_j|/N)\ln\Big(\frac{|A_i \cap B_j|/N}{(|A_i|/N)(|B_i|/N)}\Big) \text{ and } H(\mathcal{A}) = -\sum_{i}\frac{|A_i|}{N}\ln(\frac{|A_i|}{N}).$$

We note that $I(\mathcal{A}, \mathcal{B}) = H(\mathcal{A}) = H(\mathcal{B})$, when $\mathcal{A}$ and $\mathcal{B}$ have the same classification, and hence $\text{NMI}(\mathcal{A}, \mathcal{B}) = 1$. We report $\text{NMI}(\widehat{\mathcal{G}}, \mathcal{G}^0)$ for all experiments.

We select the tuning parameters λ_1 and λ_2 by minimizing the information criteria in (3.3) and (4.2), respectively. In practice, we search for the optimal tuning parameter on a 50 logarithmically spaced grids in the interval $[0, \lambda_{\max}]$, where $\lambda_{\max}$ is a tuning parameter that classifies all individuals in one group (i.e. the estimated number of groups is one). We find $\lambda_{\max}$ for each simulation by

trial and error. We choose $\rho_{1,NT} = \rho_{2,NT} = C_{NT}\ln(NT)/NT$ in (3.3) or (4.2) with $C_{NT} = c\sqrt{NT}$. We consider different values of c in our simulations, and the results suggest that the performance of our method is not sensitive to the choice of c, especially when N and/or T is large. Due to space limitation, we only report the results when $c = 0.7$, but the complete results are available in the Online Supplemental Appendix.

We set $\kappa = 2$ in the construction of the adaptive weights $\{\dot{w}_{ij} : i, j \in \{1, \ldots, N\}\}$ and $\{\ddot{w}_{ij} : i, j \in \{1, \ldots, N\}\}$, which are used for the PLS and PGMM estimations, respectively. The number of Monte Carlo simulations in all experiments is 200, and we consider all combinations of (N, T) with $N = \{50, 100, 200\}$ and $T = \{20, 40, 80\}$.

We consider the following DGPs:

- **DGP 1 (Static panel with two exogenous regressors):** The model is generated from the following static panel DGP with two exogenous regressors:

$$y_{it} = \beta_{i,1}x_{it,1} + \beta_{i,2}x_{it,2} + \mu_i + u_{it}, \quad i = 1, \ldots, N, \quad t = 1, \ldots, T,$$

  where the regressors $x_{it} = (x_{it,1}, x_{it,2})'$ are generated as $x_{it,1} = 0.2\eta_i + e_{it,1}$ and $x_{it,2} = 0.2\eta_i + e_{it,2}$ where $e_{it,1}$ and $e_{it,2}$ are both i.i.d. $\mathrm{N}(0, 1)$ and mutually independent. The fixed effects and the idiosyncratic errors follow the standard normal distribution and are mutually independent across i and t. The true number of groups is $K_0 = 3$, with the true coefficients

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

- **DGP 2:** Same as DGP 1, but $u_{it} \sim \mathrm{AR}(1)$. Specifically, for each i and t: $u_{it} = 0.5u_{i,t-1} + \epsilon_{it}$, with $\epsilon_{it} \sim$ i.i.d. $\mathrm{N}(0, 1)$.

- **DGP 3:** Same as DGP 1, but $u_{it} \sim \mathrm{GARCH}(1, 1)$. Specifically, for each i and t: $u_{it} = \sqrt{h_{it}}\epsilon_{it}$, $h_{it} = 0.05 + 0.05u_{i,t-1}^2 + 0.9h_{i,t-1}$, with $\epsilon_{it} \sim$ i.i.d. $\mathrm{N}(0, 1)$.

- **DGP 4 (Dynamic Panel AR(1) with two exogenous regressors):** The model is generated from the following equation

$$y_{it} = \beta_{i1}^0 y_{i,t-1} + \beta_{i2}^0 x_{it,1} + \beta_{i3}^0 x_{it,2} + \eta_i(1 - \beta_{i1}^0) + u_{it},$$

where the exogenous regressors $x_{it,1}$ and $x_{it,2}$ follow the standard normal distributions, mutually independent, and are independent of the error term. The initial values take the form $y_{i0} = \beta_{i2}^0 x_{it,1} + \beta_{i3}^0 x_{it,2} + \eta_i + u_{i0}$. The fixed effects and the idiosyncratic errors follow the standard normal distribution and are mutually independent across i and t. The true number of groups is $K_0 = 3$, with the true coefficients

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

- **DGP 5:** Same as DGP 1, but $K_0 = 8$ with the true coefficients

$$(\alpha_1^0, \alpha_2^0, \alpha_3^0, \alpha_4^0, \alpha_5^0, \alpha_6^0, \alpha_7^0, \alpha_8^0) = \left(\begin{bmatrix} -4 \\ 4 \end{bmatrix}, \begin{bmatrix} -3 \\ 3 \end{bmatrix}, \begin{bmatrix} -2 \\ 2 \end{bmatrix}, \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \begin{bmatrix} 2 \\ -2 \end{bmatrix}, \begin{bmatrix} 3 \\ -3 \end{bmatrix}, \begin{bmatrix} 4 \\ -4 \end{bmatrix} \right).$$

- **DGP 6:** Same as DGP 4, but $K_0 = 8$ with the true coefficients

$$(\alpha_1^0, \alpha_2^0, \alpha_3^0, \alpha_4^0, \alpha_5^0, \alpha_6^0, \alpha_7^0, \alpha_8^0) = \left(\begin{bmatrix} 0.8 \\ -4 \\ 4 \end{bmatrix}, \begin{bmatrix} 0.6 \\ -3 \\ 3 \end{bmatrix}, \begin{bmatrix} 0.4 \\ -2 \\ 2 \end{bmatrix}, \begin{bmatrix} 0.2 \\ -1 \\ 1 \end{bmatrix}, \begin{bmatrix} -0.2 \\ 1 \\ -1 \end{bmatrix}, \begin{bmatrix} -0.4 \\ 2 \\ -2 \end{bmatrix}, \begin{bmatrix} -0.6 \\ 3 \\ -3 \end{bmatrix}, \begin{bmatrix} -0.8 \\ 4 \\ -4 \end{bmatrix} \right).$$

The observations in DGPs 1–4 are drawn from three groups with the proportions $N_1/N = 0.4$, $N_2/N = 0.3$ and $N_3/N = 0.3$, and the observations in DGPs 5–6 are drawn with the proportions $N_1/N = 0.3$, $N_k/N = 0.1$ for $k = 2, \ldots, 8$.

The idiosyncratic error process in DGP 1 is strong white noise, while DGP 2 and DGP 3 allow for serial correlation and conditional heteroskedasticity, respectively. We apply PLS to estimate these models. In DGP 4, the model contains a lagged dependent variable. We use both PLS and PGMM to estimate the model, where the PGMM uses $(y_{i,t-2}, y_{i,t-3}, \Delta x_{it,2}, \Delta x_{it,3})$ as the instruments in the first-differenced model. For DGP 5, we use PLS to estimate the model, and use both PLS and PGMM to estimate the model in DGP 6. For DGP 6, the PGMM uses $(y_{i,t-2}, y_{i,t-3}, \Delta x_{it,2}, \Delta x_{it,3})$ as the instruments in the first-differenced model.

We note that when the group-specific parameters are not sufficiently separated from each other, or when the time period is relatively small that can cause the preliminary estimates to be quite different from the true parameter values, the PAGFL may produce trivial groups (groups that contain few individuals and it would be hard to decide whether these small groups are the correct groups or are generated due to misclassification). In such cases, following Park et al. (2007) and Wang et al. (2018), to eliminate the trivial groups, we consider hierarchical clustering where we allow the minimum number of observations within each group to be a certain percentage (e.g., 1% or 10%) of the total number of individuals. Furthermore, we noticed that classifying individuals i and j in a group when $\hat{\beta}_i = \hat{\beta}_j$ might be too stringent. Consequently, this may cause the method to produce trivial groups where the difference between the slope coefficients of the two groups is very small. Hence, to ensure that the individuals classified in different groups are sufficiently separated from each other, we classify individuals i and j in a group if $\|\hat{\beta}_i - \hat{\beta}_j\| \leq \epsilon_{tol}$ where ϵ_{tol} is a prescribed tolerance level (e.g., 0.001).

Table 1 reports simulation results of the selection consistency. It displays the empirical probability that a particular group size from 1 to 5 is estimated when the true number of groups is equal to three for DGPs 1–4. Additionally, it reports the empirical probability that a particular group size from 6 to 10 is estimated when the true number of groups is equal to eight for DGPs 5 and 6. We use the PLS estimation for DGPs 1, 2, 3, and 5. Since DGPs 4, and 6 are dynamic panels we report their results based on both the PLS and PGMM estimations.[4] In the following we summarize some important findings from these two tables. First, the simulations confirm that increasing N and/or T improves the selection consistency substantially, and this is true for all DGPs and both the PLS and PGMM estimations. Second, when T is small, the performance of our method is better for DGPS in which the degree of heterogeneity is larger, for instance comparing these empirical probabilities for DGP 1 and DGP 4, or the results of DGP 1 and DGP 5. Third, in DGP 6 where the model is dynamic and the number of groups is large, the PLS method performs better than the PGMM method. Fourth, for DGPs 2 and 3, where the errors, respectively, are serially correlated and conditionally heteroscedastic, the PLS performs better even at small N and T.

To measure the accuracy of classification, we report the normalized mutual information measure in Table 2. Both PLS and PGMM very accurately estimate group classification. As expected,

[4]To save space, we only report the PLS results in the main text. The PGMM results are available in the Online Supplemental Appendix.

when the sample size is large enough, and/or the difference between the slope parameters across the groups is relatively large, classification of PAGFL is accurate. In DGPs 4, and 6 PLS appears to be more accurate than PGMM. The results of classification accuracy is higher for DGPs 2 and 3 that contain serially correlated and conditionally heteroscedastic errors, relative to DGP 1 where the errors are independently identically distributed.

Table 3 provides the RMSE, and Bias of the proposed post-Lasso PAGFL, and the oracle estimator.[5] The PLS estimator for DGPs 4, and 6 is bias-corrected by using the Split-panel jackknife method of Dhaene and Jochmans (2015). The RMSEs of the PAGFL get close to the RMSEs of the oracle estimator when T increases. This demonstrates the practical relevance of the oracle property. For DGPs 4, and 6 where the model is dynamic, the PLS performs better than the PGMM.

7 Illustrations

We illustrate the PAGFL estimation and identification in an empirical application of unemployment dynamics at the U.S. state level.[6]

7.1 Unemployment Dynamics at the U.S. State Level

In this application, we apply the PAGFL estimation and identification procedure to a model of unemployment dynamics at the U.S. state level. Bun and Carree (2005) study this subject using a dynamic panel data model that relates each of the states' current unemployment rate (U_{it}) to the unemployment rate and economic growth rate (G_{it}) in the previous year. In addition to capture state specific effects, their model includes both state individual intercepts η_i, and time effect θ_t. Their model can be written as below

$$U_{it} = \gamma U_{i,t-1} + \beta G_{i,t-1} + \eta_i + \theta_t + \epsilon_{it}, \tag{7.1}$$

[5] We also compare the performance of PAGFL method with C-Lasso of Su et al. (2016) in the Online Supplemental Appendix. The results reveal that the PAGFL estimator generally outperforms the C-Lasso.

[6] We also illustrate the PAGFL estimation and identification in two additional empirical applications of a cost system of U.S. commercial banks, and forecasting output growth of 33 countries using macroeconomic and financial variables, in the Online Supplemental Appendix.

or equivalently

$$U_{it} - U_{i,t-1} = (\gamma - 1)(U_{i,t-1} - \alpha_i) + \beta(G_{i,t-1} - \delta) + \theta_t + \epsilon_{it}, \tag{7.2}$$

where $(1-\gamma)\alpha_i - \beta\delta = \eta_i$. The model in (7.2) shows that changes in unemployment rate are determined by two observable components: first, the adjustment of the unemployment rate toward a natural or equilibrium rate of unemployment, α_i, which is allowed to vary across states, second, the deviation of the economic growth rate around a constant equilibrium. In addition, in the model above, $1-\gamma$ denotes the speed of adjustment of the unemployment rate toward the natural or equilibrium rate. Further, it is expected to have $\beta < 0$, because a state that has relatively high economic growth is more likely to have reduced unemployment rates compared with states in which the economy is growing more slowly.

The model above imposes the assumption of heterogeneous intercepts and homogeneous slope coefficients across states, and as pointed out by Campello et al. (2019), estimation methods of such models can result in severely biased parameters and incorrect inferences. To avoid this issue, alternatively, we consider the following latent group structure model

$$U_{it} = \gamma_{g_i} U_{i,t-1} + \beta_{g_i} G_{i,t-1} + \eta_i + \epsilon_{it}, \tag{7.3}$$

where g_i denotes group membership of state i. The model above equivalently can be written as

$$U_{it} - U_{i,t-1} = (\gamma_{g_i} - 1)(U_{i,t-1} - \alpha_i) + \beta_{g_i}(G_{i,t-1} - \delta_{g_i}) + \epsilon_{it}. \tag{7.4}$$

The data for the unemployment rate are taken from the U.S. Bureau of Labor Statistics for 1976–2019 period, and the data for the states gross product are per capita personal income (thousands of dollars) which are obtained from the U.S. Bureau of Economic Analysis deflated by annual implicit price deflator.[7] The economic growth rate is taken to be the relative growth of the state product. Therefore, in our application $N = 51$, all U.S. states and Washington, DC, and $T = 43$ because year 1976 is taken as the starting observation.

The PAGFL divides the states in three groups, where the group memberships are presented in Table 4. Table 5 reports the estimated coefficient estimates based on full sample (ignoring

[7]Similar to Galvo and Kato (2014), we deflate the gross product data by the price deflator, hence we do not consider a time effect in the model.

parameter heterogeneity) and three groups with their corresponding standard deviations. All the estimated coefficients of γ are highly significant among the four models under 1% level. The value of γ in full sample and group 1 are almost the same and equal to 0.8, which implies an adjustment rate of around 20% per year. The adjustment rate in group 3 is smaller around 14% and that of group 2 is faster around 28%. The value of the full sample estimate of β equals -0.261, whereas the value of the estimate in group 1 and group 2 are -0.716, and -0.567 and all are significant under 1% level. This implies a somewhat stronger effect of economic growth on the change in unemployment than other states in group 3.

8 Conclusion

This paper introduces two simple and computationally efficient shrinkage procedures to jointly estimate and identify latent group structures in panel data via pairwise adaptive group fused Lasso penalties: PLS estimation for models without endogenous regressors, and PGMM estimation for models with endogeneity. Our proposed method does not require the knowledge of the true number of groups a priori, since the number of groups are estimated within the estimation procedure. This is a main advantages of our method relative to the existing methods in the literature. In addition, if information regarding the minimum number of individuals within each group is available, our method allows for hierarchical clustering to improve estimation accuracy. We develop the theoretical results and show that the proposed procedure can (1) consistently estimate the true group structure, hence (2) automatically and consistently estimate the true number of groups, (3) consistently estimate the regression coefficients. In addition, the PLS estimator asymptotically achieves the oracle property, but the PGMM oracle property is confined to some restrictive assumptions. Our proposed method is applicable to models where the number of groups is either fixed or divergent, thus our method can be applied to a large body of applications. We propose an ADMM algorithm to implement our procedure, and then derive the convergence properties of the ADMM algorithm. Monte Carlo simulations are conducted to examine the finite sample properties of the proposed method which show that the approach has good finite-sample performance. Our empirical application on the unemployment dynamics in the U.S. state level finds strong evidence that the slope coefficients are heterogenous and can be conveniently classified into three distinct groups.

References

Ando, T. and J. Bai (2016). Panel data models with grouped factor structure under unknown group membership. *Journal of Applied Econometrics* 31, 163–191.

Bai, J. (1997). Estimating multiple breaks one at a time. *Econometric Theory* 13, 315–352.

Baltagi, B. H., S. Garvin, S. Kerman, (1989). Further monte carlo evidence on seemingly unrelated regressions with unequal number of observations. *Annales D'Economie et de Statistique* 14, 103–15.

Baltagi, B. H. and Griffin, J. M. (1997). Pooled estimators vs. their heterogeneous counterparts in the context of dynamic demand for gasoline. *Journal of Econometrics* 77, 303–327.

Bester, C. A. and C. B. Hansen (2016). Grouped effects estimators in fixed effects models. *Journal of Econometrics.* 190, 197–208.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning* 3, 1–122.

Browning, M., And J. M. Carro (2007). Heterogeneity and microeconometrics modelling. *In Advances In Economics And Econometrics, Theory And Applications: Ninth World Congress Of The Econometric Society, Vol. 3, Ed. By R. Blundell, W. K. Newey, And T. Persson.* New York: Cambridge University Press, 45–74.

Bonhomme, S. and E. Manresa (2015). Grouped patterns of heterogeneity in panel data. *Econometrica* 83, 1147–1184.

Bun, M. J. G., and Carree, M. A. (2005). Bias-corrected estimation in dynamic panel data models. *Journal of Business & Economic Statistics* 23, 200–210.

Campello, M., Galvao, A. F., and Juhl, T. (2019). Testing for slope heterogeneity bias in panel data models. *Journal of Business & Economic Statistics* 37, 749–760.

Chi, E. C., and Lange, K. (2015). Splitting Methods for Convex Clustering. *Journal of Computational and Graphical Statistics* 24, 994–1013.

Deb, P. and P. K. Trivedi (2013). Finite mixture for panels with fixed effects. *Journal of Econometric Methods* 2, 35–51.

Dhaene, G., and Jochmans, K. (2015). Split-panel jackknife estimation of fixed-effect models. Review of Economic Studies. *Review of Economic Studies* 82, 991–1030.

Fan, J., Li, R., (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96, 1348–1360.

Galvao, A. F., and Kato, K. (2014). Estimation and Inference for Linear Panel Data Models Under Misspecification When Both n and T are Large. *Journal of Business & Economic Statistics* 32, 285–309.

Gourieroux, C., P. C. B. Phillips, And J. Yu (2010). Indirect inference for dynamic panel models. *Journal of Econometrics* 157, 68–77.

Gu, J. and S. Volgushev (2019). Panel data quantile regression with grouped fixed effects. *Journal of Econometrics* 213, 68–91.

Hahn, J., and G. Kuersteiner (2002). Asymptotically unbiased inference for a dynamic panel model with fixed effects when both n and t are large. *Econometrica* 70, 1639–1657.

Hahn, J., and H. R. Moon (2010). Panel data models with finite number of multiple equilibria. *Econometric Theory* 26, 863–881.

Han, C., P. C. B. Phillips and D. Sul (2014). X-Differencing and dynamic panel model estimation. *Econometric Theory* 30, 201–251.

Hoogstrate, A. J., F. C. Palm, G. A. Pfann, (2000). Pooling in dynamic panel-data models: An application to forecasting GDP growth rates. *Journal of business & Economic Studies* 18, 274–83.

Hsiao, C., And A. K. Tahmiscioglu (1997). A panel analysis of liquidity constraints and firm investment. *Journal of The American Statistical Association* 92, 455–465.

Huang, W., Jin, S., Su, L., (2020). Panel cointegration with latent group structures and an application to the PPP theory. *Econometric Theory* 36, 410–456.

Kasahara, H., And K. Shimotsu (2009). Nonparametric identification of finite mixture models of dynamic discrete choices. *Econometrica* 77, 135–175.

Ke, T., Fan, J., and Wu, Y. (2015). Homogeneity in regression. *Journal of the American Statistical Association* 110, 175–194.

Kiviet, J. F. (1995). On bias, inconsistency, and efficiency of various estimators in dynamic panel data models. *Journal of Econometrics* 68, 53–78.

Lee, Y. (2012). Bias in dynamic panel models under time series misspecification. *Journal of Econometrics* 169, 54–60.

Lin, C. C., and S. Ng (2012). Estimation of panel data models with parameter heterogeneity when group membership is unknown. *Journal of Econometric Methods* 1, 42–55.

Liu, R., Schick, A., Shang, Z., Zhang, Y., Zhou, Q., (2020). Identification and estimation in panel models with overspecified number of groups. *Journal of Econometrics* 2, 574-590.

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

Ma, S., Huang, J., (2017). A concave pairwise fusion approach to subgroup analysis. *Journal of the American Statistical Association* 517, 410–423.

Maddala, G. S., (1991). To pool or not to pool: that is the question. *Journal of Quantitative Economics* 7, 255–64.

Maddala, G. S., W. Hu, (1996). The pooling problem. In *The Econometrics of Panel Data.* Edited by L. Matyas and P. Sevestre. Advanced Studies in Theoretical and Applied Econometrics, vol 33. Springer, Dordrecht, pp. 307–22.

Park, M. Y., Hastie, T., and Tibshirani, R. (2007). Averaged gene expressions for regression. *Biostatistics* 8, 212–227.

Phillips, P. C. B., and D. Sul (2007). Bias in dynamic panel estimation with fixed effects, incidental trends and cross section dependence. *Journal of Econometrics* 137, 162–188.

Qian, J., and Su, L., (2015). Shrinkage estimation of regression models with multiple structural changes. *Econometric Theory* 32, 1376–1433.

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.

Sarafidis, V. and N. Weber (2015). A partially heterogeneous framework for analyzing panel data. *Oxford Bulletin of Economics and Statistics* 77, 274–296.

Su, L., And Q. Chen (2013). Testing homogeneity in panel data models with interactive fixed effects. *Econometric Theory* 29, 1079–1135.

Su, L. and G. Ju (2018). Identifying latent grouped patterns in panel data models with interactive fixed effects. *Journal of Econometrics* 206, 554–573.

Su, L., Z. Shi, and P. C. B. Phillips (2016). Identifying latent structures in panel. *Econometrica* 84, 2215–2264.

Su, L., Wang, X., Jin, S., (2019). Sieve estimation of time-varying panel data models with latent structures. *Journal of Business & Economics Statistics* 37, 334–349.

Sun, Y. (2005). Estimation And Inference In Panel Structure Models. *Working Paper, Dept. of Economics, UCSD.*

Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, And K. Knight (2005). Sparsity and smoothness via the fused lasso. *Journal of The Royal Statistical Society Series B*, 67, 91–108.

Wang, H., Li, R., and Tsai, C. L., (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94, 553–568.

Wang, W., Phillips, P.C., Su, L., (2018). Homogeneity pursuit in panel data models: Theory and application. *Journal of Applied Econometrics* 33, 797–815.

Wang, W., Su, L., (2021). Identifying latent group structures in nonlinear panels. *Journal of Econometrics* 2, 272–295

Yuan, M., And Y. Lin (2006). Model Selection And Estimation In Regression With Grouped Variables. *Journal of The Royal Statistical Society Series B*, 68, 49–67.

Zhang, C. (2010). Nearly Unbiased Variable Selection Under Minimax Concave Penalty. *The Annals of Statistics* 38, 894–942.

Zhang, Y., Li, R., Tsai, C. L. (2010). Regularization parameter selections via generalized information criterion. *Regularization parameter selections via generalized information criterion* 105, 312–323.

Zou, H., (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* 101, 1418–1429.

Table 1: Frequency of Selecting $K = 1, \ldots, 5$ Groups when $K_0 = 3$, and Selecting $K = 6, \ldots, 10$ Groups when $K_0 = 8$

		DGP 1					DGP 2					DGP 3				
N	T	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
50	20	0.00	0.02	**0.94**	0.04	0.00	0.00	0.00	**0.99**	0.02	0.00	0.00	0.00	**1.00**	0.01	0.00
50	40	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00
50	80	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00
100	20	0.00	0.00	**0.98**	0.03	0.00	0.00	0.00	**0.99**	0.02	0.00	0.00	0.00	**1.00**	0.01	0.00
100	40	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00
100	80	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00
200	20	0.00	0.00	**0.98**	0.02	0.00	0.00	0.00	**0.99**	0.01	0.00	0.00	0.00	**0.99**	0.01	0.00
200	40	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00
200	80	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00
		DGP 4					DGP 5					DGP 6				
N	T	1	2	3	4	5	6	7	8	9	10	6	7	8	9	10
50	20	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00	**0.95**	0.06	0.00	0.00	0.04	**0.92**	0.05	0.00
50	40	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00
50	80	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00
100	20	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00	**0.96**	0.04	0.00	0.00	0.00	**0.99**	0.02	0.00
100	40	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00
100	80	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00
200	20	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00
200	40	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00
200	80	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00

Note: This table reports the empirical probability of estimating a particular groups size from 1 to 5 when $K_0 = 3$, and 6 to 10 when $K_0 = 10$ via the PAGFL procedure. The results are based on the penalized least squares (PLS) method.

Table 2: NMI Measure of Correct Classification

N	T	DGP 1	DGP 2	DGP 3	DGP 4	DGP 5	DGP 6
50	20	0.82	0.84	0.94	1.00	0.99	0.99
50	40	0.97	0.98	1.00	1.00	1.00	1.00
50	80	1.00	1.00	1.00	1.00	1.00	1.00
100	20	0.81	0.84	0.93	0.99	0.99	0.99
100	40	0.96	0.97	0.99	1.00	1.00	1.00
100	80	1.00	1.00	1.00	1.00	1.00	1.00
200	20	0.81	0.83	0.92	0.99	0.99	0.99
200	40	0.96	0.97	0.99	1.00	1.00	1.00
200	80	1.00	1.00	1.00	1.00	1.00	1.00

Note: This table reports the NMI measure of classification accuracy. The results are based on the penalized least squares (PLS) method.

Table 3: Root Mean Squared Error and Bias of Coefficient Estimates

		DGP 1				DGP 2				DGP 3			
		PAGFL		Oracle		PAGFL		Oracle		PAGFL		Oracle	
N	T	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias
50	20	0.154	0.002	0.054	0.002	0.140	-0.003	0.052	-0.003	0.082	0.000	0.042	0.000
50	40	0.056	0.001	0.037	0.001	0.055	0.000	0.037	0.000	0.032	-0.001	0.029	-0.001
50	80	0.027	0.001	0.027	0.001	0.025	-0.001	0.025	-0.001	0.021	0.000	0.021	0.000
100	20	0.146	0.003	0.037	0.003	0.132	-0.002	0.036	-0.002	0.084	0.000	0.030	0.000
100	40	0.056	0.000	0.027	0.000	0.050	0.000	0.027	0.000	0.026	0.001	0.021	0.001
100	80	0.021	0.000	0.019	0.000	0.020	0.001	0.019	0.001	0.015	0.000	0.014	0.000
200	20	0.141	0.000	0.027	0.000	0.130	0.001	0.025	0.001	0.081	0.001	0.021	0.001
200	40	0.053	0.000	0.018	0.000	0.051	0.000	0.019	0.000	0.022	0.001	0.014	0.001
200	80	0.015	0.000	0.013	0.000	0.015	0.000	0.013	0.000	0.010	0.000	0.010	0.000
		DGP 4				DGP 5				DGP 6			
		PAGFL		Oracle		PAGFL		Oracle		PAGFL		Oracle	
N	T	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias
50	20	0.059	0.003	0.057	0.002	0.116	0.002	0.091	0.002	0.099	0.001	0.027	-0.001
50	40	0.037	0.001	0.037	0.001	0.064	0.002	0.064	0.002	0.075	0.001	0.075	0.002
50	80	0.025	0.000	0.025	0.000	0.046	-0.001	0.046	-0.001	0.052	0.001	0.052	0.000
100	20	0.044	0.004	0.037	0.004	0.091	-0.001	0.064	-0.001	0.083	-0.002	0.067	-0.001
100	40	0.024	0.001	0.024	0.001	0.045	-0.001	0.044	-0.001	0.046	0.001	0.046	0.001
100	80	0.018	0.001	0.018	0.001	0.031	0.000	0.031	0.000	0.032	0.001	0.032	0.001
200	20	0.033	0.003	0.026	0.003	0.078	-0.002	0.045	-0.002	0.061	0.000	0.043	0.000
200	40	0.018	0.002	0.018	0.002	0.033	0.000	0.032	0.000	0.030	0.000	0.030	0.000
200	80	0.012	0.000	0.012	0.000	0.022	0.000	0.022	0.000	0.021	0.000	0.021	0.000

Note: This table reports the root mean squared errors (RMSE) and Bias of the post-Lasso PAGFL, and the oracle estimator. The results are based on the penalized least squares (PLS) method.

Table 4: Group membership of the U.S. states in the Unemployment-Growth model

Group 1	Group 2	Group 3
Alabama	Alaska	Arkansas
Arizona	Kansas	Connecticut
California	Montana	Washington, DC
Colorado	New Mexico	Delaware
Florida	Oklahoma	Georgia
Hawaii	Texas	Iowa
Idaho	Utah	Illinois
Kentucky	Wyoming	Indiana
Louisiana		Massachusetts
Missouri		Maryland
Mississippi		Maine
North Dakota		Michigan
Nebraska		Minnesota
New Hampshire		North Carolina
Nevada		New Jersey
New York		Ohio
Washington		Oregon
West Virginia		Pennsylvania
		Rhode Island
		South Carolina
		South Dakota
		Tennessee
		Virginia
		Vermont
		Wisconsin

Table 5: Estimation results of the Unemployment-Growth Model

		PAGFL		
	Full Sample	Group 1	Group 2	Group 3
$\hat{\gamma}$	0.800^{***}	0.796^{***}	0.720^{***}	0.852^{***}
	(0.020)	(0.032)	(0.030)	(0.035)
$\hat{\beta}$	-0.261^{***}	-0.716^{***}	-0.567^{***}	0.028^{*}
	(0.011)	(0.018)	(0.017)	(0.019)

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