Journal of Econometrics 215 (2020) 574–590

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Identification and estimation in panel models with overspecified number of groups☆

Ruiqi Liu [a], Zuofeng Shang [a,b], Yonghui Zhang [c,*], Qiankun Zhou [d]

[a] Department of Mathematical Sciences, Indiana University-Purdue University Indianapolis, Indianapolis, IN 46202, United States of America
[b] Department of Mathematical Sciences, New Jersey Institute of Technology, Newark, NJ 07102, United States of America
[c] School of Economics and Institute of China's Economic Reform & Development, Renmin University of China, Beijing, 100872, China
[d] Department of Economics, Louisiana State University, Baton Rouge, LA 70803, United States of America

ARTICLE INFO

Article history:
Received 10 February 2018
Received in revised form 2 August 2019
Accepted 26 September 2019
Available online 12 November 2019

JEL classification:
C01
C13
C23
C33

Keywords:
Classification
Fixed effects
Group structure
K-means algorithm
Linear and nonlinear panel
M-estimation

ABSTRACT

We propose a simple and fast approach to identify and estimate the unknown group structure in panel models by adapting the M-estimation method. We consider both linear and nonlinear panel models where the regression coefficients are heterogeneous across groups but homogeneous within a group and the group membership is unknown to researchers. The main result of the paper is that under certain assumptions, our approach is able to provide uniformly consistent estimation as long as the number of groups used in estimation is not smaller than the true number of groups. We also show that, asymptotically, our method may partition some true groups into further subgroups, but cannot mix units from different groups. When the true number of groups is used in estimation, all units can be categorized correctly with probability approaching one, and we establish the limiting distribution for the estimators of the group parameters. In addition, we provide an information criterion to select the number of groups, and establish the consistency of the selection criterion under some mild conditions. Monte Carlo simulations are conducted to examine the finite sample performance of the proposed method. The findings in the simulation confirm our theoretical results in the paper. Applications to two real datasets also highlight the necessity to consider both individual heterogeneity and group heterogeneity in the model.

1. Introduction

Panel data models are widely used in empirical research of both economics and finance. A challenging problem in panel data models is how to control or model the individual-level heterogeneity. Unfortunately, most of the heterogeneity is

☆ The authors gratefully acknowledge the constructive comments on an earlier version of this paper from the Editor Oliver Linton, an associate editor, and two anonymous referees. We also acknowledge the helpful comments from Zongwu Cai, Qi Li, Liangjun Su, as well as the seminar participants at TAMU, Purdue, Xiamen, Jinan University and the Asian Meeting of Econometric Society at Seoul, Korea. We gratefully appreciate the contribution of Anton Schick to the earlier version of the paper. Shang's research is sponsored by the National Science Foundation (Grant No. DMS-1764280 and DMS-1821157). Zhang's research is sponsored by the National Natural Science Foundation of China (Grant No. 71401166, 71973141, and 71873033) and Beijing municipal fund for building world-class universities (disciplines) of Renmin University of China. Zhou's research is sponsored by the National Natural Science Foundation of China (Grant No. 71431006). Any remaining errors are the sole responsibility of the authors.

* Corresponding author.
E-mail address: yonghui.zhang@ruc.edu.cn (Y. Zhang).

https://doi.org/10.1016/j.jeconom.2019.09.008
0304-4076/

unobservable (e.g., willingness to pay for education, impact of economic policy, etc.). In practice, besides the commonly-used fixed effects to control the individual-level heterogeneity, there are two opposite approaches to deal with the heterogeneity in coefficients or parameters. The first one is to assume homogeneous coefficients across individuals (see, e.g., Lancaster, 2002; Hahn and Newey, 2004; Arellano and Bonhomme, 2009). Indeed, this approach reduces the model complexity and facilitates statistical inference. However, this common coefficients assumption might be too strong in practice and could lead to model misspecification (see, e.g., Hsiao and Tahmiscioglu, 1997; Lee et al., 1997). The other approach is to allow heterogeneous coefficients across individuals (see, e.g., Hsiao and Pesaran, 2008; Baltagi et al., 2008). This assumption helps avoid misspecification problem. However, it may lose the latent connections among individuals and the efficiency of estimation.

To allow such a possibility that a portion of the individuals share common coefficients, a mild and reasonable assumption is to impose group structures in panels. Group structures in panels refer to the regression coefficients that are the same within each group but differ across groups. Recently, group structures in panels have received considerable attention in the literature. To name a few, Lin and Ng (2012) provide two methods for estimating panel data models with group specific parameters when group membership is unknown. Under the same settings of Lin and Ng (2012), Sarafidis and Weber (2015) propose a modified k-means algorithm to determine the number of clusters and estimate the common coefficients. More recently, Su et al. (2016) propose a Classifier Lasso (C-Lasso) penalized procedure to identify and estimate nonlinear panels with latent group structures. Based on the seminal work of Su et al. (2016), Lu and Su (2017) study the determination of the number of groups in latent panel structure, Su and Ju (2018) identify the latent grouped pattern in panel models with interactive fixed effects, and Su et al. (2019) apply the C-Lasso to identify and estimate the time-varying panel data models with latent group structures. For nonparametric panel data models, Vogt and Linton (2019) develop a bandwidth-free method to identify and estimate the latent group structure of nonparametric regression curves. In addition, there is another strand of literature studying the group structure of individual fixed effects. For example, Bonhomme and Manresa (2015) consider a linear panel model with a latent group structure on the time-varying fixed effects and propose a "grouped fixed effects" estimator based on the k-means algorithm, and Bester and Hansen (2016) investigate the asymptotic properties of the group effects estimates of common parameters in nonlinear panel data models when the individual-specific fixed effects are assumed common across groups at some level. Further, Ando and Bai (2016) extend the work of Bonhomme and Manresa (2015) to linear panel data models with an unknown grouped factor structure.

Following the work by Lin and Ng (2012) and Su et al. (2016), this paper proposes a simple and straightforward method to identify and estimate panels with group structures when the true number of groups and the group membership are both unknown. The proposed method can be applied to both linear and nonlinear panels, and is computationally simple and fast. Besides the simplicity, our method has several advantages as follows.

First, the major theoretical contribution of this paper is that we show, under certain regularity conditions, that the consistency of our proposed estimation is independent of the number of groups used as long as this number is not less than the true number of groups.[1] A practical implication of this result is that a safe way to estimate the panel model with an unknown group structure is to set a slightly large number of groups. This is of crucial importance to researchers since the number of groups in the data is usually unknown. We also show that, asymptotically, our method can partition some true groups into further subgroups but cannot mis-classify individuals from different groups into the same group. When the true number of groups is used in estimation, all the individuals can be categorized correctly with probability approaching one.

Second, unlike the C-Lasso approach proposed by Su et al. (2016), relying on the choice of tuning parameters for estimation and classification, our approach is penalty-free if the number of groups is specified, which is a significant advantage for empirical applications. It is well known in the literature that Lasso-type methods are able to perform model selection consistently. However, the consistency highly relies on the right choice of the tuning parameters (e.g., Chand, 2012; Kirkland et al., 2015). Therefore, in empirical applications, the estimation results may be sensitive to the choice of tuning parameters, and how to choose the optimal tuning parameters in C-Lasso is still an open question. Consequently, it would be convenient to have a penalty-free approach to identify the group structures in panels, and our proposed method serves this purpose. Moreover, compared with C-Lasso, our method is computationally simple and much faster, especially when the model is nonlinear.

Finally, once the group membership is correctly identified and estimated, our proposed estimation performs similarly to the estimation based on the true group membership. This oracle property allows one to combine existing estimation and inference techniques with our method. For instance, for the classified group units, one can adapt the jackknife method in Hahn and Newey (2004) or Dhaene and Jochmans (2015) to reduce the bias of the fixed effects estimators in nonlinear panels.

The rest of the paper is organized as follows. In Section 2, we first introduce the fixed effects panel data model with an unknown group structure, and then propose an estimation and classification procedure. The asymptotic properties of our estimator are established in Section 3. Section 4 carries out a set of Monte Carlo simulations to investigate the finite sample performance of our method. Applications to two real datasets are provided in Section 5. The conclusion

[1] Moon and Weidner (2015) established a similar result for the linear panel data model with interactive fixed effects when the number of factors is overspecified.

is presented in Section 6. All mathematical derivations of main theorems and propositions are provided in Appendix A. An online Supplement contains the proofs of all relevant technical lemmas, additional simulation results and additional empirical studies.

Notation: For any squared matrix A, let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ be the smallest and largest eigenvalues of A. Let $\|B\|_2 = \sqrt{tr(BB')}$ denote the Frobenius norm for matrix B. Define a set of integers $[k] := \{1, 2, \ldots, k\}$ for any positive integer k. Let $(N, T) \to \infty$ denote N and T diverging to infinity jointly and $\xrightarrow{P}$, $\xrightarrow{D}$ denote convergence in probability and in distribution as $(N, T) \to \infty$, respectively.

2. Model and estimation

Let Y_{it} be a real-valued observation and $X_{it} \in \mathbb{R}^p$ be a real vector of observed covariates, both collected on the ith individual at time t for $i \in [N]$, $t \in [T]$. Assume that the N individuals actually belong to G^0 underlying groups where G^0 is unknown. In particular, $G^0 = 1$ corresponds to the traditional fixed effects model without group structures (see Hahn and Newey, 2004).

To identify the unknown group structures, a common practice is to predetermine the number of groups, denoted by G, and classify the N individuals into G groups. In practice, correctly specifying G (i.e., $G = G^0$) is difficult due to the unobservability of group patterns. A more realistic way is to pick G to be relatively large so that $G \geq G^0$. Obviously, such misspecification brings more challenges into the theoretical studies. In this paper, we propose a method for identifying group patterns under this misspecification and investigate its asymptotic properties.

For individual i, let $g_i \in [G]$ denote the group membership variable, $\beta_{g_i} \in \mathbb{K} \subset \mathbb{R}^p$ denote the unobservable group-specific parameter, and $\alpha_i \in \mathbb{A} \subset \mathbb{R}$ denote the unobservable individual-specific effect, with both $\mathbb{K}$ and $\mathbb{A}$ being compact. If individuals i and j belong to the same group, then $\beta_{g_i} = \beta_{g_j}$ (i.e., they share a common group parameter, but α_i and α_j might still be different due to individual-level heterogeneity). Let $\underline{\beta} = (\beta_1, \beta_2, \ldots, \beta_G) \in \mathbb{K}^G$ denote the tuple of G group-specific parameters, $\underline{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_N) \in \mathbb{A}^N$ denote the N-vector of individual parameters, and $\gamma_N = (g_1, g_2, \ldots, g_N) \in \Gamma_N$ denote the N-vector of group membership variables, where $\Gamma_N = [G]^N$ is the collection of all possible group assignments. Our aim is to estimate the triplet $\theta_N = (\underline{\beta}, \underline{\alpha}, \gamma_N)$, which can be achieved through the following M-estimation:

$$\widehat{\theta}_N = \underset{\theta_N = (\underline{\beta}, \underline{\alpha}, \gamma_N) \in \Theta_N}{\operatorname{argmax}} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \psi(X_{it}, Y_{it}, \beta_{g_i}, \alpha_i), \tag{2.1}$$

where $\Theta_N = \mathbb{K}^G \times \mathbb{A}^N \times \Gamma_N$ denotes the entire parameter space, $\psi(X_{it}, Y_{it}, \beta_{g_i}, \alpha_i)$ denotes the logarithm of the pseudo likelihood function of Y_{it} given X_{it} with parameters β_{g_i}, α_i. Here, we provide several examples of the explicit form of $\psi(X_{it}, Y_{it}, \beta_{g_i}, \alpha_i)$ for different panel models.

Example 2.1. Linear panel model: $Y_{it} = \beta_{g_i}' X_{it} + \alpha_i + \epsilon_{it}$, where ϵ_{it} represents the idiosyncratic error. In this case, one chooses $\psi(x, y, \beta, \alpha) = -(y - \beta' x - \alpha)^2$.

Example 2.2. Binary choice panel model: $Y_{it} = 1(\beta_{g_i}' X_{it} + \alpha_i \geq \epsilon_{it})$, where ϵ_{it} is the idiosyncratic error with common distribution function F, and $1(\cdot)$ denotes the indicator. In this case, the pseudo likelihood is $\psi(x, y, \beta, \alpha) = y \log F(\beta' x + \alpha) + (1-y) \log[1 - F(\beta' x + \alpha)]$. When F is the distribution function of standard normal (logistic), the above model becomes the probit (logit) model.

Example 2.3. Poisson panel model: Given X_{it} and (β_{g_i}, α_i), Y_{it} follows the Poisson distribution with mean $\exp(\beta_{g_i}' X_{it} + \alpha_i)$. In this case, we can choose $\psi(x, y, \beta, \alpha)$ as the logarithm of Poisson density function with mean $\exp(\beta' x + \alpha)$.

Unlike the penalized approach, such as the C-Lasso by Su et al. (2016), our M-estimation procedure (2.1) requires optimizing the objective function over the pre-regularized parameter space Θ_N where the parameters β_{g_i} intrinsically incorporate group constraint. This important feature avoids the delicate choice of penalty parameters as required by these penalization-based methods.

However, due to the complex structure of the parameter space Θ_N, it is challenging to directly solve (2.1). Instead, we introduce an efficient iterative algorithm. Before that, let us introduce some notation for future use:

$$\widehat{H}_i(\beta, \alpha) = \frac{1}{T} \sum_{t=1}^{T} \psi(X_{it}, Y_{it}, \beta, \alpha) \quad \text{and} \quad \widehat{\Psi}_N(\theta_N) = \widehat{\Psi}_N(\underline{\beta}, \underline{\alpha}, \gamma_N) = \frac{1}{N} \sum_{i=1}^{N} \widehat{H}_i(\beta_{g_i}, \alpha_i).$$

Here $\widehat{H}_i$ is the empirical pseudo likelihood function for individual i, and $\Psi_N(\theta_N)$ is the empirical pooled pseudo likelihood function considering the group variables. Consequently, (2.1) can be rewritten as follows:

$$\widehat{\theta}_N = (\widehat{\underline{\beta}}, \widehat{\underline{\alpha}}, \widehat{\gamma}_N) = \underset{\theta_N \in \Theta_N}{\operatorname{argmax}} \frac{1}{N} \sum_{i=1}^{N} \widehat{H}_i(\beta_{g_i}, \alpha_i), \tag{2.2}$$

where $\widehat{\underline{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_G)$, $\widehat{\underline{\alpha}} = (\widehat{\alpha}_1, \ldots, \widehat{\alpha}_N)$, and $\widehat{\gamma}_N = (\widehat{g}_1, \ldots, \widehat{g}_N)$.

We suggest the following iterative algorithm to solve (2.2):

(a) Choose the initial estimators $\underline{\beta}^{(0)} = (\beta_1^{(0)}, \ldots, \beta_G^{(0)})$.

(b) Update Group Membership: for each $i \in [N]$, in the $(s+1)$th iteration, find

$$g_i^{(s+1)} = \text{argmax}_{g\in[G]} \max_{\alpha\in\mathbb{A}} \widehat{H}_i(\beta_g^{(s)}, \alpha).$$

Then, update the group membership $\gamma_N^{(s+1)} = (g_1^{(s+1)}, \ldots, g_N^{(s+1)})$.

(c) Update Coefficients: given group membership $\gamma_N^{(s+1)}$, solve[2]

$$(\underline{\beta}^{(s+1)}, \underline{\alpha}^{(s+1)}) = \underset{\underline{\beta}\in\mathbb{K}^G}{\text{argmax}} \max_{\underline{\alpha}\in\mathbb{A}^N} \widehat{\Psi}_N(\underline{\beta}, \underline{\alpha}, \gamma_N^{(s+1)}).$$

(d) Repeat steps (b)-(c) until $\widehat{\Psi}_N(\underline{\beta}^{(s+1)}, \underline{\alpha}^{(s+1)}, \gamma_N^{(s+1)}) = \widehat{\Psi}_N(\underline{\beta}^{(s)}, \underline{\alpha}^{(s)}, \gamma_N^{(s)})$ or some tolerance criterion is met.[3]

The above algorithm is essentially a modified k-means algorithm. Bonhomme and Manresa (2015) also proposed a similar algorithm to study the group patterns of fixed effects in a linear panel. As discussed by Bonhomme and Manresa (2015), the modified k-means algorithm is sensitive to the choice of initial estimators $\underline{\beta}^{(0)}$, and they recommended trying different initial values. For our algorithm, we propose the following two strategies to obtain initial estimators:

Strategy 1 Randomly generate $\underline{\beta}^{(0)} = (\beta_1^{(0)}, \ldots, \beta_G^{(0)})$ around the usual fixed effects estimators to obtain the initial estimators. For this strategy, we firstly treat all the individuals as a group and compute the fixed effects estimator $\beta^* = (\beta_1^*, \ldots, \beta_p^*) \in \mathbb{R}^p$, and then, for each $g \in [G]$, we generate $\beta_g^{(0)} = \beta^* + r\delta_g$, where $r > 0$ is some prespecified constant and δ_g is a centered p-dimensional normal random vector whose variance matrix is diagonal with diagonal elements $|\beta_1^*|^2, \ldots, |\beta_p^*|^2$.[4]

Strategy 2 For each $i \in [N]$, let $\widehat{\beta}_i^{\text{ML}}$ and $\widehat{\alpha}_i^{\text{ML}}$ be the pseudo maximum likelihood estimators of β_i^0 and α_i^0 based on $\{X_{it}, Y_{it}\}_{t=1}^T$, namely, $(\widehat{\beta}_i^{\text{ML}}, \widehat{\alpha}_i^{\text{ML}}) = \text{argmax}_{\beta\in\mathbb{K},\alpha\in\mathbb{A}} \widehat{H}_i(\beta, \alpha)$. Next, apply the standard k-means algorithm with G groups to $(\widehat{\beta}_1^{\text{ML}}, \ldots, \widehat{\beta}_N^{\text{ML}})$ and denote the estimated groups as $(\beta_1^{(0)}, \ldots, \beta_G^{(0)})$. Finally, set $\underline{\beta}^{(0)} = (\beta_1^{(0)}, \ldots, \beta_G^{(0)})$ to be the initial estimators for iteration.

In the simulation studies and empirical applications, we try 21 different initial estimators, of which 20 are randomly generated by Strategy 1 and one is generated by Strategy 2. After implementing the proposed algorithm using different initial estimators, we choose the one leading to the largest pseudo likelihood. The simulation results show that our proposed procedure works very well and is also quite fast even when multiple initial values are used.[5]

3. Asymptotic theory

In this section, we prove several asymptotic results, such as estimation consistency, classification consistency for using an overspecified G, and asymptotic normality for the estimators obtained with a correctly specified G. We also provide a procedure to determine the number of groups consistently. Throughout this section, let $\theta_N^0 = (\underline{\beta}^0, \underline{\alpha}^0, \gamma_N^0)$ denote the true parameters under which the observations X_{it} and Y_{it} are generated, where $\underline{\beta}^0 = (\beta_1^0, \beta_2^0, \ldots, \beta_{G^0}^0)$, $\underline{\alpha}^0 = (\alpha_1^0, \alpha_2^0, \ldots, \alpha_N^0)$, and $\gamma_N^0 = (g_1^0, g_2^0, \ldots, g_N^0)$. Moreover, we denote $\mathcal{D}$ as the σ-field generated by the fixed effects $\{\alpha_1^0, \alpha_2^0, \ldots, \alpha_N^0\}$ and define $P_{\mathcal{D}}(\cdot)$, $E_{\mathcal{D}}(\cdot)$ to be the conditional probability and expectation given $\mathcal{D}$, respectively. Using the above notations, we define $H_i(\beta, \alpha) = E_{\mathcal{D}}(\widehat{H}_i(\beta, \alpha))$ and $\Psi_N(\theta_N) = \Psi_N(\underline{\beta}, \underline{\alpha}, \gamma_N) = \frac{1}{N}\sum_{i=1}^N H_i(\beta_{g_i}, \alpha_i)$, which are the population counterparts of $\widehat{H}_i$'s and $\widehat{\Psi}_N$, respectively.

3.1. Estimation consistency

The main result of this section is to show that the proposed M-estimation is consistent. Before stating our main theorem, let us introduce some technical conditions. To start, for each $g \in [G^0]$, we define $N_g = \sum_{i=1}^N 1(g_i^0 = g)$ (i.e., the true number of individuals from group g).

---

[2] The optimization problem can be solved as follows. First, divide all individuals into G groups based on $\gamma_N^{(s+1)}$. Next, notice for individuals in the gth group, it is a fixed effects model (without group structures) and can be solved efficiently by existing R packages (e.g., speedglm package, see Fernández-Val and Weidner, 2016). The corresponding estimators based on different groups are denoted as $\beta_g^{(s+1)}$ for $g \in [G]$ and $\alpha_i^{(s+1)}$ for $i \in [N]$. Finally, set $\underline{\beta}^{(s+1)} = (\beta_1^{(s+1)}, \ldots, \beta_G^{(s+1)})$ and $\underline{\alpha}^{(s+1)} = (\alpha_1^{(s+1)}, \ldots, \alpha_N^{(s+1)})$.

[3] The criterion we used to stop iteration is that the difference between $\underline{\beta}^{(s+1)}$ and $\underline{\beta}^{(s)}$ in Euclidean norm is less than 10^{-6}. For the above proposed modified k-means algorithm, a proof of local convergence is provided in the online Supplement S.III.

[4] Even though there is no golden principle to choose the random initial estimators, the recommended procedure with $r = 1$ performs considerably well in both simulation and empirical studies.

[5] For the robustness and computational time of the proposed strategies, we defer the simulation results to the online Supplement S.IV for additional simulation evidences.

Assumption A1.

(a) $\{(X_{it}, Y_{it}), t \in [T]\}$ are mutually conditionally independent of each other across $i \in [N]$ given $\mathcal{D}$.

(b) For each $i \in [N]$, the process $\{X_{it}, Y_{it} : t \in [T]\}$ is conditionally stationary and α-mixing given $\mathcal{D}$ with conditional mixing coefficients $\alpha_{[i]}(\cdot)$. Moreover, $\alpha(\tau) := \sup_{N\geq 1} \max_{1\leq i\leq N} \alpha_{[i]}(\tau)$ satisfies $\alpha(\tau) \leq \exp(-C_0\tau^{b_0})$ almost surely for all $\tau \geq 0$ and some constants $C_0, b_0 > 0$.

(c) For each $i \in [N]$, $H_i(\beta, \alpha)$ is uniquely maximized at $(\beta^0_{g_i^0}, \alpha_i^0)$. Furthermore, the following identification condition holds almost surely:

$$\chi(\epsilon) := \inf_{N\geq 1} \inf_{1\leq i\leq N} \inf_{\|\beta-\beta^0_{g_i^0}\|_2^2+|\alpha-\alpha_i^0|^2\geq\epsilon} [H_i(\beta^0_{g_i^0}, \alpha_i^0) - H_i(\beta, \alpha)] > 0, \qquad \text{for all } \epsilon > 0.$$

(d) $d_0 := \inf_{\tilde{g}\neq g} \|\beta^0_g - \beta^0_{\tilde{g}}\|_2 > 0$.

(e) There exists a non-negative function $Q(x, y)$ such that for all $(\beta, \alpha), (\check{\beta}, \check{\alpha}) \in \mathbb{K} \times \mathbb{A}$,

$$|\psi(x, y, \beta, \alpha) - \psi(x, y, \check{\beta}, \check{\alpha})| \leq Q(x, y)(\|\beta - \check{\beta}\|_2^2 + |\alpha - \check{\alpha}|^2)^{1/2},$$

and $|\psi(x, y, \beta, \alpha)| \leq Q(x, y)$ for all $(\beta, \alpha) \in \mathbb{K} \times \mathbb{A}$. Furthermore, there exist $b_1 \in (0, \infty]$ and $B_1 > 0$ such that the following inequality holds almost surely:

$$\sup_{N\geq 1} \sup_{1\leq i\leq N} P_{\mathcal{D}}(Q(X_{i1}, Y_{i1}) > v) \leq \exp\left(1 - (v/B_1)^{b_1}\right), \qquad \text{for all } v > 0.$$

(f) For all $g \in [G^0]$, there exists a positive constant π_g such that $N_g/N \to \pi_g$ as $N \to \infty$.

Remark 3.1. Assumption A1.(a) assumes conditional cross-sectional independence among the individuals given $\mathcal{D}$. Assumption A1.(b) imposes conditional weak dependence for the observations along time with the level of dependence controlled by an exponential bound with parameter b_0. The conditional stationarity assumption can be relaxed at the cost of introducing extra notations. The conditional stationarity and α-mixing conditions are also used in Su and Chen (2013) and Fernández-Val and Weidner (2016). Assumption A1.(c) is a regularity condition for identification, which can be verified case by case under certain mild conditions. Assumption A1.(d) assumes that the pairwise differences between the group parameters are bounded from zero. This condition is needed to ensure the identification of the group parameters. Similar conditions are also assumed by Bonhomme and Manresa (2015) and Su et al. (2016). Assumption A1.(e) states that ψ is smooth and satisfies certain exponential tail conditions with the decay rate of the tail probability characterized by b_1. When ψ is a bounded function, we can choose $B_1 = 2\|\psi\|_\infty$ and $b_1 = \infty$. Similar tail conditions are also assumed by Bonhomme and Manresa (2015) for the error term. Compared with other conditions, such as finite moment assumptions on Q, the exponential tail condition can lead to better convergence results and is still valid for many commonly-used models, such as Examples 2.1 and 2.3. Assumption A1.(f) excludes the groups with ignorable proportions. This condition is standard and necessary for panel models with a finite number of groups (e.g., see Bonhomme and Manresa, 2015; Su et al., 2016).

Let $d = b_0b_1/(b_0+b_1)$. Since b_0 and b_1 characterize the weak dependence of the observations and decay rate of the tail probability, respectively (see discussion in Remark 3.1), d can be viewed as a quantity jointly controlling both. A special case is $b_1 = \infty$ for bounded ψ, and we have $d = b_0$.

Assumption A2. $\log N = o(T^{\frac{d}{1+d}})$.

Remark 3.2. Compared with the standard assumption on the rate of N and T in the literature where the ratio of T/N being a nonzero constant (e.g., Hahn and Newey, 2004), Assumption A2 is a relatively weak condition since Assumption A2 allows N to diverge exponentially faster than T, where the ratio of T/N approaches to zero. Moreover, Assumption A2 is also quite reasonable in practice since most microeconomics datasets are with moderately large T and large N.

To prove the consistency of $\widehat{\theta}_N$, we introduce the following pseudo metric d_N on Θ_N. For any $\theta_N = (\underline{\beta}, \underline{\alpha}, \gamma_N), \widetilde{\theta}_N = (\widetilde{\underline{\beta}}, \widetilde{\underline{\alpha}}, \widetilde{\gamma}_N) \in \Theta_N$, we define:

$$d_N(\theta_N, \widetilde{\theta}_N) = \frac{1}{N}\sum_{i=1}^{N}\left(\|\beta_{g_i} - \widetilde{\beta}_{\widetilde{g}_i}\|_2 + |\alpha_i - \widetilde{\alpha}_i|\right).$$

Specifically, $d_N(\theta_N, \widetilde{\theta}_N)$ measures the average discrepancy of (β_{g_i}, α_i)'s and $(\widetilde{\beta}_{\widetilde{g}_i}, \widetilde{\alpha}_i)$'s. Theorem 1 proves consistency for $\widehat{\theta}_N$ under this pseudo metric as well as uniform consistency of $\widehat{\beta}_{\widehat{g}_i}$'s.

Theorem 1. *Suppose $G \geq G^0$ and Assumptions A1 and A2 hold. Then, it follows that*

$$d_N(\widehat{\theta}_N, \theta_N^0) \xrightarrow{P} 0 \quad \text{and} \quad \sup_{1\leq i\leq N} \|\widehat{\beta}_{\widehat{g}_i} - \beta^0_{g_i^0}\|_2 \xrightarrow{P} 0 \quad \text{as } (N, T) \to \infty.$$

Theorem 1 establishes two consistency results of the proposed estimator, namely, consistency on the average for $\widehat{\theta}_N$ and uniform consistency for $\widehat{\beta}_{\widehat{g}}$ under the assumption $G \geq G^0$. However, if $G < G^0$, the above results will be invalid since, in this scenario, some individuals from different groups need to be classified into a same group, and this will lead to inconsistency.

3.2. Detection of group structure among individuals

Detection of group structures in panel data is a fundamental problem in panels with group structure. The popular C-Lasso approach recently proposed by Su et al. (2016) requires the use of penalty for effectively classifying the individuals. In this section, we study our penalty-free grouping method and investigate its asymptotic property if G is pre-specified. Our theory and method are valid under $G \geq G^0$.

Recall that $\widehat{\gamma}_N = (\widehat{g}_1, \widehat{g}_2, \ldots, \widehat{g}_N)$ is the estimator of the group membership variables obtained in (2.2). Our grouping method is simply based on $\widehat{g}_i$'s as follows. For $g \in [G]$, define $\widehat{\mathcal{C}}_g = \{i \in [N] : \widehat{g}_i = g\}$ (i.e., $\widehat{\mathcal{C}}_g$ is the collection of the individuals belonging to the gth estimated group). We also define $\mathcal{C}_g^0 = \{i \in [N] : g_i^0 = g\}$ for $g \in [G^0]$ (i.e., $\mathcal{C}_g^0$ is the population analogy based on the true group membership variables). It is necessary to provide the conditions under which such a simple grouping method is valid, that is, for any $g \in [G]$, a $\widetilde{g} \in [G^0]$ exists such that $\widehat{\mathcal{C}}_g \subseteq \mathcal{C}_{\widetilde{g}}^0$ with probability approaching one. A formal statement of this result is provided in Theorem 2. Such a property implies that the individuals are correctly grouped.

To prove this result, we need stronger assumptions on the smoothness of ψ. To deal with partial derivatives of a multivariate function, we introduce the following multi-index notation. Let $\mathbf{k} = (k_1, k_2, \ldots, k_{p+1})$ denote a multi-index, where k_l's are non-negative integers. For any $\beta \in \mathbb{K} \subset \mathbb{R}^p$, denote $\beta = (\beta_{[1]}, \beta_{[2]}, \ldots, \beta_{[p]})$, where $\beta_{[l]}$ is the lth coordinate of β. Define the $\mathbf{k}$th order partial derivative of $\psi(x, y, \beta, \alpha)$ with respect to β, α as follows:

$$D^{\mathbf{k}}\psi(x, y, \beta, \alpha) = \frac{\partial^{|\mathbf{k}|}\psi(x, y, \beta, \alpha)}{\partial\beta_{[1]}^{k_1} \ldots \partial\beta_{[p]}^{k_p}\partial\alpha^{k_{p+1}}},$$

where $|\mathbf{k}| = k_1 + k_2 + \cdots + k_{p+1}$. We also denote the Hessian of ψ and H_i (with respect to β, α) by

$$\ddot{\psi}(x, y, \beta, \alpha) = \begin{pmatrix} \frac{\partial^2\psi(x,y,\beta,\alpha)}{\partial\beta\partial\beta'} & \frac{\partial^2\psi(x,y,\beta,\alpha)}{\partial\beta'\partial\alpha} \\ \frac{\partial^2\psi(x,y,\beta,\alpha)}{\partial\beta\partial\alpha} & \frac{\partial^2\psi(x,y,\beta,\alpha)}{\partial\alpha^2} \end{pmatrix}, \quad \ddot{H}_i(\beta, \alpha) = E_{\mathcal{D}}(\ddot{\psi}(X_{i1}, Y_{i1}, \beta, \alpha)).$$

We require the following conditions on the partial derivatives of ψ and Hessian of H_i's. Let $\mathcal{B}_i = \{(\beta, \alpha) \in \mathbb{K} \times \mathbb{A} : \|\beta - \beta_{g_i^0}^0\|_2 + |\alpha - \alpha_i^0| \leq a_0\}$ for $i \geq 1$, and $\mathcal{B} = \cup_{i\geq 1}\mathcal{B}_i$.

Assumption A3.

(a) There exist a function $J(x, y)$, a constant $a_0 > 0$ and an integer $q_0 \geq 4$ such that, for any $\mathbf{k}$ with $|\mathbf{k}| \leq 4$ and $(\beta, \alpha) \in \mathcal{B}$,

$$|D^{\mathbf{k}}\psi(x, y, \beta, \alpha)| \leq J(x, y) \quad \text{and} \quad \sup_{N\geq 1}\sup_{1\leq i\leq N} E_{\mathcal{D}}J^{q_0}(X_{i1}, Y_{i1}) < \infty, \text{ almost surely.}$$

(b) $\sup_{N\geq 1}\sup_{1\leq i\leq N}\lambda_{\max}(\ddot{H}_i(\beta_{g_i^0}^0, \alpha_i^0)) < 0$ almost surely.

Remark 3.3. Assumption A3.(a) requires higher-order smoothness and existence of finite q_0th moment on the derivatives of pseudo likelihood function ψ to guarantee the correct classification. A similar assumption has been made by Hahn and Newey (2004) and Su et al. (2016). Assumption A3.(b) requires the largest eigenvalues of Hessian matrices at the true values are uniformly bounded away from zero almost surely, which implies that Hessian matrices of the expected objective function are uniformly negative definite. This assumption is similar to the conditions on the Hessian matrices of the profiled pseudo likelihood function in Su et al. (2016). Note that both Assumptions A1.(c) and A3.(b) are imposed on the population counterpart of pseudo likelihood function. They are both standard regularity conditions for nonlinear panels, where the former is the uniqueness condition for identification of parameters of interest and the latter is a local condition to study the asymptotic distribution of the estimators. In general, neither of these two assumptions is nested in the other. However, if $H_i(\beta, \alpha)$'s are convex and $(\beta_{g_i^0}^0, \alpha_i^0)$'s are the local maximizers, Assumption A3.(b) implies Assumption A1. (c).

Below is the main theorem which provides the classification consistency of our grouping method under $G \geq G^0$.

Theorem 2. *Suppose $G \geq G^0$ and Assumptions A1–A3 hold. Then, for each $g \in [G]$, there exists a $\widetilde{g} \in [G^0]$ such that* $\lim_{(N,T)\to\infty} P\left(\widehat{\mathcal{C}}_g \subseteq \mathcal{C}_{\widetilde{g}}^0\right) = 1$.

Remark 3.4. Theorem 2 demonstrates that the proposed grouping method is valid under misspecification in the sense that, with probability approaching one, any grouped individuals asymptotically belong to a population group. This implies that any population group is either identical to a selected group or is partitioned into subgroups without any misclassification, which is possibly the best result one can expect under $G \geq G^0$. In the special case $G = G^0$, Theorem 2 naturally leads to classification consistency (i.e., up to a proper relabeling, $\widehat{\mathcal{C}}_g = \mathcal{C}_g^0$ for all $g \in [G^0]$ with probability approaching one). When $G = G^0$ the classification consistency is also established by Su et al. (2016).

Remark 3.5. Intuitively, Theorem 2 implies that, under Assumptions A1–A3 and if $G > G^0$, then asymptotically, (i) individuals from the same group may be divided into different subgroups; (ii) individuals from different groups cannot be categorized into the same group.

Remark 3.6. The implication of Theorem 2 is that, since the true number of groups is unknown in practice, it is safe to use a relatively large number of groups to classify the data and to obtain consistent estimation. Otherwise, if $G < G^0$, neither the estimation nor the classification is consistent.

3.3. Asymptotic normality

In this section, we study the asymptotic normality of $\widehat{\underline{\beta}}$ under $G = G^0$. First, we introduce the "oracle" estimator $\widetilde{\underline{\beta}}$ of $\underline{\beta}$ when the true group assignment γ_N^0 were known. Let

$$\widetilde{\underline{\beta}} = \underset{\underline{\beta} \in \mathbb{K}^{G^0}}{\operatorname{argmax}} \max_{\alpha_i \in \mathbb{A}} \frac{1}{N} \sum_{i=1}^{N} \widehat{H}_i(\beta_{g_i^0}, \alpha_i).$$

Clearly, $\widetilde{\underline{\beta}}$ is infeasible since γ_N^0 is practically unavailable. In fact, $\widetilde{\underline{\beta}}$ and $\widehat{\underline{\beta}}$ are asymptotically equivalent if $G = G^0$, as shown in the following proposition.

Proposition 1. *Suppose $G = G^0$ and Assumptions A1–A3 hold, then*

$$\lim_{(N,T)\to\infty} P(\widehat{\underline{\beta}} = \widetilde{\underline{\beta}}) = 1.$$

The above proposition roots from Theorem 2 when $G = G^0$. In this case, we can recover the true latent group structure with probability approaching one and, asymptotically, there is no difference between the oracle estimator and our proposed estimator. Thus, to derive the asymptotic normality of $\widehat{\underline{\beta}}$, it is sufficient to derive the asymptotic normality of $\widetilde{\underline{\beta}}$, i.e., given γ_N^0. To this end, we make an additional assumption, Assumption A4. Before that, let us define

$$\rho_i = E_{\mathcal{D}}^{-1}\left(\frac{\partial^2 \psi}{\partial \alpha^2}(X_{i1}, Y_{i1}, \beta_{g_i^0}^0, \alpha_i^0)\right) E_{\mathcal{D}}\left(\frac{\partial^2 \psi}{\partial \beta \partial \alpha}(X_{i1}, Y_{i1}, \beta_{g_i^0}^0, \alpha_i^0)\right),$$

$$U_i(x, y, \beta, \alpha) = \frac{\partial \psi}{\partial \beta}(x, y, \beta, \alpha) - \rho_i \frac{\partial \psi}{\partial \alpha}(x, y, \beta, \alpha), \quad R_i(x, y, \beta, \alpha) = \frac{\partial \psi}{\partial \alpha}(x, y, \beta, \alpha),$$

$$V_i(x, y, \beta, \alpha) = \frac{\partial U_i}{\partial \beta'}(x, y, \beta, \alpha), \quad \mathcal{I}_i = E_{\mathcal{D}}(V_i(X_{i1}, Y_{i1}, \beta_{g_i^0}^0, \alpha_i^0)).$$

The above notation is standard in the literature of nonlinear panel models (e.g., Hahn and Newey, 2004; Arellano and Hahn, 2007). For notational simplicity, we denote $U_i^{\alpha} = \partial U_i / \partial \alpha$, $U_i^{\alpha\alpha} = \partial^2 U_i / \partial \alpha^2$, $U_{it} = U_i(X_{it}, Y_{it}, \beta_{g_i^0}^0, \alpha_i^0)$, and $U_{it}^{\alpha} = U_i^{\alpha}(X_{it}, Y_{it}, \beta_{g_i^0}^0, \alpha_i^0)$. We also define R_{it} and R_{it}^{α} analogically. For each $i \in [N]$, let Λ_i denote the asymptotic conditional covariance matrix of $\sum_{t=1}^{T} U_{it}/\sqrt{T}$ as $T \to \infty$, which has the following expression:

$$\Lambda_i = \lim_{T\to\infty} \frac{1}{T} E_{\mathcal{D}}\left(\left[\sum_{t=1}^{T} U_{it}\right]\left[\sum_{t=1}^{T} U_{it}'\right]\right) = E_{\mathcal{D}}(U_{i1}U_{i1}') + 2\sum_{t=1}^{\infty} E_{\mathcal{D}}(U_{i1}U_{i,1+t}').$$

Convergence of the above series holds uniformly for i under Assumptions A1 and A3.

Assumption A4.

(a) There exists a constant $B_3 \in (0, 1)$ such that

$$B_3 \leq \inf_{N\geq 1} \inf_{1\leq i\leq N} \lambda_{\min}(\Lambda_i) \leq \sup_{N\geq 1} \sup_{1\leq i\leq N} \lambda_{\max}(\Lambda_i) \leq 1/B_3 \text{ almost surely.}$$

Moreover, for each $g \in [G^0]$, there exist a strictly positive definite matrix D_g and a strictly negative definite matrix W_g such that the following convergence results hold almost surely

$$\lim_{N\to\infty} \sum_{i:g_i^0=g} \Lambda_i/N_g = D_g \quad \text{and} \quad \lim_{N\to\infty} \sum_{i:g_i^0=g} \mathcal{I}_i/N_g = W_g.$$

(b) For each $g \in [G^0]$, there exists a vector $\Delta_g \in \mathbb{R}^p$ such that the following convergence holds almost surely:

$$\lim_{(N,T)\to\infty} \frac{1}{N_g} \sum_{i:g_i^0=g} E_{\mathcal{D}}\left\{ \left(\frac{\sum_{t=1}^T R_{it}}{\sqrt{T} E_{\mathcal{D}}(R_{i1}^\alpha)} \right) \left(\frac{1}{\sqrt{T}} \sum_{t=1}^T [U_{it}^\alpha - \frac{E_{\mathcal{D}}(U_{i1}^{\alpha\alpha})}{2E_{\mathcal{D}}(R_{it}^\alpha)} R_{it}] \right) \right\} = \Delta_g.$$

Remark 3.7. Assumption A4.(a) requires that the eigenvalues of the covariance matrices Λ_i are all bounded away from zero and infinity. Assumption A4.(b) is a common condition to handle the asymptotic bias; see Hahn and Newey (2004) and Arellano and Hahn (2007) for similar assumptions.

As the main result of this section, Theorem 3 shows that the elements of $\widehat{\underline{\beta}}$ are asymptotically normally distributed.

Theorem 3. *Suppose $G = G^0$ and Assumptions A1, A3 and A4 hold. If $N/T \to \kappa$ for some $\kappa \geq 0$, then as $(N, T) \to \infty$, for each $g \in [G^0]$ and conditioning on $\mathcal{D}$,*

$$\sqrt{NT}(\widehat{\beta}_g - \beta_g^0) \xrightarrow{D} N(-\sqrt{\kappa} W_g^{-1} \Delta_g, \pi_g^{-1} W_g^{-1} D_g W_g^{-1}), \text{ almost surely.}$$

Furthermore, for each $g \in [G^0]$, it follows that

$$\sqrt{\pi_g NT} D_g^{-1/2} W_g (\widehat{\beta}_g - \beta_g^0) + \sqrt{\pi_g \kappa} D_g^{-1/2} \Delta_g \xrightarrow{D} N(0, I).$$

Remark 3.8. Theorem 3 is closely related to a number of studies on panel data models with fixed effects. First, the asymptotic bias of $\widehat{\beta}_g$ is of order $\sqrt{N/T}$. For the fixed effects model, Hahn and Newey (2004) derived the same order for the asymptotic bias of the fixed effects estimator. In particular, $\widehat{\beta}_g$'s become asymptotically unbiased when $N = o(T)$. Second, when (X_{it}, Y_{it}) are conditionally independent given $\mathcal{D}$ across i and t, the bias term has following expression:

$$-\sqrt{\kappa} W_g^{-1} \Delta_g = -\sqrt{\kappa} W_g^{-1} \lim_{N\to\infty} \frac{1}{N_g} \sum_{i:g_i^0=g} \left(\frac{E_{\mathcal{D}}(R_{i1} U_{i1}^\alpha)}{E_{\mathcal{D}}(R_{i1}^\alpha)} - \frac{E_{\mathcal{D}}(U_{i1}^{\alpha\alpha}) E_{\mathcal{D}}(|R_{i1}|^2)}{2E_{\mathcal{D}}^2(R_{i1}^\alpha)} \right).$$

For the fixed effects model without group structures (i.e., $\pi_g = N_g/N = 1$), the above expression coincides with those obtained by Arellano and Hahn (2007).

Remark 3.9. For statistical inference, we need to obtain a bias-corrected estimator and the associated estimator of the variance–covariance matrix. In practice, for each subgroup, one can either use the jackknife procedure proposed by Hahn and Newey (2004) and Dhaene and Jochmans (2015), or the plug-in method suggested by Arellano and Hahn (2007) to obtain the bias-corrected estimator. For instance, let $\widehat{N}_g$ be the number of individuals in the corresponding estimated group $\widehat{\mathcal{C}}_g$, then the bias-corrected estimator using the plug-in method is given by

$$\widehat{\beta}_g^{BC} = \widehat{\beta}_g + \frac{1}{T} \widehat{W}_g^{-1} \widehat{\Delta}_g,$$

where $\widehat{W}_g = \sum_{i:g_i^0=g} \widehat{\mathcal{I}}_i/\widehat{N}_g$ with $\widehat{\mathcal{I}}_i = \widehat{E}_T(\widehat{V}_{it})$, $\widehat{V}_{it} = V_i(X_{it}, Y_{it}, \widehat{\beta}_{\widehat{g}_i}, \widehat{\alpha}_i)$, $\widehat{E}_T(\cdot) = \sum_{t=1}^T (\cdot)/T$ denoting the sample average and

$$\widehat{\Delta}_g = \frac{1}{\widehat{N}_g} \sum_{i:\widehat{g}_i=g} \widehat{E}_T\left\{ \left(\frac{\sum_{t=1}^T \widehat{R}_{it}}{\sqrt{T}\widehat{E}_T(\widehat{R}_{it}^\alpha)} \right) \left(\frac{1}{\sqrt{T}} \sum_{t=1}^T [\widehat{U}_{it}^\alpha - \frac{\widehat{E}_T(\widehat{U}_{it}^{\alpha\alpha})}{2\widehat{E}_T(\widehat{R}_{it}^\alpha)} \widehat{R}_{it}] \right) \right\},$$

with $\widehat{R}_{it} = R_i(X_{it}, Y_{it}, \widehat{\beta}_{\widehat{g}_i}, \widehat{\alpha}_i)$, $\widehat{R}_{it}^\alpha = R_i^\alpha(X_{it}, Y_{it}, \widehat{\beta}_{\widehat{g}_i}, \widehat{\alpha}_i)$, $\widehat{U}_{it}^\alpha = U_{it}^\alpha(X_{it}, Y_{it}, \widehat{\beta}_{\widehat{g}_i}, \widehat{\alpha}_i)$, $\widehat{U}_{it}^{\alpha\alpha} = U_i^{\alpha\alpha}(X_{it}, Y_{it}, \widehat{\beta}_{\widehat{g}_i}, \widehat{\alpha}_i)$. Similarly, the variance–covariance matrix can be estimated by $\widehat{\pi}_g^{-1} \widehat{W}_g^{-1} \widehat{D}_g \widehat{W}_g^{-1}$, where $\widehat{\pi}_g = \widehat{N}_g/N$ and

$$\widehat{D}_g = \frac{1}{\widehat{N}_g} \sum_{i:\widehat{g}_i=g} \widehat{\Lambda}_i,$$

with $\widehat{\Lambda}_i = T\widehat{E}_T(\widehat{U}_{it})\widehat{E}_T(\widehat{U}_{it}')$, and $\widehat{U}_{it} = U_i(X_{it}, Y_{it}, \widehat{\beta}_{\widehat{g}_i}, \widehat{\alpha}_i)$. It should be noted that in practice $\widehat{\beta}_g$ can be replaced with the bias-corrected estimator $\widehat{\beta}_g^{BC}$. For details, see Arellano and Hahn (2007).

3.4. Determination of the number of groups

Although our estimation and classification consistency results are valid for overspecified G, it is still of interest to estimate the number of groups due to following reasons. Estimating G can provide a guideline to select the number of groups for estimation and classification, while statistical inference requires a correctly specified G. In this section, we propose an efficient approach based on an information criterion to address this problem and establish its theoretical validity. Let $\widehat{\theta}_N^G$ be the estimator in (2.2) using G as the number of groups. To estimate G^0, we define the following information criterion function

$$PC(G) = \widehat{\Psi}_N(\widehat{\theta}_N^G) - \eta_{NT} G,$$

where $\eta_{NT} > 0$ is a penalty parameter that is used to exclude the extremely large and unlikely choice of G. We estimate G^0 by

$$\widehat{G} = \underset{G \in [G_{\max}]}{\operatorname{argmax}} PC(G), \tag{3.1}$$

where $G_{\max}$ is a predetermined upper bound for G. The following theorem shows that $\widehat{G}$ is a consistent estimator of G^0.

Theorem 4. *Suppose Assumptions A1 and A3 hold. If* $\log N = o(T^{\frac{d}{2(1+d)}})$, $\eta_{NT} T^{\frac{1}{2(1+d)}} \to \infty$ *and* $\eta_{NT} \to 0$, *then* $\lim_{(N,T)\to\infty} P(\widehat{G} = G^0) = 1$.

Note that the rate condition $\log N = o(T^{\frac{d}{2(1+d)}})$ in Theorem 4 is slightly stronger than Assumption A2, though both conditions allow N to grow exponentially with T. Since η_{NT} plays a crucial role in determining the number of groups, we follow Su et al. (2016) to pick η_{NT} by trying a list of candidates (see Section 4 for details).

4. Monte Carlo simulation

To investigate the finite-sample performance of our proposed procedure and compare it with C-Lasso, we consider three data generating processes (DGPs) that cover both linear and nonlinear panels of static and dynamic models. Throughout these DGPs, we generate the fixed effect α_i from standard normal distribution with a truncation of 3 standard deviations independently across i, and draw the idiosyncratic error u_{it} independently from standard normal distribution across i and t. Moreover, u_{it} is also independent of all regressors. We set the number of groups to be three (i.e., $G^0 = 3$), and the number of elements in each group are given by $N_1 = \lfloor 0.3N \rfloor$, $N_2 = \lfloor 0.3N \rfloor$ and $N_3 = N - N_1 - N_2$, where N is the total number of cross-sectional units and $\lfloor \cdot \rfloor$ denotes the integer part of "$\cdot$".

DGP 1 (Linear panel model): The data is generated as follows:

$$y_{it} = \alpha_i + X_{it}' \boldsymbol{\beta}_{g_i} + u_{it},$$

where $X_{it} = \left(0.2\alpha_i + e_{it,1}, 0.2\alpha_i + e_{it,2}\right)'$ and $e_{it,1}, e_{it,2} \sim$ I.I.D.$N(0, 1)$ across i, t and are independent of α_i. The true coefficients are $(0.4, 1.6)$, $(1, 1)$ and $(1.6, 0.4)$ for the three groups, respectively.

DGP 2 (Linear dynamic panel model): The observations are generated by the following:

$$y_{it} = \alpha_i \left(1 - \gamma_{g_i}\right) + \gamma_{g_i} y_{it-1} + X_{it}' \boldsymbol{\beta}_{g_i} + u_{it},$$

where X_{it} is a two-dimensional random vector generated in the same way as DGP 1. The true coefficients are $(0.4, 1.6, 1)$, $(0.6, 1, -1)$ and $(0.8, 0.4, 1.6)$ for the three groups, respectively.

DGP 3 (Dynamic probit panel model): The data follows the generating process below,

$$y_{it} = 1\left(\gamma_{g_i} y_{it-1} + x_{it}\beta_{1,g_i} + \beta_{2,g_i} + \alpha_i > u_{it}\right),$$

where $x_{it} = 0.1\alpha_i + e_{it}$ with $e_{it} \sim$ I.I.D.$N(0, 1)$ and is independent of all other variables. The true coefficients are $(1, -1, 0.5)$, $(0.5, 0, -0.25)$, and $(0, 1, 0)$. It should be noted that γ_{g_i} and β_{1,g_i} are identifiable in this model, whereas β_{2,g_i} is unidentifiable because it is absorbed into the individual specified effects α_i.

For all the three DGPs, we consider the combinations of (N, T) with $N = (100, 200)$ and $T = (15, 25, 50)$. During each replication, the group membership is held fixed and several numbers of groups, namely $G = 3, 4, 5$, are employed to estimate to coefficients. The number of replications is set to be $R = 1000$. Since the goal of this paper is to consistently estimate the regression coefficients, group membership, and number of groups, we consider the following three criteria to compare the finite sample performance of the proposed M-estimation and C-Lasso.

(1) Estimation consistency. For $G \geq G^0$, the consistency of estimation is evaluated using the root mean squared error (RMSE) of the estimated coefficients for each individual, which is defined as follows:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \|\widehat{\beta}_{\widehat{g}_i} - \beta_{g_i^0}^0\|_2^2}.$$

Table 1
RMSE under $G = 3, 4, 5$ with $G^0 = 3$.

		DGP 1						DGP 2						DGP 3					
	G	3		4		5		3		4		5		3		4		5	
N	T	M-Est	C-Lasso	M-Est	C-Lasso	M-Est	C-Lasso	M-Est	C-Lasso	M-Est	C-Lasso	M-Est	C-Lasso	M-Est	C-Lasso	M-Est	C-Lasso	M-Est	C-Lasso
100	15	0.190	0.189	0.217	0.225	0.234	0.238	0.115	0.245	0.178	0.341	0.224	0.415	0.296	0.366	0.512	0.623	0.571	0.732
100	25	0.113	0.120	0.140	0.145	0.157	0.165	0.055	0.195	0.117	0.234	0.154	0.352	0.190	0.187	0.256	0.242	0.277	0.268
100	50	0.036	0.028	0.068	0.066	0.083	0.088	0.036	0.092	0.131	0.166	0.109	0.258	0.119	0.142	0.173	0.163	0.182	0.179
200	15	0.188	0.186	0.214	0.221	0.233	0.238	0.111	0.201	0.174	0.324	0.201	0.428	0.286	0.278	0.381	0.532	0.399	0.503
200	25	0.109	0.122	0.136	0.150	0.153	0.163	0.048	0.157	0.167	0.235	0.184	0.248	0.185	0.170	0.240	0.302	0.261	0.256
200	50	0.032	0.046	0.065	0.090	0.080	0.106	0.028	0.035	0.074	0.158	0.101	0.143	0.116	0.110	0.162	0.174	0.176	0.168

Table 2
GRMSE of DGPs 1–3 with $G^0 = 3$. "Oracle" refers to estimation using the true group membership (i.e., g_i^0's are used).

		DGP 1			DGP 2			DGP 3		
N	T	M-Est	C-Lasso	Oracle	M-Est	C-Lasso	Oracle	M-Est	C-Lasso	Oracle
100	15	0.070	0.062	0.048	0.083	0.132	0.082	0.180	0.332	0.135
100	25	0.042	0.044	0.037	0.055	0.088	0.055	0.110	0.164	0.091
100	50	0.034	0.029	0.025	0.036	0.043	0.036	0.077	0.089	0.066
200	15	0.050	0.048	0.036	0.060	0.095	0.060	0.125	0.279	0.094
200	25	0.031	0.035	0.026	0.043	0.063	0.043	0.080	0.123	0.068
200	50	0.021	0.023	0.018	0.029	0.035	0.029	0.054	0.065	0.047

When $G = G^0$, we also consider another type RMSE similar to Su et al. (2016), which is defined as follows:

$$\text{Group RMSE} = \sqrt{\frac{1}{G^0}\sum_{g=1}^{G^0}\|\widehat{\beta}_g - \beta_g^0\|_2^2}.$$

(2) Consistency of $\widehat{G}$. The selection consistency for the number of groups is measured by the empirical percentage of selecting the true number of groups, namely, $\widehat{G} = G^0 = 3$ in our designs.[6]
(3) Classification consistency. We evaluate the percentage of correctly specifying the group membership of individuals, which is calculated as $\sum_{i=1}^{N} 1(\widehat{g}_i = g_i^0)/N$ under appropriate relabeling.

The simulation results of DGPs 1–3 are summarized in Tables 1–4. Several interesting findings can be observed. First, Table 1 provides the RMSE for the proposed M-estimation and C-Lasso[7] using various numbers of groups under $G^0 = 3$. From Table 1, we can observe that the RMSE decreases rapidly with the increase of either N or T for both M-estimation and C-Lasso regardless of the choices of G. Moreover, the estimation obtained using our approach in general has a smaller RMSE than those obtained using C-Lasso. Second, Table 2 provides the results that, if G^0 is prespecified, the Group RMSE (GRMSE) of M-estimator and C-Lasso diminishes quickly when increasing N and T, and both approaches perform similarly to the oracle estimator (e.g., knowing the true group membership), which is consistent with our findings in Theorem 1. Compared with C-Lasso, our method has better finite sample performance for both linear (DGPs 1–2) and nonlinear (DGP 3) models. It is worth mentioning that the GRMSE obtained from our proposed method is almost identical to that of the oracle one in the dynamical linear panel (DGP 2) regardless of the sample size. Third, Table 3 summarizes the accuracy of selection for the number of groups using the criterion $PC(G)$ proposed in Section 3.4 and the C-Lasso procedure. We note that, throughout all our designs of both linear and nonlinear panels, the determination of the number of groups is fairly accurate in the sense that the percentage of choosing the true number of groups is quite close to 1 when the sample size is large enough, which is comparable with that using C-Lasso. Finally, Table 4 presents the simulation results of the correct classification. For the correctness of classification, we observe that, with a large enough sample size, the classification for both linear and nonlinear panels is fairly accurate, and it is evident that the classification is consistent, as shown in Theorem 2. Furthermore, compared with C-Lasso, our method has significantly better classification accuracy in both linear and nonlinear dynamic panels (DGPs 2 and 3). In conclusion, the performance of our proposed procedure is fairly good with a finite sample, and performs comparably better than the C-Lasso method. Therefore, we can claim

[6] For the choice of the tuning parameter η_{NT} in Theorem 4 in determining the number of groups, we follow the idea in Su et al. (2016) and try various distinct values. We find that $\frac{1}{5\log(T)T^{1/8}}$ and $\frac{\log(N)^{1/8}}{5\log(T)T^{1/8}}$ have fairly good performance in linear and probit models, respectively. These settings are also used in empirical studies. For implementation of C-Lasso, we follow the same settings and principle in Su et al. (2016) to choose the tuning parameters.

[7] The authors appreciate (Su et al., 2016) for sharing their codes for C-Lasso.

Table 3
Percentage of choosing $G = 1, 2, \ldots, 5$ with $G^0 = 3$.

			DGP 1					DGP 2					DGP 3				
N	T	G	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
100	15	M-Est	0	0.004	0.976	0.020	0	0	0.002	0.473	0.372	0.153	0	0.081	0.612	0.262	0.045
		C-Lasso	0	0	0.994	0.004	0.002	0	0.249	0.743	0.003	0.005	0	0.129	0.583	0.192	0.096
100	25	M-Est	0	0	0.996	0.004	0	0	0	0.939	0.060	0.001	0	0.058	0.810	0.128	0.004
		C-Lasso	0	0	1	0	0	0	0.010	0.980	0.005	0.005	0	0.096	0.646	0.242	0.016
100	50	M-Est	0	0	0.988	0.012	0	0	0	0.995	0.005	0	0	0.007	0.895	0.098	0
		C-Lasso	0	0	1	0	0	0	0	1	0	0	0	0	0.986	0.014	0
200	15	M-Est	0	0	0.996	0.004	0	0	0	0.669	0.221	0.110	0	0.063	0.705	0.221	0.011
		C-Lasso	0	0	0.890	0.106	0.004	0	0.028	0.972	0	0	0	0.121	0.662	0.235	0.018
200	25	M-Est	0	0	1	0	0	0	0	0.963	0.037	0	0	0.011	0.881	0.106	0.002
		C-Lasso	0	0	1	0	0	0	0	1	0	0	0	0	0.986	0.014	0
200	50	M-Est	0	0	1	0	0	0	0	0.998	0.002	0	0	0.002	0.932	0.066	0
		C-Lasso	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0

Table 4
Percentage of correct classification with $G^0 = 3$.

		DGP 1		DGP 2		DGP 3	
N	T	M-Est	C-Lasso	M-Est	C-Lasso	M-Est	C-Lasso
100	15	0.902	0.889	0.995	0.943	0.752	0.703
100	25	0.934	0.958	1	0.952	0.909	0.893
100	50	0.966	0.996	1	1	0.979	0.945
200	15	0.903	0.898	0.995	0.923	0.883	0.856
200	25	0.967	0.972	0.999	0.944	0.929	0.893
200	50	0.995	0.996	1	0.981	0.980	0.955

that the simulation results confirm our theoretical findings in this paper regarding the identification and estimation for panels with unknown group structures.[8]

5. Empirical application

In this section, we apply the above estimation and classification method to two datasets, including linear and binary choice models.

5.1. Women's labor-force participation

The first dataset comes from the Panel Study of Income Dynamics (PSID) and contains 664 married women for 10 calendar years from 1979 to 1988. We consider the following dynamic probit panel model with fixed effects:[9]

$$y_{it} = 1\left(\alpha_i + \gamma_{g_i} y_{it-1} + X_{it}'\beta_{g_i} + \varepsilon_{it} > 0\right)$$

where y_{it} takes a value of one if woman i participates in period t, and zero otherwise. Moreover, α_i represents individual-specific effect and ϵ_{it} is standard normal. Other independent variables are X_{it} =(#children$_{it}$, log*income*$_{it}$, *race*, *eduwife*, *agewife*, and *agewife*^{2}), where #children$_{it}$ is the number of children aged between 0 and 17, log*income* is the log of the husband's labor income deflated by the Consumer Price Index, *race* is an indicator function and takes value 1 for black, *eduwife* is the years of education of the woman, *agewife* is the age of the woman (divided by 10) and *agewife*^{2} is the squared age. Similar variables were also considered by Hyslop (1999) and Carro (2007).

Using the classification method in the previous section, we divide the original sample into two groups (i.e., $G = 2$). The summary statistics for the full sample and the two subgroups are provided in Table 5. From Table 5, we can observe that these two groups have quite distinct observations for some variables. For example, comparatively, individuals in Group 2 have more children, lower percentage of black race, and younger age, while individuals in Group 1 have more years of education. The differences in these two groups make a difference in the estimation. Furthermore, on average, individuals from Group 2 have a much higher tendency to join the labor market compared with individuals from Group 1, e.g., the mean of labor force participation rate is 0.7898 for individuals from Group 2 and is 0.3982 for Group 1.

For the estimated group membership, we apply the fixed effects probit regression for each group and the whole sample. The estimation results are summarized in Table 6. Several interesting findings can be observed in the above estimation. First, we note that the effects of variables of the previous year's labor force participation, husband's income, and wife's age

[8] We also provide the comparison of the computational time of the proposed algorithm and C-Lasso, which is included in Supplement S.IV. As shown in the Supplement, for the nonlinear dynamic model (DGP 3), our algorithm is much faster.

[9] The estimation of model with both individual and time effects is provided in the Supplement S.V.

Table 5
Summary statistics for the original sample and two subgroups.

	Full Sample			Group 1			Group 2		
Variables	min	mean	max	min	mean	max	min	mean	max
yit	0	0.574	1	0	0.398	1	0	0.789	1
#children	0	1.76	7	0	1.691	6	0	1.841	7
logincome	5.806	10.471	13.846	5.806	10.483	12.995	6.64	10.46	13.85
race	0	0.164	1	0	0.179	1	0	0.147	1
eduwife	5	12.05	18	5	12.13	18	5	11.95	18
agewife	1.8	3.557	6.3	1.9	3.671	6.2	1.8	3.424	6.3
agewife2	3.24	13.41	39.69	3.61	14.25	38.44	3.24	12.43	39.69

Table 6
Estimation results of women's labor-force participation.

Variables	Full Sample	Group 1	Group 2
$y_{i,t-1}$	2.0504***	2.1779***	0.8835***
#children	0.0000	−0.0395	−0.0915**
logincome	−0.1933***	−0.2408***	−0.1787**
race	−0.1735**	−0.1938*	0.1215
eduwife	0.0096	0.0088	0.0257**
agewife	1.2635***	1.8465***	2.6434***
agewife2	−0.1610***	−0.2297***	−0.3097***

Note: *, **, and *** refer to significance at 10%, 5% and 1% level, respectively.

remain the same across the whole sample and two groups, even if the effects are quite different across different groups. Second, the race has negative effects on the labor force participation in the whole sample and Group 1, while race is no longer significant in Group 2. From the summary statistics, we notice that Group 2 has a relatively low percentage of black wives, which indicates that the effect of race is offset by other variables in this group. Finally, we observe that the education of the wife is not significant in the whole sample and Group 1, while it is significant in Group 2, which indicates that education indeed has a positive significant effect on the labor force participation for individuals in Group 2.[10]

5.2. Aggregate production

The second empirical application is to analyze the aggregate production function, which is important for economists to understand economic growth, technology changes, productivity difference across countries, and production efficiency. In the following, we apply our method to the Aggregate Production Data, which is extracted from version 9.0 of the Penn World Table. We keep a balanced panel dataset for 43 countries across the world during the period 1950–2014. Following Glass et al. (2016), we consider following linear model[11]:

$$y_{it} = \beta_{k,g_i} k_{it} + \beta_{l,g_i} l_{it} + \beta_{pub,g_i} pub_{it} + \beta_{xm,g_i} xm_{it} + \alpha_i + error_{it},$$

where y_{it}, k_{it}, and l_{it} are the real log GDP, capital stock, and number of people engaged in the ith country at year t, respectively. Furthermore, pub_{it} is the government/public expenditure and xm_{it} is the net trade openness defined as the exports minus imports of merchandise. Based on our estimation procedure, we divide the whole sample into three subgroups. Table 7 reports the estimated parameters and corresponding significance based on the full sample and three subgroups. All the estimated coefficients are highly significant among the four models under 1% level, except the coefficient of xm in Group 1, which is insignificant under 10% level. The capital and labor elasticity of full sample are estimated to be 0.6366 and 0.3845, approximately summing up to 1, which coincides with the convention in the Cobb–Douglas model (appropriately scaled). These two elasticities are similar between Groups 2 and 3 with capital elasticity being higher, while in Group 1, the capital elasticity is estimated as 0.4113, about 0.2 smaller than the labor elasticity. Furthermore, in terms of the coefficient for pub, Groups 1 and 3 are fairly close, with coefficients estimated to be around 1. However, the slope for pub in Group 2 is −3.4091, showing a significant difference. Finally, xm has distinct impacts on y for three groups, namely, no impact for Group 1, a negative impact for Group 2 with coefficient of −0.6887, and a positive impact for Group 1 with coefficient of 0.8006.[12]

[10] A possible reason for the heterogeneity between the two groups could be the motivation to participate the job market for married women with different background such as race, years of education, etc. However, to verify whether this is the source of heterogeneity needs more empirical evidences, which is out of the scope of this paper.

[11] The estimation of model with both individual and time effects is provided in the Supplement S.V.

[12] A possible reason for the heterogeneity among countries and groups on the capital and labor elasticities might be due to the difference of economic structures, natural endowment and distinct technological background among countries (see, e.g., Villacorta, 2017 and the references therein).

Table 7
Estimation results of aggregate production.

Variable	Full Sample	Group 1	Group 2	Group 3
k	0.6366***	0.4113***	0.6533***	0.6152***
l	0.3845***	0.6005***	0.5610***	0.5771***
pub	0.7636***	1.0805***	−3.4091***	1.0224***
xm	0.5649***	0.1609	−0.6887***	0.8006***

Note: *, **, and *** refer to significance at 10%, 5% and 1% level, respectively.

6. Conclusion

In this paper, we consider the identification and estimation of panel models with group structures when the true number of groups and the group membership are unknown to researchers. We propose an M-estimation procedure to estimate the parameters of interest and classify individuals, and an information criterion function to determine the number of groups. The method we proposed is applicable to both linear and nonlinear panels. Asymptotic properties are established for the estimation, classification and the determination of the number of groups. As a major theoretical contribution, we show that, under certain assumptions, the consistency of our proposed estimation and classification procedure is independent of the number of groups used in estimation as long as this number is not underestimated. The important practical implication of this result is that it is safe to use a relative large number of groups to estimate the model. Monte Carlo simulations are conducted to examine the finite sample properties of the proposed method, and the results confirm our theoretical findings. Applications to two real datasets also highlight the necessity to consider both individual-level and group-level heterogeneity.

Appendix A

In this appendix, we provide the proofs of main results in the paper. The proofs use some technical lemmas whose proofs are provided in the online supplement.

Before proceeding, we define some notation. We say a random variable $\zeta_{NT} = o_{P_{\mathcal{D}}}(r_{NT})$ for some real number r_{NT} if and only if for any $\epsilon > 0$, $P_{\mathcal{D}}(Z_{NT} > r_{NT}^{-1}\epsilon) = o_P(1)$ holds. Similarly, We say random variable $\zeta_{NT} = O_{P_{\mathcal{D}}}(r_{NT})$ if and only if for any $\epsilon > 0$, there exists $C_\epsilon > 0$ such that $P_{\mathcal{D}}(Z_{NT} > r_{NT}^{-1}C_\epsilon) \le \epsilon + o_P(1)$. We say an event A_{NT} holds with conditional probability given $\mathcal{D}$ approaching one, if and only if $P_{\mathcal{D}}(A_{NT}^c) = o_P(1)$. By the definitions of $o_{P_{\mathcal{D}}}$, $O_{P_{\mathcal{D}}}$ and the Dominated Convergence Theorem (DCT), we can show the following statements which reveal the relation between $P_{\mathcal{D}}$ and P:

(s.1) If $Z_{NT} = o_{P_{\mathcal{D}}}(r_{NT})$, then $Z_{NT} = o_P(r_{NT})$.
(s.2) If $Z_{NT} = O_{P_{\mathcal{D}}}(r_{NT})$, then $Z_{NT} = O_P(r_{NT})$.
(s.3) If an event A_{NT} holds with conditional probability given $\mathcal{D}$ approaching one, then A_{NT} also holds with probability approaching one.

A.1. Proof of Theorem 1

We complete the proof of Theorem 1 in three propositions. Proposition A.1 shows the first part of Theorem 1, and Propositions A.2 and A.3 prove the second part of Theorem 1. The technical lemmas used in the proofs of each proposition are provided before the corresponding proposition.

Lemma A.1. *Under Assumption A1, for every $0 < \epsilon < R$, with $R = \sup_{\beta_1,\beta_2\in\mathbb{K},\alpha_1,\alpha_2\in\mathbb{A}} \|\beta_1 - \beta_2\|_2 + |\alpha_1 - \alpha_2|$, the inequality*

$$\inf_{d_N(\theta_N,\theta_N^0)\ge\epsilon} [\Psi_N(\theta_N^0) - \Psi_N(\theta_N)] \ge \frac{\epsilon}{2R}\chi(\epsilon^2/8)$$

holds almost surely.

Lemma A.2. *Suppose Assumption A1 holds. Then there exist positive constants C_3, C_4, C_5 such that the inequality*

$$\sup_{1\le i\le N} P_{\mathcal{D}}\left(\sup_{(\beta,\alpha)\in\mathbb{K}\times\mathbb{A}}\left|\widehat{H}_i(\beta,\alpha) - H_i(\beta,\alpha)\right| > 10z\right) \le C_4\left[1 + \frac{1}{z^{2(p+2)}}\right]$$
$$\times\left[\left(1 + \frac{T^{\frac{1}{1+d}}z^2}{C_5}\right)^{-T^{\frac{d}{1+d}}/4} + \frac{1}{d}\exp\left(-C_3T^{\frac{d}{1+d}}z^d\right) + \frac{\exp(-C_3dT^{\frac{d}{1+d}}z^d)}{1-\exp(-C_3dT^{\frac{d}{1+d}}z^d)}\right]$$

holds almost surely for all $z > 0, N \ge 1$ and $T^{\frac{d}{1+d}} \ge 4(p+2)$. Furthermore, the condition $\log N = o(T^{\frac{d}{1+d}})$ implies

$$\sup_{1\le i\le N}\sup_{(\beta,\alpha)\in\mathbb{K}\times\mathbb{A}}\left|\widehat{H}_i(\beta,\alpha) - H_i(\beta,\alpha)\right| = o_{P_{\mathcal{D}}}(1) = o_P(1)$$

and

$$\sup_{\theta_N\in\Theta_N}\left|\widehat{\Psi}_N(\theta_N)-\Psi_N(\theta_N)\right|=o_{P_{\mathcal{D}}}(1)=o_P(1).$$

Lemma A.2 indicates the following concentration inequality:

$$S_{NT}:=\sup_{1\le i\le N}\sup_{\beta\in\mathbb{K},\alpha\in\mathbb{A}}|\widehat{H}_i(\beta,\alpha)-H_i(\beta,\alpha)|=o_{P_{\mathcal{D}}}(1)=o_P(1),$$

which plays an important role in our proof. Moreover, the definition of $\widehat{\Psi}$ suggests the following inequality:

$$\sup_{\theta_n\in\Theta_N}|\widehat{\Psi}_N(\theta_N)-\Psi_N(\theta_N)|=\sup_{(\underline{\beta},\underline{\alpha},\gamma_N)\in\mathbb{K}\times\mathbb{A}\times\Gamma_N}\left|\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{H}_i(\beta_{g_i},\alpha_i)-H_i(\beta_{g_i},\alpha_i)\right)\right|\le S_{NT}.$$

Proposition A.1. *Suppose $G\ge G^0$ and Assumptions A1, A2 hold. Then $d_N(\widehat{\theta}_N,\theta_N^0)=o_{P_{\mathcal{D}}}(1)=o_P(1)$.*

Proof. By the definition of $\widehat{\theta}_N$, we have the inequality

$$\Psi_N(\theta_N^0)-S_{NT}\le\widehat{\Psi}_N(\theta_N^0)\le\widehat{\Psi}_N(\widehat{\theta}_N)\le\Psi_N(\widehat{\theta}_N)+S_{NT}\le\Psi_N(\theta_N^0)+S_{NT}.$$

This inequality and Lemma A.2 establish the convergence result

$$\Psi_N(\widehat{\theta}_N)-\Psi_N(\theta_N^0)=o_{P_{\mathcal{D}}}(1).\tag{A.1}$$

Fix ϵ in the interval $(0,R)$ with $R=\sup_{\beta_1,\beta_2\in\mathbb{K},\alpha_1,\alpha_2\in\mathbb{A}}\|\beta_1-\beta_2\|_2+|\alpha_1-\alpha_2|$. Lemma A.1 yields the lower bound

$$\Psi_N(\theta_N^0)-\Psi_N(\widehat{\theta}_N)\ge[\Psi_N(\theta_N^0)-\Psi_N(\widehat{\theta}_N)]I(d_N(\widehat{\theta}_N,\theta_N^0)\ge\epsilon)\ge c_0I(d_N(\widehat{\theta}_N,\theta_N^0)\ge\epsilon)$$

with $c_0=\epsilon\chi(\epsilon^2/8)/(4R)$. From this and (A.1) we derive that $P_{\mathcal{D}}(d_N(\widehat{\theta}_N,\theta_N^0)\ge\epsilon)$ converges to zero in probability. This and (s.1) together give the desired results. □

Proposition A.1 proves the first part of Theorem 1. To prove the second part (see Proposition A.3), we need more notation. For $\beta\in\mathbb{K}$, we define

$$\widehat{\alpha}_i(\beta):=\operatorname*{argmax}_{\alpha\in\mathbb{A}}\widehat{H}_i(\beta,\alpha),$$

and for $\underline{\beta}:=(\beta_1,\beta_2,\ldots,\beta_G)\in\mathbb{K}^G$, define

$$\widehat{\gamma}_N(\underline{\beta}):=\operatorname*{argmax}_{\gamma_N\in\Gamma_N}\max_{\underline{\alpha}\in\mathbb{A}^N}\widehat{\Psi}_N(\underline{\beta},\underline{\alpha}),\tag{A.2}$$

with $(\widehat{g}_1(\underline{\beta}),\widehat{g}_2(\underline{\beta}),\ldots,\widehat{g}_N(\underline{\beta}))$ being the elements in $\widehat{\gamma}_N(\underline{\beta})$. To measure the difference between $(\widehat{\beta}_1,\widehat{\beta}_2,\ldots,\widehat{\beta}_G)$ and $(\beta_1^0,\beta_2^0,\ldots,\beta_{G^0}^0)$ with possibly $G\ne G^0$, we define a map $\sigma:[G^0]\to[G]$ by:

$$\sigma(g)=\operatorname*{argmin}_{\widetilde{g}\in[G]}\|\widehat{\beta}_{\widetilde{g}}-\beta_g^0\|_2,\ \text{ for } g\in[G^0].\tag{A.3}$$

If there are multiple minimizers, we just pick one of them.

Lemma A.3. *Under Assumption A1, the following Lipchitz condition holds almost surely:*

$$\sup_{N\ge1}\sup_{1\le i\le N}\sup_{(\beta_1,\alpha_1)\ne(\beta_2,\alpha_2)\in\mathbb{K}\times\mathbb{A}}\frac{|H_i(\beta_1,\alpha_1)-H_i(\beta_2,\alpha_2)|}{(\|\beta_1-\beta_2\|_2^2+|\alpha_1-\alpha_2|^2)^{1/2}}\le B_2,\tag{A.4}$$

where $B_2=\int_0^\infty\exp\left(1-(t/B_1)^{b_1}\right)dt$ if $0<b_1<\infty$ and $B_2=B_1$ if $b_1=\infty$.

Proposition A.2. *Suppose Assumptions A1–A2 and $G\ge G^0$ hold. Then we have*

$$\max_{g\in[G^0]}\|\widehat{\beta}_{\sigma(g)}-\beta_g^0\|_2=o_{P_{\mathcal{D}}}(1)=o_P(1).$$

Proof. Recall that $N_g=\sum_{i=1}^N I(g_i^0=g)$ denotes the size of group g. By the definition of σ, we have

$$\|\widehat{\beta}_{\sigma(g)}-\beta_g^0\|_2=\frac{1}{N_g}\sum_{i=1}^{N}I(g=g_i^0)\|\widehat{\beta}_{\sigma(g_i^0)}-\beta_{g_i^0}^0\|_2\le\frac{1}{N_g}\sum_{i=1}^{N}I(g=g_i^0)\|\widehat{\beta}_{\widehat{g}_i}-\beta_{g_i^0}^0\|_2$$

for $g \in [G^0]$, and therefore obtain the bound

$$\max_{g\in[G^0]} \|\widehat{\beta}_{\sigma(g)} - \beta^0_g\|_2 \le \max_{g\in[G^0]} \frac{N}{NN_g} \sum_{i=1}^{N} \|\widehat{\beta}_{\widehat{g}_i} - \beta^0_{g_i^0}\|_2 \le \max_{g\in[G^0]} \frac{N}{N_g}\, d_N(\widehat{\theta}_N, \theta^0_N).$$

Thus the desired result follows from Assumption A1.(f) and Proposition A.1. □

Lemma A.4. *Suppose Assumptions A1–A2 hold. Then*

$$\sup_{1\le i\le N} |\widehat{\alpha}_i(\beta^0_{g_i^0}) - \alpha^0_i| = o_{P_{\mathcal{D}}}(1) = o_P(1).$$

Furthermore, if $\{\beta_{Ti}, i \in [N]\}$ are random vectors satisfying $\sup_{1\le i\le N} \|\beta_{Ti} - \beta^0_{g_i^0}\|_2 = o_{P_{\mathcal{D}}}(1)$, then we have

$$\sup_{1\le i\le N} |\widehat{\alpha}_i(\beta_{Ti}) - \alpha^0_i| = o_{P_{\mathcal{D}}}(1) = o_P(1).$$

Proposition A.3. *Suppose $G \ge G^0$ and Assumptions A1 and A2 hold, then $\sup_{1\le i\le N} \|\widehat{\beta}_{\widehat{g}_i} - \beta^0_{g_i^0}\|_2 = o_{P_{\mathcal{D}}}(1) = o_P(1)$.*

Proof. From the inequality $\widehat{H}_i(\widehat{\beta}_{\sigma(g_i^0)}, \widehat{\alpha}(\widehat{\beta}_{\sigma(g_i^0)})) \le \widehat{H}_i(\widehat{\beta}_{\widehat{g}_i}, \widehat{\alpha}(\widehat{\beta}_{\widehat{g}_i}))$ and the definition of S_{NT}, we derive

$$H_i(\widehat{\beta}_{\sigma(g_i^0)}, \widehat{\alpha}(\widehat{\beta}_{\sigma(g_i^0)})) - S_{NT} \le \widehat{H}_i(\widehat{\beta}_{\sigma(g_i^0)}, \widehat{\alpha}(\widehat{\beta}_{\sigma(g_i^0)})) \le \widehat{H}_i(\widehat{\beta}_{\widehat{g}_i}, \widehat{\alpha}(\widehat{\beta}_{\widehat{g}_i})) \le H_i(\widehat{\beta}_{\widehat{g}_i}, \widehat{\alpha}(\widehat{\beta}_{\widehat{g}_i})) + S_{NT}$$

and therefore

$$C_{NT} := \sup_{1\le i\le N} \left(H_i(\widehat{\beta}_{\sigma(g_i^0)}, \widehat{\alpha}(\widehat{\beta}_{\sigma(g_i^0)})) - H_i(\widehat{\beta}_{\widehat{g}_i}, \widehat{\alpha}(\widehat{\beta}_{\widehat{g}_i})) \right) \le 2S_{NT}. \tag{A.5}$$

Fix $\epsilon > 0$ and set $A_{NT} = \{\sup_{1\le i\le N} \|\widehat{\beta}_{\widehat{g}_i} - \beta^0_{g_i^0}\|_2 \ge \epsilon\}$. By Assumption A1.(c), we have

$$\begin{aligned} \inf_{1\le i\le N} \left(H_i(\beta^0_{g_i^0}, \alpha^0_i) - H_i(\widehat{\beta}_{\widehat{g}_i}, \widehat{\alpha}(\widehat{\beta}_{\widehat{g}_i})) \right) &\ge \inf_{1\le i\le N} \left(H_i(\beta^0_{g_i^0}, \alpha^0_i) - H_i(\widehat{\beta}_{\widehat{g}_i}, \widehat{\alpha}(\widehat{\beta}_{\widehat{g}_i})) \right) I(A_{NT}) \\ &\ge \chi(\epsilon^2) I(A_{NT}). \end{aligned} \tag{A.6}$$

Lemma A.3 and the triangle inequality imply

$$\begin{aligned} D_{NT} &:= \sup_{1\le i\le N} \left(H_i(\beta^0_{g_i^0}, \alpha^0_i) - H_i(\widehat{\beta}_{\sigma(g_i^0)}, \widehat{\alpha}(\widehat{\beta}_{\sigma(g_i^0)})) \right) \\ &\le \sup_{1\le i\le N} B_2 \left(\|\widehat{\beta}_{\sigma(g_i^0)} - \beta^0_{g_i^0}\|_2 + |\widehat{\alpha}(\widehat{\beta}_{\sigma(g_i^0)}) - \alpha^0_i| \right). \end{aligned} \tag{A.7}$$

By Proposition A.2, we have

$$\sup_{1\le i\le N} \|\widehat{\beta}_{\sigma(g_i^0)} - \beta_{g_i^0}\|_2 \le \max_{g\in[G^0]} \|\widehat{\beta}_{\sigma(g)} - \beta_g\|_2 = o_{P_{\mathcal{D}}}(1). \tag{A.8}$$

Thus Lemma A.4 applied with $\beta_{Ti} = \widehat{\beta}_{\sigma(g_i^0)}$ yields

$$\sup_{1\le i\le N} |\widehat{\alpha}(\widehat{\beta}_{\sigma(g_i^0)}) - \alpha^0_i| = o_{P_{\mathcal{D}}}(1). \tag{A.9}$$

The inequalities (A.7)–(A.9) further imply

$$D_{NT} = o_{P_{\mathcal{D}}}(1) \tag{A.10}$$

Combining (A.5), (A.6), (A.10) and Lemma A.2, we have

$$\chi(\epsilon^2) I(A_{NT}) \le C_{NT} + D_{NT} \le 2S_{NT} + D_{NT} = o_{P_{\mathcal{D}}}(1).$$

This implies $P_{\mathcal{D}}(A_{NT}) = o_{P_{\mathcal{D}}}(1) = o_P(1)$, which is the desired result. □

Proof of Theorem 1. This is a direct consequence of Propositions A.1 and A.3. □

A.2. Proof of Theorem 2

To compare $\underline{\beta} = (\beta_1, \ldots, \beta_G) \in \mathbb{K}^G$ and $\underline{\widetilde{\beta}} = (\widetilde{\beta}_1, \ldots, \widetilde{\beta}_{\widetilde{G}}) \in \mathbb{K}^{\widetilde{G}}$ for possible $G \ne \widetilde{G}$, we introduce following Hausdorff distance:

$$d_H(\underline{\beta}, \underline{\widetilde{\beta}}) = \max\left\{ \max_{g\in[G]} \min_{\widetilde{g}\in[\widetilde{G}]} \|\beta_g - \widetilde{\beta}_{\widetilde{g}}\|_2, \max_{\widetilde{g}\in[\widetilde{G}]} \min_{g\in[G]} \|\beta_g - \widetilde{\beta}_{\widetilde{g}}\|_2 \right\}.$$

Define a η-neighborhood of $\underline{\beta}^0$ by $\mathcal{N}_\eta = \{\underline{\beta} \in \mathbb{K}^G : d_H(\underline{\beta}, \underline{\beta}^0) < \eta\}$. Also for each $\underline{\beta} \in \mathcal{N}_\eta$, we denote sets $\mathcal{A}_\eta(\underline{\beta}, g) = \{\widetilde{g} \in [G] : \|\beta_{\widetilde{g}} - \beta^0_g\|_2 < \eta\} \subset [G]$, for all $g \in [G^0]$. Here $\mathcal{A}_\eta(\underline{\beta}, \cdot)$ plays a role of relabeling that connects labels in $[G^0]$ with labels in $[G]$.

Before proving Theorem 2, we provide two useful lemmas.

Lemma A.5. *Suppose Assumptions A1–A2 and $G \geq G^0$ hold. Then we have*

$$d_H(\widehat{\underline{\beta}}, \underline{\beta}^0) = o_{P_{\mathcal{D}}}(1) = o_P(1).$$

Lemma A.6. *Suppose Assumptions A1, A3 and $G \geq G^0$ hold. Then for $\eta > 0$ small enough, we have the following:*

(i) For all $\underline{\beta} \in \mathcal{N}_\eta$, $\{\mathcal{A}_\eta(\underline{\beta}, g), g \in [G^0]\}$ is a partition of $[G]$ and each $\mathcal{A}_\eta(\underline{\beta}, g)$ is non-empty for all $g \in [G^0]$.

(ii) $P_{\mathcal{D}}\left(\sup_{\underline{\beta} \in \mathcal{N}_\eta} \sup_{1 \leq i \leq N} I(\widehat{g}_i(\underline{\beta}) \notin \mathcal{A}_\eta(\underline{\beta}, g_i^0)) > 0\right) = o_P(1)$.

(iii) If $G = G^0$, then each $\mathcal{A}_\eta(\underline{\beta}, g)$ contains exactly one element for all $g \in [G^0]$ and thus $\mathcal{A}_\eta(\underline{\beta}, \cdot)$ is a permutation of $[G^0]$. Under this permutation,

$$P_{\mathcal{D}}\left(\sup_{\underline{\beta} \in \mathcal{N}_\eta} \sup_{1 \leq i \leq N} I(\widehat{g}_i(\underline{\beta}) \neq g_i^0) > 0\right) = o_P(1).$$

Proof of Theorem 2. From the definition of d_H and Lemma A.5, we conclude that for $0 < \eta < d_0/2$, with conditional probability given $\mathcal{D}$ approaching one, $\{\mathcal{A}_\eta(\widehat{\underline{\beta}}, g), g \in [G^0]\}$ is a partition of $[G]$ and each $\mathcal{A}_\eta(\widehat{\underline{\beta}}, g)$ is non-empty for all $g \in [G^0]$.

Next by Lemma A.6 and the fact that with conditional probability given $\mathcal{D}$ approaching one, $\widehat{\underline{\beta}} \in \mathcal{N}_\eta$, we have

$$1 - P_{\mathcal{D}}\left(\widehat{g}_i \in \mathcal{A}_\eta(\widehat{\underline{\beta}}, g_i^0), \forall i \in [N]\right) = o_P(1).$$

By DCT, it follows that

$$\lim_{(N,T)\to\infty} P\left(\widehat{g}_i \in \mathcal{A}_\eta(\widehat{\underline{\beta}}, g_i^0), \forall i \in [N]\right) = 1.$$

Finally, suppose $i, j \in \widehat{\mathcal{C}}_g$ for some $g \in [G]$, then $\widehat{g}_i = \widehat{g}_j = g$. From argument above, we can see, with conditional probability given $\mathcal{D}$ approaching one, $g \in \mathcal{A}_\eta(\widehat{\underline{\beta}}, g_i^0)$ and $g \in \mathcal{A}_\eta(\widehat{\underline{\beta}}, g_j^0)$. Notice with conditional probability given $\mathcal{D}$ approaching one, $\{\mathcal{A}_\eta(\widehat{\underline{\beta}}, g), g \in [G^0]\}$ is a partition of $[G]$, so it follows that $g_i^0 = g_j^0$. Now define $\widetilde{g} = g_i^0 = g_j^0 \in [G^0]$, then $i, j \in \mathcal{C}_{\widetilde{g}}$. Therefore, with conditional probability given $\mathcal{D}$ approaching one, for each $g \in [G]$, there exists $\widetilde{g} \in [G^0]$, such that $\widehat{\mathcal{C}}_g \subset \mathcal{C}_{\widetilde{g}}$. □

A.3. Proof of Proposition 1 and Theorem 3

Proof of Proposition 1. Suppose $G = G^0$, then in the proof of Theorem 2, under appropriate relabeling, it follows that for each $g \in [G^0]$, $P_{\mathcal{D}}(\widehat{\mathcal{C}}_g \neq \mathcal{C}_g) = o_P(1)$. Further, we have $1 - P_{\mathcal{D}}(\widehat{g}_i = g_i^0, \forall i \in [N]) = o_P(1)$. Since on the event $\{\widehat{g}_i = g_i^0, \forall i \in [N]\}$, we have $\widehat{\underline{\beta}} = \widetilde{\underline{\beta}}$. Therefore, we finish the proof. □

Lemma A.7. *Suppose Assumptions A1, A3, A4 hold and $N = O(T)$, then for all $g \in [G^0]$ and conditioning on $\mathcal{D}$,*

$$\sqrt{N_g T}(\widetilde{\beta}_g - \beta_g^0) + \sqrt{N_g/T} W_g^{-1} \Delta_g \xrightarrow{D} N(0, W_g^{-1} D_g W_g^{-1}), \text{ almost surely.}$$

Moreover, for all $g \in [G^0]$,

$$\sqrt{N_g T} D_g^{-1/2} W_g(\widetilde{\beta}_g - \beta_g^0) + \sqrt{N_g/T} D_g^{-1/2} \Delta_g \xrightarrow{D} N(0, I).$$

Proof of Theorem 3. In the proof of Proposition 1, we show that $\widehat{\underline{\beta}} = \widetilde{\underline{\beta}}$ holds with conditional probability given $\mathcal{D}$ approaching one. By (s.3), it also holds that $\widehat{\underline{\beta}} = \widetilde{\underline{\beta}}$ with probability approaching one. Therefore, the asymptotic distribution follows from above asymptotic equivalence and Lemma A.7. □

A.4. Proof of Theorem 4

Before proving Theorem 4, we first supply two lemmas.

Lemma A.8. *Under Assumptions A1, A3 and $G < G^0$, there exists a constant B_4 such that the following hold almost surely:*

$$[\Psi_N(\theta_N^0) - \Psi_N(\widehat{\theta}_N)] \geq B_4 > 0.$$

Lemma A.9. *Suppose Assumptions A1, A3 and $G > G^0$ hold. Then the condition $\log N = o(T^{\frac{d}{2(1+d)}})$ implies*

$$|\widehat{\Psi}_N(\widehat{\theta}_N) - \widehat{\Psi}_N(\theta_N^0)| = O_{P_{\mathcal{D}}}(T^{-\frac{1}{4(1+d)}}).$$

Proof of Theorem 4. By (s.1), it suffices to show that

$$P_{\mathcal{D}}(PC(G) > PC(G^0)) = o_P(1) \text{ holds for each } G \neq G^0. \tag{A.11}$$

Now we consider two cases, namely $G < G^0$ and $G > G^0$.

Under-fitting case, $G < G^0$: By direct examination and Lemma A.8 , for (N, T) is large enough, it follows that

$$\begin{aligned} PC(G^0) - PC(G) &= \widehat{\Psi}_N(\widehat{\theta}_N^{G^0}) - \widehat{\Psi}_N(\theta_N^0) - \widehat{\Psi}_N(\widehat{\theta}_N^G) + \widehat{\Psi}_N(\theta_N^0) - \eta_{NT}(G^0 - G) \\ &\geq \widehat{\Psi}_N(\theta_N^0) - \widehat{\Psi}_N(\widehat{\theta}_N^G) - \eta_{NT}(G^0 - G) \\ &\geq \Psi_N(\theta_N^0) - \Psi_N(\widehat{\theta}_N^G) - \eta_{NT}(G^0 - G) - 2S_{NT} \\ &\geq B_4/2 + o_{P_{\mathcal{D}}}(1). \end{aligned} \tag{A.12}$$

Since $\eta_{NT} \to 0$, it follows from (A.12) that (A.11) holds for the case $G < G^0$.

Over-fitting case, $G > G^0$: By Lemma A.9, it follows that

$$\begin{aligned} PC(G^0) - PC(G) &= \widehat{\Psi}_N(\widehat{\theta}_N^{G^0}) - \widehat{\Psi}_N(\theta_N^0) - \widehat{\Psi}_N(\widehat{\theta}_N^G) + \widehat{\Psi}_N(\theta_N^0) + \eta_{NT}(G - G^0) \\ &= O_{P_{\mathcal{D}}}(T^{-\frac{1}{4(1+d)}}) + \eta_{NT}(G - G^0). \end{aligned} \tag{A.13}$$

Since $\eta_{NT} T^{\frac{1}{4(1+d)}} \to \infty$ and $G > G^0$, so (A.11) holds for the case when $G > G^0$. □

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2019.09.008.

References

Ando, T., Bai, J., 2016. Panel data models with grouped factor structure under unknown group membership. J. Appl. Econometrics 31 (1), 163–191.

Arellano, M., Bonhomme, S., 2009. Robust priors in nonlinear panel data models. Econometrica 77 (2), 489–536.

Arellano, M., Hahn, J., 2007. Understanding bias in nonlinear panel models: Some recent developments. In: Advances in Economics and Econometrics Theory and Applications, Ninth World Congress. Cambridge University Press, pp. 381–409, chapter 12.

Baltagi, B.H., Bresson, G., Pirotte, A., 2008. To pool or not to pool? In: The Econometrics of Panel Data. Springer-Verlag Berlin Heidelberg, pp. 517–546.

Bester, C.A., Hansen, C.B., 2016. Grouped effects estimators in fixed effects models. J. Econometrics 190 (1), 197–208.

Bonhomme, S., Manresa, E., 2015. Grouped patterns of heterogeneity in panel data. Econometrica 83 (3), 1147–1184.

Carro, J., 2007. Estimating dynamic panel data discrete choice models with fixed effects. J. Econometrics 140 (2), 503–528.

Chand, S., 2012. On tuning parameter selection of lasso-type methods-a monte carlo study. In: Proceedings of 2012 9th International Bhurban Conference on Applied Sciences and Technology, IBCAST. IEEE, pp. 120–129.

Dhaene, G., Jochmans, K., 2015. Split-panel jackknife estimation of fixed-effect models. Rev. Econom. Stud. 82 (3), 991–1030.

Fernández-Val, I., Weidner, M., 2016. Individual and time effects in nonlinear panel models with large n, t. J. Econometrics 192 (1), 291–312.

Glass, A.J., Kenjegalieva, K., Sickles, R.C., 2016. A spatial autoregressive stochastic frontier model for panel data with asymmetric efficiency spillovers. J. Econometrics 190 (2), 289–300.

Hahn, J., Newey, W., 2004. Jackknife and analytical bias reduction for nonlinear panel models. Econometrica 72 (4), 1295–1319.

Hsiao, C., Pesaran, H., 2008. Random coefficient models. In: The Econometrics of Panel Data. Springer-Verlag Berlin Heidelberg, pp. 185–213.

Hsiao, C., Tahmiscioglu, A.K., 1997. A panel analysis of liquidity constraints and firm investment. J. Amer. Statist. Assoc. 92 (438), 455–465.

Hyslop, D.R., 1999. State dependence, serial correlation and heterogeneity in intertemporal labor force participation of married women. Econometrica 67 (6), 1255–1294.

Kirkland, L.-A., Kanfer, F., Millard, S., 2015. LASSO tuning parameter selection. In: Annual Proceedings of the South African Statistical Association Conference, vol. 2015. South African Statistical Association (SASA), pp. 49–56.

Lancaster, T., 2002. Orthogonal parameters and panel data. Rev. Econom. Stud. 69 (3), 647–666.

Lee, K., Pesaran, M., Smith, R., 1997. Growth and convergence in a multi-country empirical stochastic growth model. J. Appl. Econometrics 12 (2), 357–392.

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

Lu, X., Su, L., 2017. Determining the number of groups in latent panel structures with an application to income and democracy. Quant. Econ. 8 (3), 729–760.

Moon, H.R., Weidner, M., 2015. Linear regression for panel with unknown number of factors as interactive fixed effects. Econometrica 83 (4), 1543–1579.

Sarafidis, V., Weber, N., 2015. A partially heterogeneous framework for analyzing panel data. Oxf. Bull. Econ. Stat. 77 (2), 274–296.

Su, L., Chen, Q., 2013. Testing homogeneity in panel data models with interactive fixed effects. Econometric Theory 29 (6), 1079–1135.

Su, L., Ju, G., 2018. Identifying latent grouped patterns in panel data models with interactive fixed effects. J. Econometrics 206, 554–573.

Su, L., Shi, Z., Phillips, P.C., 2016. Identifying latent structures in panel data. Econometrica 84 (6), 2215–2264.

Su, L., Wang, X., Jin, S., 2019. Sieve estimation of time-varying panel data models with latent structures. J. Bus. Econom. Statist. 37 (2), 334–349.

Villacorta, L., 2017. Estimating Country Heterogeneity in Capital-Labor Substitution Using Panel Data. Working Paper.

Vogt, M., Linton, O., 2019. Multiscale clustering of nonparametric regression curves. arXiv preprint arXiv:1903.01459.