Subspace clustering for panel data with interactive effects

Jiangtao DUAN¹,², Wei GAO²*, Hao QU², and Hon Keung Tony NG³

¹School of Economics and Statistics, Guangzhou University, Guangzhou, China
²Key Laboratory for Applied Statistics of MOE, School of Mathematics and Statistics, Northeast Normal University, Changchun, China
³Department of Statistical Science, Southern Methodist University, Dallas, Texas, USA

Key words and phrases: Factor structure models; least-squares estimation; K-means; mean-squared error; Monte Carlo simulation; panel data; subspace clustering.

MSC 2010: Primary 62H30; secondary 62H25.

Abstract: We study a statistical model for panel data with unobservable grouped factor structures which are correlated with the regressors and whose group membership can be unknown. We assume the factor loadings belong to different subspaces and consider the subspace clustering for factor loadings. We propose a method called least-squares subspace clustering (LSSC) to estimate the model parameters by minimizing the least-squares distance while simultaneously performing the subspace clustering. We establish the consistency of our proposed subspace clustering method and study the asymptotic properties of our proposed estimators under certain conditions. Monte Carlo simulation studies illustrate the advantages of our proposed methodologies. To choose the subspace dimensions consistently, we use a model selection criterion. We also outline further considerations for situations when the number of subspaces and the dimensions of factors are unknown. For illustrative purposes, our proposed methods are applied to study the linkage between income and democracy across countries. 

Résumé: Ce travail porte sur un modèle statistique de données de panel à structures factorielles agrégées, non observables, corrélées avec les régisseurs et dont l’appartenance à un groupe peut être inconnue. Sachant que les saturations factorielles associées peuvent appartenir à différents sous-espaces, une agrégation de ces sous-espaces est requise. Pour ce faire, les auteurs proposent une méthode, dite agrégation de sous espaces par moindres carrés (LSSC), qui consiste à minimiser une distance de moindres carrés tout en effectuant simultanément l’agrégation des sous-espaces. L’étude de la convergence et du comportement asymptotique de la méthode et des estimateurs proposés est effectuée sous certaines conditions. Le recours à des simulations de type Monte Carlo permet d’illustrer clairement les avantages et l’intérêt que présentent les méthodes proposées. Un critère de choix de modèle est utilisé pour déterminer les dimensions du sous-espace de manière cohérente. De plus, les situations où le nombre de sous-espaces et les dimensions des facteurs sont inconnus ont également été examinées. Enfin, en guise d’illustration de la méthode, une étude portant sur le lien entre le revenu et la démocratie dans différents pays est présentée.

© 2021 Statistical Society of Canada

Additional Supporting Information may be found in the online version of this article at the publisher’s website.

* Corresponding author: gaow@nenu.edu.cn

© 2021 Statistical Society of Canada
1. INTRODUCTION

Panel data, also known as longitudinal data, contain multi-dimensional observations that are collected over multiple periods for a sample of individuals. There is evidence to show that unobservable heterogeneity among different individuals in multiple dimensions exists in panel data; hence, suitable statistical models are needed. Among the various models for panel data, the interactive fixed-effects model uses the interactive effects that combine individual effects and time effects to reveal the effects of the common factors and thereby capture the unobserved information. The panel data model with interactive effects has been widely studied in the literature, for example, see Pesaran (2006) and Bai (2009).

Because the number of parameters in the standard fixed-effects model is the same as the number of individuals, the estimation of the fixed effects may be unreliable. Therefore, in order to investigate the individual heterogeneity, the number of parameters in the panel data model needs to be reduced. The grouped panel data model is an effective method of resolving this incidental-parameter problem. In a grouped panel data model, individuals in the same group are assumed to exhibit the same effect, which is called the group effect, and these group effects can exhibit the individual heterogeneity. Grouped panel data models have been studied in the past decade. For example, Lin & Ng (2012) investigated linear panel data models with time-varying grouped heterogeneity by using the $K$-means clustering algorithm. Bonhomme & Manresa (2015) developed the “grouped fixed-effects” (GFE) method of estimating the model parameters and derived the statistical properties of the GFE estimators when the sample sizes of the cross-section ($N$) and length of the time series ($T$) are large. Ando & Bai (2016) studied grouped panel data models with unobserved group factor structures and estimated the model parameters by minimizing the sum of least-squared errors with a shrinkage penalty. Ando & Bai (2016) also proved the consistency and asymptotic normality of these estimators when both $N$ and $T$ are large. Su, Shi & Phillips (2016) proposed using the C-Lasso method to estimate the parameters in heterogeneous linear panel data models with unknown group membership, where the slope parameters are heterogeneous across groups but homogeneous within a group. Su & Ju (2018) solved the penalized principal component method of estimation by extending the C-Lasso method to accommodate panel data with interactive fixed effects. These same authors also assumed that the individual slope parameters are heterogeneous across groups but homogeneous within a group.

In all these aforementioned studies, the unobserved heterogeneous elements are assumed to occupy the same space and belong to ball-shaped (spherical Gaussian) clusters. However, in many practical applications, a data object often has multiple attributes, many of which may belong to some low-dimensional subspaces. For example, for disease detection in newborns, various measurements such as blood chemistry and heart rate are collected on the infants, and the levels of those factors are recorded. Thus, each newborn is associated with a vector containing the values of the factors, and one can even form a newborn factor-level matrix in which each row contains the factor levels of a different newborn. Paediatricians often want to cluster groups of newborns based on the disease or diseases that they suffer from. Usually, each disease is correlated with a specific set of factors, which implies that points corresponding to newborns suffering from a given disease belong to a lower dimensional subspace (Kriegel, Kröger & Zimek, 2009). Therefore, the clustering of newborns based on their specific diseases, together with the identification of the relevant factors associated with each disease, can be formulated as a subspace clustering problem. The $K$-means algorithm clusters data around various cluster centres in the entire data space and estimates the cluster centres by minimizing the sum-of-squared distances from each data point to the nearest cluster centre. As a result, traditional clustering methods such as the $K$-means algorithm may not be entirely useful in these cases.

In this article, we characterize the unobserved effects of the $i$th unit at time $t$ as $\eta_{it} = f_t' \lambda_i$, where $f_t$ and $\lambda_i$ are each $r \times 1$ dimensional vectors and $\lambda_i, i = 1, \ldots, N$ belong to some
low-dimensional subspaces. A key feature of the low-dimensional subspaces is that they can accommodate more general data types, and the corresponding factor loadings, which also belong to the low-dimensional subspaces, can encapsulate the main direction of variation. Terada (2014) showed that subspace clustering is a more general clustering method which includes the conventional $K$-means clustering method as a special case. We propose a novel approach called least-squares subspace clustering (LSSC) to simultaneously estimate the model parameters and cluster the individual effects by using the least-squares criterion in combination with the subspace clustering principle. In the clustering of individual effects, we treat each individual effect as a vector in a high-dimensional space and then cluster these vectors into some low-dimensional subspaces. The main contributions of this article are as follows:

(i) The grouping is accomplished by limiting the data points to a specific subspace rather than according to the distance between points; this enables us to better reflect the underlying structure of the data and can also be used to model a more general data structure. In addition, we prove the consistency of the clustering procedure.
(ii) Our proposed model allows the covariates to be correlated with the factor structure. Monte Carlo simulation results show that the LSSC method performs well even when $T$ is small.
(iii) Our proposed method allows different groups to share common factors, and hence different groups of factors can be correlated. Thus, the model structure is able to capture the spatial structure of individual effects and reflect the cluster structure in real data.
(iv) We introduce a model selection criterion to choose the subspace dimensions consistently, which makes our proposed model and method of estimating its parameters more general and flexible.

The rest of this article is organized as follows: In Section 2, we describe the model and state some constraint conditions. In Section 3, we propose an algorithm for estimating the model parameters and subspace clustering simultaneously. In Section 4, we state the consistency of the subspace clustering and study the asymptotic properties of the estimators. In Section 5, we report the results of Monte Carlo simulation studies with different settings that illustrate the performance of our proposed method and compare its performance with that of the GFE method of estimation as well as the method proposed in Bai (2009). In Section 6, we explore some further considerations when the number of subspaces for factors, the dimensions of factors, and the dimensions of subspaces are unknown. For illustrative purposes, we investigate the relationship between income and democracy in Section 7 using LSSC. Section 8 consists of some concluding remarks. The proofs of all theoretical results, as well as the details concerning additional simulations, can be found in the accompanying Supplementary Material.

2. MODEL DESCRIPTION

Let $k$ be the number of subspaces, which is unknown but fixed, and $G = \{g_1, g_2, \ldots, g_N\}$ be the grouping of the cross-sectional units into the $k$ subspaces, where the subspace membership variable $g_i = j$ indicates that the $i$th unit belongs to the $j$th subspace with $g_i \in \{1, 2, \ldots, k\}$. Let $N_j$ denote the number of cross-sectional units within the $j$th subspace; the total number of units is $N = \sum_{j=1}^{k} N_j$. We consider the following panel data model with subspace factor structure:

$$y_{it} = x_{it}'\beta + f_{g_i,t}'\lambda_{g_i,t} + \epsilon_{i,t}, \quad i = 1, 2, \ldots, N, \quad t = 1, 2, \ldots, T, \quad (1)$$

where $y_{it}$ is the response variable for the $i$th unit observed at time $t$, $x_{it}$ is a $p \times 1$ observable vector, and $\beta$ is a $p \times 1$ vector of unknown regression coefficients. The factor loading vector $\lambda_{g_i,t} = (\lambda_{g_i,t}^1, \ldots, \lambda_{g_i,t}^r)'$ is an $r \times 1$ vector that represents the unobserved unit/individual effect for
the $i$th individual. Finally, $f_{g_i,t}$ is an $r \times 1$ vector of unobservable subspace-specific pervasive factors that affect only the units in the $g_i$th subspace, and $\varepsilon_{g_i,t}$ is the unit-specific error.

We assume that all the factor loadings or individual effects belong to an $r$-dimensional space, which contains some low-dimensional subspaces and these subspaces contain all the individual effects $\lambda_i$, $i = 1, 2, \ldots, N$. In mathematical terms, $\lambda_i; g_i \in S_j$ and $\bigcup_{i=1}^{N} S_j \subset \mathbb{R}^r$, where $S_j$ is the $j$th subspace. The covariate $x_{it}$ can be correlated to $\lambda_{g_i,t}$ alone, or it can be associated with both $\lambda_{g_i,t}$ and $f_{g_i,t}$ simultaneously. For example, $x_{it}$ can be a nonlinear function of $\lambda_{g_i,t}$ and $f_{g_i,t}$. Stacking the observations over $t$, we have $F = (f_{11}, f_{21}, \ldots, f_{T1}, f_{12}, f_{22}, \ldots, f_{T2}, \ldots, f_{1T}, f_{2T}, \ldots, f_{kT})$. Furthermore, if we let $\tilde{F}_j$ be the vector of factors for the $j$th subspace, then we have $F_j = (f_{g_i=j,1}, f_{g_i=j,2}, \ldots, f_{g_i=j,T})'$. Similarly, we have $A_j = (\lambda_{j,1}, \lambda_{j,2}, \ldots, \lambda_{j,N_j})'$. We also consider the constraints $F_j'F_j/T = I_r(j = 1, 2, \ldots, k)$ and $A_j'A_j(j = 1, 2, \ldots, k)$ being diagonal for the issue of identifiability as described in Bai (2009), Ando & Bai (2016), and Stock & Watson (2002). Our aim is to estimate the parameters $\beta, A_j, F_j, j = 1, 2, \ldots, k$ for each subspace, and identify the subspace membership $G = \{g_1, g_2, \ldots, g_N\}$ and the bases $B_1, B_2, \ldots, B_k$ of the orthogonal spaces of $S_1, \ldots, S_k$ (denoted as $S_1^\perp, \ldots, S_k^\perp$) simultaneously.

3. PARAMETER ESTIMATION AND CLUSTERING

In this section, we propose a method of estimating the model parameters and subspace clustering simultaneously. This method is a natural generalization and combination of both least-squares estimation and the $K$-means clustering algorithm. Thus, what we propose is innovative in comparison to existing piecemeal approaches.

3.1. Estimation Procedure

For a given number of subspaces $k$, the objective function is

$$Q(\beta, F_1, \ldots, F_k, A_1, \ldots, A_k, B_1, \ldots, B_k, G) = \sum_{j=1}^{k} \sum_{i: g_i=j} \|y_i - x_i' \beta - F_{g_i,t} \lambda_{g_i,t}\|^2,$$

where $y_i = (y_{i1}, y_{i2}, \ldots, y_{iT})$, and $x_i = (x_{i1}, x_{i2}, \ldots, x_{iT})$. The estimator of the vector of model parameters $\{\hat{\beta}, \hat{F}_1, \ldots, \hat{F}_k, \hat{A}_1, \ldots, \hat{A}_k, \hat{B}_1, \ldots, \hat{B}_k, \hat{G}\}$ is defined as

$$\{\hat{\beta}, \hat{F}_1, \ldots, \hat{F}_k, \hat{A}_1, \ldots, \hat{A}_k, \hat{B}_1, \ldots, \hat{B}_k, \hat{G}\} = \arg \min_{\beta, F_{g_i}, \lambda_{g_i}, A_j, B_j, G} Q(\beta, F_1, \ldots, F_k, A_1, \ldots, A_k, B_1, \ldots, B_k, G)$$

subject to the constraints $F_j'F_j/T = I_r(j = 1, \ldots, k)$, and $A_j'A_j(j = 1, \ldots, k)$ is diagonal, where $A_j' = (\lambda_{j,1}, \ldots, \lambda_{j,N_j})$ represents the $r \times N_j$ factor loading matrix ($j = 1, \ldots, k$) for the subspace-specific factors, which belong to $k$ different subspaces embedded in the $r$-dimensional space. In Equation (2), span$\{B_j\}$ represents the subspace spanned by the basis $B_j$, and span$\{B_j\}^\perp$ denotes the subspace orthogonal to span$\{B_j\}$. These constraints and assumptions are needed to ensure that the model is identifiable. Here, we aim to estimate the model parameters and to cluster $\lambda_i \in \mathbb{R}^r, i = 1, 2, \ldots, N$ into the $k$ different subspaces simultaneously. Unlike existing classification methods, we approach this challenging problem from a subspace clustering point of view. The major idea is to divide the space $\mathbb{R}^r$ into several subspaces and project $\lambda_i$ ($i = 1, 2, \ldots, N$) into the nearest subspace to resolve the classification problem.

The constrained minimization of the objective function $Q(\cdot)$ identified in Equation (2) can be obtained by using the following iterative algorithm:
Step 1. Initialize the starting value $\beta^{(0)}$ and set $h = 0$.

Step 2. Given the value of $\beta = \beta^{(h)}$, we define

$$y_i^* = y_i - x_i \beta = F \lambda_i + \epsilon_i,$$

which is a pure factor model. We can readily obtain $\Lambda' = (\lambda_1, \ldots, \lambda_N)$.

Step 3. Given $\Lambda' = (\lambda_1, \ldots, \lambda_N)$, using the subspace clustering method, we next derive the bases $B_1, \ldots, B_k$ of the orthogonal spaces of $S_1, \ldots, S_k$ (which we denote by $S^1_k, \ldots, S^k_k$) and

$$g_i = \arg \min_{j=1,\ldots,k} \| B_j \lambda_i \|, \quad i = 1, 2, \ldots, N. \tag{3}$$

Step 4. Given $\beta = \beta^{(h)}$ and $g_i, \ i = 1, 2, \ldots, N$, we can obtain the estimators of $F$ and $\Lambda$ in each subspace by using a method similar to the one outlined in Step 2. These estimators are denoted by $F_1, \ldots, F_k$ and $\Lambda_1, \ldots, \Lambda_k$.

Step 5. Given $\Lambda_1, \ldots, \Lambda_k, F_1, \ldots, F_k$ and $g_i, \ i = 1, \ldots, N$, we now define

$$y_i - F_{g_i} \hat{\lambda}_{g,i} = x_i \beta + \epsilon_i, \quad i = 1, \ldots, N,$$

from which we obtain the updated least-squares estimator $\beta^{(h+1)}$; then set $h = h + 1$.

Step 6. Repeat Steps 2–5 until convergence occurs.

Although the least-squares objective function is not globally convex (Bai, 2009), from the results of the Monte Carlo simulation studies that we report in Section 5, our proposed algorithm is robust with respect to the starting value under large $N$ and large $T$. In our numerical studies, we use the least-squares method to derive an initial value $\beta^{(0)}$ by ignoring the unobserved group factor structures. In our experience, this choice of initial value seems to have good convergence properties. In practice, if one is concerned that the algorithm may converge to a local optimizer, we suggest using different randomly selected starting values and then choosing the solution that yields the smallest value of the objective function if those various solutions are not of the same value.

To evaluate the complexity of our proposed iterative procedure, we investigated the complexity of Steps 2–4 in the above algorithm. In Step 2 we need to obtain the $r$ largest singular values of the $T \times N$ data matrix in order to obtain the values $F$ and $\Lambda$, which entails a complexity of $O(\delta_{NT}^2 \delta_{NT}) + O(\delta_{NT}^3)$ operations, where $\delta_{NT} = \min[N,T], \delta_{NT} = \max[N,T]$. Step 3 of the algorithm requires us to sort the $\lambda_i$ according to their projection distances to these subspaces; the corresponding computational cost is $O(Nkr)$. In Step 4, the least-squares estimation involves a complexity of size $O(p^2N)$ because of the inverse operation. Thus, the complexity of the above iterative process is $O(\delta_{NT}^2 \delta_{NT}) + O(\delta_{NT}^3) + O(Nkr) + O(p^2N) = O(\delta_{NT}^2 \delta_{NT})$ operations.

### 3.2. Subspace Clustering for Factor Loadings

In this subsection, we outline the procedure of subspace segmenting for the factor loadings $\lambda_i, \ i = 1, \ldots, N$. We assume that the $i$th factor loading $\lambda_i \in \mathbb{R}^r$ $(i = 1, \ldots, N)$ belongs to $k$ different subspaces, where the dimensions of these $k$ subspaces are $d_1, d_2, \ldots, d_k$, $0 < d_j < r, \ j = 1, 2, \ldots, k$. For the $j$th subspace $S_j \subset \mathbb{R}^r$ with dimension $d_j$, we select a basis $B_j = [b_{j,1}, \ldots, b_{j,r-d_j}] \in \mathbb{R}^{r \times (r-d_j)}$ for its orthogonal complement $S^\perp_j$. Using this notation, we can obtain the following equation for the $j$th subspace $S_j$ and the $i$th factor loading $\lambda_i$:

$$\{ \lambda_i \in S_j \} = \{ \lambda_i \in \mathbb{R}^r : B_j^T \lambda_i = 0 \} = \left\{ \lambda_i \in \mathbb{R}^r : \bigwedge_{m=1}^{r-d_j} (b_{jm}^T \lambda_i = 0) \right\}. \tag{4}$$
Since \( \lambda_i \in \mathbb{R}^r \) belongs to \( \bigcup_{j=1}^{k} S_j \) if and only if \( (\lambda_i \in S_1) \lor \ldots \lor (\lambda_i \in S_k) \), where the notation \( \lor \) represents the “or” operator, this condition is equivalent to
\[
\bigvee_{j=1}^{k} (\lambda_i \in S_j) \iff \bigwedge_{j=1}^{k} \bigvee_{m=1}^{r-d_j} (b_j^T \lambda_i = 0) \iff \bigwedge_{\sigma} \bigvee_{j=1}^{k} (b_j^T \lambda_i = 0),
\]
where the notation \( \bigwedge \) represents the “and” operator, and \( \sigma \) is a particular choice of the normal vector \( b_j^T \) from the \( j \)th basis \( B_j \). Note that the right-hand side of Equation (5) is obtained by exchanging the “and” and “or” operators using De Morgan’s laws. Since
\[
\bigvee_{j=1}^{k} (b_j^T \lambda_i = 0) \iff \prod_{j=1}^{k} (b_j^T \lambda_i = 0) \iff p_{k\sigma}(\lambda_i) = 0,
\]
which is a homogeneous polynomial of degree \( k \) in \( r \) variables, we can write each of the polynomials as
\[
p_{k\sigma}(\lambda_i) = \prod_{j=1}^{k} (b_j^T \lambda_i) = c_k^T \nu_k(\lambda_i) = \sum_{0 \leq k_1 \leq 1, \ldots, k_r \leq r} c_{k_1, k_2, \ldots, k_r} \lambda_{i_1}^{k_1} \lambda_{i_2}^{k_2} \ldots \lambda_{i_r}^{k_r} = 0, \tag{7}
\]
where \( c_k \) is a vector of polynomial coefficients, \( c_{k_1, k_2, \ldots, k_r} \) is the polynomial coefficient, \( \lambda_i = (\lambda_{i_1}, \ldots, \lambda_{i_r}) \) and \( \nu_k : \mathbb{R}^r \to \mathbb{R}^{M_k(r)} \) is the Veronese map of degree \( k \) (Fischler & Bolles, 1981), which is also known in machine learning as the polynomial embedding, defined as \( \nu_k : [\lambda_{i_1}, \ldots, \lambda_{i_r}]^T \mapsto [\lambda_{i_1}^1, \ldots, \lambda_{i_r}^k]^T \) with \( I \) being chosen in the degree-lexicographic order, \( \lambda_{i}^t = \lambda_{i_1}^{k_1} \lambda_{i_2}^{k_2} \ldots \lambda_{i_r}^{k_r} \) and the dimension \( M_k(r) = \frac{c^r - 1}{c^{k-1}} \).

Since the polynomial in Equation (7) can be satisfied by all the factor loadings \( \lambda_i, i = 1, 2, \ldots, N \), we can then use these factor loadings to obtain the subspaces. Although the polynomial equations identified in Equation (7) are nonlinear in each point \( \lambda_i \), these polynomials are actually linear in the vector of polynomial coefficients \( c_k \). Indeed, since each polynomial \( p_{k\sigma}(\lambda_i) = c_k^T \nu_k(\lambda_i) \) must be satisfied by every data point, we can obtain \( c_k^T \nu_k(\lambda_i) = 0 \) for all \( i = 1, 2, \ldots, N \).

Suppose that \( I_k \) is the space of the vector of polynomial coefficients \( c_k \) of all the homogeneous polynomials that vanish in the \( k \) subspaces; then the vector of polynomial coefficients of the factorizable polynomial defined in Equation (6) spans a (possibly proper) subspace in \( I_k \), i.e., \( \text{span}_{\sigma} \{ p_{k\sigma} \} \subseteq I_k \). As every vector \( c_k \) in \( I_k \) represents a polynomial that vanishes on all the data points (on the subspaces), the vector \( c_k \) must satisfy the system of linear equations
\[
c_k^T V_k(r) = c_k^T [\nu_k(\lambda_1), \ldots, \nu_k(\lambda_N)] = \mathbf{0}^T, \tag{8}
\]
where \( V_k(r) \in \mathbb{R}^{M_k(r) \times N} \) is the embedded data matrix. It follows that \( I_k \subseteq \text{null}(V_k(r)) \).

**Remark 1.** The zero set of each vanishing polynomial \( p_k(\lambda_i), i = 1, 2, \ldots, N \) is a surface in \( \mathbb{R}^r \); therefore, the derivative of \( p_k(\lambda_i) \) at \( \lambda_i \in S_j \), which we denote by \( \text{D}p_k(\lambda_i) \), gives a vector normal to the surface. Since a union of subspaces is locally flat, i.e., in a neighbourhood of \( \lambda_i \), this surface is merely the surface \( S_j \), and the derivative at \( \lambda_i \) lies in the orthogonal complement \( S_j^\perp \) of \( S_j \). By evaluating the derivatives of all the polynomials in \( I_k \) at the same point \( \lambda_i \), we obtain a set of normal vectors that span the orthogonal complement of \( S_j \).
Following the results stated as Theorem 3 of Vidal & Sastry (2005), we can obtain a set of polynomials \( p_k(\lambda_i), i = 1, \ldots, N \) with coefficients equal to the eigenvectors in the null space of \( V_k(r) \). By evaluating the derivatives \( Dp_k(\lambda_i) \) at each \( \lambda_i, i = 1, \ldots, N \), we can obtain a set of vectors orthogonal to the subspace in which the points lie. Note that the generalized principal components method of analysis (Vidal & Sastry, 2005, 2016) uses reliable samples from a subspace to segment the dataset; however, in the presence of noise, our sample may not be reliable. Here, for each sample, we assume that the sample could be obtained from all the candidate co-dimension classes, and the sample is voted by the dominant vectors of \( Dp_k(\lambda_i) \) as a basis. Finally, the base associated with the highest vote will be used as the normal vectors perpendicular to the subspaces as suggested by Yang et al. (2005). After obtaining the orthogonal bases of those subspaces, we can assign \( \lambda_i \) to the subspace \( j^* \), where \( j^* = \arg \min_{j=1,\ldots,k} \| B_j^T \lambda_i \| \).

### 4. ASYMPTOTIC PROPERTIES

In this section, we characterize the asymptotic properties of the estimators as \( N \) and \( T \) tend to infinity. We prove that the estimated clustering converges to the corresponding true subspaces under some conditions. In particular, we follow the method of Pollard (1981) to establish the consistency of subspace clustering.

#### 4.1. Consistency of Clustering Procedure

The proposed clustering procedure prescribes a criterion for partitioning a set of points into \( k \) subspaces. To divide the factor loadings \( \lambda_1, \ldots, \lambda_N \) in \( \mathbb{R}^r \), we first choose \( k \) (\( k \) is fixed) cluster subspaces \( S_1, S_2, \ldots, S_k \) with dimensions \( d_1, d_2, \ldots, d_k \), respectively, that minimize

\[
W_N = \frac{1}{N} \sum_{i=1}^{N} \min_{1 \leq j \leq k} \phi(\Delta(\lambda_i, S_j)),
\]

where \( \lambda_1, \ldots, \lambda_N \) can be viewed as \( N \) vectors of the sample points and \( \Delta(\lambda_i, S_j) \) represents the angle between the vector \( \lambda_i \) and the subspace \( S_j \), which is a value in \([0, \pi/2]\). Let \( \| \lambda_i \| = 1 \), \( b_1, \ldots, b_{m} \) be an orthonormal basis of \( S_j \), and \( \theta \) be the angle between \( \lambda_i \) and \( S_j \); then \( \sin(\theta) = \| \lambda_i - \sum_{\ell=1}^{m} b_{j,\ell} (b_{j,\ell}^T \lambda_i) \| = \sqrt{1 - \sum_{\ell=1}^{m} (b_{j,\ell}^T \lambda_i)^2} \). Thus, clustering by angles is equivalent to clustering by projecting \( \lambda_i \) to the nearest subspace. Here, the function \( \phi \) must satisfy some regularity conditions that ensure \( \phi \) is continuous and nondecreasing, with \( \phi(0) = 0 \). For any subspaces \( x \) and \( y \), \( \Delta(x, y) \in [0, \pi/2] \), therefore, \( \phi(\Delta(x, y)) \) must belong to a compact set.

Since \( \max_{1 \leq i \leq N} \| \hat{\lambda}_i - \lambda_i \| = o_p(1) \), we can show that the empirical distribution function converges uniformly to the corresponding true distribution function by the strong law of large numbers and the Glivenko–Cantelli theorem, i.e.,

\[
\lim_{N \to \infty} \sup_{x \in \mathbb{R}^r} |\hat{P}_N(x) - P(x)| = 0,
\]

where \( \hat{P}_N(x) = \frac{1}{N} \sum_{i=1}^{N} I_{\{ \hat{\lambda}_i \leq x \}} \), with \( I_A = 1 \) if \( A \) is true and 0 otherwise, is the empirical distribution function and \( P(x) \) is the true distribution function. Therefore, clustering for the estimate \( \hat{\lambda}_i \) obtained by using our proposed method is equivalent to clustering for the true value \( \lambda_i \) \((i = 1, 2, \ldots, N)\). Since \( \phi(\Delta(\lambda, S)) \) is an increasing function of the angle deviation which can be used to define a within-cluster sum of angle deviations, the criterion considered here minimizes the within-cluster sum of angle deviations.
We assume that \( \{ \lambda_1, \ldots, \lambda_N \} \) is a sample of independent observations on some probability measure \( P \). Here, we consider the empirical measure

\[
W(S, \hat{P}_N) = \int \min_{S \in S} \phi(\Delta(\lambda_i, S)) \hat{P}_N(d\lambda),
\]

where \( S \) is a set of subspaces. For a fixed set of subspaces \( S \), we can show that

\[
W(S, \hat{P}_N) \xrightarrow{a.s.} W(S, P) = \int \min_{S \in S} \phi(\Delta(\lambda_i, S)) P(d\lambda),
\]

where \( \lambda_1, \ldots, \lambda_N \) are the \( N \) vectors of the sample points which can be clustered by minimizing the within-cluster sum of angle deviations. Let \( S_N \) be the set of subspaces that minimizes \( W(\cdot, \hat{P}_N) \) (i.e., the set of optimal clustered subspaces based on the samples), and \( \tilde{S} \) be the set of subspaces that minimizes \( W(\cdot, P) \). Provided that \( \tilde{S} \) can be uniquely determined, we expect that \( S_N \) should lie close to \( \tilde{S} \).

For a probability measure \( Q \) on \( \mathbb{R}^r \) and a finite set of subspaces \( S \) of \( \mathbb{R}^r \), we define

\[
\Phi(S, Q) = \int \min_{S \in S} \phi(\Delta(\lambda, S)) Q(\lambda)
\]

and

\[
m_k(Q) = \inf\{ \Phi(S, Q) : S \text{ contains } k \text{ or fewer subspaces}; \text{ and } d_1, \ldots, d_k \text{ are known} \}.
\]

For a given value of \( k \), the set of optimal clustered subspaces based on the samples \( S_N = S_N(k) \) is chosen to satisfy \( \Phi(S_N, \hat{P}_N) = m_k(\hat{P}_N) \), and the set of optimal population clustered subspaces \( \tilde{S} = \tilde{S}(k) \) is chosen to satisfy \( \Phi(\tilde{S}, P) = m_k(P) \).

To define the distance measures, we have the following assumption:

**Assumption A.** Suppose that \( \int \phi(||x||) P(dx) < \infty \) and that \( m_j(P) > m_k(P) \) for \( j = 1, \ldots, k - 1 \).

Our aim here is to prove a consistency result for the cluster subspaces such that \( S_N \xrightarrow{a.s.} \tilde{S} \).

To show \( S_N \xrightarrow{a.s.} \tilde{S} \), we first consider the subspace distance defined in Wang, Wang & Feng (2006).

**Definition 1.** The symmetric distance between any \( m \)-dimensional subspace \( U \) and \( n \)-dimensional subspace \( \tilde{U} \) is defined as

\[
D(U, \tilde{U}) = \max(\overline{D}(U, \tilde{U}), \overline{D}(\tilde{U}, U)) = \sqrt{\max(m, n) - \sum_{i=1}^{m} \sum_{j=1}^{n} (\tilde{u}_j^i u_i)^2},
\]

where \( (u_1, \ldots, u_m) \) and \( (\tilde{u}_1, \ldots, \tilde{u}_n) \) are the bases of \( U \) and \( \tilde{U} \), respectively. Note that this subspace distance satisfies the triangle inequality

\[
D(U, \tilde{U}) \leq D(U, W) + D(W, \tilde{U}),
\]

where \( W \) is any nonempty subspace.
Remark 2. The angle $\Delta(\cdot, \cdot)$ used in Equation (9)

$$\Delta(U, \bar{U}) = \sqrt{\min(m, n) - \sum_{i=1}^{m} \sum_{j=1}^{n} (\bar{u}_i u_j)^2},$$

is different from the distance measure defined in Definition 1 for measuring the distance between subspaces. In fact, the angle $\Delta(\cdot, \cdot)$ projects the low-dimensional subspace onto the high-dimensional subspace, while the distance $D(\cdot, \cdot)$ projects the high-dimensional subspace onto the low-dimensional subspace. In order to avoid different dimensional subspaces being treated as the same subspace, $D(\cdot, \cdot)$ is used to characterize the distance between two subspaces.

For example, consider a two-dimensional plane and a line which is parallel to this plane as two subspaces. If we use the distance $\Delta(\cdot, \cdot)$ to characterize the distance between these two subspaces, it is likely to yield the result that these two subspaces are simply treated as the same subspace because the angle between these two subspaces is 0. However, using the distance measure $D(\cdot, \cdot)$ enables us to avoid this issue. Here, we further define a distance measure that is similar to the Hausdorff distance.

Definition 2. Let $\mathcal{X}$ and $\mathcal{Y}$ be two nonempty compact subsets that contain multiple subspaces. We define their Hausdorff distance $D_H(\mathcal{X}, \mathcal{Y})$ via

$$D_H(\mathcal{X}, \mathcal{Y}) = \max\{\sup_{x \in \mathcal{X}} \inf_{y \in \mathcal{Y}} D(x, y), \sup_{y \in \mathcal{Y}} \inf_{x \in \mathcal{X}} D(x, y)\}, \quad (14)$$

where $x \in \mathcal{X}$ is a subspace rather than a point, and the measure $D(\cdot, \cdot)$ is the distance between two subspaces defined in Definition 1.

According to Definition 2, we have $D_H(\mathcal{X}, \mathcal{Y}) < \delta$ if and only if every subspace of $\mathcal{X}$ is within the distance $\delta$ of the subspaces of $\mathcal{Y}$, and vice versa. Suppose $\mathcal{X}$ contains exactly $k$ distinct subspaces, and that $\delta$ is chosen to be a value less than half of the minimum distance between the subspaces of $\mathcal{X}$. Then, if $\mathcal{Y}$ is any set of $k$ or fewer subspaces for which $D_H(\mathcal{X}, \mathcal{Y}) < \delta$, $\mathcal{Y}$ must contain exactly $k$ distinct subspaces. Therefore, the almost sure convergence of $\mathcal{X}_N$ in the above sense of distance could be translated into the almost sure convergence of subspaces with a suitable labelling. By definition, for any two subspaces $S_1$ and $S_2$, if $D(S_1, S_2) < \delta$, then $\Delta(S_1, S_2) < \delta$. We will prove Theorem 1 (see below) to establish the consistency of our clustering procedure. Since the conclusion of the theorem is expressed in terms of almost sure convergence, there might be aberrant null sets of subspaces $S$’s for which the convergence does not hold. In order to estimate the parameters in the model identified in Equation (1) and to prove the consistency of the estimators, similar to Bonhomme & Manresa (2015) and Ando & Bai (2016), we add Assumption B, namely that each group must have a certain proportion of individuals. This assumption also guarantees the uniqueness because the null set situation is excluded.

Assumption B. All units are divided into a finite number of subspaces $k$, each of them containing $N_j$ units such that $0 < a < N_j / N < \tilde{a} < 1$.

For notational simplicity and clarity, in the following theorem we assume that $\lambda_i, i = 1, \ldots, N$ is known.

Theorem 1. Suppose Assumptions A and B hold, and for each $j = 1, 2, \ldots, k$ there exists a unique set of subspaces $\tilde{S}(j)$ that satisfies $\Phi(\tilde{S}(j), P) = m_j(P)$. Then $S_N \xrightarrow{a.s.} \tilde{S}(k)$, and $\Phi(S_N, P_N) \xrightarrow{a.s.} m_k(P)$.

The proof of Theorem 1 can be found in Section S2 of the accompanying Supplementary Material.
4.2. Consistency of the Estimators

In this subsection, we discuss the asymptotic properties of the proposed estimators. Recall that the proposed estimators can be obtained by minimizing the objective function \( Q(\cdot) \) identified in Equation (2) subject to the constraints \( g_i = \arg \min_{j \in [1, \ldots, k]} \| B_j^T \lambda_i \|, \ F_j^T F_j / T = I_j (j = 1, \ldots, k) \), and \( \Lambda_j' \Lambda_j (j = 1, \ldots, k) \) being diagonal. While the consistency of the subspace clustering procedure was discussed in Section 4.1, Theorems 2–4 establish the properties of the estimators when \( T \) and \( N \) are large. To show the various properties of the estimators, Assumptions C, D, E, and F, as presented in Bai (2009) and Ando & Bai (2016), are needed; these four assumptions are identified in Section S1 of the accompanying Supplementary Material.

Theorem 2. Suppose that Assumptions A–E hold as \( N \to \infty \) and \( T \to \infty \); then the following statements hold:

(i) \( \| \hat{\beta} - \beta^0 \| = o_p(1) \),

(ii) \( \| P_{\mathcal{F}_j} - P^0 \| = o_p(1), j = 1, \ldots, k \).

Theorem 3 establishes the consistency of the estimator of group membership.

Theorem 3. Suppose that Assumptions A–E hold; then for all \( \tau > 0 \) and \( T, N \to \infty \)

\[
P \left( \sup_{i \in [1, \ldots, N]} \left| \hat{g}_i (\hat{\beta}, \hat{\Phi}, \hat{\Lambda}, \ldots, \hat{\Lambda}_k) - g^0_i \right| \right) = o(1) + o(N/T^\tau).
\]

Theorem 4 establishes the asymptotic normality of our estimator.

Theorem 4. Suppose that Assumptions A–F hold and \( T/N \to \rho > 0 \); then

\[
\sqrt{NT}(\hat{\beta} - \beta^0) \xrightarrow{d} N(v_0, V_\beta(F_1^0, \ldots, F_k^0, B_1^0, \ldots, B_k^0)),
\]

where \( v_0 \) is the probability limit of

\[
v = \sqrt{\frac{T}{N}} \sum_{j=1}^{k} D(F_1^0, \ldots, F_k^0, B_1^0, \ldots, B_k^0)^{-1} \eta_j + \sqrt{\frac{T}{N}} \sum_{j=1}^{k} D(F_1^0, \ldots, F_k^0, B_1^0, \ldots, B_k^0)^{-1} \xi_j
\]

with

\[
\eta_j = - \frac{1}{N_j} \sum_{i : g_i^0 = j} \sum_{\varepsilon : g_i^0 = j} \frac{(x_i - V_j, \lambda_j') F_j^0}{T} \left( F_j^0 F_j^0 / T \right)^{-1} \left( \Lambda_j^0 / N_j \right)^{-1} \lambda_j^0, \varepsilon \left( \frac{\varepsilon e_i e_i}{T} \right),
\]

\[
\xi_j = \sqrt{\frac{T}{N_j T}} \sum_{i : g_i^0 = j} \sum_{\varepsilon : g_i^0 = j} x_i^j M_j^0 \Omega_k F_j^0 (F_j^0 F_j^0 / T)^{-1} (\Lambda_j^0 / N_j)^{-1} \lambda_j^0, i
\]

\[
D(F_1^0, \ldots, F_k^0, B_1^0, \ldots, B_k^0) = \frac{1}{NT} \sum_{j=1}^{k} \sum_{i : g_i^0 = j} x_i^j M_j^0 x_i
\]

\[
- \frac{1}{NT} \sum_{j=1}^{k} \left[ \frac{1}{N_j} \sum_{i : g_i^0 = j} \sum_{\varepsilon : g_i^0 = j} x_i^j M_j^0 \varepsilon \varepsilon c_j, i \right]
\]
\[ V_{\beta}(F_0^0, \ldots, F_0^k, B_0^0, \ldots, B_0^k) = D_0(F_0^0, \ldots, F_0^k, B_0^0, \ldots, B_0^k)^{-1} J_0(F_0^0, \ldots, F_0^k, B_0^0, \ldots, B_0^k) \times D_0(F_0^0, \ldots, F_0^k, B_0^0, \ldots, B_0^k)^{-1}, \]

where \( D_0(F_0^0, \ldots, F_0^k, B_0^0, \ldots, B_0^k) \) is the probability limit of \( D(F_0^0, \ldots, F_0^k, B_0^0, \ldots, B_0^k) \). The matrix \( J_0(F_0^0, \ldots, F_0^k, B_0^0, \ldots, B_0^k) \) is defined in Assumption F, \( V_{\beta} = N_j^{-1} \sum_{i=1}^{N_j} c_{i,j} \epsilon_i x_i x_i^T \), and \( M_j = \frac{1}{T} F_j B_j^{0T} F_j^T \).

The proofs of Theorems 2–4 can be found in Section S3 of the accompanying Supplementary Material.

5. MONTE CARLO SIMULATION STUDIES

In this section, we report the results of Monte Carlo simulation studies involving different settings that illustrate our proposed methodology and illuminate the finite sample properties of our proposed methods. We assume that \( \lambda_i \) (i = 1, 2, ..., \( N \)) derives from different subspaces with dimension \( d_1, d_2, \ldots, d_k \) and also that \( \lambda_i \) can follow different probability distributions over different subspaces. Furthermore, we consider that \( \lambda_i, i = 1, 2, \ldots, N \), allow moderate noise. All the results that we report are based on 100 simulations.

5.1. Setting 1

We consider the situation involving three different subspaces in \( \mathbb{R}^3 \) with known dimensions \( d_1, d_2, \) and \( d_3 \), i.e., \( k = 3 \) and \( r = 3 \). The bases of the three subspaces with dimensions \( d_1, d_2, \) and \( d_3 \) are represented by \( \alpha_1, \alpha_2, \) and \( \alpha_3 \), respectively. Let \( N_{\epsilon_1 \times \epsilon_2}(\mu, \sigma^2) \) be an \( \epsilon_1 \times \epsilon_2 \) matrix whose elements are random variables that are independent and identically distributed as normal with mean \( \mu \) and variance \( \sigma^2 \). In this setting, we generate the panel data \( y_{ij} \) in the \( j \)th subspace \( (j = 1, 2, 3; i = 1, 2, \ldots, N_j; t = 1, 2, \ldots, T) \) based on the panel data model identified in Equation (1) with \( N_1 = N_2 = N_3 = N = 100, T = 6 \), and the following structure:

- \( \lambda_i \sim N_{r \times 1}(1, 1) \) with random noise from a normal distribution with mean 0 and variance 0.1, i.e., \( \Lambda_j = N_{N_j \times d_j}(1, 1) \alpha_j' + N_{N_j \times r}(0, 0.1) \), \( \alpha_j \sim N_{r \times d_j}(0, 1) \) \( j = 1, 2, 3 \);
- \( p = 2 \) with \( \beta = (\beta_1, \beta_2)' = (1, 2)' \);
- the covariate \( X \) is a \( T \times N \times p \) array with
  \[ X_{-1} = \mu_1 + c_1 F \Lambda' + \tau \Lambda' + \eta_1, X_{-2} = \mu_2 + c_2 F \Lambda' + \tau \Lambda' + \eta_2, \]
  where \( X_{-1} \) and \( X_{-2} \) are \( T \times N \) matrices, \( \mu_1, \mu_2 \) are \( T \times N \) matrices such that all the elements are 1, and \( c_1 = c_2 = 0.5 \);
- \( F \Lambda = (F_1 \Lambda_1', F_2 \Lambda_2', F_3 \Lambda_3') \), where \( F_1, F_2, F_3 \) are \( T \times r \) matrices that satisfy \( F_j' F_j / T = I_r \), \( j = 1, 2, 3 \);
- \( \eta_1 \sim N_{T \times N}(0, 1) \) and \( \eta_2 \sim N_{T \times N}(0, 1) \);
- \( \tau \) is a \( T \times r \) matrix such that all the elements are 1;
- the random error \( \epsilon_{i,t} \sim \text{i.i.d.} N(0, 0.5) \), \( i = 1, 2, \ldots, N, t = 1, 2, \ldots, T \).

In this study, we compare the performance of our proposed LSSC method with that of the GFE method (Bonhomme & Manresa, 2015) and the estimation method proposed by Bai (2009) (BAI) in terms of the estimated biases and root-mean-squared errors (RMSEs). The simulated biases and RMSEs of the estimators obtained using the GFE, BAI, and LSSC methods of estimation for Setting 1 are summarized in Table 1. We observe that our proposed LSSC method shows better performance than both the GFE and BAI methods in terms of biases and RMSEs.

DOI: 10.1002/cjs.11642
TABLE 1: Simulated biases and root-mean-squared errors (RMSEs) of the GFE, BAI, and LSSC estimation methods for Setting 1.

| Dimension of subspaces | GFE | BAI | LSSC |
|------------------------|-----|-----|------|
|                        | Bias| RMSE| Bias| RMSE| Bias| RMSE |
| $d_1 = d_2 = d_3 = 1$  | $\beta_1$ | 0.1271 | 0.1304 | 0.0067 | 0.0121 | 0.0010 | 0.0035 |
|                        | $\beta_2$ | 0.1255 | 0.1295 | 0.0066 | 0.0116 | 0.0011 | 0.0035 |
| $d_1 = d_2 = d_3 = 2$  | $\beta_1$ | 0.2042 | 0.2082 | 0.2821 | 0.3141 | 0.0225 | 0.0518 |
|                        | $\beta_2$ | 0.2042 | 0.2084 | 0.2819 | 0.3120 | 0.0218 | 0.0503 |
| $d_1 = d_2 = 2, d_3 = 1$ | $\beta_1$ | 0.1530 | 0.1580 | 0.1871 | 0.2258 | 0.0084 | 0.0765 |
|                        | $\beta_2$ | 0.1571 | 0.1620 | 0.1884 | 0.2269 | 0.0111 | 0.0740 |
| $d_1 = 2, d_2 = d_3 = 1$ | $\beta_1$ | 0.1770 | 0.1793 | 0.0898 | 0.1272 | 0.0034 | 0.0144 |
|                        | $\beta_2$ | 0.1786 | 0.1810 | 0.0906 | 0.1262 | 0.0036 | 0.0149 |

TABLE 2: Simulated average misclassified rate of GFE and LSSC clustering methods for Setting 1.

| Dimension of subspaces | GFE | LSSC |
|------------------------|-----|------|
|                        | Bias| RMSE |
| $d_1 = d_2 = d_3 = 1$  | 0.2982 | 0.0660 |
| $d_1 = d_2 = d_3 = 2$  | 0.2713 | 0.0908 |
| $d_1 = d_2 = 2, d_3 = 1$ | 0.3428 | 0.1511 |
| $d_1 = 2, d_2 = d_3 = 1$ | 0.3314 | 0.1416 |

It is noteworthy that although the theoretical proofs of the asymptotic properties require $T$ to be large, the simulation results show that our proposed method performs well even when $T$ is small.

To compare the performance of the clustering methods, we summarize the simulated average misclassification rates of the clustering methods based on GFE and our proposed LSSC in Table 2. The simulated misclassification rates of the LSSC method are smaller than the corresponding misclassification rates of the GFE method. To verify the consistency of the proposed LSSC method, in addition to $N_1 = N_2 = N_3 = N = 100$, we considered various other sample sizes $N = 50, 200, 300, \text{ and } 500$ in order to investigate the effect of the sample size on the biases and RMSEs. We also adopted different values for the time period $T = 5, 10, 30, 50, 100$ to study the effect of the time period on the biases and RMSEs of the parameter $\beta$. These results are summarized in Tables 3 and 4. Clearly, the biases and RMSEs of the LSSC estimators decrease as the sample size $N$ or time period $T$ increases, empirically verifying the result stated in Theorem 2.

5.2. Setting 2

The method of estimation proposed by Ando & Bai (2016) is effective for large $N$ and $T$ when the regressors are not correlated with factors and factor loadings, but it does not perform well when the regressors and the factors are correlated, such as the scenario in Setting 1. To compare our proposed method of estimation with that of Ando & Bai (2016), we considered the following choices:
Table 3: Simulated biases and root-mean-squared errors (RMSEs) of the LSSC estimator with $d_1 = 2, d_2 = 2, d_3 = 2$ in Setting 1 with different sample sizes $N_1 = N_2 = N_3 = N$.

| Sample size ($N$) | $\beta_1$ Bias | $\beta_1$ RMSE | $\beta_2$ Bias | $\beta_2$ RMSE |
|-------------------|----------------|----------------|----------------|----------------|
| 50                | 0.0701         | 0.0691         | 0.0304         | 0.0304         |
| 100               | 0.0525         | 0.0524         | 0.0225         | 0.0227         |
| 200               | 0.0525         | 0.0537         | 0.0214         | 0.0212         |
| 300               | 0.0378         | 0.0381         | 0.0167         | 0.0169         |
| 500               | 0.0018         | 0.0021         | 0.0008         | 0.0011         |

Table 4: Simulated biases and root-mean-squared errors (RMSEs) of the LSSC estimator with $d_1 = 2, d_2 = 2, d_3 = 2$ in Setting 1 with different time periods $T$ and same sample sizes $N_1 = N_2 = N_3 = N$.

| Time period ($T$) | $\beta_1$ Bias | $\beta_1$ RMSE | $\beta_2$ Bias | $\beta_2$ RMSE |
|-------------------|----------------|----------------|----------------|----------------|
| 5                 | 0.0019         | 0.0049         | 0.0018         | 0.0045         |
| 10                | 0.0004         | 0.0023         | 0.0008         | 0.0020         |
| 30                | 0.0004         | 0.0010         | 0.0004         | 0.0012         |
| 50                | 0.0003         | 0.0008         | 0.0002         | 0.0008         |
| 100               | 0.0001         | 0.0006         | 0.0003         | 0.0005         |

(a) We assumed the regressors $x_{it} \sim \text{Uniform}(-2, 2)$, and chose the other settings to be the same as those used in Setting 1. Thus, the regressors were not correlated with the factors and the factor loadings. The simulated biases and RMSEs of the estimation method outlined in Ando & Bai (2016) (which are labelled Ando–Bai) and our proposed LSSC for time periods $T = 10$ and 100 are summarized in Tables 5 and 6, respectively. From Table 5, we can see that our proposed LSSC method yields smaller biases and RMSEs compared to the Ando–Bai method when $T = 10$. In Table 6, the overall performance of the Ando–Bai method when $T = 100$ is satisfactory, but that of our LSSC method is clearly superior.

(b) Next we assumed that the regressors and the factor loadings were correlated as follows:

- the covariate $X$ is a $T \times N \times p$ array with
  $$X_{.,1} = \rho \tau \Lambda' + \eta_1, \quad X_{.,2} = \rho \tau \Lambda' + \eta_2,$$
  where $X_{.,1}$ and $X_{.,2}$ are $T \times N$ matrices;
- $\eta_1 \sim \text{Uniform}_{T \times N}(-2, 2)$ and $\eta_2 \sim \text{Uniform}_{T \times N}(-2, 2)$;
- $\tau$ is a $T \times r$ matrix consisting entirely of 1s;
- $\rho$ is a constant representing the correlation between the covariate $X$ and the factor loadings.
Table 5: Simulated biases and root-mean-squared errors (RMSEs) of estimates based on LSSC and Ando–Bai methods for Setting 2(a). The true parameters are $\beta = (1, 2)$ and $N_1 = N_2 = N_3 = 100, T = 10, r = 3$. The results are based on 100 simulations for each setting.

| Dimension of subspaces | Ando–Bai | LSSC |
|------------------------|----------|------|
|                        | Bias     | RMSE | Bias     | RMSE |
| $d_1 = d_2 = d_3 = 2$  | $\beta_1$ | 0.1002 | 0.1019 | 0.0002 | 0.0035 |
|                        | $\beta_2$ | 0.2019 | 0.2027 | 0.0003 | 0.0037 |
| $d_1 = d_2 = d_3 = 1$  | $\beta_1$ | 0.1022 | 0.1036 | 0.0001 | 0.0017 |
|                        | $\beta_2$ | 0.1989 | 0.1997 | 0.0002 | 0.0020 |
| $d_1 = d_2, d_3 = 1$   | $\beta_1$ | 0.1006 | 0.1019 | 0.0063 | 0.0457 |
|                        | $\beta_2$ | 0.1990 | 0.1999 | 0.0113 | 0.0491 |
| $d_1 = 2, d_2 = d_3 = 1$ | $\beta_1$ | 0.0990 | 0.1002 | 0.0040 | 0.0343 |
|                        | $\beta_2$ | 0.2008 | 0.2015 | 0.0019 | 0.0367 |

Table 6: Simulated biases and root-mean-squared errors (RMSEs) of estimates based on LSSC and Ando–Bai methods for Setting 2(a). The parameters are $\beta = (1, 2)$ and $N_1 = N_2 = N_3 = 100, T = 100, r = 3$. The results are based on 100 simulations for each setting.

| Dimension of subspaces | Ando–Bai | LSSC |
|------------------------|----------|------|
|                        | Bias     | RMSE | Bias     | RMSE |
| $d_1 = d_2 = d_3 = 2$  | $\beta_1$ | 0.0100 | 0.0102 | 0.0000 | 0.0008 |
|                        | $\beta_2$ | 0.0198 | 0.0199 | 0.0001 | 0.0010 |
| $d_1 = d_2 = d_3 = 1$  | $\beta_1$ | 0.0098 | 0.0100 | 0.0000 | 0.0005 |
|                        | $\beta_2$ | 0.0198 | 0.0199 | 0.0000 | 0.0005 |
| $d_1 = d_2, d_3 = 1$   | $\beta_1$ | 0.0097 | 0.0099 | 0.0003 | 0.0020 |
|                        | $\beta_2$ | 0.0197 | 0.0198 | 0.0001 | 0.0013 |
| $d_1 = 2, d_2 = d_3 = 1$ | $\beta_1$ | 0.0099 | 0.0101 | 0.0003 | 0.0048 |
|                        | $\beta_2$ | 0.0199 | 0.0200 | 0.0000 | 0.0025 |

The remaining settings were the same as those adopted in Setting 1. Figure 1 displays the simulated biases and RMSEs of the estimators for $\beta_1$ and $\beta_2$ obtained from the LSSC and Ando–Bai method of estimation as $\rho$ is varied from 0 to 1. From Figure 1, we can see that the LSSC method of estimation is more stable and yields smaller biases and RMSEs compared to its Ando–Bai competitor.

6. MODEL SELECTION AND POSSIBLE EXTENSIONS

In the previous sections, we assumed that the number of subspaces and their dimensions are known. We now extend our procedure to a more general setting in which the dimensions of the
Figure 1: Simulated biases and RMSEs of the estimators of $\beta_1$ and $\beta_2$ obtained from the proposed LSSC method and the Ando–Bai method with $\rho$ varying from 0 to 1 for Setting 2(b). The true values of the parameters are $\beta = (1, 2)$ and $N_1 = N_2 = N_3 = 100, T = 100, r = 3$. The results are based on 100 simulations for setting with $d_1 = d_2 = d_3 = 2$.

factors is unknown and discuss those situations where the number of subspaces for the factors and the dimension of those subspaces are unknown.

6.1. Determining the Number of Subspaces for Factors

One of the critical aspects of cluster analysis is the problem of determining the number of subspaces empirically based on the observed data. However, for experimental data there is no “true” number of subspaces, but rather only a choice of a suitable value of $k$ which can provide stable and replicable results with a satisfactory fit to the observed data. In fact, the problem of estimating the number of subspaces is unquestionably a challenging model selection problem. Here, we do not intend to provide a detailed review of all the existing methods for establishing the number of subspaces for the factors. Instead, we aim to provide a feasible solution based on the work of Liu et al. (2013).

Liu et al. (2013) proposed a novel objective function, which they called a low-rank representation (LRR), which seeks the lowest rank representation among all the candidates that can represent the samples as linear combinations of the bases in a given dictionary. The computational solution for LRR involves solving a nuclear norm regularized problem (Fazel, 2002), which is a convex optimization problem that can be solved in polynomial time. As Liu et al. (2013) show, the estimate of the number of subspaces can be expressed as

$$\hat{k} = N - \text{int} \left[ \sum_{i=1}^{N} f_r(\sigma_i) \right],$$

(15)
where $\tau$ is a cut-off threshold, $\sigma_j$ denotes a singular value of the normalized Laplacian matrix of the affinity matrix of data, $\text{int}[a]$ is the nearest integer of a real number $a$, and $f_\tau$ is a summation function which counts different values. In particular, $f_\tau(\sigma) = 1$ if $\sigma \geq \tau$, and $f_\tau(\sigma) = \log_2(1 + \frac{\sigma^2}{\sigma^2})$ if $\sigma < \tau$, where $0 < \tau < 1$ is a parameter. Thus, we can use the following steps to obtain the number of subspaces:

**Step 1.** Given $\beta$, update $F$ and $\Lambda$ by ignoring the subspace structures;

**Step 2.** Given $F$ and $\Lambda$, update $\beta$;

**Step 3.** Repeat Steps 1 and 2 until convergence occurs. Let $y_i - x_i' \beta = F' \lambda_i + \epsilon_i$, $i = 1, \ldots, N$.

We can now compute the affinity matrix $W$ by using Algorithm 2 in Liu et al. (2013);

**Step 4.** Compute the Laplacian matrix $L = I - D^{-\frac{1}{2}} WD^{-\frac{1}{2}}$, where

$$D = \text{diag}\left(\sum_j [W]_{1j}, \ldots, \sum_j [W]_{nj}\right);$$

**Step 5.** Obtain the number of subspaces, $\hat{k}$, using the expression found in Equation (15).

To verify the performance of this algorithm in estimating the number of subspaces, we assumed that the numbers of subspaces in Settings 1 and 2 are unknown and employed the above algorithm to derive the number of subspaces based on each simulated dataset. In these simulations, we found that the algorithm outlined above yielded $\hat{k} = 3$ correctly for all 100 simulated datasets. Obviously, evaluating the performance of this algorithm for determining the number of subspaces under different settings (e.g., different sample sizes, different number of subspaces, etc.) would clearly be of interest.

6.2. Determining the Dimensions of Factors and the Dimensions of Subspaces

The problem of determining the unknown dimensions of factors (dimensions of ambient spaces) is obviously an interesting research topic. The dimensions of factors can be specified based on the particular practical problem using professional or expert knowledge. When such informed insight is not available, Bai & Ng (2019) have developed a regularization criterion to determine the number of factors; this criterion is more stable when the nominal number of factors is inflated by the presence of weak factors or large measurement noise. To choose the dimension of factors $r \in [0, r_{\text{max}}]$, their expression for this criterion is

$$\bar{r} = \min_{r=0, \ldots, r_{\text{max}}} \log \left[1 - \sum_{j=1}^r (D_{jj} - \gamma)^2\right] + kg(N, T),$$

where $g_{N, T} = \frac{N+T}{NT} \log \left(\frac{NT}{N+T}\right)$, $D_{jj}$ is the $j$th singular value of the scaled observable data, and $\gamma$ is a constant threshold. Through Monte Carlo simulation, we found that this criterion exhibits performance that is particularly satisfactory when the factors and factor loadings have the subspace structure.

When the dimensions of the subspaces are unknown, determining the dimension for each subspace is still an open and challenging problem. In this section, we suggest obtaining the solution of the optimal model selection as

$$Z^* = \arg \min_{\mathcal{A} : Z \in \mathcal{A}} \text{SSR}(\hat{Z}) + \hat{\sigma}^2 \sum_{j=1}^k d_j^2 \frac{T + N_j}{NT} \log(TN_j), \quad \text{SSR}(\hat{Z}) < \tau, \quad (16)$$
where $\text{SSR}(\hat{Z})$ represents the mean-squared error for the subspaces set $\hat{Z}$ (i.e., a measure of the data fidelity), $\tau$ is the error tolerance, $d_j$ is the dimension of the $j$th subspace, $N_j$ denotes the number of individuals in the $j$th subspace, $k$ is the number of subspaces, $\hat{\sigma}^2$ is the estimated variance, and $\sum_{i=1}^{k} d_j^2 \frac{T+N_j}{NT} \log(TN_j)$ is a penalty term that measures the model complexity under the subspaces set $\hat{Z}$. This proposed criterion can be viewed as a trade-off between how well the model fits the data and the model complexity. It can be shown that the penalty function $d_j^2 \frac{T+N_j}{NT} \log(TN_j) \to 0$ and $\min\{N, T\} d_j^2 \frac{T+N_j}{NT} \log(TN_j) \to \infty$ as $T, N \to \infty$ and $T/N$ converges to a constant.

**Theorem 5.** Suppose that Assumptions A–F listed in the accompanying Supplementary Material hold and $T/N \to \rho > 0$. Then the dimensions $\{\hat{d}_1, \ldots, \hat{d}_k\}$ of the subspaces obtained using Equation (16) converge in probability to $\{d_0^1, \ldots, d_0^k\}$, the true dimensions of the subspaces. For a proof of Theorem 5, see Section S3 of the accompanying Supplementary Material.

To investigate our proposed method for model selection, we simulated panel data from the two models with $r = 3$ and 4, and then obtained a solution to the optimal model selection problems using Equation (16) with different numbers of units in each subspace and different time periods $T$. We assumed that there was no covariate, i.e., $\beta = 0$, and used the error tolerance $\tau = 10(d_1^3 + \cdots + d_{k-1}^3)/\min\{N, T\}$. The simulated percentages for identifying the correct dimension of the subspaces (based on 1000 simulations for each setting) are summarized in Tables 7 and 8 for $r = 3$ and 4 with three and four subspaces, respectively. From these simulation results, we conclude that our proposed method of model selection exhibits reasonably

### Table 7: Simulated percentages of identifying the correct dimension of subspaces using Equation (16) with different $(N, T)$ and $N_1 = N_2 = N_3 = N/3$, and the number of factors $r = 3$. 

| Dimension of subspaces | $(N, T)$        |               |               |               |
|------------------------|----------------|---------------|---------------|---------------|
|                        | (600, 300)     | (600, 600)    | (1500, 300)   | (1500, 600)   |
| $d_1 = d_2 = d_3 = 1$  | 97.10%         | 100.00%       | 96.70%        | 100.00%       |
| $d_1 = d_2 = d_3 = 2$  | 100.00%        | 100.00%       | 100.00%       | 100.00%       |
| $d_1 = d_2 = 2, d_3 = 1$ | 79.20%       | 80.40%        | 87.00%        | 93.70%        |
| $d_1 = d_2 = 1, d_3 = 2$ | 90.20%       | 80.10%        | 96.70%        | 100.00%       |

### Table 8: Simulated percentages of identifying the correct dimension of subspaces using Equation (16) with different $(N, T)$ and $N_1 = N_2 = N_3 = N_4 = N/4$, and the number of factors $r = 4$. 

| Dimension of subspaces | $(N, T)$        |               |               |               |
|------------------------|----------------|---------------|---------------|---------------|
|                        | (800, 300)     | (800, 800)    | (2000, 300)   | (2000, 800)   |
| $d_1 = d_2 = d_3 = d_4 = 1$ | 95.00%       | 100.00%       | 92.50%        | 100.00%       |
| $d_1 = d_2 = d_3 = d_4 = 2$ | 99.00%       | 99.90%        | 99.40%        | 100.00%       |
| $d_1 = d_2 = d_3 = d_4 = 3$ | 100.00%      | 100.00%       | 100.00%       | 100.00%       |
| $d_1 = d_2 = d_3 = 2, d_4 = 1$ | 85.60%       | 84.10%        | 85.40%        | 84.40%        |
TABLE 9: BAI, GFE and LSSC methods used to obtain the parameter estimates as ($\hat{\theta}_1, \hat{\theta}_2$) and corresponding fitting errors $\hat{SSR}$, where the number of factors $r = 5$ and the number of groups $k = 3$.

| Methods | ($\hat{\theta}_1, \hat{\theta}_2$) | $\hat{SSR}$ |
|---------|-----------------|-------------|
| BAI     | (0.6023, 0.3729) | 0.0024      |
| GFE     | (0.0869, 0.1723) | 0.1823      |
| LSSC    | (0.8330, 0.3540) | 3.5714e−04 |

satisfactory performance in the case of a hyperplane, i.e., when the subspaces have the same dimensions. Compared with the case of the hyperplane, when the dimensions of the subspaces are not all the same, the estimated percentages for identifying the correct dimension can decrease to roughly 80%.

7. A REAL DATA APPLICATION

In this section, we illustrate our proposed methods using the observed data provided by Bonhomme & Manresa (2015) to study the linkage between income growth and democracy in different countries. Following Bonhomme & Manresa (2015), we use the linear dynamic model to identify group membership and the linkage between income growth and democracy in different countries, i.e.,

$$democracy_{it} = \theta_1 democracy_{i(t-1)} + \theta_2 \log GDP_{pc_{it(t-1)}} + \lambda_{g_{it},f_{g_{it},t}} + \nu_{it},$$

where $democracy_{it}$ represents the democracy index (as measured by the Freedom House indicator with values between 0 (the lowest) and 1 (the highest)) for the $i$th country at time $t$, $GDP_{pc_{it}}$ is the gross domestic product (GDP) per capita of the $i$th country at time period $t$, and $\lambda_{g_{it},f_{g_{it},t}}$ are the unobservable grouped factor loadings and factors, respectively. Here, the dependent variable is the country’s democracy index, and the explanatory variables are the first-order lagged democracy index and the income of a country as measured by the logarithm of GDP per capita.

The dataset contains a balanced panel of 90 countries and seven periods collected at a 5-year interval over the calendar period 1970–2000. First, using the information criterion suggested in Bai & Ng (2019) to estimate the number of factors, we obtain the dimension of the factor space as $r = 5$. Then, using Equation (15), we estimate the number of subspaces to be $\hat{k} = 3$. These results are consistent with those reported in Su, Shi & Phillips (2016). Next we use the criterion specified in Equation (16) to select the optimal model; our observed results suggest that the optimal model has the dimensions $d_1 = d_2 = d_3 = 4$. Finally, we use the BAI, GFE, and LSSC methods of estimation to obtain the parameter estimates as ($\hat{\theta}_1, \hat{\theta}_2$) and the corresponding fitting errors (defined as $SSR = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (democracy_{it} - \theta_1 democracy_{i(t-1)} - \theta_2 \log GDP_{pc_{it(t-1)}} - \lambda_{g_{it},f_{g_{it},t}})^2$). The estimated results are summarized in Table 9. From each fitted model, we conclude that the effect of income on democracy is positive. However, we also note that our LSSC method of estimation yields the smallest value of the fitting error.

In order to visualize the group membership results generated by our proposed method, we translate the fitted groupings onto a world map; see Figure 2 in which the countries in the same group are represented using the same colour. The detailed lists of grouped countries are as follows:

- Group 1 (45 countries): Argentina, Australia, Bangladesh, Burkina Faso, Burundi, Cameroon, Canada, Chile, Congo, Costa Rica, Denmark, Dominican Rep., Ecuador, El Salvador,
France, Gambia, Ghana, Guatemala, Honduras, Iran, Israel, Italy, Jamaica, Jordan, Kenya, Luxembourg, Malawi, Malaysia, Morocco, Nepal, New Zealand, Nicaragua, Nigeria, Norway, Paraguay, Peru, Philippines, Romania, Spain, Sweden, Togo, Trinidad and Tobago, United States, Venezuela, Zambia.

- Group 2 (24 countries): Algeria, Belgium, Bolivia, Brazil, China, Colombia, Egypt, Finland, Greece, Indonesia, Ireland, Japan, Korea, Lesotho, Mali, Netherlands, Niger, Portugal, Rwanda, South Africa, Sri Lanka, Tunisia, United Kingdom, Uruguay.
- Group 3 (21 countries): Austria, Barbados, Benin, Chad, Gabon, Guinea, Hungary, Iceland, India, Madagascar, Mauritius, Mexico, Panama, Senegal, Switzerland, Syria, Tanzania, Thailand, Turkey, Uganda, Zimbabwe.

From these groupings, it can be seen that most of the mature developed economies belong to the first group, which seem to exhibit a similarity in both political structure and economic orientation. This group includes the United States and Canada, most of the countries in continental Europe, the coastal countries of South America, and Australia. Most of the countries in the second group are developing countries in Asia, Africa, and South America that experienced rapid economic development during 1970–2000, including China, Brazil, and South Africa. Japan and South Korea also belong to the second group because they are both countries that experienced rapid economic development during this period of observation and have many similarities. Most of the countries in the third group experienced slower economic growth and development during this same observation period.

8. CONCLUDING REMARKS

In this article, we have considered a panel data model that allows the covariates and the unobservable latent variables to be correlated. We proposed a subspace clustering method for factor loadings of the panel data model that captures the grouped unobserved heterogeneity. The common regression parameters, grouped unobservable factor structure, and group membership can be estimated simultaneously using our proposed method of estimation. Asymptotic results showed that the subspace clustering and the model parameter estimators are consistent. Monte Carlo simulation results demonstrated that our proposed methods outperform the existing alternative methods of estimation in various different settings. Under the model considered in this article, we proposed a consistent model selection criterion to determine a suitable subspace dimension. We also identified some possible directions for future research with respect to the problem of determining the number of subspaces and the factor dimensions when these values
are unknown. These issues are currently under investigation, and we hope to report our progress in a future article.

ACKNOWLEDGEMENTS

We thank the Associate Editor and the referees for their helpful comments and suggestions which improved the presentation and the quality of this article. Jiangtao Duan’s research was supported by the Postdoctoral international training program for young talents in Guangdong Province, and Guangdong Basic and Applied Basic Research Foundation (No. 2019A151510448). Wei Gao’s research was supported by the National Natural Science Foundation of China (No. 11871141). H. K. T. Ng’s research was supported by a grant from the Simons Foundation (#709773 to Tony Ng).

REFERENCES

Ahn, S. C., Lee, Y. H., & Schmidt, P. (2013). Panel data models with multiple time-varying individual effects. *Journal of Econometrics*, 174, 1–14.

Amemiya, T. (1971). The estimation of the variances in a variance-components model. *International Economic Review*, 12, 1–13.

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

Bai, J. S. (2009). Panel data models with interactive fixed effects. *Econometrica*, 77, 1229–1279.

Bai, J. S. & Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70, 191–221.

Bai, J. S. & Ng, S. (2019). Rank regularized estimation of approximate factor models. *Journal of Econometrics*, 212, 78–96.

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

Fazel, M. (2002). *Matrix Rank Minimization with Applications*, Ph.D. thesis. Department of Electrical Engineering, Stanford University, Stanford.

Fischler, M. A. & Bolles, R. C. (1981). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 26, 381–395.

Gobillon, L. & Magnac, T. (2016). Regional policy evaluation: Interactive fixed effects and synthetic controls. *The Review of Economics and Statistics*, 98, 535–551.

Hsiao, C., Ching, H. S., & Wan, S. K. (2012). A panel data approach for program evaluation: Measuring the benefits of political and economic integration of Hong Kong with Mainland China. *Journal of Applied Econometrics*, 27, 705–740.

Kanatani, K. (2012). Motion segmentation by subspace separation: Model selection and reliability evaluation. *International Journal of Image and Graphics*, 2, 179–197.

Kriegel, H. P., Kröger, P., & Zimek, A. (2009). Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 3, 1–58.

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

Liu, G., Lin, Z., Yan, S., Sun, J., Yu, Y., & Ma, Y. (2013). Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35, 171–184.

Nickel, S. (1981). Biases in dynamic models with fixed effects. *Econometrica*, 49, 1417–1426.

Pesaran, H. M. (2006). Estimation and inference in large heterogeneous panels with a multi-factor error structure. *Econometrica*, 74, 967–1012.

Pollard, D. (1981). Strong consistency of k-mean clustering. *The Annals of Statistics*, 9, 135–140.

Shi, W. & Lee, L. F. (2017). Spatial dynamic panel data models with interactive fixed effects. *Journal of Econometrics*, 197, 323–347.

Stock, J. H. & Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97, 1167–1179.
Su, L. & Ju, G. S. (2018). Identifying latent grouped patterns in panel data models with interactive fixed effects. *Journal of Econometrics*, 206, 554–573.
Su, L., Shi, Z., & Phillips, P. C. B. (2016). Identifying latent structures in panel data. *Econometrica*, 84, 2215–2264.
Terada, Y. (2014). Strong consistency of reduced $k$-means clustering. *Scandinavian Journal of Statistics*, 41, 913–931.
Vidal, M. Y. R. & Sastry, S. (2005). Generalized principal component analysis (GPCA). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27, 1945–1959.
Vidal, M. Y. R. & Sastry, S. (2016). *Generalized Principal Component Analysis*. Springer-Verlag, New York.
Wallace, T. D. & Hussain, A. (1969). The use of error components models in combining cross section with time series data. *Econometrica*, 37, 55–72.
Wang, L. W., Wang, X., & Feng, J. F. (2006). Subspace distance analysis with application to adaptive Bayesian algorithm for face recognition. *Pattern Recognition*, 39, 456–464.
Yang, A., Rao, S., Wagner, A., Ma, Y., & Fossum, R. M. (2005). Hilbert functions and applications to the estimation of subspace arrangements. *ICCV’05: Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV’05)*, IEEE, New York, 158–165.

Received 27 March 2020
Accepted 5 March 2021