JOURNAL OF APPLIED ECONOMETRICS
J. Appl. Econ. **31**: 163–191 (2016)
Published online 21 May 2015 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/jae.2467

PANEL DATA MODELS WITH GROUPED FACTOR STRUCTURE UNDER UNKNOWN GROUP MEMBERSHIP

TOMOHIRO ANDO[a,b] AND JUSHAN BAI[c,d]*
[a] *Graduate School of Business, Keio University*
[b] *Melbourne Business School, Melbourne University*
[c] *Department of Economics, Columbia University, New York, NY, USA*
[d] *School of Finance, Nankai University, Tianjin, China*

SUMMARY

This paper studies panel data models with unobserved group factor structures. The group membership of each unit and the number of groups are left unspecified. We estimate the model by minimizing the sum of least squared errors with a shrinkage penalty. The number of explanatory variables can be large. The regressions coefficients can be homogeneous or group specific. The consistency and asymptotic normality of the estimator are established. We also introduce new C_p-type criteria for selecting the number of groups, the numbers of group-specific common factors and relevant regressors. Monte Carlo results show that the proposed method works well. We apply the method to the study of US mutual fund returns and to the study of individual stock returns of the China mainland stock markets. Copyright © 2015 John Wiley & Sons, Ltd.

Received 1 November 2013; Revised 12 March 2015

Supporting information may be found in the online version of this article.

1. INTRODUCTION

There is an increasing literature on panel data models with multiple unobserved common factors (e.g. Pesaran, 2006; Bai, 2009). This paper considers a grouped panel data model in which N individuals are divided into S groups, with each group having its own panel regression coefficients and its own factor structure. A key feature of the model is that group memberships are unknown and are to be estimated from the observed data. The explanatory variables in the model are allowed to be correlated with the factors, or factor loadings, or both. In addition, the number of explanatory variables can be large, and the relevant regressors are selected via a regularization (penalized) approach.

Previous studies exist for grouped factor structures, where the group memberships are known (e.g. Moench *et al.*, 2012; Diebold *et al.*, 2008; Kose *et al.*, 2008; Wang, 2010; Moench and Ng, 2011). These studies do not consider the presence of explanatory variables. There are also studies that consider the challenging problem of unknown group memberships (e.g. Bonhomme and Manresa, 2012; Lin and Ng, 2012; Sun, 2005). These papers do not consider factor error structures, although Bonhomme and Manresa's setup can be considered as a special factor model with known loading values of 0 or 1.

We consider the joint estimation of the optimal grouping of the N cross-sectional units, the regression coefficients and grouped factor structure. We derive the asymptotic properties of the proposed estimator and show that the proposed estimator is consistent as N and T go to infinity simultaneously. We provide a novel argument for consistency in the presence of unknown group memberships.

An important issue in practice is the selection of a proper model from among many candidates or, equivalently, the determination of the number of group-specific factors, the determination of the magnitude of the regularization parameter for implementing the regularization approach

* Correspondence to: Jushan Bai, Department of Economics, Columbia University, New York, NY, USA. E-mail: jushan.bai@columbia.edu

(to be introduced), and the determination of the number of groups. For this purpose we develop new C_p-type criteria for selecting a proper model from a predictive perspective. Specifically, the panel data model is evaluated from a predictive point of view, and we propose an estimator of the expected mean squared error (MSE). The criterion is developed by correcting the asymptotic bias in the MSE as an estimate of the expected MSE. With additional penalty, the criteria allow consistent estimation of the number of factors in a group, and the number of groups.

We consider both homogeneous regression coefficients, in which the slope parameters are common across groups, and group-heterogeneous coefficients, in which the slope parameters vary across the groups. The model with heterogeneous regression coefficients is applied to the analysis of the US mutual fund styles. Financial institutions manage clients' assets according to the investment style that defines the nature of the fund. We aim at grouping mutual funds and identifying their styles by analyzing the time series of past returns of individual mutual funds. We also apply our method to the analysis of the two Chinese mainland stock markets: the Shanghai and Shenzhen stock exchanges. Given the A-shares and B-shares markets in China, it is natural to conjecture that there exist at least two groups. We address the following questions. How many groups exist in the stock markets in mainland China? How many group-specific factors exist in the stock markets in mainland China? What type of observable risk factors explains the stocks in each group? Furthermore, how can the unobservable factors be understood in terms of observable variables in the economy? A number of interesting findings are reported.

The remainder of this paper is organized as follows. Section 2 states the assumptions and Section 3 describes the estimation procedure. Section 4 investigates the consistency of the proposed estimator. Its asymptotic behaviors are also investigated. Section 5 develops the model selection criterion from a predictive point of view. Section 6 reports the results of a Monte Carlo analysis. The simulations confirm that the proposed criterion performs well. Section 7 extends the results to panel data models with heterogeneous regression coefficients. Section 8 applies the procedure to the analysis of the US mutual fund styles and the Chinese mainland stock markets. Concluding remarks are provided in Section 9. Proofs are provided as supporting information in the online supplement.

2. MODEL

Let $t = 1, \dots, T$ be the time index and $i = 1, \dots, N$ be the cross-section index. Let S be the number of groups (which is unknown and fixed), and let $G = \{g_1, \dots, g_N\}$ denote the group membership such that $g_i \in \{1, \dots, S\}$. Let N_j be the number of cross-sectional units within group j $(j = 1, \dots, S)$ so that $N = \sum_{j=1}^{S} N_j$.

In this section, we assume that the response variable of the ith unit, observed at time t, y_{it}, is expressed as

$$y_{it} = \boldsymbol{x}'_{it}\boldsymbol{\beta} + \boldsymbol{f}'_{g_i,t}\boldsymbol{\lambda}_{g_i,i} + \varepsilon_{i,t}, \quad i = 1, \dots, N,\ t = 1, \dots, T \tag{1}$$

where $\boldsymbol{x}_{it}$ is a $p \times 1$ vector of observable vectors, and $\boldsymbol{f}_{g_i,t}$ is an $r_j \times 1$ vector of unobservable group-specific factors that affect the units only in group g_i. The $p \times 1$ vector $\boldsymbol{\beta}$ are the unknown regression coefficients, $\boldsymbol{\lambda}_{g_i,i}$ are the factor loadings and ε_{it} is the unit-specific error. Here $\boldsymbol{\beta}$ is common for all i. Model (1) is extended to group-dependent coefficients in Section 7.

In vector form, model (1) can be expressed as $\boldsymbol{y}_i = X_i\boldsymbol{\beta} + F_{g_i}\boldsymbol{\lambda}_{g_i,i} + \boldsymbol{\varepsilon}_i, \quad i = 1, \dots, N$ where (for $g_i = j$, $F_{g_i} = F_j$)

$$\boldsymbol{y}_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iT} \end{pmatrix}, \ X_i = \begin{pmatrix} \boldsymbol{x}'_{i1} \\ \boldsymbol{x}'_{i2} \\ \vdots \\ \boldsymbol{x}'_{iT} \end{pmatrix}, \ F_j = \begin{pmatrix} \boldsymbol{f}'_{j,1} \\ \boldsymbol{f}'_{j,2} \\ \vdots \\ \boldsymbol{f}'_{j,T} \end{pmatrix}, \ \boldsymbol{\varepsilon}_i = \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \vdots \\ \varepsilon_{iT} \end{pmatrix}$$

DOI: 10.1002/jae

Depending on applications, the unobserved factor components may be specified as an exact dynamic factor model (Geweke, 1977; Sargent and Sims, 1977), a static approximate factor model (Chamberlain and Rothschild, 1983) or a special model of the generalized dynamic factor model (Forni *et al.*, 2000; see also Forni and Lippi, 2001; Amengual and Watson, 2007; Hallin and Liska, 2007). Details of $\boldsymbol{f}'_{g_i,t}\boldsymbol{\lambda}_{g_i,i}$ will be specified in the next section.

2.1. Assumptions

We first state the assumptions and then provide comments concerning these assumptions. Throughout, the norm of matrix A is defined as $\|A\| = [\mathrm{tr}(A'A)]^{1/2}$.

Assumption A: Group-specific pervasive factors. The group-specific pervasive factors satisfy $E\|\boldsymbol{f}_{j,t}\|^4 < \infty$ $j = 1, \ldots, S$. Furthermore, $T^{-1}\sum_{t=1}^{T}\boldsymbol{f}_{j,t}\boldsymbol{f}_{j,t}{}' \to \Sigma_{F_j}$ as $T \to \infty$, where Σ_{F_j} is an $r_j \times r_j$ positive definite matrix. Although correlations between $\boldsymbol{f}_{j,t}$ and $\boldsymbol{f}_{k,t}$ $(j \neq k)$ are allowed, they are not correlated perfectly.

Assumption B: factor loadings.

B1: The factor loading matrix for the group-specific pervasive factors $\Lambda_j = [\boldsymbol{\lambda}_{j,1}, \ldots, \boldsymbol{\lambda}_{j,N_j}]'$ satisfies $E\|\boldsymbol{\lambda}^4_{j,i}\| < \infty$ and $\|N_j^{-1}\Lambda'_j\Lambda_j - \Sigma_{\Lambda_j}\| \to \mathbf{0}$ as $N_j \to \infty$, where Σ_{Λ_j} is an $r_j \times r_j$ positive definite matrix, $j = 1, \ldots, S$. We also assume that $\|\boldsymbol{\lambda}_{j,i}\| > 0$.

B2: For each i and j, $\boldsymbol{f}'_{j,t}\boldsymbol{\lambda}_{j,i}$ is strongly mixing processes with mixing coefficients that satisfy $r(t) \leq \exp(-a_1t^{b_1})$ and with tail probability $P(|\boldsymbol{f}'_{j,t}\boldsymbol{\lambda}_{j,i}| > z) \leq \exp\{1 - (z/b_2)^{a_2}\}$, where a_1, a_2, b_1 and b_2 are positive constants.

Assumption C: Error terms. The error terms $\boldsymbol{\varepsilon}_t$ of the model in equation (1) have zero mean, but may have cross-sectional dependence and heteroskedasticity. Furthermore, there exists a positive constant $C < \infty$ such that for all N and T.

C1: $E[\varepsilon_{it}] = 0$ for all i and t.

C2: $E[\varepsilon_{it}\varepsilon_{js}] = \tau_{ij,ts}$ with $|\tau_{ij,ts}| \leq |\tau_{ij}|$ for some τ_{ij} for all (t,s), and $N^{-1}\sum_{i,j=1}^{N}|\tau_{ij}| < C$; and $|\tau_{ij,ts}| \leq |\eta_{ts}|$ for some η_{ts} for all (i,j), and $T^{-1}\sum_{t,s=1}^{T}|\eta_{ts}| < C$. In addition, $(TN)^{-1}\sum_{i,j,t,s=1}|\tau_{ij,ts}| < C$.

C3: For every (s,t), $E[|N^{-1/2}\sum_{i=1}^{N}(\varepsilon_{is}\varepsilon_{it} - E[\varepsilon_{is}\varepsilon_{it}])|^4] < C$.

C4: $T^{-2}N^{-1}\sum_{t,s,u,v}\sum_{i,j}|\mathrm{cov}(\varepsilon_{is}\varepsilon_{it}, \varepsilon_{js}\varepsilon_{jt})| < C$ and $T^{-1}N^{-2}\sum_{t,s}\sum_{i,j,k,l}|\mathrm{cov}(\varepsilon_{it}\varepsilon_{jt}, \varepsilon_{ks}\varepsilon_{lt})$ $| < C$.

C5: For all i, ε_{it} is strongly mixing processes with mixing coefficients that satisfy $r(t) \leq \exp(-a_1t^{b_1})$ and with tail probability $P(|\varepsilon_{it}| > z) \leq \exp\{1 - (z/b_2)^{a_2}\}$, where a_1, a_2, b_1 and b_2 are positive constants.

C6: ε_{it} is independent of $\boldsymbol{x}_{js}$, $\boldsymbol{\lambda}_{j,i}$ and $\boldsymbol{f}_{j,s}$ for all i, j, t, s.

Assumption D: Observable predictors.

D1: Define $D_j = \frac{1}{NT}\sum_{i;g_i=j}X'_iM_{F_j}X_i$, $E_j = \mathrm{diag}\{E_{j1}, \ldots, E_{jS}\}$, $L_j = \left(L'_{j1}, \ldots, L'_{jS}\right)'$, where E_{jk}, and L_{jk} are $E_{jk} = \frac{1}{N}\sum_{i;g_i=j,g_i^0=k}(\boldsymbol{\lambda}^0_{k,i}\boldsymbol{\lambda}^0_{k,i}{}') \otimes I_T$, $L_{jk} = \sum_{i;g_i=j,g_i^0=k}\frac{1}{NT}\boldsymbol{\lambda}^0_{k,i} \otimes M_{F_j}X_i$ with g_i^0 denoting the true membership and $\boldsymbol{\lambda}^0_{k,i}$ the true factor loadings. Let $A = \{F_j : F'_jF_j/T = I, j = 1, \ldots, S\}$. The smallest eigenvalue of the matrix

$$\sum_{j=1}^{S}(D_j - L'_jE_j^{-}L_j)$$

DOI: 10.1002/jae

is greater than a positive constant c for all $(F_1, \dots, F_S) \in A$ and for all groupings with a positive fraction of membership for each group (Assumption E below), where E_j^- is a generalized inverse of E_j. Note that if some components of E_j are zero, then the corresponding components of L_j are also zero so that $L_j' E_j^{-1} L_j$ is well defined. Further comments on this assumption are given below.

D2: The vector of predictor $\boldsymbol{x}_{it}$ satisfies $\max_{1 \le i \le N} T^{-1} \|X_i\|^2 = O_p(N^\alpha)$ with $\alpha < 1/8$. We also assume $N/T^2 \to 0$.

Assumption E: Number of units in each group. All units are divided into a finite number of groups S, each of them containing N_j units such that $0 < \underline{a} < N_j/N < \bar{a} < 1$, which implies that the number of units in the jth group increases as the total number of units N grows.

Some comments on the assumptions are in order. Assumptions A and B imply the existence of r_j group-specific pervasive factors, $j = 1, \dots, S$. Assumption C imposes weak serial and cross-sectional correlations on ε_{it}. Heteroskedasticity is allowed. These assumptions are made in Bai (2009) except for C5. Assumption C5 assumes that the error term is strongly mixing with a faster than polynomial decay rate and restricts the tail property. This condition is used to bound misclassification probabilities, and is used in Bonhomme and Manresa (2012). Simulations show that the method performs well without this condition (e.g. Student-t distribution with five degrees of freedom).

Assumption D1 is similar to a condition used in Bai (2009), where only a single group exists. The assumption is used for proof of consistency. Assumption D1 is analogous to the full rank condition in standard linear regression models, but it is stronger than that due to the unobservableness of factors and the membership groupings. An alternative and weaker assumption is that $\sum_{j=1}^S (D_j - L_j' E^- L_j)$ is positive definite when evaluated at the true factors and true groupings. This will correspond to the usual full rank condition. This alternative assumption is discussed in Bai (2009) and is also used by Ando and Bai (2014), in which group memberships are known. Under this assumption, one first proves the consistency of the estimated factors and membership groupings, and then proves the consistency of the estimated beta coefficient. This argument of consistency is more involved. The current assumption allows a simpler proof of consistency of $\hat{\beta}$. Assumption D2 is a weaker condition than assuming that x_{it} has exponentially decaying tails. The regressors can be correlated with factors, factor loadings or both. This correlation is controlled for by treating both factors and factor loadings as parameters. As in usual panel data analysis, the number of cross-sectional units N can be much greater than the number of time periods T. The true number of groups, S, is assumed fixed. The case of increasing S is considered in the online supplementary document. Bester and Hansen (2012) allowed the true number of groups in both dimensions of the panel to tend to infinity. In their setup, there are individual effects but no factor structure, and the group membership is assumed known.

3. ESTIMATION

3.1. Estimation Procedure

Under a given number of groups S, number of factors $r_1, \dots, r_S$, and size of the penalty κ in $p_{\kappa,\gamma}(|\boldsymbol{\beta}|)$, the estimator $\{\hat{\boldsymbol{\beta}}, \hat{G}, \hat{F}_1, \dots, \hat{F}_S, \hat{\Lambda}_1, \dots, \hat{\Lambda}_S\}$ is defined as the minimizer of

$$L_{NT}(\boldsymbol{\beta}, G, F_1, \dots, F_S, \Lambda_1, \dots, \Lambda_S) = \sum_{j=1}^{S} \sum_{i; g_i = j} \|\boldsymbol{y}_i - X_i \boldsymbol{\beta} - F_{g_i} \boldsymbol{\lambda}_{g_i, i}\|^2 + NT \cdot p_{\kappa,\gamma}(|\boldsymbol{\beta}|)$$

subject to the constraints $F_j' F_j / T = I_{r_j}$ $(j = 1, \dots, S)$, $\Lambda_j' \Lambda_j$ $(j = 1, \dots, S)$ being diagonal. Here, $\Lambda_j = (\boldsymbol{\lambda}_{j,1}, \dots, \boldsymbol{\lambda}_{j,N_j})$ is the $r_j \times N_j$ factor loading matrix $(j = 1, \dots, S)$ for the group-specific factors. These restrictions are needed to avoid the model identification problem and are commonly used in the literature (Connor and Korajzcyk, 1986; Stock and Watson, 2002; Bai and Ng, 2002).

DOI: 10.1002/jae

The objective function here is similar to that of Bai (2009) and Moon and Weidner (2009) other than the penalty term. If the objective is to estimate $\boldsymbol{\beta}$, the CCE method of Pesaran (2006) will provide a consistent estimation (also see Kapetanios *et al.*, 2011; Pesaran and Tosetti, 2011).

For the penalty function, $p_{\kappa,\gamma}\left(|\boldsymbol{\beta}|\right)$ is designed to identify the significant components of the regression coefficients. This is important when the number of regressors (p) is large and some regressors may be irrelevant. In this paper we use the SCAD penalty of Fan and Li (2001), which is formally given as $p_{\kappa,\gamma}(|\boldsymbol{\beta}|) = \sum_{j=1}^{p} p_{\kappa,\gamma}(|\beta_j|)$ with

$$p_{\kappa,\gamma}(|\beta_j|) = \begin{cases} \kappa|\beta_j| & (|\beta_j| \le \kappa) \\ \dfrac{\gamma\kappa|\beta_j| - 0.5(\beta_j^2 + \kappa^2)}{\gamma - 1} & (\kappa < |\beta_j| \le \gamma\kappa) \\ \dfrac{\kappa^2(\gamma^2 - 1)}{2(\gamma - 1)} & (\gamma\kappa < |\beta_j|) \end{cases}$$

for $\kappa > 0$ and $\gamma > 2$. This penalty first applies the same rate of penalization as the lasso method and then reduces the rate to zero as it moves further away from zero. Fan and Li (2001) showed that the value $\gamma = 3.7$ minimizes a Bayesian risk criterion for the regression coefficients. We also use $\gamma = 3.7$. For further discussion of the shrinkage methods, we refer to Tibshirani (1996), Fan and Li (2001), Fan and Peng (2004), Zou (2006), Cheng *et al.* (2013), Lu and Su (2013), Ando and Bai (2014) and references therein.

The minimization of the objective function is obtained via iterations. Given the group membership G and the group-specific factor structures $F_j\boldsymbol{\lambda}_{j,i}$, we define the variable $\boldsymbol{y}_i^* = \boldsymbol{y}_i - F_{g_i}\boldsymbol{\lambda}_{g_i,i}$ for $i = 1,\ldots,N$. The objective function can then be viewed as $\sum_{j=1}^{S}\sum_{i;g_i=j} \|\boldsymbol{y}_i^* - X_i\boldsymbol{\beta}\|^2 + NT\cdot p_{\kappa,\gamma}(|\boldsymbol{\beta}|)$. The estimator of $\boldsymbol{\beta}$ is obtained by the SCAD approach.

Given the group membership G and the value of the regression coefficient $\boldsymbol{\beta}$, we define the variable $W_j = (\boldsymbol{w}_{j,1},\ldots,\boldsymbol{w}_{j,N_j})$ with $\boldsymbol{w}_{j,i} = \boldsymbol{y}_i - X_i\boldsymbol{\beta}$ for $g_i = j$. The original model (1) then reduces to $\boldsymbol{w}_{j,i} = F_j\boldsymbol{\lambda}_{j,i} + \boldsymbol{\varepsilon}_i$, which implies that matrix W_j has a pure factor structure. The least squares objective function without the penalty term is $\sum_{j=1}^{S}\mathrm{tr}\{(W_j - F_j\Lambda_j')(W_j - F_j\Lambda_j')'\}$. From the analysis of pure factor models estimated by the method of least squares (i.e. principal components; see Connor and Korajzcyk, 1986; Stock and Watson, 2002), we concentrate out $\Lambda_j = W_j'F_j(F_j'F_j)^{-1} = W_j'F_j/T$, then the objective function becomes

$$\sum_{j=1}^{S}\mathrm{tr}\left\{W_j'W_j\right\} - \sum_{j=1}^{S}\mathrm{tr}\left\{F_j'W_jW_j'F_j\right\}/T.$$

Noting that only N_j units are related to the factor structure F_j of the jth group and that the penalty term is not related to F_j, minimizing the objective function with respect to F_j is equivalent to maximizing $\mathrm{tr}\left\{F_j'W_jW_j'F_j\right\}$. The principal components estimate of F_j subject to the constraint, $\hat{F}_j$, is $\sqrt{T}$ times the eigenvectors corresponding to the r_j largest eigenvalues of the $T \times T$ matrix W_jW_j'. Given $\hat{F}_j$, the factor loading matrix can be obtained as $\hat{\Lambda}_j = \hat{F}_jW_j/T$ (see also Bai and Ng, 2002, pp. 197–198).

It is noted that, for any given values of $\boldsymbol{\beta}$ and $F_j\boldsymbol{\lambda}_{j,i}$ $(j = 1,\ldots,S)$, the optimal assignment for each individual unit is

$$g_i^* = \arg\min_{j\in\{1,\ldots,S\}} \|\boldsymbol{y}_i - X_i\boldsymbol{\beta} - F_j\boldsymbol{\lambda}_{j,i}\|^2$$

This is because g_i^* minimizes $L_{NT}(\boldsymbol{\beta}, G, F_1,\ldots,F_S,\Lambda_1,\ldots,\Lambda_S)$, given values of $\boldsymbol{\beta}$ and $F_{g_i}\boldsymbol{\lambda}_{g_i,i}$. In practical implementation, we need to replace $\boldsymbol{\beta}$, F_j and λ_{ji} by their estimates.

DOI: 10.1002/jae

Estimating $\boldsymbol{\beta}$ and F_j is discussed earlier; the factor loadings $\boldsymbol{\lambda}_{j,i}$ is obtained by $\boldsymbol{\lambda}_{j,i} = (F_j'F_j)^{-1}F_j'(\boldsymbol{y}_i - X_i\boldsymbol{\beta}) = F_j'(\boldsymbol{y}_i - X_i\boldsymbol{\beta})/T$ for $j = 1, \ldots, S$. The final estimator for $\hat{g}_i$ satisfies $\hat{g}_i = \arg\min_{j\in\{1,\ldots,S\}} \|\boldsymbol{y}_i - X_i\hat{\boldsymbol{\beta}} - \hat{F}_j\hat{\boldsymbol{\lambda}}_{j,i}\|^2$. We mention that some prior information can be incorporated by using the Bayesian procedure (not considered in this paper). The estimates of $\boldsymbol{\beta}$, $\{F_j, \Lambda_j;\ j = 1, \ldots, S\}$, and $G \in \{g_1, \ldots, g_N\}$ depend on each other. The estimators are obtained by using the following iterative algorithm.

Estimation algorithm.

Step 1. Fix κ, $\{r_1, \ldots, r_S\}$ and S. Initialize the unknown parameters $\boldsymbol{\beta}^{(0)}$, $\left\{F_j^{(0)}, \Lambda_j^{(0)};\ j = 1, \ldots, S\right\}$, $G^{(0)} \in \{g_1^{(0)}, \ldots, G_N^{(0)}\}$.
Step 2. Given the values of $\boldsymbol{\beta}$ and $\{F_j, \Lambda_j;\ j = 1, \ldots, S\}$, update G.
Step 3. Given the values of $\boldsymbol{\beta}$ and G, update $\{F_j, \Lambda_j\}$ for $j = 1, \ldots, S$.
Step 4. Given the values of G and $\{F_j, \Lambda_j;\ j = 1, \ldots, S\}$, update $\boldsymbol{\beta}$.
Step 5. Repeat Steps 2 and 4 until convergence.

In step 1, starting values for $\boldsymbol{\beta}$, G, and $\{F_j, \Lambda_j;\ j = 1, \ldots, S\}$ are needed. In the next section, we discuss how to prepare initial values for these parameters.

3.2. Initial Parameter Values

For fast initialization of group membership G, we use the well-known K-means algorithm (Forgy, 1965). Given the number of groups S, the algorithm finds a collection of centers of each group such that the sum of the Euclidean distances between each unit and the closest center is minimized. The K-means algorithm divides the dataset $\{\boldsymbol{y}_i;\ i = 1, \ldots, N\}$ into S clusters that correspond to the number of groups. Thus an initial estimate of the group membership $G^{(0)} \in \{g_1^{(0)}, \ldots, g_N^{(0)}\}$ is obtained this way. Given $G^{(0)}$, an initial estimate of $\boldsymbol{\beta}^{(0)}$ is obtained via the SCAD by ignoring the group-specific factor structures $\{F_j, \Lambda_j;\ j = 1, \ldots, S\}$. Finally, given the values of $\boldsymbol{\beta}^{(0)}$ and $G^{(0)}$, we obtain the starting values $\{F_j^{(0)}, \Lambda_j^{(0)}\}$ for $j = 1, \ldots, S$ by the principal components.

It is known that the least squares objective function is not globally convex (Bai, 2009). In other words, an arbitrary starting value will not necessarily provide the global optimal solution. However, under large N and large T, the final estimates are quite robust to starting values. Only under a small N or a small T, there are cases that different initial values may lead to different estimates. In our simulations, the reported results are based on a single starting value as outline above; the results are satisfactory.

4. ASYMPTOTIC PROPERTIES

In Sections 2 and 3 we described the assumptions imposed on the model and proposed an estimation procedure. This section investigates some asymptotic properties of the parameter estimates. All proofs of the theorems, described below, are given in the Appendix. We use $\{F_j^0,\ j = 1, \ldots, S\}$ to denote the true parameter values of the group-specific factors F_j obtained from the true data-generating process. As T increases, the number of elements of F_j $(j = 1, \ldots, S)$ are also increasing. We claim that the estimated factors are consistent in the sense of some averaged norm, which will be specified below. We have the following theorem.

Theorem 1. Consistency. Under Assumptions A–E, $\kappa \to 0$ and $\min\{N, T\} \times \kappa \to \infty$ as $T, N \to \infty$, and the estimator $\hat{\boldsymbol{\beta}}$ is consistent

DOI: 10.1002/jae

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\| = o_p(1)$$

where $\boldsymbol{\beta}^0$ denotes the true parameter value. In addition, $\{\hat{F}_j,\ j = 1, \ldots, S\}$ are consistent in the sense of the following norm

$$T^{-1}\|\hat{F}_j - F_j^0 H_j\|^2 = o_p(1), \quad j = 1, \ldots, S \tag{2}$$

where $H_j^{-1} = V_{j,N_jT}(F_j^0\hat{F}_j/T)^{-1}(\Lambda_j^{0\prime}\Lambda_j^0/N_j)^{-1}$, and V_{j,N_jT} satisfies

$$\left[\frac{1}{N_jT}\sum_{i;\hat{g}_i=j}^{N_j}(\boldsymbol{y}_i - X_i\hat{\boldsymbol{\beta}})(\boldsymbol{y}_i - X_i\hat{\boldsymbol{\beta}})'\right]\hat{F}_j = \hat{F}_jV_{j,N_jT}$$

The estimated individual membership satisfies $\hat{g}_i = \arg\min_{j\in\{1,\ldots,S\}}\|\boldsymbol{y}_i - X_i\hat{\boldsymbol{\beta}} - \hat{F}_j\hat{\boldsymbol{\lambda}}_{j,i}\|^2$, and thus minimizes the sum of squared residuals among the S possible groups. The estimates of $\boldsymbol{\beta}$, $\{F_j, \Lambda_j;\ j = 1, \ldots, S\}$, and $G \in \{g_1, \ldots, g_N\}$ depend on each other, and we therefore denote the estimator of group membership $\hat{g}_i$ as $\hat{g}_i(\hat{\boldsymbol{\beta}}, \hat{F}, \hat{\Lambda})$. Here, $\hat{F} = \{\hat{F}_1, \ldots, \hat{F}_S\}$ and $\hat{\Lambda} = \{\hat{\Lambda}_1, \ldots, \hat{\Lambda}_S\}$. The following theorem shows that the estimated group membership converges to the true group membership as T and N grow.

Theorem 2. Consistency of the estimator of group membership. Suppose that the assumptions in Theorem 1 hold. Then, for all $\tau > 0$ and $T, N \to \infty$, we have

$$P\left(\sup_{i\in\{1,\ldots,N\}}\left|\hat{g}_i(\hat{\boldsymbol{\beta}}, \hat{F}, \hat{\Lambda}) - g_i^0\right| > 0\right) = o(1) + o(N/T^\tau).$$

The result of Theorem 2 shows that if for some $b > 0$, $N/T^b \to 0$ as both N and T tend to infinity simultaneously, the true group membership g_i^0 and the proposed group membership estimator $\hat{g}_i$ are asymptotically equivalent. This holds because $N/T^\tau \to 0$ for $\tau > b$. Theorem 2 is similar to a result obtained by Bonhomme and Manresa (2012). Our proof for this result relies on the assumption that factor loadings $\lambda_{j,i}$ cannot be very small or zero. If individual i's factor loading is zero, then obviously this individual does not belong to any group. The uniform result holds over all individuals whose factor loadings are bounded away from zero. That is, we can always replace $\sup_{i\in\{1,2,\ldots,N\}}$ in Theorem 2 over the set of individuals satisfying $\|\lambda^0_{g_i^0,i}\| \geq a > 0$. Theorem 2 is a very strong result.

Let us define $\tilde{\boldsymbol{\beta}}, \tilde{F}_1, \ldots, \tilde{F}_S, \tilde{\Lambda}_1, \ldots, \tilde{\Lambda}_S$ as the infeasible version of our estimator where group membership G is fixed to its population G^0. It is defined as the minimum of $L_{NT}(\boldsymbol{\beta}, G^0, F_1, \ldots, F_S, \Lambda_1, \ldots, \Lambda_S)$ subject to the constraints $F_j'F_j/T = I_{r_j}$ $(j = 1, \ldots, S)$, and $\Lambda_j'\Lambda_j$ $(j = 1, \ldots, S)$ being diagonal.

Theorem 2 implies that our estimator $\{\hat{\boldsymbol{\beta}}, \hat{G}, \hat{F}_1, \ldots, \hat{F}_S, \hat{\Lambda}_1, \ldots, \hat{\Lambda}_S\}$ is asymptotically equivalent to the infeasible estimates $\{\tilde{\boldsymbol{\beta}}, \tilde{F}_1, \ldots, \tilde{F}_S, \tilde{\Lambda}_1, \ldots, \tilde{\Lambda}_S\}$ as N and T tend to infinity. More precisely, if for some $b > 0$, $N/T^b \to 0$ as both N and T tend to infinity simultaneously, the proposed estimator $\hat{\boldsymbol{\beta}}$, $\hat{F}_j$ $(j = 1, \ldots, S)$ and the infeasible estimator $\tilde{\boldsymbol{\beta}}$, $\tilde{F}_j$ $(j = 1, \ldots, S)$ with known population groups are asymptotically equivalent.

Our proposed method can identify the set of explanatory variables with non-zero coefficients. Let $\boldsymbol{\beta}^0 = (\boldsymbol{\beta}_1^{0\prime}, \boldsymbol{\beta}_2^{0\prime})'$ be the true parameter value, and $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1', \hat{\boldsymbol{\beta}}_2')'$ be the corresponding parameter estimate. Without loss of generality, assume that $\boldsymbol{\beta}_2^0 = \mathbf{0}$. We show that the estimator must possess the sparsity property, $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$. We denote $\hat{\boldsymbol{\beta}}_1$ as the parameter estimate of non-zero true coefficients $\boldsymbol{\beta}_1^0$. To show the asymptotic normality of $\sqrt{NT}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^0)$, we impose the following assumption.

DOI: 10.1002/jae

Assumption F. Let $X_{i,\beta^0\neq 0}$ be the submatrix of X_i corresponding to columns of non-zero elements of the parameter vector $\boldsymbol{\beta}^0$, and q be the number of non-zero elements of $\boldsymbol{\beta}$. For the non-random positive definite matrix $J_0\left(F_1^0,\ldots,F_S^0\right)$

$$\frac{1}{\sqrt{NT}}\sum_{j=1}^{S}\sum_{i:g_i^0=j} Z_{j,i}\left(F_j^0\right)'\boldsymbol{\varepsilon}_i \rightarrow_d N(\mathbf{0}, J_0\left(F_1^0,\ldots,F_S^0\right))$$

where $J_0\left(F_1^0,\ldots,F_S^0\right)$ is the probability limit of

$$\hat{J}\left(F_1^0,\ldots,F_S^0\right)=\frac{1}{NT}\sum_{j=1}^{S}\sum_{k=1}^{S}\sum_{i:g_i^0=j}\sum_{\ell:g_\ell^0=k} Z_{j,i}\left(F_j^0\right)' E[\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_\ell']\, Z_{k,\ell}\left(F_j^0\right)$$

with $Z_{j,i}\left(F_j^0\right) = X'_{i,\beta^0\neq 0}M_{F_j^0} - \frac{1}{N_j}\sum_{k:g_k^0=j} c_{j,ki}X'_{k,\beta^0\neq 0}M_{F_j^0}$, where $c_{j,ki} = \boldsymbol{\lambda}^{0\prime}_{g_k^0,k}\left(\Lambda_j^{0\prime}\Lambda_j^0/N_j\right)^{-1}\boldsymbol{\lambda}^0_{g_i^0,i}$.

The notation $J_0\left(F_1^0,F_2^0,\ldots,F_S^0\right)$ does not mean it still depends on the sample $\left(F_1^0,\ldots,F_S^0\right)$, but rather the limit is taken under the true factors. One could use the notation J_0 in place of $J_0\left(F_1^0,F_2^0,\ldots,F_S^0\right)$. Similar comments apply to $D_0\left(F_1^0,\ldots,F_S^0\right)$ used below.

Then we have the following theorem. Here, we emphasize that the regularization parameter κ depends on T, and thus denote it as κ_T.

Theorem 3. Asymptotic normality and variable selection consistency. Suppose that the assumptions of Theorem 1 hold, and $T/N \rightarrow \rho > 0$. Let $\hat{\boldsymbol{\beta}}_1$ as the parameter estimate of non-zero true coefficients $\boldsymbol{\beta}_1^0$. Then, $\sqrt{NT}(\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_1^0)$ is asymptotically normal with mean $\boldsymbol{v}_0$ and variance–covariance matrix $V_\beta\left(F_1^0,\ldots,F_S^0\right)$, i.e., $\sqrt{NT}(\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_1^0)\rightarrow_d N(\boldsymbol{v}_0, V_\beta\left(F_1^0,\ldots,F_S^0\right))$. Moreover, the following variable selection consistency holds:

$$P(\hat{\boldsymbol{\beta}}_2=\mathbf{0})\rightarrow 1,\quad N,T\rightarrow\infty$$

Here, $\boldsymbol{v}_0$ is the probability limit of

$$\boldsymbol{v}=\sqrt{\frac{T}{N}}\times\sum_{j=1}^{S}\hat{D}\left(F_1^0,\ldots,F_S^0,\kappa\right)^{-1}\boldsymbol{\eta}_j+\sqrt{\frac{N}{T}}\times\sum_{j=1}^{S}\hat{D}\left(F_1^0,\ldots,F_S^0,\kappa\right)^{-1}\boldsymbol{\zeta}_j$$

with

$$\boldsymbol{\eta}_j=-\frac{1}{N_jT}\sum_{i:g_i^0=j}\sum_{k:g_k^0=j}(X_{i,\beta^0\neq 0}-V_{j,i})'F_j^0\left(\frac{F_j^{0\prime}F_j^0}{T}\right)^{-1}\left(\frac{\Lambda_j^{0\prime}\Lambda_j^0}{N_j}\right)^{-1}\boldsymbol{\lambda}_{g_k^0,k}\left(\frac{E[\boldsymbol{\varepsilon}_i'\boldsymbol{\varepsilon}_k]}{T}\right) \tag{3}$$

$$\boldsymbol{\zeta}_j=-\frac{1}{N_jT}\sum_{i:g_i^0=j}\sum_{k:g_k^0=j}X'_{i,\beta^0\neq 0}M_{F_j^0}\Omega_kF_j^0\left(\frac{F_j^{0\prime}F_j^0}{T}\right)^{-1}\left(\frac{\Lambda_j^{0\prime}\Lambda_j^0}{N_j}\right)^{-1}\boldsymbol{\lambda}_{g_i^0,i} \tag{4}$$

$$\begin{aligned}\hat{D}(F_1^0,\ldots,F_S^0,\kappa_T)&=\frac{1}{NT}\sum_{j=1}^{S}\sum_{i:g_i^0=j}X'_{i,\beta^0\neq 0}M_{F_j^0}X_{i,\beta^0\neq 0}\\&\quad-\frac{1}{NT}\sum_{j=1}^{S}\frac{1}{N_j}\sum_{i:g_i^0=j}\sum_{k:g_k^0=j}X'_{i,\beta^0\neq 0}M_{F_j^0}X_{k,\beta^0\neq 0}c_{j,ki}+\frac{1}{NT}\Sigma(\kappa_T)\end{aligned}$$

where $V_{j,i} = N_j^{-1} \sum_{k:g_k^0=j} c_{j,ki} X_{k,\beta^0 \neq 0}$, $X_{i,\beta^0 \neq 0}$ is the submatrix X_i corresponding to the columns of the non-zero element of $\boldsymbol{\beta}^0$, $c_{j,ki}$ is defined in Assumption F, and $\Sigma(\kappa_T)$ is defined as $\Sigma(\kappa_T) = \text{diag}\{p'_{\kappa_T,\gamma}(|\beta_{10}|)/|\beta_{10}|, \ldots, p'_{\kappa_T,\gamma}(|\beta_{q0}|)/|\beta_{q0}|\}$ where q is the number of non-zero elements of $\boldsymbol{\beta}^0$, and $\Omega_k = E[\boldsymbol{\varepsilon}_k \boldsymbol{\varepsilon}_k']$. The asymptotic covariance matrix $V_\beta\left(F_1^0, \ldots, F_S^0\right)$ is given by

$$V_\beta\left(F_1^0, \ldots, F_S^0\right) = D_0\left(F_1^0, \ldots, F_S^0\right)^{-1} J_0\left(F_1^0, \ldots, F_S^0\right) D_0\left(F_1^0, \ldots, F_S^0\right)^{-1}$$

where $D_0\left(F_1^0, \ldots, F_S^0\right)$ is the probability limit of $\hat{D}\left(F_1^0, \ldots, F_S^0, \kappa_T\right)$.

This indicates that we can perform statistical significance tests. Note that the bias $\boldsymbol{v}_0$ can be consistently estimated as in Bai (2009), Hahn and Kuersteiner (2002), Hahn and Newey (2004 and Arellano and Hahn (2005), so bias correction can be performed. Also, the bias $\boldsymbol{v}_0$ will become zero in the absence of correlations and heteroskedasticity. In particular, $\boldsymbol{\eta}_j = \mathbf{0}$ when cross-sectional correlation and heteroskedasticity are absent in ε_{it}, and similarly $\boldsymbol{\zeta}_j = \mathbf{0}$ when serial correlation and heteroskedasticity are absent in ε_{it}. There will be no bias if ε_{it} are i.i.d. over t and over i. Thus bias correction can be simplified depending on the assumptions made on ε_{it}.

The case of $p \to \infty$ as well as $S \to \infty$ was considered in a previous version. However, their rate going to infinity is slow. The case of faster divergence of p and S would require a new argument, and is beyond the scope of this paper.

The estimation algorithm requires knowledge of the number of groups, the number of group-specific factors and the size of the regularization parameter κ. In practice, however, we have to select these quantities. In the next section, we propose a new criterion to select these parameters.

5. A C_P-TYPE CRITERION FOR MODEL SELECTION

In this section, we denote the true number of groups and the true number of group-specific factors in each group as S_0 and $\left\{r_1^0, \ldots, r_{S_0}^0\right\}$, respectively. Also, we let S be the specified number of groups and $\left\{r_1^*, \ldots, r_S^*\right\}$ be the corresponding true number of group-specific factors. Note that, if two true groups G_k and G_ℓ having the true number of group-specific factors r_k^0 and r_ℓ^0 are merged, then the resulting group (called G_j^*, say) has $r_j^* = r_k^0 + r_\ell^0$ number of factors. Thus, when S is different from S_0, $\left\{r_1^*, \ldots, r_S^*\right\}$ is needed for dealing with the notion of the true number of group-specific factors. When $S = S_0$, $\left\{r_1^*, \ldots, r_S^*\right\}$ and $\left\{r_1^0, \ldots, r_S^0\right\}$ are identical after an appropriate permutation of labeling. Our aim is to select S_0 and $\left\{r_1^0, \ldots, r_{S_0}^0\right\}$ consistently, while maintaining the predictive power that also depends on the value of regularization parameter κ.

Suppose that $z_1, \ldots, z_N$ are replicates of the response variables $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N$ given true values of the factors F_j, factor loadings Λ_j and the design matrices X_i $(i = 1, \ldots, N)$. To assess the predictive ability of the estimated model, we consider the expected MSE

$$\eta(S, k_1, \ldots, k_S, \kappa) := E_z\left[\frac{1}{NT} \sum_{j=1}^{S} \sum_{i;\hat{g}_i=j}^{N_j} \left|\left|z_i - X_i \hat{\boldsymbol{\beta}} - \hat{F}_{\hat{g}_i} \hat{\boldsymbol{\lambda}}_{\hat{g}_i,i}\right|\right|^2\right] \tag{5}$$

where $k_1, \ldots, k_S$ are the given number of group-specific factors, κ is the regularization parameter and the expectation $E_z[\cdot]$ is taken with respect to the joint distribution of $z_1, \ldots, z_N$ conditional on the true factor structure and the set of predictors X_i. The best model is chosen by minimizing the expected MSE.

A natural estimator of the expected MSE in equation (5) is the sample-based MSE:

$$\hat{\eta}(S, k_1, \ldots, k_S, \kappa) := \frac{1}{NT}\sum_{j=1}^{S}\sum_{i;\hat{g}_i=j}^{N_j}\left\|\boldsymbol{y}_i - X_i\hat{\boldsymbol{\beta}} - \hat{F}_{\hat{g}_i}\hat{\boldsymbol{\lambda}}_{\hat{g}_i,i}\right\|^2$$

This quantity is formally calculated by replacing the replicates z_i with an observed value $\boldsymbol{y}_i$. This sample-based MSE generally has some bias with respect to the expected MSE because, among other reasons, the same data are used to estimate the parameters of the model. We therefore consider a bias-corrected version of the measure.

The bias b of the sample-based MSE $\hat{\eta}$ with respect to the expected MSE η is

$$b := E_y\left[\eta(S, k_1, \ldots, k_S, \kappa) - \hat{\eta}(S, k_1, \ldots, k_S, \kappa)\right] \tag{6}$$

where the expectation $E_y[\cdot]$ is taken with respect to the joint distribution of $\boldsymbol{y}_i (i = 1, \ldots, N)$ conditional on the true factor structure and the set of predictors X_i.

In the previous section, Theorem 3 showed the asymptotic normality of the penalized estimator of regression coefficients. This claim can be obtained as long as the specified number of factors for each group is sufficiently large, or more precisely, $r_1^* \le k_1, r_2^* \le k_2, \ldots, r_S^* \le k_S$ holds. This implies that a specified model with $k_j < r_j^*$ for some j asymptotically has the larger expected MSE than that based on a model that satisfies $r_1^* \le k_1, r_2^* \le k_2, \ldots, r_S^* \le k_S$. This is because a model with insufficient number of group-specific factors has the following characteristics: (i) the underlying factor structure can not be captured; and (ii) the estimated $\hat{\boldsymbol{\beta}}$ may no longer be consistent. Thus we focus only on the situation where the specified number of factors for each group is sufficiently large. We point out that selecting true numbers (S_0 and $\left\{r_1^0, \ldots, r_{S_0}^0\right\}$) is feasible by using our C_p criterion, which will be given later. Now, we provide the following lemma.

Lemma 1. Under the assumptions of Theorem 3, if $r_1^* \le k_1, r_2^* \le k_2, \ldots, r_S^* \le k_S$, then the bias term b is of the form

$$b(S, k_1, \ldots, k_S, \kappa) = \frac{1}{NT}\mathrm{tr}\left[K_x V_\beta\left(F_1^0, \ldots, F_S^0, \kappa\right)\right] + o\left(\frac{1}{NT}\right)$$

where $K_x = 2(NT)^{-1}\sum_{i=1}^{N} X'_{i,\hat{\beta}\neq 0}X_{i,\hat{\beta}\neq 0}$ with $X_{i,\hat{\beta}\neq 0}$ being the submatrix of X_i such that the corresponding columns contain a non-vanishing component of the parameter estimate, and $V_\beta\left(F_1^0, \ldots, F_S^0, \kappa\right) = \hat{D}\left(F_1^0, \ldots, F_S^0, \kappa\right)^{-1}\hat{J}\left(F_1^0, \ldots, F_S^0\right)\hat{D}\left(F_1^0, \ldots, F_S^0, \kappa\right)^{-1}$. Here, $\hat{J}\left(F_1^0, \ldots, F_S^0\right)$ and $\hat{D}\left(F_1^0, \ldots, F_S^0, \kappa\right)$ are defined in Assumption F and Theorem 3.

As shown in the derivation, Lemma 1 provides a bias estimation caused by $\hat{\boldsymbol{\beta}}$ given the factor structure. The matrices K_x and $V_\beta\left(F_1^0, \ldots, F_S^0, \kappa\right)$ consist of the submatrix of X_i with columns containing the non-vanishing component of the parameter estimate, and thus the bias term implicitly depends on κ. A smaller value of κ leads to larger dimensions of K_x and $V_\beta\left(F_1^0, \ldots, F_S^0, \kappa\right)$. The trace $\mathrm{tr}\left[K_x V_\beta\left(F_1^0, \ldots, F_S^0, \kappa\right)\right]$ is $O(\dim\{\hat{\beta} \neq 0\})$ and is similar to the concept of the number of free parameters (e.g. Akaike, 1974).

In view of Lemma 1, we estimate b by

$$\hat{b}(S, k_1, \ldots, k_S, \kappa) = \frac{1}{NT}\mathrm{tr}\left[K_x V_\beta(\hat{F}_1, \ldots, \hat{F}_S, \kappa)\right]$$

then

$$\frac{1}{NT}\sum_{j=1}^{S}\sum_{i;\hat{g}_i=j}^{N_j}\left\|\boldsymbol{y}_i - X_i\hat{\boldsymbol{\beta}} - \hat{F}_{\hat{g}_i}\hat{\boldsymbol{\lambda}}_{\hat{g}_i,i}\right\|^2 + \hat{b}(S,k_1,\ldots,k_S,\kappa)$$

provides an estimate for the predictive mean squared error given $\{S,k_1,\ldots,k_S,\kappa\}$. The regularization parameter κ is chosen to minimize this predictive mean squared error.

As explained in the paragraph preceding Lemma 1, underestimating the number of factors is costly (see also Bai and Ng, 2002; Bai, 2009). But we also need to pay attention not to overestimate the number of factors. Additional penalty $\sum_{j=1}^{S} k_j g_j(T,N_1,\ldots,N_S)$ on the number of group-specific factors can solve this problem. Here $g_j(T,N_1,\ldots,N_S)$ satisfies (a) $g_j(T,N_1,\ldots,N_S)\to 0$ and (b) $\min\{N,T\}\times g_j(T,N_1,\ldots,N_S)\to\infty$ as $T,N\to\infty$. An example of the function $g_j(T,N_1,\ldots,N_S)$ that satisfies conditions (a) and (b) of the theorem is $g_j(T,N_1,\ldots,N_S)=\frac{N_j}{N}\times\frac{T+N_j}{TN_j}\log(TN_j)$. Theoretical derivation of these two conditions is provided in the supplementary document.

Adding together the sample-based MSE $\hat{\eta}(S,k_1,\ldots,k_S,\kappa)$, the bias term $\hat{b}(S,k_1,\ldots,k_S,\kappa)$ and the penalty on the number of factors, the final criterion is

$$\begin{aligned} C_p(S,k_1,\ldots,k_S,\kappa) &= \frac{1}{NT}\sum_{j=1}^{S}\sum_{i;\hat{g}_i=j}\|\boldsymbol{y}_i - X_i\hat{\boldsymbol{\beta}} - \hat{F}_{\hat{g}_i}\hat{\boldsymbol{\lambda}}_{\hat{g}_i,i}\|^2 \\ &\quad + \frac{1}{TN}\mathrm{tr}\left[K_x V_\beta(\hat{F}_1,\ldots,\hat{F}_S,\kappa)\right] + \sum_{j=1}^{S}k_j\hat{\sigma}^2\frac{N_j}{N}\left(\frac{T+N_j}{TN_j}\right)\log\left(TN_j\right) \end{aligned} \tag{7}$$

where $\hat{\sigma}^2$ is an estimator of $(NT)^{-1}\sum_{j=1}^{S}\sum_{i;g_i^0=j}\|\boldsymbol{y}_i - X_i\hat{\boldsymbol{\beta}} - \hat{F}_{g_i^0}\hat{\boldsymbol{\lambda}}_{g_i^0,i}\|^2$.

We can regard the proposed criterion as a generalization of the C_p criterion of Mallows (1973) in the panel data context with unobservable interactive effects. Like the C_p criterion, $\hat{\sigma}^2$ provides proper scaling for the penalty term. In applications, it can be replaced by $(NT)^{-1}\sum_{j=1}^{S}\sum_{i;\hat{g}_i=j}\|\boldsymbol{y}_i - X_i\hat{\boldsymbol{\beta}} - \hat{F}_{\hat{g}_i}\hat{\boldsymbol{\lambda}}_{\hat{g}_i,i}\|^2$, which is obtained under the maximum possible dimension of X_i, the maximum possible number of groups $S_{\max}$ and the maximum possible number of group-specific factors $r_{j,\max}$, $j=1,\ldots,S$. Finally, we provide the following theorem.

Theorem 4. Let $\hat{S}$ and $\{\hat{r}_1,\ldots,\hat{r}_{\hat{S}}\}$ be the minimizer of the proposed C_p criterion. Suppose that the assumptions in Theorem 3 hold, and the number of units satisfies $N/T\to c$ where $c\geq 0$. Then, the determined number of groups $\hat{S}$ and common factors $\{\hat{r}_1,\ldots,\hat{r}_{\hat{S}}\}$ converge in probability to the true number of groups S_0 and true number of common factors $\left\{r_1^0,\ldots,r_{S_0}^0\right\}$ as $T,N\to\infty$.

Thus the number of factors and the number of groups can be identified as the minimizer of our C_p criterion. The following is a procedure for jointly selecting the values of these quantities.

Model selection algorithm.

Step 1. Prepare a set of candidate values of the regularization parameter κ, the number of groups $S=\{1,2\ldots,S_{\max}\}$ and the number of group-specific factors $\{k_1,\ldots,k_S\}$.

Step 2. Fix the value of the number of groups S.

Step 3. Fix the value of the regularization parameter κ.

Step 4. Given the number of groups S and the regularization parameter κ, we optimize the number of group-specific factors $\{k_1,\ldots,k_S\}$.

Step 5. Repeat steps 3 and 4 under the different values of κ.

Step 6. Repeat steps 2–5 under the different number of groups S. Then select the combination of the regularization parameter κ, the number of group-specific factors $\{k_1, \ldots, k_S\}$ and the number of groups S that minimize the C_p score.

6. SIMULATION STUDY

We consider three different data-generating processes (DGP). The first DGP is $\boldsymbol{y}_i = X_i\boldsymbol{\beta} + F_{g_i}\boldsymbol{\lambda}_{g_i,i} + \boldsymbol{\varepsilon}_i$, where the r_j-dimensional group-specific pervasive factor $\boldsymbol{f}_{j,t}$ $(j = 1, \ldots, S)$ is a vector of $N(j, 1)$ variables, and each element of the factor loading matrix Λ_j follows $N(0, j)$. The N-dimensional vector $\boldsymbol{\varepsilon}_t$ has a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix I_N. The number of columns of X_i is set to $p = 80$, while the true number of predictors is $q = 3$. Each of the elements of X_i is generated from the uniform distribution over $[-2, 2]$. The non-zero true parameter values of $\boldsymbol{\beta}$ are set to be $(1, 2, 3)$. These non-zero elements are put into the first three elements of $\boldsymbol{\beta}_i$ and thus the true parameter vector is $\boldsymbol{\beta} = (1, 2, 3, 0, 0, \ldots, 0)'$. We set the number of groups $S = 3$, and the true numbers of group-specific pervasive factors are $r_1 = 3$ $r_2 = 3$, $r_3 = 3$. Setting the number of units in each group as $N_1 = N_2 = N_3$, we generated a set of T observations with various (N, T) combinations.

The second DGP is non-homoskedastic errors with cross-sectional dependence such that $\boldsymbol{y}_i = X_i\boldsymbol{\beta} + F'_{g_i}\boldsymbol{\lambda}_{g_i,i} + \boldsymbol{\varepsilon}_i$ and $\varepsilon_{it} = 0.9e^1_{it} + \delta_t 0.9e^2_{it}$, where $\delta_t = 1$ if t is odd and zero if t is even, and the N-dimensional vectors $\boldsymbol{e}^1_t = (e^1_{1t}, \ldots, e^1_{Nt})'$ and $\boldsymbol{e}^2_t = (e^2_{1t}, \ldots, e^2_{Nt})'$ follow multivariate normal distributions with mean $\mathbf{0}$ and covariance matrix $S = (s_{ij})$, with $s_{ij} = 0.3^{|i-j|}$, and $\boldsymbol{e}^1_t$ and $\boldsymbol{e}^2_t$ are independent. The noise terms are not serially correlated. The group-specific factors and their loading matrices, the design matrix X_i and the true parameter vector $\boldsymbol{\beta}$ are generated by the same method as before.

The third DGP allows the errors to have serial and cross-sectional correlations. The model is $\boldsymbol{y}_i = X_i\boldsymbol{\beta} + F_{g_i}\boldsymbol{\lambda}_{g_i,i} + \boldsymbol{\varepsilon}_i$ with $\varepsilon_{it} = 0.2\varepsilon_{i,t-1} + e_{it}$, where $t = 1, \ldots, T$, the N-dimensional vector $\boldsymbol{e}_t = (e_{1t}, \ldots, e_{Nt})'$ follows multivariate normal distributions with mean $\mathbf{0}$ and covariance matrix $S = (s_{ij})$, where $s_{ij} = 0.3^{|i-j|}$. The other variables are defined as before.

We generated 1000 replications using each of the three data-generating models. We then applied the proposed model selection criterion, C_p, to select simultaneously the number of groups, the number of group-specific pervasive factors and the size of the regularization parameter. We set the possible numbers of group-specific factors to range from zero to eight. Thus the maximum number of group-specific factors was set to eight. The number of groups ranges from two to four. Possible candidates for the regularization parameter κ are $\kappa = \{10, 1, 0.1, 0.01, 0.001\}$.

Our simulation results showed that the proposed C_p criterion works well in selecting the number of groups and the number of group-specific pervasive factors. Furthermore, the regression coefficients $\hat{\boldsymbol{\beta}}$ are estimated very well. Owing to space constraits, details are reported in the online Appendix.

7. HETEROGENEOUS GROUP-SPECIFIC COEFFICIENTS

The model (1) can be extended to the heterogeneous group-specific coefficients:

$$y_{it} = \boldsymbol{x}'_{it}\boldsymbol{\beta}_{g_i} + \boldsymbol{f}'_{g_i,t}\boldsymbol{\lambda}_{g_i,i} + \varepsilon_{i,t}, \quad i = 1, \ldots, N,\ t = 1, \ldots, T \tag{8}$$

where the $p_i \times 1$ vector $\boldsymbol{\beta}_{g_i}$ contains the unknown regression coefficients for each group. Previous research supports heterogeneous regression coefficients (e.g. Hsiao and Tahmiscioglu, 1997; Lin and Ng, 2012). Here the regression coefficients are group specific but not individual specific. It may be of interest to extend the model to individual-dependent coefficients, which is not studied in this paper. The model assumptions are the same as in Section 2.1, except we need to modify Assumption D as follows.

DOI: 10.1002/jae

Assumption D$'$: Observable predictors.

D1$'$: For the matrices D_j, E_j and L_j defined in Assumption D of Section 2.1, we assume $D_j - L_j' E_j^{-1} L_j$ is positive definite for all F_j such that $F_j' F_j / T = I$ and for all groupings with a positive fraction of membership. Assumption D2 is maintained.

D2$'$: The vector of explanatory variables $\boldsymbol{x}_{it}$ satisfies $\max_{1 \le i \le N} T^{-1} \| X_i \|^2 = O_p(N^{\alpha})$ with $\alpha < 1/16$. We also assume $N/T^2 \to 0$.

In D2$'$, we now require $\alpha < 1/16$ instead of $\alpha < 1/8$. Again, this is much weaker than assuming x_{it} has exponential tails. Assumption D$'$ ensures the existence of the asymptotic variance matrix of the estimated regression coefficients. This condition is used for the proof of consistency.

7.1. Estimation and Asymptotic Results

Under a given number of groups S, number of factors $r_1, \ldots, r_S$ and the penalization κ, the estimator $\{\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_S, \hat{G}, \hat{F}_1, \ldots, \hat{F}_S, \hat{\Lambda}_1, \ldots, \hat{\Lambda}_S\}$ is defined as the minimizer of

$$L_{NT}(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_S, G, F_1, \ldots, F_S, \Lambda_1, \ldots, \Lambda_S) = \sum_{j=1}^{S} \sum_{i; g_i = j} \| \boldsymbol{y}_i - X_i \boldsymbol{\beta}_{g_i} - F_{g_i} \boldsymbol{\lambda}_{g_i,i} \|^2 + \sum_{j=1}^{S} NT \cdot p_{\kappa,\gamma} \left(|\boldsymbol{\beta}_j| \right)$$

subject to the constraints on the factor and factor loading matrices in Section 2.

The estimation is again through iterations. Given the group membership G and the values of regression coefficient $\boldsymbol{\beta}_j$, the factor structures are estimated as described in Section 2. Given the group membership G and the factor structures, the regression coefficients $\boldsymbol{\beta}_j$ is then updated. It is easy to see that, for any given values of $\boldsymbol{\beta}_j$ and $F_j \boldsymbol{\lambda}_{j,i}$ ($j = 1, \ldots, S$), the optimal assignment for each individual unit is: $g_i^* = \arg\min_{j \in \{1,\ldots,S\}} T^{-1} \| \boldsymbol{y}_i - X_i \boldsymbol{\beta}_j - F_j \boldsymbol{\lambda}_{j,i} \|^2 + p_{\kappa,\gamma}(|\boldsymbol{\beta}_j|)$. The estimates of $\boldsymbol{\beta}$, $\{F_j, \Lambda_j;\ j = 1, \ldots, S\}$ and $G = \{g_1, \ldots, g_N\}$ depend on each other and thus iterations are needed.

Hereafter, we use $\left\{F_j^0,\ j = 1, \ldots, S\right\}$ to denote the true parameter values of the group-specific factors F_j. As N and T increase, the estimated factors are consistent in the sense of some averaged norm. We have the following theorem.

Theorem 5. Consistency. Under Assumptions A, B, C, D$'$ and E, $\kappa \to 0$ and $\min\{N_j, T\} \times \kappa \to \infty$ as $T, N \to \infty$, the estimators $\hat{\boldsymbol{\beta}}_j$ are consistent:

$$\| \hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^0 \| = o_p(1), \quad \text{for} \quad j = 1, \ldots, S$$

In addition, $\{\hat{F}_j,\ j = 1, \ldots, S\}$ are consistent in the sense of

$$T^{-1/2} \| \hat{F}_j - F_j^0 H_j \| = o_p(1)$$

where $H_j^{-1} = V_{j,N_jT} (F_j^0 \hat{F}_j / T)^{-1} (\Lambda_j^{0\prime} \Lambda_j^0 / N_j)^{-1}$, and V_{j,N_jT} satisfies

$$\left[\frac{1}{N_j T} \sum_{i; \hat{g}_i = j}^{N_j} (\boldsymbol{y}_i - X_i \hat{\boldsymbol{\beta}}_j)(\boldsymbol{y}_i - X_i \hat{\boldsymbol{\beta}}_j)' \right] \hat{F}_j = \hat{F}_j V_{j,N_jT}$$

DOI: 10.1002/jae

The estimates of $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_S, \{F_j, \Lambda_j;\ j = 1, \dots, S\}$ and $G = \{g_1, \dots, g_N\}$ depend on each other, and we therefore denote the estimator of group membership $\hat{g}_i$ as $\hat{g}_i(\hat{\boldsymbol{\beta}}_1, \dots, \hat{\boldsymbol{\beta}}_S, \hat{F}, \hat{\Lambda})$. The following theorem shows that the estimated group membership converges to the true group membership as T and N go to infinity.

Theorem 6. Consistency of the estimator of group membership. Suppose that the assumptions in Theorem 5 hold. Then, for all $\tau > 0$ and $T, N \to \infty$, we have

$$P\left(\sup_{i \in \{1,\dots,N\}} |\hat{g}_i(\hat{\boldsymbol{\beta}}_1, \dots, \hat{\boldsymbol{\beta}}_S, \hat{F}, \hat{\Lambda}) - g_i^0| > 0\right) = o(1) + o(N/T^\tau)$$

where $\hat{F} = \{\hat{F}_1, \dots, \hat{F}_S\}$ and $\hat{\Lambda} = \{\hat{\Lambda}_1, \dots, \hat{\Lambda}_S\}$.

In the next theorem, we provide the asymptotic normality and the variable selection consistency. Let $\boldsymbol{\beta}_j^0 = \left(\boldsymbol{\beta}_{j,1}^{0\ \prime}, \boldsymbol{\beta}_{j,2}^{0\ \prime}\right)'$ be the true parameter vector such that $\boldsymbol{\beta}_{j,2}^0 = \mathbf{0}$. We denote the corresponding estimate as $\hat{\boldsymbol{\beta}}_j = \left(\hat{\boldsymbol{\beta}}_{j,1}', \hat{\boldsymbol{\beta}}_{j,2}'\right)'$. We show that $P(\hat{\boldsymbol{\beta}}_{j,2} = \mathbf{0})$ will converge to 1 as $N, T \to \infty$. Also $\hat{\boldsymbol{\beta}}_{j,1}$ is asymptotically normal.

Assumption F′. Let $X_{i,\beta_j^0 \neq \mathbf{0}}$ or simply $X_{i,j}$ be the submatrix of X_i corresponding to columns of the non-zero elements of $\boldsymbol{\beta}_j^0$. Let q_j be the number of non-zero elements of $\boldsymbol{\beta}_j^0$ $(j = 1, \dots, S)$. For the non-random positive definite matrix $J_0\left(F_j^0\right)$:

$$\frac{1}{\sqrt{N_j T}} \sum_{i:g_i^0=j} Z_{j,i}\left(F_j^0\right)' \boldsymbol{\varepsilon}_i \to_d N\left(\mathbf{0}, J_0\left(F_j^0\right)\right)$$

where $Z_{j,i}\left(F_j^0\right) = X'_{i,\beta^0 \neq \mathbf{0}} M_{F_j^0} - N_j^{-1} \sum_{k:g_k^0=j} c_{j,ki} X'_{k,\beta^0 \neq \mathbf{0}} M_{F_j^0}$, with $c_{j,ki} = \boldsymbol{\lambda}_{g_k^0,k}^{0\prime} \left(\Lambda_j^{0\prime} \Lambda_j^0 / N_j\right)^{-1} \boldsymbol{\lambda}_{g_i^0,i}^0$, and $J_0\left(F_j^0\right)$ is the probability limit of

$$\hat{J}(F_j^0) = \frac{1}{N_j T} \sum_{i:g_i^0=j} \sum_{\ell:g_\ell^0=j} Z_{j,i}\left(F_j^0\right)' E\left[\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_\ell'\right] Z_{j,\ell}\left(F_j^0\right)$$

Then, we have the following theorem.

Theorem 7. Asymptotic normality and variable selection consistency. Assume that the assumptions in Theorems 5 and 6 and F' hold. Then, $\sqrt{N_j T}\left(\hat{\boldsymbol{\beta}}_{j,1} - \boldsymbol{\beta}_{j,1}^0\right)$ is asymptotically normal with mean $\boldsymbol{v}_j^0$ and variance–covariance matrix $V_\beta\left(F_j^0\right)$, i.e. $\sqrt{N_j T}\left(\hat{\boldsymbol{\beta}}_{j,1} - \boldsymbol{\beta}_{j,1}^0\right) \to_d N\left(\boldsymbol{v}_0^j, V_\beta\left(F_j^0\right)\right)$. Moreover, the following variable selection consistency holds:

$$P(\hat{\boldsymbol{\beta}}_{j,2} = \mathbf{0}) \to 1 \quad N, T \to \infty$$

for $j = 1, \dots, S$. Here, the variance–covariance matrix $V_{\beta_j}\left(F_j^0\right)$ is

$$V_\beta\left(F_j^0\right) = D_0\left(F_j^0\right)^{-1} J_0\left(F_j^0\right) D_0\left(F_j^0\right)^{-1}$$

where $D_0\left(F_j^0\right)$ is the probability limit of

$$\hat{D}(F_j^0, \kappa_T) = \frac{1}{N_j T} \sum_{i:g_i^0=j} \left[X_{i,j}' M_{F_j^0} X_{i,j} - \frac{1}{N_j} \sum_{k:g_k^0=j} c_{j,ki} X_{i,j}' M_{F_j^0} X_{k,j} \right] + \frac{1}{N_j T} \Sigma_j(\kappa_T)$$

with $\Sigma_j(\kappa_T) = \text{diag}\left\{ p_{\kappa_T,\gamma}'\left(|\beta_{j,1}^0|\right)/|\beta_{j,1}^0|, \ldots, p_{\kappa_T,\gamma}'\left(|\beta_{j,q_j}^0|\right)/|\beta_{j,q_j}^0| \right\}$, where q_j is the number of non-zero elements of $\boldsymbol{\beta}_j^0$, and $\boldsymbol{v}_0^j$ is the probability limit of

$$\sqrt{\frac{T}{N_j}} \times \hat{D}\left(F_j^0, \kappa_T\right)^{-1} \boldsymbol{\eta}_j + \sqrt{\frac{N_j}{T}} \times \hat{D}\left(F_j^0, \kappa_T\right)^{-1} \boldsymbol{\zeta}_j$$

where

$$\boldsymbol{\zeta}_j = -\frac{1}{N_j T} \sum_{i:g_i^0=j} \sum_{k:g_k^0=j} X_{i,j}' M_{\tilde{F}_j} \Omega_k F_j^0 \left(\frac{F_j^{0\prime} F_j^0}{T} \right)^{-1} \left(\frac{\Lambda_j^{0\prime} \Lambda_j^0}{N_j} \right)^{-1} \boldsymbol{\lambda}_{g_i^0,i},$$

$$\boldsymbol{\eta}_j = -\frac{1}{N_j T} \sum_{i:g_i^0=j} \sum_{k:g_k^0=j} (X_i - V_{j,i})' F_j^0 \left(\frac{F_j^{0\prime} F_j^0}{T} \right)^{-1} \left(\frac{\Lambda_j^{0\prime} \Lambda_j^0}{N_j} \right)^{-1} \boldsymbol{\lambda}_{g_k^0,k} \left(\frac{E[\boldsymbol{\varepsilon}_i' \boldsymbol{\varepsilon}_k]}{T} \right)$$

with $c_{j,ki} = \boldsymbol{\lambda}_{g_k^0,k}^{0\prime} \left(\Lambda_j^{0\prime} \Lambda_j^0 / N_j \right)^{-1} \boldsymbol{\lambda}_{g_i^0,i}^0$, and $V_{j,i} = N_j^{-1} \sum_{k:g_k^0=j} c_{j,ki} X_{k,j}$.

The proof of the theorem is given in the online supplement.

7.2. Determining the Number of Groups/Factors

Taking into account the consistency of the proposed model selection criterion, we again suggest minimization of $\frac{1}{NT} \sum_{j=1}^{S} \sum_{i;\hat{g}_i=j} \| \boldsymbol{y}_i - X_i \hat{\boldsymbol{\beta}}_{g_i} - \hat{F}_{\hat{g}_i} \hat{\boldsymbol{\lambda}}_{\hat{g}_i,i} \|^2 +$ (Penalty term). The first term measures the goodness of fit of the model, whereas the second term is a penalty on the complexity of the model. It remains to construct a proper penalty term. Again, our aim is to select the true number of groups S_0 and the true number of group-specific factors $\left\{ r_1^0, \ldots, r_{S_0}^0 \right\}$ consistently, while maintaining the predictive power that also depends on the value of regularization parameter κ.

Using the same investigations as in Section 5, we obtained the following penalty term:

$$\text{Penalty} = \sum_{j=1}^{S} \frac{1}{NT} \text{tr}\left[K_{j,x} V_\beta\left(F_j^0, \kappa\right) \right] + \sum_{j=1}^{S} k_j \times g_j(T, N_1, \ldots, N_S) \tag{9}$$

where $K_{j,x} = 2 \sum_{i;g_i=j} X_{i,\hat{\beta}_j \neq 0}' X_{i,\hat{\beta}_j \neq 0} / (N_j T)$ with $X_{i,\hat{\beta}_j \neq 0}$ being the submatrix of X_i such that the corresponding columns contain a non-vanishing component of the parameter estimate, and $V_\beta\left(F_j^0, \kappa\right) = \hat{D}\left(F_j^0, \kappa\right)^{-1} \hat{J}\left(F_j^0\right) \hat{D}\left(F_j^0, \kappa\right)^{-1}$. Here $\hat{J}\left(F_j^0\right)$ and $\hat{D}\left(F_j^0, \kappa\right)$ are defined in Assumption F′ and Theorem 7. The function $g_j(T, N_1, \ldots, N_S)$ satisfies (a) $g_j(T, N_1, \ldots, N_S) \rightarrow 0$ and (b) $\min\{N, T\} \times g_j(T, N_1, \ldots, N_S) \rightarrow \infty$ as $T, N \rightarrow \infty$.

DOI: 10.1002/jae

Specifying $g_j(T, N_1, \ldots, N_S) = \hat{\sigma}^2 \frac{N_j}{N} \left(\frac{T+N_j}{TN_j}\right) \log(TN_j)$, we propose the following C_p criterion:

$$\begin{aligned} C_p(S, k_1, \ldots, k_S, \kappa) = & \frac{1}{NT} \sum_{j=1}^{S} \sum_{i;\hat{g}_i=j} \|\boldsymbol{y}_i - X_i \hat{\boldsymbol{\beta}}_{\hat{g}_i} - \hat{F}_{\hat{g}_i} \hat{\boldsymbol{\lambda}}_{\hat{g}_i,i}\|^2 \\ & + \sum_{j=1}^{S} \frac{1}{NT} \mathrm{tr}\left[K_{j,x} V_\beta(\hat{F}_j, \kappa)\right] + \sum_{j=1}^{S} k_j \hat{\sigma}^2 \frac{N_j}{N} \left(\frac{T+N_j}{TN_j}\right) \log\left(TN_j\right) \end{aligned} \tag{10}$$

where $\hat{\sigma}^2$ is a consistent estimate of $(NT)^{-1} \sum_{j=1}^{S} \sum_{g_i^0=j} \|\boldsymbol{y}_i - X_i \hat{\boldsymbol{\beta}}_{g_i^0} - \hat{F}_{g_i^0} \hat{\boldsymbol{\lambda}}_{g_i^0,i}\|^2$. Under the criterion, the numbers of factors are consistently estimated.

Similar to the C_p criterion, $\hat{\sigma}^2$ provides proper scaling for the penalty term. In applications, it can be replaced by its consistent estimator. Finally, we provide the following theorem, which states that the true number of groups and the true number of group-specific factors can be identified as the minimizer of the proposed C_p criterion.

Theorem 8. Let $\hat{S}$ and $\{\hat{r}_1, \ldots, \hat{r}_{\hat{S}}\}$ be the minimizer of the proposed $C_p(k_1, \ldots, k_S, \kappa)$ criterion in equation (10). Suppose that assumptions of Theorems 5–7 hold. If $N/T \to c$ where $c \geq 0$ hold, then the determined number of groups, $\hat{S}$, and the determined number of group-specific factors $\{\hat{r}_1, \ldots, \hat{r}_{\hat{S}}\}$ converge in probability to the true number of groups S_0 and the true number of group-specific factors $\left\{r_1^0, \ldots, r_{S_0}^0\right\}$ as $T, N \to \infty$.

The derivation of the criterion function and the proof of Theorem 8 is provided in the supplementary document.

8. APPLICATIONS

8.1. Analysis of US Mutual Fund Styles

A mutual fund is a portfolio of financial assets managed by a professional institution on behalf of its clients. It is common that the professional institutions manage clients' assets according to a particular investment style, which defines the nature of the fund. There are well-known criteria that define the investment styles, e.g. 'Value' and 'Growth', 'Large Cap' and 'Small Cap'. To provide investors with a guide to the mutual funds market, some professional institutions issue classifications of existing mutual funds according to the investment objectives stated by the funds. Practically, one may rely on the institutional classification scheme; however, it does not always provide consistent and representative peer groups of fund styles. In this section, we aim at grouping mutual funds and identifying their styles by analyzing the time series of past returns of each mutual fund.

8.1.1. Data and Preliminary Analysis

We analyze $T = 85$ monthly returns y_{it} for $N = 536$ US mutual funds, collected from Thomson Financial Datastream database for October 2003 to October 2010. Here we focus mainly on the four mutual fund styles: Small Capital & Growth, Large Capital & Growth, Small Capital & Value, and Large Capital & Value.

We first consider the model with homogeneous slope coefficients:

$$\begin{aligned} y_{it} = & \beta_0 + \mathrm{Mkt}_t \beta_{\mathrm{Mkt}} + \mathrm{HML}_t \beta_{\mathrm{HML}} + \mathrm{SMB}_t \beta_{\mathrm{SMB}} + \mathrm{LTR}_t \beta_{\mathrm{LTR}} \\ & + \mathrm{STR}_t \beta_{\mathrm{STR}} + \mathrm{Mom}_t \beta_{\mathrm{Mom}} + \boldsymbol{f}'_{g_i,t} \boldsymbol{\lambda}_{g_i,i} + \varepsilon_{it} \end{aligned}$$

$i = 1, \ldots, N$, $t = 1, \ldots, T$,. The three factors, SMB, HML, and Mkt are suggested by Fama and French (1993), they are based on size, book-to-market ratio and market returns, respectively. We also used the long-term return reversal factor (RTR), the short-term return reversal factor (STR) and the momentum factor (Mom). These factors are obtained from the Fama and French database.

We applied the proposed model selection criterion, C_p, to select simultaneously the number of groups, the number of group-specific factors and the size of the regularization parameter. The maximum number of group-specific pervasive factors is set to five. The number of groups ranges from one to eight, i.e. $S_{\max} = 8$. Possible candidates for the regularization parameter κ are $\kappa = \{1, 0.1, 0.01, 0.001\}$. As a result, the selected number of groups is $\hat{S} = 6$. We also found that the estimated regression coefficients on the style factors $\{\hat{\beta}_{\text{Mkt}}, \hat{\beta}_{\text{HML}}, \ldots, \hat{\beta}_{\text{Mom}}\}$ are zero. This makes sense because the investment styles (i.e. a sensitivity to the set of investment style factors $(\{\text{Mkt}_t, \text{HML}_t, \text{SMB}_t, \text{LTR}_t, \text{STR}_t, \text{Mom}_t\})$ are different among the set of 536 mutual funds.

To check the slope homogeneity assumption, we used the test developed in Pesaran and Yamagata (2008). They consider a panel data model with fixed effects and heterogeneous slopes $y_{it} = \alpha_i + \boldsymbol{\beta}_i' \boldsymbol{x}_{it} + \varepsilon_{it}$, and proposed a standardized version of Swamy's test of slope homogeneity for panel data models. Here ε_{it} are mutually uncorrelated over i and t. We calculated the $\tilde{\Delta}$ statistic (equation (27) in Pesaran and Yamagata, 2008) that exploits the cross-section dispersion of individual slopes weighted by their relative precision (their $\hat{\Delta}$ statistic provides similar results). Under the null hypothesis $H_0\colon \boldsymbol{\beta} = \boldsymbol{\beta}_i$ for all i, the $\tilde{\Delta}$ statistic asymptotically follows the standard normal distribution. To remove the unobservable dependence structures, the adjusted response $y_{it}^* = y_{it} - \hat{F}_c \hat{\boldsymbol{\lambda}}_{c,i} - \hat{F}_{\hat{g}_i} \hat{\boldsymbol{\lambda}}_{\hat{g}_i}$ is used incalculating the $\tilde{\Delta}$ statistics, where $\hat{F}_c$ and $\hat{F}_{\hat{g}_i}$ are estimated common and group-specific factors, and $\hat{\boldsymbol{\lambda}}_{c,i}$ and $\hat{\boldsymbol{\lambda}}_{\hat{g}_i}$ are the corresponding factor loadings. The computed test statistic is $\tilde{\Delta} = -25.738$, which indicates that the slope homogeneity assumption is strongly rejected. Thus we apply the panel data model with heterogeneous slope coefficients (8) in the next section.

8.1.2. *Results under the Heterogeneous Slope Coefficients*

Because of the rejection of the slope homogeneity assumption, we consider a model with heterogeneous slope coefficients:

$$\begin{aligned} y_{it} &= \beta_{g_i,0} + \text{Mkt}_t \beta_{g_i,\text{Mkt}} + \text{HML}_t \beta_{g_i,\text{HML}} + \text{SMB}_t \beta_{g_i,\text{SMB}} + \text{LTR}_t \beta_{g_i,\text{LTR}} \\ &\quad + \text{STR}_t \beta_{g_i,\text{STR}} + \text{Mom}_t \beta_{g_i,\text{Mom}} + \boldsymbol{f}_{g_i,t}' \boldsymbol{\lambda}_{g_i,i} + \varepsilon_{it} \end{aligned}$$

Similar to Section 8.1.1, we selected the best model among the set of candidate models. Again, the selected number of groups was 6.

A two-way table of estimated groupings against the four mutual fund names (styles) is provided in Table I. The two classification schemes appear to be similar in several respects, although the classification based on the mutual fund names is more parsimonious than our estimated groupings. Memberships overlap considerably for the constructed groups and the classification by names. The

Table I. Scatter matrix of estimated groupings versus classification by mutual fund names (Small Capital & Growth, Large Capital & Growth, Small Capital & Value and Large Capital & Value)

	Our grouping					
Classification by name	G1	G2	G3	G4	G5	G6
Small Capital & Growth	29	8	0	52	1	57
Large Capital & Growth	2	1	5	11	101	2
Small Capital & Value	82	41	1	19	5	0
Large Capital & Value	1	3	104	5	6	0

DOI: 10.1002/jae

Table II. Statistically significant regressors $\boldsymbol{x}_t$ for each group

Variable	G1	G2	G3	G4	G5	G6
Mkt_t	−0.488***	−0.485***	−0.585***	−0.597***	−0.567***	−0.432***
HML_t	0.181***	0.000	0.000	0.242***	0.339***	0.448***
SMB_t	−3.363***	−3.247***	−1.498	−5.179***	−2.323***	−5.10***
LTR_t	−0.485***	−0.400***	−0.348	−0.471***	−0.493***	−0.549***
STR_t	0.151***	0.175***	0.023***	−0.001	0.000	0.017***
Mom_t	−0.047***	−0.087***	0.000	−0.010***	−0.043***	−0.146***

Note: Asterisks indicate that the estimated regression coefficient is statistically significant at the ***1%, **5% and *10% level. The estimated coefficients are multiplied by 100, i.e. $10^2 \times \hat{\boldsymbol{\beta}}_g \ (g = 1, \ldots, 6)$ are provided.

distribution of the funds' estimated memberships is easy to interpret according to mutual fund names. For example, the constructed Group 2 (G2), Group 3 (G3), Group (G5) and Group 6 (G6) correspond to Small Capital & Value, Large Capital & Value, Large Capital & Growth, and Small Capital & Growth, respectively. However, Small Capital & Growth mutual funds are divided into other groups. Group 1 (G1) and Group 4 (G4) contain 111 and 71 Small Capital & Growth mutual funds and the 'Small' factor is the most important characteristic. The comparisons in Table I show the potential of the proposed method. The agreements between the two schemes suggest that our procedure succeeds in recognizing the fundamental differences among funds.

Table II provides statistically significant regressors $\boldsymbol{x}_t$ for each group. In contrast to the estimated coefficients under the slope homogeneity model in equation (1), almost all of the estimated slope coefficients are statistically significant. This result also makes sense. We again apply the slope homogeneity test of Pesaran and Yamagata (2008) to each of the six groups; the procedure is described in Section 8.1.1. Here, we use the set of regressors that are statistically significant. Thus the set of explanatory variables varies over groups. The computed test statistics $\tilde{\Delta}$ are 3.603 (Group 1), 1.508 (Group 2), 3.809 (Group 3), 1.815 (Group 4), 1.498 (Group 5) and 2.795 (Group 6), respectively. These results imply that the null hypothesis is not rejected at the 5% level for Group 2, Group 4 and Group 5. Although the null hypothesis is rejected for the remaining three groups, the corresponding values of $\tilde{\Delta}$ are much smaller than under slope homogeneity (−25.738). In summary, the proposed procedure provides a useful tool for empirical analysis.

8.2. Analysis of China's Mainland Stock Markets

The relative strengths of industry versus exchange-listed effects can be of major importance for equity portfolio managers. If market-listed effects dominate, then primary consideration can be given to the market allocation decision. In contrast, if China's mainland stock market integration is reducing the distinction between the markets, then an industry-first investment process may be more appropriate. There are two stock exchange markets in mainland China: the Shanghai and Shenzhen stock exchanges.

In these markets, two types of shares are traded, namely A-shares and B-shares. Although A-shares and B-shares are listed and traded in the mainland market, the former are denominated in RMB and were originally traded only among Chinese citizens, whereas the latter are denominated in foreign currencies and were originally traded among non-Chinese citizens or among Chinese residing overseas. The Chinese government launched the qualified foreign institutional investors (QFII) policy in 2003 and introduced foreign investors into the domestic A-share market. Although Chinese mainlanders have been eligible to trade B-shares with legal foreign currency accounts since March 2001, the mainlanders may prefer to trade only in A-shares because of the currency barrier. It therefore seems plausible that the underlying asset return structure of A-shares is different from that of B-shares.

This paper investigates empirical questions such as the following: How many groups exist in the stock markets in mainland China? How many group-specific pervasive factors exist in the stock markets in mainland China? What types of observable risk factors explain the stocks in each group? Finally, how can the unobservable factors be understood in terms of observable variables in the economy?

8.2.1. Data

We use monthly excess returns of the Shanghai and Shenzhen stock exchanges from Standard & Poor (S&P)'s Datastream Database. We consider an approximately 8-year sample, covering March 2002 to October 2010, and systematically exclude stocks with missing returns data. We calculate excess returns by subtracting the interest rate on the 1-month interbank offer rate from the individual stock returns. The above filtering procedure yields 1039 A-share firms and 102 B-share firms, listed on the Shanghai stock exchange and the Shenzhen stock exchange respectively.

Numerous studies have analyzed the stock market reaction of developed countries to changes in macroeconomic variables (Fama, 1981; Chen *et al.*, 1986; Fama and French, 1989). Therefore, for the observable risk factors, we use two macroeconomic variables: macroeconomic climate leading index and the money supply. We also use commodity prices because they are a major cost factor for various economic activities in China. Therefore, commodity prices include the prices of industrial metal, aluminum, copper, crude oil, natural gas and nickel. In addition to these, we use the gold and silver prices, which affect the price of alternatives to these financial instruments. Currency movements directly affect the earnings of Chinese firms; thus various exchange rates are used, including the Chinese yuan to the US dollar, the Chinese yuan to Japanese yen, the Chinese yuan to the euro, and the Chinese yuan to the HK dollar exchange rates. Finally, international stock market conditions may affect China's mainland stock markets. Therefore, we use the S&P 500 index, the MSCI World index, the MSCI Europe index, TOPIX, the Hang Seng index, as well as the MSCI China index.

8.2.2. Result

We fit the model (8) by minimizing the objective function. Then, we applied the proposed model selection criterion, C_p, to select simultaneously the number of groups S, the number of group-specific pervasive factors, and the size of the regularization parameter κ. We set the maximum number of groups to $S_{\max} = 20$. The possible number of group-specific pervasive factors r_j range from 0 to 20. Although the maximum number of possible factors is limited to 20, this number may be enough based on the stock market analysis of other countries (see for example, Fama and French, 1993). Possible candidates for the regularization parameter κ are $\kappa = \{10, 1, 0.1, 0.01, 0.001\}$.

The estimated number of groups is $\hat{S} = 6$, which gives the smallest value of the proposed model selection criterion, C_p. This suggests that there are approximately six groups in the Chinese mainland stock markets. Hereafter we denote each of these six groups as G1–G6. As the market/industry classifications are known, a two-way table of the estimated group membership $\hat{g}_i$ against these classifications is provided in Table III. The nominal classification schemes are based on: (i) location of stock exchanges; (ii) types of share (A-share or B-share); and (iii) industry. The estimated group memberships appear to be more related to the A-share/B-share classification rather than to the other two factors. Group G5 is comprised of almost exclusively (approximately 90%) B-shares. Although group G3 also contains A-shares, we suspect that the international investors are also buying the A-shares included in group G3. This indicates that the investors may first consider the types of share (A-share/B-share) rather than the industry or stock exchanges.

The estimated numbers of group-specific factors are: three group-specific factors with respect to groups G3 and G5; two group-specific factors with respect to groups G2, G4 and G5; and one

DOI: 10.1002/jae

Table III. Scatter matrices of the estimated group membership $\hat{g}_i$ against nominal classification schemes based on (i) location of stock exchanges, (ii) types of share and (iii) industry

	Classification	G1	G2	G3	G4	G5	G6
(i)	*Location of stock exchanges*						
	Shanghai stock exchange	179	67	132	77	105	81
	Shenzhen stock exchange	125	29	94	64	95	93
(ii)	*Types of share*						
	A-shares	211	95	224	141	196	172
	B-shares	93	1	2	0	4	2
(iii)	*Category based on industry*						
	Chemicals, Construction, Manufacturing	76	15	70	36	53	49
	Food, Beverages, Personal Goods	40	14	24	21	25	13
	Gas, Metals, Mining, Oil	42	16	16	17	17	26
	Banks, Financial Services, Real Estate	30	6	25	15	23	17
	Retails	29	18	26	19	19	21
	Utilities	17	8	16	6	19	9
	Pharmaceuticals, Health	24	6	21	10	16	12
	Information Technology	27	8	21	9	19	11
	Others	11	4	4	5	7	13

Table IV. Results of regression of group-specific factors $\hat{f}_{jk,t}$ $(j = 1, \dots, S; k = 1, \dots, r_j)$ on observable economic factors z_t

		VIX	ER–A	ER–B	HML	SMB
Group 1	First	0.516	7.872***	−1.275	−2.819	7.518 ***
	SD	0.318	1.454	1.347	1.865	1.543
	Second	0.676**	−13.321***	14.922***	0.449	−1.438
	SD	0.300	1.370	1.269	1.757	1.454
Group 2	First	0.469	10.151***	−4.056***	−2.205	1.694
	SD	0.349	1.596	1.478	2.047	1.694
Group 3	First	0.599*	11.995***	−4.409***	−1.627	4.992***
	SD	0.305	1.394	1.291	1.788	1.480
	Second	0.464	−2.366	−0.618	−2.555	2.597
	SD	0.469	2.145	1.987	2.752	2.277
Group 4	First	0.105	10.20***	−3.737**	−1.960	6.618***
	SD	0.338	1.545	1.431	1.982	1.640
Group 5	First	0.425	11.039***	−4.428***	−3.519*	6.115***
	SD	0.331	1.513	1.402	1.941	1.606
	Second	0.550	0.534	0.139	1.464	−0.134
	SD	0.482	2.201	2.039	2.824	2.337
	Third	0.178	−3.424	−0.547	−5.126*	5.907***
	SD	0.453	2.071	1.918	2.657	2.199
Group 6	First	0.369	9.322***	−2.896**	−3.560*	7.086***
	SD	0.331	1.514	1.403	1.943	1.608
	Second	0.062	−3.076	1.514	−4.188	0.003
	SD	0.476	2.176	2.016	2.792	2.311

Note: The four observable risk factors $\boldsymbol{z}_t$ are market excess returns of A-shares (ER–A), market excess returns of B-shares (ER–B), the book-to-market ratio (HML) and the market capitalization (SMB). These variables are computed with Chinese data. The regression takes the form $\hat{f}_{jk,t} = \boldsymbol{z}_t'\boldsymbol{\gamma}_{jk} + e_{jk,t}$. For each factor, the first row corresponds to the estimated regression coefficients $\hat{\boldsymbol{\gamma}}_G$, whereas the second row is the corresponding standard deviations.

group-specific factor with respect to group G1. Although the group G1 is a mix of A-shares and B-shares, the number of group-specific factors of this group is smaller than that of group G5.

The estimated group-specific factors do not have an immediate economic interpretation. We therefore further explore the economic meanings of the estimated factors in each group. In this paper, we

DOI: 10.1002/jae

regress the estimated group-specific pervasive factors $\hat{f}_{jk,t}$ $(j = 1, \ldots, S; k = 1, \ldots, r_j)$ on some economic factors z_t, $\hat{f}_{jk,t} = z_t'\boldsymbol{\gamma}_{jk} + e_{jk,t}$, and then conduct statistical significance tests of the least squares estimate $\hat{\boldsymbol{\gamma}}_{jk}$.

To make a link between the estimated group-specific pervasive factors, we consider the following four observable market variables: the Chicago Board Options Exchange (CBOE) volatility index, market excess returns of A-shares, market excess returns of B-shares and two factors considered by Fama and French (1993): HML and SMB. We calculated the market excess returns of A-shares by subtracting the interest rate on the 1-month interbank offered rate from the average return of the Shanghai stock exchange A-share price index and the Shenzhen stock exchange A-share price index. The market excess returns of B-shares are calculated in the same way. The HML factor accounts for the spread in returns between value and growth stocks, and thus shows the value premium. SMB measures the historic excess returns of small caps over big caps. These variables are computed using Chinese data.

Table IV summarizes the results. For each factor, the first row corresponds to the estimated regression coefficients, whereas the second row corresponds to the standard deviations. In the table, asterisks indicate that the estimated regression coefficient is statistically significant at the ***1%, **5%, and *10% levels. For the first group-specific factors, the first element of $\boldsymbol{f}_{k,t}$ relates to the market excess returns of A-shares. This is expected because all groups contain many A-shares, and even for group G5 the number of A-shares exceeds the number of B-shares. Furthermore, the size factor SMB also relates to the first group-specific pervasive factor. Contrary to findings for the US market, the book-to-market ratio factor (HML) is weakly related to the estimated factors. As expected, the group-specific factors of group G1 relate strongly to the market excess returns of B-shares as well as A-shares. With respect to VIX, the group-specific factors of group G1 and G3 are weakly related. We suspect that the

Table V. Statistically significant regressors $\boldsymbol{x}_t$ for each group

Variable	G1	G2	G3	G4	G5	G6
China macroeconomic variables						
Macroeconomic Index (leading)	0.975***	0.000	0.160	0.679***	0.936***	0.447***
Money supply − M2	1.022	1.158***	0.370***	0.927***	2.021***	2.110***
Exchange rates						
Chinese yuan to US dollar	0.872***	0.557***	0.284*	1.296***	0.103	0.000
Chinese yuan to yen	0.000	0.000	0.000	0.000	0.018	0.043***
Chinese yuan to euro	0.000	0.000	0.000	0.000	0.000	0.000
Chinese yuan to HK dollar	0.000	0.000	0.000	0.000	0.000	0.000
Commodity price index (spot)						
S&P GSCI Industrial Metals	0.000	0.000	0.000	0.000	0.000	0.000
S&P GSCI Aluminum	0.000	−0.027***	0.000	0.000	0.000	0.000
S&P GSCI Copper	0.000	0.000	0.000	0.000	0.000	0.000
S&P GSCI Crude Oil	0.000	−0.001	0.000	0.000	0.000	0.000
S&P GSCI Gold	0.141	0.152***	0.150***	0.241***	0.073***	0.000
S&P GSCI Natural Gas	−0.007	−0.024***	−0.020***	0.000	−0.021***	0.000
S&P GSCI Nickel	0.000	0.000	0.000	0.008	0.000	0.014***
S&P GSCI Silver	0.000	0.000	0.000	0.000	0.000	0.000
Major stock market indexes						
S&P 500	0.000	0.000	0.000	0.000	−0.092***	0.000
MSCI World	0.000	0.000	0.000	0.000	0.000	0.000
MSCI Europe	0.000	0.000	0.000	0.000	0.000	0.000
TOPIX	0.000	0.000	−0.019	0.000	−0.048***	0.000
Hang Seng	0.000	0.000	0.000	0.000	0.000	0.000
MSCI China	0.291***	0.304***	0.398***	0.242***	0.390***	0.240***

Asterisks indicate that the estimated regression coefficient is statistically significant at the ***1%, **5% and *10% level, respectively.

investors in B-shares are monitoring the volatility index. Overall, we can see some differences among the group-specific pervasive factors.

From Theorem 7, we can implement a statistical significance test for the estimated regression coefficients $\hat{\boldsymbol{\beta}}_k$ $k = 1, \ldots, 6$. Thus we can check whether the regression coefficients $\hat{\boldsymbol{\beta}}_k$ for each security are statistically significant. Table V shows the statistically significant observable risk factors for each group.

Table V presents the following results. First, together with the results of Table IV, market excess returns of A-shares, and size factor SMB exist in each group. This indicates that, although the set of observable risk factors listed in the table may affect the shares in all groups, the major factors are these two extracted factors. Second, groups G2$\sim$G6 are partially explained by the money supply. Furthermore, a leading indicator of the macroeconomic climate index is one of the risk factors for groups G4–G6. Thus Chinese macroeconomic variables are important in explaining asset returns. The exchange rate of the Chinese yuan to the US dollar has a large impact on the excess returns of groups G1–G4. Third, Table V shows that the S&P 500 and TOPIX are important factors for group G5. Although other stock market indexes are not included, this does not indicate that the other markets are irrelevant. This is because these five stock market indexes are highly correlated and thus some of the indexes are sufficient for explaining the fluctuations of individual stock returns.

The empirical results show that the number of unobservable and observable factors varies across groups. Group G5 is subject to a total of 10 factors, including three group-specific pervasive factors and seven observable risk factors. In contrast, group G1 is subject to two group-specific pervasive factors and three observable risk factors.

Finally, we applied the slope homogeneity test developed in Pesaran and Yamagata (2008) to each of the six groups. The same operation described in Section 8.1.2 is employed. The test statistics $\tilde{\Delta}$ are -0.122 (Group 1), -0.187 (Group 2), 0.213 (Group 3), -0.902 (Group 4), -2.204 (Group 5) and 2.021 (Group 6), respectively. It implies that the null hypothesis is not rejected at the 1% level for all groups and thus the results support our grouping procedure.

9. CONCLUSION

The proposed panel data-modeling procedures provide a flexible yet parsimonious approach to capturing unobserved heterogeneity. The regression parameters, unobservable factor structure and group membership were all estimated jointly. The penalized method allows us to select the relevant regressors from a large set. Asymptotic normality and variance selection consistency are established in the presence of unknown group memberships. The Monte Carlo results showed that the proposed procedure performed well. The procedure enriches the toolbox for panel data analysis, which has been widely used by researchers, as is evidenced by the popularity of the monographs and textbooks of Arellano (2003), Baltagi (2008), Hsiao (2003) and Wooldridge (2010).

The proposed procedure is applied to the study of US mutual fund styles. A two-way table of the grouping output against the four mutual fund styles showed that our procedure succeeds in recognizing the fundamental differences among funds. The analysis of the two Chinese mainland stock markets—the Shanghai and Shenzhen stock exchanges—shows that there are approximately six groups in the Chinese mainland stock markets. Using the proposed variable selection procedure, the set of important predictors for each group were determined. We also found that the set of relevant predictors varied over the groups and that the number of group-specific factors and their interpretations vary over the groups.

ACKNOWLEDGEMENTS

The authors thank the editor and three anonymous referees for constructive comments. This paper was presented at the 2014 NBER Summer Institute Workshop on Forecasting and Empirical Methods, and at the 20th International Panel Data Conference, Tokyo. Tomohiro Ando's research is supported by the Japan Securities Scholarship Foundation, Japan; Jushan Bai's research is supported by the National Science Foundation (SES0962410 and SES1357598).

REFERENCES

Akaike H. 1973. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, Petroc B, Csake F (eds). Akademiai Kiado: Budapest; 267–281.

Amengual D, Watson MW. 2007. Consistent estimation of the number of dynamic factors in a large N and T panel. *Journal of Business and Economic Statistics* **25**: 91–96.

Ando T, Bai J. 2014. Asset pricing with a general multifactor structure. *Journal of Financial Econometrics*. DOI: 10.1093/jjŕnec/nbu026.

Arellano M. 2003. *Panel Data Econometrics*. Oxford University Press: Oxford.

Arellano M, Hahn J. 2005. *Understanding bias in nonlinear panel models: some recent developments*, Invited lecture. Econometric Society World Congress: London.

Bai J. 2009. Panel data models with interactive fixed effects. *Econometrica* **77**: 1229–1279.

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

Baltagi BH. 2008. *Econometric Analysis of Panel Data*. Wiley: Chichester.

Bester A, Hansen C. 2012. Grouped effects estimators in fixed effects models. *Journal of Econometrics*. DOI: 10.1016/j.jeconom.2012.08.022.

Bonhomme S, Manresa E. 2012. Grouped patterns of heterogeneity in panel data. Working paper 2012–1208. CEMFI.

Chamberlain G, Rothschild M. 1983. Arbitrage, factor structure and mean–variance analysis in large asset markets. *Econometrica* **51**: 1305–1324.

Chen NF, Roll R, Ross S. 1986. Economic forces and the stock market. *Journal of Business* **59**: 383–403.

Cheng X, Liao Z, Schorfheide F. 2013. Shrinkage estimation of high-dimensional factor models with structural instabilities. NBER working paper.

Connor G, Korajzcyk R. 1986. Performance measurement with the arbitrage pricing theory: a new framework for analysis. *Journal of Financial Economics* **15**: 373–394.

Diebold F, Li C, Yue V. 2008. Global yield curve dynamics and interactions: a dynamic Nelson–Siegel approach. *Journal of Econometrics* **146**: 315–363.

Fama EF. 1981. Stock prices, real activity, inflation and money. *American Economic Review* **71**: 545–565.

Fama EF, French KR. 1989. Business conditions and expected returns on stocks and bonds. *Journal of Financial Economics* **5**: 23–49.

Fama EF, French KR. 1993. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* **33**: 3–56.

Fan J, Li R. 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**: 1348–1361.

Fan J, Peng H. 2004. Nonconcave penalized likelihood with a diverging number of parameters. *Annals of Statistics* **32**(3): 928–961.

Forgy EW. 1965. Cluster analysis of multivariate data: efficiency vs. interpretability of classifications. *Biometrics* **21**: 768–769.

Forni M, Lippi M. 2001. The generalized factor model: representation theory. *Econometric Theory* **17**: 1113–1141.

Forni M, Hallin M, Lippi M, Reichlin L. 2000. The generalized dynamic factor model: identification and estimation. *Review of Economics and Statistics* **82**: 540–554.

Geweke J. 1977. The dynamic factor analysis of economic time series. In *Latent Variables in Socio-economic Models*, Aigner D J, Goldberger A S (eds). North-Holland: Amsterdam; 365–383.

Hahn J, Kuersteiner GM. 2002. Asymptotically unbiased inference for a dynamic panel model with fixed effects when both N and T are large. *Econometrica* **70**: 1639–1657.

Hahn J, Newey WK. 2004. Jackknife and analytical bias reduction for nonlinear panel models. *Econometrica* **72**: 1295–1319.

Hallin M, Liska R. 2007. The generalized dynamic factor model: determining the number of factors. *Journal of the American Statistical Association* **102**: 603–617.

Hsiao C. 2003. *Analysis of Panel Data* (2 edn). Cambridge University Press: Cambridge, UK.

Hsiao C, Tahmiscioglu AK. 1997. A panel analysis of liquidity constraints and firm investment. *Journal of the American Statistical Association* **92**: 455–465.

Kapetanios G, Pesaran MH, Yamagata T. 2011. Panels with non-stationary multifactor error structures. *Journal of Econometrics* **160**: 326–348.

Kose A, Otrok C, Whiteman C. 2008. Understanding the evolution of world business cycles. *International Economic Review* **75**: 110–130.

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.

Lu X, Su L. 2013. Shrinkage estimation of dynamic panel data models with interactive fixed effects. Working paper. Singapore Management University.

Mallows CL. 1973. Some comments on Cp. *Technometrics* **15**: 661–675.

Moench E, Ng S. 2011. A factor analysis of housing market dynamics in the U.S. and the regions. *Econometrics Journal* **14**: 1–24.

Moench E, Ng S, Potter S. 2012. Dynamic hierarchical factor models. *Review of Economics and Statistics* **95**: 1811–1817.

Moon HR, Weidner M. 2009. Likelihood expansion for panel regression models with factors. Working paper. Department of Economics, University of Southern California.

Pesaran MH. 2006. Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica* **74**: 967–1012.

Pesaran MH, Tosetti E. 2011. Large panels with common factors and spatial correlation. *Journal of Econometrics* **161**: 182–202.

Pesaran MH, Yamagata T. 2008. Testing slope homogeneity in large panels. *Journal of Econometrics* **142**: 50–93.

Sargent TJ, Sims CA. 1977. Business cycle modeling without pretending to have too much a priori economic theory. In *New Methods in Business Cycle Research*, Sims C, *et al.* (ed). Federal Reserve Bank of Minneapolis: Minneapolis, MN; 45–109.

Stock JH, Watson MW. 2002. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association* **97**: 1167–1179.

Sun YX. 2005. Estimation and inference in panel structure models. Working paper. Department of Economics, University of, California, San Diego.

Tibshirani R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society* **B58**: 267–288.

Wang P. 2010. Large dimensional factor models with a multi-level factor structure. Working paper. Department of Economics, HKUST.

Wooldridge JM. 2010. *Econometric Analysis of Cross Section and Panel Data* (2 edn). MIT Press: Cambridge, MA.

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

APPENDIX

We first introduce some notations. Let $G^0 = \{g_1^0, \ldots, g_N^0\}$ and $G = \{g_1, \ldots, g_N\}$ denote, respectively, the population grouping and any grouping of the cross-sectional units into S groups. Thus for each i, we have $g_i \in \{1, \ldots, S\}$. Let $\mathcal{G}$ be the collection of all such groupings. That is, $\mathcal{G} = \{(g_1, g_2, \ldots, g_N);\ g_i \in (1, 2, \ldots, S)\}$. Define $\mathcal{F}_{\mathcal{G}} = \left\{(F_{g_1}, \ldots, F_{g_N});\ (g_1, g_2, \ldots, g_N) \in \mathcal{G}, F_j' F_j / T = I_{r_j}, 1 \le j \le S\right\}$. The element of $\mathcal{G}$ is denoted by G and the element of $\mathcal{F}_{\mathcal{G}}$ is denoted by F_G. Each $G = (g_1, \ldots, g_N) \in \mathcal{G}$ is associated with an element $F_G = (F_{g_1}, \ldots, F_{g_N})$ in $\mathcal{F}_{\mathcal{G}}$. The true regression coefficient is denoted by $\boldsymbol{\beta}^0$; $F^0{}_{g_i^0}$ and $\boldsymbol{\lambda}^0_{g_i^0,i}$ are the true factor and factor loading of individual i.

PROOF OF THEOREM 1

Here, we will prove $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|^2 = O_p(T^{-1/4}) + O_p(N^{-1/4})$ and $\frac{1}{T}\|\hat{F}_{\sigma(g)} - F_g^0\|^2 = O_p(T^{-1/8}) + O_p(N^{-1/8})$, where $(\sigma(1), \sigma(2) \ldots, \sigma(S))$ is a permutation of $(1, 2, \ldots, S)$. The result $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|^2 = O_p(T^{-1/4}) + O_p(N^{-1/4})$ will be used in the proof of Lemma A.2.

Let $G = \{g_1, \ldots, g_N\}$ denote an arbitrarily given grouping of the N cross-sectional units ($g_i \in \{1, 2, \ldots, S\}$). Let N_j denote the number of cross-sectional units within the jth group ($j = 1, 2, \ldots, S$) with $N = N_1 + N_2 + \cdots + N_S$.

The estimator $\{\hat{\boldsymbol{\beta}}, \hat{G}, \hat{F}_1, \ldots, \hat{F}_S, \hat{\Lambda}_1, \ldots, \hat{\Lambda}_S\}$ is defined as the minimizer of the penalized squared loss function $L_{NT}(\boldsymbol{\beta}, G, F_1, \ldots, F_S, \Lambda_1, \ldots, \Lambda_S)$ subject to the constraints $F_j'F_j/T = I_{r_j}$ ($j = 1, \ldots, S$), $\Lambda_j'\Lambda_j$ ($j = 1, \ldots, S$) being diagonal. Here $\Lambda_j = (\boldsymbol{\lambda}_{j,1}, \ldots, \boldsymbol{\lambda}_{j,N_j})$ is the $r_j \times N_j$ factor loading matrix ($j = 1, \ldots, S$) for the group-specific factors.

We first show that $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}^0$. Without loss of generality, we assume $\boldsymbol{\beta}^0 = \mathbf{0}$ for notational simplicity and we concentrate out the factor loadings through $\Lambda_j = W_j'F_j(F_j'F_j)^{-1} = W_j'F_j/T$, where $W_j = (\boldsymbol{w}_{j,1}, \ldots, \boldsymbol{w}_{j,N_j})$ such that $\boldsymbol{w}_{j,i} = \boldsymbol{y}_i - X_i\boldsymbol{\beta}$ for $g_i = j$. Note that the set of estimates $\{\hat{\boldsymbol{\beta}}, \hat{G}, \hat{F}_1, \ldots, \hat{F}_S, \hat{\Lambda}_1, \ldots, \hat{\Lambda}_S\}$ that jointly minimizes the objective function $L_{NT}(\boldsymbol{\beta}, G, F_1, \ldots, F_S, \Lambda_1, \ldots, \Lambda_S)$, and the set of estimates $\{\hat{\boldsymbol{\beta}}, \hat{G}, \hat{F}_1, \ldots, \hat{F}_S\}$ that jointly minimizes the following concentrated and centered objective function:

$$U_{NT}(\boldsymbol{\beta}, G, F_1, \ldots, F_S) = \frac{1}{NT}\left[\sum_{i=1}^{N}(\boldsymbol{y}_i - X_i\boldsymbol{\beta})'M_{F_{g_i}}(\boldsymbol{y}_i - X_i\boldsymbol{\beta})\right] + p_{\kappa,\gamma}(|\boldsymbol{\beta}|) - \frac{1}{NT}\sum_{i=1}^{N}\boldsymbol{\varepsilon}_i'M_{F^0{}_{g_i^0}}\boldsymbol{\varepsilon}_i$$

are the same. The term $\frac{1}{NT}\sum_{i=1}^{N}\boldsymbol{\varepsilon}_i'M_{F^0{}_{g_i^0}}\boldsymbol{\varepsilon}_i$ is for the purpose of centering. It does not depend on unknown parameters.

Noting that the true data-generating process is $\boldsymbol{y}_i = F^0_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i} + \boldsymbol{\varepsilon}_i$ ($X_i\boldsymbol{\beta}^0 = \mathbf{0}$), the objective function $U_{NT}(\boldsymbol{\beta}, G, F_1, \ldots, F_S)$ is further expressed as

$$\begin{aligned} U_{NT}(\boldsymbol{\beta}, G, F_1, \ldots, F_S) = {} & \boldsymbol{\beta}'\left(\frac{1}{NT}\sum_{i=1}^{N}X_i'M_{F_{g_i}}X_i\right)\boldsymbol{\beta} + \frac{1}{NT}\sum_{i=1}^{N}\boldsymbol{\lambda}^0_{g_i^0,i}{}'F^{0\prime}{}_{g_i^0}M_{F_{g_i}}F^0{}_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i} \\ & + 2\boldsymbol{\beta}'\left[\frac{1}{NT}\sum_{i=1}^{N}X_i'M_{F_{g_i}}F^0{}_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i}\right] + 2\boldsymbol{\beta}'\left(\frac{1}{NT}\sum_{i=1}^{N}X_i'M_{F_{g_i}}\boldsymbol{\varepsilon}_i\right) \\ & + 2\frac{1}{NT}\sum_{i=1}^{N}\boldsymbol{\lambda}^0_{g_i^0,i}{}'F^{0\prime}{}_{g_i^0}M_{F_{g_i}}\boldsymbol{\varepsilon}_i + \frac{1}{NT}\sum_{i=1}^{N}\boldsymbol{\varepsilon}_i'(P_{F^0_{g_i^0}} - P_{F_{g_i}})\boldsymbol{\varepsilon}_i + p_{\kappa,\gamma}(|\boldsymbol{\beta}|) \end{aligned}$$

Lemma A1 in the online supplement implies that the fourth to sixth terms are bounded by $O_p(T^{-1/4}) + O_p(N^{-1/4})$ (assuming $\boldsymbol{\beta}$ is bounded) uniformly over the parameter space. By choosing κ to be small, we make the last penalty term also this order of magnitude. Thus we have

$$U_{NT}(\boldsymbol{\beta}, G, F_1, \ldots, F_S) = \tilde{U}_{NT}(\boldsymbol{\beta}, G, F_1, \ldots, F_S) + O_p(T^{-1/4}) + O_p(N^{-1/4}) \tag{11}$$

uniformly over the parameter space, where

$$\begin{aligned} \tilde{U}_{NT}(\boldsymbol{\beta}, G, F_1, \ldots, F_S) = {} & \boldsymbol{\beta}'\left(\frac{1}{NT}\sum_{j=1}^{S}\sum_{i;g_i=j}X_i'M_{F_{g_i}}X_i\right)\boldsymbol{\beta} \\ & + \frac{1}{NT}\sum_{j=1}^{S}\sum_{i;g_i=j}\boldsymbol{\lambda}^0_{g_i^0,i}{}'F^{0\prime}{}_{g_i^0}M_{F_{g_i}}F^0{}_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i} \\ & + 2\boldsymbol{\beta}'\left[\frac{1}{NT}\sum_{j=1}^{S}\sum_{i;g_i=j}X_i'M_{F_{g_i}}F^0{}_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i}\right] \end{aligned} \tag{12}$$

We rewrite $\tilde{U}_{NT}$ as $\tilde{U}_{NT}(\boldsymbol{\beta}, G, F_1, \ldots, F_S) = \sum_{j=1}^{S}[\boldsymbol{\beta}' D_j \boldsymbol{\beta} + \boldsymbol{\zeta}'_j E_j \boldsymbol{\zeta}_j + 2\boldsymbol{\beta}' L'_j \boldsymbol{\zeta}_j]$ where D_j, E_j, L_j and $\boldsymbol{\zeta}_j$ are

$$D_j = \frac{1}{NT} \sum_{i;g_i=j} X'_i M_{F_j} X_i, \quad E_j = \text{diag}\{E_{j1}, \ldots, E_{jS}\},$$
$$L_j = (L'_{j1}, \ldots, L'_{jS})' \quad \boldsymbol{\zeta}_j = (\boldsymbol{\zeta}'_{j1}, \ldots, \boldsymbol{\zeta}'_{jS})'$$

with E_{jk}, L_{jk} and $\boldsymbol{\zeta}_{jk}$ $(k = 1, \ldots, S)$ being

$$E_{jk} = \frac{1}{N} \sum_{i;g_i=j,g_i^0=k} \left(\boldsymbol{\lambda}^0_{k,i} \boldsymbol{\lambda}^{0\,\prime}_{k,i}\right) \otimes I_T, \quad \boldsymbol{\zeta}_{jk} = \text{vec}(M_{F_j} F^0_k),$$
$$L_{jk} = \frac{1}{NT} \sum_{i;g_i=j,g_i^0=k} \boldsymbol{\lambda}^0_{k,i} \otimes M_{F_j} X_i$$

Completing the square of $\tilde{U}_{NT}(\boldsymbol{\beta}, G, F_1, \ldots, F_S)$, we have

$$\tilde{U}_{NT}(\boldsymbol{\beta}, G, F_1, \ldots, F_S) = \boldsymbol{\beta}' \left(\sum_{j=1}^{S} D_j - \sum_{j=1}^{S} L'_j E_j^{-1} L_j \right) \boldsymbol{\beta} + \sum_{j=1}^{S} \left(\boldsymbol{\zeta}'_j + \boldsymbol{\beta}' L'_j E_j^{-1}\right) E_j (\boldsymbol{\zeta}_j + E_j^{-1} L_j \boldsymbol{\beta}) \tag{13}$$

By Assumption D, the matrix $\sum_{j=1}^{S} D_j - \sum_{j=1}^{S} L'_j E_j^{-1} L_j$ is positive definite. Also, E_j is semi-positive definite, so $\tilde{U}_{NT}(\boldsymbol{\beta}, G, F_1, \ldots, F_S) \geq 0$ for all $(\boldsymbol{\beta}, G, F_1, \ldots, F_S)$. Further note that $\tilde{U}_{NT}\left(\boldsymbol{\beta}^0, G^0, F^0_1, \ldots, F^0_S\right) = 0$. This can be easily seen from equation (12) by replacing $\boldsymbol{\beta}$ by $\boldsymbol{\beta}^0 = 0$ and $M_{F^0_j} F^0_j = 0$ for $g_i = g^0_i = j$ $(j = 1, 2, \ldots, S)$. Note that we use the notation $\boldsymbol{\beta}^0 = 0$. Otherwise, $\boldsymbol{\beta}$ should be replaced by $\boldsymbol{\beta} - \boldsymbol{\beta}^0$.

Evaluate equation (11) at $\left(\boldsymbol{\beta}^0, G^0, F^0_1, \ldots, F^0_S\right)$, and noting $\tilde{U}_{NT}\left(\boldsymbol{\beta}^0, G^0, F^0_1, \ldots, F^0_S\right) = 0$

$$\begin{aligned} O_p(T^{-1/4}) + O_p(N^{-1/4}) &= U_{NT}\left(\boldsymbol{\beta}^0, G^0, F^0_1, \ldots, F^0_S\right) \\ &\geq U_{NT}\left(\hat{\boldsymbol{\beta}}, \hat{G}, \hat{F}_1, \ldots, \hat{F}_S\right) \\ &= \tilde{U}_{NT}(\hat{\boldsymbol{\beta}}, \hat{G}, \hat{F}_1, \ldots, \hat{F}_S) + O_p(T^{-1/4}) + O_p(N^{-1/4}) \end{aligned}$$

The last equality follows from by evaluating equation (11) at $(\hat{\boldsymbol{\beta}}, \hat{G}, \hat{F}_1, \ldots, \hat{F}_S)$. Combined with $\tilde{U}_{NT}(\hat{\boldsymbol{\beta}}, \hat{G}, \hat{F}_1, \ldots, \hat{F}_S) \geq 0$, it must be

$$\tilde{U}_{NT}(\hat{\boldsymbol{\beta}}, \hat{G}, \hat{F}_1, \ldots, \hat{F}_S) = O_p(T^{-1/4}) + O_p(N^{-1/4}) \tag{14}$$

Because the two terms in $\tilde{U}_{NT}$ (see equation (13)) are both non-negative, each term must be $O_p(T^{-1/4}) + O_p(N^{-1/4})$. Thus (note we used the notation $\boldsymbol{\beta}^0 = 0$)

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|^2 = O_p(T^{-1/4}) + O_p(N^{-1/4}) \tag{15}$$

which implies that $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}^0$. As discussed in Bai (2009), we cannot deduce that $\hat{F}_j$ is consistent for $F^0_j H_j$. This is because the number of elements of F^0_j goes to infinity, so the usual consistency is not well defined. However, because $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\| = O_p(T^{-1/8}) + O_p(N^{-1/8})$, the expressions in equation (12) together with (14) imply that

$$\frac{1}{NT}\sum_{j=1}^{S}\sum_{i;\hat{g}_i=j}\left[\boldsymbol{\lambda}^0_{g_i^0,i}{}'F^{0\prime}_{g_i^0}M_{\hat{F}_j}F^0_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i}\right]=O_p(T^{-1/8})+O_p(N^{-1/8}) \tag{16}$$

We can rewrite equation (16) as the trace of the following matrix:

$$\begin{aligned}
&\left[\frac{1}{T}F_1^{0\prime}M_{\hat{F}_1}F_1^0\right]\left[\frac{1}{N}\sum_{i=1}^{N}1(\hat{g}_i=1)\boldsymbol{\lambda}^0_{1,i}\boldsymbol{\lambda}^{0\prime}_{1,i}\right]+\cdots+\left[\frac{1}{T}F_1^{0\prime}M_{\hat{F}_S}F_1^0\right]\left[\frac{1}{N}\sum_{i=1}^{N}1(\hat{g}_i=S)\boldsymbol{\lambda}^0_{1,i}\boldsymbol{\lambda}^{0\prime}_{1,i}\right]\\
+&\left[\frac{1}{T}F_2^{0\prime}M_{\hat{F}_1}F_2^0\right]\left[\frac{1}{N}\sum_{i=1}^{N}1(\hat{g}_i=1)\boldsymbol{\lambda}^0_{2,i}\boldsymbol{\lambda}^{0\prime}_{2,i}\right]+\cdots+\left[\frac{1}{T}F_2^{0\prime}M_{\hat{F}_S}F_2^0\right]\left[\frac{1}{N}\sum_{i=1}^{N}1(\hat{g}_i=S)\boldsymbol{\lambda}^0_{2,i}\boldsymbol{\lambda}^{0\prime}_{2,i}\right]\\
&\vdots\\
+&\left[\frac{1}{T}F_S^{0\prime}M_{\hat{F}_1}F_S^0\right]\left[\frac{1}{N}\sum_{i=1}^{N}1(\hat{g}_i=1)\boldsymbol{\lambda}^0_{S,i}\boldsymbol{\lambda}^{0\prime}_{S,i}\right]+\cdots+\left[\frac{1}{T}F_S^{0\prime}M_{\hat{F}_1}F_S^0\right]\left[\frac{1}{N}\sum_{i=1}^{N}1(\hat{g}_i=S)\boldsymbol{\lambda}^0_{S,i}\boldsymbol{\lambda}^{0\prime}_{S,i}\right]
\end{aligned}$$

The first line involves distributing the true group 1 individuals over S different estimated groups, the second line involves distributing true group 2 individuals into S estimated groups, and so on. Because the trace of each term is non-negative and the sum of the traces is bounded by $O_p(T^{-1/8})+O_p(N^{-1/8})$, the trace of each term cannot exceed $O_p(T^{-1/8})+O_p(N^{-1/8})$.

For ease of exposition and to be concrete, consider the case of $S=3$. Then the above becomes

$$\begin{aligned}
&\left[\frac{1}{T}F_1^{0\prime}M_{\hat{F}_1}F_1^0\right]A_{11}+\left[\frac{1}{T}F_1^{0\prime}M_{\hat{F}_2}F_1^0\right]A_{12}+\left[\frac{1}{T}F_1^{0\prime}M_{\hat{F}_3}F_1^0\right]A_{13}\\
+&\left[\frac{1}{T}F_2^{0\prime}M_{\hat{F}_1}F_2^0\right]A_{21}+\left[\frac{1}{T}F_2^{0\prime}M_{\hat{F}_2}F_2^0\right]A_{22}+\left[\frac{1}{T}F_2^{0\prime}M_{\hat{F}_3}F_2^0\right]A_{23}\\
+&\left[\frac{1}{T}F_3^{0\prime}M_{\hat{F}_1}F_3^0\right]A_{31}+\left[\frac{1}{T}F_3^{0\prime}M_{\hat{F}_2}F_3^0\right]A_{32}+\left[\frac{1}{T}F_3^{0\prime}M_{\hat{F}_3}F_3^0\right]A_{33}
\end{aligned}$$

where

$$A_{kh}=\frac{1}{N}\sum_{i=1}^{N}1(\hat{g}_i=h)\boldsymbol{\lambda}^0_{k,i}\boldsymbol{\lambda}^{0\prime}_{k,i},\quad h,k=1,2,\ldots,S$$

The earlier argument shows that

$$tr\left(\left[\frac{1}{T}F_k^{0\prime}M_{\hat{F}_h}F_k^0\right]A_{kh}\right)=O_p(T^{-1/8})+O_p(N^{-1/8}),\quad k,h=1,2,\ldots,S$$

Let A denote the matrix $A=(A_{ij})$. In the following discussion, the first row of A refers to A_{1j} (j=1,2,3), and the first column of A refers to A_{j1} (j=1,2,3), etc. Each row sum of the A_{ij} matrices converges to a positive definite matrix by assumption for example, $A_{11}+A_{12}+A_{13}=\frac{1}{N}\Lambda_1^{0\prime}\Lambda_1^0$, where Λ_1^0 is the factor loading matrix associated with true group 1 individuals. Because we require that each estimated group have a positive fraction of individuals, each column sum of these matrices also converges to a positive definite matrix. For example, the first estimated group contains the fraction of individuals $\frac{1}{N}\sum_{i=1}^{N}1(\hat{g}_i=1)\to c_1>0$. This implies

$$A_{11} + A_{21} + A_{31} = \left[\frac{1}{N}\sum_{i=1}^{N} 1(\hat{g}_i = 1)\boldsymbol{\lambda}_{1,i}^{0}\boldsymbol{\lambda}_{1,i}^{0\prime}\right] + \left[\frac{1}{N}\sum_{i=1}^{N} 1(\hat{g}_i = 1)\boldsymbol{\lambda}_{2,i}^{0}\boldsymbol{\lambda}_{2,i}^{0\prime}\right]$$
$$+ \left[\frac{1}{N}\sum_{i=1}^{N} 1(\hat{g}_i = 1)\boldsymbol{\lambda}_{3,i}^{0}\boldsymbol{\lambda}_{3,i}^{0\prime}\right] \to \Psi_1 > 0$$

(note that the limit is not required to exist, but the $\liminf_N$ being positive is sufficient. For notational simplicity, we assume the limit exists). From $A_{11} + A_{21} + A_{31} \to \Psi_1 > 0$, one of the three matrices will have a non-zero limit. Suppose the first matrix A_{11} has a non-zero limit, so that $A_{11} \to A_{11}^0 > 0$, then from $tr\left(\frac{1}{T}F_1^{0\prime}M_{\hat{F}_1}F_1^0 A_{11}\right) = O_p(T^{-1/8}) + O_p(N^{-1/8})$, we must have

$$\frac{1}{T}F_1^{0\prime}M_{\hat{F}_1}F_1^0 = O_p(T^{-1/8}) + O_p(N^{-1/8}) \tag{17}$$

because A_{11} is positive definite. This implies that

$$T^{-1}\|\hat{F}_1 - F_1^0 H_1\|^2 = O_p(T^{-1/8}) + O_p(N^{-1/8}) \tag{18}$$

for some rotation matrix H_1.[1] Once A_{11} is assumed to have a non-zero limit, then the limits of A_{21} and A_{31} must be zero. Otherwise, the same reasoning implies that $\hat{F}_1$ will also be consistent for F_2^0 and F_3^0. This is impossible since a limit is unique.

The preceding argument assumes A_{11} has a non-zero limit. In the case that A_{21} has a non-zero limit, then $\hat{F}_1$ is consistent for F_2^0 (and in this case, A_{11} and A_{31} will have a zero limit because the limit of $\hat{F}_1$ is unique). But this is just a matter of relabeling (a permutation). So without loss of generality, we assume the limit of A_{11} is non-zero, so that the limits of A_{21} and A_{31} are zero.

Next consider the second column of the A matrices. Given that A_{11} has non-zero limit, we argue that either A_{22} or A_{32} has a non-zero limit. We show this by a contradiction argument. If not, suppose that both A_{22} and A_{32} have zero limit. Then A_{23} will have a non-zero limit because the row sum for the second row has a non-zero limit (as argued earlier, each row sum has a positive definite limiting matrix). Similarly, A_{33} will also have a non-zero limit because the row sum for the third row has a non-zero limit (we already know A_{31} and A_{32} have zero limit). This implies that $\frac{1}{T}F_2^{0\prime}M_{\hat{F}_3}F_2^0 = O_p(T^{-1/8}) + O_p(N^{-1/8})$ and $\frac{1}{T}F_3^{0\prime}M_{\hat{F}_3}F_3^0 = O_p(T^{-1/8}) + O_p(N^{-1/8})$. This further implies that $\hat{F}_3$ is consistent for both F_2^0 and F_3^0. This is a contradiction since the limit is unique. So without loss of generality, we assume A_{22} has a non-zero limit. Then we have $\frac{1}{T}F_2^{0\prime}M_{\hat{F}_2}F_2^0 = O_p(T^{-1/8}) + O_p(N^{-1/8})$, or equivalently,

$$\frac{1}{T}\|\hat{F}_2 - F_2^0 H_2\|^2 = O_p(T^{-1/8}) + O_p(N^{-1/8})$$

for some notational matrix H_2. Since each column can only have a single matrix to possess a non-zero limit, this implies that A_{12} and A_{32} have zero limit.

Next consider the third column (or the third row) of the A matrices. Since we already obtain that A_{31} and A_{32} in the third row have zero limit, then A_{33} must have a non-zero limit. This implies that $\frac{1}{T}F_3^{0\prime}M_{\hat{F}_3}F_3^0 = O_p(T^{-1/8}) + O_p(N^{-1/8})$, or

[1] To be exact, equation (17) implies $\|P_{\hat{F}_1} - P_{F_1^0}\|^2 = O_p(T^{-1/8}) + O_p(N^{-1/8})$, where $P_{\hat{F}_1} = I_T - \hat{F}_1(\hat{F}_1'\hat{F}_1)^{-1}\hat{F}_1'$ and $P_{F_1^0}$ is similarly defined (see Bai, 2009, p. 1265). That is, the space spanned by $\hat{F}_1$ and F_1^0 are asymptotically the same. In fact, $\|P_{\hat{F}_1} - P_{F_1^0}\|^2 = O_p(T^{-1/8}) + O_p(N^{-1/8})$ is sufficient for our purpose, and this result is used in the proof of Lemma A2 below. A direct proof of equation (18) requires additional argument.

DOI: 10.1002/jae

$$T^{-1}\|\hat{F}_3 - F_3^0 H_3\|^2 = O_p(T^{-1/8}) + O_p(N^{-1/8})$$

for some H_3. Again, each column can only have a single matrix with a non-zero limit by the uniqueness of a limit so that the limits of A_{13} and A_{23} are zero.

The preceding analysis shows that there is a permutation $\sigma(\cdot)$ of $\{1,2,3\}$ with $\sigma(\{1,2,3\}) = \{\sigma(1), \sigma(2), \sigma(3)\}$ such that for each j we have $\frac{1}{T}\|\hat{F}_{\sigma(j)} - F_j^0 H_j\|^2 = O_p(T^{-1/8}) + O_p(N^{-1/8})$.

Using the same argument, in the general case we can show that for each $j \in \{1,2,\ldots,S\}$ there is a permutation of $\{\sigma(1),\ldots,\sigma(S)\}$ such that

$$\frac{1}{T}\|\hat{F}_{\sigma(j)} - F_j^0 H_j\|^2 = O_p(T^{-1/8}) + O_p(N^{-1/8})$$

This result is similar to that of Bonhomme and Manresa (2012, p. 51). By simple relabeling of the elements of $\sigma(j)$, we take $\sigma(j) = j$ so that

$$\frac{1}{T}\|\hat{F}_j - F_j^0 H_j\|^2 = O_p(T^{-1/8}) + O_p(N^{-1/8}) \quad ,j = 1,2,\ldots,S \tag{19}$$

This proves Theorem 1. The proofs for the remaining theorems are provided in the online supplement.

DOI: 10.1002/jae