Panel data models with grouped factor structure under unknown group membership

Bai, Jushan and Ando, Tomohiro

16 December 2013

Panel data models with grouped factor structure under unknown group membership

December, 2013

Tomohiro Ando [1] and Jushan Bai [2]

Abstract

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. The number of explanatory variables can be large. We estimate the model by minimizing the sum of least squared errors with a shrinkage penalty. 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 under homogeneous regression coefficients, and the China mainland stock market under group-specific regression coefficients.

Keywords: Clustering, penalization, lasso, SCAD, serial and cross-sectional error correlations, factor structure
JEL CODES: C23, C52

[1]The Graduate School of Business, Keio University, andoh@kbs.keio.ac.jp. Financial support from the Japan Securities Scholarship Foundation is acknowledged.

[2]Department of Economics, Columbia University, jb3064@columbia.edu. Financial support from the National Science Foundation (SES0962410) is acknowledged.

1 Introduction

Individual heterogeneity is an important issue in panel data analysis. The degree of heterogeneity increases with larger data sets (more individuals or more time periods). The latter are increasingly available with the advancement in information technology. There are already many studies devoted to large N and large T settings, for example, Arellano and Hahn (2005), Bester and Hansen (2012), Hahn and Kuersteiner (2004), Hahn and Newey (2004), Kapetanios et al. (2011), Moon and Weidner (2009), Pesaran (2006), Pesaran and Tosetti (2011). For panel data textbooks, we refer to Arellano (2003), Baltagi (2008), Hsiao (2003), and Wooldridge (2010).

This paper considers estimation of grouped panel data models with unobserved heterogeneity, which has many attractive features. First, we allow time varying individual effects (factor error structure) as opposed to the usual individual fixed effects. Second, our method allows a large number of explanatory variables. The relevant variables are selected through a lasso approach. Third, the explanatory variables are allowed to be correlated with factors or factor loadings or both. Fourth, the group membership of each unit is unknown, and will be estimated along with other parameters of the model. Finally, the number of groups is unknown and is to be determined. There are a small number of papers that study panel data models with unobserved heterogeneity when group membership is unknown. Bonhomme and Manresa (2012), Lin and Ng (2012) and Sun (2005) investigated this challenging problem. In contrast to previous models, there is a factor structure in each group.

Bai (2009) estimated panel data models with interactive effects, permitting the predictor to be correlated with unobserved heterogeneity. Incorporating this idea, we model time-varying grouped patterns of heterogeneity in panel data by assuming a group-specific pervasive factor structure. Grouped factor structures have been considered in a number of economic studies (Moench et al. (2012), Diebold et al. (2008), Kose et al. (2008), Wang (2010), Moench and Ng (2011)).

We allow the error term to be weakly correlated across units and over time; heteroskedasticity is also allowed in both dimensions. A distinctive feature of the model is that group membership is not specified. Our method jointly estimates the optimal grouping of the N cross-sectional units, the regression coefficients and grouped patterns of heterogeneity. To improve the speed of computation, the lasso method (Tibshirani (1996)) is incorporated in the estimation algorithm. As the lasso method provides estimates of zero for redundant parameters, the computational cost is considerably lower than that of traditional variable selection methods. Although the lasso method is widely used, the shrinkage introduced by the lasso results in bias toward zero for large regression coefficients. To diminish this bias, we use the smoothly clipped absolute deviation (SCAD) penalty approach (Fan and Li (2001)).

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. The proof of parameter consistency with unknown group membership is enormous difficult, we

provide a novel argument for consistency. Given consistency, we further establish that the proposed estimator is asymptotically equivalent to the infeasible version of the estimator in which the population groups are known. This latter result is similar to that of Bonhomme and Manresa (2012), who deal with special known loadings (0 or 1 values). We also develop the asymptotic distribution of the proposed estimator for the regression coefficients. We show that asymptotic bias arises under interactive effects, leading to nonzero-centered limiting distributions. However, the asymptotic bias of the limiting distribution is zero for some cases, including: Case 1: where the error terms are independently, identically distributed, or Case 2: where there is an absence of serial correlation and heteroskedasticity and where $T/N \to 0$ $(N, T \to \infty)$, and Case 3: where there is an absence of cross-sectional correlation and heteroskedasticity and where $N/T \to 0$ $(N, T \to \infty)$. In such cases, there is no need to perform higher-order bias correction.

In panel data modeling, an important issue is the selection of a proper model from among many candidates or, equivalently, determination of the number of group-specific pervasive factors, determination of the magnitude of the regularization parameter for implementing the SCAD approach (to be introduced), and determination of the number of groups. We develop a 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. To prove the consistency of the selection of the number of group-specific pervasive factors, we extend the analysis of Bai (2009). There exist several references concerning model selection of panel data models with factor structures. Ando and Tsay (2013) investigated the model selection problem for large panel data models with the interactive fixed effects of Bai (2009), where the slope coefficients are common to each unit. Ando and Bai (2013) studied the panel data model selection problem under heterogeneous slopes and hierarchical factor error structures. These results are for panel data models where group membership is known. Therefore, our problem is different, as we need to further develop the criterion for selecting the number of groups.

Panel data models with homogeneous regression coefficients between the groups involve parsimonious specifications that may be suitable for some applications. However, there is evidence that homogeneity of the parameters is rejected (see for example Hsiao and Tahmiscioglu (1997), Lin and Ng (2012)). To deal with the presence of unobserved heterogeneity, we therefore extend the proposed model to the flexible yet parsimonious approach. This approach delivers estimates of group-specific regression parameters, together with interpretable estimates of unit-specific time patterns and group membership. After we describe the model estimation procedure, the consistency and asymptotic distribution of the proposed estimator are established. To determine the number of group-specific pervasive factors, the magnitude of the regularization parameter and the number of groups, we again develop a new C_p-type criterion for selecting these quantities. The proposed panel data modeling procedures under ho-

mogeneous regression coefficients are applied to the analysis of the US mutual fund styles. It is common that the 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. The proposed panel data modeling procedures under heterogeneous regression coefficients are applied to the analysis of the two Chinese mainland stock markets, the Shanghai and Shenzhen stock exchanges. We address the following questions. 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 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 describes the model assumptions and Section 3 develops 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 Monte Carlo simulations confirm that the proposed criterion performs well. Applications to US mutual fund data are described in Section 7. Section 8 extends the developed results to the panel data models with heterogeneous regression coefficients. Section 9 applies the procedure to the analysis of Chinese mainland stock markets. Concluding remarks are provided in Section 10.

Notation. Let $\|A\| = [tr(A'A)]^{1/2}$ be the norm of matrix A, where "tr" denotes the trace of a square matrix. The equation $a_n = O(b_n)$ states that the deterministic sequence a_n is at most of order b_n, $c_n = O_p(d_n)$ states that the random variable c_n is at most of order d_n in probability, and $c_n = o_p(d_n)$ is of smaller order in probability. All asymptotic results are obtained under $N, T \to \infty$. Restrictions on the relative rates of convergence of N and T are specified in later sections.

2 Model

Let $t = 1, ..., T$ be an index for time, $i = 1, ..., N$ be an index for units. Let S be the number of groups (which is unknown and fixed), and let $G = \{g_1, ..., g_N\}$ be any grouping of the cross-sectional units into S groups. Therefore, for each i, we have $g_i \in \{1, ..., S\}$. Let N_j be the number of cross-sectional units within the group j, $j = 1, ..., S$ and thus the sum of them will equal the total number of units $N = \sum_{j=1}^{S} N_j$.

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

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

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 pervasive factors that affect the units only in group g_i.

The $p \times 1$ vector $\boldsymbol{\beta}$ is the unknown regression coefficients, $\boldsymbol{\lambda}_{g_i,i}$ is the factor loadings, and ε_{it} is the unit specific error. Our approach is useful in applications where time invariance of the fixed effects is a problematic assumption. Furthermore, the factor structure has been used frequently in recent studies. In Section 8, we extend the model (1) to the heterogeneous regression coefficients, which vary over the groups.

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

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

Depending on the researcher's view, each of the unobserved heterogeneity components may be specified as a dynamic exact 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), also see, 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 below.

Assumption A: Group-specific pervasive factors

The group-specific pervasive factors satisfy $E\|\boldsymbol{f}_{j,t}\|^4 < \infty$ $j = 1, ..., S$. Furthermore,

$$T^{-1}\sum_{t=1}^{T} \boldsymbol{f}_{j,t}\boldsymbol{f}_{j,t}{}' \to \Sigma_{F_j} \quad \text{as} \quad T \to \infty,$$

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 $E\|\boldsymbol{\lambda}_{j,i}^4\| < \infty$ 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, $j = 1, ..., S$. 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 $r(t) \leq \exp(-a_1 t^{b_1})$ 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 (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 $T^{-1}\sum_{t,s=1}^{N}|\eta_{ts}| < C$. 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): $T^{-2}N^{-1}\sum_{t,s,u,v}\sum_{i,j}|\text{cov}(\varepsilon_{is}\varepsilon_{it}, \varepsilon_{js}\varepsilon_{jt})| < C$ and $T^{-1}N^{-2}\sum_{t,s}\sum_{i,j,k,l}|\text{cov}(\varepsilon_{it}\varepsilon_{jt}, \varepsilon_{ks}\varepsilon_{lt})| < C$.

(C5): For all i, ε_{it} is strongly mixing processes with mixing coefficients that satisfy $r(t) \leq \exp(-a_1 t^{b_1})$ 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 $D_j = \frac{1}{NT}\sum_{i;g_i=j} X_i' M_{F_j} X_i$, $E_j = \text{diag}\{E_{j1}, ..., E_{jS}\}$, $L_j = (L_{j1}', ..., L_{jS}')'$, where E_{jk}, and L_{jk} are $E_{jk} = \frac{1}{N}\sum_{i;g_i=j,g_i^0=k}(\boldsymbol{\lambda}_{k,i}^0 \boldsymbol{\lambda}_{k,i}^{0\prime}) \otimes I_T$, $L_{jk} = \sum_{i;g_i=j,g_i^0=k}\frac{1}{NT}\boldsymbol{\lambda}_{k,i}^0 \otimes M_{F_j} X_i$ with g_i^0 denoting the true membership and $\lambda_{k,i}^0$ the true factor loadings. Let $A = \{F_j : F_j'F_j/T = I, j = 1, ..., S\}$. We assume the matrix

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

is positive definite for all $(F_1, ..., F_S) \in A$ 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 is given below.

(D2): The vector of predictor $\boldsymbol{x}_{it}$ satisfies $\max_{1\leq i\leq N} T^{-1}\|X_i\|^2 = O_p(N^{\alpha})$ 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 S_j-th 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, $j = 1, ..., S$. Assumption C imposes

weak serial and cross-sectional correlations on ε_{it}. Heteroskedasticity is allowed. These assumptions are made in Bai (2009) except 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).

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 $\sum_{j=1}^{S}(D_j - Lj'E^{-}L_j)$ 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 (2013), 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 (the factor and membership grouping can be treated as known). 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 the assumption 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. In this paper, the true number of groups, S, is kept fixed. 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 $r_1, ..., r_S$, and size of the penalty κ in $p_{\kappa,\gamma}(|\boldsymbol{\beta}|)$, the estimator $\{\hat{\boldsymbol{\beta}}, \hat{G}, \hat{F}_1, ..., \hat{F}_S, \hat{\Lambda}_1, ..., \hat{\Lambda}_S\}$ is defined as the minimizer of

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

subject to the constraints $F_j'F_j/T = I_{r_j}$ $(j = 1, ..., S)$, $\Lambda_j'\Lambda_j$ $(j = 1, ..., S)$ being diagonal. Here, $\Lambda_j = (\boldsymbol{\lambda}_{j,1}, ..., \boldsymbol{\lambda}_{j,N_j})$ is the $r_j \times N_j$ factor loading matrix $(j = 1, ..., S)$ 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)).

For the penalty function, $p_{\kappa,\gamma}(|\boldsymbol{\beta}|)$ 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, which is formally given as $p_{\kappa,\gamma}(|\boldsymbol{\beta}|) = \sum_{k=1}^{p} p_{\kappa,\gamma}(|\beta_k|)$ with

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

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 criteria for the regression coefficients. We also used the SCAD penalty with $\gamma = 3.7$.

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 $g_i = j$. Then the original model (1) 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 with the penalty is

$$\sum_{j=1}^{S} \mathrm{tr}\left\{ \left(W_j - F_j\Lambda_j'\right)\left(W_j - F_j\Lambda_j'\right)' \right\} + NT \cdot p_{\kappa,\gamma}\left(|\boldsymbol{\beta}|\right).$$

From the analysis of pure factor models estimated by the method of least squares (i.e., principal components; see Connor and Korajzcyk (1986) and Stock and Watson (2002)), by concentrating out $\Lambda_j = W_j'F_j(F_j'F_j)^{-1} = W_j'F_j/T$, the objective function becomes

$$\sum_{j=1}^{S} \mathrm{tr}\left\{W_j'W_j\right\} - \sum_{j=1}^{S} \mathrm{tr}\left\{F_j'W_jW_j'F_j\right\}/T + NT \cdot p_{\kappa,\gamma}\left(|\boldsymbol{\beta}|\right). \tag{2}$$

Noting that only N_j units are related to the factor structure F_j of the j-th group S_j 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, pp197∼198).

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

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

In this paper, the group membership of each unit is estimated through the observed panel data information only. We mention that some prior information can be incorporated by using the Bayesian procedure (not considered in this paper). The estimates of $\boldsymbol{\beta}$, $\{F_j, \Lambda_j;\ j = 1, ..., S\}$, and $G \in \{g_1, ..., g_N\}$ depend on each other. The estimators are obtained by using the following iterative algorithm.

Estimation algorithm

Step 1. Fix κ and $\{r_1, ..., r_S\}$. Initialize the unknown parameters $\boldsymbol{\beta}$, $\{F_j^{(0)}, \Lambda_j^{(0)};\ j = 1, ..., S\}$, $G^{(0)} \in \{g_1^{(0)}, ..., G_N^{(0)}\}$.

Step 2. Given the values of $\boldsymbol{\beta}$ and $\{F_j, \Lambda_j;\ j = 1, ..., S\}$, update G.

Step 3. Given the values of $\boldsymbol{\beta}$ and G, update $\{F_j, \Lambda_j\}$ for $j = 1, ..., S$.

Step 4. Given the values of G and $\{F_j, \Lambda_j;\ j = 1, ..., S\}$, update $\boldsymbol{\beta}$.

Step 5. Repeat Steps 2 and 4 until convergence.

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

3.2 Initial parameter values

First, we refer to the clustering literature in order to achieve fast initialization of group membership G. For this purpose, the well-known K-means algorithm (Forgy (1965)) is used. 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 data set $\{\boldsymbol{y}_i;\ i = 1, ..., N\}$ into S clusters that correspond to the number of groups. Thus an initial estimate of the group membership $G^{(0)} \in \{g_1^{(0)}, ..., g_N^{(0)}\}$ is obtained this way. Second, given the values of $G^{(0)}$, an initial estimate of $\boldsymbol{\beta}^{(0)}$ is obtained by the SCAD approach by ignoring the group-specific factor structures $\{F_j, \Lambda_j;\ j = 1, ..., S\}$. Finally, given the values of $\boldsymbol{\beta}^{(0)}$ and $G^{(0)}$, we obtain the starting values $\{F_j^{(0)}, \Lambda_j^{(0)}\}$ for $j = 1, ..., S$.

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. To maximize the chance of obtaining the global maximum, one may prepare several starting values. After convergence, one may choose the estimators that give a smaller value of the objective function. If the converged values are different, we select the one that minimizes the objective function.

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 $\{F_j^0,\ j = 1, ..., S\}$ 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 $(j = 1, ..., S)$ 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*

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

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

$$T^{-1}\|\hat{F}_j - F_j^0 H_j\|^2 = o_p(1), \quad j = 1, ..., S, \tag{3}$$

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*

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

The estimated individual membership satisfies $\hat{g}_i = \text{argmin}_{j\in\{1,...,S\}}\|\boldsymbol{y}_i - X_i\hat{\boldsymbol{\beta}} - \hat{F}_j\hat{\boldsymbol{\lambda}}_{j,i}\|^2$. The estimates of $\boldsymbol{\beta}$, $\{F_j, \Lambda_j;\ j = 1, ..., S\}$, and $G \in \{g_1, ..., g_N\}$ depend on each other, and we therefore denote the estimator of group membership $\hat{g}_i$ as $\hat{g}_i(\hat{\boldsymbol{\beta}}, \hat{F}, \hat{\Lambda})$ in the following theorem. Here, $\hat{F} = \{\hat{F}_1, ..., \hat{F}_S\}$ and $\hat{\Lambda} = \{\hat{\Lambda}_1, ..., \hat{\Lambda}_S\}$. Although the group indicator is unknown in practice and needs to be estimated, the following theorem shows that the estimated group membership converges to the true group membership as T and N goes to infinity.

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*

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

The result of Theorems 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 $\sup_{i\in\{1,2,...,N\}}$ in Theorem 2 over the set of individuals satisfying $\|\lambda_{g_i^0,i}^0\| \geq a > 0$. Theorem 2 is a very strong result.

Let us define $\tilde{\boldsymbol{\beta}}, \tilde{F}_1, ..., \tilde{F}_S, \tilde{\Lambda}_1, ..., \tilde{\Lambda}_S$ 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 $L_{NT}(\boldsymbol{\beta}, G^0, F_1, ..., F_S, \Lambda_1, ..., \Lambda_S)$ subject to the constraints $F_j'F_j/T = I_{r_j}$ $(j = 1, ..., S)$, and $\Lambda_j'\Lambda_j$ $(j = 1, ..., S)$ being diagonal.

Theorem 2 implies that our estimator $\{\hat{\boldsymbol{\beta}}, \hat{G}, \hat{F}_1, ..., \hat{F}_S, \hat{\Lambda}_1, ..., \hat{\Lambda}_S\}$ is asymptotically equivalent to the infeasible estimates $\{\tilde{\boldsymbol{\beta}}, \tilde{F}_1, ..., \tilde{F}_S, \tilde{\Lambda}_1, ..., \tilde{\Lambda}_S\}$ 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 $\hat{\boldsymbol{\beta}}$, $\hat{F}_j$ $(j = 1, ..., S)$ and the infeasible estimator $\tilde{\boldsymbol{\beta}}$, $\tilde{F}_j$ $(j = 1, ..., S)$ with known population groups are asymptotically equivalent.

Our proposed method can identify the set of explanatory variables with nonzero coefficients. Let $\boldsymbol{\beta}^0 = (\boldsymbol{\beta}_1^{0\prime}, \boldsymbol{\beta}_2^{0\prime})'$ be the true parameter value, and $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^{\prime}, \hat{\boldsymbol{\beta}}_2^{\prime})'$ 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.

Assumption F

Let $X_{i,\beta\neq 0}$ be the submatrix of X_i corresponding to columns of nonzero elements of the parameter vector $\boldsymbol{\beta}^0$, and q be the number of nonzero elements of $\boldsymbol{\beta}$. For the nonrandom positive definite matrix $J_0(F_1^0, ..., F_S^0)$,

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

where $J_0(F_1^0, ..., F_S^0)$ is the probability limit of

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

with

$$Z_{j,i}(F_j^0) = X_{i,\beta\neq 0}' M_{F_j^0} - \frac{1}{N_j} \sum_{k:g_k^0=j} c_{j,ki} X_{k,\beta\neq 0}' M_{F_j^0},$$

where $c_{j,ki} = \boldsymbol{\lambda}_{g_k^0,k}^{0\prime} (\Lambda_j^{0\prime} \Lambda_j^0 / N_j)^{-1} \boldsymbol{\lambda}_{g_i^0,i}^0$.

The notation $J_0(F_1^0, F_2^0, ..., F_S^0)$ does not mean it still depends on $(F_1^0, ..., F_S^0)$, but rather the limit is taken under the true factors. We could have used the notation J_0 in place of $J_0(F_1^0, F_2^0, ..., F_S^0)$. The same comments apply to $D_0(F_1^0, ..., F_S^0)$ (the notation D_0 could be used).

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 $T/N \to \rho > 0$. 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 $V_\beta(F_1^0, ..., F_S^0)$, i.e.,*

$\sqrt{NT}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^0) \to_d N(\boldsymbol{v}_0, V_\beta(F_1^0, ..., F_S^0))$. *Moreover, the following variable selection consistency holds:*

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

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

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

with

$$\boldsymbol{\eta}_j = -\frac{1}{N_j T} \sum_{i:g_i^0=j} \sum_{k:g_k^0=j} (X_i - V_{j,i})' F_j^0 \left(\frac{F_j^{0\prime} F_j^0}{T} \right)^{-1} \left(\frac{\Lambda_j^{0\prime} \Lambda_j^0}{N_j} \right)^{-1} \boldsymbol{\lambda}_{g_k^0,k} \left(\frac{E[\boldsymbol{\varepsilon}_i' \boldsymbol{\varepsilon}_k]}{T} \right) \quad (4)$$

$$\boldsymbol{\zeta}_j = -\frac{1}{N_j T} \sum_{i:g_i^0=j} \sum_{k:g_k^0=j} X_i' M_{F_j^0} \Omega_k F_j^0 \left(\frac{F_j^{0\prime} F_j^0}{T} \right)^{-1} \left(\frac{\Lambda_j^{0\prime} \Lambda_j^0}{N_j} \right)^{-1} \boldsymbol{\lambda}_{g_i^0,i}, \quad (5)$$

$$\hat{D}(F_1^0, ..., F_S^0, \kappa_T) = \frac{1}{NT} \sum_{j=1}^{S} \sum_{i:g_i^0=j} X_{i,\beta\neq 0}' M_{F_j^0} X_{i,\beta\neq 0}$$

$$- \frac{1}{NT} \sum_{j=1}^{S} \frac{1}{N_j} \sum_{i:g_i^0=j} \sum_{k:g_k^0=j} X_{i,\beta\neq 0}' M_{F_j^0} X_{k,\beta\neq 0} c_{j,ki} + \frac{1}{NT} \Sigma(\kappa_T),$$

where $V_{j,i} = N_j^{-1} \sum_{k:g_k^0=j} c_{j,ki} X_k$, $X_{i,\beta\neq 0}$ *is the submatrix* X_i *corresponding to the columns of the nonzero element of* $\boldsymbol{\beta}^0$, $c_{j,ki}$ *is defined in Assumption F, and* $\Sigma(\kappa_T)$ *is defined as*

$$\Sigma(\kappa_T) = \text{diag}\left\{ p'_{\kappa_T,\gamma}(|\beta_{10}|)/|\beta_{10}|, \ldots, p'_{\kappa_T,\gamma}(|\beta_{q0}|)/|\beta_{q0}| \right\},$$

where q *is the number of nonzero elements of* $\boldsymbol{\beta}^0$, *and* $\Omega_k = E[\boldsymbol{\varepsilon}_k \boldsymbol{\varepsilon}_k']$. *The asymptotic covariance matrix* $V_\beta(F_1^0, ..., F_S^0)$ *is given by*

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

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

This indicates that we can perform statistical significance tests. Notice that the bias $\boldsymbol{v}_0$ can be consistently estimated as in Bai (2009), Hahn and Kuersteiner (2002), and Hahn and Newey (2004) 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 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. An informal but frequently used approach is to plot the value of the sum of squared errors for each S, and then try to find the "screen point" at which the objective function starts to flatten. However, the sum of squared errors depends also on the number of group-specific factors, and the size of the regularization parameter κ. Thus, the determination of these quantities is not a straightforward task. In the next section, we propose a new criterion to select these parameters.

5 A new C_p-type criterion for model selection

5.1 Development of a new model selection criterion

Suppose that $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N$ 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

$$\eta(S, k_1, ..., k_S, \kappa) := E_z \left[\frac{1}{NT} \sum_{j=1}^{S} \sum_{i;\hat{g}_i=j}^{N_j} \left|\left| \boldsymbol{z}_i - X_i \hat{\boldsymbol{\beta}} - \hat{F}_{\hat{g}_i} \hat{\boldsymbol{\lambda}}_{\hat{g}_i,i} \right|\right|^2 \right], \tag{6}$$

where $k_1, ..., k_S$ are the 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 $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N$ 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 (6) is the sample-based MSE

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

This quantity is formally calculated by replacing the replicates $\boldsymbol{z}_i$ 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 given by

$$b := E_y \left[\eta(S, k_1, ..., k_S, \kappa) - \hat{\eta}(S, k_1, ..., k_S, \kappa) \right], \tag{7}$$

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. We discuss how to estimate b below. Let $\hat{b}(S, k_1, ..., k_S, \kappa)$ be an estimate of b. Taking

into account the consistency of the proposed model selection criterion, we suggest minimization of the predictive measure

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

The first term on the right-hand side measures the goodness of fit of the model whereas the second term is a penalty that depends on the complexity of the model. It remains to construct a proper estimator of the penalty term. Another contribution of this paper is the following theorem.

Theorem 4 *Under the assumptions of Theorem 3, the penalty term is*

$$\hat{b}(S,k_1,...,k_S,\kappa)=\frac{1}{NT}\text{tr}\left[K_xV_\beta(F_1^0,...,F_S^0,\kappa)\right]+\sum_{j=1}^{S}k_j\times g_j(T,N_1,...,N_S),$$

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 nonvanishing component of the parameter estimate, and* $V_\beta(F_1^0,...,F_S^0,\kappa)=\hat{D}(F_1^0,...,F_S^0,\kappa)^{-1}\hat{J}(F_1^0,...,F_S^0)\hat{D}(F_1^0,...,F_S^0,\kappa)^{-1}$. *Here,* $\hat{J}(F_1^0,...,F_S^0)$ *and* $\hat{D}(F_1^0,...,F_S^0,\kappa)$ *are defined in Assumption F and Theorem 3. The function* $g_j(T,N_1,...,N_S)$ *satisfies (a)* $g_j(T,N_1,...,N_S)\to 0$ *and (b)* $\min\{N,T\}\times g_j(T,N_1,...,N_S)\to\infty$ *as* $T,N\to\infty$. *Under the criterion, the numbers of factors are consistently estimated.*

An example of the function $g_j(T,N_1,...,N_S)$ that satisfies conditions (a) and (b) of the theorem is

$$g_j(T,N_1,...,N_S)=\frac{N_j}{N}\times\frac{T+N_j}{TN_j}\log\left(TN_j\right).$$

Note that $N_j/N=O(1)$ from the assumption E. Substituting $g_j(T,N_1,...,N_S)$ into the criterion function, we have the following criterion

$$\begin{aligned}C_p(k_1,...,k_S,\kappa)\;=\;&\frac{1}{NT}\sum_{j=1}^{S}\sum_{i;\hat{g}_i=j}\|\boldsymbol{y}_i-X_i\hat{\boldsymbol{\beta}}-\hat{F}_{\hat{g}_i}\hat{\boldsymbol{\lambda}}_{\hat{g}_i,i}\|^2\\ &+\frac{1}{TN}\text{tr}\left[K_xV_\beta(\hat{F}_1,...,\hat{F}_S,\kappa)\right]+\sum_{j=1}^{S}k_j\hat{\sigma}^2\frac{N_j}{N}\left(\frac{T+N_j}{TN_j}\right)\log\left(TN_j\right),\quad(8)\end{aligned}$$

where $\hat{\sigma}^2$ is a consistent 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) for selecting panel data models with unobservable interactive effects in a data-rich environment. 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 $r_{j,max}$, $j = 1, ..., S$. Finally, we provide the following theorem.

Theorem 5 *Let $\hat{S}$ be the minimizer of the proposed $C_p(k_1, ..., k_S, \kappa)$ criterion. Under the assumptions in Theorem 4, the determined number of groups, $\hat{S}$, will converge to the true number of groups S_0 as $T, N \to \infty$.*

Thus, the value of S can also be identified as the minimizer of our C_p criterion. The following is a procedure for selecting the value of the regularization parameter κ, the number of factors and the groups S.

5.2 Model selection algorithm

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

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 $\{k_1, ..., k_S\}$.

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

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

6 Simulation study

6.1 Data-generating processes

The first data-generating model considered 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 $\boldsymbol{f}_{j,t}$ $(j = 1, .., S)$ 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 nonzero true parameter values of $\boldsymbol{\beta}$ are set to be $(1, 2, 3)$. These nonzero elements are put into the first three elements of $\boldsymbol{\beta}_i$ and thus the true parameter vector is $\boldsymbol{\beta} = (1, 2, 3, 0, 0, ..., 0)'$. We set the number of groups $S = 3$, and the true numbers of group-specific pervasive factors are $r_1 = 3$ $r_2 = 3$, $r_3 = 3$. Set-

ting the number of units in each group as $N_1 = N_2 = N_3$, we generated a set of T observations. The variables N_j $(j = 1, 2, 3)$ and T take various values.

We next investigate the case in which the noise term is nonhomoscedastic. The second data-generating model considered is $\boldsymbol{y}_i = X_i\boldsymbol{\beta} + F_{g_i}\boldsymbol{\lambda}_{g_i,i} + \boldsymbol{\varepsilon}_i$ and $\varepsilon_{it} = 0.9e^1_{it} + \delta_t 0.9e^2_{it}$, where $\delta_t = 1$ if t is odd and is 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 pervasive factors and the loading matrices, the design matrix X_i and the true parameter vector $\boldsymbol{\beta}$ are generated by the same method as before. The key feature of the model is that the noise terms are not homoscedastic.

As a third example, we investigate the performance of the proposed method when the idiosyncratic errors have some 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.

6.2 Results

We generated 1,000 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 pervasive factors to range from zero to eight. Thus, the maximum number of group-specific pervasive factors was set to eight. The number of groups ranges from two to four. Possible candidates for the regularization parameter κ are $\kappa = \{10, 1, 0.1, 0.01, 0.001\}$.

Table 1 reports the percentage of under-, correct, and overidentified values for the proposed C_p criterion under the three data-generating models. With respect to the number of groups, we can easily calculate the percentages of under- (U), correct (C), and overidentified (O) values. The percentages with respect to the number of group-specific factors are calculated under the condition that the number of groups is correctly selected. This is because of the difficulty of the matching between the true number of group-specific factors and the selected number of group-specific factors when the selected number of groups and true number of groups are different. As shown in the tables, the proposed C_p criterion is capable of selecting the true number of groups as well as the true number of group-specific pervasive factors.

Finally, we discuss the regression coefficient estimation results. The simulation results for the parameter estimates of $\hat{\boldsymbol{\beta}}$ are reported in Table 2. Because the length of $\hat{\boldsymbol{\beta}}$ is very long, we report the estimation results for the true predictors $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)'$, and that for the first three wrong predictors $(\hat{\beta}_4, \hat{\beta}_5, \hat{\beta}_6)$. We point out that the remaining elements of $\hat{\boldsymbol{\beta}}$ (i.e., $\hat{\beta}_7, \hat{\beta}_8, ...,$) are similar to the estimation results of $(\hat{\beta}_4, \hat{\beta}_5, \hat{\beta}_6)$.

As shown in Table 2, the parameters are well estimated in the simulation studies. Furthermore, the accuracy improves as the size of the panel increases.

We point out that the proposed method implicitly assumes that each group should not be completely overlapped. Additional experiments that we have performed suggest that group separation, coming from the observable parts of $X_i\boldsymbol{\beta}$, or coming from the group-specific factor structures $\boldsymbol{f}_{g_i,t}\boldsymbol{\lambda}_{g_i}$, or both, is important. In summary, our simulation results show 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.

7 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, for example, "Value" and "Growth", "Large Cap" and "Small Cap", etc. 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.

7.1 Data and model

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.

The specified model is

$$\begin{aligned} y_{it} &= \beta_0 + \sum_{s=1}^{7} y_{i,t-s}\beta_s + \sum_{w=1}^{7} u_{i,tw}\beta_w + Mkt_t\beta_{Mkt} + HML_t\beta_{HML} + SMB_t\beta_{SMB} \\ &\quad + LTR_t\beta_{LTR} + STR_t\beta_{STR} + Mom_t\beta_{Mom} + \boldsymbol{f}'_{g_i,t}\boldsymbol{\lambda}_{g_i,i} + \varepsilon_{it}, \end{aligned}$$

$i = 1, \ldots, N,\ t = 1, \ldots, T,$, where $u_{i,tw} = \sum_{k=0}^{w} I(y_{i,t-k} > 0)$ is the number of past months such that a positive return is realized, and $I(\cdot)$ is an indicator function that takes the values 1 or 0. Therefore $u_{i,tw}$ is the cumulative sum of the months with positive monthly returns. Fama and French (1993) suggested that an asset return model on a stock index can be constructed using three different weighted averages of the portfolio values: one based on size (SMB), another based on the book-to-market

ratio (HML), and the third based on excess return (Mkt) on the market. 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.

7.2 Results

We 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 pervasive factors to range from zero to eight. Thus, the maximum number of group-specific pervasive factors was set to eight. The number of groups ranges from one to nine, i.e., $S_{\max} = 9$. Possible candidates for the regularization parameter κ are $\kappa = \{10, 1, 0.1, 0.01, 0.001\}$.

As a result, the selected number of groups is $\hat{S} = 6$. A two-way table of the grouping output against the four mutual fund styles is provided in Table 3. The two classification schemes appear to be similar in several respects, although the classification based on the mutual fund names is more parsimonious than in our grouping. Memberships overlap considerably for the constructed groups and the classification by name. The distribution of the funds' memberships is easy to interpret according to mutual fund names. For example, the constructed group 6 (G6) corresponds to Small Capital & Growth. However, Small Capital & Growth mutual funds are divided into other groups. Group 1 contains 64 Small Capital & Growth mutual funds and the "Growth" factor plays a main role. Groups 2 and 4 contain 19 and 14 Small Capital & Growth mutual funds and the "Small" factor is the most important characteristic. Group 3 mostly contains Large Capital & Growth mutual funds. Therefore, both "Large" and "Growth" factors may characterize the fund returns. Group 5 is the group in which "Large" and "Value" factors might be related. The comparisons in Table 3 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.

The selected numbers of group-specific pervasive factors are $r_1 = 4$, $r_2 = 3$, $r_3 = 3$, $r_4 = 3$, $r_5 = 2$, $r_6 = 3$, respectively. Therefore, there are group-specific pervasive factors that explain the mutual fund returns within the groups. The estimated regression coefficients $\{\hat{\beta}_s, \hat{\beta}_w | s, w = 1, ..., 7\}$ are estimated as zero, which partially implies that the return prediction from the historical information is difficult. We found that the estimated regression coefficients on the style factors $\{\hat{\beta}_{Mkt}, \hat{\beta}_{HML}, ..., \hat{\beta}_{Mom}\}$ are also zero. This result makes sense because the investment styles (i.e., a sensitivity to the set of investment style factors $\{Mkt_t, HML_t, SMB_t, LTR_t, STR_t, Mom_t\}$) are different among the set of 536 mutual funds. In Section 8, we introduce the model with heterogeneous regression coefficients that vary over the groups.

Table 4 provides the correlations between the estimated group-specific pervasive factors and the Fama and French (1993) factors (Mkt, HML, SMB), Short-Term Reversal Factor (STR), Long-Term Reversal Factor (LTR), and Momentum Factor (Mom). If

the absolute value of the correlations are larger than 0.18, 0.22, and 0.29, the corresponding significance levels are 10%, 5% and 1%, respectively. From the table, we make the following observations. First, the Mkt factors are mainly related to the first group-specific pervasive factor for all six groups. In particular, the magnitude of the correlation is the highest among those with other styles. Second, the SMB factor is highly related to the first group-specific pervasive factor of group 1, while it has a very low correlation with the factors of other groups. Third, the HML factor has high correlations with the group-specific pervasive factors of groups 1 and 2. Fourth, the momentum factor, short-term return reversal factor, and long-term return reversal factor also play an important role as the high-level correlations show. Fifth, some of the estimated group-specific pervasive factors have low correlations with the six observable investment styles (Mkt, HML, SMB, STR, LTR, and Mom). For example, the second group-specific pervasive factor of group 5 has very low correlation with these variables. It would be interesting to explore some possible investment styles that have a large correlation with such factors. Overall, the estimated group-specific pervasive factors vary over the groups.

8 Heterogeneous group-specific coefficients

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

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta}_{g_i} + \boldsymbol{f}_{g_i,t}'\boldsymbol{\lambda}_{g_i,i} + \varepsilon_{i,t}, \quad i = 1, \ldots, N,\ t = 1, \ldots, T, \tag{9}$$

where the $p_i \times 1$ vector $\boldsymbol{\beta}_{g_i}$ contains the unknown regression coefficients for each group. The regression coefficient is 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.

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 predictor $\boldsymbol{x}_{it}$ 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.

8.1 Estimation procedure

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

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

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

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 $\boldsymbol{\beta}_j$ can be updated. It is easy to see that, for any given values of $\boldsymbol{\beta}_j$ and $F_j\boldsymbol{\lambda}_{j,i}$ $(j = 1, ..., S)$, the optimal assignment for each individual unit is: $g_i^* = \text{argmin}_{j\in\{1,...,S\}} T^{-1}\|\boldsymbol{y}_i - X_i\boldsymbol{\beta}_j - F_j\boldsymbol{\lambda}_{j,i}\|^2 + p_{\kappa,\gamma}(|\boldsymbol{\beta}_j|)$. The estimates of $\boldsymbol{\beta}$, $\{F_j, \Lambda_j;\ j = 1, ..., S\}$, and $G = \{g_1, ..., g_N\}$ depend on each other, the estimators are obtained by almost the same procedures as in Section 2.

8.2 Asymptotic results

Here, we use $\{F_j^0,\ j = 1, ..., S\}$ to denote the true parameter values of the group-specific factors F_j that satisfies Assumptions A, B, C, D' and E. As N and T increase, 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 6 : 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*

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

In addition, $\{\hat{F}_j,\ j = 1, ..., S\}$ 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

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

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

Theorem 7 : Consistency of the estimator of group membership. *Suppose that the assumptions in Theorem 6 hold. Then, for all $\tau > 0$ and $T, N \to \infty$, we have*

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

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

Let us define $\tilde{\boldsymbol{\beta}}_1, ..., \tilde{\boldsymbol{\beta}}_S, \tilde{F}_1, ..., \tilde{F}_S, \tilde{\Lambda}_1, ..., \tilde{\Lambda}_S$ as the infeasible version of our estimator where group membership is fixed to its population G^0. It is defined as the minimizer of $L_{NT}(\boldsymbol{\beta}_1..., \boldsymbol{\beta}_S, G^0, F_1, ..., F_S, \Lambda_1, ..., \Lambda_S)$ subject to the usual constraints. Theorem 7 shows that, under a certain condition, our estimator $\{\hat{\boldsymbol{\beta}}_1, ..., \hat{\boldsymbol{\beta}}_S, \hat{F}_1, ..., \hat{F}_S, \hat{\Lambda}_1, ..., \hat{\Lambda}_S\}$ is asymptotically equivalent to the infeasible estimates $\tilde{\boldsymbol{\beta}}_1, ..., \tilde{\boldsymbol{\beta}}_S, \tilde{F}_1, ..., \tilde{F}_S, \tilde{\Lambda}_1, ..., \tilde{\Lambda}_S$ as N and T tend to infinity. If for some $b > 0$, $N/T^b \to 0$ as both N and T tend to infinity simultaneously, the proposed estimator $\hat{\boldsymbol{\beta}}_j$, $\hat{F}_j$ $(j = 1, ..., S)$ and the infeasible estimator $\tilde{\boldsymbol{\beta}}_j$, $\tilde{F}_j$ $(j = 1, ..., S)$ with known population groups are asymptotically equivalent.

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

Assumption F′

Let $X_{i,\beta_j\neq 0}$ be the submatrix of X_i corresponding to columns of the nonzero elements of the parameter vector $\boldsymbol{\beta}_j$. Let q_j be the number of nonzero elements of $\boldsymbol{\beta}_j$ $(j = 1, ..., S)$. For the nonrandom positive definite matrix $J_0(F_j^0)$,

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

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

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

Then, we have the following theorem.

Theorem 8 : Asymptotic normality and variable selection consistency *Assume that the assumptions in Theorems 6 and 7 and F′ hold. Then, $\sqrt{N_j T}(\hat{\boldsymbol{\beta}}_{j,1} - \boldsymbol{\beta}_{j,1}^0)$ is asymptotically normal with mean $\boldsymbol{v}_j^0$ and variance–covariance matrix $V_\beta(F_j^0)$, i.e.,*

$\sqrt{N_jT}(\hat{\boldsymbol{\beta}}_{j,1} - \boldsymbol{\beta}^0_{j,1}) \to_d N(\boldsymbol{v}^j_0, V_\beta(F^0_j))$. *Moreover, the following variable selection consistency holds:*

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

for $j = 1, ..., S$. *Here, the variance–covariance matrix* $V_{\beta_j}(F^0_j)$ *is*

$$V_\beta(F^0_j) = D_0(F^0_j)^{-1} J_0(F^0_j) D_0(F^0_j)^{-1},$$

where $D_0(F^0_j)$ *is the probability limit of*

$$\hat{D}(F^0_j, \kappa_T) = \frac{1}{N_jT} \sum_{i:g^0_i=j} \left[X_i' M_{F^0_j} X_i - \frac{1}{N_j} \sum_{k:g^0_k=j} c_{j,ki} X_i' M_{F^0_j} X_k \right] + \frac{1}{N_jT} \Sigma_j(\kappa_T),$$

with $\Sigma_j(\kappa_T) = \text{diag}\left\{ p'_{\kappa_T,\gamma}(|\beta_{j,1}|)/|\beta_{j,1}|, \ldots, p'_{\kappa_T,\gamma}(|\beta_{j,q_j}|)/|\beta_{j,q_j}| \right\}$, *where* q_j *is the number of nonzero elements of* $\boldsymbol{\beta}^0_j$, *and* $\boldsymbol{v}^j_0$ *is the probability limit of*

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

where

$$\boldsymbol{\zeta}_j = -\frac{1}{N_jT} \sum_{i:g^0_i=j} \sum_{k:g^0_k=j} X_i' M_{\tilde{F}_j} \Omega_k F^0_j \left(\frac{F^{0\prime}_j F^0_j}{T} \right)^{-1} \left(\frac{\Lambda_j' \Lambda_j}{N_j} \right)^{-1} \boldsymbol{\lambda}_{g^0_i,i},$$

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

with $c_{j,ki} = \boldsymbol{\lambda}^{0\prime}_{g^0_k,k} (\Lambda^{0\prime}_j \Lambda^0_j / N_j)^{-1} \boldsymbol{\lambda}^0_{g^0_i,i}$, *and* $V_{j,i} = N_j^{-1} \sum_{k:g^0_k=j} c_{j,ki} X_k$.

The proof of the theorem is given in the Appendix.

8.3 Determining the number of groups/factors

Taking into account the consistency of the proposed model selection criterion, we again suggest minimization of the predictive measure

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

The first term on the right-hand side measures the goodness of fit of the model whereas the second term is a penalty that depends on the complexity of the model. It remains to construct a proper estimator of the penalty term. We have the following result.

Theorem 9 *Under the assumptions of Theorems 6-8, the penalty term is*

$$\hat{b}(k_1,...,k_S,\kappa) \;=\; \sum_{j=1}^{S}\frac{1}{NT}\mathrm{tr}\left[K_{j,x}V_\beta(F_j^0,\kappa)\right]+\sum_{j=1}^{S}k_j\times g_j(T,N_1,...,N_S),$$

where $K_{j,x}=2\sum_{i;g_i=j}X'_{i,\hat{\beta}_j\neq 0}X_{i,\hat{\beta}_j\neq 0}/(N_jT)$ *with* $X_{i,\hat{\beta}_j\neq 0}$ *being the submatrix of* X_i *such that the corresponding columns contain a nonvanishing component of the parameter estimate, and* $V_\beta(F_j^0,\kappa)=\hat{D}(F_j^0,\kappa)^{-1}\hat{J}(F_j^0)\hat{D}(F_j^0,\kappa)^{-1}$*. Here* $\hat{J}(F_j^0)$ *and* $\hat{D}(F_j^0,\kappa)$ *are defined in Assumption* F' *and Theorem 8. The function* $g_j(T,N_1,...,N_S)$ *satisfies (a)* $g_j(T,N_1,...,N_S)\to 0$ *and (b)* $\min\{N,T\}\times g_j(T,N_1,...,N_S)\to\infty$ *as* $T,N\to\infty$*.*

Using the same investigations in Section 4, we have the following criterion:

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

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 value of S can also be identified as the minimizer of the preceding information criterion.

Theorem 10 *Let* $\hat{S}$ *be the minimizer of the proposed* $C_p(k_1,...,k_S,\kappa)$ *criterion in (10). The determined number of groups,* $\hat{S}$*, converges in probability to the true number of groups* S_0 *as* $T,N\to\infty$*.*

9 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 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. Because of the location of the markets, the underlying asset return structure of the Shanghai stock exchange may be different from that of the Shenzhen stock exchange.

In these markets, two types of shares are traded, namely A- and B-shares. Although A- 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?

9.1 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 eight-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 one-month interbank offer rate from the individual stock returns. The above filtering procedure yields 1,039 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 price and the silver price, which affect the price of alternatives to these financial instruments. Currency movements directly affect the earnings of Chinese firms. In this paper, we use the Chinese yuan to US dollar exchange rate, the Chinese yuan to Japanese yen exchange rate, the Chinese yuan to euro exchange rate, and the Chinese yuan to HK dollar exchange rate. 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.

9.2 Result

We fit the model (9) 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 we set the maximum number of possible factors more than 20, this number may be enough based on the stock market analysis of other countries (see for e.g., 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 $\hat{S} = 6$ because this achieved 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 5. The nominal classification schemes are based on: 1. Location of stock exchanges, 2. Types of share (A-share or B-share), and 3. 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 number of group-specific pervasive factors is: 3 group-specific pervasive factors with respect to groups G3 and G5, 2 group-specific pervasive factors with respect to groups G2, G4 and G5, and 1 group-specific pervasive factor with respect to group G1. Although the group G1 is a mix of A-shares and B-shares, the number of group-specific pervasive factors of this group is smaller than that of group G5.

The estimated group-specific pervasive 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 regress the estimated group-specific pervasive factors $\hat{f}_{jk,t}$ $(j = 1, ..., S; k = 1, ..., r_j)$ on some economic factors $\boldsymbol{z}_t$; $\hat{f}_{jk,t} = \boldsymbol{z}_t'\boldsymbol{\gamma}_{jk} + e_{jk,t}$, 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 one-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 6 summarizes the results. In Table 6, for each factor, the first row corresponds to the estimated regression coefficients, whereas the second row corresponds to the standard deviations. In the table, stars (***), (**) and (*) mean that the estimated regression coefficient is statistically significant at the 1%, 5%, and 10% levels, respectively. We can see from Table 6 that for the first group-specific pervasive factor, 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 we 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 an G3 are weakly related. We suspect that the investors in B-shares are monitoring the volatility index. Overall, we can see some differences among the group-specific pervasive factors.

From Theorem 8, we can implement a statistical significance test for the estimated regression coefficients $\hat{\boldsymbol{\beta}}_k$ $k = 1, ..., 6$. Thus, we can check whether the regression coefficients $\hat{\boldsymbol{\beta}}_k$ for each security are statistically significant. Table 7 shows the statistically significant observable risk factors for each group. In the table, stars (***), (**) and (*) mean that the observable risk factor is statistically significant at the 1%, 5%, and 10% levels, respectively.

Table 7 presents the following results. First, together with the results of Table 6, 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∼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 for explaining asset returns. The exchange rate of the Chinese yuan to the U.S. dollar has a large impact on the excess returns of groups G1∼G4. Third, table 7 shows that the S&P 500 and TOPIX are important factors for the group G5. Although other stock market indexes are not included, this does not indicate that the other markets are completely ignored. 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 ten 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.

10 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 lasso approach allows us to implement the model estimation procedure easily. We provided a novel argument of consistency, which is the most difficult part to obtain. We also proposed a C_p-type model selection criterion. The Monte Carlo results showed that the proposed procedure performed well. The proposed procedure is then applied to the study of US mutual fund style analysis. 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.

We also consider heterogeneous regression coefficients that varies over the groups. Asymptotic normality and the variable selection consistency of the estimated heterogeneous coefficients were again obtained. To determine the number of group-specific factors, the magnitude of the regularization parameter and the number of groups, we again developed a C_p-type model selection criterion for selecting these quantities.

The proposed modeling procedure was then applied to the analysis of the two Chinese mainland stock markets, the Shanghai and Shenzhen stock exchanges. The empirical result showed 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. Moreover, we provided a partial solution to the issue of how to interpret the constructed unobservable factors from an economic perspective. Again, the number of group-specific factors and their interpretations varied over the groups.

Appendix

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

Lemma A1

Under the Assumptions of Theorem 1,

$$\sup_{G\in\mathcal{G},F_G\in\mathcal{F}_\mathcal{G}} \|\frac{1}{NT}\sum_{i=1}^{N} X_i' M_{F_{g_i}} \boldsymbol{\varepsilon}_i\| = O_p(T^{-1/4}) + O_p(N^{-1/4}),$$

$$\sup_{G\in\mathcal{G},F_G\in\mathcal{F}_\mathcal{G}} \|\frac{1}{NT}\sum_{i=1}^{N} \boldsymbol{\lambda}_{g_i^0,i}^{0\prime} F_{g_i^0}^{0\prime} M_{F_{g_i}} \boldsymbol{\varepsilon}_i\| = O_p(T^{-1/4}) + O_p(N^{-1/4}),$$

$$\sup_{G\in\mathcal{G},F_G\in\mathcal{F}_\mathcal{G}} \|\frac{1}{NT}\sum_{i=1}^{N} \boldsymbol{\varepsilon}_i' P_{F_{g_i}} \boldsymbol{\varepsilon}_i\| = O_p(T^{-1/2}) + O_p(N^{-1/2})$$

where $M_{F_{g_i}} = I - F_{g_i}F_{g_i}'/T$ *and* $P_{F_{g_i}} = F_{g_i}F_{g_i}'/T$*;*

Proof of Lemma A1. The proof of Lemma A1 is similar to that of Lemma A1 of Bai (2009), also see Bonhomme and Manresa (2012). Lemma A1 is due to the boundedness of S. If S is allowed to increase, the right hand side of the equations should be multiplied by S. Because S is fixed, it is absorbed into the O_p term. First note that $T^{-1}\|F_{g_i}\|^2 = r_j$ if $g_i = j$. Thus $T^{-1/2}\|F_{g_i}\| \leq \sqrt{r_j} \leq \sqrt{r}$, where $r = \max\{r_1, r_2, ..., r_S\}$. In addition, $T^{-1}\|X_i'F_{g_i}\| \leq r^{1/2}T^{-1/2}\|X_i\| = O_p(1)$.

Consider the first claim in the lemma. From $\frac{1}{NT}\sum_{i=1}^{N} X_i'\boldsymbol{\varepsilon}_i = O_p((NT)^{-1/2})$, it is sufficient to consider $\frac{1}{NT}\sum_{i=1}^{N} X_i'P_{F_{g_i}}\boldsymbol{\varepsilon}_i = \frac{1}{NT^2}\sum_{i=1}^{N} X_i'F_{g_i}F_{g_i}'\boldsymbol{\varepsilon}_i$. Its norm is bounded by

$$\frac{1}{N}\sum_{i=1}^{N} \|T^{-1}X_i'F_{g_i}\| \cdot \|T^{-1}F_{g_i}'\varepsilon_i\| \leq \sqrt{r}\Big(\frac{1}{N}\sum_{i=1}^{N}\|T^{-1/2}X_i\|^2\Big)^{1/2}\Big(\frac{1}{N}\sum_{i=1}^{N}\|\frac{1}{T}F_{g_i}'\varepsilon_i\|^2\Big)^{1/2}$$

Next,

$$\frac{1}{N}\sum_{i=1}^{N}\|\frac{1}{T}F_{g_i}'\varepsilon_i\|^2 = \frac{1}{N}\sum_{i=1}^{N}\|\sum_{j=1}^{S}1(g_i=j)\frac{1}{T}F_j'\varepsilon_i\|^2 \leq \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{S}\|\frac{1}{T}F_j'\varepsilon_i\|^2$$

$$\leq S\sup_F \frac{1}{N}\sum_{i=1}^{N}\|\frac{1}{T}F'\varepsilon_i\|^2$$

where the supremum with respect to F is taken over F such that $F'F/T = I_r$. The latter was shown to be $O_p(N^{-1/2}) + O_p(T^{-1/2})$ by Bai (2009). Taking the squared-root gives the desired result. The proofs for the remaining two claims are similar. □

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 $(\sigma(1), \sigma(2)..., \sigma(S))$ is a permutation of $(1, 2, ..., S)$. 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 $G = \{g_1, ..., g_N\}$ denote an arbitrarily given grouping of the N cross-sectional units ($g_i \in \{1, 2, ..., S\}$). Let N_j denote the number of cross-sectional units within the jth group ($j = 1, 2, ..., S$) with $N = N_1 + N_2 + \cdots + N_S$. The true population grouping is denoted by $G^0 = (g_1^0, ..., g_N^0)$.

The estimator $\{\hat{\boldsymbol{\beta}}, \hat{G}, \hat{F}_1, ..., \hat{F}_S, \hat{\Lambda}_1, ..., \hat{\Lambda}_S\}$ is defined as the minimizer of

$$\begin{aligned} & L_{NT}(\boldsymbol{\beta}, G, F_1, ..., F_S, \Lambda_1, ..., \Lambda_S) \\ = & \sum_{i=1}^{N} \|\boldsymbol{y}_i - X_i\boldsymbol{\beta} - F_{g_i}\boldsymbol{\lambda}_{g_i,i}\|^2 + NT \cdot p_{\kappa,\gamma}(|\boldsymbol{\beta}|) \end{aligned}$$

subject to the constraints $F_j'F_j/T = I_{r_j}$ $(j = 1, ..., S)$, $\Lambda_j'\Lambda_j$ $(j = 1, ..., S)$ being diagonal. Here $\Lambda_j = (\boldsymbol{\lambda}_{j,1}, ..., \boldsymbol{\lambda}_{j,N_j})$ is the $r_j \times N_j$ factor loading matrix $(j = 1, ..., S)$ 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 $W_j = (\boldsymbol{w}_{j,1}, ..., \boldsymbol{w}_{j,N_j})$ such that $\boldsymbol{w}_{j,i} = \boldsymbol{y}_i - X_i\boldsymbol{\beta}$ for $g_i = j$. Note that the set of estimates $\{\hat{\boldsymbol{\beta}}, \hat{G}, \hat{F}_1, ..., \hat{F}_S, \hat{\Lambda}_1, ..., \hat{\Lambda}_S\}$ that jointly minimizes the objective function $L_{NT}(\boldsymbol{\beta}, G, F_1, ..., F_S, \Lambda_1, ..., \Lambda_S)$, and the set of estimates $\{\hat{\boldsymbol{\beta}}, \hat{G}, \hat{F}_1, ..., \hat{F}_S\}$ that jointly minimizes the following concentrated and centered objective function

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

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 $U_{NT}(\boldsymbol{\beta}, G, F_1, ..., F_S)$ is further expressed as

$$\begin{aligned} & U_{NT}(\boldsymbol{\beta}, G, F_1, ..., F_S) \\ = & \boldsymbol{\beta}'\left(\frac{1}{NT}\sum_{i=1}^{N}X_i'M_{F_{g_i}}X_i\right)\boldsymbol{\beta} + \frac{1}{NT}\sum_{i=1}^{N}\boldsymbol{\lambda}^0_{g_i^0,i}{}'F^{0\prime}_{g_i^0}M_{F_{g_i}}F^0_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i} \end{aligned}$$

$$+2\boldsymbol{\beta}'\left[\frac{1}{NT}\sum_{i=1}^{N}X_i'M_{F_{g_i}}F_{g_i^0}^0\boldsymbol{\lambda}_{g_i^0,i}^0\right]+2\boldsymbol{\beta}'\left(\frac{1}{NT}\sum_{i=1}^{N}X_i'M_{F_{g_i}}\boldsymbol{\varepsilon}_i\right)$$
$$+2\frac{1}{NT}\sum_{i=1}^{N}\boldsymbol{\lambda}_{g_i^0,i}^{0\prime}F_{g_i^0}^{0\prime}M_{F_{g_i}}\boldsymbol{\varepsilon}_i+\frac{1}{NT}\sum_{i=1}^{N}\boldsymbol{\varepsilon}_i'(P_{F_{g_i^0}^0}-P_{F_{g_i}})\boldsymbol{\varepsilon}_i+p_{\kappa,\gamma}\left(|\boldsymbol{\beta}|\right).$$

Lemma A1 implies that the fourth to the 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

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

uniformly over the parameter space, where

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

We rewrite $\tilde{U}_{NT}$ as

$$\tilde{U}_{NT}(\boldsymbol{\beta},G,F_1,...,F_S)=\sum_{j=1}^{S}\left[\boldsymbol{\beta}'D_j\boldsymbol{\beta}+\boldsymbol{\zeta}_j'E_j\boldsymbol{\zeta}_j+2\boldsymbol{\beta}'L_j'\boldsymbol{\zeta}_j\right]$$

where D_j, E_j, L_j and $\boldsymbol{\zeta}_j$ are

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

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

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

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

$$\begin{aligned}&\tilde{U}_{NT}(\boldsymbol{\beta},G,F_1,...,F_S)\\ =\;&\frac{1}{N}\left[\boldsymbol{\beta}'\left(\sum_{j=1}^{S}D_j-\sum_{j=1}^{S}L_j'E_j^{-1}L_j\right)\boldsymbol{\beta}+\sum_{j=1}^{S}(\boldsymbol{\zeta}_j'+\boldsymbol{\beta}'L_j'E_j^{-1})E_j(\boldsymbol{\zeta}_j+E_j^{-1}L_j\boldsymbol{\beta})\right].\end{aligned}\tag{13}$$

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 $\tilde{U}_{NT}(\boldsymbol{\beta}, G, F_1, ..., F_S) \geq 0$ for all $(\beta, G, F_1, ..., F_S)$. Further note that

$$\tilde{U}_{NT}(\boldsymbol{\beta}^0, G^0, F_1^0, ..., F_S^0) = 0$$

This can be easily seen from (12) by replacing β by $\beta^0 = 0$ and $M_{F_j^0} F_j^0 = 0$ for $g_i = g_i^0 = j$ $(j = 1, 2, ..., S)$. Note that we use the notation $\beta^0 = 0$. Otherwise, β should be replaced by $\beta - \beta^0$.

Evaluate (11) at $(\boldsymbol{\beta}^0, G^0, F_1^0, ..., F_S^0)$, and noting $\tilde{U}_{NT}(\boldsymbol{\beta}^0, G^0, F_1^0, ..., F_S^0) = 0$,

$$\begin{aligned} O_p(T^{-1/4}) + O_p(N^{-1/4}) &= U_{NT}(\boldsymbol{\beta}^0, G^0, F_1^0, ..., F_S^0) \\ &\geq U_{NT}(\hat{\boldsymbol{\beta}}, \hat{G}, \hat{F}_1, ..., \hat{F}_S) \\ &= \tilde{U}_{NT}(\hat{\boldsymbol{\beta}}, \hat{G}, \hat{F}_1, ..., \hat{F}_S) + O_p(T^{-1/4}) + O_p(N^{-1/4}). \end{aligned}$$

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

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

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

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

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 $F_j^0 H_j$. This is because the number of elements of F_j^0 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 (12) together with (14) imply that

$$\frac{1}{NT} \sum_{j=1}^{S} \sum_{i;\hat{g}_i=j} \left[\boldsymbol{\lambda}_{g_i^0,i}^{0\prime} F_{g_i^0}^{0\prime} M_{\hat{F}_j} F_{g_i^0}^0 \boldsymbol{\lambda}_{g_i^0,i}^0 \right] = O_p(T^{-1/8}) + O_p(N^{-1/8}). \tag{16}$$

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

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

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

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

where

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

The earlier argument shows that

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

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

$$A_{11} + A_{21} + A_{31} =$$

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

(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 $tr(\frac{1}{T}F_1^{0\prime}M_{\hat{F}_1}F_1^0A_{11}) = O_p(T^{-1/8}) + O_p(N^{-1/8})$, 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

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

for some rotation matrix H_1.[3] 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 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 re-labeling (a permutation). So without loss of generality, we assume the limit of A_{11} is nonzero 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 nonzero limit (as argued earlier, each row sum has a positive definite limiting matrix). Similarly, A_{33} will also have a nonzero limit because the row sum for the third row has a nonzero 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 nonzero 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 nonzero 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 nonzero 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

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

for some H_3. Again, each column can only have a single matrix with a nonzero 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 $j \in \{1,2,...,S\}$, there is a permutation of $\{\sigma(1),...,\sigma(S)\}$ such that

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

[3]To be exact, (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, page 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 (18) requires additional argument.

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

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

This proves Theorem 1. □

To proof Theorem 2, we use the following lemma.

Lemma A2

Under the Assumptions of Theorem 1, for all $j \in \{1, 2, ..., S\}$, we have

$$(a) \qquad \max_{i\in\{1,...,N\}} \frac{1}{T}(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}})' X_i' M_{\hat{F}_j} X_i (\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}) = o_p(1),$$

$$(b) \qquad \max_{i\in\{1,...,N\}} \frac{1}{T}(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}})' X_i' M_{\hat{F}_j} F_{g_i^0}^0 \boldsymbol{\lambda}_{g_i^0,i}^0 = o_p(1),$$

$$(c) \qquad \max_{i\in\{1,...,N\}} \frac{1}{T}(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}})' X_i' M_{\hat{F}_j} \boldsymbol{\varepsilon}_i = o_p(1),$$

$$(d) \qquad \max_{i\in\{1,...,N\}} \frac{1}{T}\left(X_i(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}) + F_{g_i^0}^0 \boldsymbol{\lambda}_{g_i^0,i}^0 + \boldsymbol{\varepsilon}_i\right)' (M_{\hat{F}_j} - M_{F_j^0}) \left(X_i(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}) + F_{g_i^0}^0 \boldsymbol{\lambda}_{g_i^0,i}^0 + \boldsymbol{\varepsilon}_i\right) = o_p(1),$$

Proof of Lemma A2. Consider (a). Note that $X_i' M_{\hat{F}_j} X_i \leq X_i' X_i$, thus

$$\begin{aligned} & \max_i \frac{1}{T}(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}})' X_i' M_{\hat{F}_j} X_i (\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}) \\ \leq\ & \max_i \frac{1}{T}(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}})' X_i' X_i (\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}) \\ \leq\ & \|\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}\|^2 \times \left(\frac{1}{T}\max_i \|X_i\|^2\right). \end{aligned}$$

From Assumption (D2), $\max_{1\leq i\leq N} T^{-1}\|X_i\|^2 = O(N^\alpha)$. Together with $\|\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}\|^2 = O_p(T^{-1/4}) + O_p(N^{-1/4})$, we have

$$\max_i \frac{1}{T}(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}})' X_i' M_{\hat{F}_j} X_i (\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}) \leq O_p\left(\frac{N^\alpha}{T^{1/4}}\right) + O_p\left(\frac{1}{N^{1/4-\alpha}}\right).$$

From Assumption D2 on α, both terms are $o_p(1)$. This proves part (a).

Next, consider (b). Similar to the proof of (a), we have

$$\begin{aligned} & \max_i \frac{1}{T}(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}})' X_i' M_{\hat{F}_j} F_{g_i^0}^0 \boldsymbol{\lambda}_{g_i^0,i}^0 \\ \leq\ & \|\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}\| \times \left(\frac{1}{T}\max_i \|X_i F_{g_i^0}^0\|^2\right)^{1/2} \\ \leq\ & \|\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}\| \times \left(\frac{1}{T}\|F_{g_i^0}^0\|^2\right)^{1/2} \times \left(\frac{1}{T}\max_i \|X_i\|^2\right)^{1/2} O_p(1) \end{aligned}$$

$$= \quad O_p\left(\frac{N^{\alpha/2}}{T^{1/8}}\right) + O_p\left(\frac{1}{N^{1/8-\alpha/2}}\right),$$

by Assumption D2, where we assume $\max_i \|\boldsymbol{\lambda}^0_{g_i^0,i}\|^2 \leq C < \infty$ and $T^{-1}\|X_i' F^0_{g_i^0}\| \leq T^{-1}\|X_i\| \cdot \|F^0_{g_i^0}\| = T^{-1/2}\|X_i\| \times O_p(1)$. The two terms in the last line are $o_p(1)$ and thus part (b) is proved. It is easy to relax the assumption $\max_i \|\boldsymbol{\lambda}^0_{g_i^0,i}\|^2 \leq C < \infty$ by allowing the upper bound to be increasing with N, or by considering the product $\max_i(\|X_i\| \cdot \|\boldsymbol{\lambda}^0_{g_i^0,i}\|)$ to be increasing with N.

Part (c) is proved in a similar manner. For part (d), note that $M_{\hat{F}_j} - M_{F_j^0} = P_{F_j^0} - P_{\hat{F}_j}$. We have

$$\begin{aligned}
& \frac{1}{T}\left(X_i(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}) + F^0_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i} + \boldsymbol{\varepsilon}_i\right)'(M_{\hat{F}_j} - M_{F_j^0})\left(X_i(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}) + F^0_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i} + \boldsymbol{\varepsilon}_i\right) \\
= \quad & \frac{1}{T}(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}})' X_i' M_{\hat{F}_j} X_i(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}) + \frac{1}{T}(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}})' X_i' M_{F_j^0} X_i(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}) \\
& + 2\frac{1}{T}\boldsymbol{\lambda}^{0\prime}_{g_i^0,i} F^{0\prime}_{g_i^0} M_{\hat{F}_j} X_i(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}) + 2\frac{1}{T}\boldsymbol{\lambda}^{0\prime}_{g_i^0,i} F^{0\prime}_{g_i^0} M_{F_j^0} X_i(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}) \\
& + 2\frac{1}{T}(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}})' X_i' M_{\hat{F}_j}\boldsymbol{\varepsilon}_i + 2\frac{1}{T}(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}})' X_i' M_{F_j^0}\boldsymbol{\varepsilon}_i \\
& + \frac{1}{T}\boldsymbol{\lambda}^{0\prime}_{g_i^0,i} F^{0\prime}_{g_i^0}(P_{F_j^0} - P_{\hat{F}_j}) F^0_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i} + \\
& + 2\frac{1}{T}\boldsymbol{\lambda}^{0\prime}_{g_i^0,i} F^{0\prime}_{g_i^0}(P_{F_j^0} - P_{\hat{F}_j})\boldsymbol{\varepsilon}_i + \frac{1}{T}\boldsymbol{\varepsilon}_i'(P_{F_j^0} - P_{\hat{F}_j})\boldsymbol{\varepsilon}_i \\
& = I_{1i} + I_{2i} + \cdots + I_{9i}.
\end{aligned}$$

Parts (a)-(c) of Lemma A2 imply that the first six terms are $o_p(1)$ uniformly in i. Using $\|P_{F_j^0} - P_{\hat{F}_j}\| \leq T^{-1/2}\|F_j^0 - \hat{F}_j H_j\| = O_p(T^{-1/16}) + O_p(N^{-1/16})$, we have

$$|\frac{1}{T}\boldsymbol{\lambda}^{0\prime}_{g_i^0,i} F^{0\prime}_{g_i^0}(P_{F_j^0} - P_{\hat{F}_j}) F^0_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i}| \leq \|P_{\hat{F}_j} - P_{F_j^0}\| \cdot \frac{1}{T}\|F^0_{g_i^0}\|^2 \cdot \|\lambda^0_{g_i^0,i}\|^2$$

But $\max_i \frac{1}{T}\|F^0_{g_i^0}\|^2 \leq \max_{j\leq S}\|F_j^0\|^2 = O_p(1)$ and $\max_i \|\lambda^0_{g_i^0,i}\|^2 \leq C < \infty$ by assumption, the 7th term is shown to be $O_p(T^{-1/16}) + O_p(N^{-1/16}) = o_p(1)$. The proof of the last two term being $o_p(1)$ is similar. For example,

$$|\frac{1}{T}\boldsymbol{\varepsilon}_i'(P_{F_j^0} - P_{\hat{F}_j})\boldsymbol{\varepsilon}_i\| \leq \|P_{\hat{F}_j} - P_{F_j^0}\| \cdot \frac{1}{T}\|\boldsymbol{\varepsilon}_i\|^2$$

The assumption of exponential tails on ε_{it} implies that $\frac{1}{T}\|\boldsymbol{\varepsilon}_i\|^2 = \frac{1}{T}|\sum_{t=1}^T \varepsilon_{it}|^2$ is a smaller order than $O_p(T^{1/16}) + O_p(N^{1/16})$ for large T, uniformly in i. Thus its product with $\|P_{F_j^0} - P_{\hat{F}_j}\|$ is $o_p(1)$ uniformly in i. This proves Lemma A2. □

Proof of Theorem 2

Note that $\hat{g}_i$ satisfies

$$\hat{g}_i = \text{argmin}_{j\in\{1,\ldots,S\}}\frac{1}{T}(\boldsymbol{y} - X_i\hat{\boldsymbol{\beta}})' M_{\hat{F}_j}(\boldsymbol{y} - X_i\hat{\boldsymbol{\beta}}).$$

Using $\boldsymbol{y}_i = X_i\boldsymbol{\beta}^0 + F^0_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i} + \boldsymbol{\varepsilon}_i$, we have

$$\begin{aligned}
& \frac{1}{T}(\boldsymbol{y} - X_i\hat{\boldsymbol{\beta}})' M_{\hat{F}_j}(\boldsymbol{y} - X_i\hat{\boldsymbol{\beta}}) \\
= \; & \frac{1}{T}\left(X_i(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}) + F^0_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i} + \boldsymbol{\varepsilon}_i\right)' M_{\hat{F}_j}\left(X_i(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}) + F^0_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i} + \boldsymbol{\varepsilon}_i\right) \\
= \; & \frac{1}{T}\left(X_i(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}) + F^0_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i} + \boldsymbol{\varepsilon}_i\right)' M_{F^0_j}\left(X_i(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}) + F^0_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i} + \boldsymbol{\varepsilon}_i\right) \\
& + \frac{1}{T}\left(X_i(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}) + F^0_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i} + \boldsymbol{\varepsilon}_i\right)' (M_{\hat{F}_j} - M_{F^0_j})\left(X_i(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}) + F^0_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i} + \boldsymbol{\varepsilon}_i\right).
\end{aligned}$$

By Lemma A2, the last expression is $o_p(1)$ uniformly in i. The terms involves $(\hat{\beta} - \beta^0)$ are also $o_p(1)$ uniformly in i, again by Lemma A2. Thus

$$\frac{1}{T}(\boldsymbol{y} - X_i\hat{\boldsymbol{\beta}})' M_{\hat{F}_j}(\boldsymbol{y} - X_i\hat{\boldsymbol{\beta}}) = \frac{1}{T}\left(F^0_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i} + \boldsymbol{\varepsilon}_i\right)' M_{F^0_j}\left(F^0_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i} + \boldsymbol{\varepsilon}_i\right) + o_p(1)$$

Expanding the right hand side, we rewrite the above as

$$\begin{aligned}
& \frac{1}{T}(\boldsymbol{y} - X_i\hat{\boldsymbol{\beta}})' M_{\hat{F}_j}(\boldsymbol{y} - X_i\hat{\boldsymbol{\beta}}) \\
= \; & \begin{cases} \frac{1}{T}{\boldsymbol{\lambda}^0_{g_i^0,i}}' {F^0_{g_i^0}}' M_{F^0_j} F^0_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i} + \frac{2}{T}{\boldsymbol{\lambda}^0_{g_i^0,i}}' {F^0_{g_i^0}}' M_{F^0_j}\boldsymbol{\varepsilon}_i + \frac{1}{T}\boldsymbol{\varepsilon}_i' M_{F^0_j}\boldsymbol{\varepsilon}_i + o_p(1) & (g_i^0 \neq j) \\ \frac{1}{T}\boldsymbol{\varepsilon}_i' M_{F^0_{g_i^0}}\boldsymbol{\varepsilon}_i + o_p(1) & (g_i^0 = j) \end{cases}
\end{aligned}$$

where $o_p(1)$ is uniform in i. We have used the fact that $M_{F^0_j}F^0_{g_i^0} = 0$ if $g_i^0 = j$.

To compare $\frac{1}{T}(\boldsymbol{y} - X_i\hat{\boldsymbol{\beta}})' M_{\hat{F}_j}(\boldsymbol{y} - X_i\hat{\boldsymbol{\beta}})$ for $j \neq g_i^0$ and $j = g_i^0$, define the event A_{ij} such that

$$A_{ij} = \left\{ \frac{1}{T}{\boldsymbol{\lambda}^0_{g_i^0,i}}' {F^0_{g_i^0}}' M_{F^0_j} F^0_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i} + \frac{2}{T}{\boldsymbol{\lambda}^0_{g_i^0,i}}' {F^0_{g_i^0}}' M_{F^0_j}\boldsymbol{\varepsilon}_i + \frac{1}{T}\boldsymbol{\varepsilon}_i' M_{F^0_j}\boldsymbol{\varepsilon}_i < \boldsymbol{\varepsilon}_i' M_{F^0_{g_i^0}}\boldsymbol{\varepsilon}_i + o_p(1) \right\}.$$

Then

$$\mathbf{1}(\hat{g}_i \neq g_i^0) = \sum_{j=1; j\neq g_i^0}^{S} \mathbf{1}(A_{ij}).$$

Now, we can show ,

$$\frac{1}{T}\boldsymbol{\varepsilon}_i' M_{F^0_j}\boldsymbol{\varepsilon}_i - \frac{1}{T}\boldsymbol{\varepsilon}_i' M_{F^0_{g_i^0}}\boldsymbol{\varepsilon}_i = \boldsymbol{\varepsilon}_i'(P_{F^0_{g_i^0}} - P_{F^0_j})\boldsymbol{\varepsilon}_i = o_p(1)$$

where $o_p(1)$ is uniformly over i. This means for any small $\delta > 0$ and $\eta > 0$, under large N and T,

$$\begin{aligned}
& P\left(\max_{i\in\{1,\ldots,N\}} \left|\frac{1}{T}\boldsymbol{\varepsilon}_i' M_{F^0_j}\boldsymbol{\varepsilon}_i - \frac{1}{T}\boldsymbol{\varepsilon}_i' M_{F^0_{g_i^0}}\boldsymbol{\varepsilon}_i\right| > \delta\right) \\
\leq \; & P\left(\max_{i\in\{1,\ldots,N\}} \left|\frac{1}{T}\boldsymbol{\varepsilon}_i' P_{F^0_j}\boldsymbol{\varepsilon}_i\right| > \frac{\delta}{2}\right) + P\left(\max_{i\in\{1,\ldots,N\}} \left|\frac{1}{T}\boldsymbol{\varepsilon}_i' P_{F^0_{g_i^0}}\boldsymbol{\varepsilon}_i\right| > \frac{\delta}{2}\right) < \eta
\end{aligned}$$

In addition, for the $o_p(1)$ term inside A_{ij}, which is uniform in i, $P(|o_p(1)| > \delta) \leq \eta$. Thus,

$$P(A_{ij}) \leq 2\eta + P\left(\frac{2}{T}\boldsymbol{\lambda}^0_{g_i^0,i}{}'F^{0\prime}_{g_i^0}M_{F_j^0}\boldsymbol{\varepsilon}_i < -\frac{1}{T}\boldsymbol{\lambda}^0_{g_i^0,i}{}'F^{0\prime}_{g_i^0}M_{F_j^0}F^0_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i} + 2\delta\right).$$

Suppose $g_i^0 = k$. For $j \neq g_i^0 = k$, the minimum eigenvalue of $\frac{1}{T}F^{0\prime}_{g_i^0}M_{F_j^0}F^0_{g_i^0} = \frac{1}{T}F^{0\prime}_k M_{F_j^0}F^0_k$ is positive. So for individuals with $\|\boldsymbol{\lambda}^0_{g_i^0,i}\|^2 > a > 0$, we have

$$\boldsymbol{\lambda}^0_{g_i^0,i}{}'(\frac{1}{T}F^{0\prime}_{g_i^0}M_{F_j^0}F^0_{g_i^0})\boldsymbol{\lambda}^0_{g_i^0,i} \geq ca > 0$$

for some $c > 0$. Choose δ small enough such that $2\delta < ca/2$, then

$$P\left(\frac{2}{T}\boldsymbol{\lambda}^0_{g_i^0,i}{}'F^{0\prime}_{g_i^0}M_{F_j^0}\boldsymbol{\varepsilon}_i < -\frac{1}{T}\boldsymbol{\lambda}^0_{g_i^0,i}{}'F^{0\prime}_{g_i^0}M_{F_j^0}F^0_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i} + 2\delta\right)$$

$$\leq P\left(\frac{2}{T}\boldsymbol{\lambda}^0_{g_i^0,i}{}'F^{0\prime}_{g_i^0}M_{F_j^0}\boldsymbol{\varepsilon}_i < -ca/2\right) = O(T^{-\tau})$$

for any given $\tau > 0$, for $j \neq g_i^0$. The last equality follows from the assumption of tail probability for $\boldsymbol{\varepsilon}_i$ and the same argument of Bonhomme and Manresa (2012). In summary, we have for $j \neq g_i^0$,

$$P(A_{ij}) \leq 2\eta + O(T^{-\tau}).$$

Since S is finite, this implies that

$$P(\hat{g}_i \neq g_i^0) \leq 2S\eta + O(T^{-\tau}),$$

where the right hand side is uniform in i. It follows that the average over i is also bounded by the above, that is

$$\frac{1}{N}\sum_{i=1}^{N} P(\hat{g}_i \neq g_i^0) = o(1) + O(T^{-\tau}).$$

We next further show that

$$P\left(\sup_{i\in\{1,\ldots,N\}} \mathbf{1}(\hat{g}_i \neq g_i^0) > 0\right) = o(1) + NO(T^{-\tau}).$$

Let us define

$$\begin{aligned}
A^*_{ij} &= \mathbf{1}\left(A_{ij}\right), \\
B_{ij} &= \mathbf{1}\left(\frac{2}{T}\boldsymbol{\lambda}^0_{g_i^0,i}{}'F^{0\prime}_{g_i^0}M_{F_j^0}\boldsymbol{\varepsilon}_i < -\frac{1}{T}\boldsymbol{\lambda}^0_{g_i^0,i}{}'F^{0\prime}_{g_i^0}M_{F_j^0}F^0_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i} + 2\delta\right), \\
C_{ij} &= \mathbf{1}\left(\left|\frac{1}{T}\boldsymbol{\varepsilon}_i'M_{F_j^0}\boldsymbol{\varepsilon}_i - \frac{1}{T}\boldsymbol{\varepsilon}_i'M_{F^0_{g_i^0}}\boldsymbol{\varepsilon}_i\right| + |o_p(1)| > 2\delta\right).
\end{aligned}$$

Then, $A_{ij}^* \leq B_{ij} + C_{ij}$ and thus

$$\sup_{i\in\{1,...,N\}} A_{ij}^* \leq \sup_{i\in\{1,...,N\}} B_{ij} + \sup_{i\in\{1,...,N\}} C_{ij}.$$

Give the assumption of tail probability for $\boldsymbol{\varepsilon}_i$, $\sup_{i\in\{1,...,N\}} C_{ij} = o_p(1)$. Also $0 \leq C_{ij} \leq 1$ is bounded so by the dominated convergence theorem,

$$E\left[\sup_{i\in\{1,...,N\}} C_{ij}\right] = o(1).$$

However, for $j \neq g_i^0$,

$$E\left[\sup_{i\in\{1,...,N\}} B_{ij}\right] \leq NE[B_{ij}] = NO(T^{-\tau}).$$

In summary,

$$P\left(\sup_{i\in\{1,...,N\}, j\neq g_i^0} 1(A_{ij})\right) = o(1) + NO(T^{-\tau}),$$

which implies

$$P\left(\sup_{i\in\{1,...,N\}} \mathbf{1}(\hat{g}_i \neq g_i^0) > 0\right) = o(1) + NO(T^{-\tau}).$$

This completed the proof of Theorem 2. □

Let $\tilde{\boldsymbol{\beta}}, \tilde{F}_1, ..., \tilde{F}_S, \tilde{\Lambda}_1, ..., \tilde{\Lambda}_S$ be the infeasible version of our estimator where group membership is fixed to its population G^0. It is defined as the minimizer of the objective function $L_{NT}(\boldsymbol{\beta}, G^0, F_1, ..., F_S, \Lambda_1, ..., \Lambda_S)$ subject to the constraints $F_j'F_j/T = I_{r_j}$ $(j = 1, ..., S)$, $\Lambda_j'\Lambda_j$ $(j = 1, ..., S)$ being diagonal. Because the group membership is known, a special case prior proof shows consistency of $\tilde{\boldsymbol{\beta}}, \tilde{F}_1, ..., \tilde{F}_S$. For completeness and for useful notation introduced, we state additional intermediate result.

Lemma A3

Under Assumptions A–E, $\kappa \to 0$ and $\min\{N, T\} \times \kappa \to \infty$ as $T, N \to \infty$, the infeasible estimator $\tilde{\boldsymbol{\beta}}$ is consistent $\tilde{\boldsymbol{\beta}} \to_p \boldsymbol{\beta}^0$. Also,

$$T^{-1/2}\|\tilde{F}_j - F_j^0 H_j\| = o_p(1), j = 1, ..., S,$$

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

$$\left[\frac{1}{N_jT}\sum_{i;g_i^0=j}^{N_j}(\boldsymbol{y}_i - X_i\tilde{\boldsymbol{\beta}})(\boldsymbol{y}_i - X_i\tilde{\boldsymbol{\beta}})'\right]\tilde{F}_j = \tilde{F}_j V_{j,N_jT}.$$

Proof of Lemma A3. If G^0 is known, the proof is similar to that of Bai (2009). Without loss of generality, we here assume that $\boldsymbol{\beta}^0 = \mathbf{0}$ (for notational simplicity). We also concentrate out the factor loadings as we can express them as $\Lambda_j = W_j' F_j (F_j' F_j)^{-1} = W_j' F_j / T$ where $W_j = (\boldsymbol{w}_{j,1}, ..., \boldsymbol{w}_{j,N_j})$ such that $\boldsymbol{w}_{j,i} = \boldsymbol{y}_i - X_i\boldsymbol{\beta}$ and $g_i^0 = j$. Noting that the true data generating process is $\boldsymbol{y}_i = F^0_{g_i^0} \boldsymbol{\lambda}^0_{g_i^0,i} + \boldsymbol{\varepsilon}_i$ ($X_i\boldsymbol{\beta}^0 = \mathbf{0}$), $\{\tilde{\boldsymbol{\beta}}, \tilde{F}_1, ..., \tilde{F}_S\}$ is also expressed as the minimizer of

$$
\begin{aligned}
& \frac{1}{NT} S_{NT}(\boldsymbol{\beta}, F_1, ..., F_S) \\
= \; & \frac{1}{NT} \left[\sum_{j=1}^{S} \sum_{i; g_i^0 = j} (\boldsymbol{y}_i - X_i\boldsymbol{\beta})' M_{F_{g_i^0}} (\boldsymbol{y}_i - X_i\boldsymbol{\beta}) \right] + p_{\kappa,\gamma}(|\boldsymbol{\beta}|) - \frac{1}{NT} \sum_{j=1}^{S} \sum_{i; g_i^0 = j} \boldsymbol{\varepsilon}_i' M_{F^0_{g_i^0}} \boldsymbol{\varepsilon}_i \\
= \; & \boldsymbol{\beta}' \left(\frac{1}{NT} \sum_{j=1}^{S} \sum_{i; g_i^0 = j} X_i' M_{F_{g_i^0}} X_i \right) \boldsymbol{\beta} + \sum_{j=1}^{S} \operatorname{tr} \left\{ \left(\frac{F_j^{0\prime} M_{F_j} F_j^0}{T} \right) \left(\frac{\Lambda_j^{0\prime} \Lambda_j^0}{N} \right) \right\} \\
& + 2\boldsymbol{\beta}' \left[\frac{1}{NT} \sum_{j=1}^{S} \sum_{i; g_i^0 = j} X_i' M_{F_{g_i^0}} F^0_{g_i^0} \boldsymbol{\lambda}^0_{g_i^0,i} \right] + o_p(1) \\
= \; & \boldsymbol{\beta}' \sum_{j=1}^{S} A_j \boldsymbol{\beta} + \sum_{j=1}^{S} \boldsymbol{\eta}_j' B_j \boldsymbol{\eta}_j + 2\boldsymbol{\beta}' \sum_{j=1}^{S} C_j' \boldsymbol{\eta}_j + o_p(1) \\
= \; & \boldsymbol{\beta}' \left(\sum_{j=1}^{S} A_j - \sum_{j=1}^{S} C_j' B_j^{-1} C_j \right) \boldsymbol{\beta} + \sum_{j=1}^{S} (\boldsymbol{\eta}_j' + \boldsymbol{\beta}' C_j' B_j^{-1}) B_j (\boldsymbol{\eta}_j + B_j^{-1} C_j \boldsymbol{\beta}) + o_p(1) \\
= \; & \frac{1}{NT} \tilde{S}_{NT}(\boldsymbol{\beta}, F_1, ..., F_S) + o_p(1),
\end{aligned}
$$

where the $o_p(1)$ term follows from Lemma A.1 and A_j, B_j, C_j and η_j are defined as

$$
A_j = \frac{1}{NT} \sum_{i; g_i^0 = j} X_i' M_{F_{g_i^0}} X_i, \quad B_j = \frac{\Lambda_j^{0\prime} \Lambda_j^0}{N} \otimes I_T
$$

$$
C_j = \frac{1}{NT} \sum_{i; g_i^0 = j} \boldsymbol{\lambda}_{g_i^0,i} \otimes M_{F_{g_i^0}} X_i, \quad \boldsymbol{\eta}_j = \operatorname{vec}(M_{F_j} F_j^0).
$$

Similar to the discussion in Bai (2009), the leading term in the last equation achieves unique minimum at $\boldsymbol{\beta}_0$ (where $\boldsymbol{\beta}_0 = 0$) and $\{F_1^0, ..., F_S^0\}$. Clearly, $\tilde{S}_{NT}(\boldsymbol{\beta}, F_1, ..., F_S) = 0$ at $\boldsymbol{\beta}_0$ and $\{F_1^0, ..., F_S^0\}$. On the other hand, for $\|\boldsymbol{\beta}\| > 0$, $\tilde{S}_{NT}(\boldsymbol{\beta}, F_1, ..., F_S) > 0$. This implies that $\tilde{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}^0$, i.e., $\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\| = o_p(1)$. Using Proposition A.1 of Bai (2009), we can also show that $T^{-1}\|F_j^0 H_j - \tilde{F}_j\|^2 = o_p(1)$, for $j = 1, ..., S$. The remaining claim also follows from Bai (2009). This completes the proof. □

Proof of Theorem 3

By Theorem 2, $P(\sup_i |\hat{g}_i - g_i^0| > 0) = o(1)$ when $N/T^\tau \to 0$. This implies that $P(\hat{g}_1 = g_1^0, \hat{g}_2 = g_2^0, ..., \hat{g}_N = g_N^0) \to 1$. Thus to prove Theorem 3, it is sufficient to

assume that the group membership is known. We first investigate the convergence rate for the estimated factors $\tilde{F}_j$ under the true group membership. We use the following facts. $T^{-1}\|X_i\|^2 = T^{-1}\sum_{t=1}^{T}\|\boldsymbol{x}_{it}\|^2 = O_p(1)$, or $T^{-1/2}\|X_i\| = O_p(1)$. Averaging over i, $(TN)^{-1}\sum_{i=1}^{N}\|X_i\|^2 = O_p(1)$. Similarly, $T^{-1/2}\|F_j\| = O_p(1)$, $T^{-1}\|X_i'F_j\| = O_p(1)$, and so forth.

Using

$$\left[\frac{1}{N_jT}\sum_{i;g_i^0=j}(\boldsymbol{y}_i - X_i\tilde{\boldsymbol{\beta}})(\boldsymbol{y}_i - X_i\tilde{\boldsymbol{\beta}})'\right]\tilde{F}_j = \tilde{F}_jV_{j,NT}$$

and $\boldsymbol{y}_i = X_i\boldsymbol{\beta}^0 + F^0_{g_i^0,i}\boldsymbol{\lambda}^0_{g_i^0,i} + \boldsymbol{\varepsilon}_i$, we have

$$\begin{aligned}
\tilde{F}_jV_{j,NT} &= \frac{1}{N_jT}\sum_{i;g_i^0=j}X_i(\boldsymbol{\beta}^0-\tilde{\boldsymbol{\beta}})(\boldsymbol{\beta}^0-\tilde{\boldsymbol{\beta}})'X_i'\tilde{F}_j + \frac{1}{N_jT}\sum_{i;g_i^0=j}X_i(\boldsymbol{\beta}^0-\tilde{\boldsymbol{\beta}})\boldsymbol{\lambda}^0_{g_i^0,i}F_j^{0\prime}\tilde{F}_j \\
&\quad + \frac{1}{N_jT}\sum_{i;g_i^0=j}X_i(\boldsymbol{\beta}^0-\tilde{\boldsymbol{\beta}})\boldsymbol{\varepsilon}_i'\tilde{F}_j + \frac{1}{N_jT}\sum_{i;g_i^0=j}F_j^0\boldsymbol{\lambda}^0_{j,i}(\boldsymbol{\beta}^0-\tilde{\boldsymbol{\beta}})'X_i'\tilde{F}_j \\
&\quad + \frac{1}{N_jT}\sum_{i;g_i^0=j}\boldsymbol{\varepsilon}_i(\boldsymbol{\beta}^0-\tilde{\boldsymbol{\beta}})'X_i'\tilde{F}_j + \frac{1}{N_jT}\sum_{i;g_i^0=j}F_j^0\boldsymbol{\lambda}^0_{j,i}\boldsymbol{\varepsilon}_i'\tilde{F}_j + \frac{1}{N_jT}\sum_{i;g_i^0=j}\boldsymbol{\varepsilon}_i\boldsymbol{\lambda}^{0\prime}_{j,i}F_j^{0\prime}\tilde{F}_j \\
&\quad + \frac{1}{N_jT}\sum_{i;g_i^0=j}\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i'\tilde{F}_j + \frac{1}{NT}\sum_{i=1}^{N}F_j^0\boldsymbol{\lambda}^0_{g_i^0,i}\boldsymbol{\lambda}^{0\prime}_{g_i^0,i}F_j^{0\prime}\tilde{F}_j \\
&= I_1^j + \cdots + I_9^j.
\end{aligned}$$

Multiplying $(F_j^{0\prime}\tilde{F}_j/T)^{-1}(\Lambda_j^{0\prime}\Lambda_j^0/N_j)^{-1}$ on each side of the prior formula, and then using the results of Bai (2009, Equation (43)) and Assumption (E) we have

$$\begin{aligned}
&T^{-1/2}\|\tilde{F}_jV_{j,NT}(F_j^{0\prime}\tilde{F}_j/T)^{-1}(\Lambda_j^{0\prime}\Lambda_j^0/N_j)^{-1} - F_j^0\| \\
&= O_p(\|\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\|) + O_p(1/\min\{\sqrt{N},\sqrt{T}\}),
\end{aligned}$$

which implies

$$T^{-1/2}\|\tilde{F}_j - F_j^0H_j\| = O_p\left(\|\boldsymbol{\beta}^0-\tilde{\boldsymbol{\beta}}\|\right) + O_p\left(\frac{1}{\min\{N^{1/2},T^{1/2}\}}\right),$$

where $H_j^{-1} = V_{j,NT}(F_j^0\tilde{F}_j/T)^{-1}(\Lambda_j^{0\prime}\Lambda_j^0/N_j)^{-1}$. Here we used the property that $V_{j,NT}$ invertible (See Bai (2009)).

The proof for the variable selection consistency $P(\hat{\boldsymbol{\beta}}_{20} = \mathbf{0}) \to 1$ as $N,T\to\infty$, is provided later. We next prove the asymptotic normality of $\hat{\boldsymbol{\beta}}_{10}$. For notational simplicity, we denote the non-zero true coefficient $\boldsymbol{\beta}_1^0$ as $\boldsymbol{\beta}^0$, and denote X_i as the corresponding columns of design matrix.

We again consider the estimator where group membership is fixed to its population G^0, in view of Theorem 2. We denote $\tilde{\boldsymbol{\beta}}$ as the parameter estimate of the non-zero

element of the true parameter $\boldsymbol{\beta}^0$ and the corresponding sub-matrix $X_{i,\beta\neq 0}$ of X_i as X_i. An alternative expression for the solution of the non-zero component of regression coefficients of $\boldsymbol{\beta}^0$ is

$$\tilde{\boldsymbol{\beta}} = \left(\sum_{j=1}^{S} \sum_{i;g_i^0=j} X_i M_{\tilde{F}_j} X_i + \Sigma(\kappa) \right)^{-1} \sum_{j=1}^{S} \sum_{i;g_i^0=j} X_i M_{\tilde{F}_j} \boldsymbol{y}_i,$$

where $M_{\tilde{F}} = I_T - \tilde{F}\tilde{F}'/T$, and $\Sigma(\kappa) = \text{diag}\{p'_{\kappa,\gamma}(|\tilde{\beta}_1|)/|\tilde{\beta}_1|, ..., p'_{\kappa,\gamma}(|\tilde{\beta}_q|)/|\tilde{\beta}_q|\}$.

Noting that $\boldsymbol{y}_i = X_i\boldsymbol{\beta}^0 + F^0_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i} + \boldsymbol{\varepsilon}_i$, we have

$$\begin{aligned}
&\frac{1}{NT}\left(\sum_{j=1}^{S} \sum_{i;g_i^0=j} X_i M_{\tilde{F}_j} X_i + \Sigma(\kappa) \right)(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) + \frac{1}{NT}\Sigma(\kappa)\boldsymbol{\beta}^0 \\
&= \frac{1}{NT}\sum_{j=1}^{S} \sum_{i;g_i^0=j} X_i M_{\tilde{F}_j} F_j^0 \boldsymbol{\lambda}_{j,i}^0 + \frac{1}{NT}\sum_{j=1}^{S} \sum_{i;g_i^0=j} X_i M_{\tilde{F}_j} \boldsymbol{\varepsilon}_i.
\end{aligned}$$

Using $\tilde{F}_j V_{j,NT} = I_1^j + \cdots + I_9^j$, we have

$$\frac{1}{NT}\sum_{j=1}^{S} \sum_{i;g_i^0=j} X_i M_{\tilde{F}_j} F_j^0 \boldsymbol{\lambda}_{j,i}^0 = -\frac{1}{NT}\sum_{j=1}^{S} \sum_{i;g_i^0=j} X_i' M_{\tilde{F}_j} \left[\sum_{k=1}^{8} I_k^j \right] (F_j^{0\prime}\tilde{F}_j/T)^{-1}(\Lambda_j^{0\prime}\Lambda_j^0/N_j)^{-1}\boldsymbol{\lambda}_{j,i}^0,$$

where we used $M_{\tilde{F}_j}\tilde{F}_j H_j^{-1} = 0$ and I_k^j is defined earlier. Each of the components in the right hand side of equation above are evaluated. For term involving I_1^j,

$$\begin{aligned}
&\frac{1}{NT}\sum_{j=1}^{S} \sum_{i;g_i^0=j} X_i' M_{\tilde{F}_j} I_1^j (F_j^{0\prime}\tilde{F}_j/T)^{-1}(\Lambda_j^{0\prime}\Lambda_j^0/N_j)^{-1}\boldsymbol{\lambda}_{j,i}^0 \\
&= \frac{1}{NT}\sum_{j=1}^{S} \sum_{i;g_i^0=j} X_i' M_{\tilde{F}_j} \left[\frac{1}{N_j T} \sum_{k:g_k^0=j} X_k(\boldsymbol{\beta}^0 - \tilde{\boldsymbol{\beta}})(\boldsymbol{\beta}^0 - \tilde{\boldsymbol{\beta}})' X_k' \tilde{F}_j \right] (F_j^{0\prime}\tilde{F}_j/T)^{-1}(\Lambda_j^{0\prime}\Lambda_j^0/N_j)^{-1}\boldsymbol{\lambda}_{j,i}^0 \\
&= o_p(1) \times (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0).
\end{aligned}$$

Next, we have

$$\begin{aligned}
&\frac{1}{NT}\sum_{j=1}^{S} \sum_{i;g_i^0=j} X_i' M_{\tilde{F}_j} I_2^j (F_j^{0\prime}\tilde{F}_j/T)^{-1}(\Lambda_j^{0\prime}\Lambda_j^0/N_j)^{-1}\boldsymbol{\lambda}_{j,i}^0 \\
&= \frac{1}{NT}\sum_{j=1}^{S} \sum_{i;g_i^0=j} X_i' M_{\tilde{F}_j} \left[\frac{1}{N_j T} \sum_{k:g_k^0=j} X_k(\boldsymbol{\beta}^0 - \tilde{\boldsymbol{\beta}})\boldsymbol{\lambda}_{j,k}^{0\prime}(\Lambda_j^{0\prime}\Lambda_j^0/N_j)^{-1} \right] \boldsymbol{\lambda}_{j,i}^0 \\
&= \frac{1}{T}\sum_{j=1}^{S} \frac{N_j}{N} \left[\frac{1}{N_j}\frac{1}{N_j} \sum_{i;g_i^0=j} \sum_{k:g_k^0=j} X_i' M_{\tilde{F}_j} X_k \boldsymbol{\lambda}_{j,k}^{0\prime}(\Lambda_j^{0\prime}\Lambda_j^0/N_j)^{-1}\boldsymbol{\lambda}_{j,i}^0 \right] (\boldsymbol{\beta}^0 - \tilde{\boldsymbol{\beta}}).
\end{aligned}$$

The third term can be evaluated as

$$\begin{aligned}
&\frac{1}{NT}\sum_{j=1}^{S}\sum_{i:g_i^0=j} X_i' M_{\tilde{F}_j} I_3^j (F_j^{0\prime}\tilde{F}_j/T)^{-1}(\Lambda_j^{0\prime}\Lambda_j^0/N_j)^{-1}\boldsymbol{\lambda}_{j,i}^0 \\
= &\frac{1}{NT}\sum_{j=1}^{S}\sum_{i:g_i^0=j} X_i' M_{\tilde{F}_G,\tilde{F}_j}\left[\frac{1}{N_jT}\sum_{k:g_k^0=j} X_k(\boldsymbol{\beta}^0-\tilde{\boldsymbol{\beta}})\boldsymbol{\varepsilon}_k'\tilde{F}_j\right](F_j^{0\prime}\tilde{F}_j/T)^{-1}(\Lambda_j^{0\prime}\Lambda_j^0/N_j)^{-1}\boldsymbol{\lambda}_{j,i}^0 \\
= &\ o_p(1)\times(\boldsymbol{\beta}^0-\tilde{\boldsymbol{\beta}}).
\end{aligned}$$

The next two terms are also

$$\begin{aligned}
&\frac{1}{NT}\sum_{j=1}^{S}\sum_{i:g_i^0=j} X_i' M_{\tilde{F}_j} I_4^j (F_j^{0\prime}\tilde{F}_j/T)^{-1}(\Lambda_j^{0\prime}\Lambda_j^0/N_j)^{-1}\boldsymbol{\lambda}_{j,i}^0 = o_p(1)\times(\boldsymbol{\beta}^0-\tilde{\boldsymbol{\beta}}), \\
&\frac{1}{NT}\sum_{j=1}^{S}\sum_{i:g_i^0=j} X_i' M_{\tilde{F}_j} I_5^j (F_j^{0\prime}\tilde{F}_j/T)^{-1}(\Lambda_j^{0\prime}\Lambda_j^0/N_j)^{-1}\boldsymbol{\lambda}_{j,i}^0 = o_p(1)\times(\boldsymbol{\beta}^0-\tilde{\boldsymbol{\beta}}).
\end{aligned}$$

Next, using the result of Bai (2009), we have

$$\begin{aligned}
&\frac{1}{NT}\sum_{j=1}^{S}\sum_{i:g_i^0=j} X_i' M_{\tilde{F}_j} I_6^j (F_j^{0\prime}\tilde{F}_j/T)^{-1}(\Lambda_j^{0\prime}\Lambda_j^0/N_j)^{-1}\boldsymbol{\lambda}_{j,i}^0 \\
= &\frac{1}{NT}\sum_{j=1}^{S}\sum_{i:g_i^0=j} X_i' M_{\tilde{F}_j}\left[\frac{1}{N_jT}\sum_{k:g_k^0=j}(F_j^0-\tilde{F}_jH_j^{-1})\boldsymbol{\lambda}_{j,k}^0\boldsymbol{\varepsilon}_k'\tilde{F}_j\right](F_j^{0\prime}\tilde{F}_j/T)^{-1}(\Lambda_j^{0\prime}\Lambda_j^0/N_j)^{-1}\boldsymbol{\lambda}_{j,i}^0.
\end{aligned}$$

Using

$$\begin{aligned}
\frac{1}{N_jT}\sum_{k:g_k^0=j}\boldsymbol{\lambda}_{j,k}^0\boldsymbol{\varepsilon}_k'\tilde{F}_j &= \frac{1}{N_jT}\sum_{k:g_k^0=j}\boldsymbol{\lambda}_{j,k}^0\boldsymbol{\varepsilon}_k'F_j^0H_j + \frac{1}{N_jT}\sum_{k:g_k^0=j}\boldsymbol{\lambda}_{j,k}^0\boldsymbol{\varepsilon}_k'(\tilde{F}_j-F_j^0) \\
&= O_p\left(\frac{1}{\sqrt{N_jT}}\right)+O_p\left(\frac{1}{N}\right)+N^{-1/2}O_p\left(\frac{1}{\min\{N,T\}}\right)
\end{aligned}$$

and

$$\begin{aligned}
&\frac{1}{N_jT}\sum_{k:g_k^0=j} X_i' M_{\tilde{F}_j}(\tilde{F}_j-F_j^0H_j)(F_j^{0\prime}\tilde{F}_j/T)^{-1}(\Lambda_j^{0\prime}\Lambda_j^0/N_j)^{-1}\boldsymbol{\lambda}_{j,i}^0 \\
&= O_p\left(\boldsymbol{\beta}^0-\tilde{\boldsymbol{\beta}}\right)+O_p\left(\frac{1}{\min\{N,T\}}\right)
\end{aligned}$$

which can be derived from Lemma A3 and Lemma A4 of Bai (2009). We have

$$\frac{1}{NT}\sum_{j=1}^{S}\sum_{i:g_i^0=j} X_i' M_{\tilde{F}_j} I_6^j (F_j^{0\prime}\tilde{F}_j/T)^{-1}(\Lambda_j^{0\prime}\Lambda_j^0/N_j)^{-1}\boldsymbol{\lambda}_{j,i}^0$$

$$= o_p(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) + o_p\left(\frac{1}{\sqrt{NT}}\right) + \frac{1}{N}O_p\left(\frac{1}{\min\{N,T\}}\right) + \frac{1}{N^{1/2}}O_p\left(\frac{1}{\min\{N^2,T^2\}}\right).$$

Next, we have

$$\begin{aligned}
&\frac{1}{NT}\sum_{j=1}^{S}\sum_{i:g_i^0=j} X_i' M_{\tilde{F}_j} I_7^j (F_j^{0\prime}\tilde{F}_j/T)^{-1}(\Lambda_j^{0\prime}\Lambda_j^0/N_j)^{-1}\boldsymbol{\lambda}_{j,i}^0 \\
&= \frac{1}{NT}\sum_{j=1}^{S}\sum_{i:g_i^0=j} X_i' M_{\tilde{F}_G,\tilde{F}_j}\left[\frac{1}{N_jT}\sum_{k:g_k^0=j}\boldsymbol{\varepsilon}_k\boldsymbol{\lambda}_{j,k}^{0\prime}F_j^{0\prime}\tilde{F}_j\right](F_j^{0\prime}\tilde{F}_j/T)^{-1}(\Lambda_j^{0\prime}\Lambda_j^0/N_j)^{-1}\boldsymbol{\lambda}_{j,i}^0 \\
&= \sum_{j=1}^{S}\frac{N_j}{N}\times\frac{1}{N_j^2T}\sum_{i:g_i^0=j}\sum_{k:g_k^0=j}\boldsymbol{\lambda}_{j,k}^{0\prime}(\Lambda_j^{0\prime}\Lambda_j^0/N_j)^{-1}\boldsymbol{\lambda}_{j,i}^0 X_i' M_{\tilde{F}_j}\boldsymbol{\varepsilon}_k.
\end{aligned}$$

Defining $E[\boldsymbol{\varepsilon}_k\boldsymbol{\varepsilon}_k'] = \Omega_k$, we have

$$\begin{aligned}
&\frac{1}{NT}\sum_{j=1}^{S}\sum_{i:g_i^0=j} X_i' M_{\tilde{F}_j} I_8^j (F_j^{0\prime}\tilde{F}_j/T)^{-1}(\Lambda_j^{0\prime}\Lambda_j^0/N_j)^{-1}\boldsymbol{\lambda}_{j,i}^0 \\
= &\frac{1}{NT}\sum_{j=1}^{S}\sum_{i:g_i^0=j} X_i' M_{\tilde{F}_j}\left[\frac{1}{N_jT}\sum_{k:g_k^0=j}\boldsymbol{\varepsilon}_k\boldsymbol{\varepsilon}_k'\tilde{F}_j\right](F_j^{0\prime}\tilde{F}_j/T)^{-1}(\Lambda_j^{0\prime}\Lambda_j^0/N_j)^{-1}\boldsymbol{\lambda}_{j,i}^0 \\
= &\sum_{j=1}^{S}\frac{N_j}{N}\times\frac{1}{N_j^2T^2}\sum_{i:g_i^0=j}\sum_{k:g_k^0=j} X_i' M_{\tilde{F}_j}(\boldsymbol{\varepsilon}_k\boldsymbol{\varepsilon}_k' - \Omega_k)\tilde{F}_j(F_j^{0\prime}\tilde{F}_j/T)^{-1}(\Lambda_j^{0\prime}\Lambda_j^0/N_j)^{-1}\boldsymbol{\lambda}_{j,i}^0 \\
&+\sum_{j=1}^{S}\frac{N_j}{N}\times\frac{1}{N_j^2T^2}\sum_{i:g_i^0=j}\sum_{k:g_k^0=j} X_i' M_{\tilde{F}_j}\Omega_k\tilde{F}_j(F_j^{0\prime}\tilde{F}_j/T)^{-1}(\Lambda_j^{0\prime}\Lambda_j^0/N_j)^{-1}\boldsymbol{\lambda}_{j,i}^0 \\
= &\sum_{j=1}^{S}\frac{N_j}{N}\times\frac{1}{N_j^2T^2}\sum_{i:g_i^0=j}\sum_{k:g_k^0=j} X_i' M_{\tilde{F}_j}\Omega_k\tilde{F}_j(F_j^{0\prime}\tilde{F}_j/T)^{-1}(\Lambda_j^{0\prime}\Lambda_j^0/N_j)^{-1}\boldsymbol{\lambda}_{j,i}^0 \\
&+o_p(1)\times O_p\left(\|\boldsymbol{\beta}^0 - \tilde{\boldsymbol{\beta}}\|^2\right) + \frac{1}{\sqrt{NT}}O_p\left(\frac{1}{\min\{N^{1/2},T^{1/2}\}}\right) + \frac{1}{\sqrt{N}}O_p\left(\frac{1}{\min\{N,T\}}\right),
\end{aligned}$$

which follows from Bai (2009). Then, we have

$$\begin{aligned}
&\left[\frac{1}{NT}\sum_{j=1}^{S}\sum_{i:g_i=j} X_i' M_{\tilde{F}_j} X_i - \frac{1}{T}\sum_{j=1}^{S}\frac{N_j}{N}\frac{1}{N_j^2}\sum_{i:g_i^0=j}\sum_{k:g_k^0=j} X_i' M_{\tilde{F}_j} X_k c_{j,ki} + \frac{1}{T}\Sigma(\kappa)\right](\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) \\
= &\sum_{j=1}^{S}\frac{N_j}{N}\sum_{i:g_i^0=j}\frac{1}{N_jT}\left[X_i' M_{\tilde{F}_j} - \frac{1}{N_j}\sum_{k:g_k^0=j} c_{j,ki} X_k' M_{\tilde{F}_j}\right]\boldsymbol{\varepsilon}_i \\
&+\sum_{j=1}^{S}\frac{N_j}{N}\times\frac{1}{N_j^2T^2}\sum_{i:g_i^0=j}\sum_{k:g_k^0=j} X_i' M_{\tilde{F}_j}\Omega_k\tilde{F}_j(F_j^{0\prime}\tilde{F}_j/T)^{-1}(\Lambda_j^{0\prime}\Lambda_j^0/N_j)^{-1}\boldsymbol{\lambda}_{j,i}^0
\end{aligned}$$

$$+o_p((NT)^{-1/2}) + \frac{1}{\sqrt{N}} O_p\left(\frac{1}{\min\{N,T\}}\right),$$

where $c_{j,ki} = \boldsymbol{\lambda}^{0'}_{g^0_k,k}(\Lambda^{0\prime}_j\Lambda^0_j/N_j)^{-1}\boldsymbol{\lambda}^0_{g^0_i,i}$. Using the Lemmas A.8 and A.9 of Bai (2009), we have the following expression

$$\begin{aligned}
&\frac{1}{\sqrt{NT}}\sum_{i:g^0_i=j}\left[X_i'M_{\tilde{F}_j} - \frac{1}{N_j}\sum_{k:g^0_k=j} c_{j,ki}X_k'M_{\tilde{F}_j}\right]\boldsymbol{\varepsilon}_i = \frac{1}{\sqrt{NT}}\sum_{i:g^0_i=j}\left[X_i'M_{F^0_j} - \frac{1}{N_j}\sum_{k:g^0_k=j} c_{j,ki}X_k'M_{F^0_j}\right]\boldsymbol{\varepsilon}_i \\
&+\sqrt{\frac{T}{N}}\times\left[-\frac{1}{N_j}\sum_{i:g^0_i=j}\sum_{k:g^0_k=j}\frac{(X_i-V_i)'F^0_j}{T}\left(\frac{F^{0\prime}_jF^0_j}{T}\right)^{-1}\left(\frac{\Lambda^{0\prime}_j\Lambda^0_j}{N_j}\right)^{-1}\boldsymbol{\lambda}_{g^0_k,k}\left(\frac{\boldsymbol{\varepsilon}_i'\boldsymbol{\varepsilon}_k}{T}\right)\right] + o_p(1) \\
&= \frac{1}{\sqrt{NT}}\sum_{i:g^0_i=j}\left[X_i'M_{F^0_j} - \frac{1}{N_j}\sum_{k:g^0_k=j} c_{j,ki}X_k'M_{F^0_j}\right]\boldsymbol{\varepsilon}_i + \sqrt{\frac{T}{N}}\boldsymbol{\eta}_j + o_p(1)
\end{aligned}$$

with $V_{j,i} = N_j^{-1}\sum_{k:g^0_k=j} c_{j,ki}X_k$, and $\boldsymbol{\eta}_j$ is defined in Theorem 3. Also,

$$\begin{aligned}
&\frac{1}{N_jT}\sum_{i:g^0_i=j} X_i'M_{\tilde{F}_j}X_i - \frac{1}{T}\frac{1}{N_j^2}\sum_{i:g^0_i=j}\sum_{k:g^0_k=j} X_i'M_{\tilde{F}_j}X_kc_{j,ki} \\
&= \frac{1}{N_jT}\sum_{i:g^0_i=j} X_i'M_{F^0_j}X_i - \frac{1}{T}\frac{1}{N_j^2}\sum_{i:g^0_i=j}\sum_{k:g^0_k=j} X_i'M_{F^0_j}X_kc_{j,ki} + o_p(1).
\end{aligned}$$

Then, we have

$$\begin{aligned}
&\hat{D}(F^0_1,...,F^0_S,\kappa)\left[\sqrt{NT}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)\right] \\
= \;& \frac{1}{\sqrt{NT}}\sum_{j=1}^S\sum_{i:g^0_i=j}\left[X_i'M_{F^0_j} - \frac{1}{N_j}\sum_{k:g^0_k=j} c_{j,ki}X_k'M_{F^0_j}\right]\boldsymbol{\varepsilon}_i \\
&+\sqrt{NT}\sum_{j=1}^S\frac{N_j}{N}\frac{1}{N_j^2T^2}\sum_{i:g^0_i=j}\sum_{k:g^0_k=j} X_i'M_{\tilde{F}_j}\Omega_k\tilde{F}_j\left(\frac{F^{0\prime}_j\tilde{F}_j}{T}\right)^{-1}\left(\frac{\Lambda^{0\prime}_j\Lambda^0_j}{N_j}\right)^{-1}\boldsymbol{\lambda}^0_{g^0_i,i} + o_p(1) \\
= \;& \frac{1}{\sqrt{NT}}\sum_{j=1}^S\sum_{i:g^0_i=j} Z_{j,i}(F^0_j)'\boldsymbol{\varepsilon}_i + \sqrt{\frac{T}{N}}\sum_{j=1}^S\boldsymbol{\eta}_j + \sqrt{\frac{N}{T}}\sum_{j=1}^S\boldsymbol{\zeta}_j + o_p(1),
\end{aligned}$$

where $Z_{j,i}(F^0_j)$, $\hat{D}(F^0_1,...,F^0_S,\kappa)$, $\boldsymbol{\eta}_j$ and $\boldsymbol{\zeta}_j$ are defined in Theorem 3. This leads to the limit of covariance matrix of $\sqrt{NT}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)$ as

$$V_\beta(F^0_1,...,F^0_S) = D_0(F^0_1,...,F^0_S)^{-1}J_0(F^0_1,...,F^0_S)D_0(F^0_1,...,F^0_S)^{-1},$$

where $D_0(F^0_1,...,F^0_S)$ is the probability limit of $\hat{D}(F^0_1,...,F^0_S)$ and $J_0(F^0_1,...,F^0_S)$ is defined in Assumption F. By the preceding asymptotic representation and Assumption F, we have

$$\sqrt{NT}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^0) \to N\left(\boldsymbol{v}_0, V_\beta(F^0_1,...,F^0_S)\right),$$

where $\boldsymbol{v}_0$ is defined in Theorem 3.

Finally, we prove the variable selection consistency $P(\hat{\boldsymbol{\beta}}_2 = \mathbf{0}) \to 1$ as $N, T \to \infty$. This part is almost identical to the proof of Fan and Li (2001). It is sufficient to show that with probability tending to 1 as $N, T \to \infty$, for some small $\delta_{N,T} = C/\sqrt{NT}$ with a constant C, and for each element of $\boldsymbol{\beta}_2 = (\beta_{21}, ..., \beta_{2,p-q})$, we have

$$\frac{\partial L_{NT}(\boldsymbol{\beta}, G, F_1, ..., F_S, \Lambda_1, ..., \Lambda_S)}{\partial \beta_{2k}} > 0 \quad (0 < \beta_{2k} < \delta_{N,T}),$$
$$\frac{\partial L_{NT}(\boldsymbol{\beta}, G, F_1, ..., F_S, \Lambda_1, ..., \Lambda_S)}{\partial \beta_{2k}} < 0 \quad (-\delta_{N,T} < \beta_{2k} < 0),$$

for $k = 1, ..., p-q$. Let $X_{i,2}$ be the set of $p-q$ columns of X_i, corresponding to $\boldsymbol{\beta}_2$. So, $X_{i,2}$ is $T \times (p-q)$ dimensional matrix. By the first derivative of $L_{NT}(\boldsymbol{\beta}, G, F_1, ..., F_S)/(NT)$ with respect to $\boldsymbol{\beta}_2 = (\beta_{21}, ..., \beta_{2,p-q})$, we have

$$\begin{aligned}
&\frac{1}{NT} \cdot \frac{\partial L_{NT}(\hat{\boldsymbol{\beta}}, \hat{G}, \hat{F}_1, ..., \hat{F}_S, \hat{\Lambda}_1, ..., \hat{\Lambda}_S)}{\partial \boldsymbol{\beta}_2} \\
= & -\frac{2}{NT} \sum_{i=1}^{N} X_{i,2}'(\boldsymbol{y}_i - X_i \hat{\boldsymbol{\beta}} - \hat{F}_{\hat{g}_i} \hat{\boldsymbol{\lambda}}_{\hat{g}_i,i}) + \frac{\partial p_{\kappa,\gamma}\left(|\hat{\boldsymbol{\beta}}_2|\right)}{\partial \boldsymbol{\beta}_2} \\
= & -\frac{2}{NT} \sum_{i=1}^{N} X_{i,2}' \left(X_i(\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}) + (F^0_{g_i^0} \boldsymbol{\lambda}^0_{g_i^0,i} - \hat{F}_{\hat{g}_i} \hat{\boldsymbol{\lambda}}_{\hat{g}_i,i}) + \boldsymbol{\varepsilon}_i \right) + \frac{\partial p_{\kappa,\gamma}\left(|\hat{\boldsymbol{\beta}}_2|\right)}{\partial \boldsymbol{\beta}_2} \\
= & -\frac{2}{NT} \sum_{i=1}^{N} X_{i,2}' X_i(\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}) - \frac{2}{NT} \sum_{i=1}^{N} X_{i,2}'(F^0_{g_i^0} \boldsymbol{\lambda}^0_{g_i^0,i} - \hat{F}_{\hat{g}_i} \hat{\boldsymbol{\lambda}}_{\hat{g}_i,i}) \\
& + \frac{2}{NT} \sum_{i=1}^{N} X_{i,2}' \boldsymbol{\varepsilon}_i + \frac{\partial p_{\kappa,\gamma}\left(|\hat{\boldsymbol{\beta}}_2|\right)}{\partial \boldsymbol{\beta}_2} \\
= & \ I_1 + I_2 + I_3 + \frac{\partial p_{\kappa,\gamma}\left(|\hat{\boldsymbol{\beta}}_2|\right)}{\partial \boldsymbol{\beta}_2}.
\end{aligned}$$

The third term I_3 is $O_p((NT)^{-1/2})$. Together with the result of Theorem 1, we know that $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 = O_p((NT)^{-1/2})$. Thus, the first term I_1 is $O_p((NT)^{-1/2})$. The second term is

$$\begin{aligned}
& \frac{1}{NT} \sum_{i=1}^{N} X_{i,2}'(F^0_{g_i^0} \boldsymbol{\lambda}^0_{g_i^0,i} - \hat{F}_{\hat{g}_i} \hat{\boldsymbol{\lambda}}_{\hat{g}_i,i}) \\
= & \ \frac{1}{NT} \sum_{i=1}^{N} X_{i,2}'(F^0_{g_i^0} - \hat{F}_{\hat{g}_i}) \boldsymbol{\lambda}^0_{g_i^0,i} + \frac{1}{NT} \sum_{i=1}^{N} X_{i,2}' \hat{F}_{\hat{g}_i}(\boldsymbol{\lambda}^0_{g_i^0,i} - \hat{\boldsymbol{\lambda}}_{\hat{g}_i,i}),
\end{aligned}$$

which is $O_p(1/\min\{N, T\})$. Each element of the first derivative $\partial p_{\kappa,\gamma}\left(|\hat{\boldsymbol{\beta}}_2|\right)/\partial \boldsymbol{\beta}_2$ is $p'_{\kappa,\gamma}\left(|\hat{\beta}_{2k}|\right) \operatorname{sign}(\beta_{2k})$ for $k = 1, ..., p-q$. Finally, we have

$$\frac{\partial L_{NT}(\hat{\boldsymbol{\beta}}, \hat{G}, \hat{F}_1, ..., \hat{F}_S, \hat{\Lambda}_1, ..., \hat{\Lambda}_S)}{\partial \beta_{2k}}$$

$$= \quad NT \cdot \kappa \left[\frac{1}{\kappa} p'_{\kappa,\gamma} \left(|\hat{\beta}_{2k}| \right) \text{sign}(\hat{\beta}_{2k}) + O_p \left(1/(\min\{N,T\} \cdot \kappa) \right) \right].$$

Thus the sign of $\hat{\beta}_{2k}$ determines the sign of $\partial L_{NT}(\hat{\boldsymbol{\beta}}, \hat{G}, \hat{F}_1, ..., \hat{F}_S, \hat{\Lambda}_1, ..., \hat{\Lambda}_S)/\partial \beta_{2k}$. Hence, this result implies the sign claim. This completes the proof. □

Proof of Theorem 4

We divide the proof of Theorem 4 into two steps. In step 1, we develop an estimator of the expected mean squared error, which can be used to select the number of predictors $\boldsymbol{x}$ under no factor structure. In step 2, we derive an additional penalty term that penalizes the model complexity caused by the factor structures.

Step 1: We decompose the bias b as

$$b = B_1 + B_2 + B_3 + B_4 + B_5,$$

where

$$B_1 = E_y \left[\frac{1}{NT} \sum_{i=1}^{N} \|\boldsymbol{y}_i - X_i \tilde{\boldsymbol{\beta}} - \tilde{F}_{g_i^0} \tilde{\boldsymbol{\lambda}}_{g_i^0,i}\|^2 - \frac{1}{NT} \sum_{i=1}^{N} \|\boldsymbol{y}_i - X_i \hat{\boldsymbol{\beta}} - \hat{F}_{\hat{g}_i} \hat{\boldsymbol{\lambda}}_{\hat{g}_i,i}\|^2 \right],$$

$$B_2 = E_y \left[\frac{1}{NT} \sum_{i=1}^{N} \|\boldsymbol{y}_i - X_i \boldsymbol{\beta}^0 - \tilde{F}_{g_i^0} \tilde{\boldsymbol{\lambda}}_{g_i^0,i}\|^2 - \frac{1}{NT} \sum_{i=1}^{N} \|\boldsymbol{y}_i - X_i \tilde{\boldsymbol{\beta}} - \tilde{F}_{g_i^0} \tilde{\boldsymbol{\lambda}}_{g_i^0,i}\|^2 \right],$$

$$B_3 = E_y \left[E_z \{ \frac{1}{NT} \sum_{i=1}^{N} \|\boldsymbol{z}_i - X_i \boldsymbol{\beta}^0 - \tilde{F}_{g_i^0} \tilde{\boldsymbol{\lambda}}_{g_i^0,i}\|^2 \} - \frac{1}{NT} \sum_{i=1}^{N} \|\boldsymbol{y}_i - X_i \boldsymbol{\beta}^0 - \tilde{F}_{g_i^0} \tilde{\boldsymbol{\lambda}}_{g_i^0,i}\|^2 \right]$$

$$B_4 = E_y \left[E_z \left[\frac{1}{NT} \sum_{i=1}^{N} \|\boldsymbol{z}_i - X_i \tilde{\boldsymbol{\beta}} - \tilde{F}_{g_i^0} \tilde{\boldsymbol{\lambda}}_{g_i^0,i}\|^2 \right] - E_z \left[\frac{1}{NT} \sum_{i=1}^{N} \|\boldsymbol{z}_i - X_i \boldsymbol{\beta}^0 - \tilde{F}_{g_i^0} \tilde{\boldsymbol{\lambda}}_{g_i^0,i}\|^2 \right] \right],$$

$$B_5 = E_y \left[E_z \left[\frac{1}{NT} \sum_{i=1}^{N} \|\boldsymbol{z}_i - X_i \hat{\boldsymbol{\beta}} - \hat{F}_{\hat{g}_i} \hat{\boldsymbol{\lambda}}_{\hat{g}_i,i}\|^2 \right] - E_z \left[\frac{1}{NT} \sum_{i=1}^{N} \|\boldsymbol{z}_i - X_i \tilde{\boldsymbol{\beta}} - \tilde{F}_{g_i^0} \tilde{\boldsymbol{\lambda}}_{g_i^0,i}\|^2 \right] \right],$$

where the expectations $E_y[\cdot]$ and $E_z[\cdot]$ are taken with respect to the joint distribution of $\{\boldsymbol{y}_1, ..., \boldsymbol{y}_N\}$ and $\{\boldsymbol{z}_1, ..., \boldsymbol{z}_N\}$ given the predictors and factor structures. Define

$$\ell_y(\boldsymbol{\beta}, F_1, ..., F_S, \Lambda_1, ..., \Lambda_S) = \frac{1}{NT} \sum_{j=1}^{S} \sum_{i;g_i=j} \|\boldsymbol{y}_i - X_i \boldsymbol{\beta} - F_{g_i} \boldsymbol{\lambda}_{g_i,i}\|^2,$$

$$\ell_z(\boldsymbol{\beta}, F_1, ..., F_S, \Lambda_1, ..., \Lambda_S) = E_z \left[\frac{1}{NT} \sum_{j=1}^{S} \sum_{i;g_i=j} \|\boldsymbol{z}_i - X_i \boldsymbol{\beta} - F_{g_i} \boldsymbol{\lambda}_{g_i,i}\|^2 \right].$$

It can be shown that B_1, B_3, and B_5 are dominated by B_2 and B_4, thus can be ignored. We next evaluate B_2. Noting that

$$\frac{\partial}{\partial \boldsymbol{\beta}} \left\{ \ell_y(\boldsymbol{\beta}, \tilde{F}_1, ..., \tilde{F}_S, \tilde{\Lambda}_1, ..., \tilde{\Lambda}_S) + p_{\kappa,\gamma}(|\boldsymbol{\beta}|) \right\} \Big|_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}} = \mathbf{0},$$

the Taylor expansion of $\ell_y(\boldsymbol{\beta}^0, \tilde{F}_1, ..., \tilde{F}_S, \tilde{\Lambda}_1, ..., \tilde{\Lambda}_S)$ around $\tilde{\boldsymbol{\beta}}$ gives

$$\begin{aligned}\ell_y(\boldsymbol{\beta}^0, \tilde{F}_1, ..., \tilde{F}_S, \tilde{\Lambda}_1, ..., \tilde{\Lambda}_S) &= \ell_y(\tilde{\boldsymbol{\beta}}, \tilde{F}_1, ..., \tilde{F}_S, \tilde{\Lambda}_1, ..., \tilde{\Lambda}_S) - \partial p_{\kappa,\gamma}(|\tilde{\boldsymbol{\beta}}|)/\partial \boldsymbol{\beta}'(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) \\ &+ \frac{1}{2}\frac{1}{NT}\sqrt{NT}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)' K_x \sqrt{NT}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) + o_p\left(N^{-1}T^{-1}\right),\end{aligned}$$

where $K_x = \frac{1}{NT}\sum_{i=1}^{N} X_i' X_i$. For small κ, $\partial p_{\kappa,\gamma}(|\tilde{\boldsymbol{\beta}}|)/\partial \boldsymbol{\beta}'(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) = o_p(1/(NT)$. The covariance matrix of $\sqrt{NT}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)$ is $V_\beta(F_1^0, ..., F_S^0, \kappa)$. Thus, we can write B_2 as

$$B_2 = \frac{1}{2NT}\text{tr}\left[K_x V_\beta(F_1^0, ..., F_S^0, \kappa)\right] + o(N^{-1}T^{-1}).$$

Using the same augment for B_2,

$$B_4 = \frac{1}{2NT}\text{tr}\left[K_x V_\beta(F_1^0, ..., F_S^0, \kappa)\right] + o(N^{-1}T^{-1}).$$

Finally, summing up all terms $B_1 \sim B_5$, the bias, contributed by the estimated observable structure $X_i\hat{\boldsymbol{\beta}}$ (the penalty on the estimated factor structures will be investigated in Step 2), becomes

$$\frac{1}{NT}\text{tr}\left[K_x V_\beta(F_1^0, ..., F_S^0, \kappa)\right] + o(N^{-1}T^{-1}).$$

Therefore, in the first step, the expected mean squared error can be approximated as

$$\frac{1}{NT}\sum_{j=1}^{S}\sum_{i;\hat{g}_i=j}\left\|\boldsymbol{y}_i - X_i\hat{\boldsymbol{\beta}} - \hat{F}_{\hat{g}_i}\hat{\boldsymbol{\lambda}}_{\hat{g}_i,i}\right\|^2 + \frac{1}{NT}\text{tr}\left[K_x V_\beta(F_1^0, ..., F_S^0, \kappa)\right]. \tag{20}$$

which is bias-corrected only for the estimated structure of $X_i\hat{\boldsymbol{\beta}}$.

Step 2: Under no factor structure, (20) can be used for selecting the regularization parameter κ. We need an additional penalty term that penalizes the model complexity caused by the factor structures. The overall criterion for selecting κ and k_j group-specific factors ($j = 1, ..., S$) has the form

$$\frac{1}{NT}\sum_{j=1}^{S}\sum_{i;\hat{g}_i=j}\left\|\boldsymbol{y}_i - X_i\hat{\boldsymbol{\beta}} - \hat{F}_{\hat{g}_i}\hat{\boldsymbol{\lambda}}_{\hat{g}_i,i}\right\|^2 + \frac{1}{NT}\text{tr}\left[K_x V_\beta(F_1^0, ..., F_S^0, \kappa)\right] + \sum_{j=1}^{S} k_j g(N_j, T)$$

and we will determine $g(N_j, T)$ such that this criterion can consistently estimate the factor structure.

The proof of this step for selecting the number of factors consistently uses the similar augment as that employed in Bai (2009). We focus on the selection of the true number of group-specific factors r_j. We first assume that $r_j \leq k_j$, where k_j is the given number of group-specific factors in the estimation process. Under $r_j \leq k_j$, we have $\hat{\boldsymbol{\beta}}(k_j) - \boldsymbol{\beta} = O(1/\sqrt{NT})$, where the script k_j indicates k_j factor models are estimated. Then it is shown that, for the data within the j-th group,

$$\boldsymbol{y}_i - X_i\hat{\boldsymbol{\beta}}(k_j) = F_j(k_j)\boldsymbol{\lambda}_{j,i} + \boldsymbol{\varepsilon}_i + O_p\left(1/\sqrt{NT}\right).$$

Thus, $O_p(1/\sqrt{NT})$ error term will not affect the analysis of Bai and Ng (2002), as mentioned in Bai (2009). This indicates that

$$\frac{1}{N_j T}\sum_{g_i=j}\left\|\boldsymbol{y}_i - X_i\hat{\boldsymbol{\beta}}(k_j) - \hat{F}_j(k_j)\hat{\boldsymbol{\lambda}}_{j,i}\right\|^2$$
$$-\frac{1}{N_j T}\sum_{g_i=j}\left\|\boldsymbol{y}_i - X_i\hat{\boldsymbol{\beta}}(r_j) - \hat{F}_j(r_j)\hat{\boldsymbol{\lambda}}_{j,i}\right\|^2 = O_p\left(\frac{1}{\min\{N,T\}}\right).$$

If $k_j < r_j$, it is then, for some $c > 0$, not depending on N_j and T, we have

$$\frac{1}{N_j T}\sum_{g_i=j}\left\|\boldsymbol{y}_i - X_i\hat{\boldsymbol{\beta}}(k_j) - \hat{F}_j(k_j)\hat{\boldsymbol{\lambda}}_{j,i}\right\|^2$$
$$-\frac{1}{N_j T}\sum_{g_i=j}\left\|\boldsymbol{y}_i - X_i\hat{\boldsymbol{\beta}}(r_j) - \hat{F}_j(r_j)\hat{\boldsymbol{\lambda}}_{j,i}\right\|^2 > c.$$

This implies that any penalty function that converges to zero but is of greater magnitude than $O_p(1/\min\{N,T\})$ will lead to consistent estimation of the number of factors. The term $(T+N_j)/TN_j \times \log(TN_j)$ satisfies these conditions. This completes the proof of Theorem 4. □

Proof of Theorem 5

We first assume that $S_0 < S$, where S_0 is the true number of groups and S is the number of groups set by researcher. Under $S_0 < S$, at least for one particular group, say the group j, the set of units within the j-th group will be divided into two (or more) sub-groups, while they are within the same group. Suppose that the data within the j-th group are divided into two-groups, j_1 and j_2. Depending upon the setting of the number of groups S, the data within the j-th group may be divided into more than two sub-groups. Or, in addition to the j-th group, some other groups may be divided into several sub-groups. Even for such cases, the argument below applies in the same manner.

Let N_{j_1} and N_{j_2} be the number of units that belong to the sub-groups j_1 and j_2. First, if $N_{j_a}/N = o(1)$ for one of j_a, the model can not be estimable and such model setting will be deleted automatically. Next, consider the case: $N_{j_a}/N = O(1)$ for $a = 1, 2$. Although the j-th group is divided two sub-groups j_1 and j_2, we have

$$\boldsymbol{y}_i - X_i\hat{\boldsymbol{\beta}}(r_j) - F_{j_a}(r_j)\boldsymbol{\lambda}_{j_a,i} - \boldsymbol{\varepsilon}_i = O_p\left(\frac{1}{\sqrt{NT}}\right),$$

for $a = 1, 2$, which implies that

$$\frac{1}{NT}\sum_{i;g_i=j}\|\boldsymbol{y}_i - X_i\boldsymbol{\beta}(r_j) - F_j(r_j)\boldsymbol{\lambda}_{j,i}\|^2$$
$$-\frac{1}{NT}\sum_{i;g_i=j} I(i \in j_1)\left\|\boldsymbol{y}_i - X_i\hat{\boldsymbol{\beta}}(r_j) - \hat{F}_j(r_j)\hat{\boldsymbol{\lambda}}_{j,i}\right\|^2$$

$$-\frac{1}{NT}\sum_{i;g_i=j} I(i \in j_2) \left\| \boldsymbol{y}_i - X_i\hat{\boldsymbol{\beta}}(r_j) - \hat{F}_j(r_j)\hat{\boldsymbol{\lambda}}_{j,i} \right\|^2 = O_p\left(\frac{1}{\min\{N,T\}}\right),$$

where $I()$ is the indicator function. To avoid over identification, we need a penalty function that is of greater magnitude than $O_p(1/(\min\{N,T\}))$.

If $S < S_0$, it is then, some different group(s), j and k are merged into one group, say ℓ, while they are originally not in the same group. Then, for a positive constant $c > 0$, which does not depend on N and T, we have

$$\begin{aligned}
&\frac{1}{NT}\sum_{i;g_i=k} \left\| \boldsymbol{y}_i - X_i\hat{\boldsymbol{\beta}}(r_k) - \hat{F}_k(r_k)\hat{\boldsymbol{\lambda}}_{k,i} \right\|^2 \\
&+\frac{1}{NT}\sum_{i;g_i=j} \left\| \boldsymbol{y}_i - X_i\hat{\boldsymbol{\beta}}(r_j) - \hat{F}_j(r_j)\hat{\boldsymbol{\lambda}}_{j,i} \right\|^2 \\
&-\frac{1}{NT}\sum_{i;g_i=\ell} \left\| \boldsymbol{y}_i - X_i\hat{\boldsymbol{\beta}}(r_\ell) - \hat{F}_\ell(r_\ell)\hat{\boldsymbol{\lambda}}_{\ell,i} \right\|^2 > c.
\end{aligned}$$

To avoid under identification, we need a penalty function that converges to zero.

Applying the above argument to each group, any penalty function that converges to zero but is of greater magnitude than $O_p(1/(\min\{N,T\}))$ will lead to consistent estimation of the number of groups S. The penalty term $\sum_{j=1}^{S} k_j\hat{\sigma}^2\frac{N_j}{N}\left(\frac{T+N_j}{TN_j}\right)\log(TN_j)$ in the proposed criterion $C_p(k_1,...,k_S,\kappa)$ in (8) satisfies these conditions. This completes the proof. □

Proofs of Theorem 6

Let $G^0 = \{g_1^0,...,g_N^0\}$ and $G = \{g_1,...,g_N\}$ denote the population grouping and any grouping of the cross-sectional units into S groups. First, we note that the estimator $\{\hat{\boldsymbol{\beta}}_1...,\hat{\boldsymbol{\beta}}_S,\hat{G},\hat{F}_1,...,\hat{F}_S,\hat{\Lambda}_1,...,\hat{\Lambda}_S\}$ is defined as the minimizer of

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

subject to the constraints imposed in Section 8.1.

Consistency of $\hat{\boldsymbol{\beta}}_j$ can be obtained by modifying the proof of Theorem 1. Without loss of generality, we again assume that $\boldsymbol{\beta}_j^0 = \mathbf{0}$ and concentrate out the factor loadings as we can express them as $\Lambda_j = W_j'F_j(F_j'F_j)^{-1} = W_j'F_j/T$ where $W_j = (\boldsymbol{w}_{j,1},...,\boldsymbol{w}_{j,N_j})$ such that $\boldsymbol{w}_{j,i} = \boldsymbol{y}_i - X_i\boldsymbol{\beta}_j$ and $g_i = j$. Note again that the set of estimates $\{\hat{\boldsymbol{\beta}}_1,...,\hat{\boldsymbol{\beta}}_S,\hat{G},\hat{F}_1,...,\hat{F}_S,\hat{\Lambda}_1,...,\hat{\Lambda}_S\}$ that jointly minimizes the objective function $L_{NT}(\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_S,G,F_1,...,F_S,\Lambda_1,...,\Lambda_S)$, and the set of estimates $\{\hat{\boldsymbol{\beta}}_1,...,\hat{\boldsymbol{\beta}}_S,\hat{G},\hat{F}_1,...,\hat{F}_S\}$ that jointly minimizes the following concentrated and centered objective function

$$U_{NT}(\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_S,G,F_1,...,F_S)$$

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

are equal. Note that the group membership is not fixed to its population.
Noting that the true data generating process is $\boldsymbol{y}_i = F^0_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i} + \boldsymbol{\varepsilon}_i$ $(X_i\boldsymbol{\beta}^0_j = \mathbf{0})$, the estimator, the objective function $U_{NT}(\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_S, G, F_1, ..., F_S)$ is further expressed as

$$\begin{aligned}
& U_{NT}(\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_S, G, F_1, ..., F_S) \\
= & \left(\frac{1}{NT}\sum_{j=1}^{S}\sum_{i;g_i=j}\boldsymbol{\beta}_j' X_i' M_{F_{g_i}} X_i\boldsymbol{\beta}_j\right) + \frac{1}{NT}\sum_{j=1}^{S}\sum_{i;g_i=j}\boldsymbol{\lambda}^0_{g_i^0,i}{}' F^{0\prime}_{g_i^0} M_{F_{g_i}} F^0_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i} \\
& + 2\left[\frac{1}{NT}\sum_{j=1}^{S}\sum_{i;g_i=j}\boldsymbol{\beta}_j' X_i' M_{F_{g_i}} F^0_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i}\right] + 2\left(\frac{1}{NT}\sum_{j=1}^{S}\sum_{i;g_i=j}\boldsymbol{\beta}_j' X_i' M_{F_{g_i}}\boldsymbol{\varepsilon}_i\right) \\
& + \frac{2}{NT}\sum_{j=1}^{S}\sum_{i;g_i=j}\boldsymbol{\lambda}^0_{g_i^0,i}{}' F^{0\prime}_{g_i^0} M_{F_{g_i}}\boldsymbol{\varepsilon}_i + \frac{1}{NT}\sum_{i=1}^{N}\boldsymbol{\varepsilon}_i'(M_{F_{g_i}} - M_{F^0_{g_i^0}})\boldsymbol{\varepsilon}_i + \sum_{j=1}^{S} p_{\kappa,\gamma}\left(|\boldsymbol{\beta}_j|\right) \\
= & \sum_{j=1}^{S}\boldsymbol{\beta}_j'\left(\frac{1}{NT}\sum_{i;g_i=j} X_i' M_{F_{g_i}} X_i\right)\boldsymbol{\beta}_j + \frac{1}{NT}\sum_{j=1}^{S}\sum_{i;g_i=j}\boldsymbol{\lambda}^0_{g_i^0,i}{}' F^{0\prime}_{g_i^0} M_{F_{g_i}} F^0_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i} \\
& + 2\sum_{j=1}^{S}\boldsymbol{\beta}_j'\left[\frac{1}{NT}\sum_{i;g_i=j} X_i' M_{F_{g_i}} F^0_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i}\right] + O_p(T^{-1/4}) + O_p(N^{-1/4}) \\
= & \frac{1}{N}\sum_{j=1}^{S}\left[\boldsymbol{\beta}_j' D_j\boldsymbol{\beta}_j + \boldsymbol{\zeta}_j' E_j\boldsymbol{\zeta}_j + 2\boldsymbol{\beta}_j' L_j'\boldsymbol{\zeta}_j\right] + O_p(T^{-1/4}) + O_p(N^{-1/4}) \\
= & \frac{1}{NT}\tilde{U}_{NT}(\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_S, G, F_1, ..., F_S) + O_p(T^{-1/4}) + O_p(N^{-1/4}),
\end{aligned}$$

where we have used Lemma A.1, and D_j, E_j, L_j and $\boldsymbol{\zeta}_j$ are defined in the proof of Theorem 1.

Completing the square of $\tilde{U}_{NT}(\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_S, G, F_1, ..., F_S)$ in terms of $\boldsymbol{\beta}_j$ and then using Assumption D$'$ and the argument in the proof of Theorem 1, we obtain

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

The argument for $T^{-1}\|\hat{F}_j - F^0_j H_j\|^2 = O_p(T^{-1/8}) + O_p(N^{-1/8})$ is the same as in Theorem 1. The details are omitted. This proves Theorem 6. □

Proof of Theorem 7

Note that $\hat{g}_i$ satisfies

$$\hat{g}_i = \text{argmin}_{j\in\{1,...,S\}}\left[\frac{1}{T}(\boldsymbol{y} - X_i\hat{\boldsymbol{\beta}}_j)' M_{\hat{F}_j}(\boldsymbol{y} - X_i\hat{\boldsymbol{\beta}}_j) + p_{\kappa,\gamma}(|\hat{\boldsymbol{\beta}}_j|)\right].$$

Using $\boldsymbol{y}_i = X_i\boldsymbol{\beta}^0_{g_i^0} + F^0_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i} + \boldsymbol{\varepsilon}_i$, we have

$$\frac{1}{T}(\boldsymbol{y} - X_i\hat{\boldsymbol{\beta}}_j)' M_{\hat{F}_j}(\boldsymbol{y} - X_i\hat{\boldsymbol{\beta}}_j)$$

$$
\begin{aligned}
&= \frac{1}{T}\left(X_i(\boldsymbol{\beta}^0_{g_i^0} - \hat{\boldsymbol{\beta}}_j) + F^0_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i} + \boldsymbol{\varepsilon}_i\right)' M_{\hat{F}_j}\left(X_i(\boldsymbol{\beta}^0_{g_i^0} - \hat{\boldsymbol{\beta}}_j) + F^0_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i} + \boldsymbol{\varepsilon}_i\right) \\
&= \frac{1}{T}\left(X_i(\boldsymbol{\beta}^0_{g_i^0} - \hat{\boldsymbol{\beta}}_j) + F^0_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i} + \boldsymbol{\varepsilon}_i\right)' M_{F^0_j}\left(X_i(\boldsymbol{\beta}^0_{g_i^0} - \hat{\boldsymbol{\beta}}_j) + F^0_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i} + \boldsymbol{\varepsilon}_i\right) \\
&\quad + \frac{1}{T}\left(X_i(\boldsymbol{\beta}^0_{g_i^0} - \hat{\boldsymbol{\beta}}_j) + F^0_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i} + \boldsymbol{\varepsilon}_i\right)' (M_{\hat{F}_j} - M_{F^0_j})\left(X_i(\boldsymbol{\beta}^0_{g_i^0} - \hat{\boldsymbol{\beta}}_j) + F^0_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i} + \boldsymbol{\varepsilon}_i\right) \\
&= a1 + a2
\end{aligned}
$$

We first show $a2 = o_p(1)$ uniformly in i whether $g_i^0 = j$ or $g_i \neq j$. From $T^{-1}\|\hat{F}_j - F^0_j\|^2 = O_p(T^{-1/8}) + O_p(N^{-1/8})$ $(j = 1, ..., S)$, we have $\|M_{\hat{F}_j} - M_{F^0_j}\|^2 = \|P_{\hat{F}_j} - P_{F^0_j}\|^2 = O_p(T^{-1/8}) + O_p(N^{-1/8})$. If $g_i^0 = j$, then $\|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}^0_j\|^2 = O_p(T^{-1/4}) + O_p(N^{-1/4})$, and $a2 = o_p(1)$ follows from Lemma A2. Suppose $g_i^0 = k \neq j$, then $\|\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}^0_j\|^2 = O_p(1)$ $(k \neq j)$, we need a different argument. Consider the term

$$
\frac{1}{T}\| \left(X_i(\boldsymbol{\beta}^0_{g_i^0} - \hat{\boldsymbol{\beta}}_j)\right)' (M_{\hat{F}_j} - M_{F^0_j})\left(X_i(\boldsymbol{\beta}^0_{g_i^0} - \hat{\boldsymbol{\beta}}_j)\right) \| \leq O_p(1)\|P_{\hat{F}_j} - P_{F^0_j}\|^2 \frac{1}{T}\|X_i\|^2
$$

By Assumption D2′, $\max_i \frac{1}{T}\|X_i\|^2 = O_p(N^{\alpha})$, the above is bounded by

$$
[O_p(T^{-1/8}) + O_p(N^{-1/8})]O_p(N^{\alpha}) = o_p(1)
$$

because $\alpha < 1/16$ and $N/T^2 \to 0$. The remaining terms in $a2$ are all $o_p(1)$ by similar argument (some are already covered by Lemma A2(d)). Thus

$$
\frac{1}{T}(\boldsymbol{y} - X_i\hat{\boldsymbol{\beta}}_j)' M_{\hat{F}_j}(\boldsymbol{y} - X_i\hat{\boldsymbol{\beta}}_j) = a1 + o_p(1)
$$

The behavior of $a1$ is different for $g_i^0 = j$ and $g_i^0 \neq j$. If $g_i^0 \neq j$

$$
\begin{aligned}
&\frac{1}{T}(\boldsymbol{y} - X_i\hat{\boldsymbol{\beta}}_j)' M_{\hat{F}_j}(\boldsymbol{y} - X_i\hat{\boldsymbol{\beta}}_j) \\
= &\frac{1}{T}\left(X_i(\boldsymbol{\beta}^0_{g_i^0} - \hat{\boldsymbol{\beta}}_j) + F^0_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i}\right)' M_{F^0_j}\left(X_i(\boldsymbol{\beta}^0_{g_i^0} - \hat{\boldsymbol{\beta}}_j) + F^0_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i}\right) \\
&+ \frac{2}{T}\left(X_i(\boldsymbol{\beta}^0_{g_i^0} - \hat{\boldsymbol{\beta}}_j) + F^0_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i}\right)' M_{F^0_j}\boldsymbol{\varepsilon}_i + \frac{1}{T}\boldsymbol{\varepsilon}_i' M_{F^0_j}\boldsymbol{\varepsilon}_i + o_p(1). \\
= &\frac{1}{T}\left(X_i(\boldsymbol{\beta}^0_{g_i^0} - \boldsymbol{\beta}^0_j) + F^0_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i}\right)' M_{F^0_j}\left(X_i(\boldsymbol{\beta}^0_{g_i^0} - \boldsymbol{\beta}^0_j) + F^0_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i}\right) \\
&+ \frac{2}{T}\left(X_i(\boldsymbol{\beta}^0_{g_i^0} - \boldsymbol{\beta}^0_j) + F^0_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i}\right)' M_{F^0_j}\boldsymbol{\varepsilon}_i + \frac{1}{T}\boldsymbol{\varepsilon}_i' M_{F^0_j}\boldsymbol{\varepsilon}_i + o_p(1).
\end{aligned}
$$

In the last equality, we replace $\hat{\boldsymbol{\beta}}_j$ by $\boldsymbol{\beta}^0_j$. This is permissible because $\hat{\boldsymbol{\beta}}_j$ is for $\boldsymbol{\beta}^0_j$ and because $\|\boldsymbol{\beta}^0_j - \hat{\boldsymbol{\beta}}_j\|^2 \frac{1}{T}\|X_i\|^2 = o_p(1)$, $\|\boldsymbol{\beta}^0_j - \hat{\boldsymbol{\beta}}_j\| \frac{1}{T}\|X_i\| \|F^0_{g_i^0}\lambda^0_{g_i^0,i}\|^2 = o_p(1)$, etc, with $o_p(1)$ being uniform in i.

On the other and, if $g_i^0 = j$,

$$
\frac{1}{T}(\boldsymbol{y} - X_i\hat{\boldsymbol{\beta}}_j)' M_{\hat{F}_j}(\boldsymbol{y} - X_i\hat{\boldsymbol{\beta}}_j) = \frac{1}{T}\boldsymbol{\varepsilon}_i' M_{F^0_{g_i^0}}\boldsymbol{\varepsilon}_i + o_p(1),
$$

which follows by noting that $M_{F^0_j}F^0_{g_i^0} = 0$ and $\hat{\boldsymbol{\beta}}_j$ is consistent for $\boldsymbol{\beta}^0_{g_i^0}$ for $g_i^0 = j$

Similar to the proof of Theorem 2, let us define the event A_{ij} such that

$$\begin{aligned} A_{ij} &= \left\{ \frac{1}{T}\left(X_i(\boldsymbol{\beta}^0_{g_i^0}-\boldsymbol{\beta}^0_j)+F^0_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i}\right)' M_{F^0_j}\left(X_i(\boldsymbol{\beta}^0_{g_i^0}-\boldsymbol{\beta}^0_j)+F^0_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i}\right)\right. \\ &\quad \left. +\frac{2}{T}\left(X_i(\boldsymbol{\beta}^0_{g_i^0}-\boldsymbol{\beta}^0_j)+F^0_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i}\right)' M_{F^0_j}\boldsymbol{\varepsilon}_i+\frac{1}{T}\boldsymbol{\varepsilon}_i' M_{F^0_j}\boldsymbol{\varepsilon}_i < \boldsymbol{\varepsilon}_i' M_{F^0_{g_i^0}}\boldsymbol{\varepsilon}_i + o_p(1)\right\}, \end{aligned}$$

where $o_p(1)$ is uniform in i and j, we have also used $p_{\kappa,\gamma}(|\hat{\boldsymbol{\beta}}_j|)=o_p(1)$. Thus

$$\mathbf{1}(\hat{g}_i \neq g_i^0) = \sum_{j=1;j\neq g_i^0}^{S} \mathbf{1}(A_{ij})$$

From the proof of Theorem 1, we also know that

$$P\left(\max_{i\in\{1,...,N\}}\left|\frac{1}{T}\boldsymbol{\varepsilon}_i' M_{F^0_j}\boldsymbol{\varepsilon}_i-\frac{1}{T}\boldsymbol{\varepsilon}_i' M_{F^0_{g_i^0}}\boldsymbol{\varepsilon}_i\right|>\delta\right)<\eta.$$

Now suppose that $g_i^0=k$, as long as $[X_i,F^0_k]$ has full column rank, which is necessary for identification of $\boldsymbol{\beta}^0_k$ anyway, then $X_i(\boldsymbol{\beta}^0_{g_i^0}-\boldsymbol{\beta}^0_j)+F^0_{g_i^0}\boldsymbol{\lambda}^0_{g_i^0,i}\neq 0$. Given the assumption of tail probability for $\boldsymbol{\varepsilon}_i$, and using the same argument in Theorem 2, we have for $j\neq g_i^0$,

$$P(A_{ij})\leq \eta + O(T^{-\tau}).$$

Since S is finite, this implies that

$$P(\hat{g}_i\neq g_i^0)\leq S\eta+O(T^{-\tau}),$$

where the right hand side is uniform in i. It follows that the average over i is also bounded by the above, that is

$$\frac{1}{N}\sum_{i=1}^{N}P(\hat{g}_i\neq g_i^0)=o(1)+O(T^{-\tau}).$$

Using the same argument as in Theorem 2, we can further show that

$$P\left(\sup_{i\in\{1,...,N\}}\mathbf{1}(\hat{g}_i\neq g_i^0)>0\right)=o(1)+NO(T^{-\tau}).$$

This completes the proof of Theorem 7. □

Proof of Theorem 8

The proof of Theorem 8 is almost same as that of Theorem 3. By Theorem 7, $P(\sup_i|\hat{g}_i-g_i^0|>0)=o(1)$ when $N/T^\tau\to 0$. This implies that $P(\hat{g}_1=g_1^0,\hat{g}_2=g_2^0,...,\hat{g}_N=g_N^0)\to 1$. Thus to prove Theorem 8, it is sufficient to assume that the group membership is known.

First, similar to the proof of Theorem 3, we can show

$$T^{-1/2}\|\tilde{F}_j - F_j^0 H_j\| = O_p\left(\sum_{j=1}^{S}\|\boldsymbol{\beta}_j^0 - \tilde{\boldsymbol{\beta}}_j\|\right) + O_p\left(\frac{1}{\min\{N_j^{1/2}, T^{1/2}\}}\right),$$

where $\tilde{\boldsymbol{\beta}}_j$ is the infeasible version of our estimator where group membership is fixed to its population G^0, $H_j^{-1} = V_{j,NT}(F_j^0\tilde{F}_j/T)^{-1}(\Lambda_j^{0\prime}\Lambda_j^0/N_j)^{-1}$.

We can prove the variable selection consistency by using the same argument in the proof of Theorem 3. For a simplicity of notation, we denote the non-zero true coefficient of j-th group as $\boldsymbol{\beta}_j^0$, and denote X_i as the corresponding columns of design matrix. Then the asymptotic normality part is proved as follows. We have

$$\begin{aligned}
&\frac{1}{N_jT}\left(\sum_{i:g_i^0=j} X_i M_{\tilde{F}_j} X_i + \Sigma_j(\kappa)\right)(\tilde{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^0) + \frac{1}{N_jT}\Sigma_j(\kappa)\boldsymbol{\beta}_j^0 \\
&= \frac{1}{N_jT}\sum_{i:g_i^0=j} X_i M_{\tilde{F}_j} F_j^0 \boldsymbol{\lambda}_{g_i^0,i}^0 + \frac{1}{N_jT}\sum_{i:g_i^0=j} X_i M_{\tilde{F}_j}\boldsymbol{\varepsilon}_i,
\end{aligned}$$

where $\Sigma_j(\kappa)$ is defined in Theorem 8. Using the same argument of Theorem 3, it then follows

$$\begin{aligned}
&\left[\frac{1}{N_jT}\sum_{i:g_i^0=j} X_i' M_{\tilde{F}_j} X_i - \frac{1}{TN_j}\sum_{i:g_i^0=j}\sum_{k:g_k^0=j} X_i' M_{\tilde{F}_j} X_k c_{j,ki} + \frac{1}{T}\Sigma_j(\kappa)\right](\tilde{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^0) \\
= \;& \frac{1}{N_jT}\sum_{i:g_i^0=j}\left[X_i' M_{\tilde{F}_j} - \frac{1}{N_j}\sum_{k:g_k^0=j} c_{j,ki} X_k' M_{\tilde{F}_j}\right]\boldsymbol{\varepsilon}_i \\
&+ \frac{1}{N_jT^2}\sum_{i:g_i^0=j}\sum_{k:g_k^0=j} X_i' M_{\tilde{F}_j}\Omega_k\tilde{F}_j(F_j^{0\prime}\tilde{F}_j/T)^{-1}(\Lambda_j^{0\prime}\Lambda_j^0/N_j)^{-1}\boldsymbol{\lambda}_{g_i^0,i}^0 \\
&+ o_p((N_jT)^{-1/2}) + \frac{1}{\sqrt{N}_j} O_p\left(\frac{1}{\min\{N_j, T\}}\right).
\end{aligned}$$

Similar to the proof of Theorem 3, replacing the $\tilde{F}_j$ by F_j^0 in the above formula leads

$$\begin{aligned}
\hat{D}(F_j^0,\kappa)\left[\sqrt{N_jT}(\tilde{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^0)\right] \;=\; & \frac{1}{\sqrt{N_jT}}\sum_{i:g_i^0=j}\left[X_i' M_{F_j^0} - \frac{1}{N_j}\sum_{k:g_k^0=j} c_{j,ki} X_k' M_{F_j^0}\right]\boldsymbol{\varepsilon}_i \\
&+ \sqrt{\frac{T}{N_j}}\boldsymbol{\eta}_j + \sqrt{\frac{N_j}{T}}\boldsymbol{\zeta}_j + o_p(1),
\end{aligned}$$

where $\boldsymbol{\zeta}_j$ and $\boldsymbol{\eta}_j$ are defined in Theorem 8. This leads to the limit of covariance matrix of $\sqrt{N_jT}(\tilde{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^0)$ given as $V_\beta(F_j^0) = D_0(F_j^0)^{-1}J_0(F_j^0)D_0(F_j^0)^{-1}$, where $D_0(F_j^0)$ and $J_0(F_j^0)$ are defined in Theorem 8. □

Proofs of Theorems 9 and 10

The proof of Theorem 9 is similar to that of Theorem 4. There are two steps. In step 1, we decompose the bias b as $b = B_1 + B_2 + B_3 + B_4 + B_5$, where

$$B_1 = E_y\Big[\frac{1}{NT}\sum_{i=1}^{N}\|\boldsymbol{y}_i - X_i\tilde{\boldsymbol{\beta}}_{g_i^0} - \tilde{F}_{g_i^0}\tilde{\boldsymbol{\lambda}}_{g_i^0,i}\|^2 - \frac{1}{NT}\sum_{i=1}^{N}\|\boldsymbol{y}_i - X_i\hat{\boldsymbol{\beta}}_{\hat{g}_i} - \hat{F}_{\hat{g}_i}\hat{\boldsymbol{\lambda}}_{\hat{g}_i,i}\|^2\Big],$$

$$B_2 = E_y\Big[\frac{1}{NT}\sum_{i=1}^{N}\|\boldsymbol{y}_i - X_i\boldsymbol{\beta}^0_{g_i^0} - \tilde{F}_{g_i^0}\tilde{\boldsymbol{\lambda}}_{g_i^0,i}\|^2 - \frac{1}{NT}\sum_{i=1}^{N}\|\boldsymbol{y}_i - X_i\tilde{\boldsymbol{\beta}}_{g_i^0} - \tilde{F}_{g_i^0}\tilde{\boldsymbol{\lambda}}_{g_i^0,i}\|^2\Big],$$

$$B_3 = E_y\Big[E_z[\frac{1}{NT}\sum_{i=1}^{N}\|\boldsymbol{z}_i - X_i\boldsymbol{\beta}^0_{g_i^0} - \tilde{F}_{g_i^0}\tilde{\boldsymbol{\lambda}}_{g_i^0,i}\|^2] - \frac{1}{NT}\sum_{i=1}^{N}\|\boldsymbol{y}_i - X_i\boldsymbol{\beta}^0_{g_i^0} - \tilde{F}_{g_i^0}\tilde{\boldsymbol{\lambda}}_{g_i^0,i}\|^2\Big]$$

$$B_4 = E_y\Big[E_z\Big[\frac{1}{NT}\sum_{i=1}^{N}\|\boldsymbol{z}_i - X_i\tilde{\boldsymbol{\beta}}_{g_i^0} - \tilde{F}_{g_i^0}\tilde{\boldsymbol{\lambda}}_{g_i^0,i}\|^2\Big] - E_z\Big[\frac{1}{NT}\sum_{i=1}^{N}\|\boldsymbol{z}_i - X_i\boldsymbol{\beta}^0_{g_i^0} - \tilde{F}_{g_i^0}\tilde{\boldsymbol{\lambda}}_{g_i^0,i}\|^2\Big]\Big],$$

$$B_5 = E_y\Big[E_z\Big[\frac{1}{NT}\sum_{i=1}^{N}\|\boldsymbol{z}_i - X_i\hat{\boldsymbol{\beta}}_{g_i} - \hat{F}_{\hat{g}_i}\hat{\boldsymbol{\lambda}}_{\hat{g}_i,i}\|^2\Big] - E_z\Big[\frac{1}{NT}\sum_{i=1}^{N}\|\boldsymbol{z}_i - X_i\tilde{\boldsymbol{\beta}}_{g_i} - \tilde{F}_{g_i^0}\tilde{\boldsymbol{\lambda}}_{g_i^0,i}\|^2\Big]\Big],$$

where the expectations $E_y[\cdot]$ and $E_z[\cdot]$ are taken with respect to the joint distribution of $\{\boldsymbol{y}_1, ..., \boldsymbol{y}_N\}$ and $\{\boldsymbol{z}_1, ..., \boldsymbol{z}_N\}$ conditioned on the design matrix X_i and the factor structure. Also, we denote

$$\ell_y(\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_S, F_1, ..., F_S, \Lambda_1, ..., \Lambda_S) = \frac{1}{NT}\sum_{i=1}^{N}\|\boldsymbol{y}_i - X_i\boldsymbol{\beta}_{g_i} - F_{g_i}\boldsymbol{\lambda}_{g_i,i}\|^2,$$

$$\ell_z(\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_S, F_1, ..., F_S, \Lambda_1, ..., \Lambda_S) = E_z\Big[\frac{1}{NT}\sum_{i=1}^{N}\|\boldsymbol{z}_i - X_i\boldsymbol{\beta}_{g_i} - F_{g_i}\boldsymbol{\lambda}_{g_i,i}\|^2\Big].$$

The same argument as in the proof of Theorem 4 applies. This gives Theorem 9. The proof of Theorem 10 is almost identical to that of Theorem 5. The details are omitted. This completes the proof. □

References

[1] Amengual, D., Watson, M. W. (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.

[2] Ando, T. and Bai, J. (2013). Multifactor asset pricing with a large number of observable risk factors and unobservable common and group-specific factors. Working paper. Available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2289201

[3] Ando, T. and Tsay, R. (2013). On model selection for large panel data with interactive effects. Working paper. Booth School of Business, University of Chicago.

[4] Arellano, M. (2003). Panel Data Econometrics. Oxford University Press.

[5] Arellano, M., and Hahn, J. (2005). Understanding Bias in Nonlinear Panel Models: Some Recent Developments. Invited Lecture, Econometric Society World Congress, London.

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

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

[8] Baltagi, B.H. (2008). Econometric Analysis of Panel Data Wiley, John & Sons.

[9] Bester, A. and Hansen, C. (2012). Grouped effects estimators in fixed effects models. *Journal of Econometrics*, Forthcoming.

[10] Bonhomme, S. and Manresa, E. (2012). Grouped patterns of heterogeneity in panel data. CEMFI, Working paper, 2012-1208.

[11] Chamberlain, G. and Rothschild, M. (1983). Arbitrage, factor structure and mean-variance analysis in large asset markets. *Econometrica* 51, 1305–1324.

[12] Chen, N. F., Roll, R., and Ross, S. (1986). Economic forces and the stock market. *Journal of Business*, 59, 383–403.

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

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

[15] Fama, E. F. (1981). Stock prices, real activity, inflation and money. *American Economic Review*, 71, 545–65.

[16] Fama, E. F. and French, K. R. (1989). Business conditions and expected returns on stocks and bonds. *Journal of Financial Economics*, 5, 23–49.

[17] Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33, 3–56.

[18] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. of the American Statistical Association* 96, 1348–1361.

[19] Forgy, E. W. (1965). Cluster analysis of multivariate data: Efficiency vs. Interpretability of classifications. *Biometrics*, 21, 768–769.

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

[21] Forni, M. and Lippi, M. (2001). The generalized factor model: representation theory. *Econometric Theory*, 17, 1113–1141.

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

[23] Hahn, J. and Kuersteiner, G. M. (2002). Asymptotically unbiased inference for a dynamic panel model with fixed effects when both N and T are large. *Econometrica*, 70, 1639–1657.

[24] Hahn, J. and Newey, W. K. (2004). Jackknife and analytical bias reduction for nonlinear panel models. *Econometrica*, 72, 1295–1319.

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

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

[27] Hsiao, C. (2003). Analysis of Panel Data, 2nd edition. Cambridge University Press

[28] Kapetanios, G., Pesaran, M.H. and Yamagata, T. (2011). Panels with nonstationary multifactor error structures. *Journal of Econometrics*, 160, 326–348.

[29] Kose, A., Otrok, C., Whiteman, C. (2008). Understanding the evolution of world business cycles. *International Economic Review*, 75, 110–130.

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

[31] Mallows, C. L. (1973). Some comments on *Cp*. *Technometrics* 15, 661–675.

[32] Moench, E. and Ng, S. (2011). A Factor Analysis of Housing Market Dynamics in the U.S. and the Regions. *Econometrics Journal*, 14, 1–24.

[33] Moench, E., Ng, S., Potter, S. (2012). Dynamic hierarchical factor models. *Review of Economics and Statistics*, forthcoming. Available at Staff Reports 412, Federal Reserve Bank of New York.

[34] Moon, H. R. and Weidner, M. (2009). Likelihood expansion for panel regression models with factors. Working Paper. Department of Economics, University of Southern California.

[35] Pesaran, M. H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica*, 74, 967–1012.

[36] Pesaran, M. H. and Tosetti, E. (2011). Large panels with common factors and spatial correlation *Journal of Econometrics*, 161, 182–202.

[37] Sargent, T. J., Sims, C. A. (1977). Business cycle modeling without pretending to have too much a priori economic theory. In: Sims C. et al. (eds), New Methods in Business Cycle Research. Federal Reserve Bank of Minneapolis, Minneapolis.

[38] Stock, J. H. and Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97, 1167–1179.

[39] Sun, Y. X. (2005). Estimation and inference in panel structure models. Working paper, Department of Economics, University of California, San Diego.

[40] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, B58, 267–288.

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

[42] Wooldridge, J.M. (2010) Econometric analysis of cross section and panel data, second edition. MIT Press.

[43] Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 38, 894–942.

Table 1: The percentages of under- (U), correct (C), and overidentification (O) of the various factor selections over 1,000 replicates for the data generated by the three data-generating processes considered in the text. The data shown are the number of selected groups S and the number of selected group-specific factors (r_j, $j = 1, .., 3$). The number of units in each group is $N_1 = N_2 = N_3 = N/3$.

Data 1

		S			r_1			r_2			r_3		
T	N	U	C	O	U	C	O	U	C	O	U	C	O
100	300	0	96	4	0	91	9	0	89	11	0	92	8
200	300	1	95	4	0	90	10	0	90	10	1	84	15
100	600	0	85	15	0	93	7	0	91	9	0	94	6
200	600	0	89	11	0	92	8	0	85	15	0	89	11

Data 2

		S			r_1			r_2			r_3		
T	N	U	C	O	U	C	O	U	C	O	U	C	O
100	300	0	87	13	0	89	11	0	86	14	0	87	13
200	300	0	85	15	0	84	16	0	84	16	0	85	15
100	600	0	83	17	0	87	13	0	88	12	0	85	15
200	600	0	84	16	0	85	15	0	85	15	0	87	13

Data 3

		S			r_1			r_2			r_3		
T	N	U	C	O	U	C	O	U	C	O	U	C	O
100	300	0	96	4	0	91	9	0	92	8	0	92	8
200	300	0	95	5	0	92	8	0	90	10	0	92	8
100	600	0	83	17	0	92	8	0	93	7	0	91	9
200	600	0	92	8	0	86	14	0	91	9	0	89	11

Table 2: Simulation results of the parameter estimates for $\hat{\boldsymbol{\beta}}$ based on 1,000 repetitions. We report the mean and standard deviation (Std.Dev.) of the parameter estimates. Because $\hat{\boldsymbol{\beta}}$ is a long vector (80×1), we report the estimation results only for the true predictors $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)'$ with true value $(\beta_1, \beta_2, \beta_3) = (1, 2, 3)'$, and for the first three irrelevant predictors, which are $(\hat{\beta}_4, \hat{\beta}_5, \hat{\beta}_6)'$ with true value $(\beta_4, \beta_5, \beta_6) = (0, 0, 0)'$. The remaining elements of $\hat{\boldsymbol{\beta}}$ are similar to $(\hat{\beta}_4, \hat{\beta}_5, \hat{\beta}_6)'$.

Data 1

T	N		$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$
100	300	Mean	0.986	1.980	2.970	0.000	0.000	0.000
		Std.Dev.	0.024	0.024	0.025	0.005	0.005	0.004
200	300	Mean	0.989	1.985	2.973	0.000	0.000	0.000
		Std.Dev.	0.016	0.017	0.017	0.003	0.003	0.004
100	600	Mean	0.995	1.990	2.984	0.000	0.000	0.000
		Std.Dev.	0.017	0.016	0.016	0.003	0.002	0.003
200	600	Mean	0.996	1.992	2.986	0.000	0.000	0.000
		Std.Dev.	0.012	0.011	0.011	0.002	0.001	0.001

Data 2

T	N		$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$
100	300	Mean	0.991	1.979	2.969	0.000	0.000	0.000
		Std.Dev.	0.024	0.023	0.024	0.006	0.004	0.005
200	300	Mean	0.991	1.980	2.970	0.000	0.000	0.000
		Std.Dev.	0.018	0.017	0.016	0.005	0.004	0.003
100	600	Mean	0.992	1.987	2.984	0.000	0.000	0.000
		Std.Dev.	0.017	0.016	0.016	0.004	0.005	0.004
200	600	Mean	0.994	1.989	2.985	0.000	0.000	0.000
		Std.Dev.	0.011	0.011	0.012	0.002	0.003	0.002

Data 3

T	N		$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$
100	300	Mean	0.988	1.980	2.967	0.000	0.000	0.000
		Std.Dev.	0.023	0.023	0.023	0.005	0.004	0.004
200	300	Mean	0.989	1.986	2.969	0.000	0.000	0.000
		Std.Dev.	0.016	0.017	0.016	0.004	0.003	0.003
100	600	Mean	0.990	1.987	2.980	0.000	0.000	0.000
		Std.Dev.	0.017	0.016	0.017	0.003	0.002	0.003
200	600	Mean	0.995	1.991	2.985	0.000	0.000	0.000
		Std.Dev.	0.011	0.012	0.011	0.001	0.001	0.001

Table 3: Scatter matrix of our grouping vs. the classification by mutual fund names (Small Capital & Growth, Large Capital & Growth, Small Capital & Value, and Large Capital & Value.)

	Our grouping					
Classification by names	G1	G2	G3	G4	G5	G6
Small Capital & Growth	64	19	0	14	0	50
Large Capital & Growth	68	7	42	5	0	0
Small Capital & Value	2	95	1	49	1	0
Large Capital & Value	1	0	5	5	108	0

Table 4: The correlations between the estimated group-specific pervasive factors and the Fama and French (1993) factors (Mkt, HML, SMB), Short-Term Reversal Factor (STR), Long-Term Reversal Factor (LTR), and Momentum Factor (Mom). If the absolute values of the correlations are larger than 0.18, 0.22, and 0.29, the corresponding significance levels are 10%, 5% and 1%, respectively.

		Observable 6 styles					
Group	Estimated factor	Mkt_t	SMB_t	HML_t	LTR_t	STR_t	Mom_t
G_1	First	0.43	-0.22	-0.28	-0.15	-0.17	0.20
	Second	-0.16	-0.07	0.07	0.19	0.18	0.02
	Third	0.10	-0.03	-0.14	-0.04	-0.07	0.15
	Fourth	-0.24	0.04	0.23	-0.21	0.19	-0.02
G_2	First	0.37	-0.08	-0.29	-0.08	0.39	0.11
	Second	0.05	-0.19	0.01	0.11	0.09	-0.08
	Third	-0.15	0.01	-0.15	0.16	-0.32	-0.17
G_3	First	0.46	-0.01	0.04	-0.18	0.02	-0.01
	Second	0.03	0.08	-0.03	0.13	0.10	-0.04
	Third	-0.11	-0.03	0.00	0.20	0.02	0.04
G_4	First	0.36	-0.09	0.18	-0.06	0.1	-0.26
	Second	0.07	0.03	-0.11	0.11	0.04	0.09
	Third	-0.13	0.16	-0.07	0.18	-0.16	0.17
G_5	First	0.46	-0.01	-0.02	0.06	0.02	-0.06
	Second	0.14	0.06	-0.13	0.04	0.13	0.14
G_6	First	0.33	0.14	-0.05	-0.19	-0.02	0.21
	Second	0.11	0.07	0.07	0.07	-0.11	0.00
	Third	-0.09	-0.03	0.00	0.26	0.01	0.10

Table 5: Scatter matrices of the estimated group membership $\hat{g}_i$ against nominal classification schemes based on 1. Location of stock exchanges, 2. Types of share, and 3. Industry.

	Classification	$G1$	$G2$	$G3$	$G4$	$G5$	$G6$
1	Location of stock exchanges						
	Shanghai stock exchange	179	67	132	77	105	81
	Shenzhen stock exchange	125	29	94	64	95	93
2	Types of share						
	A-shares	211	95	224	141	196	172
	B-shares	93	1	2	0	4	2
3	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 6: The results of regression of group-specific pervasive factors $\hat{f}_{jk,t}$ $(j = 1, ..., S; k = 1, ..., r_j)$ on some economic factors $\boldsymbol{z}_t$; $\hat{f}_{jk,t} = \boldsymbol{z}_t'\boldsymbol{\gamma}_{jk} + e_{jk,t}$, and then conduct the statistical significance test of the least squared estimate $\hat{\boldsymbol{\gamma}}_{jk}$. The four observable market 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. 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. (***), (**) and (*) means that the estimated regression coefficient is statistically significant at the 1%, 5%, and 10% levels, respectively.

		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	6.444***
	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

Table 7: Statistically significant regressors $\boldsymbol{x}_t$ for each group. (***), (**) and (*) means that the estimated regression coefficient is statistically significant at 1%, 5%, and 10% level, respectively.

Variables	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	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	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***