Clustering huge number of financial time series: A panel data approach with high-dimensional predictors and factor structures [1]

May 13, 2016

Tomohiro Ando and Jushan Bai

Abstract

This paper introduces a new procedure for clustering a large number of financial time series based on high-dimensional panel data with grouped factor structures. The proposed method attempts to capture the level of similarity of each of the time series based on sensitivity to observable factors as well as to the unobservable factor structure. The proposed method allows for correlations between observable and unobservable factors and also allows for cross-sectional and serial dependence and heteroskedasticities in the error structure, which are common in financial markets. In addition, theoretical properties are established for the procedure. We apply the method to analyze the returns for over 6,000 international stocks from over 100 financial markets. The empirical analysis quantifies the extent to which the U.S subprime crisis spilled over to the global financial markets. Furthermore, we find that nominal classifications based on either listed market, industry, country or region are insufficient to characterize the heterogeneity of the global financial markets.

Keywords: Clustering; Factor structure; Heterogeneous panel; Lasso; Serial and cross-sectional error correlations.

[1]Tomohiro Ando is Associate Professor, Melbourne Business School, University of Melbourne, Australia (E-mail: T.Ando@mbs.edu). Jushan Bai is Professor of Economics, Columbia University, and School of Finance, Nankai University (E-mail: jb3064@columbia.edu). Ando's research is supported by Research Grant from Melbourne Business School, and Bai's research is supported by the National Science Foundation (SES1357198).

1 Introduction

The U.S. subprime crisis of 2007, which was triggered by the collapse of the U.S. housing market, subsequently spilled over to the entire U.S. and the European financial markets, resulting in bankruptcies, forced mergers, and bailouts for many large firms.[2] These financial shocks further spread to the global financial markets, and led to massive declines in worldwide asset values.

Thus, identifying the sources of the co-movement of international stock returns is one of the most important issues in finance. Portfolio managers explore investment opportunities not only in applicable domestic markets but also in foreign financial markets. In the field of asset pricing, researchers have been searching for those factors that explain the cross-sectional variations in global stock returns (See, e.g., Griffin (2002), Hou. et al. (2011) and references therein). Previous studies that have attempted to identify the factors influencing the determination of international stock returns have arrived at mixed results. Fama and French (1998) emphasized a more globally integrated market, while Griffin (2002) argued that only local, country-specific factors matter in explaining global stock returns. This paper attempts to address these important questions by introducing a new statistical modeling procedure for building an empirical asset pricing model.

An internationally diversified portfolio often requires that the most influential factors in a variety of regions, countries, markets and industries be assessed and evaluated (See for e.g., Heston and Rouwenhorst (1994)). Thus, comparing a geographic diversification approach and its alternatives – such as country-, industry-, and market-based diversification strategies, to name a few – is also an important issue for international portfolio managers. Moreover, it is sometimes difficult to assign a nationality to a multinational company. In other words, industry classification can be a subjective undertaking, particularly for large conglomerates.

Motivated by these important issues, this study analyzes a large number of financial industry stock returns, i.e., over 6,000 returns from more than 100 international stock markets. In particular, we seek to answer the following empirical questions:
(1) How many groups are there among the large number of assets returns?
(2) In each group, how many group-specific common factors are there that explain the

[2]The firms affected include AIG, Bear Stearns, BNP Paribas, Fannie Mae, Freddie Mac, Lehman Brothers, and Merrill Lynch.

cross-sectional and time-series variations in stock returns?

(3) Do the co-movements within a market, industry, country, or region constitute the only sources of cross-sectional and time series variations in the stock markets?

(4) What are the different characteristics of the markets that can be observed during the recent financial crisis?

To address these questions, we introduce a new procedure for clustering a large number of financial time series based on high-dimensional panel data with grouped factor structures. Clustering is based on similarity measures using past stock returns. In particular, we introduce panel data models with heterogeneity, and these models have many attractive features. First, heterogeneity is captured by using a factor error structure and heterogeneous regression coefficients. Second, our method allows for a large number of observable factors, while the set of relevant observable factors are selected automatically. Third, observable factors can be correlated with unobservable factors or factor loadings or both. Fourth, the group membership of each unit is unknown, and these memberships will be estimated from historical stock returns. Finally, the number of groups remains unknown and is to be determined using a novel model selection criterion. We note that our asymptotic theory is developed for the optimal solution, which is obtained by minimizing the penalized objective function. In terms of computation in practice, the exact optimal solution can be time consuming due to the nature of the NP-hard problem. We therefore consider an algorithm that quickly searches approximate solutions.

Our empirical analysis indicates that the country-specific factor is one of the sources of co-movement in the cross-sectional and time-series variations of stock returns. This result is consistent with that of Fama and French (2012), who reported that global models fare poorly, while local versions of their three- and four-factor models for each of four regions – North America, Europe, Japan, and Asia Pacific – capture local average returns rather well.

The remainder of this paper is organized as follows. In the next section, we introduce asset pricing model and its assumptions. We also briefly review the related literature. Section 3 lays out the proposed statistical framework, discusses the parameter estimation, and explains how the best model is chosen to capture the underlying market structures. Section 4 establishes new asymptotic results, including the consistency of the proposed estimator and its asymptotic behaviors. A number of theoretical

results are established. Additional theoretical results are provided in the Appendix. The Appendix also contains simulation results that demonstrate that the proposed method works well. Section 5 describes the dataset. Empirical results are given in Section 6. Finally, Section 7 provides some concluding remarks.

2 Asset pricing model and related literature

2.1 Asset pricing model

Let $t = 1, ..., T$ be the time index and $i = 1, ..., N$ be the index of financial asset. Let S be the number of asset groups (which is unknown, finite and independent of N and T), and let $G = \{g_1, ..., g_N\}$ denote the group membership such that $g_i \in \{1, ..., S\}$. A distinctive feature of the model is that group membership is not specified. Let N_j be the number of cross-sectional assets within group G_j $(j = 1, ..., S)$ such that $N = \sum_{j=1}^{S} N_j$. To capture the underlying market characteristics, we assume that the return of the i-th asset, observed at time t, y_{it}, is expressed as

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

where $\boldsymbol{x}_{it}$ is a $p_i \times 1$ vector of observable factors, $\boldsymbol{f}_{c,t}$ is an $r \times 1$ vector of unobservable common factors that affect the returns of all securities in all groups, $\boldsymbol{\lambda}_{c,i}$ is the corresponding factor loading; $\boldsymbol{f}_{g_i,t}$ is an $r_{g_i} \times 1$ vector of unobservable group-specific factors that affect the returns only of asset group g_i, $\boldsymbol{\lambda}_{g_i,i}$ is the unknown sensitivity to unobservable group-specific factors, and ε_{it} is the asset-specific error. The $p_i \times 1$ vector $\boldsymbol{\beta}_i$ represents the unknown sensitivity to the explanatory variables (observable factors). It is assumed that ε_{it} is independent of $\boldsymbol{f}_{c,t}$, $\boldsymbol{f}_{g_i,t}$ and $\boldsymbol{x}_{it}$. Depending on applications, the unobserved factor components may be specified as an exact dynamic factor model, as a static approximate factor model, or as a special case of the generalized dynamic factor model. Technical assumptions are specified in Section 2.3. The factor structures in the model (1) considered here is similar to Hallin and Liška (2011), Wang (2010). However, the group membership here is unknown.

The unobservable term, $\boldsymbol{f}_{c,t}'\boldsymbol{\lambda}_{c,i} + \boldsymbol{f}_{g_i,t}'\boldsymbol{\lambda}_{g_i,i} + \varepsilon_{i,t}$, is typically treated as the overall error term that embodies cross-sectional/serially dependence and heteroskedasticity. However, ignoring the unobservable factor structure, $\boldsymbol{f}_{c,t}'\boldsymbol{\lambda}_{c,i} + \boldsymbol{f}_{g_i,t}'\boldsymbol{\lambda}_{g_i,i}$, in general, does not work due to the endogeneity problem (the regressors are correlated with the factors and factor loadings). Under such circumstances, the dependency between the

regressors and unobservable factor structures should be captured simultaneously. Our model building procedure takes this important issue into account.

When building the model, there are two issues to address. One is how to estimate the unknown parameters, including the regression coefficients $\{\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_N\}$, the unobservable factor structure $\{F_c, F_1, ..., F_S\}$, its corresponding loadings $\{\Lambda_c, \Lambda_1, ..., \Lambda_S\}$, and the unknown group membership of each of the N units. Here, $F_c = (\boldsymbol{f}_{c,1}, \boldsymbol{f}_{c,2}, ..., \boldsymbol{f}_{c,T})'$ is the $T \times r$ matrix of common factors, $\Lambda_c = (\boldsymbol{\lambda}_{c,1}, ..., \boldsymbol{\lambda}_{c,N})'$ is the $N \times r$ matrix of factor loadings for F_c, $F_j = (\boldsymbol{f}_{j,1}, \boldsymbol{f}_{j,2}, ..., \boldsymbol{f}_{j,T})'$ is the $T \times r_j$ matrix of factors for group G_j, and $\Lambda_j = (\boldsymbol{\lambda}_{j,1}, ..., \boldsymbol{\lambda}_{j,N_j})'$ is the $N_j \times r_j$ matrix of factor loadings for group G_j. These quantities can be determined under the given values of the number of groups S and the dimension of the unobservable factors structure. Given S and the dimension of factor structure for each group, our method jointly estimates the optimal grouping of the cross-sectional stocks, the regression coefficients, the factors and the factor loadings. This part of the modeling is called the estimation problem. To improve the speed of computation, the shrinkage method is incorporated into the estimation algorithm.

The other issue is determining the number of groups (S) and the dimension of factors in each group. We will also determine the relevant explanatory variables (observable factors). We refer to this part of the modeling as the model selection problem.

2.2 Related literature

Our modeling procedure and empirical investigation are related to several disparate strands in the statistics and finance literature. In terms of methodology, we propose a new modeling procedure for asset pricing. The proposed statistical inference procedure is a combination of high-dimensional grouped factor analysis and shrinkage methods. In particular, a large panel data model is considered for cases in which both the number of stocks, N, and the length of time periods, T, are large.

Factor models have attracted substantial research interest in recent years. In the econometric and statistical literature, a number of studies have been devoted to factor models that analyze high-dimensional data, including a dynamic exact factor model (Geweke, 1977; Sargent and Sims, 1977), a static approximate factor model (Chamberlain and Rothschild, 1983), and a generalized dynamic factor model (Forni et al., 2000; Forni and Lippi, 2001; Amengual and Watson, 2007; Hallin and Liska, 2007), among others.

Previous studies have considered panel data models with factor error structures to address cross-sectional and serial dependence, including Bai (2009) and Pesaran (2006). Bai (2009) estimated panel data models with interactive effects, permitting the predictor to be correlated with unobserved heterogeneity. These papers considered homogeneous regression coefficients over cross-sectional units, which is somewhat restrictive. Ando and Bai (2015) relaxed Bai (2009)'s model to allow for heterogeneous regression coefficients that vary for each stock. In addition, some previous studies featured "grouped" factor structures with known group membership, including Moench et al. (2012), Diebold et al. (2008), Kose et al. (2008), Hallin and Liška (2011), Moench and Ng (2011), and Wang (2010). Under "grouped" factor structures, each group is subject to its own unobservable factors that vary by group. In the context of asset pricing models, each asset group is exposed to unobservable factors, which are group-specific. Although these methods are useful under known group memberships, the group membership of each unit is often unknown,

Some studies consider unknown group memberships without factor structures, e.g., Lin and Ng (2012), Su et al., (2014) and Sun (2005). The model by Bonhomme and Manresa (2015) may be considered as a special factor model with a single factor and known factor loadings being one. Ando and Bai (2016) considered unknown group memberships with factor structure under the common or group-heterogeneous coefficients, in which the slope parameters are either the same or vary only across the groups. Although their procedure captures underlying market structure well, group-heterogeneous coefficients remain restrictive. In fact, these authors applied their method to the analysis of the two Chinese mainland stock markets – the Shanghai and Shenzhen stock exchanges – and found that group-heterogeneous coefficients are acceptable for some groups, whereas other groups show that group-heterogeneous coefficients are a too strong assumption. In this paper, we allow heterogeneous regression coefficients that vary for each stock.

Recently, various types of shrinkage methods have been proposed, including the lasso method (Tibshirani, 1996) and its variants (Zou, 2006; Yuan and Lin, 2006, Park and Casella, 2008), least-angle regression (Efron et al., 2004), elastic net (Zou and Hastie, 2005), the smoothly clipped absolute deviation approach (SCAD; Fan and Li, 2001), the minimax concave penalty method (MCP; Zhang 2010), and the Dantzig selector (Candes and Tao, 2007), among many others. As with these studies, we aim

to select an appropriate set of observable factors among the huge number of possible variables. More specifically, we use the smoothly clipped absolute deviation (SCAD) penalty approach (Fan and Li, 2001). Thus, the non-zero coefficients are estimated as if the zero coefficients were known and were imposed (the so called "oracle property'). This result is obtained despite the existence of many unobservable factors.

In empirical finance, identifying the sources of international stock returns' co-movements is of central importance. Some studies are based on factors that can explain the cross-sectional variation in global stock returns (See for e.g., Griffin (2002), Hou et al. (2011) and references therein). There is a substantial body of literature that has attempted to identify the influential factors that determine international stock returns, but these studies have yielded only mixed results. Griffin (2002) argues that country-specific factors are important to explaining global stock returns and voiced doubts about the benefits of extending the Fama and French (1993) three-factor model to a global context. By contrast, Fama and French (1998) demonstrated the applicability of the global version of multifactor models. In regards to the influences of country and industry factors, a number of studies emphasize the dominance of the country factor over the industry factor (Heston and Rouwenhorst (1994, 1995), Beckers et al. (1996), Griffin and Karolyi (1998), Kuo and Satchell (2001)), whereas Roll (1992) reported that industry factors are the most important. Baca et al. (2000) and Cavaglia et al. (2000) argued that the relative influence of the country factor and the industry factor depend on the time period. In this paper, we investigate this important issue by analyzing the impact of the U.S. financial crisis on international financial markets.

Our theoretical and empirical contributions are summarized as follows. First, the model to be introduced in the next section is new and very general. Under unknown group membership, the model allows heterogeneous regression coefficients that vary with each stock (asset-dependence coefficients). Moreover, the number of regressors can increase as the size of the panel increases. In the context of a cross sectional regression, Fan and Peng (2004) and Lam and Fan (2008) considered the case of increasing number of regressors. However, this is the first study to consider a divergent number of regressors under the panel data models with a grouped factor structure.

Second, a number of theoretical results – including consistency, asymptotic normality, oracle property, and model selection consistency – are established. Because of the more general model structure, establishing the inferential theory requires non-

trivial arguments. Although Ando and Bai (2015) and Ando and Bai (2014) considered heterogeneous regression coefficients, they assumed that group membership is known and also that the number of regressors is fixed. In contrast, the group membership is unknown in this paper. While Ando and Bai (2016) considered unknown group membership, the regression coefficients in their model only vary over the groups, and the number of regressors is fixed. Here in this paper, the group membership is unknown, the regression coefficients are asset dependent, and the number of regressors can increase with the sample size. We provide a theoretical analysis for these results. To our knowledge, this is the first study that investigates a divergent number of regressors for "heterogeneous" regression coefficients in panel data with a factor structure.

This paper also makes an empirical contribution in analyzing the recent U.S. financial crisis (Longstaff (2010), Diebold and Yilmaz (2014) and so on). The results provide insightful information on the grouping of financial assets and its evolution prior to and during the financial crisis.

2.3 Assumptions

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

The true regression coefficient is denoted by $\boldsymbol{\beta}_i^0$. Further, $F_c^0 = (\boldsymbol{f}_{c,1}^0, ..., \boldsymbol{f}_{c,T}^0)'$ and $\boldsymbol{\lambda}_{c,i}^0$ are the true common factor and its factor loading of individual i, and $F_{g_i^0}^0 = (\boldsymbol{f}_{g_i^0,1}^0, ..., \boldsymbol{f}_{g_i^0,T}^0)'$ and $\boldsymbol{\lambda}_{g_i^0,i}^0$ are the true factor and factor loading of individual i with true group membership g_i^0.

Assumption A: Common and group-specific factors

The common factors satisfy $E\|\boldsymbol{f}_{c,t}^0\|^4 < \infty$ and $T^{-1}\sum_{t=1}^T \boldsymbol{f}_{c,t}^0 \boldsymbol{f}_{c,t}^{0\,\prime} \to \Sigma_{F_c}$ as $T \to \infty$, where Σ_{F_c} is an $r \times r$ positive definite matrix. The group-specific factors satisfy $E\|\boldsymbol{f}_{j,t}^0\|^4 < \infty$ $j = 1, ..., S$. Furthermore, $T^{-1}\sum_{t=1}^T \boldsymbol{f}_{j,t}^0 \boldsymbol{f}_{j,t}^{0\,\prime} \to \Sigma_{F_j}$ as $T \to \infty$, where Σ_{F_j} is an $r_j \times r_j$ positive definite matrix. Although correlations between $\boldsymbol{f}_{j,t}^0$ and $\boldsymbol{f}_{k,t}^0$

$(j \neq k)$ are allowed, they are not perfectly correlated. Also, we assume orthogonality between the common and group-specific factors $\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{f}_{c,t}^0 \boldsymbol{f}_{j,t}^{0\prime} = 0$ for $j = 1, ..., S$.

Assumption B: Factor loadings

(B1): The factor loading matrix for the common factors $\Lambda_c^0 = [\boldsymbol{\lambda}_{c,1}^0, \ldots, \boldsymbol{\lambda}_{c,N}^0]'$ satisfies $E\|\boldsymbol{\lambda}_{c,i}^{0\ 4}\| < \infty$ and $\|N^{-1}\Lambda_c^{0\prime}\Lambda_c^0 - \Sigma_{\Lambda_c}\| \to \mathbf{0}$ as $N \to \infty$, where Σ_{Λ_c} is an $r \times r$ positive definite matrix. The factor loading matrix for the group-specific factors $\Lambda_j^0 = [\boldsymbol{\lambda}_{j,1}^0, \ldots, \boldsymbol{\lambda}_{j,N_j}^0]'$ satisfies $E\|\boldsymbol{\lambda}_{g_i,i}^{0\ 4}\| < \infty$ and $\|N_j^{-1}\Lambda_j^{0\prime}\Lambda_j^0 - \Sigma_{\Lambda_j}\| \to \mathbf{0}$ 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 $\|\boldsymbol{\lambda}_{g_i^0,i}^0\| > 0$.

(B2): For each i and j, $\boldsymbol{f}_{j,t}^{0\ \prime}\boldsymbol{\lambda}_{g_i^0,i}^0$ is strongly mixing processes with mixing coefficients that satisfy $r(t) \leq \exp(-a_1 t^{b_1})$ and with tail probability $P(|\boldsymbol{f}_{j,t}^{0\ \prime}\boldsymbol{\lambda}_{g_i,i}^0| > z) \leq \exp\{1 - (z/b_2)^{a_2}\}$, where a_1, a_2, b_1 and b_2 are positive constants.

Assumption C: Error terms

(C1): $E[\varepsilon_{it}] = 0$, $\text{var}(\varepsilon_{it}) = \sigma_i^2$, and ε_{it} is independent over i and over t.

(C2): A positive constant, $C < \infty$, exists such that $E[|\varepsilon_{it}|^8] < C$ for all i and t.

(C3): ε_{it} is independent of $\boldsymbol{x}_{ks}$, $\boldsymbol{\lambda}_{c,\ell}^0$, $\boldsymbol{\lambda}_{g,\ell}^0$, $\boldsymbol{f}_{c,s}^0$ and $\boldsymbol{f}_{j,s}^0$ $(j = 1, ..., S)$ for all i, k, ℓ, t, s.

Assumption D: Observable factors

(D1): 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 $N/T^2 \to 0$.

(D2): Define $M_{F_c,F_j} = I - F_c'(F_c'F_c)^{-1}F_c - F_j'(F_j'F_j)^{-1}F_j$. Let $X_{i,\beta_i^0\neq 0}$ be the submatrix of X_i, corresponding to the columns of nonzero elements of the true parameter vector $\boldsymbol{\beta}_i^0$. We use q_i to denote the number of nonzero elements of $\boldsymbol{\beta}_i^0$. Suppose that the i-th financial asset belongs to the g-th group (i.e., $g_i^0 = g$). We assume the $q_i \times q_i$ matrix

$$\frac{1}{T}\left[X_{i,\beta_i^0\neq 0}' M_{F_c^0,F_g^0} X_{i,\beta_i^0\neq 0}\right]$$

is positive definite.

(D3): Define $A_i = \frac{1}{T}X_i' M_{F_c,F_{g_i}} X_i$, $C_i = (C_{ci}, C_{gi})$,

$$B_i = \begin{pmatrix} B_{ci} & B_{cgi} \\ B_{cgi}' & B_{gi} \end{pmatrix},$$

with

$$B_{ci} = (\boldsymbol{\lambda}_{c,i}^0 \boldsymbol{\lambda}_{c,i}^{0\prime}) \otimes I,\ B_{gi} = (\boldsymbol{\lambda}_{g_i^0,i}^0 \boldsymbol{\lambda}_{g_i^0,i}^{0\prime}) \otimes I,\ B_{cgi} = (\boldsymbol{\lambda}_{c,i}^0 \boldsymbol{\lambda}_{g_i^0,i}^{0\prime}) \otimes I,$$
$$C_{ci} = \frac{1}{\sqrt{T}} \boldsymbol{\lambda}_{c,i}^{0\prime} \otimes X_i' M_{F_c,F_{g_i}},\ C_{gi} = \frac{1}{\sqrt{T}} \boldsymbol{\lambda}_{g_i^0,i}^{0\prime} \otimes X_i' M_{F_c,F_{g_i}}.$$

Let $\mathcal{A}$ be the collection of (F_c, F_g) such that $\mathcal{A} = \{(F_c, F_g) : F_c'F_c/T = I,\ F_g'F_g/T = I\}$. We assume, for $j = 1, ..., S$,

$$\inf_{F_c,F_j \in \mathcal{A}} \left[\frac{1}{N} \sum_{i;g_i^0=j} E_i(F_c, F_j) \right] \quad \text{is positive definite,}$$

where $E_i(F_c, F_j) = B_i - C_i' A_i^- C_i$ and A_i^- is a generalized inverse of A_i.

Assumption E: Number of units in each group

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

Some comments on the assumptions are provided. Assumptions A and B are usual and imply the existence of r common factors and r_j group-specific factors, $j = 1, ..., S$. The last part of the assumption A assumes that the common factors $\boldsymbol{f}_{c,t}^0$ and the group-specific factors $\boldsymbol{f}_{g,t}^0$ are orthogonal. This assumption is needed to separately identify the common and the group-specific factors (Wang (2010)). In Assumption C, heteroskedasticity is allowed. Although it is outside the scope of this paper, the errors are also allowed to have cross-sectional correlation, serial correlation, or both. This allows us to address various types of dependency. However, it will require more technical conditions such as those in Bai (2009), thus its discussion is omitted. Assumption D1 requires some moment conditions on the observable factors. The observable factors can be correlated with group-specific factors, factor loadings or both. The number of cross-sectional units N can be much greater than the number of time periods T. In practice, most of the applications are carried out in the case of $N > T$, as the number of assets N is much larger than the length of the time series. However, N should grow less than T^2 and especially $N = O(\exp(T))$ is not allowed. The true number of groups, S, is assumed to be finite and independent of N and T. Assumption D2 is analogous to the full rank condition in standard linear regression models and that is made in Ando and Bai (2015). Assumption D3 is similar to a condition used in Bai (2009), where only

a single group exists. The assumption is used for proof of consistency when factor and factor loadings are also estimated. Note that if $A_i = 0$, then C_i must be zero because $X_i' M_{F_c,F_g} = 0$. As a result, the term $C_i' A_i^- C_i$ becomes $C_i' A_i^- C_i = 0$. Thus, $C_i' A_i^- C_i$ is well defined even if $A_i = 0$. This assumption is also used in Ando and Bai (2015).

3 Model building

In this section, we describe our modeling framework, which involves identifying the number of groups and group-specific factors, and estimating model parameters from a large panel data. The goal of our procedure is to identify the underlying market structure.

3.1 Estimation

Given the number of groups S, the number of common factors r, the number of factors in group r_j $(j = 1, 2, ..., S)$, and the size of penalty κ_i in $p_i(\beta_i) = p_{\kappa_i,\gamma}(\boldsymbol{\beta}_i)$, the estimator $\{\hat{\boldsymbol{\beta}}_1, ..., \hat{\boldsymbol{\beta}}_N, \hat{G}, \hat{F}_c, \hat{F}_1, ..., \hat{F}_S, \hat{\Lambda}_c, \hat{\Lambda}_1, ..., \hat{\Lambda}_S\}$ is defined as the minimizer of

$$\begin{aligned} & L(\boldsymbol{\beta}_1, .., \boldsymbol{\beta}_N, G, F_c, F_1, ..., F_S, \Lambda_c, \Lambda_1, ..., \Lambda_S) \\ = & \sum_{i=1}^{N} \|\boldsymbol{y}_i - X_i \boldsymbol{\beta}_i - F_c \boldsymbol{\lambda}_{c,i} - F_{g_i} \boldsymbol{\lambda}_{g_i,i}\|^2 + T \sum_{i=1}^{N} p_i(\boldsymbol{\beta}_i), \end{aligned} \tag{2}$$

subject to normalization restrictions on F_c and F_j $(j = 1, ..., S)$, to be discussed below. Here, $\Lambda_c = (\boldsymbol{\lambda}_{c,1}, ..., \boldsymbol{\lambda}_{c,N})'$ is the $N \times r$ factor loading matrix for the common factors F_j, $\Lambda_j = (\boldsymbol{\lambda}_{j,1}, ..., \boldsymbol{\lambda}_{j,N_j})'$ is the $N_j \times r_j$ factor loading matrix $(j = 1, ..., S)$ for the group-specific factors F_j (Connor and Korajzcyk (1986), Stock and Watson (2002), Bai and Ng (2002)). The first term is the fitness term and the second term is the penalty.

For the penalty function $p_i(\boldsymbol{\beta}_i)$ in (2), we can consider the ridge penalty, the lasso penalty (Tibshirani, 1996) and its variants (Zou, 2006; Yuan and Lin, 2006), the elastic net (Zou and Hastie, 2005), the minimax concave penalty (Zhang 2010) the SCAD penalty of Fan and Li (2001), and so on. To explain the estimation procedure, we here use the SCAD penalty of Fan and Li (2001), which is given by

$$p_i(\boldsymbol{\beta}_i) \equiv p_{\kappa_i,\gamma}(\boldsymbol{\beta}_i) = \sum_{j=1}^{p_j} p_{\kappa_i,\gamma}(|\beta_{i,j}|)$$

with

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

for $\kappa_i > 0$ and $\gamma > 2$. Following Fan and Li (2001), we use the value $\gamma = 3.7$, which minimizes a Bayesian risk criteria for the regression coefficients. The regularization parameter κ_i controls the size of the penalty and varies over the cross-sectional asset. In contrast to the previous studies where the common regularization parameter is employed for each of the cross-sectional units (for e.g., Ando and Bai (2015)), this paper allows the flexibility to the regularization parameter. This flexibility makes sense because some assets may be subject to a small number of factors, whereas the other group may be influenced by a large number of factors. To our best knowledge, this is the first panel study that allows the regularization parameter to vary over the cross-sectional units. Later in this section, we provide an algorithm that searches the best values of the regularization parameters $\{\kappa_1, ..., \kappa_N\}$ over a pre-specified candidate space.

To obtain the minimizer of $L(\boldsymbol{\beta}_1, .., \boldsymbol{\beta}_N, G, F_c, F_1, ..., F_S, \Lambda_c, \Lambda_1, ..., \Lambda_S)$, we can use an iterative scheme. Given the group membership G, the common factor structures $F_c\boldsymbol{\lambda}_{c,i}$, and the group-specific factor structures $F_j\boldsymbol{\lambda}_{g_i,i}$, we define the variable $\boldsymbol{y}_i^* = \boldsymbol{y}_i - F_c\boldsymbol{\lambda}_{c,i} - F_{g_i}\boldsymbol{\lambda}_{g_i,i}$ for $i = 1, ..., N$. Then, the objective function for $\boldsymbol{\beta}_i$ can be viewed as $\|\boldsymbol{y}_i^* - X_i\boldsymbol{\beta}_i\|^2 + Tp_{\kappa_i,\gamma}(\boldsymbol{\beta}_i)$. Thus, the estimator of $\boldsymbol{\beta}_i$ can be obtained by the SCAD approach.

Given the group membership G, the common factor structures $F_c\boldsymbol{\lambda}_{c,i}$, and the value of the regression coefficient $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_N$, we define the variable $Z_j = (\boldsymbol{z}_{j,1}, \ldots, \boldsymbol{z}_{j,N_j})$ with $\boldsymbol{z}_{j,i} = \boldsymbol{y}_i - X_i\boldsymbol{\beta}_i - F_c\boldsymbol{\lambda}_{c,i}$ for $g_i = j$. Then, model (1) reduces to $\boldsymbol{z}_{j,i} = F_j\boldsymbol{\lambda}_{g_i,i} + \boldsymbol{\varepsilon}_i$. Because this implies that matrix Z_j has a pure factor structure, we can use the previously established factor analysis methods. We can obtain the principal components' estimate of F_j, subject to the normalization $F_j'F_j/T = I_{r_j}$, is $\sqrt{T}$ times the eigenvectors corresponding to the r_j largest eigenvalues of the $T \times T$ matrix $Z_j'Z_j$. Given $\hat{F}_j$, the factor loading matrix can be obtained as $\hat{\Lambda}_j = Z_j\hat{F}_j/T$. See also Bai and Ng (2002, pp197∼198), Connor and Korajzcyk (1986) and Stock and Watson (2002).

Given the group membership G, the group-specific factor structures $F_{g_i}\boldsymbol{\lambda}_{g_i,i}$, and the value of the regression coefficient $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_N$, we define the variable $Z_c = (\boldsymbol{z}_{c,1}, \ldots, \boldsymbol{z}_{c,N})$

with $\boldsymbol{z}_{c,i} = \boldsymbol{y}_i - X_i\boldsymbol{\beta}_i - F_{g_i}\boldsymbol{\lambda}_{g_i,i}$ for $g_i = j$. The model (1) reduces to $\boldsymbol{z}_{c,i} = F_c\boldsymbol{\lambda}_{c,i} + \boldsymbol{\varepsilon}_i$. The principal components' estimate of F_c subject to the normalization $F_c'F_c/T = I_r$, is $\sqrt{T}$ times the eigenvectors corresponding to the r largest eigenvalues of the $T \times T$ matrix $Z_c'Z_c$. Given $\hat{F}_c$, the factor loading matrix can be obtained as $\hat{\Lambda}_c = Z_c\hat{F}_c/T$.

For any given values of $\boldsymbol{\beta}_i$, $F_c\boldsymbol{\lambda}_{c,i}$ and $F_{g_i}\boldsymbol{\lambda}_{g_i,i}$ $(j = 1, ..., S)$, the optimal assignment for each individual unit is given as $g_i^* = \text{argmin}_{j\in\{1,...,S\}}\|\boldsymbol{y}_i - X_i\boldsymbol{\beta}_i - F_c\boldsymbol{\lambda}_{c,i} - F_j\boldsymbol{\lambda}_{j,i}\|^2$. The final estimated individual membership satisfies

$$\hat{g}_i = \text{argmin}_{j\in\{1,...,S\}}\|\boldsymbol{y}_i - X_i\hat{\boldsymbol{\beta}}_i - \hat{F}_c\hat{\boldsymbol{\lambda}}_{c,i} - \hat{F}_j\hat{\boldsymbol{\lambda}}_{j,i}\|^2, \tag{3}$$

which minimizes the sum of squared residuals among the S possible groups. Here $\hat{\boldsymbol{\lambda}}_{j,i} = \hat{F}_j'(\boldsymbol{y}_i - X_i\hat{\boldsymbol{\beta}}_i - \hat{F}_c\hat{\boldsymbol{\lambda}}_{c,i})/T$.

Because the estimates of $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_N$, $\{F_c, \Lambda_c\}$, $\{F_j, \Lambda_j;\ j = 1, ..., S\}$, and $G \in \{g_1, ..., g_N\}$ depend on one another, we update the set of parameters sequentially. Moreover, we have to employ this strategy to capture the dependency between the regressors and unobservable factor structures simultaneously due to the endogeneity problem.

Estimation algorithm

Step 1. Fix κ_1,...,κ_N, r, $\{r_1, ..., r_S\}$ and S. Initialize the unknown parameters $\boldsymbol{\beta}_1^{(0)}, ..., \boldsymbol{\beta}_N^{(0)}$, $\{F_c^{(0)}, \Lambda_c^{(0)}\}$, $\{F_j^{(0)}, \Lambda_j^{(0)};\ j = 1, ..., S\}$, $G^{(0)} \in \{g_1^{(0)}, ..., G_N^{(0)}\}$.

Step 2. Given the values of $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_N$, $\{F_c, \Lambda_c\}$, and $\{F_j, \Lambda_j;\ j = 1, ..., S\}$, update g_i for $i = 1, ..., N$ based on (3).

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

Step 4. Given the values of $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_N$, $\{F_j, \Lambda_j;\ j = 1, ..., S\}$ and G, update $\{F_c, \Lambda_c\}$.

Step 5. Given the values of $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_N$, $\{F_c, \Lambda_c\}$ and G, update $\{F_j, \Lambda_j\}$ for $j = 1, ..., S$.

Step 6. Repeat Steps 2 $\sim$ 5 until convergence. Then we obtain the estimators $\{\hat{\boldsymbol{\beta}}_1, .., \hat{\boldsymbol{\beta}}_N, \hat{G}, \hat{F}_c, \hat{F}_1, ..., \hat{F}_S, \hat{\Lambda}_c, \hat{\Lambda}_1, ..., \hat{\Lambda}_S\}$.

Remark 1 In Step 1, starting values are needed. To obtain initial group membership $G^{(0)}$, we use the well-known K-means algorithm (Forgy (1965)) that divides the data set $\{\boldsymbol{y}_i;\ i = 1, ..., N\}$ into S clusters that correspond to the number of groups. An initial estimate of $\boldsymbol{\beta}_i^{(0)}$ $(i = 1, ..., N)$ is obtained via the SCAD approach by ignoring the factor structures. Given $G^{(0)}$ and $\boldsymbol{\beta}_i^{(0)}$ $(i = 1, ..., N)$, the common factor structure

$\{F_c^{(0)}, \Lambda_c^{(0)}\}$ is then obtained. Finally, given the values of $\boldsymbol{\beta}^{(0)}$, $\{F_c^{(0)}, \Lambda_c^{(0)}\}$ and $G^{(0)}$, we obtain the starting values $\{F_j^{(0)}, \Lambda_j^{(0)}\}$ for $j = 1, ..., S$ by the principal components.

Remark 2 Our asymptotic theory is developed for the optimal solution $\{\hat{\boldsymbol{\beta}}_1, .., \hat{\boldsymbol{\beta}}_N, \hat{G}, \hat{F}_c, \hat{F}_1, ..., \hat{F}_S, \hat{\Lambda}_c, \hat{\Lambda}_1, ..., \hat{\Lambda}_S\}$. In terms of computation in practice, the exact optimal solution can be time consuming because the possible combinations of the group membership G is large. Approximate solutions are relatively quick to obtain. Our computation is approximate, as is mostly done in practice for clustering analysis. Group membership G is updated sequentially in the algorithm instead of brute-force enumeration.

Remark 3 In a simulation study, we also consider the standard lasso

$$p_{\kappa_i}(\beta_{i,j}) = \kappa_i |\beta_{i,j}|$$

and the minimax concave penalty (MCP)

$$p_{\kappa_i,\gamma}(\beta_{i,j}) = \begin{cases} \kappa_i \beta_{i,j} - \beta_{i,j}^2/(2\gamma) & (\beta_{i,j} \leq \gamma\kappa_i) \\ \kappa_i^2 \gamma/2 & (\beta_{i,j} > \gamma\kappa_i) \end{cases}$$

for $\gamma > 1$. Breheny and Huang (2011) suggested that MCP and SCAD are worthwhile alternatives to the lasso. The supplemental document contains more details for comparison.

Remark 4 As discussed in Bai and Ng (2002, p.198), we can consider either one of the following two procedures when estimating the factor structure. To explain the idea, we focus on extracting the common factor structure from the matrix Z_c. In procedure 1, we first estimate F_c from the $T \times T$ matrix $Z_c' Z_c$ to obtain $\hat{F}_c$ subject to the normalization of $F_c' F_c/T = I$. Then, the corresponding factor loading is obtained $\hat{\Lambda}_c = Z_c \hat{F}_c/T$. As an alternative procedure, we can first extract eigenvectors of the $N \times N$ matrix $Z_c Z_c'$ to obtain the estimate $\tilde{\Lambda}_c$ subject to the normalization of $\Lambda_c' \Lambda_c/N = I$. This normalization implies $\tilde{F}_c = Z_c' \tilde{\Lambda}_c/N$. Even when $N > T$, the spiked eigenvalues and related eigenvectors are consistently estimated (Fan et al., 2013). As suggested by Bai and Ng (2002, p.198), the first procedure is computationally less intensive when $T < N$, which is the case in our application. The second procedure is preferred when $N < T$ because the computation is less costly. Thus, the size of panel will be useful in determining the procedure for extracting factor structures. Either procedure produces the same common components (the multiplication of factor and factor loadings).

3.2 Model selection

In practice, however, the number of groups, S, and the number of common factors, r, the number of group-specific factors, $\{r_1, ..., r_S\}$, are unknown. Moreover, we have to select the size of the regularization parameters such that the relevant observable factors are included, while excluding irrelevant observable factors. We propose a new criterion to select these quantities.

$$\begin{aligned} & \mathrm{PIC}^C(S, k, k_1, ..., k_S, \kappa_1, ..., \kappa_N) \\ = \; & \frac{1}{NT} \sum_{j=1}^{S} \sum_{i;g_i=j} \left\| \boldsymbol{y}_i - X_i \hat{\boldsymbol{\beta}}_i - \hat{F}_c \hat{\boldsymbol{\lambda}}_{c,i} - \hat{F}_{\hat{g}_i} \hat{\boldsymbol{\lambda}}_{\hat{g}_i,i} \right\|^2 + C \times \frac{1}{N} \sum_{i=1}^{N} \hat{\sigma}^2 \log(T) \hat{p}_i \\ & + C \times k \times \hat{\sigma}^2 \left(\frac{T+N}{TN} \right) \log(TN) + \sum_{j=1}^{G} C \times k_j \times \hat{\sigma}^2 \left(\frac{T+N_j}{TN_j} \right) \log(TN_j) \quad (4) \end{aligned}$$

where $\hat{p}_i$ is the number of non-zero elements of $\hat{\boldsymbol{\beta}}_i$, C is some positive constant and $\hat{\sigma}^2$ is an estimate of $\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} E(\varepsilon_{it}^2)$. Note that the effect of the regularization parameters κ_i $(i = 1, ..., N)$ is measured through $\hat{p}_i$. Too large of a regularization parameter will lead $\hat{p}_i = 0$ for $i = 1, ..., N$, while too small of a regularization parameter causes an over-fitting problem. In the next section, we show that our proposed panel information criterion, PIC, can select the true non-zero regression coefficients. By minimizing PIC, we can choose the number of groups S, the number of common factors k, the number of group-specific factors k_j $(j = 1, ..., S)$, and the size of the regularization parameters κ_1,...,κ_N.

The criterion has incorporated the procedure of Hallin and Liška (2007), i.e., the criterion (4) has the theoretical property that a penalty function on the number of common factors $k \times \hat{\sigma}^2 (\frac{T+N}{TN}) \log(TN)$ leads to a consistent estimate of the true number of common factors, r, even when multiplying the penalty by some positive constant C. However, for given finite N and T, the value of C affects the model selection result. A very large value of C over penalizes the number of common factors and vise versa. Similar arguments also apply to the penalty functions on the dimension of regression coefficients as well as to the number of group-specific factors.

To optimize the value of C, we use the suggestion of Hallin and Liška (2007) and Alessi et al. (2010). We investigate the asymptotic behavior of the selected number of groups S, number of common factors r, and number of group-specific factors $k_1, ..., k_S$, from considering subsamples of sizes $(N^{(a)}, T^{(a)})$ with $a = 1, ..., A$ such that $0 < N^{(1)} <$

$N^{(2)} < \cdots < N^{(A)} = N$ and $0 < T^{(1)} < T^{(2)} < \cdots < T^{(A)} = T$ For any $(N^{(a)}, T^{(a)})$ and C, we can compute the number of groups $S^C(N^{(a)}, T^{(a)})$, number of common factors $r^C(N^{(a)}, T^{(a)})$, and number of group-specific factors $k_1^C(N^{(a)}, T^{(a)}), ..., k_S^C(N^{(a)}, T^{(a)})$. Hallin and Liška (2007) pointed out that between the extreme small values and too large values, there exist a range of moderate values of C such that the selected model is a stable function of the subsample size $(N^{(a)}, T^{(a)})$. They measured the stability with respect to sample size by the empirical variance of the selected values of $S^C(N^{(a)}, T^{(a)})$, $r^C(N^{(a)}, T^{(a)})$, and $k_1^C(N^{(a)}, T^{(a)}), ..., k_S^C(N^{(a)}, T^{(a)})$. In our case, it is measured by

$$\begin{aligned} V_C^2 &= \frac{1}{A}\sum_{a=1}^{A}\left(r^C(N^{(a)}, T^{(a)}) - A^{-1}\sum_{b=1}^{A} r^C(N^{(b)}, T^{(b)})\right)^2 \\ &\quad + \sum_{j=1}^{S_{\max}}\left[\frac{1}{A}\sum_{a=1}^{A}\left(r_j^C(N^{(a)}, T^{(a)}) - A^{-1}\sum_{b=1}^{A} r_j^C(N^{(b)}, T^{(b)})\right)^2\right], \end{aligned} \tag{5}$$

where the first term measures the variability of selected common factors, and the second term measures the variability of selected number of groups as well as the number of group-specific factors.

Under the given value of C, $N^{(a)}$, $T^{(a)}$, the following provides a model search algorithm.

Model search algorithm

Step 1. Prepare the candidate values of regularization parameters $\{\kappa_1, ..., \kappa_N\}$, the number of groups S and the numbers of group-specific factors $\{k_1, ..., k_S\}$.

Step 2. Fix S and initialize the values of regularization parameters $\{\kappa_1, ..., \kappa_N\}$, the numbers of common factors k, and the numbers of group-specific factors $\{k_1, ..., k_S\}$.

Step 3. Given the current values of S, k and $\{k_1, ..., k_S\}$, optimize each of the regularization parameters, κ_i $(i = 1, ..., N)$ by minimizing PIC

Step 4. Given the values of S, the numbers of group-specific factors $\{k_1, ..., k_S\}$ and κ_i $(i = 1, ..., N)$, optimize the numbers of common factors k by minimizing PIC

Step 5. Given the values of S, the numbers of common factors k, and κ_i $(i = 1, ..., N)$, optimize the numbers of group-specific factors $\{k_1, ..., k_S\}$ by minimizing PIC

Step 6. Repeat Steps 3 and 5 until convergence. Then, store the value of PIC.

Step 7. Change the value of S and implement Steps $2 \sim 6$.

Step 8. Select the best model based on the stored values of PIC.

Using the above algorithm, we calculate the stability measure V_C^2 in (5). The final model is obtained under the optimized C, which is a moderate value such that the selected model is stable. Simulation results show that the proposed algorithm perform well. For more details, see supplementary materials.

4 Theoretical Analysis

In this section, we consider the asymptotic analysis. In particular, we derive the asymptotic properties of the proposed estimator and show that the proposed estimator is consistent, as N and T go to infinity simultaneously. We also develop the variable selection consistency of the proposed estimator for the regression coefficients. Ando and Bai (2015) established an oracle property for the finite parameter case, under the known group membership. In the context of cross-sectional regression, Fan and Li (2001) demonstrated that penalized likelihood estimators based on SCAD are asymptotically as efficient as the oracle estimator. In this paper, the group membership is unknown and thus the establishment of variable selection consistency is a challenge.

We also consider the situation in which the number of predictors tends to infinity. In Appendix B, we provide the consistency of the estimated regression coefficients, the consistency of the estimated group membership, and the variable selection consistency under the diverging number of parameters.

We use $F_c^0, \{F_j^0,\ j = 1, ..., S\}$ to denote the true parameter values of the common and group-specific factors from the data-generating process. As T increases, the number of elements in F_c and F_j $(j = 1, ..., S)$ is also increasing. We first show that the estimated factors are consistent in the sense of some averaged norm, which will be specified below. First, we have the following theorem.

Theorem 1 : Consistency. *Under Assumptions A–E,* $\kappa = \max\{\kappa_1, ..., \kappa_N\} \to 0$ *and* $T \times \kappa_i \to \infty$ $(i = 1, ..., N)$ *as* $T, N \to \infty$. *The estimators* $\hat{F}_c$ *and* $\{\hat{F}_j,\ j = 1, ..., S\}$ *are consistent in the sense of the following norm*

$$T^{-1}\|\hat{F}_c - F_c^0 H_c\|^2 = o_p(1), \quad T^{-1}\|\hat{F}_j - F_j^0 H_j\|^2 = o_p(1), \quad j = 1, ..., S, \tag{6}$$

where $H_c^{-1} = V_{c,NT}(F_c^0 \hat{F}_c/T)^{-1}(\Lambda_c^{0\prime}\Lambda_c^0/N)^{-1}$, $H_j^{-1} = V_{j,N_jT}(F_j^0 \hat{F}_j/T)^{-1}(\Lambda_j^{0\prime}\Lambda_j^0/N_j)^{-1}$,

and $V_{c,NT}$ *and* V_{j,N_jT} *satisfy*

$$\Big[\frac{1}{NT}\sum_{i=1}^{N}(\boldsymbol{y}_i - X_i\hat{\boldsymbol{\beta}}_i - \hat{F}_{\hat{g}_i}\hat{\boldsymbol{\lambda}}_{\hat{g}_i,i})(\boldsymbol{y}_i - X_i\hat{\boldsymbol{\beta}}_i - \hat{F}_{\hat{g}_i}\hat{\boldsymbol{\lambda}}_{\hat{g}_i,i})'\Big]\hat{F}_c = \hat{F}_c V_{c,NT},$$

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

The following theorem shows that the estimated group membership converges to the true group membership as T and N grow.

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

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

Theorem 2 implies 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. Theorem 2 is similar to a result obtained by Bonhomme and Manresa (2015) and Ando and Bai (2016). But these studies do not allow heterogeneous regression coefficients.

Next, we establish the variable selection consistency of the estimated regression coefficients. Let $\boldsymbol{\beta}_i^0 = (\boldsymbol{\beta}_{i10}{}', \boldsymbol{\beta}_{i20}{}')'$ be the true parameter value, and $\hat{\boldsymbol{\beta}}_i = (\hat{\boldsymbol{\beta}}_{i1}', \hat{\boldsymbol{\beta}}_{i2}')'$ be the corresponding parameter estimate. Without loss of generality, we assume that $\boldsymbol{\beta}_{i20} = \mathbf{0}$. We also assume that the dimension of $\boldsymbol{\beta}_{i10}$ is small (uniformly bounded over i) but the dimension of $\boldsymbol{\beta}_{i20}$ can be large. We show that the estimator possesses the sparsity property, $\hat{\boldsymbol{\beta}}_{i2} = \mathbf{0}$. We denote $\hat{\boldsymbol{\beta}}_{i1}$ as the parameter estimate of non-zero true coefficients $\boldsymbol{\beta}_{i10}$.

Before we provide the theorem that establishes the oracle property of our estimator, we introduce the following assumption.

Assumption F: Regression coefficients

Each element of $\boldsymbol{\beta}_{i10}$ satisfies

$$\min|\beta_{i10,k}|/\kappa_i \to \infty \quad \text{as} \quad T \to \infty$$

for $k = 1, ..., q_i$ with q_i being the dimension of $\boldsymbol{\beta}_{i10}$. This assumption is required for obtaining the variable selection consistency.

Theorem 3 : Variable selection consistency. *Suppose that the assumptions A∼F hold. Let* $\hat{\boldsymbol{\beta}}_{i2}$ *be the parameter estimate of zero true coefficients* $\boldsymbol{\beta}_{i20}$*. The following variable selection consistency holds:*

$$P(\hat{\boldsymbol{\beta}}_{i2} = \mathbf{0}) \to 1, \quad N, T \to \infty.$$

Also, the parameter estimate of non-zero true coefficients $\boldsymbol{\beta}_{i10}$*,* $\hat{\boldsymbol{\beta}}_{i1}$*, is consistent,* $\|\boldsymbol{\beta}_{i10} - \hat{\boldsymbol{\beta}}_{i1}\| \to 0$.

In Theorems 1∼3, we assumed that the number of observable factors is fixed. We note that similar results still hold even when the number of observable factors goes to infinity, as $N, T \to \infty$. All theoretical proofs relating to these arguments are provided in Appendix B.

Finally, we must determine the number of groups, S, the number of group-specific factors, k_j $(j = 1, ..., S)$, and the size of the regularization parameters, κ_1,...,κ_N. The following theorem justifies the use of our proposed PIC in (4) for selecting these quantities.

Theorem 4 : Consistent model selection. *Suppose that the assumptions of Theorem 3 hold. Moreover, the difference between the diverging speed of* N *and* T *are not so extreme such that* $(T+N)/(TN)\log(TN) \to 0$ *and* $\min\{T, N\}\times(T+N)/(TN)\log(TN) \to \infty$*. Then, the proposed PIC provides a consistent estimation of the true number of groups, the true number of group-specific factors, and the set of true observable factors.*

Note that the conditions $(T + N)/(TN)\log(TN) \to 0$ and $\min\{T, N\} \times (T + N)/(TN)\log(TN) \to \infty$ are not strong. As discussed in Bai and Ng (2002), examples such as $N = \exp(T)$ or $T = \exp(N)$ are rare situations that will violate the conditions.

5 Data

The data employed in this paper cover publicly traded firms and firms traded at over-the-counter trading markets. The firms in our dataset belong to the following industries: Banking, Life Insurance, Nonlife Insurance, Financial Services, and Real Estate Investment and Services. All data are collected from the Datastream database, and we followed the industry assignment for each firm from this database.

We analyze the stock returns of over 6,000 firms from more than 100 financial markets. Following Forbes and Rigobon (2002), market returns are calculated as the rolling average, two-day returns of each of the firms. We use the two-day average returns because the worldwide financial markets do not have the same trading hours. For example, the business hours of the New York Stock Exchange (NYSE) and stock markets in East Asia (Tokyo, Hong Kong, Shanghai, etc.) do not overlap at all. Notably, the performance of financial markets in the Asia Pacific region may influence the financial markets in North America. Conversely, the U.S. stock exchange often influences the next day's performance of the Tokyo Stock Exchange, as described in Ohno and Ando (2014). Following Forbes and Rigobon (2002), we calculate stock returns in U.S. dollars. To study the dynamic characteristics of the worldwide stock market during the subprime financial crisis, we analyze the following 5 periods, in addition to the whole period (July 1, 2006 to November 31, 2009).

Period 1: July 1, 2006 to December 31, 2006
Period 2: July 1, 2007 to December 31, 2007
Period 3: February 1, 2008 to August 31, 2008
Period 4: October 1, 2008 to March 31, 2009
Period 5: May 1, 2009 to November 31, 2009

Based on the information summarized by Reuters and Federal Reserve Bank of St. Louis, Table 1 provides a timeline of the U.S. subprime crisis. Longstaff (2010) investigated the pricing of subprime asset-backed collateralized debt obligations and contagion effects arising from the U.S. sub-prime market in a worldwide framework. In that study, the sample period is divided into three distinct periods: the 2006 pre-crisis period, the 2007 subprime-crisis period, and the 2008 global financial crisis period. During Period 1, particularly during late 2006, the US housing markets had peaked, and delinquency rates for subprime mortgages were on the rise, setting up the subprime crisis. Period 1, Period 2, and Period 3 are considered the pre-crisis period, the subprime-crisis period, and the global financial crisis period, respectively. As shown in Table 1, conditions worsened during Period 4, including the Lehman Brothers bankruptcy. So Period 4 is also a global financial crisis period. In October, 2009, the Dow Jones Industrial Average closed above 10,000 for the first time since October 3, 2008. Thus, Period 5 contains the recovery of the U.S. financial markets.

Stocks with missing returns are excluded from our analysis. In addition, stocks with no variation at all were deleted from our sample. This operation leads to the final

sample for each period as follows, Period 1: $N = 6066$ firms, Period 2: $N = 6100$ firms, Period 3: $N = 6087$ firms, Period 4: $N = 6010$ firms and Period 5: $N = 6003$ firms, respectively. For the whole period, there are $N = 5813$ firms without missing values. Table 2 shows the distribution of our sample stocks across the markets. Financial markets with fewer than 50 stocks in our sample are merged together and denoted as "Others" in Table 2. Table 3 presents the distribution of our sample stocks across markets and industries during the whole period. The distribution of our sample stocks across markets and industries for the other periods are similar to Table 3 and thus omitted.

For each of the 31 markets (with more than 50 stocks), we compute the average return. Then, the lagged average returns $\boldsymbol{x}_{\text{lag},t} = (x_{\text{lag},1t}, x_{\text{lag},2t}, ..., x_{\text{lag},31t})'$ are used as the predictors $\boldsymbol{x}_{it}$. Also, their interactions (products) are also added to the predictors. Thus, the dimension of the observable factors $\boldsymbol{x}_{it}$ is $p_i = 31 + (31 \times 30)/2 = 496$, which is more than three times larger than the length of time series T for period 1 $\sim$ period 5. Figure 1 shows the correlation matrix of the set of 31 lagged average returns $\boldsymbol{x}_{\text{lag},t}$, and the magnitude of correlation increases as time passes. In particular, the correlation in Period 4 exhibits the highest dependency among the 5 periods. The magnitudes of correlation decrease in Period 5.

6 An empirical analysis

We estimate the model parameters in (1) by minimizing the objective function. Then, we apply the proposed model-selection criterion, PIC, to simultaneously select the number of groups, S, the number of group-specific pervasive factors, and the size of the regularization parameters, $\{\kappa_1, ..., \kappa_N\}$. We set the maximum number of groups to $S_{\max} = 30$. The possible number of group-specific pervasive factors r_j ranges from 0 to 16. Possible candidates for the regularization parameter, κ_i $i = 1, ..., N$ are $\kappa_i = 10^{1-(k-1)/2}$ with $k = 1, ..., 11$. To determine the value of C in PIC, we prepared its candidate values as $C = 0.1 \times k$ for $k = 1, ..., 20$. When we calculate the V_C^2 score in (5), we prepared the subsamples of sizes $(N^{(a)}, T^{(a)}) = (N - a \times 100, T - a \times 10)$ with $a = 0, 1, ..., 5$. We then optimized the value of C by minimizing V_C^2.

Figure 2 shows the behavior of V_C^2 as a function of C under the period 1. Noting that too small C (i..e, C) leads $V_C^2 \approx 0$ (because the maximum model will be identified) and too large C leads also $V_C^2 \approx 0$ (because the model with no factor structure will

be identified), Figure 2 indicates that the stable range of V_C^2 is around $C = [0.9, 1.4]$. Similarly we can identify the stable range of V_C^2 for the other periods. Using the value of C that achieves the minimum value of V_C^2 under the stable range, the proposed criterion PIC selects the best model among the set of candidate models.

6.1 Grouping results

The estimated number of groups, the number of common factors, and the number of group-specific factors for each of the periods are summarized in Table 4. In this table, the number of groups in Period 1 is determined to be $S = 7$ because it achieved the smallest value of the proposed model-selection criterion, PIC, which suggests that there are approximately $S = 7$ asset groups in Period 1. The table shows that the number of groups is increasing as time goes by. In addition, the total number of group-specific unobservable factors ($\sum_{j=1}^{7} r_j$) in Period 1 is much smaller than that in other periods, which implies that the degree of market heterogeneity has increased during the financial crisis. Thus, investors' behaviors may be more varied due to the increase in the degree of uncertainty of future events. The empirical results also show that the number of factors varies across groups. There exist one common factors in each period. To explore the economic meanings of the constructed common factors, we regress the extracted common factor on Thomson Reuters Global Financial Index. We found that the factor is relating to the index; the estimated regression coefficients are statistically significant at the 1% level.

We investigated how much variation is left in the error term $\hat{\varepsilon}_{it} = y_{it} - \boldsymbol{x}_{it}\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{f}}_c\hat{\boldsymbol{\lambda}}_{c,i} - \hat{\boldsymbol{f}}_{\hat{g}_i}\hat{\boldsymbol{\lambda}}_{\hat{g}_i,i}$. Here, we reported the following two ratios:

$$R_1 = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(\hat{\varepsilon}_{it})^2}{\sum_{i=1}^{N}\sum_{t=1}^{T} y_{it}^2}, \quad \text{and} \quad R_2 = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(\hat{\varepsilon}_{it})^2}{\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - \boldsymbol{x}_{it}\hat{\boldsymbol{\beta}}_i)^2}, \tag{7}$$

where R_2 adjusts the contribution by the observable factors. Table 4 also reports the values of R_1 and R_2 for each period. It can be seen that period 4 has larger R_1 values among the five sub-periods, while R_2 achieves the smallest. This implies that the relative explanatory power of unobservable factor structure increased in period 4. We can also see that R_1 and R_2 under the whole period is much larger than the other sub-periods. One of the possible reasons may be that the number of groups S is changing over the sub-periods, a single common factor structure is too restrictive for the entire period.

The size of the groups is summarized in Table 5. In Period 1, Group G7 is subject to a total of two factors. In contrast, Group G5 is subject to one group-specific factor. Notably, the size of G5 is more than 1,100 and is two times larger than G7. This implies that Group G7's degree of heterogeneity is much larger than that of G5. Such contrast can also be observed in Period 2. Group G3 is subject to a single group-specific factor, while Group G8 is subject to three group-specific factors.

Because the industry classifications and listed markets are known, a two-way table of the estimated group membership $\hat{g}_i$ against these classifications is investigated. Figure 3 shows the distribution of the firms in each sector. An (i, j)-th element denotes % of firms in industry i such that they belong to j-th group. More specifically, let $n_{i,j}$ denote the number of firms that belongs to sector i and to group j. Then, the (i, j)-th element m_{ij} is calculated as $m_{ij} = n_{i,j}/\{\sum_{k=1}^{S} n_{i,k}\}$. In Period 1 and Period 2, Banking and Insurance (Life, Non-life) sectors seem to be in large clusters. However, due to the subprime crisis, stocks in the real estate investment sector diverged more in Period 3. In particular, we find a large cluster in Period 4, whereas we found a large cluster in the Banking, Life insurance and Real estate sectors in Period 5. These investigations imply that the industry factor by itself is an important factor pre-financial crisis or during the financial crisis, but industry may not matter after a large shock. This finding, derived form our general procedure, is a useful insight for institutional investors.

Figure 4 shows the distribution of the firms in each of the stock exchanges. An (i, j)-th element denotes % of firms listed at stock exchange i such that they belong to the j-th group. We can thus make the following observations. First, the magnitudes of similarity between the New York Stock Exchange and the NASDAQ are stable over the periods. In contrast, non-NASDAQ OTC market exhibits different behavior. This implies that the investor's behaviors at the New York Stock Exchange and the NASDAQ are different from those at non-NASDAQ OTC market. Thus, investors should consider such market characteristics although all three markets are located in the U.S. Second, an increase in the dissimilarity of the Shanghai and Shenzen stock exchanges from the rest of the exchanges is observed after Period 3. Third, the stocks listed on the Tokyo stock exchange tend to be in the same group, even in Period 1, which indicates that it is difficult to diversity the portfolio risk that consists purely of stocks listed on the Tokyo stock exchange. Portfolio managers who are willing to diversify their portfolio should consider taking positions in stocks in different exchanges. These

results imply that country and listed markets are important factors to be considered. However, we also note that these factors are not able to fully capture the underlying market behavior.

When we compare the distribution of firms in each of the stock exchanges in Figure 4, period 4 indicates co-movements owing to financial crisis because many stocks in different stock exchanges congregate in the same group G4. In fact, G4 is the biggest cluster in table 5, which indicates stronger comovements as a result of the financial crisis. On the other hand, this phenomenon is relatively much weaker in the other sub-periods. Thus, our estimation result is consistent with the idea that in crisis co-movements are stronger.

We also implemented Fisher's Exact Test for the independence between our grouping results and these two nominal classifications. These two null hypothesis of independence were rejected for all periods. Although this rejection implies that the our grouping results are relating to industry and listed exchange, the nominal classifications are not sufficient to capture the complicated market characteristics. In summary, these investigations imply that while industry, market, country, and region are sources of the co-movement of the cross-sectional and time-series variations in stock returns, they are not the only sources of the co-movement.

6.2 Price of risk

In the APT framework, the expected returns on assets are approximatively linear in their sensitivities to the factors $E[r] = \nu_0 + \boldsymbol{\gamma}'\boldsymbol{\nu}$, where ν_0 is a constant, $\boldsymbol{\nu}$ is a vector of factor risk premiums, and $\boldsymbol{\gamma}$ is a vector of factor sensitivities. In our model, factor sensitivities correspond to the regression coefficients $\boldsymbol{\beta}_i$, and factor loadings $\boldsymbol{\lambda}_{c,i}$ and $\boldsymbol{\lambda}_{g_i,i}$ for $i = 1, ..., N$. Here, we partition the excess returns into the identified groups and investigate the subset pricing relations based on the Fama and MacBeth (1973) approach. This subgroup two-stage approach was also used in Goyal et al. (2008) and Ando and Bai (2015) for example.

Through the model building process, we have already estimated the matrix of factor sensitivities $\hat{\Lambda}_c$ (common factors), $\hat{\Lambda}_j$ (group-specific factors with respect to j-th group). Following Ando and Bai (2015), we then run the following cross-sectional regression for each group:

$$\hat{\boldsymbol{r}}_j = \nu_{0,j}\mathbf{1} + \hat{\Lambda}_{c,j}\boldsymbol{\nu}_{c,j} + \hat{\Lambda}_j\boldsymbol{\nu}_{g,j} + \boldsymbol{\xi}_j, \quad (j = 1, 2, ..., S),$$

where $\mathbf{1}$ is a vector of ones, $\boldsymbol{\xi}_j$ is a vector of pricing errors, $\hat{\Lambda}_{c,j}$ corresponds to the sub-element of $\hat{\Lambda}_c$ associated with the j-th group, $\hat{\boldsymbol{r}}_j$ is a vector of average excess returns, which are observable-risk adjusted, i.e., for the i-th security, $T^{-1}\sum_{t=1}^{T}(y_{it} - \boldsymbol{x}_{it}'\hat{\boldsymbol{\beta}}_i)$ is used. Table 6 reports the results of this cross-sectional regression. The estimates for the risk premium on the common and group-specific factors are statistically significant in each group. Almost all factors seem to be priced. This indicates that our method extracted factors that are priced.

One of the main characteristics of the proposed method is selecting the set of relevant observable factors. Figure 5 provides the histogram of the percentages (%) of non-zero estimated regression coefficients for each of the observable factors $\sum_{i=1}^{N} I(\hat{\beta}_{ik} \neq 0)/N$ for $k = 1, ..., 496$. That is, the histogram is based on 496 values of percentages. Because the histograms under the periods 2, 3, and 5 are similar to that under period 1, these figures are omitted. We can see that each of the observable factors are relevant for at most 5% of stocks' returns for period 1. The relevance is further reduced for the whole period, implying that the sensitivities to the observable factor are likely to be time varying (because the percentage becomes smaller under longer time series T).

It is also interesting to see whether these selected observable factors are priced in the cross-section of asset returns. Similar to the above analysis, we run the following cross-sectional regression for the observable factors:

$$\hat{\boldsymbol{r}} = \nu_0 \mathbf{1} + \hat{\Lambda}_\beta \boldsymbol{\nu}_\beta$$

where $\mathbf{1}$ is a vector of ones, $\hat{\Lambda}_\beta$ is the matrix of sensitivities to the observable factors and $\hat{\boldsymbol{r}}$ is a vector of average excess returns, which are unobservable common/group-specific factors adjusted, i.e., for the i-th security, $T^{-1}\sum_{t=1}^{T}(y_{it} - \hat{\boldsymbol{f}}_{c,t}'\hat{\boldsymbol{\lambda}}_{c,i} - \hat{\boldsymbol{f}}_{\hat{g}_i,t}'\hat{\boldsymbol{\lambda}}_{\hat{g}_i,i})$ is used. Figure 5 also provides the histogram of the p-values of estimated $\boldsymbol{\nu}_\beta$. Again, the histograms under the periods 2, 3, and 5 are similar to that under period 1, and thus omitted. When we set the critical level as $\alpha = 0.05$, more than 50% of observable factors are priced under the period 1. This indicates that our method detected relevant observable factors for explaining stock returns. In contrast, 5% of observable factors are priced under the whole period. As shown in Figure 5, our proposed method estimated almost all of regression coefficients as zero. This indicates that most of observable factors are priced in a short time period, while it may not be true in the whole sample period. In contrast, the extracted unobservable factor structures are priced even for the whole sample period. Thus, the unobservable factor structure, $\boldsymbol{f}_{c,t}'\boldsymbol{\lambda}_{c,i} + \boldsymbol{f}_{g_i,t}'\boldsymbol{\lambda}_{g_i,i}$,

plays an important role in modeling asset returns.

6.3 Robustness check

Finally, the five periods split the sample into pre-crisis, and during-crisis subsets. Different specifications can be used for these sub-periods. However, we note that the different sub-period specifications still lead to similar results.

To treat the differences in international market trading hours, we used the rolling two-day average of returns. It is also possible to implement the proposed modeling procedure for daily returns instead of a two-day rolling average. Again, similar results are obtained. Thus our results are robust in different subperiod specifications and in using either daily returns or two-day averages.

7 Conclusion

This paper proposed a novel and a general approach that simultaneously implements the following features: (1) detecting a set of relevant observable factors, (2) extracting unobservable common and group-specific factors, (3) automatically determining the number of groups, and (4) clustering a huge number of assets.

To study the global financial crisis caused by the collapse of the U.S. sub-prime mortgage market, we analyzed the daily stock returns of several industries related to financial services for over 6,000 stocks from more than 70 countries and over 100 financial markets. We found that the number of groups during the financial crisis is much larger than during the pre-crisis period. We also found that industry, market, country, and region are sources – but not the only sources – of the co-movement of cross-sectional and time-series variations in stock returns during the financial crisis and that other sources of co-movement extend beyond these usual classifications.

Although grouping stocks based on nominal classifications (industry, market, country, and region) is convenient and simple, the market structure is not simple enough for portfolios to be well diversified based on these nominal classifications alone. We recommend that investors looking for global investment opportunities consider diversifying their portfolios broadly based on our grouping results. Our empirical findings may be useful for paired trading, valuation of firms, etc.

This paper analyzes stocks relating to the financial industry. Our method is also applicable to stocks of other industries. In addition, the proposed methods can be ap-

plied to the analysis of international bond markets and to the analysis of high-frequency trading data. Because the proposed method can handle hundreds and thousands of asset returns simultaneously, the scope of its applicability is wide.

Acknowledgments The authors would like to thank the Co-editor, the associate editor and two anonymous reviewers for constructive and helpful comments that improved the quality of the paper considerably. We also thank comments from participants at the 28th Australasian Finance and Banking Conference. Ando's research is supported by Research Grant from Melbourne Business School, and Bai's research is supported by the National Science Foundation (SES1357198).

References

Alessi, L., Barigozzi, M. and Capasso, M. (2010), "Improved penalization for determining the number of factors in approximate static factor models," *Statistics and Probability Letters*, 80, 1806–1813.

Amengual, D., and 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.

Ando, T., and Bai, J. (2014), "Selecting the regularization parameters in high-dimensional panel data models: consistency and efficiency," *Econometric Reviews*, forthcoming.

Ando, T., and Bai, J. (2015), "Asset pricing with a general multifactor structure," *Journal of Financial Econometrics*, 13, 556–604.

Ando, T., and Bai, J. (2016), "Panel data models with grouped factor structure under unknown group membership," *Journal of Applied Econometrics*, 136, 163–191.

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

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

Baca, S.P., Garbe, B.L., and Weiss, R.A. (2000), "The rise of sector effects in major equity markets," *Financial Analysts Journal*, 56, 34–40.

Beckers, S., Connor, G., and Curds, R. (1996), "National versus global influences on equity returns," *Financial Analysts Journal*, 52, 31–39.

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

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

Candes, E., and Tao, T. (2007), "The Dantzig selector: statistical estimation when p is much larger than n," *Annals of Statistics*, 35, 2313–2351.

Cavaglia, S., Brightman, C., and Aked, M. (2000), "The increasing importance of industry factors," *Financial Analysts Journal*, 56, 41–54.

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

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.

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.

Diebold F.X., and Yilmaz, K. (2014), "On the network topology of variance decompositions: Measuring the connectedness of financial firms," *Journal of Econometrics*, 182, 119–134.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004), "Least angle regression," *Annals of Statistics*, 32, 407–499.

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

Fama, E. F., and French, K. R. (1998), "Value versus Growth: The International Evidence," *Journal of Finance*, 53, 1975–1999.

Fama, E. F., and French, K. R. (2012), "Size, value, and momentum in international stock returns," *Journal of Finance*, 105, 457–472.

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

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

Fan, J., Liao, Y., and Mincheva, M. (2013). "Large covariance estimation by thresholding principal orthogonal complements." *Journal of the Royal Statistical Society*, B75, 603–680.

Forbes, K. J., and Rigobon, R. (2002), "No contagion, only interdependence: Measuring stock market comovements," *Journal of Finance*, 57, 2223–2261.

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.

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

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

Griffin, J. M. (2002), "Are the Fama and French Factors Global or Country Specific," *Review of Financial Studies*, 15, 783–803.

Griffin, J.M., and Karolyi, G.A., (1998), "Another look at the role of the industrial structure of markets for international diversification strategies," *Journal of Financial Economics*, 50, 351–373.

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

Hallin, M., and R. Liška (2011), "Dynamic factors in the presence of blocks," *Journal of Econometrics*, 163, 29–41.

Heston, S.L., and Rouwenhorst, K.G., (1994), "Does industrial structure explain the benefits of industrial diversification," *Journal of Financial Economics*, 36, 3–27.

Heston, S.L., and Rouwenhorst, K.G., (1995), "Industry and country effects in international stock returns," *Journal of Portfolio Management*, 21, 53–58.

Hou, K., Karolyi, G.A., and Kho, B.-C. (2011), "What Factors Drive Global Stock Returns," *Review of Financial Studies*, 24, 2527–2574.

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

Kuo, W., and Satchell, S.E.,(2001), "Global equity styles and industry effects: the pre-eminence of value relative to size," *Journal of International Financial Markets, Institutions and Money*, 11, 1–28.

Lam, C., and Fan, J. (2008), "Profile-kernel likelihood inference with diverging number of parameters," *Annals of Statistics*, 36, 2232–2260.

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.

Longstaff, F.A. (2010), "The subprime credit crisis and contagion in financial markets," *Journal of Financial Economics*, 97, 436–450.

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

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.

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

Ohno, S. and Ando, T. (2015), "Stock return predictability: A factor-augmented predictive regression system with shrinkage method," *Econometric Reviews*, forthcoming.

Park, T., and Casella, G. (2008), "The Bayesian Lasso," *Journal of the American Statistical Association*, 103, 681–686.

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

Roll, R., (1992), "Industrial structure and the comparative behavior of international stock market indices," *Journal of Finance*, 47, 3–42.

Sargent, T. J. and C. A. Sims. (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.

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

Su, L., Shi, Z. and Phillips, P. (2014), "Identifying latent structures in panel data," Working paper.

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

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

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

Yuan, M. and Lin, Y. (2006), "Model selection and estimation in regression with

grouped variables, " *Journal of the Royal Statistical Society*, B68, 49–67

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

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

Zou, H. and Hastie, T. (2005), "Regularization and variable selection via the Elastic Net," *Journal of the Royal Statistical Society* B67, 301–320.

Table 1: Timeline of the U.S. subprime crisis, summarized from Reuters and Federal Reserve of St.Louis.

Time	Events
Late 2006	The U.S. housing market slows down and delinquency rates on U.S. subprime mortgages are on the rise.
February 2007	HSBC announces that it must manage bad debts in U.S. subprime lending portfolios.
April 2007	California's New Century Financial Corp files for Chapter 11 bankruptcy protection.
June 2007	Standard and Poor's and Moody's Investor Services downgrad over 100 bonds backed by subprime mortgages.
	Two Bear Stearns funds must manage losses after making bad bets on securities backed by subprime loans.
	S&P slashes ratings on some top-rated mortgage bonds by eight notches.
August 2007	American Home Mortgage Investment Corporation files for Chapter 11 bankruptcy protection.
	BNP Paribas freezes $2.2 billion worth of funds by citing subprime problems.
	Barclays Bank borrows 314 million pounds from Bank of England's standing lending facility.
October 2007	UBS indicates it will write down $3.4 billion of assets.
	Merrill Lynch announces $8.4 billion of losses and writedowns in CDOs, subprime and leveraged loans in Q3.
	Citigroup announces $6.5 billion of losses and writedowns on subprime-related debt and loans in Q3.
November 2007	Citigroup announces a further $8-11 billion of writedowns.
March 2008	JP Morgan acquires Bear Stearns in rescue.
July 2008	Federal Reserve Bank of New York is authorized to lend to the Federal National Mortgage Association (Fannie Mae) and the Federal Home Loan Mortgage Corporation (Freddie Mac).
September 2008	Federal Housing Finance Agency (FHFA) places Fannie Mae and Freddie Mac in government conservatorship.
	Bank of America announces its intent to purchase Merrill Lynch & Co. for $50 billion.
	Lehman Brothers files for Chapter 11 bankruptcy protection.
	Federal Reserve Bank of New York is authorized to lend up to $85 billion to American International Group.
	FOMC expands/authorizes swap lines with the Bank of Japan, Bank of England, and Bank of Canada.
	FOMC establishes new swap lines with the Reserve Bank of Australia, the Sveriges Riksbank, the Danmarks National bank and the Norges Bank.
	FOMC increases swap lines with the European Central Bank and the Swiss National Bank.
	FDIC announces that Citigroup will purchase the banking operations of Wachovia Corporation.
October 2008	The U.S. Congress passes the Emergency Economic Stabilization Act to establish the $700 billion Troubled Asset Relief Program.
December 2008	The U.S. Treasury purchases preferred stock ($1.91 billion) from 7 U.S. banks under the Capital Purchase Program.
May 2009	Freddie Mac reports a first quarter 2009 loss of $9.9 billion.
October 2009	The Dow Jones Industrial Average closes above 10,000 for the first time since October 3, 2008.

Table 2: Distributions of the number of listed financial firms. Period 1: July 1, 2006 to December 31, 2006. Period 2: July 1, 2007 to December 31, 2007. Period 3: February 1, 2008 to August 31, 2008. Period 4: October 1, 2008 to March 31, 2009. Period 5: May 1, 2009 to November 31, 2009. Whole period: July 1, 2006 to November 31, 2009. NYSE: New York Stock Exchange. PSE: Philippine Stock Exchange.

	Period 1	Period 2	Period 3	Period 4	Period 5	Whole
Amman	67	68	68	66	66	64
Australian	99	97	97	96	100	92
Bangkok	87	87	85	84	85	81
Berlin	145	146	148	142	144	135
BSE Ltd	285	283	282	283	278	262
Dhaka	68	69	69	68	69	67
Euronext.liffe Paris	80	80	81	82	81	73
Frankfurt	530	537	536	524	533	509
Hong Kong	197	202	202	200	198	188
Indonesia	75	76	75	75	75	73
Karachi	56	55	55	55	54	52
Korea Stock Exchange	52	51	53	52	53	50
Kuala Lumpur	98	99	99	99	99	98
Kuwait City	75	75	74	72	71	69
London	145	146	143	141	145	135
NASDAQ	367	371	372	370	374	368
National India	101	103	102	100	101	96
NYSE	184	185	182	182	180	178
Non NASDAQ OTC	1265	1277	1287	1244	1230	1233
PSE	81	83	83	82	80	77
Shanghai	73	73	73	73	73	73
Shenzen	55	55	55	54	52	51
SIX Swiss	68	68	69	68	68	68
Stuttgart	62	64	62	63	60	58
Taiwan	60	61	61	61	60	59
Tel Aviv	107	106	104	105	106	102
Thailand	58	59	58	57	57	52
Tokyo Stock Exchange	176	177	175	176	176	173
Toronto	80	78	79	78	79	73
TSX Ventures	64	63	64	62	60	55
XETRA	82	81	82	84	83	79
Others	1124	1125	1112	1112	1113	1070
Total	6066	6100	6087	6010	6003	5813

Table 3: Distributions by market and industry: Whole period: July 1, 2006 to November 31, 2009. NYSE: New York Stock Exchange. PSE: Philippine Stock Exchange.

	Bank	Life Insurance	Non Life Insurance	Financial Services	Real Estate
Amman	11	0	21	18	14
Australian	7	3	3	54	25
Bangkok	10	1	14	20	36
Berlin	54	6	13	36	26
BSE Ltd	0	0	0	232	30
Dhaka	31	4	23	8	1
Euronext.liffe Paris	19	1	4	23	26
Frankfurt	165	35	66	153	90
Hong Kong	10	4	3	54	117
Indonesia	20	1	10	18	24
Karachi	14	3	9	26	0
Korea Stock Exchange	5	0	10	34	1
Kuala Lumpur	10	1	6	12	69
Kuwait City	9	0	8	26	26
London	11	7	12	67	38
NASDAQ	300	5	23	31	9
National India	34	1	1	47	13
NYSE	58	18	45	48	9
Non NASDAQ OTC	471	21	23	647	71
PSE	13	2	0	22	40
Shanghai	4	0	0	6	63
Shenzen	1	0	0	4	46
SIX Swiss	25	1	6	11	25
Stuttgart	10	1	3	19	25
Taiwan	17	5	5	5	27
Tel Aviv	10	1	7	29	55
Thailand	8	1	6	11	26
Tokyo Stock Exchange	78	3	3	35	54
Toronto	11	7	5	34	16
TSX Ventures	2	0	3	40	10
XETRA	28	5	13	23	10
Others	395	30	89	336	220

Table 4: Selected number of groups, the number of common factors and the number of group-specific factors. Period 1: July 1, 2006 to December 31, 2006. Period 2: July 1, 2007 to December 31, 2007. Period 3: February 1, 2008 to August 31, 2008. Period 4: October 1, 2008 to March 31, 2009. Period 5: May 1, 2009 to November 31, 2009. Whole period: July 1, 2006 to November 31, 2009. R_1 and R_2 in (7) measure how much variations left in the error term $\hat{\varepsilon}_{it} = y_{it} - \boldsymbol{x}_{it}\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{f}}_c\hat{\boldsymbol{\lambda}}_{c,i} - \hat{\boldsymbol{f}}_{\hat{g}_i}\hat{\boldsymbol{\lambda}}_{g_i,i}$.

	Period 1	Period 2	Period 3	Period 4	Period 5	Whole
S	7	8	8	11	8	13
r	1	1	1	1	1	1
r_1	1	3	3	3	1	2
r_2	1	4	4	8	5	9
r_3	3	1	2	3	3	12
r_4	3	3	1	10	1	6
r_5	1	3	4	3	2	3
r_6	2	3	1	1	3	4
r_7	2	1	4	7	3	9
r_8		3	5	8	3	9
r_9				1		16
r_{10}				2		10
r_{11}				10		15
r_{12}						4
r_{13}						11
R_1	0.0744	0.1362	0.1383	0.1835	0.1697	0.5647
R_2	0.1988	0.3049	0.2917	0.2612	0.3562	0.5685

Table 5: Size of each clusters. Period 1: July 1, 2006 to December 31, 2006. Period 2: July 1, 2007 to December 31, 2007. Period 3: February 1, 2008 to August 31, 2008. Period 4: October 1, 2008 to March 31, 2009. Period 5: May 1, 2009 to November 31, 2009. Whole period: July 1, 2006 to November 31, 2009.

	Period 1	Period 2	Period 3	Period 4	Period 5	Whole
N_1	1301	964	877	135	843	481
N_2	187	934	1179	504	1182	622
N_3	1125	1161	191	459	487	579
N_4	1014	472	862	2616	220	429
N_5	1196	597	857	297	183	262
N_6	703	528	272	19	928	272
N_7	540	999	537	421	1555	450
N_8		445	1312	567	605	558
N_9				156		668
N_{10}				144		163
N_{11}				692		672
N_{12}						71
N_{13}						586

Table 6: Factor risk premiums for the common and group-specific factors. For each group,we run the following cross-sectional regression: $\hat{\boldsymbol{r}}_j = \nu_{0,j}\mathbf{1} + \hat{\Lambda}_{c,j}\boldsymbol{\nu}_{c,j} + \hat{\Lambda}_j\boldsymbol{\nu}_{g,j} + \boldsymbol{\xi}_j$, $(j = 1, 2, ..., S)$. Details on this model are described in Section 6.2. The number of common and group-specific factors that are priced at the significance level $\alpha = 0.05$. For integers a^* and b^*, the ratio a^*/b^* means that a^* factors are priced among a set of b^* factors.

	Period 1		Period 2		Period 3		Period 4		Period 5		Whole	
	$\boldsymbol{f}_c$	$\boldsymbol{f}_g$	$\boldsymbol{f}_c$	$\boldsymbol{f}_g$	$\boldsymbol{f}_c$	$\boldsymbol{f}_g$	$\boldsymbol{f}_c$	$\boldsymbol{f}_g$	$\boldsymbol{f}_c$	$\boldsymbol{f}_g$	$\boldsymbol{f}_c$	$\boldsymbol{f}_g$
G_1	1/1	1/1	1/1	2/3	1/1	3/3	1/1	2/3	1/1	1/1	1/1	1/2
G_2	0/1	1/1	1/1	4/4	1/1	4/4	1/1	8/8	1/1	5/5	1/1	8/9
G_3	1/1	3/3	1/1	1/1	1/1	2/2	1/1	3/3	1/1	3/3	1/1	12/12
G_4	1/1	3/3	1/1	3/3	1/1	1/1	1/1	10/10	1/1	1/1	1/1	5/6
G_5	0/1	1/1	1/1	3/3	1/1	4/4	1/1	3/3	1/1	2/2	1/1	3/3
G_6	0/1	2/2	1/1	3/3	1/1	1/1	1/1	1/1	1/1	3/3	0/1	4/4
G_7	1/1	1/2	1/1	1/1	1/1	4/4	1/1	6/7	1/1	3/3	1/1	5/9
G_8			1/1	3/3	1/1	5/5	1/1	7/8	1/1	3/3	1/1	6/9
G_9							1/1	1/1			1/1	14/16
G_{10}							1/1	1/2			1/1	9/10
G_{11}							1/1	8/10			1/1	14/15
G_{11}											0/1	3/4

Period 1.

Period 2.

Period 3.

Period 4.

Period 5.

Figure 1: Correlation matrix of the set of observable factors (see the text for explanation). Period 1: July 1, 2006 to December 31, 2006. Period 2: July 1, 2007 to December 31, 2007. Period 3: February 1, 2008 to August 31, 2008. Period 4: October 1, 2008 to March 31, 2009. Period 5: May 1, 2009 to November 31, 2009.

Figure 2: The behavior of V_C^2 as a function of C under the period 1.

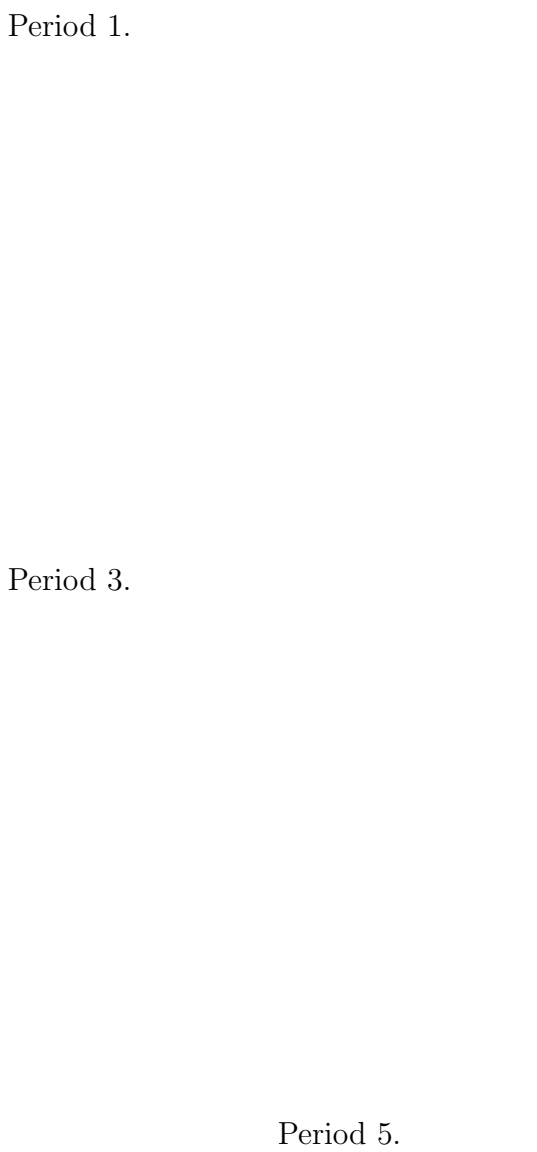

Figure 3: Distribution of firms in each of the sectors. An (i, j)-th element denotes % of firms in industry i such that they belong to j-th group. Period 1: July 1, 2006 to December 31, 2006. Period 2: July 1, 2007 to December 31, 2007. Period 3: February 1, 2008 to August 31, 2008. Period 4: October 1, 2008 to March 31, 2009. Period 5: May 1, 2009 to November 31, 2009.

Period 1.

Period 2.

Period 3.

Period 4.

Period 5.

Figure 4: Distribution of firms in each of the stock exchanges An (i, j)-th element denotes % of firms listed in a stock exchange i such that they belong to j-th group. Period 1: July 1, 2006 to December 31, 2006. Period 2: July 1, 2007 to December 31, 2007. Period 3: February 1, 2008 to August 31, 2008. Period 4: October 1, 2008 to March 31, 2009. Period 5: May 1, 2009 to November 31, 2009.

Period 1 (Non-zero estimated β_{ik})

Period 1 (Price of risk: p-value).

Period 4 (Non-zero estimated β_{ik})

Period 4 (Price of risk: p-value).

Whole period
(Non-zero estimated β_{ik})

Whole period
(Price of risk: p-value).

Figure 5: Left column: Histogram of the percentages (%) of non-zero estimated regression coefficients for each of the observable factors $\sum_{i=1}^{N} I(\hat{\beta}_{ik} \neq 0)/N$ for $k = 1, ..., 496$. Right column: Histogram of the p-values of price of risk. Period 1: May 1 2007 to December 31, 2008, Period 4: September 1 2008 to March 31, 2009, Whole May 1 2007 to December 31, 2009.