Identification of Unknown Common Factors: Leaders and Followers*

Jason Parker
Michigan State University

Donggyu Sul
University of Texas at Dallas

January, 2015

Abstract

This paper has the following contributions. First, this paper develops a new criterion for identifying whether or not a particular time series variable is a common factor in the conventional approximate factor model. Second, by modeling observed factors as a set of potential factors to be identified, this paper reveals how to easily pin down the factor without performing a large number of estimations. This allows the researcher to check whether or not each individual in the panel is the underlying common factor and, from there, identify which individuals best represent the factor space by using a new clustering mechanism. Asymptotically, the developed procedure correctly identifies the factor when N and T jointly approach infinity under the minimal assumptions of Bai and Ng (2002). The procedure is shown to be quite effective in the finite sample by means of Monte Carlo simulation. The procedure is then applied to an empirical example, demonstrating that the newly-developed method identifies the unknown common factors accurately.

JEL Classifications: C33

Key words: *Asymptoticially Weak Factor, Dominant Leader, Cross Section Dependence, Principal Component Analysis, Common Factor Model*

*We thank Ryan Greenaway-McGrevy, Chirok Han, and the two anonymous referees for their helpful comments.

1 Introduction

In the last two decades there has been rapid development in analyzing cross-sectional dependence by using the approximate common factor structure. Among many others, Ahn and Horenstein (2013), Ahn and Perez (2010), Amengual and Watson (2007), Bai and Ng (2002), Hallin and Liska (2007), Harding (2013), Kapetanios (2010), and Onatski (2009) suggest consistent estimation procedures for the number of common factors, while Bai (2003, 2004), Bates, Plagborg-Møller, Stock, and Watson (2013), Choi (2012), Forni, Hallin, Lippi, and Reichlin (2000, 2005), and Stock and Watson (2002a, 2002b) propose consistent estimators for the common factors.

However, the most thorny challenge in this literature is the identification of these unknown common factors. Without identification, a common factor model of an economic phenomenon is fundamentally incomplete. Presently, empirical researchers have two general identification strategies. First, some researchers are forced to settle for simply describing the factors using their shape, correlation to observed series, and factor loadings (e.g. Ludwigson and Ng 2007, Reis and Watson 2010). The problem with this approach is that the factor is only described, not pinned down. Sometimes researchers propose a name for the factor, but such a name is completely arbitrary. The other approach is to directly compare a $(m \times 1)$ vector of potentially true factors P_t with the $(r \times 1)$ vector of unknown latent factors G_t. Of course, the true factors G_t are not observable, so Bai and Ng (2006) propose several tests to check whether or not a linear combination among the principal component (PC, hereafter) estimates of G_t is identical to the potential factors P_t. Their methods require some restrictive assumptions. As we will show later, when the potential factors are slightly different from the true factors, even for only one time period, the tests proposed by Bai and Ng (2006) fail as the number cross-sectional units (N) and time series observations (T) go to infinity. Furthermore, the Bai and Ng (2006) tests cannot identify which components of P_t align with particular components of G_t. The solution to this problem does not seem to exist unless it can be assumed that the estimated factors are identical to the true factors.

The purpose of this paper is to provide a novel and intuitive approach to identify whether or not an observed time series is asymptotically equal to an unobserved true factor. The newly suggested identification strategy does not require any identification restrictions for the PC estimators. The underlying logic is based on the notion of an 'asymptotically weak factor.' When a panel data set has only asymptotically weak factors, the estimated number of common factors in the panel declines to zero with probability one as both N and T go to infinity. For example, the PC estimates of the idiosyncratic components have asymptotically weak factors. Obviously, conventional factor number estimation such as Bai and Ng (2002) or Hallin and Liska (2007) will estimate a factor number of zero with panel data which only has asymptotically weak factors. We are utilizing this principle to identify whether or not the vector of the potentially true factors, P_t, is indeed a linear combination

of the true latent factors, G_t. Let P_{jt} be the jth element of P_t. Then, it is easy to show that the regression residuals from the regression of one of the potential factors and any $(r-1)$ vector of the estimated common factors have only asymptotically weak factors, so the conventional factor number estimators can be used to examine whether or not a potential factor is the true common factor. Of course, if P_{jt} is not a true factor, then the regression residuals must have at least one strong factor. This simple but novel idea does not require any identification restrictions either on the PC estimators, the latent factors, or the latent factor loadings.

Moreover, this paper models the factor as potentially being one particular individual which appears in the panel. When one individual is exactly equal to the factor, we call this individual a 'dominant leader.' If the individual is not a factor in the finite sample, but becomes a factor as N and T go to infinity, the individual is called an 'approximate dominant leader.'

This leadership model has a powerful interpretation: one or more of the individuals act(s) as the source(s) of the cross-sectional dependence in the factor model, spreading its/their influence over the other individuals in accordance with the factor loadings. Consider the following hypothetical example of leadership: In industrial organization, one or a few dominant firms can set a price for their product, and the rest of firms in the market more or less adopt that price. This type of causal relationship can be observed in many areas including the social, agricultural, and behavioral sciences. In natural science, earthquakes and the spread of viruses are potential examples of this pattern. In such situations, a few individuals or locations become leaders or sources of epidemic events. Therefore, an important task is to identify the leaders from a set of individuals.

Since the factor identification strategy above must be performed separately for each individual, there could be some failure probability when N is large. To control this probability, we provide a method based on ranking R^2 values from regressions of the estimated PC factors on each individual time series separately. Individuals with high R^2 are considered 'leader candidates' to be considered as the potential factors or leaders, P_{jt}.

It is worth mentioning that two papers have already used our identification strategy. Gaibulloev, Sandler and Sul (2013) find that Lebanon is the main determinant of transnational terrorism. Greenaway-McGrevy, Mark, Sul and Wu (2014) utilize our method to find three key currencies as the main determinants for local exchange rates.

The remainder of the paper is organized as follows. Section 2 provides information about the setting as well as the definition of weak factors. Section 3 discusses leadership modeling and testing. Detailed asymptotic analyses are also provided. Section 4 demonstrates the finite sample performance of our test and also compares our results with Bai and Ng (2006). Section 5 provides an empirical example to show the effectiveness of our test. Section 6 concludes. Mathematical proofs are provided in the Appendix. Gauss code for the procedures as well as extra Monte Carlo

simulations are available on the authors' website.

2 Preliminary

Before we proceed, we define the variables that are used in the paper. y_{it} is the panel data of interest where the cross-sectional dependence can be expressed in a static common factor representation. G_t is the $r \times 1$ vector of potentially correlated, latent common factors. $\hat{F}_t$ is the $r \times 1$ vector of the PC estimator, $\#(y_{it})$ is the true number of common factors of y_{it} and $\hat{\#}(y_{it})$ is the estimated number of common factors of y_{it}. y_{it}^o is the idiosyncratic component to y_{it}.

To provide an intuitive explanation of how factor number estimation can be used to identify the true factors, we consider the following static factor structure with two factors ($r = 2$) as an example.

$$y_{it} = \alpha_{1i} G_{1t} + \alpha_{2i} G_{2t} + y_{it}^o, \tag{1}$$

where α_{ji} is the true factor loading coefficient for the ith individual and to the jth factor. We define y, G, $\hat{F}$, and $\mathcal{A}$ as the $T \times N$ matrix of y values, the $T \times r$ matrix of latent factors, the $T \times r$ matrix of estimated factors (when the factor number is known), and the $N \times r$ matrix of true factor loadings, respectively. We also define $H = (\mathcal{A}'\mathcal{A}/N)\left(G'\hat{F}/T\right)V_{NT}^{-1}$ to be the $r \times r$ rotation/rescaling matrix as defined in Bai (2003), where V_{NT} is the $r \times r$ diagonal matrix of the first r eigenvalues of $(NT)^{-1} y'y$ in decreasing order. In (1) and in the equations that follow, we exclude any non-zero constant terms for notational simplicity. Including constant terms does not change the results at all.

The number of factors in y_{it}^o is naturally zero, even if y_{it}^o has some weak cross-sectional dependence. Interestingly, the estimate of y_{it}^o – the panel of regression residuals of $\hat{y}_{it}^o$ from running y_{it} on G_t – does not include any significant common factor either, as long as the least squares estimator for the factor loading coefficients is consistent. That is,

$$\hat{y}_{it}^o = y_{it}^o + (\alpha_{1i} - \hat{\alpha}_{1i}) G_{1t} + (\alpha_{2i} - \hat{\alpha}_{2i}) G_{2t} = y_{it}^o + O_p\left(T^{-1/2}\right), \tag{2}$$

where $\hat{\alpha}_{1i}$ and $\hat{\alpha}_{2i}$ are the least squares estimates for α_{1i} and α_{2i}, respectively. Even though $\hat{y}_{it}^o$ has two common factors in the finite sample, asymptotically $\hat{y}_{it}^o$ does not have any common factors since the common components vanish asymptotically. We call such factors 'asymptotically weak factors.'

Let x_{it}^o be the random variables which satisfy Bai and Ng (2002)'s Assumption C for the idiosyncratic components. Define $x_{it} = \psi_i' Z_t + x_{it}^o$, where ψ_i and Z_t are factor loadings and common factors of x_{it}, respectively. Then, formally, the asymptotically weak factor can be defined as

Definition: (Asymptotically Weak Factors) *x_{it} has asymptotically weak factors if and only if $\psi_i' Z_t = O_p\left(C_{NT}^{-1}\right)$ where $C_{NT} = \min\left[\sqrt{N}, \sqrt{T}\right]$.*

Note that Chudik and Pesaran (2013) use the terminology of 'weak factor' to define the cross-sectionally weak dependence where the common factor is $O_p(1)$ but the factor loadings are $O_p\left(N^{-1/2}\right)$. Hence, the notion of asymptotically weak factors used in this paper is weaker than the concept of 'weak factor.'

Next, the following lemma can be directly established. Recall that in the beginning of this section we defined $\#(x_{it})$ as the true factor number of x_{it} and $\hat{\#}(x_{it})$ as the estimator for the factor number of x_{it}.

Lemma 1 (Asymptotic Factor Number for Weak Factors) As $N, T \to \infty$ jointly,

$$\lim_{N,T\to\infty} \Pr\left[\hat{\#}(x_{it}) = 0\right] = 1. \tag{3}$$

See the Appendix for the proof. Intuitively, if x_{it} has only asymptotically weak factors, then asymptotically the cross-sectional dependence among x_{it} is equivalent to that among x_{it}^o, which leads to (3). According to Lemma 1, it becomes clear that $\Pr\left[\hat{\#}(\hat{y}_{it}^o) = 0\right] \to 1$ as $N, T \to \infty$. Hence, if G_t are true factors of y_{it}, then the regression residuals $\hat{y}_{it}^o$ should not have any strong factors. However, the opposite is not true in general. Consider a variable W_t which is not correlated with y_{it} at all. Then, as long as W_t is included as a regressor with G_t, the new regression residuals will only have asymptotically weak factors. That is, consider the following regression:

$$y_{it} = \alpha_{1i} G_{1t} + \alpha_{2i} G_{2t} + \alpha_{3i} W_t + y_{it}^o,$$

and define the new residuals as

$$\hat{y}_{it}^o = y_{it} - \hat{\alpha}_{1i} G_{1t} - \hat{\alpha}_{2i} G_{2t} - \hat{\alpha}_{3i} W_t = y_{it}^o + o_p(1),$$

where $\hat{\alpha}_{3i}$ is the least squares estimate for α_{3i}. Since $\hat{\alpha}_{3i} \to^p 0$ as $T \to \infty$, W_t becomes an asymptotically weak factor of $\hat{y}_{it}^o$. However, the asymptotically weak factor status of W_t does not imply that it is a common factor of y_{it}.

Accordingly, our interest becomes the identification of variables one at a time. If the latent factors were known, such identification could be achieved. Note that G_t is not observable but can be estimated by $H'^{-1}\hat{F}_t$, where H is the invertible 2×2 rotation/rescaling matrix defined above (in this example). Let $\hat{\lambda}_i$ be the PC estimators for α_i. Rewrite (1) as

$$y_{it} = \hat{\lambda}_i' \hat{F}_t + \hat{y}_{it}^o, \tag{4}$$

where the residual, $\hat{y}_{it}^{o}$, is defined as

$$\hat{y}_{it}^{o} = y_{it}^{o} - \left(\hat{\lambda}_i' - \alpha_i' H'^{-1}\right)\hat{F}_t - \alpha_i' H'^{-1}\left(\hat{F}_t - H' G_t\right). \tag{5}$$

It is well-known that under the suitable conditions given in Bai (2003), $\hat{\lambda}_i' - \alpha_i' H'^{-1} = O_p\left(T^{-1/2}\right)$ and $\left(\hat{F}_t - H' G_t\right) = O_p\left(N^{-1/2}\right)$. In other words, the regression residuals of $\hat{y}_{it}^{o}$ in (5) also have only asymptotically weak factors.

Here, we rewrite $\hat{F}_t$ as a linear function of G_t and error terms.

$$\begin{bmatrix} \hat{F}_{1t} \\ \hat{F}_{2t} \end{bmatrix} = \begin{bmatrix} h_{1,1} & h_{1,2} \\ h_{2,1} & h_{2,2} \end{bmatrix} \begin{bmatrix} G_{1t} \\ G_{2t} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix},$$

where from Bai (2003) Theorem 1, ε_{1t} and ε_{2t} are $O_p\left(N^{-1/2}\right)$ if $\sqrt{N}/T$ approaches zero as $N, T \to \infty$ or $O_p\left(T^{-1}\right)$ otherwise and where $H = [h_{i,j}]$. In order to perform the identification one at a time, consider regressing y_{it} on both G_{1t} and $\hat{F}_{2t}$ in the following equation.

$$y_{it} = \alpha_{1i}^{*} G_{1t} + \alpha_{2i}^{*} \hat{F}_{2t} + u_{it}^{*}, \tag{6}$$

where $\alpha_{1i}^{*} = \alpha_{1i} - h_{2,2}^{-1} h_{2,1} \alpha_{2i}$ and $\alpha_{2i}^{*} = h_{2,2}^{-1} \alpha_{2i}$ and where $H = [h_{i,j}]$. Because $\varepsilon_{2t} = O_p\left(N^{-1/2}\right)$, $\hat{\alpha}_{1i}^{*}$ and $\hat{\alpha}_{2i}^{*}$ are consistent as $N, T \to \infty$. Hence, as long as $h_{2,2}$ doesn't approach zero, $\hat{u}_{it}$ will have an asymptotically weak factor structure and $\Pr\left[\hat{\#}\left(\hat{u}_{it}\right) = 0\right] \to 1$ as $N, T \to \infty$. Interestingly, a similar result can be found if $\hat{F}_{2t}$ is replaced by $\hat{F}_{1t}$ as long as $h_{1,2}$ doesn't approach zero. Hence, this strategy will not allow us to separately identify which latent factor corresponds to which estimated factor since the estimated factors are only estimated up to a rotation. However, this process will allow us to identify which time series is a latent factor, and this identification is the primary goal of the paper.

Next, we consider the following alternative case. Define $L_{1t} = G_{1t} + v_t$, where $v_t = O_p(1)$. Due to the random error of v_t, L_{1t} is not a true factor. Similar to (6), we can consider the following regression.

$$y_{it} = \alpha_{1i}^{o} L_{1t} + \alpha_{2i}^{o} \hat{F}_{2t} + u_{it}^{o}. \tag{7}$$

It is straightforward to see that $\hat{\alpha}_{1i}^{o} \nrightarrow_p \alpha_{1i}$ as $N, T \to \infty$. Hence, $\hat{u}_{it}^{o}$ will have a factor structure which is not asymptotically weak, and $\Pr\left[\hat{\#}\left(\hat{u}_{it}^{o}\right) = 0\right] \to 0$ as $N, T \to \infty$.

In sum, as long as we are interested in identifying whether or not an observed time series is one of the true factors, we do not need any identification restriction on the rotation matrix H. See Bai and Ng (2013) for restriction conditions under which the latent factors can be estimated asymptotically without rotation.

We formally present the identification procedure in the next section.

3 Definitions and Identification Procedure

Before we start to provide identification procedures and strategies, we provide conceptual definitions of the empirical true factors: Dominant and approximate dominant leaders.

3.1 Definitions

Let $P_t = [P_{1t}, ..., P_{mt}]'$ be the $m \times 1$ vector of potential true factors which researchers want to examine. Note that m is not necessarily equal to r. We will discuss the reason shortly. If P_t are the true factors, then the inclusion of P_t into the panel data $y_t = [y_{1t}, ..., y_{Nt}]$ always leads to more accurate estimation of the common factors (See Boivin and Ng, 2006). Also, it is possible that a few leaders are the true common factors of the panel data. An example of this endogenous estimation appears in Gaibulloev, Sandler and Sul (2013), which finds that transnational terrorism in Lebanon is the main determinant of transnational terrorism for the rest of the world. Hence, without loss of generality, we can include P_t as a part of the panel data $\{y_{it}\}$ and re-order them as $\{y_{1t}, ..., y_{mt}, y_{m+1,t}, ..., y_{N+m,t}\}$ so that the first m individuals are the potential true common factors to $\{y_{it}\}$.

Definition (Dominant Leaders): *The jth unit is an exact dominant leader if and only if $y_{jt} = G_{jt}$.*

In general, the maximum number of dominant leaders should be the same as the number of true common factors. However, sometimes the number of leaders can be larger than the number of the factors, especially when there are many approximate dominant leaders. These can be defined as,

Definition (Approximate Dominant Leaders): *The jth unit becomes an approximate dominant leader for the jth true factor if and only if $G_{jt} = y_{jt} + \zeta_{jt}$ for $j = 1, ..., r$ where $\zeta_{jt} = \epsilon_{jt}/\sqrt{T}$ and $Var(\epsilon_{jt}) = \sigma_j^2$ where $\bar{\sigma}^2 = \max \sigma_j^2$ and $0 < \bar{\sigma}^2 < \infty$ even as $N, T \to \infty$.*

When $\sigma_j^2 = 0$, the jth unit of $\{y_{1t}, ..., y_{N+m,t}\}$ becomes a dominant leader. The non-zero variance of σ_j^2 implies that the jth unit may lose his leadership temporarily for a fixed set of time periods, $\mathcal{T}$. That is,

$$y_{jt} = \begin{cases} G_{jt} & \text{if } t \notin \mathcal{T} \\ G_{jt} + \epsilon_{jt}^* & \text{if } t \in \mathcal{T} \end{cases} \quad \text{for Var}\left(\epsilon_{jt}^*\right) = \sigma_j^2 > 0.$$

The number of elements of $\mathcal{T}$ will be denoted as p, which is fixed as $N, T \to \infty$. Thus y_{jt} is not the leader for p time periods. Then the variance of the deviation between the common factor and

the dominant leader, $y_{jt} - G_{jt}$, becomes

$$\sigma_{j,T}^2 = \mathrm{E}\left[\frac{1}{T}\sum_{t=1}^{T}(y_{jt} - G_{jt})^2\right] = \frac{p\sigma_j^2}{T} \text{ for a small constant } p > 1.$$

When there are approximate dominant leaders, then the number of these leaders can be larger than the number of true common factors. Also note that it is impossible to asymptotically distinguish approximate dominant and dominant leaders.

3.2 Identification Procedures

The identification procedures differ depending on whether or not potential leaders are given or selected by the researcher. We first consider the simplest case, where potential leaders are given or known. We also assume that the number of the true factors is known. This assumption is fairly reasonable since Bai and Ng (2002)'s criteria perform fairly well when the panel data are rather homogeneous.[1] Note that we are identifying whether or not a time series is either a dominant or approximate dominant leader for $\hat{F}_{jt}$ for $j = 1, ..., r$, since the true statistical factors are unknown.

For clarity, we use a case where $r = 2$ but $m = 3$ throughout this section. That is, G_t and $\hat{F}_t$ are 2×1 vectors but the potential factor, P_t, is the 3×1 vector.

Even when H is an identity matrix, the first PC estimator $\hat{F}_{1t}$ can be G_{2t} depending on the values of factor loadings in (1) and the variance-covariance matrix of G_t, Ω. However, regardless of the ordering, the point of interest becomes whether or not $\hat{F}_{1t}$ can be identified by P_{st} for $s = 1, .., m$. Let

$$G_{jt} - P_{st} = \gamma_{js}\epsilon_{st}/\sqrt{T} + \delta_{js}\xi_{st}, \tag{8}$$

where $\mathrm{Var}(\epsilon_{st}) = \sigma_{\epsilon,s}^2$ and $\mathrm{Var}(\xi_{st}) = \sigma_{\xi,s}^2$ for positive and finite constants $\sigma_{\epsilon,s}^2$ and $\sigma_{\xi,s}^2$. By definition, if $\delta_{js} = 0$ but $\gamma_{js} \neq 0$, then P_{st} becomes the approximate dominant leader for G_{jt}. Note that all P_{st} could be approximate dominant leaders for only G_{1t} or G_{2t}. Alternatively, some of P_{st} (for example, P_{1t} and P_{2t}) are the approximate dominant leaders for G_{1t} and the other P_{st} (for example, P_{3t}) become the approximate dominant leaders for G_{2t}.

To identify whether or not P_{st} is a dominant or approximate dominant leader, we suggest examining whether or not the regression residuals from the following regressions have any strong

[1] When we refer to panel homogeneity in this paper, we are specifically refering to the panel being constructed of one central variable, such as state-level unemployment rates over time. In terms of the factor structure, homogeneity appears when the order of intergration is the same across cross-secitonal units and when the idiosyncratic variances are not seriously heterogeneous.

common factors.

$$y_{it} = \beta_{2,si} P_{st} + \alpha^*_{2,si} \hat{F}_{2t} + y^o_{2s,it}, \tag{9}$$

$$y_{it} = \beta_{1,si} P_{st} + \alpha^*_{1,si} \hat{F}_{1t} + y^o_{1s,it}, \tag{10}$$

Suppose that $P_{1t} = G_{1t}$ exactly. Even in this case it is possible that $P_{1t} \neq \hat{F}_{1t}$ but instead P_{1t} becomes a linear function of $\hat{F}_{1t}$ and $\hat{F}_{2t}$. Depending on the values α_{1i} and α_{2i}, the estimated number of the common factors in either one or both of $\hat{y}^o_{2s,it}$ and $\hat{y}^o_{1s,it}$ becomes zero. For $r \geq 2$, (9) and (10) can be written as

$$y_{it} = \beta_{j,si} P_{st} + \alpha^{*\prime}_{i,-j} \hat{F}_{-j,t} + y^o_{js,it} \text{ for } j = 1, ..., r, \tag{11}$$

where $\hat{F}_{-j,t} = \left[\hat{F}_{1t}, ..., \hat{F}_{j-1,t}, \hat{F}_{j+1,t}, ..., \hat{F}_{rt}\right]$ and $\alpha^*_{s,-j} = \left[\alpha^*_{i1}, ..., \alpha^*_{ij-1}, \alpha^*_{ij+1}, ..., \alpha^*{}_{irt}\right]$. When $r = 1$, $\hat{F}_{-j,t}$ and $\alpha^{*\prime}_{i,-j}$ are not present in (11). Thus more formally, we have

Theorem 1 (Identification of Estimated Factors: Known Potential Leaders) *Under the assumptions in Bai and Ng (2002),*

(i) If $\delta_{ij} = 0$, *then*

$$\lim_{N,T\to\infty} \Pr\left[\hat{\#}\left(\hat{y}^o_{1j,it}\right) = 0 \text{ or } \hat{\#}\left(\hat{y}^o_{2j,it}\right) = 0 \text{ or,..., } \hat{\#}\left(\hat{y}^o_{rj,it}\right) = 0\right] = 1. \tag{12}$$

(ii) If $\delta_{ij} \neq 0$, *then*

$$\lim_{N,T\to\infty} \Pr\left[\hat{\#}\left(\hat{y}^o_{1j,it}\right) = 0 \text{ or } \hat{\#}\left(\hat{y}^o_{2j,it}\right) = 0 \text{ or,..., } \hat{\#}\left(\hat{y}^o_{rj,it}\right) = 0\right] = 0 \tag{13}$$

Many times when leaders are unknown and N is large, applying our criterion to each individual in the panel could lead to over-estimation of the number of approximate dominant leaders, since the 'size' of the procedure is non-zero. One solution to this problem is to run the following regression:

$$\hat{F}_{st} = c_{ss} P_{jt} + c_{s,-s} \hat{F}_{-s,t} + \varepsilon^*_{st} \text{ for each } P_j \text{ and for each } s = 1, .., r, \tag{14}$$

where c_{ss} and $c_{s,-s}$ are regression coefficients. Next, obtain the R^2-statistics. For each factor, $\hat{F}_{st}$, the individuals, P_j, with high R^2 values have high estimated partial correlation to the factor. Choosing to test only these individuals avoids over-estimation of the number of approximate dominant leaders. It is easy to show that this procedure is consistent as N and T go to infinity.

By running (9) and (10) for $r = 2$, or more generally (11) for any $r > 2$, approximate dominant leaders can be identified for any G_{st}, but the dominant leaders for a particular G_{st} are not known. To distinguish the leaders, we suggest the following method to cluster approximate dominant leaders to each G_{st}.

3.3 Clustering Method

For clear exposition, we continue to use the above example where the number of approximate dominant leaders is three and the number of true common factors is two. Since the true factors are unknown, it is impossible to identify which approximate dominant leader P_{jt} for $j = 1, 2, 3$ is associated with G_{st} for $s = 1, 2$. However, there is a way to cluster P_{jt} into two groups. Let $P_{t,(1,2)} = [P_{1t}, P_{2t}]'$, $P_{t,(1,3)} = [P_{1t}, P_{3t}]'$, and $P_{t,(2,3)} = [P_{2t}, P_{3t}]'$. Consider the following regressions.

$$\begin{aligned} y_{it} &= \lambda^{*\prime}_{i,(1,2)} P_{t,(1,2)} + y^{o}_{it,(1,2)}, \\ y_{it} &= \lambda^{*\prime}_{i,(1,3)} P_{t,(1,3)} + y^{o}_{it,(1,3)}, \\ y_{it} &= \lambda^{*\prime}_{i,(2,3)} P_{t,(2,3)} + y^{o}_{it,(2,3)}. \end{aligned} \tag{15}$$

where $\lambda^{*}_{i,(l1,l2)}$ is the 2×1 vector of the regression coefficients for $l1 = 1, 2$ and $l2 = 2, 3$ but $l1 \neq l2$. If P_{1t} and P_{2t} are the approximate dominant leaders for the same common factor (either G_{1t} or G_{2t}), then as $N, T \to \infty$, the estimated number of the common factors of the regression residuals of $\hat{y}^{o}_{it,(1,2)}$ becomes 1. That is,

$$\lim_{N,T\to\infty} \Pr\left[\#\left(\hat{y}^{o}_{it,(1,2)}\right) = 1\right] = 1. \tag{16}$$

However, if P_{1t} and P_{2t} are the approximate dominant leaders for G_{1t} and G_{2t}, respectively, then

$$\lim_{N,T\to\infty} \Pr\left[\#\left(\hat{y}^{o}_{it,(1,2)}\right) = 0\right] = 1. \tag{17}$$

For example, suppose that P_{1t} and P_{2t} are the approximate dominant leaders for G_{1t} and P_{3t} is the approximate dominant leader for G_{2t}, then as $N, T \to \infty$,

$$\begin{aligned} &\lim_{N,T\to\infty} \Pr\left[\#\left(\hat{y}^{o}_{it,(1,2)}\right) = 1\right] = 1, \\ &\lim_{N,T\to\infty} \Pr\left[\#\left(\hat{y}^{o}_{it,(1,3)}\right) = 0\right] = 1, \\ &\lim_{N,T\to\infty} \Pr\left[\#\left(\hat{y}^{o}_{it,(2,3)}\right) = 0\right] = 1. \end{aligned} \tag{18}$$

This method is easily extended pairwise to cases where there are more or fewer leaders. Note that it is impossible to identify whether or not each P_{jt} is a specific G_{it} without further identifying restrictions since G_{it} is unknown.

3.4 Comparison to Extant Testing Method

Bai and Ng (2006) consider a similar factor identification problem. Their test is originally designed to examine whether or not observed vectors of variables, P_t, are true factors, G_t. They do not explicitly discuss whether P_t can be members of $\{y_{it}\}$ and consider only "outside" macro and

financial variables in their empirical application. However, it is straightforward to extend their method to identify dominant leaders in the leadership model.

Their test is based on the following. If P_{jt} is one of the exact dominant leaders, then it should hold that

$$P_{jt} = A_j G_t + \pi_{jt}, \text{ for } \pi_{jt} = 0 \text{ all } t. \tag{19}$$

Their test is examining whether or not π_{jt} are statistically zero for all t. $\hat{P}_{jt}$ is constructed to be the fitted values from the following regression equation

$$P_{jt} = B_j \hat{F}_t + \pi_{jt}. \tag{20}$$

Hence, their test is based on the following statistic

$$\tau_t(j) = \frac{\hat{P}_{jt} - P_{jt}}{\sqrt{V\left(\hat{P}_{jt}\right)}}. \tag{21}$$

In their Monte Carlo simulation, they found that the performance of the max τ_t test works well. The max τ_t test is defined as

$$M(j) = \max_{1 \le t \le T} |\hat{\tau}_t(j)|, \tag{22}$$

where $\hat{\tau}_t(j)$ is obtained with the estimate of $V\left(\hat{P}_{jt}\right)$.

In the case of approximate dominant leaders, the above test fails. Let $\pi_{jt} = \epsilon_{jt}/\sqrt{T}$ where $\text{Var}(\epsilon_{jt}) > 0$ even as $N, T \to \infty$. Thus,

$$P_{jt} = A_j H^{-1\prime} \hat{F}_t - A_j H^{-1\prime} \left[\hat{F}_t - H' G_t\right] + T^{-1/2} \epsilon_{jt}.$$

To find $\hat{A}_j$, just use the standard formula for least squares, or

$$\hat{B}_j = T^{-1} \hat{F}' P_j = H^{-1} A_j - T^{-1} \hat{F}' \left[\hat{F} - GH\right] H^{-1} A_j + T^{-3/2} \hat{F}' \epsilon_j.$$

Interestingly,

$$\sqrt{N}\left(\hat{B}_j - H^{-1} A_j\right) = -N^{1/2} T^{-1} \hat{F}' \left[\hat{F} - GH\right] H^{-1} A_j + N^{1/2} T^{-3/2} \hat{F}' \epsilon_j = O_p\left(\sqrt{N}/T\right),$$

which is exactly the same as in the exact dominant leaders case, and accordingly, both the exact and the approximate cases require the condition that $\sqrt{N}/T \to 0$ as $N, T \to \infty$. However, unique to the approximate dominant leaders case is that $\sqrt{N}\left(\hat{P}_{jt} - P_{jt}\right)$ resolves into

$$\sqrt{N}\left(\hat{P}_{jt} - P_{jt}\right) = \sqrt{N}\left(A_j H^{-1\prime}\left[\hat{F}_t - H' G_t\right]\right) - \sqrt{N}\left(A_j H^{-1\prime} - \hat{B}_j\right)\hat{F}_t + \epsilon_{jt}\sqrt{N/T},$$

which can be further reduced to[2]

$$\sqrt{N}\left(\hat{P}_{jt} - P_{jt}\right) = \sqrt{N} A_j H^{-1\prime}\left[\hat{F}_t - H' G_t\right] + O_p\left(\sqrt{N/T}\right). \tag{23}$$

[2] We thank an anonymous referee for pointing out this problem with Bai and Ng (2006)'s tests.

$\sqrt{N}A_jH^{-1\prime}\left[\hat{F}_t - H'G_t\right]$ is the term for which the variance is estimated in (21). However, the $O_p\left(\sqrt{N/T}\right)$ term must approach zero in order for $\tau_t(j)$ to converge to $N(0,1)$, which is a very restrictive condition. Therefore, in order to detect approximate dominant leaders, the $M(j)$ test requires that N/T approaches zero as $N,T \to \infty$, and even in the exact dominant leaders case, N/T^2 must approach zero as $N,T \to \infty$.

3.5 Utilizing Other Factor Number Estimation Methods

In this paper, the Bai and Ng (2002) criterion is used to estimate the number of factors in the residuals. Recently, other procedures have been suggested for estimating the number of factors in panel data sets. For instance, Hallin and Liska (2007), Onatski (2009, 2010), and Ahn and Horenstein (2013) have suggested alternative methods. Here we briefly discuss other factor number estimation methods.

Let ϱ_i be the ith largest eigenvalue of the $(NT)^{-1}\hat{y}^{o\prime}\hat{y}^o$ matrix where $\hat{y}^o$ is the $T \times N$ matrix of the residual of $\hat{y}^o_{it}$. The Bai and Ng (2002) criteria maximize the following statistic.

$$Q_{BN} = \underset{0\leq k\leq k_{\max}}{\arg\min}\left[\ln\left(\sum\nolimits_{i=k+1}^{h}\varrho_i\right) + k \times p(N,T)\right],$$

where $p(N,T)$ is a penalty or threshold function, $h = \min[T,N]$, and $k_{\max}$ is the maximum factor number usually assigned by a practitioner.

Onatski (2009, 2010) propose two methods for consistent factor number estimation. Onatski (2009)'s test is for the general dynamic factor model but can be used for the static factor model under the assumption of Gaussian idiosyncratic errors. Onatski (2010)'s criteria is for the approximate static factor model which does not require Gaussianality. Onatski (2010)'s estimator is defined as

$$Q_{Onat} = \max\{i \leq k_{\max} : \varrho_i - \varrho_{i+1} \geq \delta\},$$

where δ is a fixed positive number. See Onatski (2010) for a detailed procedure of how to calibrate δ.

Hallin and Liska (2006) propose modified criteria for the factor number in the general dynamic factor model based on Bai and Ng (2002) criteria. However, their method can be directly utilized in the static factor model, as demonstrated in Alessi, Barigozzi and Capasso (2010). The Hallin and Liska (2006) criterion uses subsamples of the panel, $0 < N_1 < N_2 < \cdots < N_L = N$ and $0 < T_1 < T_2 < \cdots < T_M = T$. Denote $\varrho_i^{(l,m)}$ as the eigenvalues of $(N_lT_m)^{-1}\hat{y}^{o\prime}\hat{y}^o$ computed using only the values of $\hat{y}^o_{it}$ where $i \leq N_l$ and $t \leq T_m$. Let

$$\hat{k}_{BN}(c,l,m) = \underset{0\leq k\leq k_{\max}}{\arg\min}\left[\ln\left(\sum\nolimits_{i=k+1}^{h}\varrho_i^{(l,m)}\right) + k \times c \times p(N_l,T_m)\right].$$

This is the subsample, scaled analog of the Bai and Ng (2002) IC criterion. $\hat{k}_{BN}$ is a function of a positive constant c which controls the sensitivity of the estimator. When c is small, $\hat{k}_{BN}$ does not penalize extra factors, so the estimator finds $k_{\max}$ as the number of factors. When c is large, $\hat{k}_{BN}$ over-penalizes the factors, so $\hat{k}_{BN}$ finds zero factors in the residual. The $S(c)$ function is defined by

$$S(c) = \left(\frac{1}{LM} \sum\nolimits_{l,m} \left(\hat{k}_{BN}(c,l,m) - \frac{1}{LM} \sum\nolimits_{l,m} \hat{k}_{BN}(c,l,m) \right)^2 \right)^{1/2}.$$

Under Bai and Ng (2002) there is no control for the coefficient, c, before the penalty function. In other words, Bai and Ng (2002) choose the value: $\hat{k}_{BN} = \hat{k}_{BN}(1, L, M)$. Hallin and Liska (2007) use subsamples to choose $c = c_o$ in a region where S vanishes. Because of this control, the penalty function is less sensitive to the size of N and T. $\hat{k}_{HL}$ is equal to $\hat{k}_{BN}$ evaluated at c_o, that is to say $\hat{k}_{HL} = \hat{k}_{BN}(c_o, L, M)$.

Ahn and Horenstein (2013)'s criteria are free from the choice of the threshold function $p(N,T)$. They proposed the following two criteria

$$Q_{AH,ER} = \max_{0<k\leq k_{\max}} \varrho_k / \varrho_{k+1},\ Q_{AH,GR} = \max_{0<k\leq k_{\max}} \ln\left[V_{k-1}/V_k\right] / \ln\left[V_k/V_{k+1}\right],$$

where $V_k = \sum_{i=k+1}^{h} \varrho_i$.

4 Practical Suggestions and Monte Carlo Studies

This section summarizes the procedure we discussed in earlier sections and reports the results of Monte Carlo simulations.

4.1 Identifying Procedures

Here we present a step-by-step procedure used in the Monte Carlo simulation and empirical exercises.

Step 1 (Estimation of Factor Number and Common Factors):

Before estimating the factor number and common factors, one should standardize each time series by its standard deviation. In our empirical examples, we usually take logs, difference, and then standardize the sample. The factor number, r, and the common factors should then be estimated. For the factor number estimation, we use Bai and Ng (2002)'s IC_2 because it performs better than the other Bai and Ng criteria. We don't report the simulation results with Hallin and Liska's IC_2 here but note that all detailed results are available on the authors' website.

Step 2 (Identifying Potential Leaders by using R^2):

Here we select potential leaders by using an R^2 criterion. From (14), the following regressions can be performed for the case of $r = 2$.

$$\hat{F}_{1t} = c_{11}P_{jt} + c_{12}\hat{F}_{2t} + \varepsilon^*_{1t} \tag{24}$$

$$\hat{F}_{2t} = c_{21}P_{jt} + c_{22}\hat{F}_{1t} + \varepsilon^*_{2t} \tag{25}$$

If $P_{jt} = G_{1t}$, or $P_{jt} = G_{1t} + \zeta_{jt}$ for $\zeta_{jt} = \epsilon_{jt}/\sqrt{T}$, then as $N, T \to \infty$, the variance of ε^*_{1t} approaches zero. Alternatively, if $P_{jt} = G_{1t} + \epsilon_{jt}$ where ϵ_{jt} has a finite variance, then the variance of ε^*_{1t} should not be close to zero. Similarly, if $P_{jt} = G_{2t}$ or $P_{jt} = G_{2t} + \zeta_{jt}$, then the variance of ε^*_{2t} also goes to zero as $N, T \to \infty$. By utilizing this fact, we select the first m potential leaders for which R^2 is the highest. The size of m must be selected to be larger than r since each factor could have multiple approximate dominant leaders. We commonly select m to be around 10% of the size of N. Later we will show that this R^2 method detects potential leaders very precisely.

Step 3 (Identifying Approximate Dominant Leaders):

Run (9) or (10) and obtain the regression residuals. Check whether or not the estimated factor number is zero. Following Theorem 1, identify whether or not P_{jt} is an approximate dominant leader. Collect all of the dominant leaders found.

Step 4 (Clustering Approximate Dominant Leaders):

Calculate the correlation matrix among approximate dominant leaders if they are many. Group the leaders by running (15).

See the next section for a demonstration of the above four steps.

4.2 Monte Carlo Results

By means of Monte Carlo simulation, we verify our theoretical claims and investigate the finite sample performance. All pseudo data used in simulations are generated from the following data generating process with some restrictions.

$$y_{it} = \alpha_{1i}G_{1t} + \alpha_{2i}G_{2t} + \sqrt{\theta}y^o_{it}, \tag{26}$$

where θ controls the signal to noise ratio. As θ increases, the signal-to-noise ratio also increases.

The common factors are serially correlated and also dependent on one another. Specifically, we generate G_t as follows.

$$G_t = \Lambda W_t,\ W_{st} = \rho_s W_{st-1} + g_{st},\ \text{for } s = 1, 2,$$

where $g_{st} \sim iid\mathcal{N}\left(0, 1-\rho_s^2\right)$ and Λ is the upper triangular matrix of the Cholesky decomposition of Ω, which is the covariance and variance matrix of G_t. That is,

$$\Omega = \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{12} & \Omega_{22} \end{bmatrix}.$$

The factor loadings are generated from

$$\alpha_{ki} \sim iidN\left(\mu_{\alpha k}, \sigma_{\alpha k}^2\right) \text{ for } k = 1, 2.$$

For the case of asymptotically weak factors, we set the variance of $\sigma_{\alpha k}^2$ to be dependent on either T or N. The value of $\mu_{\alpha k}$ is set to one for the asymptotically weak factors case, whereas it is set to zero for the other cases.

The idiosyncratic errors are generated from

$$y_{it}^o = \phi y_{it-1}^o + v_{it} + \beta \sum_{j\neq 0, j=-J}^{J} v_{i+j,t},$$

where $v_{it} \sim iidN\left(0, \sigma_i^2\left[1-\phi^2\right] / \left[1+2J\beta^2\right]\right)$. To avoid the impact of the initial observation and boundary condition, we generate $(T+100)\times(N+20)$ pseudo variables. Next we discard the first T_o observations over time and select the middle of N cross-sectional units.

Note that the data generating process is similar to Bai and Ng (2002), Onatski (2010) and Ahn and Horenstein (2013). Here we allow the common factors to be correlated with one another. For all simulations, we set the simulation size to be 2,000 with $N = [25, 50, 100, 200]$ and $T = [25, 50, 100, 200]$. The signal-to-noise ratio, θ, is set to be unity; otherwise the second factor is not well-estimated. While the above data generating process is quite general, we will be considering three restrictions on the idiosyncratic errors.

Case I (Independent, Identically Distributed Errors): The idiosyncratic errors have no serial correlation ($\phi = 0$) and no cross-sectional dependence ($\beta = J = 0$), and the errors are homoskedastic with unit variance ($\sigma_i^2 = 1$ for all i).

Case II (Serially Correlated Errors): The idiosyncratic errors have serial correlation with $\phi = 0.5$ and no cross-sectional dependence ($\beta = J = 0$), and the errors are homoskedastic with unit variance ($\sigma_i^2 = 1$ for all i).

Case III (Serially Correlated and Cross-Sectionally Dependent Errors): The idiosyncratic errors have serial correlation with $\phi = 0.5$ and cross-sectional dependence $\beta = 0.1$ and $J = 4$, and the errors are homoskedastic with unit variance ($\sigma_i^2 = 1$ for all i).

In Bai and Ng (2002), cross-sectional dependence in the errors is essentially set so that $\beta = 0.2$ and $J = 10$, and in Ahn and Horenstein (2013), it was set so that $\beta = 0.2$ and $J = 8$. The data generating process described in Case III is slightly different because we consider small T and N cases. As N and T grow, the finite sample performance is less influenced by the values of β and J. Unless otherwise stated, we will be assuming that $\rho_1 = \rho_2 = 0.5$.

4.2.1 The Factor Number Estimation of Asymptotically Weak Factors

Here, we verify Lemma 1 first, and then we examine the finite sample performance of factor number estimation on a panel of data which contains only asymptotically weak factors. We consider only the case where there is a single factor. An asymptotically weak factor is generated by setting $\alpha_{1i} \sim iidN(1, 1/T)$ or $\alpha_{1i} \sim iidN(1, 1/N)$. To save space, we report the former case only. Even though the variance of α_{1i} is decreasing over time, the mean of α_{1i} is not equal to zero. Hence the estimated factor number of y_{it}, $\hat{\#}(y_{it})$ should be one asymptotically. Meanwhile, the cross-sectionally demeaned series $\tilde{y}_{it} = y_{it} - N^{-1}\sum_{i=1}^{N} y_{it}$ has only an asymptotically weak factor. Hence by Lemma 1, the estimated factor number of $\tilde{y}_{it}$, $\hat{\#}(\tilde{y}_{it})$, should be zero as $N, T \to \infty$. We also consider the case where the variance of the factor loadings, $\sigma_{\alpha 1}^2$, is small. That is, we set $\alpha_{1i} \sim iidN(1, 0.2)$. Even though $\sigma_{\alpha 1}^2$ is small, the cross-sectionally demeaned series $\tilde{y}_{it}$ is found to have a single strong factor as $N, T \to \infty$.

Table 1 reports the estimated probabilities. The first three columns in Table 1 show the estimated probability of $\hat{\#}(y_{it}) = 1$. For Case I, even with small N and T, the factor number is very accurately estimated. As we introduce serial dependence in the idiosyncratic error (Case II), the performance of IC_2 deteriorates, especially when T is small. However, as T increases, the factor number is estimated more accurately. Note that as Bai and Ng (2002) showed, the factor number is over-estimated. When the idiosyncratic error has both serial and cross-sectional dependence (Case III), the IC_2 performs badly either with small N or T. Again, as both N and T increase, the accuracy of IC_2 is restored. The next three columns in Table 1 display the frequencies of finding that $\hat{\#}(\tilde{y}_{it}) = 0$. As Lemma 1 shows, the number of the common factors should be zero since the cross-sectionally demeaned series, $\tilde{y}_{it}$, has only an asymptotically weak factor. Overall the factor number is accurately estimated except when T is small. The last three columns report the estimated probability of $\hat{\#}(\tilde{y}_{it}) = 0$ with $\sigma_{\alpha 1}^2 = 0.2$. When both N and T are small, the probability of false estimation is moderately large. However, as either N or T increases, the false estimation probability decreases quickly. Interestingly, the false estimation probability in Case I is higher

than that in Case III. This happens because the factor number is over-estimated more when the idiosyncratic errors are more serially and cross-sectionally dependent.

We considered the case of $\beta = 0.2$ but don't report the result here. Overall the performance is worse when N and T are small, but when N and T are large ($N \geq 100$ and $T \geq 100$), the performance is improved. We also examined the performance of HL_2 and found that it is similar but slightly worse (better) than that of IC_2 when N and T are large (small).

4.2.2 Identifying True Factors with Given Potential Factors

Here we assume that the potential factors and the number of factors are known and evaluate how often the leadership method identifies the potential factor as a leader (either correctly or incorrectly) when the potential factor is truly an exact, approximate, or false leader. We also report the performance of Bai and Ng (2006)'s $M(j)$ test under the same conditions for comparison.

We set $\alpha_{ki} \sim iidN(0,1)$ (i.e., $\mu_{\alpha k} = 0$, $\sigma^2_{\alpha k} = 1$). We also set $[\Omega_{11}, \Omega_{12}, \Omega_{22}] = [2, 0.5, 1]$, $\theta = 1$, and $\rho_s = 0.5$ for $s = 1, 2$. In the 'Exact' case, the potential factors are defined to be $P_{kt} = G_{kt}$ so that both P_{1t} and P_{2t} are exact dominant leaders. In the 'Approximate' case, the potential factors are defined to be $P_{kt} = G_{kt} + \epsilon_{kt}/\sqrt{T}$ so that both P_{1t} and P_{2t} are approximate dominant leaders. In the 'False' case, the potential factors are defined to be $P_{kt} = G_{kt} + \epsilon_{kt}$ so that both P_{1t} and P_{2t} are not leaders. In both the 'Approximate' and 'False' cases, we let $\epsilon_{kt} \sim iidN(0,1)$.

Table 2 reports the estimated probability of how often the IC_2 criterion identifies the first potential factor as a leader. The results seem quite good in Case I as the method correctly identifies both 'Exact' and 'Approximate' factors almost perfectly and only frequently misidentifies 'False' factors when N and T are both greater than 50. Case II tells a similar story; however here correct identification of 'Exact' and 'Approximate' factors happens very often only when N and T are both 50 or greater. Case III is more dramatic and here correct identification of 'Exact' and 'Approximate' factors is only quite likely when N and T are both greater than 50. Note that in Cases II and III, identification of 'False' factors performs almost perfectly, with the possible exception of $N = T = 25$ since the factor number is usually over-estimated in Cases II and III. While also not reported, HL_2's performance initially seems better for small sample sizes. However, when N and T are both large HL_2 has some trouble identifying 'Exact' and 'Approximate' factors when there is cross-sectional dependence in the errors (Case III). These unreported results are all available on the authors' website.

Table 3 reports the size and power for Bai and Ng (2006)'s $M(j)$ test under the same conditions for comparison. Note that the $M(j)$ test requires there to be no serial and cross-sectional dependence. In Case I, the $M(j)$ test works properly, but as serial dependence is introduced (Case II), the $M(j)$ test suffers from somewhat serious size distortion. Under serial and cross-sectional

dependence (Case III), the size distortion of the $M(j)$ test is much worse than that in Case II. As predicted from (23), the $M(j)$ test fails to identify approximate leaders when N is greater than or equal to T. It should be noted that we used the heteroskedastic variance estimator in our estimation. If you control for cross correlated errors instead of heteroskedastic, there would be some improvement to the size in Case III when N is smaller than T. However, since there is also serial correlation here, Case II would not improve by using cross correlated standard errors, and Case III would not improve beyond Case II and would certainly have serious problems as N becomes large. In all cases and for all the selected values of N and T, the power of the $M(j)$ test is almost perfect.

4.2.3 Identifying True Factors when Potential Leaders are Unknown

When the potential leaders are unknown, they must be estimated. To do this, we use the R^2 criterion as discussed in Section 3. The DGP is as defined in (26) with $[\Omega_{11}, \Omega_{12}, \Omega_{22}] = [1, 0.2, 1]$. We generate two approximate dominant leaders for each factor as follows.

$$P_{jt} = G_{1t} + \epsilon_{jt}/\sqrt{T} \text{ for } j = 1, 2, \text{ and } P_{jt} = G_{2t} + \epsilon_{jt}/\sqrt{T} \text{ for } j = 3, 4.$$

The first four approximate factors are included in the panel data y_{it}. That is, $y_{it} = P_{it}$ for $i = 1, 2, 3, 4$. For the remaining y_{it}, we impose the following restriction on the idiosyncratic variance.

$$y_{it}^o \sim iidN\left(0, \sigma_i^2\right), \text{ for } \sigma_i^2 = \frac{1}{T}\sum_{t=1}^{T} C_{it}^2, \; C_{it} = \alpha_i' G_t. \tag{27}$$

Without imposing this restriction, there is always a chance that some y_{it} for $i > 4$ becomes an approximate common factor with high α_{1i} and α_{2i}. We also generate α_is from a uniform distribution without imposing the restriction in (27), but since the results are very similar, we only report this case.

We choose four potential leaders, each by maximizing the R^2 statistics from (24) and (25). Thus, the maximum number of potential leaders becomes eight. Next, we check whether or not each potential leader is truly a common factor by estimating the number of common factors of the regression residuals in (9) and (10).

Table 4 reports the frequencies with which approximate dominant leaders are selected as the true factors by combining the sieve method with the R^2 criterion together. The first column reports the correct inclusion rate that all four approximate dominant leaders are selected as the true factors. The second column shows the frequency that any false dominant leader is selected as a true factor. Evidently, the sieve method with the R^2 criterion suggested in Section 3 works very well. When T is moderately large ($T \geq 50$), the suggested method selects only correct approximate dominant leaders as the true factors.

Finally, Table 5 reports the clustering results. As it shown in Table 4, the accuracy of identifying the true factor is fairly sharp. Hence we assume that the approximate dominant leaders are given. We use the criteria in (16) and (17) since the selected leaders are only four. The first column in Table 5 reports the frequency that the clustering algorithm selects correct members. The second column shows the false inclusion rate. Obviously the criteria in (16) and (17) demonstrate pinpoint accuracy.

5 Empirical Examples: Fama-French Three Factor Model

One of the most popular examples in factor analysis is the Fama-French portfolio theory. Fama and French (1993) found three key factors for portfolio returns, denoted as follows: 'Market', 'SMB', and 'HML'. Hence, if our method is accurate, these three factors should be identified as the true factors. Note that Bai and Ng (2006) failed to identify Market as one of the true factors by using the annual data from 1960 to 1996. We will show shortly that we find that all three factors are well identified by using our proposed method. While these time series are provided on Fama's website, they are constructed so that Market is correlated with the fluctuation of the overall stock market, SMB (small and medium business) is correlated with the fluctuation of small market capitalization stocks, and HML (high minus low) is correlated with the fluctuation of stocks with high book-to-market ratios. We are interested in testing if these empirical factors are actually the underlying unobserved. This is not a leadership model; rather, this is an exogenous factor test. The famous Fama-French three factor model is given by

$$y_{it} = r_{ft} + \alpha_{1i}\left(\mathrm{Mkt}_t - r_{ft}\right) + \alpha_{2i}\mathrm{SMB}_t + \alpha_{3i}\mathrm{HML}_t + y_{it}^o,$$

where r_{ft} is the risk-free return rate. For our analysis, we use annual average value portfolio returns for 96 portfolios plus the three Fama-French factors from 1964 to 2008. The data available on Kenneth French's website begins in 1927 and ends in 2012. Our sample is chosen to avoid missing data before 1964 and structural market changes from the 2008 financial collapse. The returns are demeaned (by time series averages) and standardized. The maximum factor number is set to be 10. Before cross-sectionally demeaning, we estimate the number of common factors in y_{it}. We find 3 factors here, and this result is robust across subsamples. This result is different from Bai and Ng (2006). They considered only 89 portfolios due to missing data and chose rather the largest estimated factor number by using Bai and Ng (2002)'s PC_p criteria where $PC_p = V_k + k \times p\left(N, T\right)$.

According to the theory, y_{it} shares the same common factor of r_{ft}. Hence we take off the cross-sectional average first, and then choose Mkt, SMB and HML as the known potential factors. We estimate the number of common factors after standardizing the sample. We denote this sample as $\tilde{y}_{it}$ to distinguish it from the sample y_{it} where the cross sectional averages are not taken off. Table

6 reports the summary of the results. First, both IC_2 and HL_2 estimate the factor number as two, surprisingly. This result does not change at all depending on the different ending years. This result implies that if the Fama-French three factor model is indeed correct, then one of the three factor loadings must be homogeneous. Moreover, if we don't take off the cross-sectional average, then we find three factors.

To identify the factors, we ran (15). That is, the panel is regressed on each hypothesized factor, Mkt, SMB, and HML:

$$\tilde{y}_{it} = \gamma_{j,i} P_{j,t} + e_{j,it},$$

where $j = Mkt$, SMB, and HML separately, one at a time. The factor number is estimated for each of the residual panels from these regressions. SMB and HML are found to be leaders. 'Market' is not found to be a leader. While this result may be initially surprising, it should be noted that it in no way contradicts the literature. Many others have found that the loadings on the market factor are almost constant (e.g., Ahn, Perez, and Gadarowski; 2013). Furthermore, Fama and French (1992) explicitly finds that the market factor does not affect stock returns. The important note here is that these are stock portfolios. Fama and French (1993) found that the market factor affects bond portfolios, but could not find that it affects stock portfolios.

Next, the following regression is performed to see if SMB and HML become the same or different factors:

$$\tilde{y}_{it} = \gamma^*_{1,i} P_{SMB,t} + \gamma^*_{2,i} P_{HML,t} + e^*_{it}.$$

The estimated factor number of the residual is found to be zero. Hence, SMB and HML must account for different factors.

To see whether or not the factor loading coefficients on Mkt are homogeneous, we estimate the first common factor without taking off the cross-sectional average, y_{it}, and then we regress $\hat{F}_{1t}$ on the risk free rate, r_{ft}, and $Mkt_t - r_{ft}$. This regression identifies the weighting coefficients between r_{ft} and $Mkt_t - r_{ft}$. The fitted value, $\tilde{F}_{1t}$, is approximated with the following weights.

$$\tilde{F}_{1t} = -0.0227 r_{ft} - 0.0032 \left(Mkt_t - r_{ft} \right).$$

Next, in Figure 1 we plot the estimated first common factor, the fitted value of $\tilde{F}_{1t}$, and $Mkt_t - r_{ft}$ together after standardization of each series. Evidently, the fitted value, $\tilde{F}_{1t}$, is more similar to the estimated factor than the fit by only market. After 1972, Mkt-Rft beats the fitted values by Rft and Mkr-Rf in only 3 time periods. From this empirical evidence, we conclude that the factor loadings on Mkt are almost homogeneous across each portfolio.

Table 6: Fama French Subsample Analysis for Leadership
Estimation by Ending Year (Starting Year: 1964; $N = 99$)

Ending Year	Estimated Factor Number (IC_2)					
	with	with	Regressand: $\tilde{y}_{it}$, Regressor			
	y_{it}	$\tilde{y}_{it}$	Mkt	SMB	HML	SMB & HML
2008	3	2	2	1	1	0
2007	3	2	2	1	1	0
2006	3	2	2	1	1	0
2005	3	2	2	1	1	0
2004	3	2	2	1	1	0
2003	3	2	2	1	1	0
2002	3	2	2	1	1	0
2001	3	2	2	1	1	0
2000	3	2	2	1	1	0

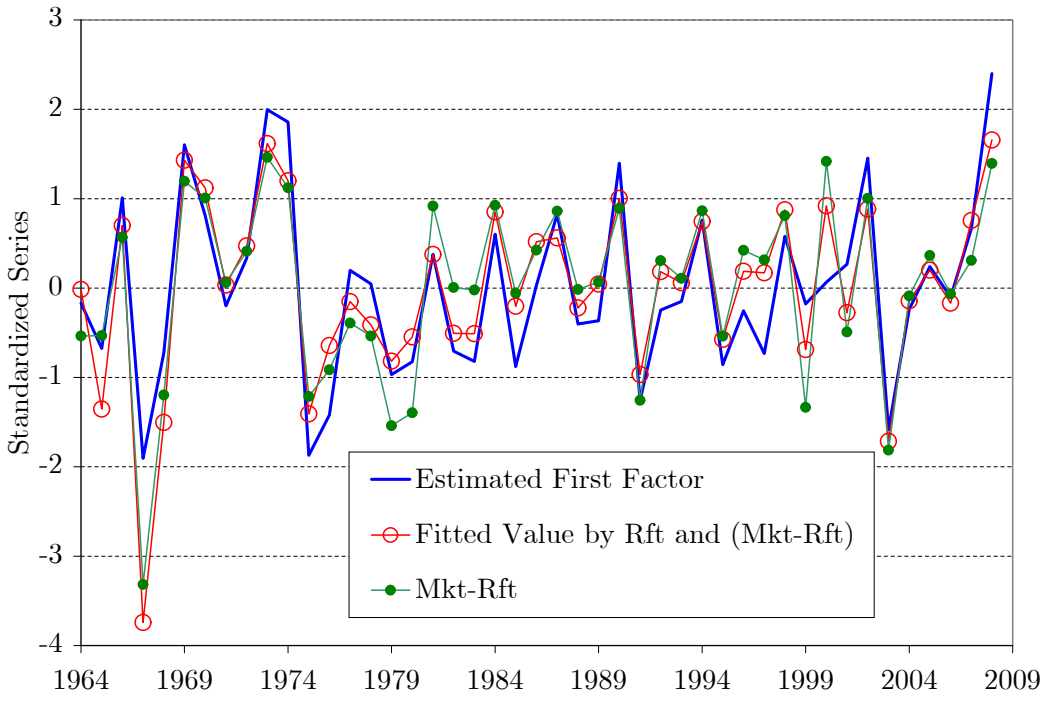

Figure 1: Missing Factor by Taking off Cross-Sectional Average

6 Conclusion

Factor analysis has become an increasingly popular tool in empirical research. Because there is no well-defined strategy for factor identification, researchers have been forced to choose between two outcomes. A researcher can either ignore any economic interpretation of the factor (perhaps the most important part of their model), or the researcher can speculate about the determinant without any concrete justification.

While this testing issue is a central problem, the thorny issue is identifying which particular time series to claim is the determinant. In many contexts, after estimating a factor the investigator is left with a time series which could be any marcoeconomic variable. A strategy is needed for selecting which variables could be an underlying factor. This issue become even more complicated when there are multiple factors because PC estimation yields a rotation of the underlying factors. Any particular variable could be a determinant even when such a variable has somewhat low correlation with each estimated factor.

This paper provides simple and effective solutions to these problems. First, a new method is described for testing if a particular variable is the common factor. Second, by modeling endogenous common factors, a strategy is developed for picking which variables could be determinants without requiring researchers to pore over the universe of exogenous variables. The performance of these methods is studied both in theory and in practice. Theoretically, the developed procedure correctly identifies the leader when N and T jointly approach infinity under the minimal assumptions of Bai and Ng (2002). Monte Carlo simulation shows that the procedure performs quite well in the finite sample. The procedure is then applied to an empirical example. The resulting estimation performs very well in practice.

References

[1] Ahn, S.C. and A. R. Horenstein (2013): "Eigenvalue ratio test for the number of factors." *Econometrica*, 1203-1227.

[2] Ahn, S.C. and M.F. Perez, (2010): "GMM estimation of the number of latent factors: with application to international stock markets." Journal of Empirical Finance, 17(4), 783-802.

[3] Ahn, S.C., M.F. Perez, and C. Gadarowski (2013): "Two-pass estimation of risk premiums with multicollinear and near-invariant betas." *Journal of Empirical Finance*, 20, 1-17.

[4] Alessi, L., M. Barigozzi, and M. Capasso (2010): "Improved penalization for determining the number of factors in approximate factor models," *Statistics & Probability Letters*, 80, 1806-1813.

[5] Amengual, D. and M. Watson (2007): "Consistent estimation of the number of dynamic factors in a large N and T panel.", *Journal of Business & Economic Statistics* 25, 91-96.

[6] Bai, J. (2003): "Inferential theory for factor models of large dimensions." *Econometrica* 71, 135-171.

[7] Bai, J. (2004): "Estimating cross-section common stochastic trends in nonstationary panel data." *Journal of Econometrics* 122, 137-183.

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

[9] Bai, J. and S. Ng (2006): "Evaluating latent and observed factors in macroeconomics and finance." *Journal of Econometrics* 131, 507-537.

[10] Bai, J. and S. Ng (2013): "Principal components estimation and identification of static factors." *Journal of Econometrics* 176, 18-29.

[11] Bates, B., M. Plagborg-Møller, J. Stock and M. Watson (2013): "Consistent factor estimation in dynamic factor models with structural instability." forthcoming in *Journal of Econometrics*.

[12] Choi, I. (2012): "Efficient estimation of factor models." *Econometric Theory*, 274–308.

[13] Fama, E.F. and K.R. French (1992): "The cross-section of expected stock returns." *Journal of Finance*, 47(2), 427-465.

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

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

[16] Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2005): "The generalized dynamic-factor model: one-sided estimation and forecasting." *Journal of the American Statistical Association* 471, 830-840.

[17] Gaibulloev, K., T. Sandler and D. Sul (2013): "Common drivers of transnational terrorism: principal component analysis." *Economic Inquiry. 707-721.*

[18] Greenaway-McGrevy, R., N. Mark, D. Sul, and J. Wu (2014): "Exchange rates as exchange rate common factors." mimeo, University of Texas at Dallas.

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

[20] Harding, M. (2013): "Estimating the number of factors in large dimensional factor models." mimeo, Stanford University.

[21] Kapetanios, G. (2010): "A testing procedure for determining the number of factors in approximate factor models with large datasets." *Journal of Business & Economic Statistics*, 28, 397–409.

[22] Ludwigson, S. and S. Ng (2007): "The empirical risk–return relation: a factor analysis approach." *Journal of Financial Economics* 83, 171-222.

[23] Onatski, A. (2009): "Testing hypotheses about the number of factors in large factor models." *Econometrica* 77, 1447-1479.

[24] Onatski, A. (2010):"Determining the number of factors from empirical distribution of eigenvalues." *Review of Economics and Statistics* 92(4), 1004-1016.

[25] Reis, R. and M. Watson (2010): "Relative goods' prices, pure inflation, and the Phillips correlation." *American Economic Journal: Macroeconomics* 2, 128-157.

[26] Stock J. and M. Watson (2002a): "Forecasting using principal components from a large number of predictors." *Journal of American Statistical Association* 97, 1167-1179.

[27] Stock J. and M. Watson (2002b): "Macroeconomic forecasting using diffusion indexes." *Journal of Business and Statistics* 20, 147-162.

Appendix

Proof of Lemma 1 (Asymptotic Factor Number for Weak Factors) The only requirement is to show that the difference in the residual sum of squares between using no factors and using k factors is small as $N, T \to \infty$; i.e.,

$$V(0) - V(k) = O_p\left(C_{NT}^{-2}\right), \tag{28}$$

for the following reason. The criterion function, IC_p, is defined by: $IC_p(k) = \ln[V(k)] + kp_{N,T}$, where $p_{N,T} \to 0$ and $C_{NT}^2 p_{N,T} \to \infty$ as $N, T \to \infty$. Hence,

$$IC_p(0) - IC_p(k) = \ln\left[\frac{V(0)}{V(k)}\right] - kp_{N,T}.$$

From Bai and Ng (2002), $V(0) - V(k) = O_p\left(C_{NT}^{-2}\right)$ implies $\ln[V(0)/V(k)] = O_p\left(C_{NT}^{-2}\right)$ whereas the penalty goes to infinity when multiplied by C_{NT}^2. Therefore, as $N, T \to \infty$, the penalty dominates $\ln[V(0)/V(k)]$ no matter which $k > 0$ is chosen. Hence, it is only necessary to show (28).

The eigenvalues of a rank k matrix A are denoted as $\varrho_1(A), \ldots, \varrho_k(A)$, ordered from largest to smallest. Proceed by expressing the difference in eigenvalue form,

$$\begin{aligned} V(0) - V(k) &= \sum_{l=1}^{N} \varrho_l\left(\frac{x'x}{NT}\right) - \sum_{l=k+1}^{N} \varrho_l\left(\frac{x'x}{NT}\right) \\ &= \sum_{l=1}^{k} \varrho_l\left(\frac{x'x}{NT}\right) \leq k\varrho_1\left(\frac{x'x}{NT}\right). \end{aligned} \tag{29}$$

Now use the model of x_{it} to show

$$\begin{aligned} \varrho_1\left(\frac{x'x}{NT}\right) &= \varrho_1\left(\frac{(\Psi Z' + x^{o\prime})(Z\Psi' + x^o)}{NT}\right) \\ &= \varrho_1\left(\frac{\Psi Z' Z\Psi'}{NT} + \frac{\Psi Z' x^o + x^{o\prime} Z\Psi'}{NT} + \frac{x^{o\prime}x^o}{NT}\right). \end{aligned}$$

From the Hermitian matrix eigenvalue inequality,

$$\begin{aligned} \varrho_1\left(\frac{x'x}{NT}\right) &\leq \varrho_1\left(\frac{\Psi Z' Z\Psi'}{NT}\right) + \varrho_1\left(\frac{\Psi Z' x^o + x^{o\prime} Z\Psi'}{NT}\right) + \varrho_1\left(\frac{x^{o\prime}x^o}{NT}\right) \\ &= I + II + III. \end{aligned} \tag{30}$$

Note that $III = \varrho_1\left(x^{o\prime}x^o/NT\right)$ follows $O_p\left(C_{NT}^{-2}\right)$ from the regularity conditions described above.

To bound the II term, re-express the quantity using the L_2-norm[3]:

$$\varrho_1\left(\frac{\Psi Z' x^o + x^{o\prime} Z\Psi'}{NT}\right) = \left\|\frac{\Psi Z' x^o + x^{o\prime} Z\Psi'}{NT}\right\|_2.$$

Next the triangle inequality can be applied:

$$\varrho_1\left(\frac{\Psi Z' x^o + x^{o\prime} Z\Psi'}{NT}\right) \leq \left\|\frac{\Psi Z' x^o}{NT}\right\|_2 + \left\|\frac{x^{o\prime} Z\Psi'}{NT}\right\|_2.$$

First consider:

$$\left\|\frac{\Psi Z' x^o}{NT}\right\|_2 = \sqrt{\varrho_1\left(\left(NT\right)^{-2} x^{o\prime} Z\Psi'\Psi Z' x^o\right)}.$$

Then,

$$\varrho_1\left(\left(NT\right)^{-2} x^{o\prime} Z\Psi'\Psi Z' x^o\right) \leq \operatorname{tr}\left[\left(NT\right)^{-2} x^{o\prime} Z\Psi'\Psi Z' x^o\right] = \operatorname{tr}\left[N^{-1}T^{-2} x^{o\prime} Z\left(\frac{1}{N}\sum_{i=1}^{T}\psi_i\psi_i'\right) Z' x^o\right]$$

Without loss of generality, assume that the loadings follow $\psi_i = O_p\left(C_{NT}^{-1}\right)$.[4] Hence,

$$\begin{aligned}\varrho_1\left(\left(NT\right)^{-2} x^{o\prime} Z\Psi'\Psi Z' x^o\right) &= O_p\left(C_{NT}^{-2}\right)\operatorname{tr}\left[N^{-1}T^{-2} x^{o\prime} Z Z' x^o\right] \\ &= O_p\left(\frac{1}{TC_{NT}^2}\right)\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t x_{it}^o\right\|_2^2,\end{aligned}$$

so that,

$$\left\|\frac{\Psi Z' x^o}{NT}\right\|_2 = O_p\left(\frac{1}{T^{1/2}C_{NT}}\right)\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{1}{T}\sum_{t=1}^{T} Z_t x_{it}^o\right\|_2^2}.$$

This is straightforward to bound because Bai and Ng (2002) makes the following exogeneity assumption regarding the factors and errors:

$$E\left[\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{1}{T}\sum_{t=1}^{T} Z_t x_{it}^o\right\|_2^2\right] \leq M.$$

Therefore, $\left\|\left(NT\right)^{-1}\Psi Z' x^o\right\|_2 = O_p\left(T^{-1/2}\times C_{NT}^{-1}\right) = O_p\left(C_{NT}^{-2}\right)$. Also $\left\|\left(NT\right)^{-1} x^{o\prime} Z\Psi'\right\|_2$ can be similarly bounded. Hence, $II = O_p\left(C_{NT}^{-2}\right)$.

[3] In Bai and Ng (2002), the commonly used norm is the Frobenius norm, $\|A\|_F = \operatorname{tr}[A'A]^{1/2}$. In our proof, we use the L2-norm, $\|A\|_2 = \varrho_1\left(A'A\right)^{1/2}$. For any matrix A, $\|A\|_2 \leq \|A\|_F$ with equality if and only if $\operatorname{rank}(A) = 1$. Hence, for any vector, $\|A\|_2 = \|A\|_F$. We occasionally state a Bai and Ng (2002) assumption in terms of the L2-norm, which is valid because of this equality for vectors.

[4] We know that $\psi_i' Z_t = O_p\left(C_{NT}^{-1}\right)$ so we can always rewrite $\psi_i' Z_t = \psi_i' I_r Z_t = \psi_i' A' A^{-1\prime} Z_t$ so that $\|A\psi_i\|_2 = O_p\left(C_{NT}^{-1}\right)$ and $\left\|A^{-1\prime} Z_t\right\|_2 = O_p\left(1\right)$.

Now, it is enough to show that $I = O_p\left(C_{NT}^{-2}\right)$. If $\psi_i' Z_t = O_p\left(C_{NT}^{-1}\right)$, we can assume without loss of generality that $Z_t = O_p\left(C_{NT}^{-1}\right)$ while $\psi_i = O_p(1)$. From here, it is straightforward to bound I :

$$\begin{aligned}\varrho_1\left(\frac{\Psi Z' Z \Psi'}{NT}\right) &= \varrho_1\left(\frac{1}{N}\Psi\left(\frac{1}{T}\sum_{t=1}^{T} Z_t Z_t'\right)\Psi'\right) \\ &= \varrho_1\left(O_p\left(C_{NT}^{-2}\right)\frac{\Psi\Psi'}{N}\right) \\ &= O_p\left(C_{NT}^{-2}\right)\rho_1\left(\frac{\Psi\Psi'}{N}\right),\end{aligned}$$

and the loadings can be bounded by,

$$\varrho_1\left(\frac{\Psi\Psi'}{N}\right) \leq \operatorname{tr}\left[\frac{\Psi\Psi'}{N}\right] = \frac{1}{N}\sum_{i=1}^{N}\psi_i'\psi_i = \frac{1}{N}\sum_{i=1}^{N}\|\psi_i\|_2^2.$$

Since the loadings are absolutely bounded ($\max_i \|\psi_i\|_2 = \bar{\psi}$),

$$\varrho_1\left(\frac{\Psi\Psi'}{N}\right) \leq \frac{1}{N}\sum_{i=1}^{N}\bar{\psi}^2 = \bar{\psi}^2,$$

and because any bounded variable is $O_p(1)$,

$$\rho_1\left(\frac{\Psi Z' Z \Psi'}{NT}\right) = O_p\left(C_{NT}^{-2}\right)\rho_1\left(\frac{\Psi\Psi'}{N}\right) = O_p\left(C_{NT}^{-2}\right).$$

Therefore, from (29) and (30), (28) is clear, and the lemma is proven.□

Proof of Theorem 1 (Identification of Estimated Factors: Known Potential Leaders): Begin under the conditions for (i). P_{jt} must be correlated with at least one of the latent factors, G_{st}. It is only necessary to show that $\hat{y}_{sj,it}^o$ has a weak factor structure. Express $\hat{y}_{sj,it}^o$ as

$$\hat{y}_{sj,it}^o = y_{it} - \hat{\beta}_{s,ji}P_{jt} - \hat{\alpha}_{i,-s}^{*\prime}\hat{F}_{-s,t} = \alpha_i' G_t - \hat{\beta}_{s,ji}P_{jt} - \hat{\alpha}_{i,-s}^{*\prime}\hat{F}_{-s,t} + y_{sj,it}^o.$$

If P_{jt} is a leader, then there exists a rotation of the latent factors, H^*, which aligns $\left[P_{jt}, \hat{F}_{-s,t}'\right]'$ with the latent factors,

$$H^* = \left(G'G\right)^{-1}G'\left[P_j, \hat{F}_{-s}\right],$$

There is a 'better' rotation (in terms of minimizing the sum of squared residuals), but $\hat{y}_{sj,it}^o$ has a weak factor structure using only the 'poor' rotation H^*, as is shown here. $\|H^*\|_2 = O_p(1)$ and $\left\|H^{*-1}\right\|_2 = O_p(1)$ follows from Stock and Watson (1998) and Bai and Ng (2002). Thus,

$$\begin{aligned}\hat{y}_{sj,it}^o &= \alpha_i' G_t - \left[\hat{\beta}_{s,ji}, \hat{\alpha}_{i,-s}^{*\prime}\right]\left(\left[P_{jt}, \hat{F}_{-s,t}'\right]' - H^{*\prime}G_t\right) - \left[\hat{\beta}_{s,ji}, \hat{\alpha}_{i,-s}^{*\prime}\right]H^{*\prime}G_t + y_{sj,it}^o \\ &= \left(\alpha_i' H^{*-1\prime} - \left[\hat{\beta}_{s,ji}, \hat{\alpha}_{i,-s}^{*\prime}\right]\right)H^{*\prime}G_t - \left[\hat{\beta}_{s,ji}, \hat{\alpha}_{i,-s}^{*\prime}\right]\left(\left[G_t, \hat{F}_{-s,t}'\right]' - H^{*\prime}G_t\right) + y_{sj,it}^o\end{aligned}$$

Using the proper alignment H^*,

$$\left\| \left[P_{jt}, \hat{F}'_{-s,t} \right]' - H^{*\prime} G_t \right\|_2 = O_p\left(N^{-1/2} \right)$$

follows from Bai and Ng (2003). Since, the factor is accurately estimated,

$$\left\| \alpha_i' H^{*-1\prime} - \left[\hat{\beta}_{s,ji}, \hat{\alpha}^{*\prime}_{i,-s} \right] \right\|_2 = O_p\left(T^{-1/2} \right),$$

follows from a simple least-squares analysis. Recognizing that $\left\| \left[\hat{\beta}_{s,ji}, \hat{\alpha}^{*\prime}_{i,-s} \right] \right\|_2 = O_p(1)$ and $\left\| H^{*\prime} G_t \right\|_2 = O_p(1)$, it is evident that $\hat{y}^o_{sj,it}$ has a weak factor structure. Therefore, $\hat{\#}\left(\hat{y}^o_{sj,it} \right) \to 0$ as $N, T \to \infty$.

Under the conditions for (ii), it is clear that the factors in $\hat{y}^o_{sj,it}$ do not vanish as $N, T \to \infty$ (the proof is straightforward).□

Table 1: Detecting Asymptotically Weak Factors (IC_2)

$y_{it} = \alpha_{1i} G_{1t} + y^o_{it}$, $G_{1t} = \rho G_{1t} + g_{1t}$, $g_{1t} \sim iidN(0,1)$

$y^o_{it} = \phi y^o_{it-1} + v_{it} + \beta \sum_{j \neq 0, j=-J}^{J} v_{i+j,t}$, $v_{it} \sim iidN(0,1)$,

		$\Pr\left[\hat{\#}(y_{it}) = 1\right]$			$\Pr\left[\hat{\#}(\tilde{y}_{it}) = 0\right]$			$\Pr\left[\hat{\#}(\tilde{y}_{it}) = 0\right]$		
		$\alpha_{1i} \sim iidN(1, 1/T)$			$\alpha_{1i} \sim iidN(1, 1/T)$			$\alpha_{1i} \sim iidN(1, 0.2)$		
T	N	I	II	III	I	II	III	I	II	III
25	25	1.00	0.91	0.58	1.00	0.97	0.79	0.80	0.73	0.55
25	50	1.00	0.78	0.44	1.00	0.93	0.71	0.54	0.40	0.27
25	100	1.00	0.64	0.38	1.00	0.86	0.66	0.34	0.20	0.14
25	200	1.00	0.52	0.30	1.00	0.81	0.65	0.20	0.10	0.08
50	25	1.00	0.99	0.60	1.00	1.00	0.75	0.49	0.55	0.32
50	50	1.00	1.00	0.92	1.00	1.00	0.97	0.33	0.41	0.33
50	100	1.00	1.00	0.93	1.00	1.00	0.98	0.05	0.07	0.06
50	200	1.00	1.00	0.97	1.00	1.00	0.99	0.01	0.01	0.01
100	25	1.00	1.00	0.61	1.00	1.00	0.73	0.24	0.41	0.19
100	50	1.00	1.00	0.94	1.00	1.00	0.98	0.04	0.10	0.06
100	100	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.01	0.00
100	200	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00
200	25	1.00	1.00	0.66	1.00	1.00	0.74	0.13	0.30	0.13
200	50	1.00	1.00	0.98	1.00	1.00	0.99	0.00	0.03	0.01
200	100	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00
200	200	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00

Note:

I: $\rho = 0.5$,$\phi = \beta = J = 0$.

II: $\rho = \phi = 0.5$,$\beta = J = 0$.

III: $\rho = \phi = 0.5$,$\beta = 0.1$,$J = 4$.

Table 2: Estimated Probability of Detecting True Factors (IC_2)

$$y_{it} = \alpha_{1i}G_{1t} + \alpha_{2i}G_{2t} + y_{it}^{o},\ G_{st} = \rho G_{st} + g_{st},$$
$$y_{it}^{o} = \phi y_{it-1}^{o} + v_{it} + \beta \sum_{j\neq 0, j=-J}^{J} v_{i+j,t},\ \alpha_{ki} \sim iidN\left(0,1\right),\ v_{it} \sim iidN\left(0,1\right),$$
$$\begin{bmatrix} g_{1t} \\ g_{2t} \end{bmatrix} \sim iidN\left(0,\Omega\right),\ \ \Omega = \begin{bmatrix} 2 & 0.5 \\ 0.5 & 1 \end{bmatrix}.$$

		Exact			Approximate			False		
		$P_{1t} = G_{1t}$			$P_{1t} = G_{1t} + \epsilon_{1t}/\sqrt{T}$			$P_{1t} = G_{1t} + \epsilon_{1t}$		
T	N	I	II	III	I	II	III	I	II	III
25	25	0.96	0.49	0.15	0.96	0.52	0.17	0.22	0.09	0.02
25	50	1.00	0.45	0.15	1.00	0.46	0.16	0.16	0.04	0.02
25	100	1.00	0.34	0.14	1.00	0.35	0.14	0.10	0.01	0.00
25	200	1.00	0.24	0.13	1.00	0.24	0.13	0.07	0.01	0.00
50	25	0.94	0.66	0.07	0.94	0.67	0.07	0.11	0.06	0.00
50	50	1.00	0.98	0.68	1.00	0.98	0.69	0.11	0.07	0.04
50	100	1.00	0.98	0.74	1.00	0.98	0.74	0.05	0.03	0.02
50	200	1.00	0.99	0.87	1.00	0.99	0.87	0.03	0.01	0.01
100	25	0.91	0.84	0.02	0.92	0.84	0.02	0.06	0.05	0.00
100	50	1.00	1.00	0.54	1.00	1.00	0.54	0.05	0.03	0.01
100	100	1.00	1.00	0.98	1.00	1.00	0.98	0.03	0.02	0.01
100	200	1.00	1.00	1.00	1.00	1.00	1.00	0.01	0.00	0.00
200	25	0.89	0.92	0.01	0.89	0.92	0.01	0.05	0.04	0.00
200	50	1.00	1.00	0.45	1.00	1.00	0.45	0.03	0.02	0.01
200	100	1.00	1.00	0.97	1.00	1.00	0.97	0.01	0.00	0.00
200	200	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00

Note:

I: $\rho = 0.5$,$\phi = \beta = J = 0$.

II: $\rho = \phi = 0.5$,$\beta = J = 0$.

III: $\rho = \phi = 0.5$,$\beta = 0.1$,$J = 4$.

Table 3: Rejection Rates of Bai and Ng's $\max_t \tau_t$ Test (Size: 5%)

$$y_{it} = \alpha_{1i} G_{1t} + \alpha_{2i} G_{2t} + y_{it}^o,\ G_{st} = \rho G_{st} + g_{st},$$
$$y_{it}^o = \phi y_{it-1}^o + v_{it} + \beta \sum_{j \neq 0, j=-J}^{J} v_{i+j,t},\ \alpha_{ki} \sim iidN(0,1),\ v_{it} \sim iidN(0,1),$$
$$\begin{bmatrix} g_{1t} \\ g_{2t} \end{bmatrix} \sim iidN(0, \Omega),\ \ \Omega = \begin{bmatrix} 2 & 0.5 \\ 0.5 & 1 \end{bmatrix}.$$

		Size (5%)						Power		
		Exact			Approximate			False		
		$P_{1t} = G_{1t}$			$P_{1t} = G_{1t} + \epsilon_t/\sqrt{T}$			$P_{1t} = G_{1t} + \epsilon_t$		
T	N	I	II	III	I	II	III	I	II	III
25	25	0.11	0.15	0.17	0.39	0.39	0.42	0.99	0.99	0.99
25	50	0.09	0.13	0.15	0.67	0.65	0.67	1.00	1.00	1.00
25	100	0.07	0.12	0.13	0.92	0.91	0.91	1.00	1.00	1.00
25	200	0.06	0.16	0.17	1.00	1.00	1.00	1.00	1.00	1.00
50	25	0.12	0.13	0.17	0.30	0.28	0.32	1.00	1.00	1.00
50	50	0.09	0.11	0.12	0.49	0.44	0.47	1.00	1.00	1.00
50	100	0.07	0.08	0.09	0.80	0.75	0.77	1.00	1.00	1.00
50	200	0.05	0.08	0.09	0.98	0.98	0.98	1.00	1.00	1.00
100	25	0.15	0.15	0.16	0.26	0.25	0.26	1.00	1.00	1.00
100	50	0.09	0.10	0.12	0.32	0.29	0.31	1.00	1.00	1.00
100	100	0.07	0.10	0.09	0.57	0.49	0.49	1.00	1.00	1.00
100	200	0.06	0.07	0.08	0.89	0.82	0.83	1.00	1.00	1.00
200	25	0.18	0.18	0.20	0.23	0.24	0.24	1.00	1.00	1.00
200	50	0.11	0.11	0.12	0.22	0.21	0.22	1.00	1.00	1.00
200	100	0.08	0.08	0.10	0.34	0.29	0.30	1.00	1.00	1.00
200	200	0.07	0.08	0.08	0.64	0.54	0.52	1.00	1.00	1.00

Note:

I: $\rho = 0.5$,$\phi = \beta = J = 0$.

II: $\rho = \phi = 0.5$,$\beta = J = 0$.

III: $\rho = \phi = 0.5$,$\beta = 0.1$,$J = 4$.

Table 4: Identifying Potential Leaders by using R^2

T	N	All P_{jt} Selected	False Selection Rate
25	25	0.98	0.52
25	50	0.94	0.42
25	100	0.93	0.32
25	200	0.94	0.35
50	25	0.99	0.03
50	50	0.99	0.00
50	100	0.97	0.00
50	200	0.98	0.00
100	25	1.00	0.00
100	50	1.00	0.00
100	100	0.99	0.00
100	200	0.99	0.00
200	25	1.00	0.00
200	50	1.00	0.00
200	100	1.00	0.00
200	200	1.00	0.00

Table 5: Clustering Frequency

T	N	Correctly Clustered	False Inclusion Rate
25	25	0.99	0.00
25	50	1.00	0.00
25	100	1.00	0.00
25	200	1.00	0.00
50	25	1.00	0.00
50	50	1.00	0.00
50	100	1.00	0.00
50	200	1.00	0.00
100	25	1.00	0.00
100	50	1.00	0.00
100	100	1.00	0.00
100	200	1.00	0.00
200	25	1.00	0.00
200	50	1.00	0.00
200	100	1.00	0.00
200	200	1.00	0.00