Identifying observed factors in approximate factor models: estimation and hypothesis testing

Chen, Liang

Universidad Carlos III de Madrid

20 March 2012

Identifying Observed Factors in Approximate Factor Models: Estimation and Hypothesis Testing

Liang Chen

Universidad Carlos III de Madrid
(This version: March, 2012)

Abstract

Despite their popularities in recent years, factor models have long been criticized for the lack of identification. Even when a large number of variables are available, the factors can only be consistently estimated up to a rotation. In this paper, we try to identify the underlying factors by associating them to a set of observed variables, and thus give interpretations to the orthogonal factors estimated by the method of Principal Components. We first propose a estimation procedure to select a set of observed variables, and then test the hypothesis that true factors are exact linear combinations of the selected variables. Our estimation method is shown to able to correctly identity the true observed factor even in the presence of mild measurement errors, and our test statistics are shown to be more general than those of Bai and Ng (2006). The applicability of our methods in finite samples and the advantages of our tests are confirmed by simulations. Our methods are also applied to the returns of portfolios to identify the underlying risk factors.

Keywords: Factor Models, Observed Factors, Estimation, Hypothesis Testing

1. Introduction

Factor models (FM henceforth) are becoming an increasingly important tool for both theoretical and empirical research. For example, in macroeconomics, the solutions of DSGE models can be written in the form of FM when these models allow for measurement errors (Altug 1989 and Sargent 1989), so that the structure of FM can help solve these models even when a large number of variables are considered (Boivin and Giannoni 2006, Kryshko 2011); in structural analysis, the factors estimated from large panel datasets can be combined with Structural Vector Autoregressions (SVAR) to identify the effects of fundamental shocks (Bernanke et al 2005), and solve the problem of *non-fundamentalness* (Forni et al 2009). Moreover, the estimated factors can significantly improve the forecasts of macro variables (Stock and Watson 2002a). In microeconomics, the demand systems are shown to have a factor structure (Lewbel 1991), and in some recent studies, FM are used to characterize the unobservable cross-sectional dependencies in panel data models (Pesaran 2006 and Bai 2009). Finally, in finance, the key assumption underlying the Arbitrage Pricing Theory (APT) is the multi-factors structure for the security returns.

As is well known, the popularity of FM is mainly due to their capability of summarizing the co-movements of a large number of variables (N) by a much smaller number of common factors ($r << N$). Moreover, the rapidly increasing dimensions of available data sets allow us to depart

from the restrictive assumptions of the classical factor analysis, and estimate the factor models consistently using the method of Principal Components (PC hereafter) (Bai and Ng 2002, Bai 2003, Stock and Watson 2002a).

Yet, it is well recognized that FM suffer from identification problems. Consider a factor model: $\mathbf{x}_t = \mathbf{\Lambda f}_t + \mathbf{e}_t$, where $\mathbf{x}_t$ is the vector of observed variables, $\mathbf{\Lambda}$ is the matrix of factor loadings, $\mathbf{f}_t$ is the unobservable factors, and $\mathbf{e}_t$ is the vector of idiosyncratic errors. Since only $\mathbf{x}_t$ can be observed, the above model is observably equivalent to: $\mathbf{x}_t = (\mathbf{\Lambda H}^{-1})(\mathbf{Hf}_t) + \mathbf{e}_t$, where $\mathbf{H}$ is any $r \times r$ nonsingular matrix. Therefore, unless one imposes $r \times r$ prior restrictions, the factors can only be identified up to a rotation, and thus the estimated factors usually lack a direct interpretation.[1].

In some situations, the object of interest is the conditional mean of some observed variables, so that the interpretation of the factors is not important. For example, in panel data models, one only needs to consistently estimate the common parts ($\mathbf{\Lambda f}_t$) of the unobservable effects, and thus the indeterminacy of the factors rotation does not matter for the results.

However, there are other instances where the direct object of interest are the factors themselves and thus a clear interpretation of them can have important implications for structural analysis. In financial economics, a large body of empirical research is concerned with identifying the factors that determine the prices of the securities. Chen et al (1986) and Shanken and Weinstein (2006) are examples of such work that try to interpret the underlying forces in the stock market in terms of some observed macro variables. Instead of using macro variables, Fama and French (1993) identify three observed factors related to the market returns and firm characteristics, which can explain most volatilities of the assets returns. In the solutions of DSGE models, the state variables and exogenous shocks (e.g., preference shocks or technology shocks) play the role of common factors, so that the interpretation of the factors is equivalent to identifying the sources of business cycles. In factor-based forecasts, not all the estimated factors necessarily have prediction power for the target variables (Tu and Lee 2011), and hence the forecasting can be further improved if some interpretational contents are attached to the factors. For example, the predictions of inflation rates could be more accurate if the factors associated with monetary policy shocks are given more weight than other factors identified as productivity changes (For more examples see Bai and Ng, 2006).

The goal of this paper is to identify the factors by relating them to some observed variables. The point of departure is the assumption that the common factors can be well approximated by (or linear functions of) some observed variables. Under this assumption, we will denote these observed variables as *observed factors.* We focus on the approximate factor models (Chamberlain and Rothschild 1983, Bai and Ng 2002) which allow for quite general assumptions about the data generating processes (DGP henceforth). More importantly, the space of the factors can be consistently estimated using the method of PC under the assumption of large N (the number of variables).

To the best of our knowledge, Bai and Ng (2006) is the only work that has addressed this issue.[2]. These authors consider the null hypothesis: $\mathbf{g}_t = \mathbf{Lf}_t$ for a $m \times r$ matrix $\mathbf{L}$ and a list of $m(> r)$ observed variables $\mathbf{g}_t$, suggested by some economic reasoning. They develop test

[1]The conventionally adopted identification assumptions for the estimation of factors using PC are that: (i) the factors are orthogonal and (ii) the covariance matrix of the factor loadings is diagonal.

[2]Bai and Ng (2011) study the identification of factors from a statistical point of view, i.e., by imposing restrictive assumptions on the data generating processes of the factors and factor loadings.

statistics for each of the observed variables g_{kt} as well as for the whole set of variables $\mathbf{g}_t$, based on the regressions of $\mathbf{g}_t$ on the estimated factors.

In practice, however, the list of observed factors is not always available, or those suggested by economic theory may not span the same space of the underlying factors. In view of these caveats, we propose here to first *estimate* (in the precise sense defined below) a list of observed factors from a much larger set of variables, and then test the null hypothesis that the underlying factors are exact linear combinations of observed variables selected in the first step.

In the estimation part, the estimated factors are regressed on different subsets of observed variables, and we label as the *estimated observed factors* that subset of variables that minimizes the Residual Sum of Squares (RSS) in these regressions. We differentiate two cases of observed factors: the *directly observed factors* (DOFs henceforth) and the *indirectly observed factors* (IOFs). In the first case, the latent factors in the FMs are directly approximated by the observed factors, i.e., there is a one-to-one correspondence between the r factors and r observed variables. In the second case, by contrast, the r factors are linear functions of m observed variables with $m \geq r$. Notice that this second setup includes the first one as a special case, but we will show that, for DOFs, our estimation method is much simpler and allows for larger measurement errors (i.e., the difference between the latent factors and the observed factors).

In the testing procedure, we consider the null hypothesis: $\mathbf{f}_t = \mathbf{B}\mathbf{x}_{1:m,t}$ for a $r \times m$ matrix $\mathbf{B}$ and a list of m observed variables $\mathbf{x}_{1:m,t}$. This hypothesis covers both cases discussed above, and is shown to be more general than the hypothesis considered by Bai and Ng (2006). We derive two types of test statistics based on the residuals in the regressions of estimated factors on $\mathbf{x}_{1:m,t}$. The advantages of our tests are that: (i) each of the proposed tests can be viewed as a test for the whole set of $\mathbf{x}_{1:m,t}$, rather than each element of $\mathbf{x}_{1:m,t}$; (ii) though Bai and Ng (2006) also proposed a test for the whole set of $\mathbf{x}_{1:m,t}$, our tests are derived under less restrictive conditions; (iii) since we consider a more general hypothesis, the tests of Bai and Ng (2006) tend to reject the null in the case of IOFs, while our tests still perform well.

The rest of the paper is organized as follows: Section 2 defines the basic notations and discusses the assumptions that define the approximate factor models. In section 3 we define the *directly observed factors*, and show how to identify them through estimation. The definition and identification of *indirectly observedfactors* are analyzed in Section 4, where we also discuss how to implement the method in practice. In section 5, we define the null hypothesis of observed factors and propose several test statistics whose asymptotic distributions are also derived. Section 6 studies the finite sample properties of the estimation methods and the test statistics, paying particular attention to their performance relative to the tests of Bai and Ng (2006). In Section 7 we apply our method to identify the risk factors for the returns of portfolios. Finally, Section 8 concludes. All the proofs are collected in the Appendices.

2. Models, Notations and Assumptions

Throughout this paper, we use the following standard notation. We define the matrix norm: $\|\mathbf{A}\| = \sqrt{\mathrm{Tr}(\mathbf{A}'\mathbf{A})}$, and use $\mathbf{A}_{1:m}$ to denote the 1st to $m-$th rows of a matrix (or a vector) $\mathbf{A}$. Further, $\mathbf{A} > 0$ (≥ 0) means that the matrix $\mathbf{A}$ is positive (semi) definite.

The following approximate factor models are considered:

$$\mathbf{x}_t = \mathbf{\Lambda}\mathbf{f}_t + \mathbf{e}_t, \tag{1}$$

where $\mathbf{x}_t = (x_{1t}, \ldots, x_{Nt})'$ is a $N \times 1$ vector of observed variables, $\mathbf{\Lambda} = (\lambda_1, \ldots, \lambda_N)'$ is a $N \times r$ vector of factor loadings, $\mathbf{f}_t = (f_{1t}, \ldots, f_{rt})'$ is a $r \times 1$ vector of common factors, and $\mathbf{e}_t = (e_{1t}, \ldots, e_{Nt})'$ is a $N \times 1$ vector of idiosyncratic errors. Unlike the classical factor analysis, we allow the number of variables N to go to infinity and the errors $\{e_{it}\}$ to be both temporarily and cross-sectionally correlated.

We assume that m among the observed variables $\mathbf{x}_t$ are *observed factors*, in a sense to be defined in the following sections, where m is a fixed number that does not increase as N goes to infinity. Without loss of generality, we assume these m *observed factors* are ordered as the first m variables of $\mathbf{x}_t$. The main issue is how to find the these m observed variables in the available set of size N . Given that the m observed factors are always placed in the first m rows, this issue becomes equivalent to finding out the first m variables out of N randomly ordered observed variables $\mathbf{x}_t$.

We consider two cases: DOFs and IOFs. In either case, the m observed factors have the following form:

$$\mathbf{x}_{1:m,t} = \mathbf{\Lambda}_{1:m}\mathbf{f}_t + \mathbf{e}_{1:m,t}. \tag{2}$$

To single out the observed factors, we have to impose some restrictions on $\mathbf{\Lambda}_{1:m}$ and $\mathbf{e}_{1:m,t}$, which will be discussed in Sections 3 and 4. Roughly speaking, for the DOFs, $\mathbf{\Lambda}_{1:m}$ should be a full-rank matrix and $\mathbf{e}_{1:m,t}$ should go to zero as N and T go to infinity; for IOFs, a necessary condition is that the covariance matrix of $\mathbf{e}_{1:m,t}$ has reduced rank.

Next we impose some assumptions for $\mathbf{\Lambda}$, $\mathbf{f}_t$ and $\mathbf{e}_t$. The following assumptions are necessary for the consistency of estimated factors using PC. Further, it should be noted that the assumptions to be imposed in Sections 3 and 4, when defining the observed factors, do not contradict with the following ones.

Let M denote a finite constant, we assume that:

Assumption 1. *(i)* $E\|\mathbf{f}_t\|^4 < M$ *for* $t = 1, \ldots, T$, *and* $\frac{1}{T}\sum_1^T \mathbf{f}_t\mathbf{f}_t' \to \mathbf{\Sigma}_F > 0$ *as* $N, T \to \infty$*; (ii)* $E\|\lambda_i\|^4 < M$ *for* $i = 1, \ldots, N$, *and* $\frac{1}{N}\sum_1^N \lambda_i\lambda_i' \to \mathbf{\Sigma}_\Lambda > 0$ *as* $N, T \to \infty$*; (iii) The* r *eigenvalues of* $\mathbf{\Sigma}_\Lambda\mathbf{\Sigma}_F$ *are different.*

Assumption 2. *(i)* $E(e_{it}) = 0$, $E(e_{it})^8 \le M$*;*
(ii) For $i, j = 1, \ldots, N$ *and* $s, t = 1, \ldots, T$, $E(e_{it}e_{js}) = \tau_{ij,ts}$, $|\tau_{ij,ts}| \le \tau_{ij}$ *for all* (t, s)*, and* $|\tau_{ij,ts}| \le \gamma_{ts}$ *for all* (i, j). $\frac{1}{N}\sum_{i,j}\tau_{ij} \le M$, $\frac{1}{T}\sum_{t,s}\gamma_{ts} \le M$, $\frac{1}{NT}\sum_{i,j,t,s}|\tau_{ij,st}| \le M$, *and* $\sum_s \gamma_{st}^2 \le M$*;*
(iii) For any (t, s), $E|N^{-1/2}\sum_{i=1}^N [e_{is}e_{it} - E(e_{is}e_{it})]|^4 \le M$.

Assumption 3. $\{\lambda_i\}$, $\{\mathbf{f}_t\}$ *and* $\{e_{it}\}$ *are three independent groups.*

These Assumptions are quite general in the sense that they allow heteroskedasticitiy, temporal and cross-sectional correlations in the factors and idiosyncratic terms. For more discussion on these Assumptions, see Bai (2003). Under Assumptions 1 to 3, the Information Criteria (IC) proposed by Bai and Ng (2002) can consistently estimate the number of factors, so that we can proceed as if this number was known. The effect of misspecification of the factor numbers is discussed in Section 4.

Define the $T \times r$ matrix $\tilde{\mathbf{F}} = (\tilde{\mathbf{f}}_1, \ldots, \tilde{\mathbf{f}}_T)'$ as $\sqrt{T}$ times the eigenvectors corresponding to the r largest eigenvalues of the $T \times T$ matrix $\mathbf{X}\mathbf{X}'$, where the $T \times N$ matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)'$. Then, denoting $\min[\sqrt{N}, \sqrt{T}]$ as $\delta_{N,T}$, the following result holds:

Lemma 1. *(Bai and Ng 2002) Under Assumptions 1 to 3,* $\delta_{N,T}\|\tilde{\mathbf{f}}_t - \mathbf{H}'\mathbf{f}_t\| = O_p(1)$ *for* $t = 1, \ldots, T$, *where* $\mathbf{H} = (\frac{\mathbf{\Lambda}'\mathbf{\Lambda}}{N})(\frac{\mathbf{F}'\tilde{\mathbf{F}}}{T})\mathbf{V}_{NT}^{-1}$, *and* $\mathbf{V}_{NT}$ *is a diagonal matrix containing the r largest eigenvalues of* $(NT)^{-1}\mathbf{X}\mathbf{X}'$.

Lemma 1 is a key result underlying our identification method for observed factors. It implies that the estimated factors are consistent for the space spanned by the true factors and hence for the observed factors. This relationship between the estimated factors and observed factors can be explored to identify the latter. For the identification of IOFs, the convergence rate $\delta_{N,T}$ is important to design an appropriate objection function, as will be shown in Section 4.

However, it is worth stressing that we do not consider a *weak factors* structure as in Onatski (2009a), in which the PC estimator of the factors are not consistent if their explanatory power is small relative to the idiosyncratic terms. In our setup, factors are strong whenver Assumption 1 is satisfied.

3. Directly Observed Factos

In this section, we deal with the identification of the DOFs. To give the precise definition of DOFs, the following assumptions are made:

Assumption 4. *(i)* $m = r$, $\mathbf{\Lambda}_{1:r}$ *has full rank, and* $e_{it} = \kappa_{N,T}\varepsilon_{it}$ *for* $i = 1, \ldots, r$, *where* $\kappa_{N,T} \to 0$ *as* $N, T \to \infty$*;*
(ii) $\varepsilon_t = (\varepsilon_{1t}, \ldots, \varepsilon_{rt})'$, $\frac{1}{T}\sum_{t=1}^{T}\varepsilon_t\varepsilon_t' \to \mathbf{\Sigma}_\varepsilon$ *and* $\|\mathbf{\Sigma}_\varepsilon\| \leq M$*;*
(iii) Let $\mathbf{e}_{N_1:N_r,t} = (e_{N_1,t}, \ldots, e_{N_r,t})'$ *for* $r+1 \leq N_1 < N_2 \ldots < N_r \leq N$, *then* $\frac{1}{T}\sum_{t=1}^{T}\mathbf{e}_{N_1:N_r,t}\mathbf{e}'_{N_1:N_r,t} \to \mathbf{\Sigma}^e_{N_1:N_r} > 0$.

Assumption 4(i) states that the first r variables span the space of the common factors $\mathbf{f}_t$ asymptotically: $\mathbf{x}_{1:r,t} \to \mathbf{\Lambda}_{1:r}\mathbf{f}_t$ as $N, T \to \infty$. When $\mathbf{\Lambda}_{1:r} = \mathbf{I}_r$, it simply means the common factors are directly measured by the first r observed variables with neglectable measurement errors. Notice that the nonsingular matrix $\mathbf{\Lambda}_{1:r}$ is just a normalization, and hence we can define the new factors as $\mathbf{g}_t = \mathbf{\Lambda}_{1:r}\mathbf{f}_t$ which are directly measured by $\mathbf{x}_{1:r,t}$, because for the remaining variables we have:

$$\begin{aligned}\mathbf{x}_{m+1:N,t} &= \mathbf{\Lambda}_{m+1:N}\mathbf{f}_t + \mathbf{e}_{m+1:N,t} \\ &= (\mathbf{\Lambda}_{m+1:N}\mathbf{\Lambda}_{1:r}^{-1})(\mathbf{\Lambda}_{1:r}\mathbf{f}_t) + \mathbf{e}_{m+1:N,t} \\ &= \mathbf{\Lambda}^*_{m+1:N}\mathbf{g}_t + \mathbf{e}_{m+1:N,t}\end{aligned}$$

Therefore, we label the first r observed variables *Directly Observed Factors*. Notice Bai and Ng (2006) identify the observed factors by constructing some test statistics under the assumption of an exact relationship between the observed variables and the factors, i.e., $\mathbf{e}_{1:r,t} = 0$ for $t = 1, \ldots, T$. By contrast, we allow for small measurement errors in the case of DOFs. We will show that the larger these measurement errors, the more difficult is the identification of the DOFs. Indeed, when $\kappa_{N,T} = 1$, there is no differences between the first m variables and the remaining $N - m$ ones.

Assumption 4(iii) rules out (asymptotic) multi-collinearity between any set of r observed variables, such that $\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_{N_1:N_r,t}\mathbf{x}'_{N_1:N_r,t} \to \mathbf{\Sigma}^x_{N_1:N_r} > 0$ for $1 \leq N_1 < \ldots < N_r \leq N$.

From Lemma 1 and Assumption 4 we can derive an approximate linear relationship between the estimated factors and the DOFs:

$$\tilde{\mathbf{f}}_t = \mathbf{H}\mathbf{f}_t + o_p(1) = \mathbf{H}\mathbf{\Lambda}_{1:r}^{-1}\mathbf{x}_{1:rt} + o_p(1) = \mathbf{A}\mathbf{x}_{1:r,t} + o_p(1), \tag{3}$$

where $\mathbf{A} = \mathbf{H}\mathbf{\Lambda}_{1:r}^{-1}$. As will be defined shortly, our method of identification is based on the regressions of the r estimated factors on r observed variables (in contrast to Bai and Ng 2006 where the observed variables are regressed on the estimated factors). The intuition for our approach is that, if $\tilde{\mathbf{f}}_t$ are regressed on the right set of observed variables: $\mathbf{x}_{1:m,t}$, the OLS estimator $\hat{\mathbf{A}}$ will converge to $\mathbf{A}$ and the residuals will be $o_p(1)$, so that RSS/ T will converge to 0. If the regressors are chosen as a set of r observed variables different from $\mathbf{x}_{1:r,t}$, we show that RSS$/T$ will instead converge to some positive numbers. As a result, we can identify the DOFs by comparing the RSS in the regression of $\tilde{\mathbf{f}}_t$ on different sets of observed variables.

Let $N_1 : N_r = [N_1, \ldots, N_r]$ denote a set of r indices such that $1 \leq N_1 < N_2 < \ldots N_r \leq N$, and let $\mathbf{x}_{N_1:N_r,t} = \mathbf{\Lambda}_{N_1:N_r}\mathbf{f}_t + \mathbf{e}_{N_1:N_r,t}$ be the corresponding observed variables. By defining:

$$S(N_1 : N_r, \mathbf{A}) = \frac{1}{T}\sum_{t=1}^{T}\left\|\tilde{\mathbf{f}}_t - \mathbf{A}\mathbf{x}_{N_1:N_r,t}\right\|^2, \tag{4}$$

and

$$[\hat{N}_1, \hat{N}_2, \ldots, \hat{N}_r] = \underset{N_1:N_r}{\arg\min}\Big(\min_{\mathbf{A}} S(N_1 : N_r, \mathbf{A})\Big), \tag{5}$$

then $\mathbf{x}_{\hat{N}_1:\hat{N}_r,t}$ is the vector of DOFs identified by our method.

Notice that

$$\frac{1}{T}\sum_{t=1}^{T}\left\|\tilde{\mathbf{f}}_t - \mathbf{A}\mathbf{x}_{N_1:N_r,t}\right\|^2 = \frac{1}{T}\sum_{k=1}^{r}\sum_{t=1}^{T}\left(\tilde{f}_{kt} - \mathbf{a}_k'\mathbf{x}_{N_1:N_r,t}\right)^2,$$

and therefore

$$\min_{\mathbf{A}} S(N_1 : N_r, \mathbf{A}) = S(N_1 : N_r, \hat{\mathbf{A}}),$$

where $\hat{\mathbf{A}}' = [\hat{\mathbf{a}}_1, \hat{\mathbf{a}}_2, \ldots, \hat{\mathbf{a}}_r]$, and $\hat{\mathbf{a}}_k$ is the OLS estimator of $\mathbf{a}_k$. This procedure can be simply implemented as follows: we first choose r observed variables, then calculate RSS_k in the OLS regression of $\tilde{f}_{kt}$ on these chosen variables, and get $\text{RSS} = \sum_{k=1}^{r}\text{RSS}_k$, where the set of variables that yield the smallest RSS are the identified DOFs.

The following theorem states that, using our method, the probability of correctly identifying the DOFs goes to 1 as N and T go to infinity.

Theorem 1. *Under Assumption 1 to 4,* $\mathbb{P}\Big([\hat{N}_1, \hat{N}_2, \ldots, \hat{N}_r] = [1, 2, \ldots, r]\Big) \to 1$ *as* $N, T \to \infty$.

This result holds as long as $\kappa_{N,T} = o(1)$. However, with finite samples, the DOFs may not be easily distinguishable from the remaining variables, due to either large measurement errors ($\kappa_{N,T}$) or large estimation errors of the PC. The finite sample properties of our identification procedure will be studied in Section 5 using simulations.

4. Indirectly Observed Factors

4.1. Definitions and comparison with Bai and Ng (2006)

In the previous section, we have studied the simple case where the common factors are directly observed, i.e., $\mathbf{f}_t = \mathbf{x}_{1:r,t}$ for $t = 1, \ldots, T$. However, in practice it is quite likely that the common factors are well approximated by the linear combinations of some observed variables, i.e., $\mathbf{f}_t = \mathbf{B}\mathbf{x}_{1:m,t}$ for a $r \times m$ matrix B with full row rank. For example, one of the macro variables considered by Chen et al (1986) is the spread of interest rates. When $m = r$, we have shown in the previous section that this case is equivalent to DOFs. Further, when $m < r$, the space spanned by the factors has rank m, instead of r, and so we should get m factors using Bai and Ng's IC. Hence, without loss of generality, we focus on the case: $m > r$ throughout this section.

We impose the following assumption to define the IOFs:

Assumption 5. *(i)* $\mathbf{f}_t = \mathbf{B}\mathbf{x}_{1:m,t}$ *for* $t = 1, \ldots, T$, *the* $r \times m$ *matrix* $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_m)$ *has full row rank, and* $\|\mathbf{b}_k\|^2 \neq 0$ *for* $k = 1, \ldots, m$;
(ii) $\mathbf{x}_{1:m,t} = \mathbf{\Lambda}_{1:m}\mathbf{f}_t + \mathbf{e}_{1:m,t}$, *and* $\mathbf{e}_{1:m,t} = \mathbf{C}_1\epsilon_t$, *where* $\mathbf{C}_1$ *is a* $m \times (m - r)$ *matrix such that* $\mathbf{C} = [\mathbf{\Lambda}_{1:m}, \mathbf{C}_1]$ *is a full rank matrix.*
(iii) For any constant number k, and any set of indices $1 \leq N_1 < \ldots < N_k \leq N$, $\frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_{N_1:N_k}\mathbf{x}'_{N_1:N_k} \xrightarrow{p} \mathbf{\Sigma}^x_{N_1:N_k} > 0$.
(iv) For any set of k indices: $m + 1 \leq N_1 < \ldots < N_k \leq N$, $\frac{1}{T}\sum_{t=1}^{T} \mathbf{e}_{N_1:N_k,t}\mathbf{e}'_{N_1:N_k,t} \xrightarrow{p} \mathbf{\Sigma}^e_{N_1:N_k} > 0$.

Although Assumption 5(ii) implies that the relation $\mathbf{f}_t = \mathbf{B}\mathbf{x}_{1:m,t}$ does exist, Assumptions 5 (i) (iii) and (iiii) guarantee that the IOFs $\mathbf{x}_{1:m,t}$ are uniquely determined. To see this notice that, from Assumption 5(ii), we can write $\mathbf{x}_{1:m,t} = \mathbf{C}\begin{pmatrix}\mathbf{f}_t \\ \epsilon_t\end{pmatrix}$. It follows that $\mathbf{C}^{-1}_{1:r}\mathbf{x}_{1:m,t} = \mathbf{f}_t$, which yields the expression in 5(i) with $\mathbf{B} = \mathbf{C}^{-1}_{1:r}$. Yet, Assumption 5(ii) alone is not enough to define a unique set of IOFs. For example, when $\mathbf{C} = \mathbf{I}$, we have $\mathbf{C}^{-1}_{1:r} = (\mathbf{I}_r, \mathbf{0})$, and thus $\mathbf{f}_t = \mathbf{x}_{1:r,t}$, which reduces to the case of DOFs. Besides, if $\mathbf{x}_{1:m,t}$ are IOFs, $\mathbf{x}_{1:m+1,t}$ will also be IOFs, since $\mathbf{f}_t = (\mathbf{B}, \mathbf{0})\mathbf{x}_{1:m+1,t}$. Therefore, the second part of Assumption 5(i) is necessary to exclude these undesirable cases. Moreover, Assumption 5(iii) excludes multi-collinearity among the element of any subset of observed variables. Together with 5(iv), it rules out the existence of IOFs formed by linear functions of other variables.

Note that the assumption $\mathbf{f}_t = \mathbf{B}\mathbf{x}_{1:m,t}$ is essential. To see this, recall that the hypothesis of interest in Bai an Ng (2006) is that $\mathbf{g}_t = \mathbf{L}\mathbf{f}_t$ for a $m \times r$ matrix $\mathbf{L}$, so that their tests are based on the regressions of the observed variables: $\mathbf{g}_t$ on the estimated factors: $\tilde{\mathbf{f}}_t$. On the contrary, as mentioned above, we regress the estimated factors on the observed variables. The difference is trivial for the case of DOFs since, given that the observed variables span the same space of the factors and that the estimated factors are consistent for the true factor space, then both regressions will produce neglectable residuals. However, this difference becomes nontrivial for the case of IOFs.

We use a simple example to illustrate the difference for IOFs. Consider a factor model with only one factor: $f_t = x_{1t} - x_{2t}$ for $t = 1, \ldots, T$, where x_{1t} and x_{2t} are two observed variables. The null hypothesis considered by Bai and Ng (2006) is:

$$\begin{pmatrix}x_{1t} \\ x_{2t}\end{pmatrix} = \begin{pmatrix}c \\ c - 1\end{pmatrix} f_t, \tag{6}$$

where c is any real number. Suppose now there is an estimator $\tilde{f}_t$ such that $f_t = \tilde{f}_t + o_p(1)$, one can write

$$x_{1t} = c\tilde{f}_t + o_p(1), \tag{7}$$

and the residuals in the regression of x_{1t} on $\tilde{f}_t$ will be $o_p(1)$ (note that the result is similar for x_{2t}). Their test statistics are based on exploring the exact order of the $o_p(1)$ term, namely $O_p(1/\sqrt{N})$ when $\sqrt{N}/T \to 0$. Now suppose there is another observed variable: $x_{3t} = f_t + e_{3t}$ with $\text{Var}(e_{3t}) = \sigma^2 > 0$. Then, since the residuals in the regression of x_{3t} on $\tilde{f}_t$ will be larger than $o_p(1)$ because we can write $x_{3t} = \tilde{f}_t + e_{3t} + o_p(1)$, their tests have power to reject x_{3t} as a member of $\mathbf{g}_t$.

In our definition, only $f_t = x_{1t} - x_{2t}$ is required, whereas x_{1t} and x_{2t} are allowed have the following FM representation:

$$\begin{pmatrix} x_{1t} \\ x_{2t} \end{pmatrix} = \begin{pmatrix} c \\ c-1 \end{pmatrix} f_t + \begin{pmatrix} 1 \\ 1 \end{pmatrix} \varepsilon_t, \tag{8}$$

for any real number c and random process ε_t. Note that (6) is a special case of (8) with $\varepsilon_t = 0$ but with x_{1t} and x_{2t} being defined as in (8), the tests of Bai and Ng (2006) will reject the null if ε_t is a process with positive variance, despite being true that $f_t = x_{1t} - x_{2t}$.

To summarize, the null hypothesis considered by Bai and Ng (2006) is equivalent to the definition of DOFs without measurement errors. Hence,it is less general the definition of IOFs considered here.

4.2. Identifying the IOFs

The idea for identifying the IOFs is similar to the identification of DOFs. If the number of IOFs: m is a priori known, one can use the method in Section 3 to select the m out N observed variables that yield the smallest RSS, where the probability of correctly selecting the m IOFs goes to 1 as N and T goes to infinity.

However, when m is not known in practice, one is faced with the choices of both m and $\mathbf{x}_{1:m,t}$. Let $\hat{m}$ be an estimator of m. If $\hat{m} < m$, then any $\hat{m}$ selected variables cannot span the space of the r factors, since otherwise Assumption 5(i) will be violated. Then the sum of RSSs (divided by T) in the regressions of the estimated factors on the selected observed variables will be positive. If $\hat{m} = m$, the sum of RSSs (divided by T) will converge to 0 if $\mathbf{x}_{1:m,t}$ are selected. However, when $\hat{m} > m$ and $\mathbf{x}_{1:m,t}$ are among the selected variables, the sum of RSSs (divided by T) will also converge to 0 because adding more regressors never increases the RSSs. To solve this problem, we need to impose some penalty functions for adding extra regressors.

To do so, let us define:

$$[\hat{m}, \hat{N}_1, \hat{N}_2, \ldots, \hat{N}_{\hat{m}}] = \underset{r \le k \le k_{max}, N_1:N_k}{\arg\min} \Big(S(N_1 : N_k, \hat{\mathbf{A}}) + k \cdot p(N,T) \Big), \tag{9}$$

where $S(N_1 : N_k, \hat{\mathbf{A}})$ is as defined in Section 3, k_{max} is a predetermined constant, and $p(N,T)$ is a penalty function depending on N and T. The following theorem constitutes the main result of this paper:

Theorem 2. *Under Assumptions 1, 2, 3 and 5,*

$$\mathbb{P}[\hat{m} = m, (\hat{N}_1, \ldots, \hat{N}_{\hat{m}}) = (1, \ldots, m)] \to 1$$

as $N, T \to \infty$*, if* $k_{max} \ge m$*,* $p(N,T) \to 0$ *and* $\delta_{N,T}^2 p(N,T) \to \infty$ *as* $N, T \to \infty$*.*

The estimation procedure in Section 3 is repeated for different values of k, and we add a penalty term to the object function. Theorem 2 implies that one can identify the number of IOFs and the IOFs simultaneously with probability approaching to 1 as N and T increase.

4.3. The choice of penalty functions

Since the penalty functions in our procedure and those considered by Bai and Ng (2002) have to satisfy the same conditions, we can use some of their choices that have been proved successful in determining the number of factors. Particularly, we consider the following three penalty functions:

$$p_1(N,T) = \left(\frac{N+T}{NT}\right)\ln\left(\frac{NT}{N+T}\right),$$

$$p_2(N,T) = \left(\frac{N+T}{NT}\right)\ln(\delta_{N,T}^2),$$

$$p_3(N,T) = \frac{\ln\delta_{N,T}^2}{\delta_{N,T}^2}.$$

These penalty functions have the same asymptotic properties but may perform differently in finite samples (see Bai and Ng, 2002) for a detailed discussion). The finite sample properties of our method using these functions are studied in the Section 6.

4.4. Practical implementation

In the previous discussion, we have assumed that the number of factors (r) is known or correctly estimated. However, in practice, the estimated number of factors using different methods usually differ for the same data set. For example, if one applies the test of Onatski (2009b) to the U.S macro data set used in Stock and Watson (2009), 2 factors can be found; but if one uses the 6 different information criteria of Bai and Ng (2002) to the same data, the estimated numbers of factors range from 2 to 6. Actually, it is very rare in practice that the number of factors can be uniquely determined by different methods. Therefore, a discussion on how to implement our methods in practice becomes necessary when the number of factors cannot be correctly specified.

When the estimated number of factors $\hat{r}$ is larger than the true one r, Lemma 1 does not hold, so that the above-mentioned methods will fail to identify the IOFs (or DOFs). When $\hat{r} < r$, Lemma 1 continues to hold, but our methods will not necessarily identify all of the IOFs. To see this, we first write:

$$\tilde{\mathbf{f}}_t = \mathbf{H}'\mathbf{B}\mathbf{x}_{1:m,t} + o_p(1) = \mathbf{A}\mathbf{x}_{1:m,t} + o_p(1)$$

by Lemma 1 and Assumption 5(i), where the matrix $\mathbf{A} = \mathbf{H}'\mathbf{B}$ is $r \times m$. Let $\mathbf{a}_k$ be the kth row of $\mathbf{A}$, then $\tilde{f}_{kt} = \mathbf{a}_k\mathbf{x}_{1:m,t} + o_p(1)$. If we apply our procedure to each of the $\tilde{f}_{kt}$ for $k = 1, \ldots, r$, then $\tilde{f}_{kt}$ can only identify those variables corresponding to the non-zero elements of $\bar{\mathbf{a}}_k = \text{plim } \mathbf{a}_k$. For example, if $\mathbf{a}_1 \xrightarrow{p} (1, 0, \ldots, 0)$, $\tilde{f}_{1t}$ can only identify x_{1t}. However, Theorem 2 guarantees that the union of the variables identified by $\tilde{f}_{1t}$ to $\tilde{f}_{rt}$ is equal to the IOFs. The reason is that, since $\mathbf{H}$ (also plim$\mathbf{H}$) is nonsingular and $\mathbf{B}$ has no zero columns (Assumption 5(i)), $\mathbf{A}$ (also plim$\mathbf{A}$) does not have zero columns.

The previous discussion suggests that we can implement our procedure as follows: Instead of regressing all the estimated factors on the observed variables, we run the regression for each of the estimated factors, starting with the first one: $\tilde{f}_{1t}$. For each $\tilde{f}_{kt}$, define:

$$[\hat{m}_k, \hat{N}_1, \hat{N}_2, \ldots, \hat{N}_{\hat{m}_k}] = \underset{r \le h \le k_{max}, N_1:N_h}{\arg\min} \Big(\frac{1}{T} \sum_{t=1}^{T} (\tilde{f}_{kt} - \hat{\mathbf{a}}_k \mathbf{x}_{N_1:N_h,t})^2 + h \cdot p(N,T) \Big), \tag{10}$$

where $\hat{\mathbf{a}}_k$ is the OLS estimator and $p(N,T)$ is as defined above. The key here is when to stop the process. If one stops when $k < r$, the union of the selected variables may be a subset of the IOFs; if one stops when $k > r$, some of the selected variables will not belong to the IOFs. The practitioner can combine the results with some economic theory to judge the appropriateness of the selected variables. If some obvious irrelevant variables are selected for some large k, one should stop the process and restrict attention to the variables already selected.

The main advantage of this procedure is that one can at least identify all of the IOFs, at the cost of identifying some non-IOF variables. While the result in Theorem 2 is much simpler, one bears the risk that none of the selected variables belong to the IOFs when the estimated number of factors is larger than r.

Another practical issue is that the computational cost of our method tend to explode as N, r, m and k_{max} increase. As will be shown in the simulations, when $N = 100$, $r = 2$, $m = 3$ and $k_{max} = 4$, the searching process takes about 1 hour. [3] In practice, N is at least around 100 in most cases, and can be as large as thousands in financial data sets. Since the number of factors r usually ranges from 2 to 8 in many applications, if we were to search in the whole set of variables for those cases, the computational cost could be huge.

To solve this problem, we can restrict our attention to a subset of n variables with $n < N$. Theorems 1 and 2 should still hold if these n variables contain the observed factors (DOFs or IOFs). In practice, a list of n candidate variables can be selected by prior knowledge and/or economic reasoning. In theory, with large samples, our methods should correctly select the observed factors as long as they are contained in the n variables. However, in practice, the accuracy of our approach with finite samples will depend on n: the smaller n, the less time the computation takes, and the more likely that the observed factors are identified. But a smaller n means that one has to exclude more variables and thus it becomes more likely to miss the IOFs. To reach a balance, we should make n as large as possible whenever the computation cost is affordable. The finite sample performances of our methods when selection is restricted to n variables are studied in Section 6.

Another shortcut that can significantly reduce the computation cost is to start the searching process with a large number of regressors, l. In the proof of Theorem 2 (See Appendix B), it is shown that if $l > m$, we will select the IOFs ($\mathbf{x}_{1:m,t}$) with other $l - m$ variables with probability approaching 1 as N and T go to infinity. In the next step, we only need to search among the l selected variables in the first step. By a simple conditional probability argument, this modified procedure should have the same asymptotic property as the procedure in (9). The computation cost will be greatly reduced since the second step is really easy to calculate. Moreover, the variables selected in the first step can be combined with other variables to form a list of n variables. In this case the computation cost mainly depends on r_{max} (the maximum of l) and N (or n).

[3] The calculations are implemented with Matlab 2009

5. Hypothesis Testing

So far we have assumed the existence of observed factors. Nevertheless, it is possible that the factors cannot be approximated by any observed variables, such as the potential GDP and the natural rate of unemployment. In such a case, it is necessary to design some tests for the null hypothesis $H_0 : \mathbf{f}_t = \mathbf{B}\mathbf{x}_{1:m,t}$ when some observed factors have been selected by our estimation methods. In this section, we propose several test statistics for the H_0 based on both individual and multiple regressions. Notice that the H_0 here covers both DOFs and IOFs, because DOFs can be viewed as a special case of IOFs with B being a $r \times r$ nonsingular matrix. We differentiate these two cases in the estimation because the method for identifying DOFs is simpler, although the method for identifying IOFs includes DOFs as a special case.

The key result underlying our tests is the following lemma proved by Bai (2003):

Lemma 2. $\sqrt{N}(\tilde{\mathbf{f}}_t - \mathbf{H}'\mathbf{f}_t) \xrightarrow{d} N(0, \mathbf{\Omega}_t)$ *if* $\sqrt{N}/T \to 0$ *as* $N, T \to \infty$, *where* $\mathbf{\Omega}_t = \mathbf{V}^{-1}\mathbf{Q}\mathbf{\Gamma}_t\mathbf{Q}'\mathbf{V}^{-1}$, *and* $\mathbf{\Gamma}_t = \lim_{N\to} \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} E(\lambda_i\lambda_j' e_{it}e_{jt})$.

The matrices $\mathbf{V}$ and $\mathbf{Q}$ are defined in Appendix A. It follows that:

$$\sqrt{N}(\tilde{f}_{kt} - \mathbf{h}_k'\mathbf{f}_t) \xrightarrow{d} N(0, \sigma_{t,k}^2), \tag{11}$$

where $\mathbf{h}_k$ is the kth column of $\mathbf{H}$ and $\sigma_{t,k}^2 = \mathbf{\Omega}_t(k,k)$. Our tests are based on the residuals in the regression of the estimated factors on the selected observed variables. Lemma 1 and the null hypothesis imply that $\tilde{\mathbf{f}}_t = \mathbf{H}'\mathbf{f}_t + v_t = \mathbf{A}\mathbf{x}_{1:m,t} + \mathbf{v}_t$, where $\mathbf{v}_t = \tilde{\mathbf{f}}_t - \mathbf{H}'\mathbf{f}_t$ and $\mathbf{A} = \mathbf{H}'\mathbf{B}$. Let $\hat{\mathbf{A}}$ denote the OLS estimator of $\mathbf{A}$, then:

$$\tilde{\mathbf{f}}_t = \hat{\mathbf{A}}\mathbf{x}_{1:m,t} + (\mathbf{A} - \hat{\mathbf{A}})\mathbf{x}_{1:m,t} + \mathbf{v}_t = \hat{\mathbf{A}}\mathbf{x}_{1:m,t} + \hat{\mathbf{v}}_t,$$

where $\hat{\mathbf{v}}_t = (\mathbf{A} - \hat{\mathbf{A}})\mathbf{x}_{1:m,t} + \mathbf{v}_t$. It follows that $\sqrt{N}\hat{\mathbf{v}}_t - \sqrt{N}\mathbf{v}_t = \sqrt{N}(\mathbf{A} - \hat{\mathbf{A}})\mathbf{x}_{1:m,t}$. Therefore $\sqrt{N}\hat{\mathbf{v}}_t$ should converge to the same distribution of $\sqrt{N}\mathbf{v}_t$ because $\sqrt{N}(\mathbf{A} - \hat{\mathbf{A}}) = o_p(1)$. To see this, we can write:

$$\hat{\mathbf{A}} - \mathbf{A} = \Big(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_{1:m,t}\mathbf{x}_{1:m,t}'\Big)^{-1}\Big(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_{1:m,t}\mathbf{v}_t'\Big).$$

By Assumption 5, $\Big(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_{1:m,t}\mathbf{x}_{1:m,t}'\Big) \xrightarrow{p} \mathbf{\Sigma}_{1:m}^{x} > 0$, and

$$\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_{1:m,t}\mathbf{v}_t' = \mathbf{\Lambda}_{1:m}\frac{1}{T}\sum_{t=1}^{T}\mathbf{f}_t\mathbf{v}_t' + \frac{1}{T}\sum_{t=1}^{T}\mathbf{e}_{1:m,t}\mathbf{v}_t'.$$

By Lemma B1 and B2 of Bai (2003), $\frac{1}{T}\sum_{t=1}^{T}\mathbf{f}_t\mathbf{v}_t'$ and $\frac{1}{T}\sum_{t=1}^{T}\mathbf{e}_{1:m,t}\mathbf{v}_t'$ are both $O_p(\delta_{N,T}^{-2})$, whereby it follows that $\sqrt{N}(\mathbf{A}-\hat{\mathbf{A}}) = O_p(\frac{\sqrt{N}}{\min[N,T]})$, which is $o_p(1)$ under the condition that $\sqrt{N}/T \to 0$. As a result of Lemma 2 and the previous analysis, the distribution of the residuals $\hat{\mathbf{v}}_t$ in the regressions of $\tilde{\mathbf{f}}_t$ on $\mathbf{x}_{1:m,t}$ can be derived as follows:

$$N\hat{\mathbf{v}}_t'\mathbf{\Omega}_t^{-1}\hat{\mathbf{v}}_t \xrightarrow{d} \chi_r^2, \tag{12}$$

$$N\Big(\frac{\hat{v}_{kt}}{\sigma_{t,k}}\Big)^2 \xrightarrow{d} \chi_1^2, \tag{13}$$

where $\hat{v}_{kt}$ is the kth element of $\hat{\mathbf{v}}_t$, i.e., the residuals in the regression of $\tilde{f}_{kt}$ on $\mathbf{x}_{1:m,t}$.

Based on these results, we can construct two types of tests. The first type is similar to the $A(j)$ test statistics of Bai and Ng (2006). First, we define:

$$\hat{\rho}_t = N\hat{\mathbf{v}}_t'\hat{\mathbf{\Omega}}_t^{-1}\hat{\mathbf{v}}_t \ , \hat{\rho}_{t,k} = N\Big(\frac{\hat{v}_{kt}}{\hat{\sigma}_{t,k}}\Big)^2, \tag{14}$$

and

$$\mathcal{A} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{1}(\hat{\rho}_t > \Phi_{r,\alpha}) \tag{15}$$

$$\mathcal{A}_k = \frac{1}{T}\sum_{t=1}^{T}\mathbf{1}(\hat{\rho}_{t,k} > \Phi_{1,\alpha}) \text{ for } k = 1,\dots,r. \tag{16}$$

where $\Phi_{r,\alpha}$ and $\Phi_{1,\alpha}$ are two constants such that $\mathbb{P}[\chi_r^2 \geq \Phi_{r,\alpha}] = \mathbb{P}[\chi_1^2 \geq \Phi_{1,\alpha}] = \alpha$, and $\hat{\mathbf{\Omega}}_t$ is a consistent estimate of $\mathbf{\Omega}_t$[4].

By the results in (12) and (13), $E(\mathbf{1}(\hat{\rho}_t > \Phi_{r,\alpha})) = \mathbb{P}[\hat{\rho}_t > \Phi_{r,\alpha}] \to \alpha$. Then, using the argument behind the Law of Large Numbers (LLN) we can prove the following result:[5]

Proposition 1. *Under Assumptions 1 to 3 and the hypothesis that* $\mathbf{f}_t = \mathbf{B}\mathbf{x}_{1:m,t}$ *for* $t = 1,\dots,T$, *then* $\mathcal{A} \xrightarrow{p} \alpha$ *and* $\mathcal{A}_k \xrightarrow{p} \alpha$ *for* $k = 1,\dots,r$ *if* $\sqrt{N}/T \to 0$ *as* $N, T \to \infty$.

Notice once more that the $A(j)$ test of Bai and Ng (2006) is based on individual regressions of the observed variables on the estimated factors (regress each of $\mathbf{x}_{1:m,t}$ on $\tilde{\mathbf{f}}_t$), while we do the opposite here (regress each of $\tilde{\mathbf{f}}_t$ on $\mathbf{x}_{1:m,t}$). As discussed in Section 4, the advantage of our procedure is that it allows us to consider more general relations between the factors and observed variables. Moreover, it allows us to construct test statistics not only for the individual regressions, but also for multiple regressions as in (15).

It should be noted that the test statistics defined in (15) and (16) cannot be used in a strict sense, because although their probability limits are derived, their distributions remain unknown. However, since they should not be too far away from their limit values under the null, they can still provide useful information to help us evaluate the hypothesis.

The second type of test are constructed by pooling those statistics defined in (14). Specifically, let us define:

$$\mathcal{P} = \frac{\sum_{t=1}^{T}\hat{\rho}_t - Tr}{\sqrt{2Tr}}, \tag{17}$$

$$\mathcal{P}_k = \frac{\sum_{t=1}^{T}\hat{\rho}_{t,k} - T}{\sqrt{2T}} \text{ for } k = 1\dots,r. \tag{18}$$

The sums of the statistics are standardized by their means and variances, and the following proposition gives their asymptotic distributions.

[4] See Bai and Ng (2006) for discussions on the estimation of $\mathbf{\Omega}_t$.

[5] The proof is omitted because given the results in (12) and (13), it is very similar to the proof of Proposition 1 in Bai and Ng (2006).

Proposition 2. *Under Assumptions 1 to 3 and the hypothesis that* $\mathbf{f}_t = \mathbf{B}\mathbf{x}_{1:m,t}$ *for* $t = 1, \ldots, T$, *then* $\mathcal{P} \xrightarrow{d} N(0,1)$ *and* $\mathcal{P}_k \xrightarrow{d} N(0,1)$ *for* $k = 1, \ldots, r$, *if* $\sqrt{N}/T \to 0$ *as* $N, T \to \infty$ *and* $\{e_{it}\}$ *are serially uncorrelated for all* $i = 1, \ldots, N$.

Unlike the statistics $\mathcal{A}$, the statistics $\mathcal{P}$ has a known limiting distribution and thus can be used for testing the null hypothesis. However, the conditions are more restrictive since the error terms are required to be serially uncorrelated.

Bai and Ng (2006) also proposed some statistics for testing the null hypothesis for a group of observed variables using the theory of canonical correlations, but the limiting distribution of their tests are known only under very restrictive conditions, e.g., $\mathbf{f}_t$ is i.i.d normal (or elliptically) distributed. Our test statistics can also be viewed as tests for a group of observed variables, but their limiting distributions are known under more general conditions.

6. Simulations

6.1. Directly Observed Factors

In this section, we study the finite sample performance of our method for identifying the DOFs. The following DGP is used: $x_{it} = \lambda_i \mathbf{f}_t + e_{it}$ for $i = 1, \ldots, N$ and $t = 1, \ldots, T$, where $\mathbf{f}_t$ are i.i.d multivariate normal vectors with mean 0 and $E(\mathbf{f}_t \mathbf{f}_t') = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$, λ_{ik} and e_{it} are i.i.d random variables drawn from standard normal distributions for $i = r+1, \ldots, N$, $t = 1, \ldots, T$ and $k = 1, \ldots, r$. Moreover, we let $r = 2$, $\mathbf{\Lambda}_{1:2} = \mathbf{I}_2$, and the first two variables are generated as $\mathbf{x}_{1:2,t} = \mathbf{f}_t + \kappa \varepsilon_t$, where ε_{it} are also i.i.d standard normal variables. As has been discussed earlier, the larger the parameter κ, the more difficult is to identify the DOFs.

In the simulations, we report the probability of correctly identifying the DOFs(i.e., the first two variables: $\mathbf{x}_{1:2,t}$) out of 1000 replications using the method proposed in Section 3, for sample sizes $N, T = 50, 100, 150, 200$, and for 4 different specifications of κ: $\kappa = 0$, $\kappa = \delta_{N,T}^{-2}$, $\kappa = \delta_{N,T}^{-1}$ and $\kappa = \delta_{N,T}^{-2/3}$. Recall that $\delta_{N,T} = \min[\sqrt{N}, \sqrt{T}]$. The results are summarized in Table 1.

It can be observed that our method can identify the DOFs correctly with very high probabilities for $\kappa = 0, \delta_{N,T}^{-2}$ and $\delta_{N,T}^{-1}$, even for $N, T = 50$. However, when κ increases to $\delta_{N,T}^{-2/3}$, the probabilities decrease dramatically to less than 30% for $N = 50$ or $T = 50$. Note that $\delta_{N,T}^{-2/3} = 0.27$ when $N = 50$ or $T = 50$, representing a big measurement error. The probabilities increase to more than 50% when $\min[N, T] = 100$ and to more than 80% when $\min[N, T] = 150$.

To study the finite sample properties of the test statistics proposed in Section 5 and to compare them to those of Bai and Ng (2006), we generate the simulated data as above except that now κ is fixed to 0. As discussed in Section 4, for the DOFs our tests should perform closely to those of Bai and Ng (2006). The simulation results from 1000 replications are summarized in Table 2.

Columns 3 to 5 report the averaged statistics defined in (17) and (18), while columns 6 to 8 display the empirical sizes of the tests defined in (19) and (20). [6] Finally, the last two columns show the $A(j)$ statistics of Bai and Ng (2006). It can be seen that all the reported numbers are close to their limiting values (5%), although the $\mathcal{P}_k$ tests tend to be oversized in small sample sizes.

[6] We use the 2.5% critical value of a standard normal distribution

Table 1: Probabilities of Correctly Identifying DOFs.

N	T	$\kappa = 0$	$\kappa = \delta_{N,T}^{-2}$	$\kappa = \delta_{N,T}^{-1}$	$\kappa = \delta_{N,T}^{-2/3}$
50	50	100	98	74	10
50	100	100	100	87	16
50	150	100	99	92	21
50	200	100	100	84	23
100	50	100	100	95	14
100	100	100	100	100	60
100	150	100	100	100	58
100	200	100	100	100	67
150	50	100	100	93	10
150	100	100	100	100	55
150	150	100	100	100	88
150	200	100	100	100	93
200	50	100	100	94	5
200	100	100	100	100	57
200	150	100	100	100	82
200	200	100	100	100	98

DGP: $x_{it} = \sum_{k=1}^{2} \lambda_{ki} f_{kt} + e_{it}$, where $\mathbf{f}_t = (f_{1t}, f_{2t})'$ is multivariate normal with $E(f_{kt}) = 0$, $E(f_{kt}^2) = 1$, and $E(f_{1t} f_{2t}) = 0.5$. $\mathbf{x}_{1:2,t} = \mathbf{f}_t + \kappa \varepsilon_t$. ε_{jt}, e_{it}, and λ_{ki} are all i.i.d standard normal variables. $\delta_{N,T} = \min[\sqrt{N}, \sqrt{T}]$. The reported numbers are the probabilities of correctly identifying the DOFs: $\mathbf{x}_{1:2,t}$ out of 100 replications.

Table 2: Test with DOFs

N	T	$\mathcal{A}_1$	$\mathcal{A}_2$	$\mathcal{A}$	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}$	$A(1)$	$A(2)$
50	50	0.051	0.058	0.056	0.051	0.100	0.080	0.058	0.059
50	100	0.052	0.054	0.054	0.062	0.067	0.068	0.057	0.057
50	150	0.051	0.054	0.053	0.051	0.073	0.073	0.054	0.056
50	200	0.052	0.053	0.052	0.055	0.082	0.075	0.056	0.056
100	50	0.053	0.057	0.055	0.055	0.071	0.065	0.056	0.056
100	100	0.051	0.054	0.053	0.052	0.076	0.067	0.054	0.054
100	150	0.050	0.053	0.052	0.065	0.068	0.070	0.053	0.053
100	200	0.051	0.053	0.053	0.063	0.072	0.075	0.053	0.054
150	50	0.048	0.057	0.053	0.047	0.089	0.072	0.054	0.054
150	100	0.050	0.053	0.052	0.051	0.064	0.054	0.052	0.053
150	150	0.049	0.053	0.051	0.062	0.066	0.064	0.052	0.052
150	200	0.050	0.052	0.051	0.055	0.062	0.074	0.052	0.052
200	50	0.049	0.057	0.054	0.048	0.080	0.069	0.053	0.054
200	100	0.051	0.053	0.052	0.052	0.067	0.062	0.052	0.053
200	150	0.051	0.052	0.052	0.057	0.057	0.070	0.052	0.052
200	200	0.050	0.053	0.052	0.056	0.064	0.052	0.052	0.052

Note: The DGPs are the same as in Table 1 except that $\kappa = 0$. In Columns 3 to 5 are the averaged values of $\mathcal{A}_k$ from 1000 replications. In Columns 6 to 8 are the empirical sizes of the tests $\mathcal{P}_k$ corresponding to the 5% critical value. In Columns 9 to 10 are the averaged values of the $A(j)$ tests of Bai and Ng (2006).

Table 3: Probabilities of Correctly Identifying IOFs

N	T	p_1			p_2			p_3		
		$n = 10$	$n = 20$	$n = 30$	$n = 10$	$n = 20$	$n = 30$	$n = 10$	$n = 20$	$n = 30$
50	50	72.2	63.0	52.2	62.2	44.4	34.6	91.6	90.6	84.4
50	100	86.2	80.8	74.4	81.4	75.2	67.4	92.4	90.2	87.8
50	150	89.8	84.4	78.2	87.4	80.8	74.6	94.6	91.0	87.8
50	200	90.4	86.0	85.6	88.6	83.8	81.8	94.0	92.6	91.2
100	50	91.0	87.8	83.0	87.6	82.6	74.8	96.0	95.2	93.8
100	100	97.6	97.0	94.8	96.8	93.0	90.4	99.4	100	99.6
100	150	99.0	99.0	98.2	98.4	97.6	96.4	99.8	99.4	100
100	200	99.2	99.2	98.4	99.0	99.0	98.2	99.8	100	99.8
150	50	95.6	90.8	90.4	94.0	88.4	87.0	97.8	96.4	95.2
150	100	98.4	99.0	98.8	97.6	97.8	97.4	99.8	100	100
150	150	99.8	99.0	9.3	99.8	99.8	98.6	100	100	99.8
150	200	99.8	100	99.4	99.6	100	99.4	100	100	100
200	50	95.2	94.8	92.0	94.4	92.4	89.8	97.4	96.8	95.4
200	100	99.4	98.4	98.8	99.0	98.4	98.2	99.8	100	100
200	150	100	99.8	100	99.8	99.8	99.8	100	100	100
200	200	100	99.6	99.8	100	99.4	99.8	100	100	100

DGP: $x_{it} = \sum_{k=1}^{2} \lambda_{ki} f_{kt} + e_{it}$, where $\mathbf{f}_t = (f_{1t}, f_{2t})'$ is multivariate normal with $E(f_{kt}) = 0$, $E(f_{kt}^2) = 1$, and $E(f_{1t} f_{2t}) = 0.5$. $f_{1t} = x_{1t} - x_{2t}$, $f_{2t} = x_{3t}$, e_{it}, and λ_{ki} are all i.i.d standard normal variables. The reported numbers are the probabilities of correctly identifying the IOFs: $\mathbf{x}_{1:3,t}$ out of 500 replications for $n = 10, 20, 30$ and 3 different penalty functions p_1, p_2 and p_3.

6.2. Indirectly observed factors

Now we generate data sets with 2 latent factors and 3 observed factors, i.e., $r = 2$ and $m = 3$. The first latent factor is the difference of the first two observed variables: $f_{1t} = x_{1t} - x_{2t}$, and the second latent factors is equal to the third observed variables: $f_{2t} = x_{3t}$. Therefore we can write:

$$\mathbf{f}_t = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \mathbf{x}_{1:3,t}.$$

The other parts of the models are generated as in Section 6.1. We use the method described in Section 4 to identify the IOFs, with $r_{max} = 4$. To reduce the computation cost, we restrict the search to subsets of the variables that contain the IOFs. As discussed in Section 4, the less variables we consider, the more likely that the IOFs are identified. The results from 500 replications for $n = 10, 20, 30$ are reported in Table 3, which shows the probabilities of correctly identifying both the number of IOFs ($m = 3$) and the IOFs.

Several conclusions can be drawn. First, our method performs well in most cases, with high probabilities (more than 80%) of correct identification. Secondly, $p_1(N, T)$ performs best among the three penalty functions we consider. Thirdly, the probabilities of correct identification decrease as we increase the number of variables (n) that include the IOFs, but the reductions are not sharp. For most cases, they are less than 2% when we include 10 extra variables in the searching process.

Next we compare our test statistics proposed in Section 5 to those of Bai and Ng (2006). The discussions in Section 4 implies that for the DGPs considered here, the tests of Bai and Ng (2006) will identify x_{3t} as an observed factor but will reject the null hypothesis for x_{1t} and x_{2t}, while our test should identify all of the three observed factors. The simulation results from 1000 replications are reported in Table 4.

Table 4: Test with IOFs

N	T	$\mathcal{A}_1$	$\mathcal{A}_2$	$\mathcal{A}$	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}$	$A(1)$	$A(2)$	$A(3)$
50	50	0.042	0.048	0.044	0.071	0.066	0.071	0.640	0.705	0.065
50	100	0.044	0.047	0.044	0.069	0.070	0.093	0.644	0.710	0.059
50	150	0.045	0.046	0.045	0.086	0.073	0.112	0.650	0.711	0.059
50	200	0.045	0.046	0.044	0.094	0.098	0.119	0.646	0.707	0.059
100	50	0.043	0.053	0.047	0.044	0.061	0.053	0.752	0.794	0.058
100	100	0.047	0.048	0.045	0.050	0.050	0.054	0.753	0.794	0.055
100	150	0.046	0.048	0.046	0.059	0.066	0.068	0.750	0.803	0.054
100	200	0.047	0.048	0.047	0.066	0.071	0.087	0.755	0.780	0.054
150	50	0.045	0.051	0.048	0.052	0.067	0.070	0.803	0.839	0.055
150	100	0.047	0.050	0.048	0.043	0.054	0.044	0.780	0.845	0.053
150	150	0.047	0.048	0.047	0.042	0.050	0.057	0.780	0.835	0.053
150	200	0.047	0.049	0.047	0.052	0.058	0.058	0.799	0.832	0.052
200	50	0.045	0.052	0.048	0.058	0.066	0.053	0.826	0.857	0.054
200	100	0.046	0.050	0.048	0.049	0.057	0.044	0.827	0.859	0.052
200	150	0.048	0.050	0.048	0.058	0.058	0.067	0.828	0.861	0.052
200	200	0.048	0.050	0.048	0.055	0.049	0.044	0.826	0.855	0.052

Note: The DGPs are the same as in Table 3. In Columns 3 to 5 are the averaged values of $\mathcal{A}_k$ from 1000 replications. In Columns 6 to 8 are the empirical sizes of the tests $\mathcal{P}_k$ corresponding to the 5% critical value. In Columns 9 to 11 are the averaged values of the $A(j)$ tests of Bai and Ng (2006).

It can be seen that our tests (columns 3 to 8) still perform good for the all the sample sizes considered, although our $\mathcal{P}_k$ tests are oversized for small sample sizes. However, the $A(j)$ tests of Bai and Ng (2006) (columns 9 to 12), based on the regressions of IOFs on the estimated factors, fail to converge to 5% for the first two observed factors because they are not directly observed ($x_{1t} - x_{2t} = f_{1t}$). Their tests can only identify x_{3t} as an observed factor because it directly approximates f_{2t}. Hence, these simulation results confirm the superiority of our test in the case of IOFs.

7. Applications

7.1. Data sets

In this part, we use our method to identify the underlying factors that determine the excess returns of portfolios. It is well known that the Fama-French (FF henceforce) 3 factors, including Market excess return (Market), Small Minus Big (SMB) and High Minus Low (HML), are good approximates of the unobservable risk factors, in the sense that they can explain well the variances of the returns. The purpose of the application is to see that, given that the FF 3 factors are the observed counterpart of the underlying risk factors, and that the estimated factors using PC are consistent for the underlying factors, if our method can successfully identify these 3 factors among a panel of other observed variables. One the other hand, if our method fails to identify the FF 3 factors, we should question the consistency of the estimated factors, or the validity of the FF 3 factors as approximations of the underlying risk factors.

We use two data sets in our empirical study. The first data set consists of the monthly returns of 100 portfolios formed on Size and Book-to-Market, which can be downloaded from the webpage of Kenneth French together with the FF 3 factors. The second data set consists of 151

monthly macro series taken from Stock and Watson (2002b), including variables such as industrial production, employment, prices, interest rates, and exchange rates. The macro variables are transformed to achieve stationarity, and the transformation methods for each variable can be found in Stock and Watson (2002b). Both data sets range from 1960 to 1997 (T=444).

We first estimate the factors from the panel of portfolio returns, and then identify the observed factors from the macro data set and FF 3 factors. Beside the FF 3 factors, it is widely believed that asset returns are also commonly affected by some macro fundamentals. The use of the macro data set allows us to find the possible connections between macro variables and financial markets.

7.2. The number of factors

Before estimating the factors, an important question is how many factors are there. We use two different methods to determine the number of factors for both data sets. The first one is the information criteria (IC) method of Bai and Ng (2002), which penalizes extra factors in a proper way such that the penalty functions help to choose the right number of factors. The second one is Onatski (2010), which is based on the fact that in a factor model with r factors, only the largest r eigenvalues of the covariance matrix explode as the number of variables go to infinity, while the remaining eigenvalues are bounded. The method of Bai and Ng (2002) is usually criticized for overestimating the number of factors, the method of Onatski (2010) is shown to have better finite sample performance when there are non-trivial cross sectional correlations between the idiosyncratic errors.

The estimation results for the number of factors are reported in Table 1. It can be seen that for the panel of portfolio returns, 3 to 5 factors are found using different ICs of Bai and Ng (2002), while Onatskis method identifies 3 factors. For the macro data set, the estimated numbers using ICs are all 10, much larger compared to the number (3) found by Onatskis method.

Table 5: The estimated number of factors using the information criteria of Bai and Ng (2002) and the method of Onatski (2010), with $rmak = 10$.

	Samples	PC_1	PC_2	PC_3	IC_1	IC_2	IC_3	Onatski	T	N
Portfolios	1960-1996	5	4	5	3	3	3	3	444	94
	1960-1980	5	4	6	3	3	4	4	240	94
	1980-1996	5	4	7	4	3	5	3	204	94
Macro	1960-1996	10	10	10	10	10	10	3	444	153
Variables	1960-1980	10	9	10	10	8	10	4	240	153
	1980-1996	10	10	10	10	10	10	3	204	153

We then split the sample by 1980 (for reasons discussed below) and estimate the number factors for each subsamples. The results from Onatski (2009) is the same for both data sets: 4 factors for samples from 1960 to 1980 and 3 factors from 1980 to 1997. The results from Bai and Ng (2002) are less consistent: for the financial data set, the estimated numbers range from 3 to 7 for the two subsamples, and the estimated numbers from subsamples are usually larger than those from the full sample. For the macro data set, the selected numbers of factors using ICs are almost all 10 for each subsamples.

As discussed in Chen Dolado and Gonzalo (2011), the differences in the numbers of factors between subsamples and full sample usually imply structural breaks in the factor model, e.g., the

breaks in the factor loadings or the change of factor numbers. However, the number of factors in the full sample should be no less than the number of factors in the subsamples, if the number of factors are correctly estimated. Therefore, the differences of the estimated factors between subsamples and full sample are more likely due to the estimation errors of the two methods in finite samples. Finally, the results in Table 1 strongly favors the specification of 3 factors for both data sets in the full sample.

7.3. Testing for structural instability

Structural instabilities are common features of financial and macro data sets, see Stock and Watson (2003) for examples. There are enough reasons to expect that the factor models considered here are subject to some sort of structural instabilities. Breitung and Eickmeier (2011) is the first paper that proposes formal test statistics to test the null hypothesis of constant factor loadings in large dimensional factor models, but their test is shown to suffer from several shortcomings by Chen, Dolado and Gonzalo (2011), who propose a new test procedure for the same null hypothesis, which is shown to have power only against big breaks in the factor loadings. It is also argued that the small breaks, which are of order $1/\sqrt{NT}$, will not affect the estimation and inference of factor models based on PC methods. A similar test is also proposed by Han and Inoue (2011) independently.

We apply the test of Chen, Dolado and Gonzalo (2011) to both data sets. Their test can be easily implement in two steps. In the first step, the number of factors is chosen or estimated (one can apply the test with different chosen number of factors as we will do here), and the factors are estimated using PC. In the second step, the first estimated factors are regressed on the remaining ones, and a Sup type test of Andrews (1993) is used to test the constancy of the coefficients in this regression. If the null is rejected in the second step, we can conclude that there are big structural breaks in the factor loadings; otherwise there are only mild instabilities that can be safely ignored.

The results with the Sup Wald tests are reported in figures 1 and 2. The chosen numbers of factors range from 3 to 6, consistent with previous findings. For the financial data set, the trimming $[0.3, 0.7]$ is used, while for the macro data set, we use $[0.05, 0.95]$. The red dotted lines are critical values (5% for the financial data and 1% for the macro data) of the Sup type test tabulated by Andrews (1993), and the black dotted lines are the critical values of the χ^2_{r-1} (for a known breaking date, the Wald test converges to χ^2_{r-1}).

The results for the portfolio returns data indicate the existence of a break around 1980, for all the numbers of factors we consider. For the macro data set, the Wald tests strongly reject the null of no structural breaks around 1966 and 1973, implying that there may be multiple breaks during this period. However, the Wald tests for the macro data set can not reject the null at 1% significant level after 1980, even when the tests are compared to the critical values of the χ^2 distributions.

It should be noted that the tests of Chen Dolado and Gonzalo (2011) and Han and Inoue (2011) are based on the relationships between estimated factors. Therefore, these tests also have powers against the breaks in the dynamics of the true factors, which can not be differentiated from breaks in the factor loadings from the tests. However, as pointed out by Chen Dolado and Gonzalo (2011), the breaks in the factor loadings and in the factors can be differentiated by comparing the estimated number of factors in the subsamples and the full sample. If the breaks happen in the factors, the estimated number of factors should be the same before and after the break; if the breaks happen in the factor loadings, the estimated number of factors using

the full sample will usually be larger than that using subsamples. This observation, combined with the results of estimated numbers of factors from previous subsection, provides evidences for the presence of breaks in the factor dynamics rather than the factor loadings. To check the consistency of our results, we apply our method in the following subsection to the full sample and the two subsamples. One should keep in mind that for the macro data, the second subsample (post 1980) is more stable according to our testing results.

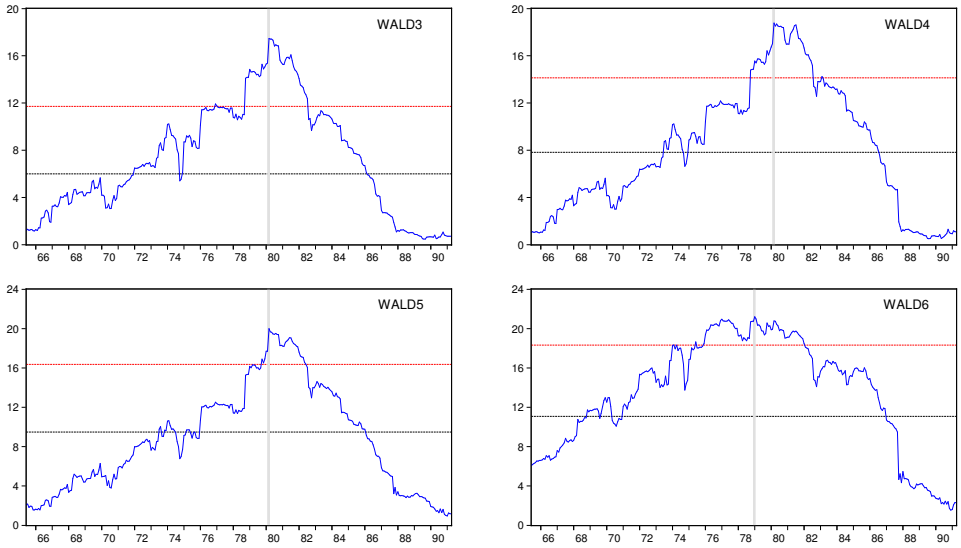

Figure 1: Wald tests for the returns of portfolios, with trimming $[0.3, 0.7]$. Red dotted lines: 5% critical values for the Sup type tests. Black dotted lines: 5% critical values for χ^2 distributions.

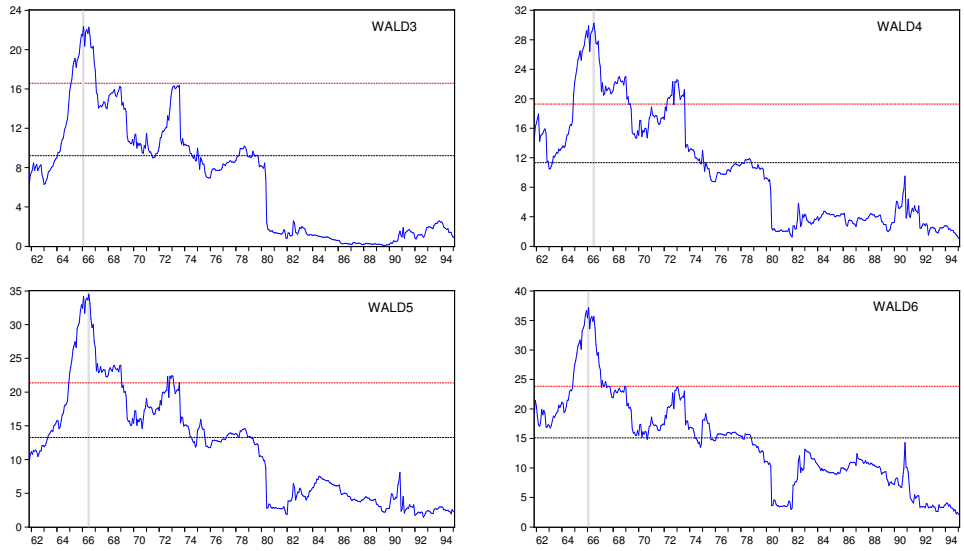

Figure 2: Wald tests for the macro variables, with trimming $[0.05, 0.95]$. Red dotted lines: 1% critical values for the Sup type tests. Black dotted lines: 1% critical values for χ^2 distributions.

7.4. Empirical results

In this subsection, we apply our identification method to find the observed factors in the portfolio returns. We first estimate the factors from the panel of returns, and then form a list of 50 candidates for the observed factors from the panel of macro variables and FF 3 factors, based on their correlations with the estimated factors and their economic meanings. As discussed in Section 4, by creating such a list of candidates, we can significantly reduce the computation cost to an affordable level. The full list of these 50 candidates including their short names, full names and transformation codes are given in the appendix. Besides the FF 3 factors, these 50 candidates include the usual macro variables such as industry production, various interest rates, monetary measures, inflations and consumptions, which have often been considered as the main economic factors that affect the financial market in previous studies.

Finally, we identify the observed factors with each of the estimated factors, starting with the first one, and apply our two type of test statistics to each set of identified observed factors. The results are reported in Table 6.

Table 6: Identification of observed factors for the returns of portfolios

	1960 – 1996				1960 – 1980				1980 – 1996			
	$m=1$	$m=2$	$m=3$	$m=4$	$m=1$	$m=2$	$m=3$	$m=4$	$m=1$	$m=2$	$m=3$	$m=4$
$\hat{f}_{1t}$	Market	Market SMB	Market SMB HML	FYGT10 Market SMB HML	Market	Market SMB	Market SMB HML	HSBR Market SMB HML	Market	Market SMB	Market SMB HML	FYGT10 Market SMB HML
p_1	0.1660	**0.1438**	0.1939	0.2497	0.1709	**0.1540**	0.2064	0.2684	0.1685	**0.1641**	0.2273	0.2910
p_2	0.1685	**0.1488**	0.2013	0.2596	0.1758	**0.1637**	0.2211	0.2880	**0.1744**	0.1758	0.2449	0.3145
p_3	0.1583	**0.1283**	0.1706	0.2187	0.1569	**0.1259**	0.1643	0.2123	0.1521	**0.1313**	0.1781	0.2255
$\mathcal{A}$	0.7410	0.3491	0.2117	0.2117	0.7375	0.4125	0.2208	0.2250	0.6912	0.2745	0.1912	0.2010
$\mathcal{P}$	739.4618	81.3144	31.9472	31.9472	613.5771	73.8801	20.4605	19.653	370.3764	27.6731	16.1182	16.0050
$\hat{f}_{2t}$	SMB	Market SMB	Market SMB HML	FSPIN Market SMB HML	SMB	Market SMB	Market SMB HML	FYGT10 FYBAAC Market SMB	SMB	Market SMB	FYGT10 Market SMB	FSPCOM FSPIN Market SMB
p_1	0.5119	**0.2502**	0.2718	0.3243	0.6100	**0.2688**	0.3227	0.3777	0.3651	**0.2611**	0.3101	0.3655
p_2	0.5144	**0.2551**	0.2793	0.3342	0.6149	**0.2786**	0.3373	0.3973	0.3710	**0.2728**	0.3277	0.3891
p_3	0.5042	**0.2347**	0.2486	0.2933	0.5959	**0.2408**	0.2806	0.3215	0.3487	**0.2283**	0.2609	0.3000
$\mathcal{A}$	0.4955	0.2613	0.1937	0.1914	0.4958	0.2417	0.2167	0.4667	0.5735	0.5049	0.4363	0.4260
$\mathcal{P}$	139.1675	48.0188	28.7555	27.3977	156.1892	34.3494	31.8496	82.7160	91.9813	64.1346	53.9793	48.2843
$\hat{f}_{3t}$	HML	Market HML	Market SMB HML	FYGT10 Market SMB HML	HML	FYGT1 HML	Market SMB HML	FYGT1 Market SMB HML	HML	Market HML	Market SMB HML	FYBAAC Market SMB HML
p_1	0.2912	**0.2864**	0.3110	0.3594	**0.2495**	0.2977	0.3318	0.3895	0.3535	**0.3115**	0.3279	0.3779
p_2	0.2937	**0.2914**	0.3184	0.3694	**0.2544**	0.3075	0.3465	0.4091	0.3594	**0.3233**	0.3456	0.4015
p_3	0.2834	**0.2709**	0.2877	0.3284	**0.2355**	0.2696	0.2897	0.3333	0.3371	**0.2788**	0.2788	0.3124
$\mathcal{A}$	0.3446	0.2072	0.1351	0.1351	0.2000	0.1750	0.1167	0.1208	0.4265	0.4265	0.0833	0.0833
$\mathcal{P}$	51.3162	29.7754	16.4461	15.2645	18.5583	16.4663	11.5260	11.0523	76.2243	67.4627	5.2939	5.2040
$\hat{f}_{4t}$	PWFSA	FSPIN FSPCAP	FSPIN PSPCAP PWFSA	FSPIN FSPCAP PSFSA PUNEW	FYGT5	FSPIN FSPCAP	PMNV FSPIN FSPCAP	PMNV FSPIN FSPCAP FYGT5	SMB	FSPIN SMB	FSNCOM FYGT10 SMB	FSNCOM FYGT10 GMDC SMB
p_1	**1.0458**	1.0797	1.1254	1.1721	1.0259	**0.9661**	1.0100	1.0564	1.0121	0.9928	**0.9561**	0.9917
p_2	**1.0483**	1.0847	1.1328	1.1820	1.0308	**0.9759**	1.0247	1.0760	1.0180	1.0046	**0.9738**	1.0152
p_3	**1.0380**	1.0642	1.1021	1.1411	1.0118	**0.9381**	0.9679	1.0002	0.9957	0.9601	**0.9069**	0.9262
$\mathcal{A}$	0.1441	0.1329	0.1306	0.1509	0.1917	0.1708	0.1708	0.1750	0.3284	0.3333	0.3186	0.3235
$\mathcal{P}$	66.7837	66.1321	65.5915	63.9249	76.6534	74.9080	74.6898	73.1556	50.7730	46.0358	40.3980	39.2865

For each of the estimated factor, we report the minimized object function in (9) with $m = 1, \ldots, 4$ and all the three penalty functions considered in Section 4.3. Several interesting results are worth noting: (1) When assuming 3 factors and the existence of DOFs, almost all the 3 estimated factors identify the FF 3 factors as the observed factors, except for the second estimated factor in the second subsample. (2) when we consider the case of IOFs, the first 2 estimated factors identify Market and SMB as observed factors, and the third estimated factors identify HML in addition to Market. (3) If a forth factor is estimated, the observed factors identified by it are mainly interests variables except for the stock market indices, and the minimized values are much higher than those of the first 3 estimated factors, implying the existence of only 3 underlying factors. (4) The results are robust for the whole sample and the two subsamples, which implies that the breaks found in the previous subsection are in the factor dynamics since such breaks will not affect the consistency of the estimated factors.

We also report the two type of test statistics for the null hypothesis of exact observed factors defined in Section 5, but almost all the tests strongly reject the null, except for the third estimated factor in the second subsample when FF 3 factors are tested. However, the testing results do

not necessarily invalidate our identified observed factors. Because for the case of DOFs, we show that our estimation method can identify the observed factors even with small measurement errors, while the test proposed statistics converge only when the measurement errors are zero. Therefore, a rejection of the hypothesis of exact observed factors does not contradict with the identification of the observed factors with measurement errors. This is also an advantage of our method compared to that of Bai and Ng (2006).

To provide a rough estimate of the size of the measurement errors, recall that we can write:

$$\begin{aligned}\hat{\mathbf{f}}_t &= \mathbf{H}\mathbf{f}_t + \mathbf{v}_t \\ &= \mathbf{H}\mathbf{f}_t + \mathbf{H}\mathbf{B}\mathbf{x}_{1:m,t} - \mathbf{H}\mathbf{B}\mathbf{x}_{1:m,t} + \mathbf{v}_t \\ &= \mathbf{A}\mathbf{x}_{1:m,t} + \mathbf{H}(\mathbf{f}_t - \mathbf{B}\mathbf{x}_{1:m,t}) + \mathbf{v}_t \\ &= \hat{\mathbf{A}}\mathbf{x}_{1:m,t} + (\mathbf{A} - \hat{\mathbf{A}})\mathbf{x}_{1:m,t} + \mathbf{H}(\mathbf{f}_t - \mathbf{B}\mathbf{x}_{1:m,t}) + \mathbf{v}_t,\end{aligned}$$

where $\mathbf{A} = \mathbf{HB}$, and $\hat{\mathbf{A}}$ is the OLS estimate of $\mathbf{A}$. Define $\eta_t = \mathbf{f}_t - \mathbf{B}\mathbf{x}_{1:m,t}$ as the measurement errors, and $\hat{\mathbf{u}}_t = (\mathbf{A} - \hat{\mathbf{A}})\mathbf{x}_{1:m,t} + \mathbf{H}(\mathbf{f}_t - \mathbf{B}\mathbf{x}_{1:m,t}) + \mathbf{v}_t$ as the residuals in the OLS regressions. It is shown in Section 5 that, if $\sqrt{N}/T \to 0$, then $\mathbf{A} - \hat{\mathbf{A}} = o_p(1/\sqrt{N})$ and $\mathbf{v}_t = O_p(1/\sqrt{N})$. Therefore, the proposed test statistics should converge to the same limit distribution as long as $\eta_t = o_p(1/\sqrt{N})$.

Given the stochastic orders of $\mathbf{A} - \hat{\mathbf{A}}$ and $\mathbf{v}_t$, we can get information about the stochastic order of η_t from the residuals $\hat{\mathbf{u}}_t$. Suppose $\hat{\mathbf{u}}_t = O_p(N^\alpha)$, then a simple estimator of α can be given as:

$$\hat{\alpha} = \log(T^{-1}\sum_{t=1}^{T}\hat{\mathbf{u}}_t^2)/2\log(N)$$

because $T^{-1}\sum_{t=1}^{T}\hat{\mathbf{u}}_t^2 = O_p(N^{2\alpha})$. The OLS estimation results of the first 3 estimated factors on the FF 3 factors and the estimated αs for each regression are reported in Table 7. It is obvious that the estimated α for the $\hat{f}_{2t}$ and $\hat{f}_{3t}$ are much larger than $-1/2$, but still less than 0, and the estimated α for the $\hat{f}_{1t}$ are close to $-1/2$. Given the sizes of $\mathbf{A} - \hat{\mathbf{A}}$ and $\mathbf{v}_t$, and the fact that $\hat{f}_{2t}$ and $\hat{f}_{3t}$ put most weights on SMB and HML respectively, it is clear that the factors SMB and HML have larger measurement errors than the Market factor, and these measurement errors cause the tests to reject the null of exact observed factors. However, it should be noted that since all estimated α are less than 0, our estimation method should correctly identify the observed factors despite the measurement errors.

Table 7: Regressions of estimated factors on observed factors.

		Market	SMB	HML	R^2	$\sum \hat{u}_t^2$	$\hat{\alpha}$
60- 96	$\hat{f}_{1t}$	0.1965	0.1239	0.0367	0.9926	3.2460	-0.5364
	$\hat{f}_{2t}$	0.1195	-0.3207	-0.1093	0.8444	68.8791	-0.2032
	$\hat{f}_{3t}$	0.0915	-0.0890	0.3787	0.8698	55.3457	-0.2271
60- 80	$\hat{f}_{1t}$	0.1826	0.1222	0.0452	0.9932	1.6118	-0.5456
	$\hat{f}_{2t}$	0.1584	-0.2966	-0.0491	0.7654	55.7797	-0.1591
	$\hat{f}_{3t}$	0.0492	-0.0618	0.3892	0.8577	32.8917	-0.2167
80- 96	$\hat{f}_{1t}$	0.2123	0.1201	0.0217	0.9941	1.1168	-0.5679
	$\hat{f}_{2t}$	0.1320	-0.3037	0.1852	0.8430	31.1217	-0.2016
	$\hat{f}_{3t}$	0.0381	0.2416	0.3677	0.9201	15.8113	-0.2789

8. Conclusion

In this paper, we have studied the identification of the factors in large dimensional FM. The observed variables that can span the space of the true factors are called observed factors. To identify these observed factors and thus provide interpretations to the orthogonal factors estimated by the method of PC, the estimated factors are regressed on some subsets of the observed variables, and the identified observed factors are those which minimize the RSS in the regressions. We show that, if the observed factors exist, this estimation procedure should identify them with probability approaching 1 as N and T go to infinity. To test the the assumption that the selected observed factors are indeed observed factors, we propose some test statistics based on individual regressions as well as multiple regressions. We show that our test statistics are more general than those of Bai and Ng (2006). The finite sample performance of our methods are studied through simulations.

Appendix A. Proof of Theorem 1

Lemma 3. *(Bai 2003) Let* $\tilde{\mathbf{F}} = (\tilde{\mathbf{F}}_1, \ldots, \tilde{\mathbf{f}}_T)'$*, and* $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_T)'$*, then under Assumptions 1 to 3,*

$$\tilde{\mathbf{F}}'\mathbf{F}/T \xrightarrow{p} \mathbf{Q},$$

where $\mathbf{Q} = \mathbf{V}^{1/2}\mathbf{\Gamma}'\mathbf{\Sigma}_\Lambda^{-1/2}$*,* $\mathbf{V}$ *is a diagonal matrix consisting of the eigenvalues of* $\mathbf{\Sigma}_\Lambda^{1/2}\mathbf{\Sigma}_F\mathbf{\Sigma}_\Lambda^{1/2}$ *in decreasing order and* $\mathbf{\Gamma}$ *consists of the corresponding eigenvectors.*

Lemma 4. *When* $\mathbf{Q}$ *is defined as in Lemma 2, then* $\mathbf{Q}'\mathbf{Q} = \mathbf{\Sigma}_F$*.*

Proof. By definition, $\mathbf{Q}'\mathbf{Q} = \mathbf{\Sigma}_\Lambda^{-1/2}\mathbf{\Gamma}\mathbf{V}^{1/2}\mathbf{V}^{1/2}\mathbf{\Gamma}'\mathbf{\Sigma}_\Lambda^{-1/2} = \mathbf{\Sigma}_\Lambda^{-1/2}\mathbf{\Gamma}\mathbf{V}\mathbf{\Gamma}'\mathbf{\Sigma}_\Lambda^{-1/2}$. Also we have $\mathbf{\Sigma}_\Lambda^{1/2}\mathbf{\Sigma}_F\mathbf{\Sigma}_\Lambda^{1/2} = \mathbf{\Gamma}\mathbf{V}\mathbf{\Gamma}'$, so $\mathbf{Q}'\mathbf{Q} = \mathbf{\Sigma}_\Lambda^{-1/2}(\Sigma_\Lambda^{1/2}\mathbf{\Sigma}_F\mathbf{\Sigma}_\Lambda^{1/2})\mathbf{\Sigma}_\Lambda^{-1/2} = \mathbf{\Sigma}_F$. □

Now let's consider a set of indices $N_1 : N_r = (N_1, \ldots, N_r)$, and the corresponding observed variables $\mathbf{x}_{N_1:N_r,t} = (x_{N_1,t}, \ldots, x_{N_r,t})'$. We can write:

$$\mathbf{x}_{N_1:N_r,t} = \mathbf{\Lambda}_{N_1:N_r}\mathbf{f}_t + \mathbf{e}_{N_1:N_r,t}.$$

We have seen that $\min_{\mathbf{\Lambda}} S(N_1 : N_r, \mathbf{\Lambda}) = S(N_1 : N_r, \hat{\mathbf{\Lambda}})$, where $\hat{\mathbf{\Lambda}}' = [\hat{\mathbf{a}}_1, \hat{\mathbf{a}}_2, \ldots, \hat{\mathbf{a}}_r]$, and $\hat{\mathbf{a}}_k$ is the OLS estimator of $\mathbf{a}_k$. For simplicity, we use $S(N_1 : N_r)$ to denote $S(N_1 : N_r, \hat{\mathbf{\Lambda}})$ in the sequel which is equal to:

$$\frac{1}{T}\mathrm{Tr}\Big[\tilde{\mathbf{F}}'(\mathbf{I}_T - \mathbf{X}_{N_1:N_r}(\mathbf{X}'_{N_1:N_r}\mathbf{X}_{N_1:N_r})^{-1}\mathbf{X}'_{N_1:N_r})\tilde{\mathbf{F}}\Big], \tag{A.1}$$

where $\mathbf{X}_{N_1:N_r} = (\mathbf{x}_{N_1:N_r,1}, \ldots, \mathbf{x}_{N_1:N_r,T})'$. The following result is key to prove Theorem 1.

Lemma 5. *Under Assumptions 1 to 4:*

$$S(N_1 : N_r) \xrightarrow{p} Tr\Big[(\mathbf{\Lambda}_{N_1:N_r}\mathbf{\Sigma}_F\mathbf{\Lambda}'_{N_1:N_r} + \mathbf{\Sigma}^e_{N_1:N_r})^{-1}\mathbf{\Sigma}^e_{N_1:N_r}\Big] \tag{A.2}$$

where $\mathbf{\Sigma}^e_{N_1:N_r} = plim\frac{1}{T}\sum_{t=1}^T \mathbf{e}_{N_1:N_r,t}\mathbf{e}'_{N_1:N_r,t}$*.*

Proof. We have

$$\begin{aligned}
&\frac{1}{T}\tilde{\mathbf{F}}'(\mathbf{I}_T - \mathbf{X}_{N_1:N_r}(\mathbf{X}'_{N_1:N_r}\mathbf{X}_{N_1:N_r})^{-1}\mathbf{X}'_{N_1:N_r})\tilde{\mathbf{F}} \\
= \;& \frac{1}{T}\tilde{\mathbf{F}}'\tilde{\mathbf{F}} - \Big(\frac{\tilde{\mathbf{F}}'\mathbf{X}_{N_1:N_r}}{T}\Big)\Big(\frac{\mathbf{X}'_{N_1:N_r}\mathbf{X}_{N_1:N_r}}{T}\Big)^{-1}\Big(\frac{\mathbf{X}'_{N_1:N_r}\tilde{\mathbf{F}}}{T}\Big) \\
= \;& \mathbf{I}_r - \Big(\frac{\tilde{\mathbf{F}}'\mathbf{X}_{N_1:N_r}}{T}\Big)\Big(\frac{\mathbf{X}'_{N_1:N_r}\mathbf{X}_{N_1:N_r}}{T}\Big)^{-1}\Big(\frac{\mathbf{X}'_{N_1:N_r}\tilde{\mathbf{F}}}{T}\Big).
\end{aligned}$$

One can write $\mathbf{X}_{N_1:N_r} = \mathbf{F}\mathbf{\Lambda}'_{N_1:N_r} + \mathbf{E}_{N_1:N_r}$, where $\mathbf{E}_{N_1:N_r} = (\mathbf{e}_{N_1:N_r,1}, \ldots, \mathbf{e}_{N_1:N_r,T})'$. Then:

$$\begin{aligned}
\frac{\tilde{\mathbf{F}}'\mathbf{X}_{N_1:N_r}}{T} &= \frac{\tilde{\mathbf{F}}'\mathbf{F}}{T}\mathbf{\Lambda}'_{N_1:N_r} + \frac{\tilde{\mathbf{F}}'\mathbf{E}_{N_1:N_r}}{T} \\
&= \frac{\tilde{\mathbf{F}}'\mathbf{F}}{T}\mathbf{\Lambda}'_{N_1:N_r} + \mathbf{H}'\frac{\mathbf{F}'\mathbf{E}_{N_1:N_r}}{T} + \frac{(\tilde{\mathbf{F}} - \mathbf{F}\mathbf{H})'\mathbf{E}_{N_1:N_r}}{T}.
\end{aligned}$$

Firstly, $\frac{\tilde{\mathbf{F}}'\mathbf{F}}{T}$ converges in probability to $\mathbf{Q}$ by Lemma 2. Secondly, $\frac{\mathbf{F}'\mathbf{E}_{N_1:N_r}}{T} = \frac{1}{T}\sum_{t=1}^T \mathbf{f}_t\mathbf{e}'_{N_1:N_r,t}$. If $1 \le i \le r$, then $\frac{1}{T}\sum_{t=1}^T f_{kt}e_{it} = o_p(1)$ by Assumption 4; if $i \ge r$, then $E|\frac{1}{\sqrt{T}}\sum_{t=1}^T f_{kt}e_{it}|^2 =$

$\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}E(f_{ks}f_{kt})E(e_{it}e_{is}) \leq \frac{C}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\gamma_{is} \leq CM$ by Assumptions 1, 2 and 3, where C is a finite constant, and thus $\frac{1}{T}\sum_{t=1}^{T}f_{kt}e_{it}$ is $o_p(1)$. Moreover, we have $\|\mathbf{H}\| = O_p(1)$ (See Bai 2003). Therefore $\mathbf{H}'\frac{\mathbf{F}'\mathbf{E}_{N_1:N_r}}{T} = o_p(1)$. Finally, the last term is $o_p(1)$ by Lemma 1 and thus we have:

$$\frac{\tilde{\mathbf{F}}'\mathbf{X}_{N_1:N_r}}{T} \xrightarrow{p} \mathbf{Q}\mathbf{\Lambda}'_{N_1:N_r}. \tag{A.3}$$

Using similar arguments, we can show that:

$$\frac{\mathbf{X}'_{N_1:N_r}\mathbf{X}_{N_1:N_r}}{T} \xrightarrow{p} \mathbf{\Sigma}^X_{N_1:N_r} = \mathbf{\Lambda}_{N_1:N_r}\mathbf{\Sigma}_F\mathbf{\Lambda}'_{N_1:N_r} + \mathbf{\Sigma}^e_{N_1:N_r}, \tag{A.4}$$

and it is easy to show that $\mathbf{\Sigma}^X_{N_1:N_r} > 0$ and thus is invertible by Assumptions 1(i), 4(i) and 4(iii). Combining the above results we have:

$$\begin{aligned} & \frac{1}{T}\tilde{\mathbf{F}}'(\mathbf{I}_T - \mathbf{X}_{N_1:N_r}(\mathbf{X}'_{N_1:N_r}\mathbf{X}_{N_1:N_r})^{-1}\mathbf{X}'_{N_1:N_r})\tilde{\mathbf{F}} \\ \xrightarrow{p} \quad & \mathbf{I}_r - \mathbf{Q}\mathbf{\Lambda}'_{N_1:N_r}(\mathbf{\Lambda}_{N_1:N_r}\mathbf{\Sigma}_F\mathbf{\Lambda}'_{N_1:N_r} + \mathbf{\Sigma}^e_{N_1:N_r})^{-1}\mathbf{\Lambda}_{N_1:N_r}\mathbf{Q}', \end{aligned}$$

and

$$\begin{aligned} & S(N_1:N_r) \\ \xrightarrow{p} \quad & \mathrm{Tr}\Big[\mathbf{I}_r - \mathbf{Q}\mathbf{\Lambda}'_{N_1:N_r}(\mathbf{\Lambda}_{N_1:N_r}\mathbf{\Sigma}_F\mathbf{\Lambda}'_{N_1:N_r} + \mathbf{\Sigma}^e_{N_1:N_r})^{-1}\mathbf{\Lambda}_{N_1:N_r}\mathbf{Q}'\Big] \\ = \quad & \mathrm{Tr}\Big[\mathbf{I}_r - (\mathbf{\Lambda}_{N_1:N_r}\mathbf{\Sigma}_F\mathbf{\Lambda}'_{N_1:N_r} + \mathbf{\Sigma}^e_{N_1:N_r})^{-1}\mathbf{\Lambda}_{N_1:N_r}\mathbf{Q}'\mathbf{Q}\mathbf{\Lambda}'_{N_1:N_r}\Big] \\ = \quad & \mathrm{Tr}\Big[\mathbf{I}_r - (\mathbf{\Lambda}_{N_1:N_r}\mathbf{\Sigma}_F\mathbf{\Lambda}'_{N_1:N_r} + \mathbf{\Sigma}^e_{N_1:N_r})^{-1}\mathbf{\Lambda}_{N_1:N_r}\mathbf{\Sigma}_F\mathbf{\Lambda}'_{N_1:N_r}\Big] \\ = \quad & \mathrm{Tr}\Big[(\mathbf{\Lambda}_{N_1:N_r}\mathbf{\Sigma}_F\mathbf{\Lambda}'_{N_1:N_r} + \mathbf{\Sigma}^e_{N_1:N_r})^{-1}\mathbf{\Sigma}^e_{N_1:N_r}\Big], \end{aligned}$$

as desired. □

To prove Theorem 1, notice that when the DOFs are selected, $N_1 : N_r = 1 : r = (1, 2, \ldots, r)$, and $\mathbf{\Sigma}^e_{1:r} = 0$ by Assumption 4(i). Therefore, we have

$$\text{plim } S(1:r) = 0.$$

While when we select the wrong set of variables, $\mathbf{\Sigma}^e_{N_1:N_r}$ is either positive definite or positive semi-definite. It is positive semi-definite when part of the selected variables belong to the first r variables, i.e., when there exists at least one N_l such that $1 \leq N_l \leq r$, but it cannot be 0 as long as one of the selected variables does nott belong to DOFs. Then, by the fact that $\mathbf{\Lambda}_{N_1:N_r}\mathbf{\Sigma}_F\mathbf{\Lambda}'_{N_1:N_r} + \mathbf{\Sigma}^e_{N_1:N_r} > 0$, we have

$$\text{plim } S(N_1:N_r) > 0.$$

Then Theorem 1 follows easily.

Appendix B. Proof of Theorem 2

Lemma 6. $\mathbf{Q}\mathbf{\Sigma}_F^{-1}\mathbf{Q}' = \mathbf{I}_r$

Proof. By definition, $\mathbf{Q}\boldsymbol{\Sigma}_F^{-1}\mathbf{Q}' = \mathbf{V}^{1/2}\boldsymbol{\Gamma}'\boldsymbol{\Sigma}_\Lambda^{-1/2}\boldsymbol{\Sigma}_F^{-1}\boldsymbol{\Sigma}_\Lambda^{-1/2}\boldsymbol{\Gamma}\mathbf{V}^{1/2} = \mathbf{V}^{1/2}\boldsymbol{\Gamma}'(\boldsymbol{\Gamma}\mathbf{V}\boldsymbol{\Gamma}')^{-1}\boldsymbol{\Gamma}\mathbf{V}^{1/2} = \mathbf{I}_r$ □

Lemma 7. *For any $k > r$ and $m + 1 \leq N_1 < N_2 \ldots < N_k \leq N$, we have*

$$plim\, S(N_1 : N_k) > 0.$$

Proof. By definition we can write:

$$\mathbf{x}_{N_1:N_k,t} = \boldsymbol{\Lambda}_{N_1:N_k}\mathbf{f}_t + \mathbf{e}_{N_1:N_k,t}.$$

Using the same arguments in Lemma 4, we can show that:

$$S(N_1 : N_r) \xrightarrow{p} \mathrm{Tr}\Big[\mathbf{I}_r - \mathbf{Q}\boldsymbol{\Lambda}'_{N_1:N_k}(\boldsymbol{\Lambda}_{N_1:N_k}\boldsymbol{\Sigma}_F\boldsymbol{\Lambda}'_{N_1:N_r} + \boldsymbol{\Sigma}^e_{N_1:N_k})^{-1}\boldsymbol{\Lambda}_{N_1:N_k}\mathbf{Q}'.\Big] \tag{B.1}$$

But now we can proceed as in the proof of Lemma 4 because $\boldsymbol{\Lambda}_{N_1:N_k}$ is not $r \times r$. Instead, by Lemma 5, we can write the matrix in the right hand side of (B.1)as:

$$\begin{aligned} & \mathbf{Q}\boldsymbol{\Sigma}_F^{-1}\mathbf{Q}' - \mathbf{Q}\boldsymbol{\Lambda}'_{N_1:N_k}(\boldsymbol{\Lambda}_{N_1:N_k}\boldsymbol{\Sigma}_F\boldsymbol{\Lambda}'_{N_1:N_r} + \boldsymbol{\Sigma}^e_{N_1:N_k})^{-1}\boldsymbol{\Lambda}_{N_1:N_k}\mathbf{Q}' \\ = \;& \mathbf{Q}\Big(\boldsymbol{\Sigma}_F^{-1} - \boldsymbol{\Lambda}'_{N_1:N_k}(\boldsymbol{\Lambda}_{N_1:N_k}\boldsymbol{\Sigma}_F\boldsymbol{\Lambda}'_{N_1:N_r} + \boldsymbol{\Sigma}^e_{N_1:N_k})^{-1}\boldsymbol{\Lambda}_{N_1:N_k}\Big)\mathbf{Q}'. \end{aligned}$$

By Assumption 5(iii) and matrix inverse equation we have:

$$\begin{aligned} & (\boldsymbol{\Lambda}_{N_1:N_k}\boldsymbol{\Sigma}_F\boldsymbol{\Lambda}'_{N_1:N_r} + \boldsymbol{\Sigma}^e_{N_1:N_k})^{-1} \\ = \;& (\boldsymbol{\Sigma}^e_{N_1:N_k})^{-1} - (\boldsymbol{\Sigma}^e_{N_1:N_k})^{-1}\boldsymbol{\Lambda}_{N_1:N_r}\Big(\boldsymbol{\Sigma}_F^{-1} + \boldsymbol{\Lambda}'_{N_1:N_r}(\boldsymbol{\Sigma}^e_{N_1:N_k})^{-1}\boldsymbol{\Lambda}_{N_1:N_r}\Big)^{-1}\boldsymbol{\Lambda}'_{N_1:N_r}(\boldsymbol{\Sigma}^e_{N_1:N_k})^{-1}. \end{aligned}$$

Define $\mathbf{C} = \boldsymbol{\Lambda}'_{N_1:N_r}(\boldsymbol{\Sigma}^e_{N_1:N_k})^{-1}\boldsymbol{\Lambda}_{N_1:N_r}$, then we have:

$$\begin{aligned} & \boldsymbol{\Sigma}_F^{-1} - \boldsymbol{\Lambda}'_{N_1:N_k}(\boldsymbol{\Lambda}_{N_1:N_k}\boldsymbol{\Sigma}_F\boldsymbol{\Lambda}'_{N_1:N_r} + \boldsymbol{\Sigma}^e_{N_1:N_k})^{-1}\boldsymbol{\Lambda}_{N_1:N_k} \\ = \;& \boldsymbol{\Sigma}_F^{-1} - \Big(\mathbf{C} - \mathbf{C}(\boldsymbol{\Sigma}_F^{-1} + \mathbf{C})^{-1}\mathbf{C}\Big) \\ = \;& \boldsymbol{\Sigma}_F^{-1} - \Big(\mathbf{C}(\boldsymbol{\Sigma}_F^{-1} + \mathbf{C})^{-1}(\boldsymbol{\Sigma}_F^{-1} + \mathbf{C}) - \mathbf{C}(\boldsymbol{\Sigma}_F^{-1} + \mathbf{C})^{-1}\mathbf{C}\Big) \\ = \;& \boldsymbol{\Sigma}_F^{-1} - \mathbf{C}(\boldsymbol{\Sigma}_F^{-1} + \mathbf{C})^{-1}\boldsymbol{\Sigma}_F^{-1} \\ = \;& (\boldsymbol{\Sigma}_F^{-1} + \mathbf{C})(\boldsymbol{\Sigma}_F^{-1} + \mathbf{C})^{-1}\boldsymbol{\Sigma}_F^{-1} - \mathbf{C}(\boldsymbol{\Sigma}_F^{-1} + \mathbf{C})^{-1}\boldsymbol{\Sigma}_F^{-1} \\ = \;& \boldsymbol{\Sigma}_F^{-1}(\boldsymbol{\Sigma}_F^{-1} + \mathbf{C})^{-1}\boldsymbol{\Sigma}_F^{-1}. \end{aligned}$$

Finally we have:

$$\begin{aligned} & S(N_1 : N_r) \\ \xrightarrow{p} \;& \mathrm{Tr}\Big[\mathbf{Q}\boldsymbol{\Sigma}_F^{-1}(\boldsymbol{\Sigma}_F^{-1} + \mathbf{C})^{-1}\boldsymbol{\Sigma}_F^{-1}\mathbf{Q}'\Big] \\ = \;& \mathrm{Tr}\Big[\boldsymbol{\Sigma}_F^{-1}(\boldsymbol{\Sigma}_F^{-1} + \mathbf{C})^{-1}\Big] \end{aligned}$$

by Lemma 3. Then the result follows by the fact that both $\boldsymbol{\Sigma}_F$ and $\boldsymbol{\Sigma}_F^{-1} + \mathbf{C}$ are positive definite. □

Lemma 8. *If* $\mathbf{e}'\mathbf{e}$ *is the sum of squared residuals when* $\mathbf{y}$ *is regressed on* $\mathbf{X}$ *and* $\mathbf{u}'\mathbf{u}$ *is the sum of squared residuals when* $\mathbf{y}$ *is regressed on* $\mathbf{X}$ *and* $\mathbf{z}$*, then*

$$\mathbf{u}'\mathbf{u} = \mathbf{e}'\mathbf{e} - c^2(\mathbf{z}_*'\mathbf{z}_*) \leq \mathbf{e}'\mathbf{e},$$

where c *is the coefficient on* $\mathbf{z}$ *in the long regression and* $\mathbf{z}_* = [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{z}$ *is the vector of residuals when* $\mathbf{z}$ *is regressed on* $\mathbf{X}$*.*

Proof. See Green (2002). □

Lemma 7 implies that in OLS regressions, adding regressors never increases the RSS.

Lemma 9. $S(1:m) = O_p(\delta_{N,T}^{-2})$.

Proof. By Lemma 1 and Assumption 5(i), we have:

$$\tilde{\mathbf{f}}_t = \mathbf{H}\mathbf{f}_t + \mathbf{v}_t = \mathbf{H}\mathbf{B}\mathbf{x}_{1:m,t} + \mathbf{v}_t = \mathbf{A}\mathbf{x}_{1:m,t} + \mathbf{v}_t, \tag{B.2}$$

where $\mathbf{A} = \mathbf{H}\mathbf{B}$ and $\mathbf{v}_t = O_p(\delta_{N,T}^{-1})$. Then we can write:

$$\tilde{\mathbf{f}}_t = \hat{\mathbf{A}}\mathbf{x}_{1:m,t} + (\mathbf{A} - \hat{\mathbf{A}})\mathbf{x}_{1:m,t} + \mathbf{v}_t.$$

Since

$$\mathbf{A} - \hat{\mathbf{A}} = \Big(T^{-1}\sum_{t=1}^{T}\mathbf{x}_{1:m,t}\mathbf{x}_{1:m,t}'\Big)^{-1}\Big(T^{-1}\sum_{t=1}^{T}\mathbf{x}_{1:m,t}\mathbf{v}_t'\Big) = O_p(\delta_{N,T}^{-1})$$

by Assumption 5(ii). It follows that

$$||\tilde{\mathbf{f}}_t - \hat{\mathbf{A}}\mathbf{x}_{1:m,t}||^2 = O_p(\delta_{N,T}^{-2})$$

and the result follows. □

The following lemma states that if the IOFs are selected together with some other variables, the RSS divided by T also goes to 0.

Lemma 10. *Let* $[1:m, N_1:N_l] = [1, 2\ldots, m, N_1, \ldots, N_l]$ *with* $m < N_1 < \ldots < N_l \leq N$*, then* $S(1:m, N_1:N_l) = O_p(\delta_{N,T}^{-2})$ *for any constant* $l \geq 0$*.*

Proof. The result follows directly from Lemma 7 and Lemma 8. □

Lemma 6 considers the case where none of the selected variables belong to the IOFs. In the following Lemma, we consider the case where only part of IOFs are selected. Without loss of generality, we assume that among the m IOFs, the kth to the last IOFs are selected.

Lemma 11. $S(k:m, N_1:N_l) \geq S(2:m, N_1:N_l)$ *for* $1 < k < m$*, and*

$$plim\, S(2:m, N_1:N_l) > 0.$$

Proof. The first part follows directly from Lemma 7. For the second part, let $\mathbf{y}_t = (\mathbf{x}'_{2:m,t}, \mathbf{x}'_{N_1:N_l,t})'$, and $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_T)'$. Recall that:

$$S(2:m, N_1:N_l) = \sum_{j=1}^{r} S_j(2:m, N_1:N_l)$$

where $S_j(2:m, N_1:N_l) = \frac{1}{T}\sum_{t=1}^{T}(\tilde{f}_{jt} - \hat{\mathbf{a}}'_j\mathbf{y}_t)^2$. Then by Lemma 7 we have:

$$S_j(2:m, N_1:N_l) = S_j(1:m, N_1:N_l) + b^2 T^{-1}\sum_{t=1}^{T}\hat{x}_{1t}^2,$$

where $\hat{x}_{1t}$ are the residuals in the regression of x_{1t} on $\mathbf{y}_t$, and b is the coefficient of x_{1t} in the OLS regression of $\tilde{f}_{jt}$ on x_{1t} and $\mathbf{y}_t$.

By (B.2) we have

$$\tilde{\mathbf{f}}_t = \mathbf{H}\mathbf{B}\mathbf{x}_{1:m,t} + o_p(1) = \mathbf{A}\mathbf{x}_{1:m,t} + o_p(1).$$

If we write $\mathbf{A} = (\mathbf{a}_1, \dots, \mathbf{a}_m)$, then $\|\mathbf{a}_k\|^2 > 0$ for $k = 1, \dots, m$ by Assumption 5(i) and the fact that $\mathbf{H}$ is nonsingular (Bai and Ng 2002). In the vector $\mathbf{a}_1$ there must exit an element $a_{1j} \neq 0$. Thus we can write

$$\tilde{f}_{jt} = a_{1j}x_{1t} + \mathbf{c}\mathbf{y}_t + o_p(1),$$

where $\mathbf{c} = [a_{2j}, \dots, a_{mj}, 0 \dots, 0]$. It follows that $b^2 \xrightarrow{p} a_{1j}^2 > 0$.

Finally, we prove that plim $T^{-1}\sum_{t=1}^{T}\hat{x}_{1t}^2 > 0$ by contradiction. Suppose plim $T^{-1}\sum_{t=1}^{T}\hat{x}_{1t}^2 = 0$, define $\mathbf{z}_t = (\mathbf{x}'_{1:m,t}, \mathbf{x}'_{N_1:N_l,t})'$, and write $x_{1t} = \hat{\mathbf{d}}'\mathbf{y}_t + \hat{x}_{1t}$, where $\hat{\mathbf{d}}$ is the OLS estimator. Then:

$$\begin{aligned}
T^{-1}\sum_{t=1}^{T}\mathbf{z}_t\mathbf{z}'_t &= \begin{pmatrix} T^{-1}\sum_{t=1}^{T}x_{1t}^2 & T^{-1}\sum_{t=1}^{T}\hat{\mathbf{d}}'\mathbf{y}_t\mathbf{y}'_t \\ T^{-1}\sum_{t=1}^{T}\mathbf{y}_t\mathbf{y}'_t\hat{\mathbf{d}} & T^{-1}\sum_{t=1}^{T}\mathbf{y}_t\mathbf{y}'_t \end{pmatrix} \\
&= \begin{pmatrix} \hat{\mathbf{d}}'(T^{-1}\sum_{t=1}^{T}\mathbf{y}_t\mathbf{y}'_t)\hat{\mathbf{d}}' + (T^{-1}\sum_{t=1}^{T}\hat{x}_{1t}^2) & \hat{\mathbf{d}}'(T^{-1}\sum_{t=1}^{T}\mathbf{y}_t\mathbf{y}'_t) \\ (T^{-1}\sum_{t=1}^{T}\mathbf{y}_t\mathbf{y}'_t)\hat{\mathbf{d}} & T^{-1}\sum_{t=1}^{T}\mathbf{y}_t\mathbf{y}'_t \end{pmatrix} \\
&\xrightarrow{p} \begin{pmatrix} \mathbf{d}'\mathbf{\Sigma}_Y\mathbf{d} & \mathbf{d}'\mathbf{\Sigma}_Y \\ \mathbf{\Sigma}_Y\mathbf{d} & \mathbf{\Sigma}_Y \end{pmatrix}.
\end{aligned}$$

The last matrix is a singular matrix, which is a contradiction with Assumption 5(ii). Therefore we have:

$$\begin{aligned}
\text{plim } S_j(2:m, N_1:N_l) &= \text{plim } S_j(1:m, N_1:N_l) + \text{plim }(b^2 T^{-1}\sum_{t=1}^{T}\hat{x}_{1t}^2) \\
&= 0 + a_{1j}^2 \text{plim }(T^{-1}\sum_{t=1}^{T}\hat{x}_{1t}^2) > 0.
\end{aligned}$$

And thus plim $S(2:m, N_1:N_l) \geq$ plim $S_j(2:m, N_1:N_l) > 0$.

□

Proof of Theorem 2

Proof. Since

$$\begin{aligned} & \mathbb{P}[\hat{m} = m, (\hat{N}_1, \ldots, \hat{N}_{\hat{m}}) = (1, \ldots, m)] \\ = \quad & \mathbb{P}[\hat{m} = m] \cdot \mathbb{P}[(\hat{N}_1, \ldots, \hat{N}_{\hat{m}}) = (1, \ldots, m) | \hat{m} = m], \end{aligned}$$

and it is obvious that $\mathbb{P}[(\hat{N}_1, \ldots, \hat{N}_{\hat{m}}) = (1, \ldots, m) | \hat{m} = m] \to 1$ as $N, T \to \infty$ by Lemma 6, 8 and 10, it suffices to prove that $\mathbb{P}[\hat{m} = m] \to 1$.

When $l < m$:

$$\begin{aligned} & \mathbb{P}[\hat{m} = l] \\ = \quad & \mathbb{P}[\min S(N_1 : N_l) + l \cdot p(N, T) > \min S(N_1 : N_m) + m \cdot p(N, T)] \\ = \quad & \mathbb{P}[\min S(N_1 : N_l) - \min S(N_1 : N_m) > (m - l) \cdot p(N, T)] \end{aligned}$$

By Lemma 6 and 10, we have $\operatorname{plim\,inf} S(N_1 : N_l) = \tau_l > 0$, and $\operatorname{plim} \min S(N_1 : N_l) = 0$. Then we have $\mathbb{P}[\hat{m} = l] \to 0$ because $p(N, T) \to 0$.

Similarly, when $l > m$:

$$\begin{aligned} & \mathbb{P}[\hat{m} = l] \\ = \quad & \mathbb{P}[\min S(N_1 : N_l) + l \cdot p(N, T) < \min S(N_1 : N_m) + m \cdot p(N, T)] \\ = \quad & \mathbb{P}[\min S(N_1 : N_m) - \min S(N_1 : N_l) > (l - m) \cdot p(N, T)] \end{aligned}$$

From Lemma 8,9 and 10 we know that $\min S(N_1 : N_m) - \min S(N_1 : N_l) = O_p(\delta_{N,T}^{-2})$. By the assumption that $\delta_{N,T}^{-2} p(N, T) \to \infty$ as $N, T \to \infty$, $\cdot p(N, T)$ goes to zero slower than $\min S(N_1 : N_m) - \min S(N_1 : N_l)$, therefore $\mathbb{P}[\hat{m} = l] \to 0$ as $N, T \to \infty$. The desired result then follows easily. □

Appendix C. Proof of Proposition 2

Proof. Following the argument of Bai and Ng (2006), $\{\sqrt{N}\hat{\mathbf{v}}_t\}$ are asymptotic normal and independent under the assumption that $\{e_{it}\}$ are serially uncorrelated. It follows that $\hat{\rho}_t$ and $\hat{\rho}_{t,k}$ are also independent. Then the results follow easily from Central Limit Theorem. □

Appendix D. Tables and Figures

Table D.8: Candidates for Observed Factors

	Short Name	Long Name	T code
1	IP	industrial production: total index	5
2	IPMFG	industrial production: manufacturing	5
3	IPXMCA	capacity util rate: manufacturing, total	1
4	LHELX	employment: ratio; help-wanted ads:no. unemployment clf	4
5	LHUR	unemployment rate: all workers, 16 years over	1
6	LPNAG	employment on nonag. payrolls: total	5
7	LEHCC	avg hr earnings of constr wkrs: construction	6
8	LEHM	avg hr earings of prod wrks: manufacturing	6
9	HSFR	housing starts: nonfarm (1947-58) ; total farm & nonfarm(1959-)	4
10	HSBR	housing authorized by build: total new priv housing units	4
11	MSMTQ	manufacturing & trade: total	5
12	MSMQ	manufacturing & trade: manufacturing; total	5
13	WTQ	merchant wholesalers: total	5
14	RTQ	retail trade:total	5
15	IVMTQ	manufacturing & trade inventories: total	5
16	PMI	purchasing managers' index	1
17	PMP	napm production index	1
18	PMNO	napm new orders index	1
19	PMNV	napm inventories index	1
20	PMEMP	napm employment index	1
21	MO	mfg new orders: all manufacturing industries, total	5
22	MDO	mfg new orders: durable goods industries, total	5
23	FM2	money stock: m2	6
24	FMFBA	monetary base, adj for reserve requirement changes	6
25	FSNCOM	NYSE common stock price index: composite	5
26	FSPCOM	S&P common stock price index: composite	5
27	FSPIN	S&P common stock price index: industries	5
28	FSPCAP	S&P common stock price index: capital goods	5
29	FYFF	interest rate: federal funds	2
30	FYCP90	interest rate: 90 day commercial paper	2
31	FYGM3	interest rate: U.S. treasury bills, sec mkt, 3-m0	2
32	FYGM6	interest rate: U.S. treasury bills, sec mkt, 3-m0	2
33	FYGT1	interest rate: U.S. treasury const maturities, 1-yr	2
34	FYGT5	interest rate: U.S. treasury const maturities, 5-yr	2
35	FYGT10	interest rate: U.S. treasury const maturities, 10-yr	2
36	FYAAAC	bond yield: moody's aaa corporate	2
37	FYBAAC	bond yield: moody's baa corporate	2
38	FYFHA	secondary market yields on fha mortgages	2
39	EXRUS	United States effective exchange rate	5
40	EXRGER	foreign exchange rate: Germany	5
41	EXRJAN	foreign exchange rate: Japan	5
42	EXRUK	foreign exchange rate: United Kingdom	5
43	EXRCAN	foreign exchange rate: Canada	5
44	PWFSA	producer price index: finished goods	6
45	PUNEW	cpi-u: all items	6
46	PUC	cpi-u: commodities	6
47	GMDC	pce, impl pr defl: pce	6
48	Market	Market minus risk free rate	1
49	SMB	small minus big	1
50	HML	high minus low	1

References

Altug, S. (1989). Time-to-build and aggregate fluctuations: some new evidence. *International Economic Review*, 889–920.

Andrews, D. (2003). Tests for parameter instability and structural change with unknown change point: a corrigendum. *Econometrica*, 395–397.

Andrews, D. W. K. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica 61*(4), pp. 821–856.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica 71*(1), 135–171.

Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica 77*(4), 1229–1279.

Bai, J. and S. Ng (2002). Determining the number of factors in approximate factor models. *Econometrica 70*(1), 191–221.

Bai, J. and S. Ng (2006). Evaluating latent and observed factors in macroeconomics and finance. *Journal of Econometrics 131*(1-2), 507–537.

Bernanke, B., J. Boivin, and P. Eliasz (2005). Measuring the effects of monetary policy: a factor-augmented vector autoregressive (favar) approach. *The Quarterly Journal of Economics 120*(1), 387.

Boivin, J. and M. Giannoni (2006). Dsge models in a data-rich environment. Technical report, National Bureau of Economic Research.

Breitung, J. and S. Eickmeier (2011). Testing for structural breaks in dynamic factor models. *Journal of Econometrics 163*(1), 71 – 84.

Chamberlain, G. and M. Rothschild (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica 51*(5), 1281–1304.

Chen, L., J. Dolado, and J. Gonzalo (2011). Detecting big structural breaks in large factor models. *manuscript, Universidad Carlos III de Madrid*.

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

Fama, E. and K. French (1993). Common risk factors in the returns on stocks and bonds. *Journal of financial economics 33*(1), 3–56.

Forni, M., D. Giannone, M. Lippi, and L. Reichlin (2009). Opening the black box: Structural factor models with large cross-sections. *Econometric Theory 25*(5), 1319–1347.

Han, X. and A. Inoue (2011). Tests for parameter stability in dynamic factor models. *manuscript, NC State Univserity*.

Kryshko, M. (2011). Data-rich dsge and dynamic factor models.

Lewbel, A. (1991). The rank of demand systems: theory and nonparametric estimation. *Econometrica: Journal of the Econometric Society*, 711–730.

Onatski, A. (2009a). Asymptotics of the principal components estimator of large factor models with weak factors. *manuscript, Columbia University*.

Onatski, A. (2009b). Testing hypotheses about the number of factors in large factor models. *Econometrica 77*(5), 1447–1479.

Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics 92*(4), 1004–1016.

Pesaran, M. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica 74*(4), 967–1012.

Sargent, T. (1989). Two models of measurements and the investment accelerator. *The Journal of Political Economy*, 251–287.

Shanken, J. and M. Weinstein (2006). Economic forces and the stock market revisited. *Journal of Empirical Finance 13*(2), 129–144.

Stock, J. and M. Watson (2002a). Forecasting using principal components from a large number of predictors. *Journal of the American statistical association 97*(460), 1167–1179.

Stock, J. and M. Watson (2003). Has the business cycle changed and why?

Stock, J. and M. Watson (2009). Forecasting in dynamic factor models subject to structural instability. *The Methodology and Practice of Econometrics. A Festschrift in Honour of David F. Hendry*, 173–205.

Stock, J. H. and M. W. Watson (2002b). Macroeconomic forecasting using diffusion indexes. *Journal of Business and Economic Statistics 20*(2), 147–162.