Determining the Number of Factors from Empirical Distribution of Eigenvalues

Alexei Onatski
Economics Department, Columbia University

March 31, 2005

Abstract

We develop a new consistent and simple to compute estimator of the number of factors in the large dimensional approximate factor models. The main advantage of our estimator relative to the previously proposed ones is that it works well in realistically small samples when the amount of cross-sectional and time-series correlation in the idiosyncratic terms is relatively large. It also improves upon the existing methods when the portion of the observed variance attributed to the factors is small relative to the variance due to the idiosyncratic term. These advantages arise because the estimator is based on a Law-of-Large-Numbers type regularity for the idiosyncratic components of the data, as opposed to the estimators based on the assumption that a significant portion of the variance is explained by the systematic part. We apply the new estimation procedure to determine the number of pervasive factors driving returns of stocks traded on NYSE, AMEX, and NASDAQ and the number of pervasive factors influencing dynamics of a large set of macroeconomic variables.

1 Introduction

Factor models with large cross-section and time-series dimensions have recently attracted an increasing amount of attention from researchers in finance and macroeconomics. Approximate factor models, where the idiosyncratic components may be weakly correlated and the common factors non-trivially affect a large number of the cross-sectional units are particularly useful in applications. In finance, such models are at the heart of Chamberlain and Rothchild's (1983) and Ingersol's (1984) extension of the arbitrage pricing theory. In macroeconomics, the models are used to identify economy-wide and global shocks, to construct coincident indexes, to forecast individual macroeconomic time series, to study relationship between microeconomic and aggregated macroeconomic dynamics, and to augment information in the VAR models used for monetary policy analysis (see, for example, Forni and Reichlin (1998), Forni, Hallin, Lippi, and Reichlin (2000), Stock and Watson (1999), Forni and Lippi (1999), and Bernanke, Boivin, and Eliasz (2004)).

An important question to be addressed by researchers using the approximate factor models is how many factors are there. This question is directly related to the behavior of the eigenvalues of the data's covariance matrix as the number of the cross-sectional units tends to infinity. By definition of the approximate factor models, the eigenvalues of the covariance matrix of the systematic components of the data must increase without bound. At the same time, the eigenvalues of the covariance matrix of the idiosyncratic components must stay bounded. For the data's covariance matrix this translates into the first r eigenvalues, where r is the number of factors, increasing without bound and the rest of the eigenvalues staying bounded. Unfortunately, as has been noted by Trzcinka (1986) and Luedecke (1984) among many others, testing whether some eigenvalues increase unboundedly whereas the other remain bounded is not a well-posed problem with a finite number of data points. Forni et al (2000 p.547) describe the problem particularly clearly: "there is no way a slowly diverging sequence (divergence under the model can be arbitrarily slow) can be told from an eventually bounded sequence (for which the bound can be arbitrarily large)".

To distinguish the diverging sequence from the bounded sequence, this paper restricts the approximate factor models by imposing some structure on the idiosyncratic components of the data. Precisely, we assume that the vector of the idiosyncratic terms is a linear transformation of a vector with i.i.d. components. The linear transformation is left relatively unconstrained so that a wide range of heteroskedasticity and cross-sectional serial correlation patterns for the idiosyncratic terms is allowed. Using this assumption and recent results from the large dimensional random matrix literature (see Z. Bai (1999) for a review), the paper shows how to estimate an upper bound on the eigenvalues of the "idiosyncratic part" of the sample covariance matrix. Counting the eigenvalues of the sample covariance matrix that are above the bound gives our estimator of the number of the factors.

In more detail, the central fact underlying our estimator is that the empirical distribution of eigenvalues of the sample covariance matrix converges to a non-random distribution when both the time series and cross-sectional dimensions of the data grow. The limiting distribution has bounded support and known functional form in the vicinity of the upper boundary u of the support. We show that, asymptotically, the first r eigenvalues of the sample covariance matrix are almost surely larger than u, where r is the true number of factors. However, the $r + 1$-th eigenvalue almost surely converges to u. To estimate u, we choose the parameters of the known functional form of the limiting distribution so that it fits a small rightmost portion of the empirical distribution of eigenvalues of the sample covariance matrix well. Finally, we count the number of eigenvalues of the sample covariance matrix that lie above our estimate of u.

We show that this estimator is consistent and use numerical simulations to demonstrate that it has good finite sample properties in many empirically-relevant situations. In particular, although the estimator is developed under an assumption that the idiosyncratic terms are cross-sectionally correlated but independent across time, our Monte Carlo results suggest that it works well for relatively large amount of cross-sectional and time-series dependence simultaneously present in the idiosyncratic terms. We informally discuss some theoretical reasons to expect such a good performance in the conclusion section.

The constraint that we impose on the approximate factor models conceptually differs from the restriction considered by many previous studies, including Connor and Korajczyk (1993), Stock and Watson (1999), and Bai and Ng (2002). These studies require that the eigenvalues of the covariance matrix of the systematic part of the data increase fast, i.e. proportionately to the number of the cross-sectional units. The fast growth assumption guarantees that the average variability explained by the factors stays away from zero even when the dimensionality of data increases to infinity. At the same time, as the eigenvalues of the idiosyncratic covariance matrix remain bounded, the average variability explained by any function of the idiosyncratic terms tends to zero as the data size grows. This creates a possibility for using model selection criteria for the number of factors determination.

Obviously, the fast growth restriction rules out situations when some of the factors are weak in the sense that although the cumulative effect of these factors grows without limit as the data size increases, the average effect vanishes. From a theoretical point of view, "weak" factors may be important for the approximate asset pricing formula of Chamberlain and Rothschild (1983). Indeed, the formula includes betas corresponding to factors with the cumulative effect, measured by the sum of the squared factors loadings, increasing to infinity not necessarily as fast as the number of the stocks in the data set.

More importantly, there exist several empirically relevant finite sample situations poorly approximated by asymptotics implied by the fast growth restriction. On one hand, the amount of the serial correlation in the idiosyncratic terms may be relatively large and the data size relatively small so that certain linear combinations of the idiosyncratic terms will have sizable effect on a nontrivial portion of observations. This will create a problem for model selection criteria because the explanatory power of some idiosyncratic shocks may be too large to distinguish them from the factors. On the other hand, it may be the case that the variability of the idiosyncratic components is large relative to the variability of the factors. Then, the portion of the variability explained by factors may be too close to zero for information criteria to distinguish them from the idiosyncratic components.

Our Monte Carlo simulations show that the above situations, corresponding to reasonable data sizes and reasonable amount of dependence in the idiosyncratic terms, indeed result in poor performance of Bai and Ng (2002) information criteria estimators. In contrast, our estimator does a good job for a wide range of cross-sectional and time-series correlation patterns and "signal-to-noise" ratios.

Since our approach is explicitly based on the investigation of behavior of the eigenvalues of the data's covariance matrix, it is related to the earlier literature exploiting the information contained in the eigenvalues. Trzcinka (1986) investigates the question of the number of factors in Chamberlain and Rothchild's (1983) extension of the arbitrage pricing theory by inspecting growth patterns of the eigenvalues of the sample covariance matrix as the number of assets in the data set increases. According to the theory, the eigenvalues of the covariance matrix that correspond to the systematic component of the data should grow without limit whereas the rest of the eigenvalues should be bounded. Trzcinka's informal analysis has been criticized from several perspectives. Brown (1989) points out that in an economy with r equally important factors the largest eigenvalue of the covariance matrix will grow much faster than the other $r-1$ eigenvalues creating a "single factor illusion". Connor and Korajczyk (1993) explain that although the eigenvalues corresponding to the idiosyncratic component of the population covariance matrix should be bounded, all eigenvalues of the sample covariance matrix will grow without limit as the number of cross-sectional units grow faster than the number of observations across time.

Since our estimate of the number of factors does not rely on a visual inspection of any graphs, Brown's criticism does not apply. As to Connor and Korajczyk's argument, we assume in the paper that the ratio of the time series dimension to the cross-sectional dimension tends to a non-zero number. Therefore, the sample eigenvalues corresponding to the idiosyncratic part of the data remain bounded. It is still true that the bounds on the population and sample eigenvalues will be different, but it is the bound on the sample eigenvalues that we estimate in this paper. Hence, our number of factors determination procedure uses the correct bound. As Monte Carlo simulations show, our estimate of the number of factors remains good even for small ratio of the time series to cross-sectional dimensions, a situation particularly relevant for applications.

There has been at least one recent study of the number of factors determination exploiting ideas from the large dimensional random matrix theory. Kapetanios (2004) proposes a consistent criterion based on the explicit calculation of the bound for the eigenvalues corresponding to the idiosyncratic component of the data. Kapetanios' bound depends only on the ratio of the time series to cross-sectional dimension of the data. Unfortunately, the bound's validity requires relatively restrictive assumption on the cross-sectional serial correlation of the idiosyncratic terms, which significantly narrows the range of applications of the method. In contrast, we estimate our bound from the data. The bound can vary from application to application and allows for relatively unrestricted form of the heteroskedasticity and cross-sectional correlation of the idiosyncratic terms.

We apply the newly developed estimation procedure to estimate the number of factors in the arbitrage pricing theory and the number of factors driving a large set of the US macroeconomic time series. For the arbitrage pricing theory, we find evidence that there exist eight pervasive factors. Bai and Ng's (2002) estimators suggest the existence of 3 to 6 pervasive factors for our data set. One possible explanation of the difference is that some important factors do not have sufficiently widespread influence on the returns or have widespread but weak influence, which makes the Bai-Ng method relegate them to the idiosyncratic component. For the large set of the US macroeconomic time series, we find 6 pervasive factors. In contrast, the Bai-Ng estimators suggest that there are 12 pervasive factors. It is possible that the amount of dependence in the idiosyncratic terms in our data set is too large for the Bai-Ng information criteria to distinguish some idiosyncratic shocks from factors.

The rest of the paper is organized as follows. In section 2 we describe the approximate factor model. Section 3 develops the new method of the number of factors determination. In section 4 we do Monte Carlo simulations to compare the performance of our method with that of Bai and Ng (2002). Section 5 uses the new method to estimate the number of factors in the arbitrage pricing theory and in a large macroeconomic panel. Section 6 concludes.

2 Approximate factor model

In this paper, we study approximate, in the sense of Chamberlain and Rothschild (1983), factor models of the form

$$X_t = \Lambda F_t + e_t, \tag{1}$$

where X_t is an $n \times 1$ vector of the cross-sectional observations at time period t and ΛF_t and e_t are unobserved systematic and idiosyncratic components of this vector respectively. The systematic part is a product of an $n \times r$ matrix of factor loadings Λ and an $r \times 1$ vector of factors F_t, which are common for all cross-sectional units but may change over time. We are interested in estimating the unknown number of factors r in (1).

Our baseline case is when the unknown number of factors is fixed, that is it does not change with the dimensionality of the data. In macroeconomic applications, the pervasive factors, arguably, should correspond to some economy-wide structural shocks. It is tempting to think that such structural shocks can be traced down to a few important sources of fluctuations. From this perspective, the requirement that the number of factors is fixed does not seems too restrictive. Recall that in the approximate factor models, the idiosyncratic components of the data can be correlated. If one is willing to model the idiosyncratic components using a traditional factor model, the number of factors in such a model is free to rise with the dimensionality of data. It is only the number of the pervasive factors, $\dim(F_t)$, that we want to bound. Anyway, after getting the results for the baseline case, we extend our analysis to the case of the slowly growing number of pervasive factors. For both cases, we assume that the true number of factors is capped by $r_{\max}$, the smallest integer larger than $\min(n^\alpha, T^\alpha)$, where $0 < \alpha < 1$.

We assume that both cross-sectional (n) and time-series (T) dimension of the data available for the estimation is large. Precisely, we make the following

Assumption 1. *n and T tend to infinity so that $\frac{n}{T} \to c$, where $c \in (0, \infty)$.*

The assumption differs from those made in the previous literature. Connor and Korajczyk (1993) develop their number of factors estimation method using sequential limit asymptotics when first n tends to infinity and then T tends to infinity. Stock and Watson (1999) assume that $\sqrt{n}/T$ goes to infinity and Bai and Ng (2002) allow n and T to go to infinity simultaneously and without any restrictions on the relative growth rates. Assumption 1 is however standard in the statistical literature on large dimensional random matrices and we adopt it here. Note that the limit c may be any positive number, so the asymptotics is consistent with a variety of empirically relevant finite sample situations.

In contrast to the exact factor models (see Anderson 1984), the covariance matrix of the idiosyncratic vector e_t does not need to be diagonal. The identification of the systematic part of the data is based on the assumption that the largest eigenvalue of the covariance matrix for the idiosyncratic vector is bounded, whereas all eigenvalues of the covariance matrix of the systematic part ΛF_t tend to infinity. For the systematic part of the data, we assume

Assumption 2. $\min \mathrm{eval}\,(\Lambda'\Lambda) \to \infty,\ B_1 < \mathrm{eval}\left(\frac{1}{T}\sum_{t=1}^{T} F_t F_t'\right) < B_2$ *almost surely for some fixed* $0 < B_1 \leq B_2 < \infty$

Here $\min \mathrm{eval}(M)$ denotes the smallest eigenvalue of matrix M. Intuitively, assumption 2 implies that factors F_t non-trivially affect an increasing number of cross-sectional units. We therefore will call the factors pervasive. Note that we do not require stationarity of F_t and do not impose any convergence restrictions so that $\frac{1}{n}\Lambda'\Lambda$ and $\frac{1}{T}\sum_{t=1}^{T} F_t F_t'$ do not need to converge to any limits. Moreover, we do not require factors be independent from the idiosyncratic terms. Connor and Korajczyk (1993), Stock and Watson (1999), and Bai and Ng (2002) make stronger assumptions on the factors and factor loadings. In particular, their assumptions imply that $\min \mathrm{eval}\,(\Lambda'\Lambda) > an$, for some $a > 0$ and large enough n. Loosely speaking, we allow for weaker pervasive factors than Connor and Korajczyk, Stock and Watson, and Bai and Ng do.

Relaxing the Stock-Watson and Bai-Ng assumptions on factor loadings has a practical value. As discussed in the introduction, the "weaker" pervasive factors can be a good approximation to the finite sample situations when the amount of dependence in the idiosyncratic terms is relatively large and/or the portion of the variation in the data explained by the factors is low relative to the variation due to the idiosyncratic term.

The flip side of our flexibility in definition of the systematic part of the data is more stringent restrictions on the idiosyncratic part. In this paper we assume that the idiosyncratic vector e_t is a linear transformation of an $n \times 1$ vector ε_t with i.i.d. components. Precisely, our next assumption is:

Assumption 3. *There exists an* $n \times n$ *random matrix* S_n*, such that*

$$e_t = S_n \varepsilon_t, \tag{2}$$

where $\varepsilon_t = (\varepsilon_{1t}, ..., \varepsilon_{nt})'$, $E\varepsilon_{it} = 0$, $E\varepsilon_{it}^2 = 1$, $E\varepsilon_{it}^4 < \infty$, ε_{it} *are i.i.d. for* $1 \leq i \leq n$, $1 \leq t \leq T$ *and* S_n *and* ε_t *are independent.*

The assumption implies that the covariance matrix of e_t is equal to $S_n S_n'$, which does not need to be diagonal. Therefore, we allow for cross-sectional serial correlation and heteroskedasticity in the idiosyncratic terms. However, we require e_t to have no serial correlation over the time dimension. This requirement is technical and is likely not necessary for the consistency of the estimator proposed below. In the conclusion section, when describing our plans for future work, we outline a possible way to relax the requirement.

Without any restrictions on S_n, the covariance matrix of e_t may have unbounded eigenvalues and thus disagree with the definition of the idiosyncratic component. We, therefore, will assume that the eigenvalues of $S_n S_n'$ are bounded. Moreover, we will require the distribution of the eigenvalues to converge in the following sense. Let $\lambda_1 \geq ... \geq \lambda_n$ be the eigenvalues of a generic $n \times n$ positive semi-definite matrix A. We define the eigenvalue distribution function for A, or as we will call it the empirical spectral distribution of A, as

$$F^A(x) = 1 - \frac{1}{n} \# \left\{ i \leq n : \lambda_i > x \right\}, \tag{3}$$

where $\# \{\cdot\}$ denotes the number of elements in the set indicated. Note that $F^A(x)$ is a valid cumulative probability distribution function (cdf). Further, for a generic probability distribution having a bounded support and cdf $G(x)$, let $u(G)$ be the upper bound of the support, that is

$$u(G) = \min \left\{ x : G(x) = 1 \right\}.$$

We will make the following

Assumption 4. *i)* $F^{S_n' S_n} \to H$ *almost surely, where* H *is a fixed cumulative distribution function with bounded support and the convergence is the weak convergence of distributions;*
ii) $u(F^{S_n' S_n}) \to u(H)$ *almost surely;*
iii) $c \displaystyle\int \frac{t^2 dH(t)}{\left(u(H) - t\right)^2} > 1$ *if the integral exists*

Part i) of the assumption is needed to insure convergence of the spectral distribution of the sample covariance matrix of the idiosyncratic term $\frac{1}{T} \sum_{t=1}^{T} e_t e_t'$ to a distribution with a bounded support. The idea is to estimate the upper bound of this support and use it as a threshold above which the eigenvalues of the data's sample covariance matrix $\frac{1}{T} \sum_{t=1}^{T} X_t X_t'$ correspond to the systematic part of the data. Of course, the weak convergence of a distribution to a distribution with bounded support does not imply the supports converge. For example, $N(0, 1/n)$ converges to a mass at zero, but has an unbounded support. That is why we need part ii) of the assumption. It guarantees that for large n the largest eigenvalue of $\frac{1}{T} \sum_{t=1}^{T} e_t e_t'$ will converge to the upper bound of the limiting spectral distribution. Finally, assumption iii) does not like limiting spectral distributions with thin tail.[1] Indeed, for inequality in iii) to be violated the limiting spectral distribution must have density and the first derivative of this density vanishing at $u(H)$. Intuitively, this can be the case when a handful of linear combinations of ε_t explain a disproportionately large part of the variation in the idiosyncratic term, which makes these combinations look very much like common factors for the components of X_t. Our estimation method will break down in this case.

[1] Assumption 4 iii) is used in the proof of lemma 2 below.

In our opinion, assumption 4 is not very restrictive. For example, a common way to model a vector of serially correlated observations e_t is to assume that $e_t = S_n \varepsilon_t$, where S_n is a symmetric matrix constant along the diagonals (a Hermitian Toeplitz matrix). It can be shown (see, for example, Bottcher and Silbermann, 1998, pp.138-143) that the spectral distribution of Hermitian Toeplitz matrices converges to a distribution with bounded support as the size of the matrix increases. Moreover, the density of the limiting distribution will actually explode near the boundary of the support. For purely heteroskedastic series, parts i) and ii) of our assumption will be guaranteed if the variances of the observations are drawn from the limiting spectral distribution, which does not seem counterintuitive. As to part iii), it should be viewed as a basic identification assumption. Without this part, we are back to the problem of not being able to separate slowly increasing sequences from eventually bounded sequences with an arbitrary large bound.

3 New Estimator

Now, we are ready to describe our estimator of the number of factors. Let X, F, and e be the $n \times T$, $r \times T$ and $n \times T$ matrices with t-th columns equal to X_t, F_t and e_t respectively. Then (1) can be rewritten as

$$X = \Lambda F + e. \tag{4}$$

Let λ_i be the i-th largest eigenvalue of the data's sample covariance matrix $\frac{1}{T}XX'$. We define a family of estimators:

$$\hat{r}_\delta = \# \left\{ i \leq n : \lambda_i > (1+\delta)\hat{u} \right\}, \tag{5}$$

indexed by a positive number δ, where $\hat{u} = w\lambda_{r_{\max}+1} + (1-w)\,\lambda_{2r_{\max}+1}$ and $w = 2^{2/3}/\left(2^{2/3}-1\right)$. Below, we will prove strong consistency of the estimator for the case when δ is fixed, and will conjecture consistency of the estimator when δ slowly decreases to 0 as $n \to \infty$.

The estimator is based on two facts. First, as n becomes large, exactly r eigenvalues of the data's sample covariance matrix $\frac{1}{T}XX'$ will be above the largest eigenvalue of the sample covariance matrix $\frac{1}{T}ee'$ of the idiosyncratic terms. This fact follows from our assumption 2 and the singular value analog of Weyl's eigenvalue inequalities (see formula (6) below). Second, as shown by Bai and Silverstein (1998), the largest eigenvalue of $\frac{1}{T}ee'$ will be almost surely below any number *larger* than u as $n \to \infty$, where u is the upper boundary of the limiting spectral distribution of $\frac{1}{T}ee'$.

The term $\hat{u}$ in the estimator is a strongly consistent estimator of u. Parameter δ plays a role of the markup over the $\hat{u}$, which is needed because the largest eigenvalue of $\frac{1}{T}ee'$ is only guaranteed to be below any number *larger* than u. If δ is fixed, the strong consistency of $\hat{u}$ will imply the strong consistency of $\hat{r}_\delta$. If δ is decreasing with n, the consistency of $\hat{r}_\delta$ will depend on whether the rate of convergence of $\hat{u}$ is fast enough so that $(1+\delta)\hat{u}$ almost surely becomes larger than u as $n \to \infty$.

Our estimator of u exploits the fact, established by Silverstein and Choi (1995), that the limiting spectral distribution of $\frac{1}{T}ee'$ has density $f(x)$ of the form $a\sqrt{u-x}\,(1+o(1))$, where a is some positive constant. Had we observed e, we would have been able to estimate u from the relatively large eigenvalues of $\frac{1}{T}ee'$. Although the spectral distribution of $\frac{1}{T}ee'$ is unobservable, it is well approximated (see proposition 1 below) by the spectral distribution of $\frac{1}{T}XX'$. Therefore, our estimator $\hat{u}$ corresponds to a particular way to fit the density $f(x)$ to the range of the empirical spectral distribution of $\frac{1}{T}XX'$ contained in between $\lambda_{2r_{\max}+1}$ and $\lambda_{r_{\max}+1}$. Such a choice of the range insures that the eigenvalues are in the neighborhood of u, where $f(x)$ is well approximated by $a\sqrt{u-x}$.

Let us denote the j-th largest eigenvalue of $\frac{1}{T}ee'$ as μ_j. Proposition 1 below formally establishes conditions that our estimator builds upon.

Proposition 1. *Under assumptions 1-4, we have:*
i) The spectral distribution of $\frac{1}{T}ee'$ weakly converges to a distribution G with bounded support almost surely.
ii) For any i such that $\frac{i}{n} \to 0$ as $n \to \infty$, the i-th eigenvalue of $\frac{1}{T}ee'$, μ_i, converges almost surely to the upper boundary u of the support of G.
iii) The spectral distribution of $\frac{1}{T}XX'$ weakly converges to G almost surely.
iv) For any $i > r$ and such that $\frac{i}{n} \to 0$ as $n \to \infty$, the i-th eigenvalue of $\frac{1}{T}XX'$, λ_i, converges almost surely to u.

Proof: The proof of the proposition is in the Appendix.

The fact that $\hat{u}$ converges to u almost surely immediately follows from statement iv) of the proposition. Indeed, since by assumption $r_{\max} \sim \min(n^{\alpha}, T^{\alpha})$ caps r and $\alpha < 1$ so that $\frac{r_{\max}}{n} \to 0$, $\lambda_{r_{\max}+1}$ and $\lambda_{2r_{\max}+1}$ are both converging to u almost surely as $n \to \infty$. But $\hat{u}$ is a fixed-weight linear combination of $\lambda_{r_{\max}+1}$ and $\lambda_{2r_{\max}+1}$. Hence, $\hat{u} \to u$ almost surely as $n \to \infty$. We use this fact to prove consistency of $\hat{r}_\delta$ for fixed $\delta > 0$.

Proposition 2. *Under assumptions 1-4, for any fixed $\delta > 0$, $\hat{r}_\delta \to r$ almost surely as $n \to \infty$.*

Proof: Since $\hat{u} \to u$ almost surely as $n \to \infty$, by statement iv) of proposition 1, we have $\lambda_i < (1+\delta)\,\hat{u}$ almost surely for large enough n and $i > r$. Therefore, $\hat{r}_\delta = \#\{i \leq n : \lambda_i > (1+\delta)\hat{u}\} \leq r$ almost surely for large enough n. Below we will prove that $\lambda_r > (1+\delta)\hat{u}$ almost surely for large n and, hence, that $\hat{r}_\delta \to r$.

According to singular value analog of Weyl's eigenvalue inequalities (see theorem 3.3.16 of Horn and Johnson (1991)), for any $n \times T$ matrices A and B, we have:

$$\sigma_{i+j-1}(A+B) \leq \sigma_i(A) + \sigma_j(B), \tag{6}$$

where $1 \leq i, j \leq \min(n, T)$ and $\sigma_i(A)$ denotes the i-th largest singular value of matrix A, which is another name for the square root of the i-th largest eigenvalue of matrix AA'. Substituting $A = \frac{1}{\sqrt{T}}X$ and $B = \frac{-1}{\sqrt{T}}e$ into the inequality and denoting the j-th largest eigenvalue of $\frac{1}{T}\Lambda FF'\Lambda'$ as ν_j, we get:

$$\lambda_r^{\frac{1}{2}} \geq \nu_r^{\frac{1}{2}} - \mu_1^{\frac{1}{2}}.$$

Statement ii) of proposition 1 implies that $\mu_1^{\frac{1}{2}} \to u^{\frac{1}{2}}$ almost surely. Hence, we only need to show that $\nu_r^{\frac{1}{2}} \to \infty$ almost surely. According to the product inequality for singular values (see Theorem 3.3.16 of Horn and Johnson, 1991), for any $n \times r$ and $r \times r$ matrices A and B

$$\sigma_i(AB) \leq \sigma_i(A)\sigma_1(B).$$

for $i \leq \min(n, r)$ (that is for $i \leq r$ for large enough n). Let $A = \Lambda \left(\frac{1}{T}FF'\right)^{\frac{1}{2}}$ and $B = \left(\frac{1}{T}FF'\right)^{-\frac{1}{2}}$, where $\frac{1}{T}FF'$ is invertible by assumption 2. Then, the above inequality implies:

$$\nu_r \geq \frac{\min \operatorname{eval}(\Lambda\Lambda')}{\max \operatorname{eval}\left(\left(\frac{1}{T}FF'\right)^{-1}\right)} = \min \operatorname{eval}(\Lambda\Lambda') \min \operatorname{eval}\left(\left(\frac{1}{T}FF'\right)\right) \to \infty$$

almost surely as $n \to \infty$ by assumption 2.□

Note that the above proof of the strong consistency of our estimator does not rely on the relatively sophisticated form of $\hat{u}$. For example, if we substitute $\hat{u}$ by $\lambda_{r_{\max}+1}$ in (5), we would get a simpler estimator

$$\tilde{r}_\delta = \#\left\{i \leq n : \lambda_i > (1+\delta)\lambda_{r_{\max}+1}\right\},$$

which converges to r almost surely by virtue of proposition 1 and the proof of proposition 2. We use the more sophisticated estimator as a mean to improve the finite sample properties of $\tilde{r}_\delta$. In finite samples, performance of both $\tilde{r}_\delta$ and $\hat{r}_\delta$ will critically depend on the choice of δ. To reduce the underestimation risk, we would like to have δ small. How small δ can be? Clearly, to avoid overestimation risk, δ should be large enough to cover up the gap between $\lambda_{r_{\max}+1}$ and μ_1 in the case of $\tilde{r}_\delta$, and the gap between $\hat{u}$ and μ_1 in the case of $\hat{r}_\delta$. As we conjecture below, the latter gap will be decreasing with n much faster than the former. Therefore, δ can be chosen much smaller for $\hat{r}_\delta$ than for $\tilde{r}_\delta$, making the finite sample properties of $\hat{r}_\delta$ better.

As will be seen shortly, the magnitude of the above mentioned gaps depends on how fast $F^{\frac{1}{T}ee'}$ converges to G and how fast the largest eigenvalue of $\frac{1}{T}ee'$, μ_1, converges to u. At the moment, we will not take stand on these rates of convergence, and will simply assume that

Assumption 5: *$\left\| F^{\frac{1}{T}ee'} - G \right\| \overset{p}{\sim} n^{-\beta_1}$, and $|\mu_1 - u| \overset{p}{\sim} n^{-\beta_2}$, where $0 < \beta_1, \beta_2 \leq 1$*

Later, we will conjecture that $\beta_1 = 1$ and $\beta_2 = \frac{2}{3}$ and will provide arguments in favor of this conjecture.

Let us define

$$g(\alpha, \beta) = \left\{ \begin{array}{ll} \frac{4}{3}(1-\alpha) & \text{if } \frac{5}{3}(1-\alpha) < \beta \leq 1 \\ \beta - \frac{1}{3}(1-\alpha) & \text{if } 1-\alpha < \beta \leq \frac{5}{3}(1-\alpha) \\ \frac{2}{3}\beta & \text{if } 0 < \beta \leq 1-\alpha \end{array} \right\}$$

and

$$h(\alpha, \beta) = \frac{2}{3} \min\{\beta, 1 - \alpha\}$$

We have the following

Proposition 3: *Let assumptions 1-5 hold, then*
i) $\hat{u} - \mu_1 = O_p(n^{-\min\{g(\alpha,\beta_1),\beta_2\}})$,
ii) $\lambda_{r_{\max}+1} - \mu_1 = O_p(n^{-\min\{h(\alpha,\beta_1),\beta_2\}})$

Proof: The proof of the proposition is in the Appendix.

Figure 1 shows the rates of convergence of $\hat{u}-\mu_1$ and $\lambda_{r_{\max}+1}-\mu_1$ as functions of α for fixed β_1, β_2. For illustration purposes, we chose $\beta_1 = \frac{11}{12}$ and $\beta_2 = \frac{2}{3}$ because this combination corresponds to a rich pattern of dependence of the rates of convergence on α. As α increases from 0, both rates stay at $\frac{2}{3}\beta_1$ until $1 - \alpha$ becomes equal to β_1. After that, the convergence rate of $\lambda_{r_{\max}+1} - \mu_1$ starts to drop as $\frac{2}{3}(1-\alpha)$ (dashed line). On the contrary, the convergence rate of $\hat{u} - \mu_1$ start to increase as $\beta_1 - \frac{1}{3}(1-\alpha)$ until $\beta_1 - \frac{1}{3}(1-\alpha) = \beta_2$, then it stays the same until $1 - \alpha$ is equal to $\frac{3}{4}\beta_2$, and only after that, decreases to 0 as $\frac{4}{3}(1-\alpha)$ (solid line). In general, $\hat{u} - \mu_1$ converges to zero no slower, and, for many combinations of α, β_1, and β_2, much faster than $\lambda_{r_{\max}+1} - \mu_1$..

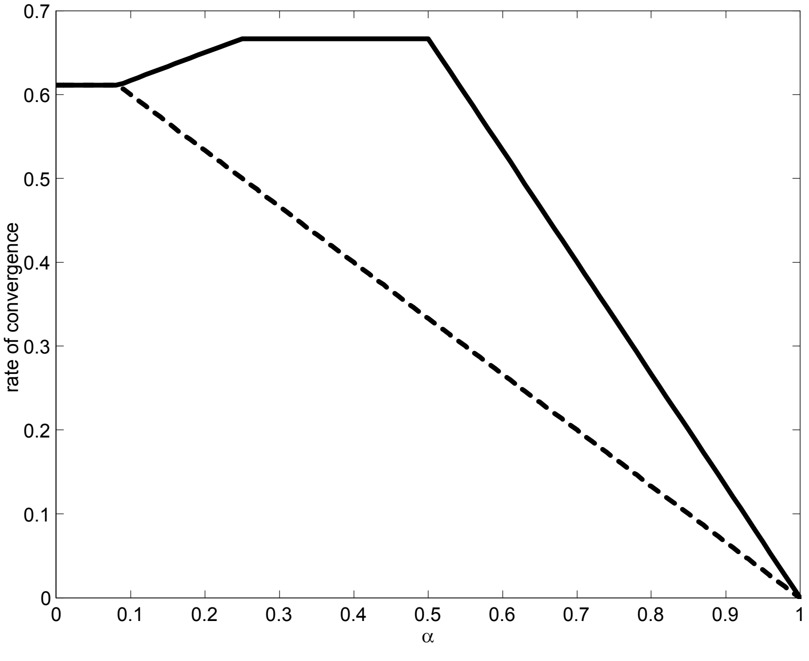

Figure1: The negative of the exponential in the rate of convergence of $\hat{u} - \mu_1$ (solid line) and $\lambda_{r_{\max}+1} - \mu_1$ (dashed line) as functions of α. β_1 is fixed at 11/12, β_2 is fixed at 2/3.

The non-monotonic dependence of the convergence rate of $\hat{u}-\mu_1$ on α can be understood as follows. First, note that $\hat{u} - \mu_1 = (\hat{u} - u) + (u - \mu_1)$. The second term converges to zero as $n^{-\beta_2}$ by assumption 5, hence β_2 is the upper limit on the rate of convergence of $\hat{u}-\mu_1$. Consider now the first term, $\hat{u}-u$. Intuitively, our trying to fit the functional form $1-a\,(u-x)^{\frac{3}{2}}$ to the empirical spectral distribution of $\frac{1}{T}XX'$ to get a better estimate of u than simply $\lambda_{r_{\max}+1}$ is not productive as long as the variation of $F^{\frac{1}{T}XX'}$ in the range $x\in[\lambda_{2r_{\max}+1},\lambda_{r_{\max}+1}]$ (which is decreasing as $\frac{r_{\max}}{n}$, that is with rate $1-\alpha$) is small relative to the error of approximation of G by $F^{\frac{1}{T}XX'}$ (decreasing with the rate β_1). Therefore, as $\beta_1<1-\alpha$, the rate of convergence of $\hat{u}$ to u is equal to the rate of convergence of the "primitive" estimator $\lambda_{r_{\max}+1}$. When α becomes such that $\beta_1>1-\alpha$, the variation in $F^{\frac{1}{T}XX'}$ becomes large enough to exploit the functional form fitting idea and the convergence rate improves. As α becomes too large, the discrepancy between G and $1-a\,(u-x)^{\frac{3}{2}}$ (which is small only in the neighborhood of u) starts to be large and fitting the functional form does not produce good results any more.

Proposition 3 suggests that the optimal choice of α in $r_{\max}\sim\min\left(n^{\alpha},T^{\alpha}\right)$ in the sense of optimizing the rate of convergence of $\hat{u}-\mu_1$ to zero depends on both β_1 and β_2. Precisely, when $\frac{4}{5}\beta_1\leq\beta_2$, the optimal α is equal to $1-\frac{3}{5}\beta_1$; when $\frac{2}{3}\beta_1\leq\beta_2<\frac{4}{5}\beta_1$, any α from the segment $[1-3(\beta_1-\beta_2),1-\frac{3}{4}\beta_2]$ is optimal; finally, when $\beta_2<\frac{2}{3}\beta_1$, any α from the segment $[0,1-\frac{3}{4}\beta_2]$ is optimal.

Unfortunately, the true values of β_1 and β_2 are not known. When ε_t in (2) is a vector of i.i.d. normal variables and S_n is the identity matrix, it is known (Johnstone (2000)) that $\beta_2=\frac{2}{3}$. For β_1, in case when S_n is the identity matrix, the standard conjecture (see Bai 1999, p.658-659) is that it is equal to 1. As Silverstein (1999) points out, this conjecture is substantiated by extensive simulations and some theoretical results. If S_n is not identity, but converges to the limiting distribution H very fast, and if the limiting distribution does not have any peculiarities, such as those eliminated by our assumption 4 iii), one may expect that the rates of convergence should be the same as with S_n equal to the identity matrix. In what follows, we therefore conjecture that $\beta_2=\frac{2}{3}$ and $\beta_1=1$. If the conjecture is correct, then any α from the range $[0,\frac{1}{2}]$ is optimal.

Given proposition 3, it is easy to prove consistency of $\hat{r}_\delta$ for decreasing δ.

Proposition 4: *Let assumptions 1-5 hold, then $\hat{r}_\delta$ is consistent for r when $\delta\sim n^{-\gamma}$, for any γ such that* $0\leq\gamma<\min\left\{g(\alpha,\beta_1),\beta_2\right\}.$

Proof: Recall that for a fixed δ, as was shown in the proof of proposition 2, $\lambda_r>(1+\delta)\hat{u}$ almost surely for large enough n. For δ local to zero, the inequality holds "even stronger". Therefore, to prove the consistency of $\hat{r}_\delta$, we only need to show that the probability that $\lambda_{r+1}<(1+\delta)\,\hat{u}$ goes to 1 for large enough n. By (10), it is enough to prove that the probability that $\mu_1<(1+\delta)\,\hat{u}$ goes to 1 for large enough n. We have:

$$(1+\delta)\,\hat{u}-\mu_1=\delta\hat{u}+(\hat{u}-\mu_1) \tag{7}$$

The second term in the above sum is $O_p\left[n^{-\min\{g(\alpha,\beta_1),\beta_2\}}\right]$ by proposition 3, the first term decays as fast as δ. Therefore, with probability going to 1, the first term will dominate the second one as $n\to\infty$ if $\delta\sim n^{-\gamma}$, for any γ, such that $0 \leq \gamma < \min\{g(\alpha,\beta_1),\beta_2\}$. Hence $\Pr(\mu_1 < (1+\delta)\hat{u}) \to 1$ as $n \to \infty$, which completes the proof.□

If, as the standard conjecture is, $\beta_1 = 1, \beta_2 = \frac{2}{3}$ and, as was suggested above, $\alpha = \frac{2}{5}$, proposition 4 says that δ can be chosen to decrease only marginally slower than $n^{-\frac{2}{3}}$ without hurting consistency of $\hat{r}_\delta$. The Monte Carlo analysis below shows that the choice $\delta = \max\left(n^{-\frac{2}{5}}, T^{-\frac{2}{5}}\right)$ corresponds to very good finite sample performance of our estimator. Such a rate of decay is sufficiently slower than $n^{-\frac{2}{3}}$ for $\delta\hat{u}$ to strongly dominate $\hat{u} - \mu_1$ in (7). Hence, overestimation of the true number of factors is not likely. At the same time, the rate is fast enough for $(1+\delta)\hat{u}$ not to significantly overshoot u and, therefore, the underestimation is unlikely too. Hence,

As was mentioned above, we can relax the assumption of fixed number of factors. In fact, the proof of strong consistency of $\hat{r}_\delta$ when δ is fixed only requires $r \leq r_{\max}$. Hence, for fixed δ, $\hat{r}_\delta$ remains consistent even if the true number of factors is increasing as n^α when $n \to \infty$. It can be shown[2] that if, as the standard conjecture is, $\beta_1 = 1, \beta_2 = \frac{2}{3}$, and $r = O(n^\theta)$ for some $\theta \leq \alpha$, then $\hat{r}_\delta$ remains consistent as long as $\delta \sim n^{-\gamma}$ for any $\gamma < \min\left\{q(\alpha,\theta), \frac{2}{3}\right\}$, where

$$q(\alpha,\theta) = \left\{ \begin{array}{ll} \frac{4}{3}(1-\alpha) & \text{if } 0 \leq \theta < \frac{5}{3}\alpha - \frac{2}{3} \\ 1 - \theta - \frac{1}{3}(1-\alpha) & \text{if } \frac{5}{3}\alpha - \frac{2}{3} \leq \theta \leq \alpha \end{array} \right\}.$$

For example, for our preferred choice of $\alpha = \frac{2}{5}$ and $\delta = \max\left(n^{-2/5}, T^{-2/5}\right)$, $\hat{r}_\delta$ will consistently estimate the number of factors rising only marginally slower than $n^{2/5}$.

In conclusion of this section, let us note that to develop our estimator we used regularity of the limiting spectral distribution *local* to the upper boundary of its support. The local nature of the regularity we exploit is the price we pay for allowing rather rich pattern of the cross-sectional serial correlation and heteroskedasticity in the idiosyncratic term.[3] Had we assumed that the idiosyncratic terms are cross-sectionally i.i.d., the limiting spectral distribution would have been of the so called Marčenko-Pastur form (see Bai, 1999) and we would have been able to use the information from all the eigenvalues to estimate u. Kapetanios (2004) explains how the i.i.d. assumption can be somewhat relaxed so that the limiting distribution is still of the Marčenko-Pastur form and proposes a consistent method of the number of factors estimation based on the implied eigenvalue threshold. However, restrictions that Kapetanios makes on the serial correlation pattern and heteroskedasticity of the idiosyncratic components remain very stringent. The main methodological contribution of this paper relative to Kapetanios (2004) is that we essentially lift those restrictions.

[2] The proof of this fact is available from me upon request.

[3] The information about this serial correlation and heteroskedasticity can be backed out from the observed empirical distribution of the eigenvalues.

4 Monte Carlo Analysis

In this section we use Monte Carlo simulation experiments to study empirical performance of our estimator and compare it to the performance of Bai and Ng (2002) estimator. The setting of the experiments is as in Bai and Ng (2002):

$$X_{it} = \sum_{k=1}^{r} \Lambda_{ik} F_{kt} + \sqrt{\theta} e_{it},$$

where the factor matrix F and the factor loadings matrix Λ are $r \times T$ and $n \times r$ matrices of independent $N(0,1)$ variables respectively and

$$e_{it} = \rho e_{i,t-1} + v_{it} + \sum_{j=-J}^{J} \beta v_{i-j,t},\ J = \min\left\{\frac{N}{20}, 10\right\},\ v_{ij} \sim IIDN(0,1). \quad (8)$$

For their experiments, Bai and Ng choose the true number of factors r equal 1,3, or 5 and consider parameters ρ, β, and θ from the sets $\{0, 0.5\}, \{0, 0.2\}$, and $\{1, 2\}$ respectively. We will consider the same choice of factors, but a wider range of parameters ρ, β, and θ. Being more flexible with respect to the choice of ρ and β allows us to study the effect of changing the degree of dependence in the idiosyncratic terms on the performance of different estimators. Considering a wider range for θ helps us to analyze the quality of estimators when the “signal-to-noise" ratio varies substantially.

We chose to study several combinations of n and T. The combinations $n = 200, T = 60$ and $n = 1000, T = 60$ are meant to represent five years of monthly financial data on 200 and 1000 stock returns. We also consider a combination $n = 1000, T = 250$ because in the stock returns application below we have 21 years of monthly data on 1148 stocks traded on NYSE, AMEX, and NASDAQ. The rest of the combinations that we consider correspond to plausible macroeconomic data sizes. Combinations $n = 150, T = 500$; $n = 150, T = 150$; $n = 100, T = 100$; and $n = 40, T = 100$ roughly correspond to the sizes of monthly Stock and Watson (1999) data, quarterly Stock and Watson (1999) data, and hypothetical data extracted from Summers and Heston (1991) tables.

We start from the case when the idiosyncratic terms are i.i.d $N(0,1)$ variables. Table 1 reports the averages of the Bai-Ng estimators and three versions of our estimator $\hat{r}_\delta$, corresponding to $\delta = 0, \delta = \max\left(n^{-\frac{1}{2}}, T^{-\frac{1}{2}}\right)$, and $\delta = \max\left(n^{-\frac{2}{5}}, T^{-\frac{2}{5}}\right)$, over 1000 replications of the data generating process. The standard errors in the majority of the experiments are low and we do not report them. There are, however, few cases when standard errors are larger than 1 but smaller than 2, and even a handful of cases when standard erros are larger than 2 but smaller than 3. We indicate such situations by marking the corresponding average estimates by one and two asterisks respectively. The true number of factors in the experiments is assumed to be $r = 1, 3$, and 5. We set $r_{\max}$ equal to the smallest integer larger than $1.55 \min\left(T^{2/5}, n^{2/5}\right)$ so that in realistic small samples, when T is equal to 60, the maximum number of factors is 8, a standard choice in the literature.[4] In all the experiments, prior to computation of the eigenvectors, each series is demeaned and standardized to have unit variance.

Table 1: DGP: $X_{it} = \sum_{k=1}^{r} \Lambda_{ik}F_{kt} + \sqrt{\theta}e_{it}$, $\theta = r$, $\rho = \beta = 0$

n	T	r	$r_{\max}$	PC$_{p1}$	PC$_{p2}$	PC$_{p3}$	IC$_{p1}$	IC$_{p2}$	IC$_{p3}$	$\hat{r}_0$	$\hat{r}_{n^{\frac{-1}{2}}}$	$\hat{r}_{n^{\frac{-2}{5}}}$
200	60	1	8	1.00	1.00	1.00	1.00	1.00	1.00	1.39	1.01	1.00
1000	60	1	8	1.00	1.00	1.00	1.00	1.00	1.00	1.18	1.00	1.00
1000	250	1	15	1.00	1.00	1.00	1.00	1.00	1.00	1.40	1.00	1.00
150	500	1	12	1.00	1.00	1.00	1.00	1.00	1.00	1.32	1.00	1.00
150	150	1	12	1.00	1.00	7.99	1.00	1.00	1.12	1.69	1.06	1.00
100	100	1	10	1.02	1.00	9.18	1.00	1.00	3.44**	1.72	1.09	1.02
40	100	1	7	2.57	1.74	5.57	1.00	1.00	1.23	1.42	1.03	1.00
200	60	3	8	3.00	3.00	3.00	3.00	3.00	3.00	3.03	3.00	3.00
1000	60	3	8	3.00	3.00	3.00	3.00	3.00	3.00	3.01	3.00	3.00
1000	250	3	15	3.00	3.00	3.00	3.00	3.00	3.00	3.09	3.00	3.00
150	500	3	12	3.00	3.00	3.00	3.00	3.00	3.00	3.30	3.01	3.00
150	150	3	12	3.00	3.00	7.34	3.00	3.00	3.06	3.24	3.02	3.00
100	100	3	10	3.00	3.00	8.63	3.00	3.00	4.32*	3.22	3.01	3.00
40	100	3	7	3.30	3.05	5.25	3.00	3.00	3.17	3.13	3.01	3.00
200	60	5	8	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00
1000	60	5	8	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00
1000	250	5	15	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00
150	500	5	12	5.00	5.00	5.00	5.00	5.00	5.00	5.10	5.00	5.00
150	150	5	12	5.00	5.00	7.16	5.00	5.00	5.01	5.00	5.00	5.00
100	100	5	10	5.00	5.00	8.49	5.00	5.00	5.79	5.02	5.00	5.00
40	100	5	7	5.00	4.98	5.65	4.87	4.70	5.11	4.98	4.86	4.78

As can be seen from table 1, when $\theta = r$, all estimators, except PC_{p3}, which substantially overestimates the true number of factors when $n = T = 150$, $n = T = 100$ and $n = 40, T = 100$, work very well and their performance is comparable. These results confirm findings reported in tables 1-3 of Bai and Ng (2002). We see that choosing δ equal to zero have a potential to overestimate the true number of factors, which is consistent with the theory. Choosing δ decreasing as $\max\left(n^{-\frac{1}{2}}, T^{-\frac{1}{2}}\right)$ or $\max\left(n^{-\frac{2}{5}}, T^{-\frac{2}{5}}\right)$ corrects the overestimation.

Table 2 increases the variance of the idiosyncratic term relative to the variance of the systematic component. We perform the same simulation experiment as above, except now $\theta = 9r$, that is the standard deviation of the idiosyncratic component is 3 times the standard deviation of the systematic component. Although there is no much change relative to table 1 for $r = 1$, the change is very substantial when the true number of factors equals to 3 or 5. The Bai-Ng

[4] For the data sizes used in the Monte Carlo experiments, such a choice of $r_{\max}$ would produce almost the same bound on the number of factors as $r_{\max}$ equal to the integer closest to $\min\left(T^{1/2}, n^{1/2}\right)$ would do.

estimators start to significantly underestimate the number of factors. The underestimation is more pronounced for IC estimators than for PC estimators, especially for relatively small sample sizes. In contrast, our estimator still works very well for $T > 40$, except for the case $r = 5$ for small sample sizes.

The deterioration in performance of the Bai-Ng estimators in this situation is what we would expect, because the factors now explain much smaller portion of the variance in the data. Estimators based on the model selection principles will therefore have hard time distinguishing the factors from idiosyncratic components. Since our method of estimation relies more on the structural properties of the idiosyncratic process, which do not change when θ is increased, its performance is less sensitive to the increase in the idiosyncratic variance.

From table 2 we also see that the deterioration of performance of the Bai-Ng estimators is larger the larger the true number of factors is. This is an artifact of our Monte Carlo setting which makes the variance of the systematic component equal to the true number of factors. Since the variance of the idiosyncratic terms is assumed to be proportional to the variance of the systematic component, the larger the true number of factors, the larger the idiosyncratic variance is.

Table 2: $\theta = 9r, \rho = \beta = 0$

n	T	r	$r_{\max}$	PC$_{p1}$	PC$_{p2}$	PC$_{p3}$	IC$_{p1}$	IC$_{p2}$	IC$_{p3}$	$\hat{r}_0$	$\hat{r}_{n^{\frac{-1}{2}}}$	$\hat{r}_{n^{\frac{-2}{5}}}$
200	60	1	8	1.00	1.00	1.00	1.00	1.00	1.00	1.30	1.00	1.00
1000	60	1	8	1.00	1.00	1.00	1.00	1.00	1.00	1.17	1.00	1.00
1000	250	1	15	1.00	1.00	1.00	1.00	1.00	1.00	1.28	1.00	1.00
150	500	1	12	1.00	1.00	1.00	1.00	1.00	1.00	1.32	1.00	1.00
150	150	1	12	1.00	1.00	2.49	1.00	1.00	1.00	1.53	1.01	1.00
100	100	1	10	1.00	1.00	5.22	1.00	1.00	1.00	1.47	1.03	1.00
40	100	1	7	1.00	1.00	1.96	1.00	0.99	1.00	1.36	1.01	1.00
200	60	3	8	1.45	1.23	2.32	1.00	1.00	1.16	2.98	2.75	2.54
1000	60	3	8	1.36	1.30	1.53	1.00	1.00	1.01	3.02	3.00	3.00
1000	250	3	15	3.00	3.00	3.00	3.00	3.00	3.00	3.04	3.00	3.00
150	500	3	12	2.90	2.77	3.00	2.17	1.80	2.98	3.04	3.00	3.00
150	150	3	12	2.41	1.72	3.55	1.12	1.01	2.99	3.07	3.00	3.00
100	100	3	10	1.94	1.28	5.68	1.01	1.00	2.87	3.04	2.85	2.71
40	100	3	7	1.54	1.21	2.81	0.95	0.85	1.11	2.28	1.58	1.36
200	60	5	8	1.02	1.00	1.67	1.00	0.99	1.00	3.28	2.24	1.81
1000	60	5	8	1.00	1.00	1.01	1.00	1.00	1.00	4.97	4.56	4.15
1000	250	5	15	4.04	3.62	4.97	2.51	1.96	4.65	5.00	5.00	5.00
150	500	5	12	1.67	1.30	3.41	1.01	1.00	1.70	5.00	5.00	5.00
150	150	5	12	1.31	1.01	5.08	1.00	1.00	3.77	4.78	4.32	3.94
100	100	5	10	1.20	1.00	6.21	1.00	0.99	2.73	3.39	2.57	2.17
40	100	5	7	1.15	1.03	2.78	0.92	0.74	1.01	1.99	1.23	1.08

To explore in more detail the differences in workings of the Bai-Ng estimators and our estimators when the “signal to noise” ratio varies, we perform the following experiment. We set the true number of factors at 3 and vary $\frac{\theta}{r}$ on a grid 0.5:0.5:25. Figure 2 reports the average (across 1000 Monte Carlo replications) estimates of the number of factors produced by PC_{p1}, IC_{p1} and $\hat{r}_{n^{-\frac{2}{5}}}$ for $\frac{\theta}{r} \in [0.5, 25]$ and $n = 150, T = 100$. We see that our estimator is relatively insensitive to the increase in the size of noise relative to the size of the systematic component. The average estimate of the number of factors remains above 2.5 (and thus closer to the true number of factors than to any other integer) until the variance of noise becomes 13 times larger than the the variance of the systematic component. In contrast, the average estimate produced by IC_{p1} goes below 2.5 as soon as the noise-signal variance ratio is 5. PC_{p1} works somewhat better. The average estimate drops below 2.5 when the noise-signal ratio is 7.5.

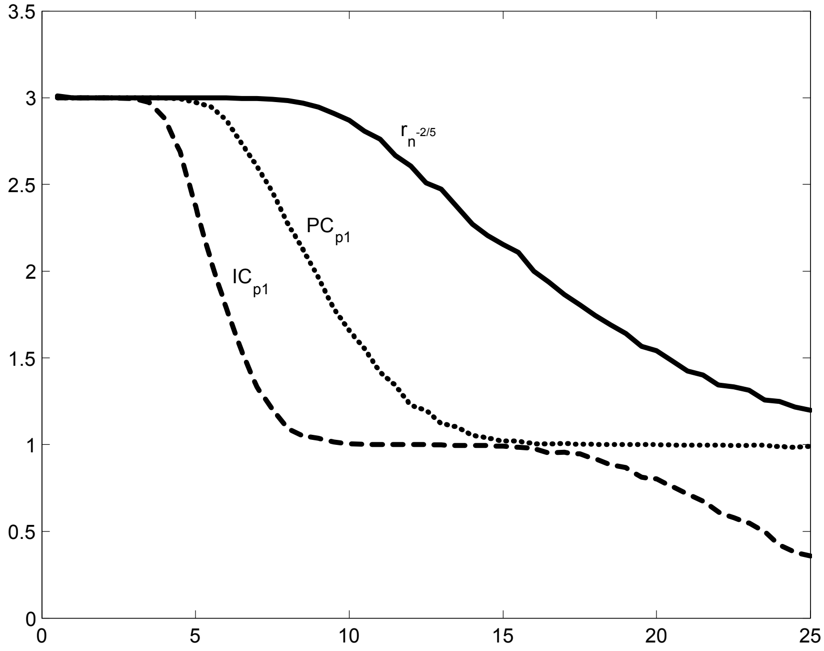

Figure 2: Average estimated number of factors according to PC_{p1}, IC_{p1}, and $\hat{r}_{n^{-2/5}}$. Horizontal axis: θ/r.

Interestingly, we see that IC_{p1} estimator temporarily stabilizes at the estimate of 1 factor, when θ/r grow. This is related to the "single factor illusion" phenomenon described by Brown (1989), who points out that in an economy with r equally important factors the largest eigenvalue of the covariance matrix will grow much faster than the other $r - 1$ eigenvalues.

Our next step is to introduce cross-sectional and time-series serial correlation to the idiosyncratic terms. We first consider the case of the time-series dependence only: $\rho = 0.5$ and $\beta = 0$ (table 3). Then, we add the cross-section dimension of the dependence: $\rho = 0.5$ and $\beta = 0.2$ (table 4). In the time-series-dependence-only case, we see that all PC estimators start to overestimate the true number of factors. When the cross-sectional dimension of the dependence is added, the amount of the overestimation dramatically increases. Now, all the Bai-Ng estimators work very poorly. In contrast, our estimators still work well except when the true number of factors is equal to 1. Even for the case $r = 1$, the deterioration in performance of our estimators relative to that of the Bai-Ng estimators is minor.

Table 3: $\theta = r, \rho = 0.5, \beta = 0$

n	T	r	$r_{\max}$	PC_{p1}	PC_{p2}	PC_{p3}	IC_{p1}	IC_{p2}	IC_{p3}	$\hat{r}_0$	$\hat{r}_{n^{\frac{-1}{2}}}$	$\hat{r}_{n^{\frac{-2}{5}}}$
200	60	1	8	4.79	3.91	7.45	1.01	1.00	3.78**	1.69	1.10	1.02
1000	60	1	8	4.36	4.07	5.16	1.00	1.00	1.00	1.02	1.00	1.00
1000	250	1	15	1.00	1.00	1.00	1.00	1.00	1.00	1.24	1.00	1.00
150	500	1	12	1.00	1.00	1.03	1.00	1.00	1.00	1.68	1.04	1.00
150	150	1	12	3.02	1.24	12.00	1.00	1.00	12.00	2.40	1.52	1.20
100	100	1	10	4.66	2.59	10.00	1.00	1.00	10.00	2.46	1.58	1.28
40	100	1	7	4.32	3.38	6.77	1.02	1.00	4.20**	1.87	1.17	1.07
200	60	3	8	5.25	4.50	7.69	3.01	3.00	5.90*	3.07	3.01	3.00
1000	60	3	8	4.66	4.40	5.44	3.00	3.00	3.00	3.00	3.00	3.00
1000	250	3	15	3.00	3.00	3.00	3.00	3.00	3.00	3.03	3.00	3.00
150	500	3	12	3.00	3.00	3.00	3.00	3.00	3.00	3.21	3.01	3.00
150	150	3	12	3.68	3.01	12.00	3.00	3.00	12.00	3.41	3.06	3.01
100	100	3	10	5.04	3.43	10.00	3.00	3.00	10.00	3.42	3.09	3.03
40	100	3	7	4.53	3.80	6.65	3.01	2.99	5.19*	3.16	3.01	3.00
200	60	5	8	5.90	5.37	7.86	5.02	4.98	7.31	4.98	4.86	4.75
1000	60	5	8	5.26	5.16	5.80	5.00	5.00	5.00	5.00	4.97	4.94
1000	250	5	15	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00
150	500	5	12	5.00	5.00	5.00	5.00	5.00	5.00	5.01	5.00	5.00
150	150	5	12	5.08	5.00	12.00	5.00	5.00	12.00	5.04	5.00	5.00
100	100	5	10	5.81	5.03	10.00	4.99	4.96	10.00	5.01	5.00	5.00
40	100	5	7	5.26	4.98	6.82	4.64	4.13	6.35	4.53	4.07	3.83

Table 4: $\theta = r, \rho = 0.5, \beta = 0.2$

n	T	r	$r_{\max}$	PC$_{p1}$	PC$_{p2}$	PC$_{p3}$	IC$_{p1}$	IC$_{p2}$	IC$_{p3}$	$\hat{r}_0$	$\hat{r}_{n^{\frac{-1}{2}}}$	$\hat{r}_{n^{\frac{-2}{5}}}$
200	60	1	8	8.00	8.00	8.00	8.00	8.00	8.00	3.53	2.81	2.49
1000	60	1	8	6.92	6.81	7.51	2.90*	2.48*	4.98**	1.96	1.26	1.10
1000	250	1	15	9.89	8.54	15.00	1.41	1.06	14.98	1.92	1.21	1.04
150	500	1	12	12.00	12.00	12.00	12.00	12.00	12.00	4.88*	3.89*	3.28*
150	150	1	12	11.99	11.47	12.00	11.98	10.78*	12.00	5.88*	5.11*	4.61*
100	100	1	10	9.96	9.10	10.00	9.84	6.60*	10.00	4.64	3.91	3.50
40	100	1	7	6.98	6.24	7.00	5.34*	2.34*	7.00	2.53	1.71	1.46
200	60	3	8	8.00	8.00	8.00	8.00	8.00	8.00	3.73	3.14	2.89
1000	60	3	8	7.35	7.26	7.82	5.29*	4.83*	6.91*	3.13	3.00	2.99
1000	250	3	15	10.37	9.08	15.00	3.44	3.07	15.00	3.25	3.02	3.00
150	500	3	12	12.00	12.00	12.00	12.00	12.00	12.00	4.07	3.39	3.18
150	150	3	12	12.00	11.86	12.00	12.00	11.61	12.00	5.47	4.77	4.35
100	100	3	10	10.00	9.46	10.00	9.96	8.06*	10.00	4.87	4.27	3.96
40	100	3	7	6.99	6.34	7.00	6.04*	4.18*	7.00	3.27	2.96	2.88
200	60	5	8	8.00	8.00	8.00	8.00	8.00	8.00	2.77	2.02	1.73
1000	60	5	8	7.62	7.63	7.96	6.90*	6.53*	7.69	3.83	3.09	2.74
1000	250	5	15	10.90	9.66	15.00	5.44	5.07	15.00	5.03	5.00	5.00
150	500	5	12	12.00	12.00	12.00	12.00	12.00	12.00	4.96	4.82	4.65
150	150	5	12	12.00	12.00	12.00	12.00	11.97	12.00	5.29	4.98	4.84
100	100	5	10	10.00	9.76	10.00	10.00	8.92*	10.00	4.75	4.16	3.83
40	100	5	7	7.00	6.49	7.00	6.43	4.99*	7.00	3.52	2.84	2.61

The observed overestimation happens because when idiosyncratic terms become dependent, some linear combinations of the idiosyncratic terms start having a non-trivial effect on a sizable portion of the data. Hence, the explanatory power of such linear combination rises and the model selection based estimators have difficulty distinguishing these combinations from the factors.

We explore the relationship between the quality of the estimators and the amount of the dependence in the idiosyncratic terms in more detail by performing the following experiment. We set the true number of factors at 3 and vary ρ on a grid 0:0.1:0.9 and β on a grid 0:0.05:0.3. Figure 3 reports the average (across 1000 Monte Carlo replications) estimates of the number of factors produced by PC_{p1}, IC_{p1} and $\hat{r}_{n^{-\frac{2}{5}}}$ for ρ and β on the grid. The corresponding combination[5] of n and T is $n = 150, T = 150$.

[5] We also considered combinations $n = 200, T = 60$ and $n = 40, T = 100$. The results for these combinations were similar to those reported.

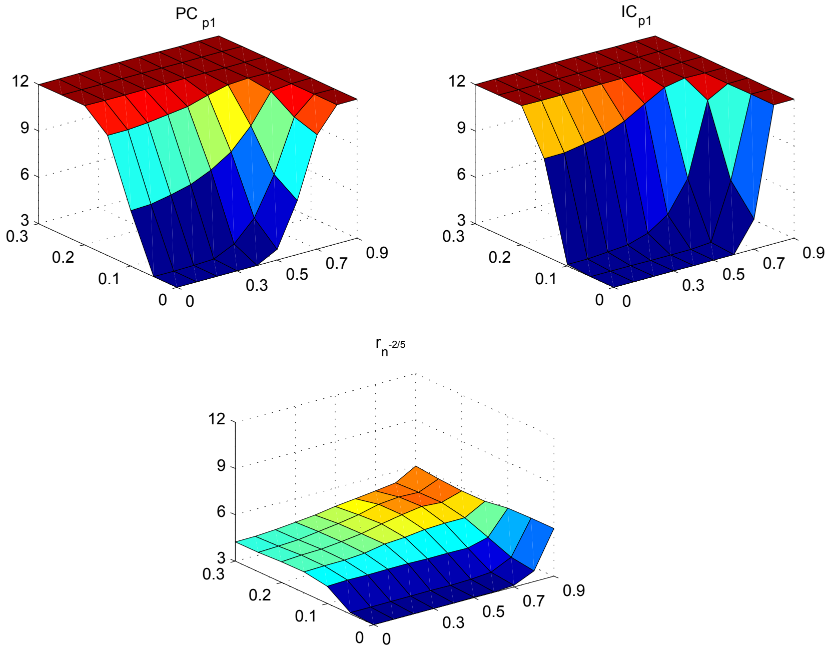

Figure 3: Estimates of the number of factors produced by PC_{p1}, IC_{p1}, and $\hat{r}_{n^{-2/5}}$. Left-hand axis: β. Right hand axis: ρ.

We see that our estimator dramatically outperforms both PC_{p1} and IC_{p1} when the amount of dependence in the idiosyncratic terms increases. As ρ and β rise, PC_{p1} and IC_{p1} quickly start to overestimate the true number of factors. Deterioration of the performance of these estimators is particularly striking in β direction. PC_{p1} starts to report twice the true number of factors as soon as β becomes equal to 0.1. IC_{p1} overestimates the true number of factors by 100% when β becomes equal to 0.125. In contrast, the maximum average estimate $\hat{r}_{n=2/5}$ on the whole range of β that we explore is only 45% larger than the true number of factors. Moreover, the deterioration reaches 100% only when, in addition to the cross-sectional dependence, a very strong time-series dependence ($\rho = 0.9$) is introduced.

The observation that the performance of the Bai-Ng estimators deteriorates in β direction faster than in ρ direction is easy to understand from the theoretical point of view. Note that the Bai-Ng information criteria would have difficulty in distinguishing idiosyncratic components from factors when the eigenvalues of the covariance matrix of the idiosyncratic terms are relatively large. It can be shown that the largest eigenvalue of the covariance matrix of the idiosyncratic terms with cross-sectional correlation only (of the form considered in our Monte Carlo experiments) rises as $1+2J\beta$. On the other hand, the largest eigenvalue of the covariance matrix of the idiosyncratic terms having time series correlation only rises as $\frac{1}{1-\rho}$. For the data size that we consider in the experiment, $J = 7$. Therefore, the largest eigenvalue in the case of the cross-sectional dependence is equal to the largest eigenvalue corresponding to the time-series dependence when $\beta = \frac{\rho}{14(1-\rho)}$. Hence, we may expect that the performance deterioration of the information criteria will be about the same for the cases $\rho = 0.8, \beta = 0$ and $\rho = 0, \beta = \frac{2}{7}$. Such a situation is close to what we observe in the experiment.

Note that the performance of our estimator is not monotonically related to the amount of dependence in the idiosyncratic terms. It is because the size of the largest eigenvalue of the covariance matrix of the idiosyncratic terms is not directly related to the quality of our estimator. Its quality depends mostly on how well the density of the limiting spectral empirical distribution of the "idiosyncratic" covariance matrix is approximated by the square root functional form for a given sample size. Intuitively, this will depend on the entire shape of the limiting spectral distribution of the idiosyncratic covariance matrix (distribution H in the terminology of section 2).

Interestingly, and importantly, even though the consistency of our estimator was conjectured only for the case of the idiosyncratic terms independent across time (or, because of the symmetry of the problem, dependent across time but independent cross-sectionally), the Monte Carlo experiments show that the estimator works relatively well in cases when both cross-sectional and time-series dependence are present. We comment on this observation in the conclusion section below.

5 Applications

We apply the new estimation procedure to determine the number of pervasive factors driving stock returns and the number of pervasive factors influencing dynamics of a large set of macroeconomic variables. First, we estimate the number of factors in the approximate factor model of the stock returns. Chamberlain and Rothchild (1983) show that if the data can be described by such a model, the mean returns on different stocks are approximately linear functions of the factor loadings. The factors in the approximate factor model are defined to have pervasive effect, which means that the sum of squared loadings for a given factor, the sum being taken over all stocks in the sample, increases without bound when the size of the sample rises. However, the rate of this increase can be slow. In particular, it is possible that the average variance of the data explained by the factors is small. In such a circumstance, as was shown above, our estimator works better than Bai and Ng's (2002) estimators. We, therefore, hope to improve upon the estimate $r = 2$, reported in their paper.

Our data consists of monthly returns on 1148 stocks traded on the NYSE, AMEX, and NASDAQ during the period from January 1983 to December 2003. Hence, the time dimension of our data is T=252 and $r_{\max} = 15$. We obtained the data from CRSP data base. Our criterion for inclusion of a stock in the data set was that the stock was traded during the whole sample period.

Our estimators corresponding to the three different choices of δ investigated in the previous section, all estimate the number of pervasive factors to be 8. The $PC_{p1}, PC_{p2}, IC_{p1}, IC_{p2}$ estimators that Bai and Ng (2002) describe as their preferred ones, estimate the number to be $6, 5, 4$, and 3 respectively. These differ from the estimate $r = 2$, obtained by Bai and Ng (2002) for their data set. Perhaps, the difference is due to our including much more time periods (252 vs. 60) in our sample.

Connor and Korajcyk (1993) find evidence for between one and six pervasive factors in the stock returns. Trzcinka (1986) finds some support to the existence of 5 pervasive factors. Five seems also to be a preferred number for Roll and Ross (1980) and Reinganum (1981). A study by Brown and Weinstein (1983) also suggested that the number of factors is unlikely to be greater than 5.[6] Huang and Jo (1995) identify only 2 common factors. The common feature of all these studies, is that they try to find the number of common components that significantly help explaining variation in the data. Therefore, the estimation procedures that these studies use may work poorly in the situations when the factors' explanatory power is relatively weak. On the contrary, our estimation procedure exploits Law-of-Large-Numbers type regularity for the idiosyncratic terms to determine the upper limit on variation that can be attributed to the idiosyncratic terms. Components that explain just a little more variation are classified as the pervasive factors. Hence, we can expect our approach reveal "less pervasive" or "weaker" factors that can be difficult to detect using the other approaches.

Our second application of the newly developed estimation method concerns determining the number of pervasive factors influencing dynamics of a large set of macroeconomic variables. The pervasive factors can be viewed as corresponding to the basic macroeconomic shocks driving the economy. Existence of such shocks is in the spirit of modern dynamic stochastic general equilibrium macroeconomic models.

The dataset we use is the same as in Watson (2003). It includes 215 monthly time series for the United States from 1959:1 to 1998:12. The series represent 14 main categories of macroeconomic time series: real output and income; employment and hours; real retail, manufacturing, and trade sales; real inventories and inventory-sales ratios; orders and unfilled orders; stock prices; exchange rates; interest rates; money and credit quantity aggregates; price indexes; average hourly earnings; and miscellaneous. The data were downloaded from Mark Watson's web site. The variables in the dataset were transformed, standardized and screened for outliers as described in Stock and Watson (2002). The determination of the number of factors was based on the data subset of the transformed and screened 148 variables available for the full sample period.

We set the upper bound on the true number of factors, $r_{\max}$, at 12. Our estimators corresponding to the non-zero choices of δ investigated in the previous section, estimate the number of pervasive factors to be 6. The estimator corresponding to $\delta = 0$ finds 7 pervasive factors. According to the Monte Carlo analysis, the latter estimator tends to overestimate the true number of factors, so we settle at 6 pervasive factors driving the macroeconomic variables in the

[6] Dhrymes, Friend, and Gultekin (1984) find that the estimated number of factors grows with the sample size. However, their setting was the classical factor model as opposed to the approximate factor model.

data set. All estimators proposed by Bai and Ng (2002) estimate the number of factors to be 12, the maximally possible number of factors. Interestingly, Stock and Watson (1999) find that the first six factors account for 39% of the variance in the full data set, and the first 12 factors together account for 53% of the variance. This suggests that the idiosyncratic noise component is very volatile for the macroeconomic time series considered, which may negatively affect performance of the Bai-Ng estimators. Perhaps more importantly, we can expect a relatively strong dependence among the idiosyncratic components of the macroeconomic panel. As our Monte Carlo analysis shows, this is a situation when the Bai-Ng estimators are likely to overestimate the true number of factors.

6 Conclusion

In this paper we develop a new consistent estimator for the number of factors in the approximate factor models. The main advantage of our estimator relative to the previously proposed ones is that it works well in realistically small samples when the amount of cross-sectional and time-series correlation in the idiosyncratic terms is relatively large. It also improves upon the existing methods when the portion of the observed variance attributed to the factors is small relative to the variance due to the idiosyncratic term. These advantages arise because the estimator is based on a Law-of-Large-Numbers type regularity for the idiosyncratic components of the data, as opposed to the estimators based on the assumption that a significant portion of the variance is explained by the systematic part. In contrast to the majority of the previous studies, we do not require the eigenvalues of the covariance matrix of the systematic part of the data to rise proportionately to the sample size, and hence do not rely on the factors explaining substantial portion of variation in the data.

Monte Carlo simulations show that our estimator indeed works better than the information criteria estimators proposed by Bai and Ng (2002) when the variance of the idiosyncratic component of the data is large relative to the variance of the systematic component and/or when the amount of dependence in the idiosyncratic terms is relatively large. This finding is robust across several empirically relevant sample size situations and different patterns of serial correlation in the idiosyncratic term. The better workings of our estimator does not come at the expense of the more complicated structure. The proposed estimator is a simple function of the eigenvalues of the sample covariance matrix and it is very easy to compute.

Our appeal to the Law-of-Large-Numbers type regularity for the idiosyncratic terms is based on a restrictive assumption about these terms. Precisely, we assume that the vector of the idiosyncratic components at a particular point in time is a relatively general linear transformation of an i.i.d. vector of the same size. The idiosyncratic components are assumed to be independent across time. Our Monte Carlo analysis suggests, however, that the latter assumption is not essential for the good performance of the estimator.

In the future work, we plan to relax the assumption of the independence across time. One way to do this is to represent the matrix of idiosyncratic components e as a sum of two matrices:

$$e = Z + \varepsilon,$$

where matrix ε would consist of the cross-sectionally and time-series independent terms, possibly representing measurement errors, and Z would be a matrix of the cross-sectionally and time-series dependent components. Silverstein and Dozier (2004) showed that, as long as the spectral distribution of $\frac{1}{T}ZZ'$ converges to a probability distribution with bounded support, the limiting spectral distribution of $\frac{1}{T}ee'$ will have bounded support and will have the square-root type density near the upper boundary of the support, which is the key condition for the applicability of our estimator. As shown by Hachem et al. (2005), if Z is a stationary Gaussian random field, so that

$$Z_{ij} = \sum_{k,s} h(i-k, j-s)\xi(k,s),$$

where $\xi(k,s)$ are i.i.d. normal random variables and $\sum_{i,j} |h(i,j)| < \infty$, the spectral distribution of $\frac{1}{T}ZZ'$ will converge. Whether the idiosyncratic components arising in macroeconomic or financial applications can be usefully modeled by a stationary Gaussian random field is an open question.

The Large Dimensional Random Matrix theory is a terrain relatively unknown by econometricians. It is likely that many existing findings in this area can be put to an immediate use by the profession. Recently, some second order results were obtained for the largest eigenvalues of large random matrices (see Johnstone, 2000). We conjecture that the results may be relevant for designing statistical tests for the number of factors in the approximate factor models.

7 Appendix

Proof of Proposition 1: Theorem 1.1 of Silverstein (1995) implies[7] that the spectral distribution of $\frac{1}{T}ee'$ weakly converges to a distribution G as $n \to \infty$. That G must have bounded support can be established using Horn's inequality relating singular values of two matrices with singular values of their product (see theorem 3.3.4 of Horn and Johnson, 1991). The inequality implies that the largest eigenvalue of $\frac{1}{T}ee'$ is smaller or equal to the product of the largest eigenvalues of S_nS_n' and $\frac{1}{T}\sum_{t=1}^{T}\varepsilon_t\varepsilon_t'$. By assumption 4 ii), the largest eigenvalue of S_nS_n' is bounded almost surely. As to the largest eigenvalue of $\frac{1}{T}\sum_{t=1}^{T}\varepsilon_t\varepsilon_t'$, Bai, Silverstein and Yin (1988) showed that, under assumption 3, it converges

[7] Condition a) of that theorem is implied by our assumption 3, condition b) is our assumption 1, condition c) is implied by our assumption 4 i), and condition d) is guaranteed by our assumption 3.

to $(1+\sqrt{c})^2$ almost surely. Hence, the largest eigenvalue of $\frac{1}{T}ee'$ is bounded almost surely and therefore, G should not have positive mass above the bound.

Turning to the proof of ii), let j be such that $\frac{j}{n} \to 0$ as $n \to \infty$. We will show that for any $\delta > 0$, $u - \delta < \mu_j < u + \delta$ almost surely as n becomes large. The rightmost inequality is an immediate consequence of theorem 1.1 of Bai and Silverstein (1998).[8] As to the other inequality, suppose it does not hold. Then with positive probability, for any N there exists $n > N$ such that $\mu_j \leq u - \delta$. Let $x_0 \in (u-\delta, u)$ be a point of continuity of G. By definition of u, we must have $G(x_0) < 1$. Now, choose N so large that for any $n > N$, $F^{\frac{1}{T}ee'}(\mu_j) \equiv 1 - \frac{j-1}{n}$ is larger than $\frac{1+G(x_0)}{2}$. Since, by statement i) of the proposition, $F^{\frac{1}{T}ee'} \to G$ almost surely, we must have:

$$\left| F^{\frac{1}{T}ee'}(x_0) - G(x_0) \right| \to 0 \tag{9}$$

as $n \to \infty$ almost surely. However, by our assumption, with positive probability, there exist however large n, such that

$$F^{\frac{1}{T}ee'}(x_0) \geq F^{\frac{1}{T}ee'}(u-\delta) \geq F^{\frac{1}{T}ee'}(\mu_j) > \frac{1+G(x_0)}{2}$$

which contradicts (9).

To prove iii) we will use the rank inequality (see Bai 1999, Lemma 2.6) saying that for any two $n \times T$ matrices A and B,

$$\left\| F^{AA'} - F^{BB'} \right\| \leq \frac{1}{n} \operatorname{rank}(A - B),$$

where $\|\cdot\|$ denotes a standard supremum distance between two functions. Taking $A = \frac{1}{\sqrt{T}}X$, $B = \frac{1}{\sqrt{T}}e$ and using the rank inequality we have:

$$\left\| F^{\frac{1}{T}XX'} - F^{\frac{1}{T}ee'} \right\| \leq \frac{1}{n} \operatorname{rank}(\Lambda F) = \frac{r}{n} \to 0$$

and hence, $F^{\frac{1}{T}XX'}$ must converge to the same limit as $F^{\frac{1}{T}ee'}$.

Finally, let us now denote the j-th largest eigenvalue of $\frac{1}{T}\Lambda F F' \Lambda'$ as ν_j. Inequality (6) implies

$$\begin{aligned} \lambda_i^{1/2} &\leq \mu_{i-r}^{1/2} + \nu_{r+1}^{1/2}, \quad i = r+1, ..., n \\ \lambda_i^{1/2} &\geq \mu_{i+r}^{1/2} - \nu_{r+1}^{1/2}, \quad i = 1, ..., n-r \end{aligned}$$

where the first inequality follows by taking $A = \frac{1}{\sqrt{T}}e$ and $B = \frac{1}{\sqrt{T}}\Lambda F$ and the second inequality follows by taking $A = \frac{1}{\sqrt{T}}X$ and $B = \frac{-1}{\sqrt{T}}\Lambda F$.

Since the rank of $\frac{1}{T}\Lambda F F' \Lambda'$ is equal to r, $\nu_{r+1}^{1/2}$ must be equal to zero so that we have:

[8] Conditions a) and e) of their theorem are satisfied by our assumption 3, condition b) is equivalent to assumption 1, conditions c) and d) follow from assumption 4 i), condition f) follows from assumption 4 ii).

$$\begin{aligned} \lambda_i &\leq \mu_{i-r}, \text{ for } i = r+1, ..., n & (10) \\ \lambda_i &\geq \mu_{i+r}, \text{ for } i = 1, ..., n-r & (11) \end{aligned}$$

Therefore, if i is such that $i > r$ and $\frac{i}{n} \to 0$, λ_i is sandwiched by two terms each of which almost surely lies inside interval $(u-\delta, u+\delta)$ for large enough n. Hence, $\lambda_i \to u$ almost surely, which completes the proof of the proposition.□

To prove proposition 3, we will need the following lemma:

Lemma 1: *Under assumptions 1,3 and 4, there exists a constant $a > 0$, such that*

$$G(x) = 1 - a\,(u-x)^{\frac{3}{2}}\,(1 + O(u-x))$$

as $x \uparrow u$.

Proof: First, note that, since the spectra of $\frac{1}{T}ee'$ and $\frac{1}{T}e'e$ differ only by $|n-T|$ zero eigenvalues, the distribution G is related to the limiting distribution of $F^{\frac{1}{T}e'e}$, which we denote as P, by equation

$$P = (1-c)I_{[0,\infty)} + cG.$$

In particular, P and G have the same upper boundaries of their supports, and their densities (where they exist) are proportional. Therefore, it is enough to establish lemma 2 for P. For G, it will follow from the above equality.

Silverstein (1995) established the fact that, under assumptions equivalent to our assumption 1,3, and 4 i), $F^{\frac{1}{T}e'e}$ converges to a limiting distribution P, whose Stieltjes transform m, defined as

$$m_P(z) \equiv \int \frac{1}{\lambda - z} dP(\lambda), \qquad z \in C^+ \equiv \{z \in C : \operatorname{Im} z > 0\},$$

is the unique solution in C^+ to

$$m = -\left(z - c\int \frac{\tau dH(\tau)}{1+\tau m}\right)^{-1}. \tag{12}$$

Silverstein and Choi (1995) study properties of distributions with the Stieltjes transforms satisfying the above equation. They show that P has continuous density $p(x)$, which has form (see formula (5.3) of Silverstein and Choi (1995)):

$$p(x) = a\,(u-x)^{\frac{1}{2}}\,(1 + o(1)) \tag{13}$$

in the neighborhood of u, the upper boundary of P's support. We would like to strengthen this formula by establishing that

$$p(x) = a\,(u-x)^{\frac{1}{2}}\,(1 + O\,(u-x))\,.$$

Silverstein and Choi prove (13) under the assumption that $-m_u^{-1}$, where m_u is defined as $\lim_{z \in C^+ \to u} m_P(z)$, is strictly larger than the upper boundary of support of H (see the discussion at p.307 of their paper). They point out that this assumption would not hold only if $-m_u^{-1}$ is the upper boundary of H's support and, in addition, $\lim_{m\downarrow m_u}\displaystyle\int\frac{\lambda^2 dH(\lambda)}{(1+\lambda m)^2}$ exists, and $\frac{1}{m^2}-c\displaystyle\int\frac{\lambda^2 dH(\lambda)}{(1+\lambda m)^2}$ is positive on $(m_u, m_u+\delta]$ for some $\delta>0$. It is straightforward to verify that our assumption 4 iii) rules out such a possibility.

To prove (13), Silverstein and Choi, first, show (their theorem 1.1) that the limit $\lim_{z\in C^+\to x} m_P(z)\equiv m_1(x)+im_2(x)$ (here i denotes the imaginary unit) exists, that $p(x)=\frac{1}{\pi}m_2(x)$, and that $m_1(x)$ and $m_2(x)$ are analytic in the neighborhood of any x such that $m_2(x)>0$. Moreover, for these x, $m_1(x)$ and $m_2(x)$ constitute the unique solution (subject to the requirement $m_2(x)>0$) of the system:

$$x = c\int\frac{\lambda dH(\lambda)}{(1+\lambda m_1)^2+\lambda^2 m_2^2} \tag{14}$$

$$0 = \frac{1}{m_1^2+m_2^2}-c\int\frac{\lambda^2 dH(\lambda)}{(1+\lambda m_1)^2+\lambda^2 m_2^2}. \tag{15}$$

Implicitly differentiating the above two equations with respect to x, Silverstein and Choi find that

$$m_2 m_2' = \frac{m_1 A_2+(m_1^2-m_2^2)A_3}{(A_2+A_3 m_1)^2+A_3^2 m_2^2}$$

for $x\in(u-\varepsilon,u)$ for some $\varepsilon>0$, where $A_j = 2c\int\frac{\lambda^j dH(\lambda)}{\left((1+\lambda m_1)^2+\lambda^2 m_2^2\right)^2}$. Using this formula, they show that $2m_2(x)m_2'(x)$ tends to a finite negative number when $x\uparrow u$. Formula (13) then follows from a simple observation that $2m_2(x)m_2'(x)=\frac{d}{dx}m_2^2(x)$ and the fact (following from the continuity of $m_2(x)$) that $m_2(u)=0$.

We now show that not only $2m_2(x)m_2'(x)$ tends to a finite negative number when $x\uparrow u$, but also the derivative of this function is bounded on $x\in(u-\varepsilon,u)$. This is equivalent to saying that $\left(m_2^2(x)\right)'$ is well approximated by a linear function with finite slope on $x\in(u-\varepsilon,u)$, which in turn is equivalent to the statement of our lemma.

Let us first show that $m_1'(x)$, $A_2'(x)$ and $A_3'(x)$ are bounded on $x\in(u-\varepsilon,u)$ for some $\varepsilon>0$. Indeed, differentiating (14) implicitly with respect to x and rearranging, we get

$$m_1' = \frac{-1-A_3 m_2 m_2'}{A_2+A_3 m_1}.$$

It is easy to see that the denominator $A_2+A_3m_1 = 2c\int\frac{\lambda^2(1+\lambda m_1)dH(\lambda)}{\left((1+\lambda m_1)^2+\lambda^2 m_2^2\right)^2}$ is a continuous function of x. Moreover, since by assumption 4 iii) $m_1(u)=m_u$ lies outside the support of H, the denominator is not equal to zero for $x=u$. Let us choose ε so small that it stays away from zero for $x\in(u-\varepsilon,u)$. Then, since as shown by Silverstein and Choi m_2m_2' is bounded on $x\in(u-\varepsilon,u)$, m_1' must be bounded on $x\in(u-\varepsilon,u)$.

For A_2 and A_3 we have

$$A_j' = -4c \int \frac{\lambda^j \left(2\lambda m_1' + 2\lambda^2 \left(m_1 m_1' + m_2 m_2'\right)\right) dH(\lambda)}{\left((1+\lambda m_1)^2 + \lambda^2 m_2^2\right)^3}$$

which is bounded on $x \in (u-\varepsilon, u)$ because $m_2 m_2'$ and m_1' are bounded.

Finally,

$$(m_2 m_2')' = \frac{x}{\left[(A_2 + A_3 m_1)^2 + A_3^2 m_2^2\right]^2},$$

where

$$\begin{aligned} x &= \left[A_2' m_1 + A_2 m_1' + 2 m_1 m_1' A_3 - 2 m_2 m_2' A_3 + \left(m_1^2 - m_2^2\right) A_3'\right] \cdot \\ &\cdot \left[(A_2 + A_3 m_1)^2 + A_3^2 m_2^2\right] - \left[A_2 m_1 + \left(m_1^2 - m_2^2\right) A_3\right] \cdot \\ &\cdot \left[2\left(A_2 + A_3 m_1\right)\left(A_2' + A_3' m_1 + A_3 m_1'\right) + 2 A_3 A_3' m_2^2 + 2 A_3^2 m_2 m_2'\right]. \end{aligned}$$

The boundedness of m_1', A_2', A_3', and $m_2 m_2'$ on $x \in (u-\varepsilon, u)$ implies the boundedness of x. As to the denominator $\left[(A_2 + A_3 m_1)^2 + A_3^2 m_2^2\right]^2$, it stays away from zero since $A_2 + A_3 m_1$ stays away from zero for $x \in (u-\varepsilon, u)$.□

Proof of proposition 3: Note that $\left\| F^{\frac{1}{T}XX'} - G \right\| \leq \left\| F^{\frac{1}{T}XX'} - F^{\frac{1}{T}ee'} \right\| + \left\| F^{\frac{1}{T}ee'} - G \right\|$. From the proof of propostition 1, we know that $\left\| F^{\frac{1}{T}XX'} - F^{\frac{1}{T}ee'} \right\| \leq \frac{r}{n}$. By assumption 5, $\left\| F^{\frac{1}{T}ee'} - G \right\| \sim n^{-\beta_1}$. Therefore,

$$\left\| F^{\frac{1}{T}XX'} - G \right\| = O_p(n^{-\beta_1}) + O(n^{-1}) = O_p(n^{-\beta_1}), \tag{16}$$

where the second equality follows from the assumption that $\beta_1 \leq 1$.

Further, according to lemma 2, $a(u-x)^{\frac{3}{2}} = (1 - G(x))\,(1 + O(u-x))$. This implies that

$$a(u-x)^{\frac{3}{2}} = (1 - G(x)) \left(1 + O\left[(1 - G(x))^{\frac{2}{3}}\right]\right). \tag{17}$$

From (16), we have:

$$1 - G(\lambda_{r_{\max}+1}) = 1 - F^{\frac{1}{T}XX'}(\lambda_{r_{\max}+1}) + O_p(n^{-\beta_1}) = \frac{r_{\max}}{n} + O_p(n^{-\beta_1}).$$

Substituting this into (17) and rearranging, we obtain

$$a(u - \lambda_{r_{\max}+1})^{\frac{3}{2}} = \frac{r_{\max}}{n} + O_p(n^{-\beta_1}) + O_p\left(n^{-\frac{5(1-\alpha)}{3}}\right), \tag{18}$$

where the terms $O_p(n^{-\beta_1}) O_p\left(\left(\frac{r_{\max}}{n}\right)^{2/3}\right)$ and $O_p(n^{-\beta_1}) O_p\left(n^{-\frac{2\beta_1}{3}}\right)$ are subsumed by $O_p(n^{-\beta_1})$, and the term $\frac{r_{\max}}{n} O_p\left(n^{-\frac{2\beta_1}{3}}\right)$ is subsumed by $O_p(n^{-\beta_1})$ if $1 - \alpha \geq \beta_1$ and by $O_p\left(n^{-\frac{5(1-\alpha)}{3}}\right)$ if $1 - \alpha < \beta_1$. Similarly, we have

$$a(u - \lambda_{2r_{\max}+1})^{\frac{3}{2}} = 2\frac{r_{\max}}{n} + O_p(n^{-\beta_1}) + O_p\left(n^{-\frac{5(1-\alpha)}{3}}\right) \tag{19}$$

Dividing (19) by (18) and taking the both sides of the resulting equality into power $\frac{2}{3}$, we get

$$\frac{u-\lambda_{2r_{\max}+1}}{u-\lambda_{r_{\max}+1}}=\left[\frac{2\frac{r_{\max}}{n}+O_p(n^{-\beta_1})+O_p\left(n^{-\frac{5(1-\alpha)}{3}}\right)}{\frac{r_{\max}}{n}+O_p(n^{-\beta_1})+O_p\left(n^{-\frac{5(1-\alpha)}{3}}\right)}\right]^{\frac{2}{3}}. \tag{20}$$

Now, consider first the case $\beta_1 \leq 1-\alpha$. Then, the right hand side of (20) can be represented in the form $2^{\frac{2}{3}}(1+O_p(1))$, and we have:

$$u=w\lambda_{r_{\max}+1}+(1-w)\lambda_{2r_{\max}+1}+\zeta, \tag{21}$$

where $w=2^{\frac{2}{3}}/\left(2^{\frac{2}{3}}-1\right)$ and $\zeta=(u-\lambda_{r_{\max}+1})\,O_p(1)$. Note that, for $\beta_1\leq 1-\alpha$, (18) implies that $u-\lambda_{r_{\max}+1}=O_p\left(n^{-\frac{2\beta_1}{3}}\right)$ and therefore $\zeta=O_p\left(n^{-\frac{2\beta_1}{3}}\right)$.

If $1-\alpha<\beta_1\leq\frac{5}{3}(1-\alpha)$, then the right hand side of (20) is $2^{\frac{2}{3}}(1+O_p(n^{-\beta_1+(1-\alpha)}))$. In addition, (18) implies that $u-\lambda_{r_{\max}+1}=O_p\left(n^{-\frac{2(1-\alpha)}{3}}\right)$. Therefore, (21) holds with $\zeta=O_p\left(n^{-\beta_1+\frac{1}{3}(1-\alpha)}\right)$.

Finally, if $\frac{5}{3}(1-\alpha)<\beta_1$, then the right hand side of (20) is $2^{\frac{2}{3}}\left(1+O_p\left(n^{-\frac{2(1-\alpha)}{3}}\right)\right)$ and $u-\lambda_{r_{\max}+1}=O_p\left(n^{-\frac{2(1-\alpha)}{3}}\right)$. Hence, (21) holds with $\zeta=O_p\left(n^{-\frac{4}{3}(1-\alpha)}\right)$.

Summarizing the three cases, we have:

$$\begin{aligned}\hat{u}-u &= O_p\left(n^{-g(\alpha,\beta_1)}\right),\\ \lambda_{r_{\max}+1}-u &= O_p\left(n^{-h(\alpha,\beta_1)}\right).\end{aligned}$$

Finally, note that $\hat{u}-\mu_1=(\hat{u}-u)+(u-\mu_1)$ and $\lambda_{r_{\max}+1}-\mu_1=(\lambda_{r_{\max}+1}-u)+(u-\mu_1)$. The second term in these equalities decays as $n^{-\beta_2}$ by assumption 5. And, therefore, the rate of convergence of $\hat{u}-\mu_1$ is $\min\{g(\alpha,\beta_1),\beta_2\}$, and the rate of convergence $\lambda_{r_{\max}+1}-\mu_1$ is $\min\{h(\alpha,\beta_1),\beta_2\}$.□

References

[1] Anderson, T.W. (1984) An Introduction to Multivariate Statistical Analysis, John Wiley & Sons, New York, Chichester, Brisbane, Toronto, Singapore.

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

[3] Bai., Z. (1999) "Methodologies in spectral analysis of large dimensional random matrices, a review", Statistica Sinica, 9, 611-677.

[4] Bai, Z. and Silverstein, J. (1998) “No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices”, The Annals of Probability, 26, pp 316-345.

[5] Bai, Z., Silverstein, J., and Yin, Y. (1988) “A note on the largest eigenvalue of a large dimensional sample covariance matrix”, Journal of Multivariate Analysis, 26, pp 166-168.

[6] Bernanke, B., Boivin, J., and Eliasz, P. (2004) “Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach”, NBER Working Paper 10220

[7] Bottcher, A. and Silbermann, B., (1998) Introduction to Large Truncated Toeplitz Matrices, Springer, New York, Berlin, Heidelberg, Barcelona, Hong Kong, London, Milan, Paris, Singapore, Tokyo.

[8] Brown, S. (1989) “The number of factors in security returns”, The Journal of Finance, 64, pp 1247-1262.

[9] Brown, S. and Weinstein, M., (1983) “A new approach to testing asset pricing models: the bilinear paradigm”, Journal of Finance, 38, pp 711-743.

[10] Chamberlain, G. and Rothchild, M. (1983) “Arbitrage, factor structure, and mean-variance analysis on large asset markets”, Econometrica, 51, pp.1281-1304.

[11] Connor, G. and Korajczyk, R. (1993) “A test for the number of factors in an approximate factor model”, The Journal of Finance, 58, pp. 1263-1291

[12] Dhrymes, P., Friend. I., and Gultekin, N. (1984) “A critical reexamination of the empirical evidence on the arbitrage pricing theory”, The Journal of Finance, 39, pp 323-346.

[13] Dozier, B. and Silverstein, J. (2004) “Analysis of the limiting spectral distribution of large dimensional information-plus-noise type matrices”, manuscript, Mathematics Department, North Carolina State University.

[14] Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2000) “The generalized dynamic-factor model: identification and estimation”, The Review of Economics and Statistics, 82, pp 540-554,

[15] Forni, M. and Lippi, M. (1999) “Aggregation of linear dynamic microeconomic models”, Journal of Mathematical Economics, 31, pp 131-158.

[16] Forni, M. and Reichlin, L. (1998) “Let’s get real: a factor analytical approach to disaggregated business cycle dynamics”, Review of Economic Studies, 65, pp 453-473.

[17] Horn, R.A. and Johnson C.R. (1991) Topics in Matrix Analysis, Cambridge University Press, Cambridge, New York, Port Chester, Melbourne, Sydney.

[18] Huang, R. and Jo, H. (1995). “Data frequency and the number of factors in stock returns”, Journal of Banking and Finance, 19, pp 987-1003.

[19] Ingersol, J., (1984) “Some results in the theory of arbitrage pricing”, The Journal of Finance, 39, pp 1021-1039.

[20] Johnstone, I (2000). “On the distribution of the largest principal component”, manuscript, Department of Statistics, Stanford University

[21] Hachem, W., Loubaton, P., and Najim J. (2005), “The empirical eigenvalue distribution of a Gram matrix: from independence to stationarity”, Manuscript, http://arxiv.org/abs/math.PR/0502535

[22] Kapetanios, G. (2004) “A new method for determining the number of factors in factor models with large datasets”, Department of Economics, Queen Mary University of London, working paper 525.

[23] Luedecke, B. (1984) An empirical investigation into arbitrage and approximate K-factor structure of large asset markets. Doctoral dissertation, Department of Economics, University of Wisconsin.

[24] Marcellino, M., Stock, J.. H., and Watson M. W. (2003) "Macroeconomic Forecasting in the Euro Area: Country Specific versus Area-Wide Information", European Economic Review, 47, pp. 1-18.

[25] Reinganum, M., (1981) “The arbitrage pricing theory: some empirical results”, Journal of Finance, 36, pp 313-321.

[26] Roll, R., and Ross S. (1980) “An empirical investigation of the arbitrage pricing theory”, Journal of Finance, 5, pp 1073-1103

[27] Ross, S. (1976) “The arbitrage theory of capital asset pricing”, Journal of Economic Theory, 13, pp.341-360.

[28] Silverstein, J. (1995) “Strong convergence of the empirical distribution of large dimensional random matrices”, Journal of Multivariate Analysis, 55, pp 331-339.

[29] Silverstein, J. (1999) “Comment on “Methodologies in spectral analysis of large dimensional random matrices, a review” by Z. Bai”, Statistica Sinica, 9, 611-677.

[30] Silverstein, J. and Choi, S. (1995) “Analysis of the limiting spectral distribution of large dimensional random matrices”, Journal of Multivariate Analysis, 54, pp 295-309

[31] Stock, J. and Watson, M. (1999) “Diffusion Indexes”, manuscript, Economics Department, Harvard University.

[32] Stock, J. and Watson, M. (2002) “Macroeconomic Forecasting Using Diffusion Indexes”, Journal of Business and Economic Statistics, 20, pp. 147-162.

[33] Summers, R. and Heston, A. (1991). “The Penn World Table (Mark 5): an expanded set of international comparisons, 1950-1988”, Quarterly Journal of Economics, 106(2), May, 327-68.

[34] Trzcinka, C. (1986) “On the number of factors in the arbitrage pricing model”, The Journal of Finance, XLI, pp 347-368.

[35] Watson, M.W. (2003) “Macroeconomic Forecasting Using Many Predictors", in M. Dewatripont, L. Hansen and S. Turnovsky (eds), Advances in Economics and Econometrics, Theory and Applications, Eight World Congress of the Econometric Society,” Vol. III, page 87-115.