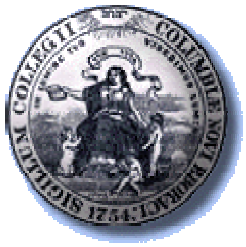

Columbia University

Department of Economics
Discussion Paper Series

Determinimg the Number of Factors from Empirical Distribution of Eigenvalues

Alexei Onatski

Discussion Paper No.: 0405-19

Department of Economics
Columbia University
New York, NY 10027

March 2005

Determinimg the Number of Factors from Empirical Distribution of Eigenvalues

Alexei Onatski
Economics Department, Columbia University

March 6, 2005

Abstract

We develop a new consistent and simple to compute estimator of the number of factors in the approximate factor models of Chamberlain and Rothchild (1983). Our setting requires both time series and cross-sectional dimensions of the data to be large. The main theoretical advantage of our estimator relative to the previously proposed ones is that it works well even in the situation when the portion of the observed variance attributed to the factors is small relative to the variance due to the idiosyncratic term. This advantage arises 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. Extensive Monte Carlo analysis shows that our estimator outperforms the recently proposed Bai and Ng (2002) estimators in finite samples when the "signal-to-noise" ratio is relatively small. We apply the new estimation procedure to determine the number of pervasive factors driving stock returns for the companies traded on NYSE, AMEX, and NASDAQ in the period from 1983 to 2003. Our estimate is equal to 8.

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 (Chamberlain and Rothchild (1983) and Ingersol (1984)), 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 the arbitrage pricing theory of Ross (1976). 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)".

The rout taken by many studies, including Connor and Korajczyk (1993), Stock and Watson (1999), and Bai and Ng (2002), is, therefore, to restrict the approximate factor model so that the eigenvalues "corresponding" to the factors increase fast, i.e. proportionately to the number of the cross-sectional units. Even with this assumption, when the coefficient of proportionality is small and the bound on "idiosyncratic" eigenvalues is large, it will be difficult to tell the growing and the bounded sequences apart. This intuition suggests that for the fast growth asymptotics to be useful for the inference in samples of moderate size, such as for example the sample of all world's countries, the average effect of the factors over the cross-sectional units must be comparable in magnitude to the variation due to the idiosyncratic components. This may or may not be so in applications. It is conceivable, for example, that the effect of a global productivity shock on each particular country can be marred in the idiosyncratic noise. Although the cumulative economic effect of such a shock will be large, it will not necessarily overwhelmingly dominate the effects of some local shocks. In general, potential existence of pervasive factors which have weak average effect relative to the idiosyncratic component undermines usefulness of the fast growth assumption in finite samples.

This paper considers an alternative restriction on the approximate factor model that helps to distinguish the diverging sequence from the bounded sequence. Instead of requiring the divergence rate to be fast, we impose some structure on the idiosyncratic component. 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 relatively arbitrary heteroskedasticity and cross-sectional serial correlation of 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. We find that our estimator substantially outperforms Bai and Ng (2002) estimators in the situations when the portion of the observed variance attributed to the factors is small relative to the variance due to the idiosyncratic component.

Our approach is explicitly based on the investigation of behavior of the eigenvalues of the data's covariance matrix and, thus, 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 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. 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.

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 those of Bai and Ng (2002). Section 5 uses the new method to estimate the number of factors in the arbitrage pricing theory. 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 $r_{\max} = \min(n^\alpha, T^\alpha)$, 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. $\min \operatorname{eval}(\Lambda'\Lambda) \to \infty$, $B_1 < \operatorname{eval}\left(\frac{1}{T}\sum_{t=1}^{T} F_t F_t'\right) < B_2$ *almost surely for some fixed* $0 < B_1 \leq B_2 < \infty$

Intuitively, this assumption 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. 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 $\min \operatorname{eval}(\Lambda'\Lambda) > an$, 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 portion of the variation in the data explained by the factors is low relative to the variation due to the idiosyncratic term. In addition, the "weaker" pervasive factors may be important for the approximate asset pricing formula of Chamberlain and Rothschild (1983). Indeed, the formula includes betas corresponding to i-th factors with $\min \text{eval}\left(\Lambda_i'\Lambda_i\right)$ increasing to infinity not necessarily as fast as n.

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 $n \times n$ matrix S_n, such that*

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

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

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 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) $c \int \frac{t^2 dH(t)}{(u(H)-t)^2} > 1$ *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.

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 to be 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

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

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

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:

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

where $\hat{u} = w\lambda_{r_{\max}+1} + (1-w)\,\lambda_{2r_{\max}+1}$ and $w = 2^{2/3}/\left(2^{2/3}-1\right)$, indexed by a positive number δ. 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 (8) 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 number. 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 sample 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 now turn to the formal proof of the strong consistency of $\hat{r}_\delta$. We will first show that $\hat{u}$ is strongly consistent for u. The strong consistency is a consequence of the following proposition:

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) The spectral distribution of $\frac{1}{T}XX'$ weakly converges to G almost surely.
iii) For any $i > r$ and such that $\frac{i}{n} \to 0$ as $n \to \infty$, $\lambda_i \to u$ almost surely, where u denotes the upper boundary of the support of G.

Proof: Theorem 1.1 of Silverstein (1995) implies[2] that the spectral distribution of $\frac{1}{T}ee'$ weakly converges to a distribution G as $n \to \infty$. That G must have

[2] Condition a) of that theorem is implied by our assumption 3, condition b) is our assump-

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 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.

To prove ii) 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 $F^{\frac{1}{T}ee'}$.

Turning to the proof of iii), let us denote the j-th largest eigenvalues of $\frac{1}{T}ee'$ as μ_j, and let j be such that $\frac{j}{n} \to 0$ as $n \to \infty$. We will first 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).[3] 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:

$$\left| F^{\frac{1}{T}ee'}(x_0) - G(x_0) \right| \to 0 \tag{6}$$

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 (6).

tion 1, condition c) is implied by our assumption 4 i), and condition d) is obviously satisfied because we require S_n to be non-random.

[3] 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).

Let us now denote the j-th largest eigenvalue of $\frac{1}{T}\Lambda FF'\Lambda'$ as ν_j. 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:

$$\sigma_{i+j-1}(A+B) \leq \sigma_i(A) + \sigma_j(B), \tag{7}$$

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 equivalent to say that $\sigma_i(A)$ is a square root of the i-th largest eigenvalue of matrix AA'. This inequality 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 $\frac{1}{T}\Lambda FF'\Lambda'$ is equal to r, $\nu_{r+1}^{1/2}$ must be equal to zero so that we have:

$$\lambda_i \leq \mu_{i-r}, \text{ for } i = r+1, ..., n \tag{8}$$

$$\lambda_i \geq \mu_{i+r}, \text{ for } i = 1, ..., n-r \tag{9}$$

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.□

The fact that $\hat{u}$ converges to u almost surely immediately follows from statement iii) of the proposition. Indeed, since by assumption $r_{\max} = \min(n^\alpha, T^\alpha)$ 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:

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 iii) 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 $\hat{r}_\delta \to r$.

Substituting $A = \frac{1}{\sqrt{T}}X$ and $B = \frac{-1}{\sqrt{T}}e$ into inequality (7), we get:

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

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:

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

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 u in the case of $\tilde{r}_\delta$, and the gap between $\hat{u}$ and u 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 gap between $\hat{u}$ and u 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: $\left\|F\frac{1}{T}ee' - G\right\| = O_p(n^{-\beta})$, *and* $|\mu_1 - u| = O_p(n^{-\beta})$, *where* $0 < \beta \leq 1$

Later, we will conjecture that $\beta = 1$ and will provide arguments in favor of this conjecture.

Let us define

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

We will prove the following

Proposition 3: *Let assumptions 1-5 hold, then*

i) $\hat{r}_\delta$ *is consistent for* r *when* $\delta \sim n^{-\gamma}$, *for any* γ *s.t.* $0 \leq \gamma < g(\alpha, \beta)$,
ii) $\tilde{r}_\delta$ *is consistent for* r *when* $\delta \sim n^{-\gamma}$, *for any* γ *s.t.* $0 \leq \gamma < \frac{2}{3}\min(\beta, 1-\alpha)$

Remark: An immediate consequence of the proposition is that $\hat{r}_\delta$ is consistent for a wider range of the δ's rate of decrease than $\tilde{r}_\delta$.

To proof proposition 3, we will need the following two lemmas:

Lemma 1. $\left\| F^{\frac{1}{T}XX'} - F^{\frac{1}{T}ee'} \right\| \leq \frac{r}{n}$

Lemma 2: *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 $x \uparrow u$.

The proofs of the lemmas are in the Appendix section. Turning to the proof of proposition 3, note that $\left\| F^{\frac{1}{T}XX'} - G \right\| \leq \left\| F^{\frac{1}{T}XX'} - F^{\frac{1}{T}ee'} \right\| + \left\| F^{\frac{1}{T}ee'} - G \right\|$. Assumption 5 and lemma 1 then imply that

$$\left\| F^{\frac{1}{T}XX'} - G \right\| = O_p(n^{-\beta}) + O(n^{-1}) = O_p(n^{-\beta}), \tag{10}$$

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

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

$$a(u-x)^{\frac{3}{2}} = (1 - G(x)) \left(1 + O\left[(1 - G(x))^{\frac{2}{3}}\right]\right). \tag{11}$$

From (10), we have:

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

Substituting this into (11) and rearranging, we obtain

$$a(u - \lambda_{r_{\max}+1})^{\frac{3}{2}} = \frac{r_{\max}}{n} + O_p(n^{-\beta}) + O_p\left(n^{-\frac{5(1-\alpha)}{3}}\right), \tag{12}$$

where the terms $O_p(n^{-\beta})O_p\left(\left(\frac{r_{\max}}{n}\right)^{2/3}\right)$, $\frac{r_{\max}}{n}O_p\left(n^{-\frac{2\beta}{3}}\right)$, and $O_p(n^{-\beta})O_p\left(n^{-\frac{2\beta}{3}}\right)$ are all subsumed by the them $O_p(n^{-\beta})$. Similarly, we have

$$a(u - \lambda_{2r_{\max}+1})^{\frac{3}{2}} = 2\frac{r_{\max}}{n} + O_p(n^{-\beta}) + O_p\left(n^{-\frac{5(1-\alpha)}{3}}\right) \tag{13}$$

Dividing (13) by (12) and taking the both sides of the resulting equality into power $\frac{2}{3}$, we get

$$\frac{u - \lambda_{2r_{\max}+1}}{u - \lambda_{r_{\max}+1}} = \left[\frac{2\frac{r_{\max}}{n} + O_p(n^{-\beta}) + O_p\left(n^{-\frac{5(1-\alpha)}{3}}\right)}{\frac{r_{\max}}{n} + O_p(n^{-\beta}) + O_p\left(n^{-\frac{5(1-\alpha)}{3}}\right)}\right]^{\frac{2}{3}}. \tag{14}$$

Now, consider first the case $\beta \leq 1-\alpha$. Then, the right hand side of (14) can be represented in the form $2^{\frac{2}{3}}(1+O_p(1))$, and we have:

$$u = w\lambda_{r_{\max}+1} + (1-w)\lambda_{2r_{\max}+1} + \zeta, \tag{15}$$

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 $\beta \leq 1-\alpha$, (12) implies that $u-\lambda_{r_{\max}+1} = O_p\left(n^{-\frac{2\beta}{3}}\right)$ and therefore $\zeta = O_p\left(n^{-\frac{2\beta}{3}}\right)$.

If $1-\alpha < \beta \leq \frac{5}{3}(1-\alpha)$, then the right hand side of (14) is $2^{\frac{2}{3}}(1+O_p(n^{-\beta+(1-\alpha)}))$. In addition, (12) implies that $u-\lambda_{r_{\max}+1} = O_p\left(n^{-\frac{2(1-\alpha)}{3}}\right)$. Therefore, (15) holds with $\zeta = O_p\left(n^{-\beta+\frac{1}{3}(1-\alpha)}\right)$.

Finally, if $\frac{5}{3}(1-\alpha) < \beta$, then the right hand side of (14) 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, (15) holds with $\zeta = O_p\left(n^{-\frac{4}{3}(1-\alpha)}\right)$.

Summarizing the three cases, we have:

$$u-\hat{u} = O_p\left(n^{-g(\alpha,\beta)}\right), \tag{16}$$

$$u-\lambda_{r_{\max}+1} = O_p\left(n^{-\frac{2}{3}\min(\beta,1-\alpha)}\right) \tag{17}$$

The above formulas show the rates of convergence of our "sophisticated" estimator of u and a "primitive" estimator $\lambda_{r_{\max}+1}$.

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 (8), it is enough to prove that the probability that $\mu_1 < (1+\delta)\,\hat{u}$ goes to 1 for large enough n. We have:

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

The second term in the above sum is $O_p\left[n^{-g(\alpha,\beta)}\right]$ by (16), the third term is $O_p(n^{-\beta})$ by assumption, the first term decays as fast as δ. Note that $g(\alpha,\beta) < \beta$. Therefore, with probability going to 1, the first term will dominate the other two as $n \to \infty$ if $\delta \sim n^{-\gamma}$, for any γ such that $0 \leq \gamma < g(\alpha,\beta)$, and hence $\Pr\left(\mu_1 < (1+\delta)\,\hat{u}\right) \to 1$ as $n \to \infty$, which completes the proof of statement i) of the proposition.

Similarly,

$$(1+\delta)\,\lambda_{r_{\max}+1} - \mu_1 = \delta\lambda_{r_{\max}+1} + (\lambda_{r_{\max}+1}-u) + (u-\mu_1),$$

and according to (17), with probability going to 1, the first term will dominate the other two as $n \to \infty$ if $\delta \sim n^{-\gamma}$, for any γ such that $0 \leq \gamma < \frac{2}{3}\min(\beta, 1-\alpha)$. Hence, $\Pr\left(\mu_1 < (1+\delta)\,\lambda_{r_{\max}+1}\right) \to 1$ and $\tilde{r}_\delta$ is consistent, which completes the proof of ii). □

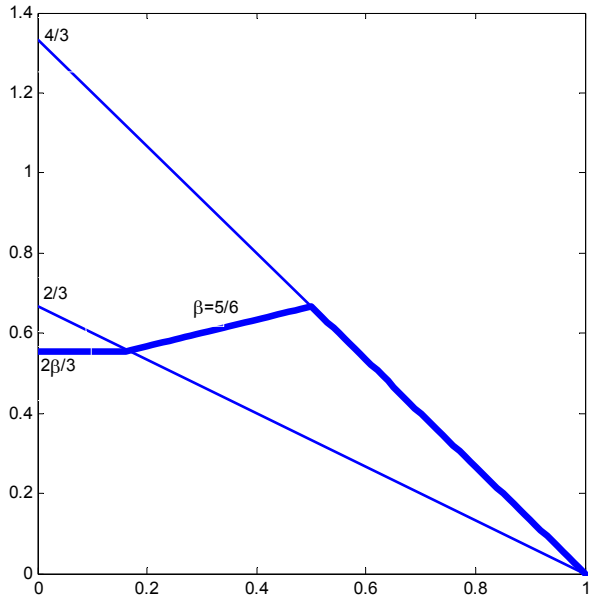

Figure 1: The negative of the exponential in the rate of convergence of $\hat{u}$ as a function of α. β is fixed at 5/6.

Figure 1 shows the rates of convergence of $\hat{u}$ as as a function of α for a range of $0 < \beta \leq 1$. As α increases from 0, the rate stays at $\frac{2}{3}\beta$ until $1-\alpha$ becomes equal to β, then increases until $1-\alpha$ is $\frac{3}{5}\beta$, and finally decreases to 0 as $1-\alpha = 0$. Intuitively, our trying to fit the functional form $1 - a\left(u-x\right)^{\frac{3}{2}}$ 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). Therefore, as $\beta < 1-\alpha$, the rate of convergence of $\hat{u}$ is equal to the rate of convergence of the "primitive" estimator $\lambda_{r_{\max}+1}$. When α becomes such that $\beta > 1-\alpha$, the variation in $F^{\frac{1}{T}XX'}$ becomes large enough to exploit the functional form fitting idea. As α becomes too large, the discrepancy between G and $1 - a\left(u-x\right)^{\frac{3}{2}}$ (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 2 suggests that the optimal choice of α in $r_{\max} = \min\left(n^{\alpha}, T^{\alpha}\right)$, in the sense that it optimizes the rate of convergence of $\hat{u}$ to u, is $1 - \frac{3}{5}\beta$. Unfortunately, not much is known about the true rates of convergence of $\left\| F^{\frac{1}{T}ee'} - G \right\|$

and $|\mu_1 - u|$. The standard conjecture (see Bai 1999, p.658-659), for the case when S_n in (2) is equal to the identity matrix, is that $\beta = 1$. 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 rate of convergence should still be n^{-1}. If $\beta = 1$, then the optimal choice of α is $\frac{2}{5}$. Therefore, in the Monte Carlo simulations and the application below we will use $r_{\max} = k \min\left(n^{2/5}, T^{2/5}\right)$. We choose to use $k = 1.55$ so that in realistic small samples, when T is equal to 60, the maximum number of factors is 8.

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[4] that if, as the standard conjecture is, $\beta = 1$, 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 $\gamma < h(\alpha, \theta)$, where

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

For example, for our choice of $\alpha = \frac{2}{5}$, $\hat{r}_\delta$ will consistently estimate the number of factors rising as $n^{-\theta}$, $\theta \leq \frac{2}{5}$ as long as $\delta \sim n^{-\gamma}$ for some $0 \leq \gamma < \frac{4}{5} - \theta$.

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.[5] 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.

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

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

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

(2002) estimator. We replicate and extend Bai and Ng's experiments. Their most general setting is as follows:

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

where the factors 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

$$e_{it} = \rho e_{i,t-1} + v_{it} + \sum_{j=-J}^{J} \beta v_{i-j,t}, \; J = \min\left\{\frac{N}{20}, 10\right\}, \; v_{ij} \sim IIDN(0,1). \quad (18)$$

In most of their exercises Bai and Ng assume that the idiosyncratic component $\sqrt{\theta} e_{it}$ has the same variance as the common component $\sum_{k=1}^{r} \Lambda_{ik} F_{kt}$, that is $\theta = r$. We consider less configurations of n and T than Bai and Ng do, trying to choose the most representative ones.[6] However, we experiment more with the relative size of θ and r. Increasing the size of θ relative to r is a particularly interesting exercise for us because large θ relative to r corresponds to the situation when our new estimator should improve upon that of Bai and Ng (2002).

Table 1 reports the averages of Bai-Ng estimators and three versions of our estimator $\hat{r}_\delta$, corresponding to $\delta = \max\left(n^{-\frac{1}{2}}, T^{-\frac{1}{2}}\right), \delta = 0$, and $\delta = \max\left(n^{-\frac{2}{3}}, T^{-\frac{2}{3}}\right)$, over 1000 replications of the data generating process. The true number of factors is assumed to be $r = 1, 3$, and 5. We take e_{it} homoskedastic independent $N(0,1)$ variables. We set $r_{\max} = 1.55 \min\left(T^{2/5}, n^{2/5}\right)$. 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}	$\hat{r}_{n^{-\frac{1}{2}}}$	$\hat{r}_0$	$\hat{r}_{n^{-\frac{2}{3}}}$
100	40	**1**	**7**	1.02	1.00	2.77	1.00	1.00	1.00	1.02	1.40	1.01
200	60	**1**	**8**	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.35	1.05
1000	60	**1**	**8**	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.19	1.00
100	100	**1**	**10**	1.00	1.00	7.32	1.00	1.00	1.11	1.06	1.57	1.23
100	40	**3**	**7**	3.00	3.00	3.62	3.00	2.99	3.00	3.00	3.03	3.00
200	60	**3**	**8**	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.03	3.00
1000	60	**3**	**8**	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.01	3.00
100	100	**3**	**10**	3.00	3.00	7.33	3.00	3.00	3.09	3.00	3.10	3.03
100	40	**5**	**7**	4.98	4.93	5.06	4.75	4.36	4.99	4.91	4.99	4.95
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
100	100	**5**	**10**	5.00	5.00	7.68	5.00	4.99	5.10	5.00	5.01	5.00

[6] Most of Bai and Ng's (2002) results do not change much until n and T change dramatically.

As can be seen from table 1, when $\theta = r$, all estimators, except PC_{p3}, which substantially overestimates the true number of factors when $n = T = 100$, work very well and their performance is comparable. These results confirm Bai and Ng's (2002) findings that they report in tables 1-3 of their paper. 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 $\delta = \max\left(n^{-\frac{1}{2}}, T^{-\frac{1}{2}}\right)$ and $\delta = \max\left(n^{-\frac{2}{3}}, T^{-\frac{2}{3}}\right)$ does not lead to substantially different results.

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 $\theta = 4r$, that is the standard deviation of the idiosyncratic component is 2 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 for $r = 5$. The Bai-Ng estimators start to significantly underestimate the number of factors. A particularly striking deterioration of performance happens for IC_{p1} and IC_{p2}, which essentially say the data generating process does not have any factors. Note that our estimator still works very well for $T > 40$. 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. Since our method of estimation relies more on the structural properties of the idiosyncratic process (which does not change when θ is increased), its performance turns out to be less sensitive to the increase in the idiosyncratic variance.

Table 2: $\theta = 4r, \rho = \beta = 0$

n	T	r	$r_{\max}$	PC_{p1}	PC_{p2}	PC_{p3}	IC_{p1}	IC_{p2}	IC_{p3}	$\hat{r}_{n^{\frac{-1}{2}}}$	$\hat{r}_0$	$\hat{r}_{n^{\frac{-2}{3}}}$
100	40	**1**	**7**	1.00	1.00	1.98	1.00	0.99	1.00	1.01	1.37	1.07
200	60	**1**	**8**	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.32	1.03
1000	60	**1**	**8**	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.17	1.00
100	100	**1**	**10**	1.00	1.00	5.33	1.00	1.00	1.00	1.03	1.46	1.16
100	40	**3**	**7**	2.66	2.30	3.17	0.66	0.24	2.24	2.60	2.94	2.78
200	60	**3**	**8**	2.90	2.78	3.00	1.91	1.39	2.83	3.00	3.03	3.00
1000	60	**3**	**8**	2.95	2.94	2.98	2.55	2.48	2.73	3.00	3.01	3.00
100	100	**3**	**10**	2.99	2.82	5.79	2.14	1.00	3.00	3.00	3.07	3.01
100	40	**5**	**7**	2.31	1.72	3.99	0.03	0.00	1.03	2.06	3.05	2.50
200	60	**5**	**8**	2.63	2.14	3.93	0.22	0.07	1.77	4.34	4.83	4.61
1000	60	**5**	**8**	2.73	2.62	3.07	0.46	0.37	0.83	5.00	5.00	5.00
100	100	**5**	**10**	3.33	2.20	6.55	0.39	0.02	4.91	4.54	4.87	4.76

When we make $\theta = 9r$ (see table 3), the deterioration of the performance of the Bai-Ng estimators and relative robustness of our estimators is even more striking.

Table 3: $\theta = 9r, \rho = \beta = 0$

n	T	r	$r_{\max}$	PC_{p1}	PC_{p2}	PC_{p3}	IC_{p1}	IC_{p2}	IC_{p3}	$\hat{r}_{n^{\frac{-1}{2}}}$	$\hat{r}_0$	$\hat{r}_{n^{\frac{-2}{3}}}$
100	40	**1**	**7**	1.00	0.99	1.77	0.59	0.38	0.96	1.01	1.30	1.05
200	60	**1**	**8**	1.00	1.00	1.00	0.96	0.92	1.00	1.00	1.28	1.03
1000	60	**1**	**8**	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.19	1.00
100	100	**1**	**10**	1.00	1.00	4.84	0.99	0.85	1.00	1.03	1.45	1.14
100	40	**3**	**7**	1.01	0.55	2.63	0.00	0.00	0.13	1.10	2.01	1.50
200	60	**3**	**8**	0.90	0.56	1.99	0.01	0.00	0.25	2.57	2.96	2.79
1000	60	**3**	**8**	0.71	0.63	0.93	0.01	0.01	0.02	3.00	3.01	3.00
100	100	**3**	**10**	1.42	0.56	5.45	0.01	0.00	2.69	2.69	3.00	2.87
100	40	**5**	**7**	0.47	0.13	2.47	0.00	0.00	0.01	0.52	1.57	0.88
200	60	**5**	**8**	0.06	0.01	1.03	0.00	0.00	0.00	1.81	3.02	2.37
1000	60	**5**	**8**	0.01	0.00	0.02	0.00	0.00	0.00	4.41	4.96	4.78
100	100	**5**	**10**	0.41	0.02	6.01	0.00	0.00	1.96	2.16	3.13	2.65

Our next step is to introduce time series serial correlation to the idiosyncratic terms. For this, we take $\rho = 0.5$ in (18). Table 4 reports our results for the cases $\theta = r$ and $\theta = 4r$. Interestingly, all PC estimators and IC_{p3} are overestimating the number of factors now. The overestimation is not substantial (except for IC_{p3}) for $r = 5$, but it is very sizable for $r = 1$ and $r = 3$. This phenomenon cannot be seen from Bai and Ng (2002) because they chose to analyze only the case $r = 5$ when cross-sectional serial correlation was introduced. As to the IC_{p1} and IC_{p2}, we again see that their performance deteriorates dramatically when $\theta = 4r$. The performance of our estimators deteriorates significantly for $r = 5$, but not as much for $r = 3$, and much less dramatically than that of IC_{p1} and IC_{p2}.

To complete the series of experiments based on the Bai-Ng setup, we study performance of the estimators in case when both cross-sectional and time series serial correlation are introduced. Precisely, we take $\rho = 0.5$ and $\beta = 0.2$ in (18). Table 5 reports our results for $\theta = r$ and $\theta = 4r$. All estimators work relatively poorly, except perhaps our estimators for $r = 3$ and $\theta = r$. The Bai-Ng estimators overestimate the true number of factors very substantially. Our estimators are doing more reasonable job. When θ jumps by a factor of 4 the overestimation of the Bai-Ng estimators is still very substantial, except for the case $n = 1000, T = 60$, $r = 3, 5$, when the IC estimators swing from the overestimation to the underestimation extreme. For this case, when $\theta = 4r$, our estimators also severely underestimate the true number of factors. However, they appear to be much more stable relative to the case $\theta = r$.

Table 4a: $\theta = r, \rho = 0.5, \beta = 0$

n	T	r	$r_{\max}$	PC_{p1}	PC_{p2}	PC_{p3}	IC_{p1}	IC_{p2}	IC_{p3}	$\hat{r}_{n^{\frac{-1}{2}}}$	$\hat{r}_0$	$\hat{r}_{n^{\frac{-2}{3}}}$
100	40	**1**	**7**	5.42	4.59	6.98	1.21	1.02	6.82	1.37	2.18	1.67
200	60	**1**	**8**	4.49	3.61	7.16	1.00	1.00	2.76	1.09	1.62	1.26
1000	60	**1**	**8**	4.18	3.91	5.02	1.0	1.00	1.00	1.00	1.02	1.00
100	100	**1**	**10**	3.91	1.93	10.00	1.00	1.00	10.00	1.47	2.33	1.86
100	40	**3**	**7**	5.88	5.09	7.00	3.36	3.02	6.97	2.97	3.15	3.03
200	60	**3**	**8**	5.03	4.25	7.51	3.01	3.00	5.15	3.00	3.07	3.01
1000	60	**3**	**8**	4.46	4.22	5.25	3.00	3.00	3.00	3.00	3.00	3.00
100	100	**3**	**10**	4.54	3.17	10.00	3.00	3.00	10.00	3.06	3.34	3.16
100	40	**5**	**7**	6.22	5.58	7.00	4.92	3.94	6.99	3.16	3.84	3.49
200	60	**5**	**8**	5.70	5.22	7.78	4.99	4.96	6.99	4.74	4.94	4.85
1000	60	**5**	**8**	5.18	5.09	5.68	5.00	5.00	5.00	4.97	5.00	4.99
100	100	**5**	**10**	5.51	5.01	10.00	4.99	4.87	10.00	5.00	5.01	5.00

Table 4b: $\theta = 4r, \rho = 0.5, \beta = 0$

n	T	r	$r_{\max}$	PC_{p1}	PC_{p2}	PC_{p3}	IC_{p1}	IC_{p2}	IC_{p3}	$\hat{r}_{n^{\frac{-1}{2}}}$	$\hat{r}_0$	$\hat{r}_{n^{\frac{-2}{3}}}$
100	40	**1**	**7**	5.16	4.29	6.97	1.08	0.98	6.64	1.25	2.01	1.53
200	60	**1**	**8**	4.12	3.26	6.95	1.00	1.00	1.88	1.05	1.53	1.18
1000	60	**1**	**8**	3.96	3.66	4.83	1.00	1.00	1.00	1.00	1.01	1.00
100	100	**1**	**10**	3.14	1.36	10.00	1.00	1.00	10.00	1.40	2.25	1.78
100	40	**3**	**7**	5.26	4.38	6.99	0.57	0.15	6.73	1.14	2.08	1.50
200	60	**3**	**8**	4.51	3.79	7.16	1.35	0.79	3.75	1.81	2.51	2.12
1000	60	**3**	**8**	4.16	3.91	4.93	2.03	1.89	2.39	1.84	2.48	2.17
100	100	**3**	**10**	3.87	2.88	10.00	1.35	0.28	10.00	2.79	3.17	2.97
100	40	**5**	**7**	4.92	3.98	6.95	0.10	0.01	6.21	0.72	1.72	1.13
200	60	**5**	**8**	4.25	3.46	6.99	0.10	0.01	2.50	0.80	1.76	1.19
1000	60	**5**	**8**	4.16	3.96	4.81	0.13	0.10	0.35	0.52	1.44	0.93
100	100	**5**	**10**	4.12	2.50	10.00	0.09	0.00	10.00	2.18	3.05	2.63

Table 5a: $\theta = r, \rho = 0.5, \beta = 0.2$

n	T	r	$r_{\max}$	PC_{p1}	PC_{p2}	PC_{p3}	IC_{p1}	IC_{p2}	IC_{p3}	$\hat{r}_{n^{\frac{-1}{2}}}$	$\hat{r}_0$	$\hat{r}_{n^{\frac{-2}{3}}}$
100	40	**1**	**7**	6.98	6.79	7.00	6.85	5.67	7.00	2.35	3.11	2.66
200	60	**1**	**8**	8.00	8.00	8.00	8.00	8.00	8.00	2.69	3.42	3.05
1000	60	**1**	**8**	6.92	6.71	7.46	2.66	2.24	4.55	1.21	1.94	1.48
100	100	**1**	**10**	9.96	9.07	10.00	9.83	6.63	10.00	3.93	4.71	4.34
100	40	**3**	**7**	6.99	6.90	7.00	6.96	6.45	7.00	2.66	3.32	2.94
200	60	**3**	**8**	8.00	8.00	8.00	8.00	8.00	8.00	2.63	3.40	2.99
1000	60	**3**	**8**	7.35	7.19	7.76	5.06	4.64	6.66	2.99	3.15	3.04
100	100	**3**	**10**	10.00	9.58	10.00	9.98	8.25	10.00	3.95	4.58	4.26
100	40	**5**	**7**	7.00	6.92	7.00	6.93	6.00	7.00	1.72	2.61	2.10
200	60	**5**	**8**	8.00	8.00	8.00	8.00	8.00	8.00	1.37	2.25	1.77
1000	60	**5**	**8**	7.62	7.48	7.89	6.48	6.10	7.45	2.64	3.52	3.05
100	100	**5**	**10**	10.00	9.77	10.00	10.00	8.69	10.00	3.45	4.17	3.82

Table 5b: $\theta = 4r, \rho = 0.5, \beta = 0.2$

n	T	r	$r_{\max}$	PC_{p1}	PC_{p2}	PC_{p3}	IC_{p1}	IC_{p2}	IC_{p3}	$\hat{r}_{n^{\frac{-1}{2}}}$	$\hat{r}_0$	$\hat{r}_{n^{\frac{-2}{3}}}$
100	40	1	7	6.96	6.69	7.00	6.72	5.00	7.00	2.17	2.96	2.49
200	60	1	8	8.00	8.00	8.00	8.00	7.99	8.00	2.50	3.29	2.91
1000	60	1	8	6.86	6.67	7.39	2.42	2.01	4.38	1.21	1.91	1.46
100	100	1	10	9.94	8.94	10.00	9.74	6.17	10.00	3.88	4.67	4.29
100	40	3	7	6.93	6.47	7.00	6.09	2.95	7.00	1.58	2.50	1.98
200	60	3	8	8.00	8.00	8.00	8.00	7.95	8.00	2.09	2.89	2.46
1000	60	3	8	6.64	6.45	7.23	0.97	0.71	2.60	0.46	1.44	0.85
100	100	3	10	9.94	8.86	10.00	9.69	3.57	10.00	2.83	3.73	3.30
100	40	5	7	6.78	6.17	7.00	5.01	2.03	7.00	1.70	2.58	2.07
200	60	5	8	8.00	7.95	8.00	7.98	7.75	8.00	2.44	3.30	2.85
1000	60	5	8	5.99	5.79	6.62	0.37	0.25	1.07	0.45	1.45	0.85
100	100	5	10	9.71	8.18	10.00	8.55	2.13	10.00	2.92	3.77	3.36

5 Application to Stock Returns

We apply the newly developed technique to 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 limit when the size of the sample rises. However, the rate of this increase can be slow. In particular, it is possible that the variance of the data explained by the factors is small relative to the variance due to the idiosyncratic component. 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 21*12=252 and $r_{\max} = [1.55 * 252^{2/5}] + 1 = 15$. 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.[7] 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 signal-to-noise ratio is relatively small. 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.

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 the new estimator

[7] 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 mode.

relative to the previously proposed ones is that it works well in the situation when the systematic component of the data represented by the factors explain only a small portion of the variation. This advantage arises 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 proportion 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.

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. 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. Monte Carlo analysis and some theoretical results suggest, 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. It is likely that, using recent results in Silverstein and Dozier (2004), one can show that our estimator remains consistent in this case as long as the spectral distribution of ZZ' converges to a probability distribution with bounded support. This convergence is likely to be present for mild cross-section and time-series correlation of the components in Z. Establishing conditions under which such convergence exists is our immediate plan.

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 approximate factor models.

7 Appendix

Proof of Lemma 1: We will consider only the case when $n > T$. The proof in the other case is similar. If $n > T$, then $n - T$ eigenvalues of $\frac{1}{T}XX'$ and $\frac{1}{T}ee'$ are equal to zero and the rest of eigenvalues are almost surely distinct and not equal to zero. If $x \leq \lambda_n = ... = \lambda_{T+1} = 0$, then $F^{\frac{1}{T}XX'}(x) = F^{\frac{1}{T}ee'}(x)$ and in particular

$$\left| F^{\frac{1}{T}XX'}(x) - F^{\frac{1}{T}ee'}(x) \right| \leq \frac{r}{n} \tag{19}$$

Let $\lambda_{i+1} \leq x < \lambda_i$ for some $i = 0, ..., n$, where we define $\lambda_0 = +\infty$ for convenience. Then,

$$F^{\frac{1}{T}XX'}(x) = F^{\frac{1}{T}XX'}(\lambda_{i+1}) = 1 - \frac{i}{n} \tag{20}$$

by definition of the spectral distribution function (3).

Consider, first, the case when $0 \leq i < r + 1$. Then, by (9)[8], $x \geq \mu_{i+r+1}$ and $F^{\frac{1}{T}ee'}(x) \geq F^{\frac{1}{T}ee'}(\mu_{i+r+1}) = 1 - \frac{i+r}{n}$. This inequality together with (20) implies (19).

Now, let $T - r < i \leq T$. Then, by (8), $x < \mu_{i-r}$ and $\frac{n-T}{n} = F^{\frac{1}{T}ee'}(0) \leq F^{\frac{1}{T}ee'}(x) \leq F^{\frac{1}{T}ee'}(\mu_{i-r}) - \frac{1}{n} = 1 - \frac{i-r}{n}$, which, coupled with (20), implies (19).

Finally, let $r + 1 \leq i \leq T - r$. Then $\mu_{i+r+1} \leq x < \mu_{i-r}$ and $1 - \frac{i+r}{n} \leq F^{\frac{1}{T}ee'}(x) \leq 1 - \frac{i-r}{n}$. Again, taking into account (20), we get (19).□

Proof of Lemma 2: 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

$$m = -\left(z - c \int \frac{\tau dH(\tau)}{1 + \tau m} \right)^{-1}. \tag{21}$$

Silverstein and Choi (1995) study properties of distributions with the Stieltjes transforms satisfying the above equation. They show that P has continuous

[8] Here we implicitly assumed that $2r < n$, which will hold for large enough n.

density $p(x)$, which has form (see formula (5.3) of Silverstein and Choi (1995)):

$$p(x) = a\,(u-x)^{\frac{1}{2}}\,(1+o(1)) \tag{22}$$

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 (22) 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, $\lim_{m\downarrow m_u}\int \frac{\lambda^2 dH(\lambda)}{(1+\lambda m)^2}$ exists, and $\frac{1}{m^2} - c\int \frac{\lambda^2 dH(\lambda)}{(1+\lambda m)^2}$ 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 (22), 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:

$$x = c\int \frac{\lambda dH(\lambda)}{(1+\lambda m_1)^2 + \lambda^2 m_2^2} \tag{23}$$

$$0 = \frac{1}{m_1^2+m_2^2} - c\int \frac{\lambda^2 dH(\lambda)}{(1+\lambda m_1)^2+\lambda^2 m_2^2}. \tag{24}$$

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 (22) 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 (23) 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 $A_2 + A_3 m_1 = 2c \int \frac{\lambda^2(1+\lambda m_1)dH(\lambda)}{\left((1+\lambda m_1)^2+\lambda^2 m_2^2\right)^2}$ 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 $m_2 m_2'$ 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

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

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 $x \in (u-\varepsilon, u)$.□

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] 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.

[22] 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.

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

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

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

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

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

[28] 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

[29] Stock, J. and Watson, M. (1999) "Diffusion Indexes", manuscript, Economics Department, Harvard University.

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