Estimating the Number of Factors in Large Dimensional Factor Models[1]

Matthew Harding[2]

April 16, 2013

Abstract

This paper develops a new spectral approach to the estimation of the number of latent factors in large dimensional factor models. It shows that by imposing restrictions on the error terms we can derive a consistent procedure with improved finite sample performance in the presence of weak factors. The paper uses free probability theory to derive analytic expressions for the limiting moments of the spectral distribution, which greatly simplifies the computational burden. The new approach performs very well in a series of Monte Carlo simulations against the leading competing approaches. This procedure is also shown to provide realistic estimates of the number of latent factors in applications of factor models to asset pricing and data reduction of macroeconomic indicators.

JEL: C33, G11, C46

Keywords: factor models, random matrix theory, free probability theory, free independence

[1]I would like to thank Victor Chernozhukov, Alan Edelman, Jerry Hausman, Whitney Newey, and Raj Rao for their encouragement and support. Gourab Mukherjee and Marcel Priebsch provided excellent research assistance. An earlier version of this paper was presented at Boston College, Boston University, Caltech, Cambridge, Chicago, Harvard, MIT, Stanford, UCL, UCSD, USC and Wharton, where seminar participants provided many useful comments.

[2]Department of Economics, Stanford University, 579 Serra Mall, Stanford, CA 94305; Phone: (650) 723-4116; Fax: (650) 725-5702; Email: mch@stanford.edu

1. Introduction

Factor models are an important tool of modern economics and finance and relate observed data to a small set of unobserved variables, which are then estimated. While factor models have been used for almost a century, classical econometric methods were developed under the assumption that the time dimension grows large while the cross-section dimension is small and bounded. In applications where both the number of individuals and the number of time periods is large, standard econometric approaches become unreliable for the estimation of the number of factors. As a result the development of new econometric procedures for the estimation of high-dimensional factor models is an active area of research (Bai and Ng, 2002; Stock and Watson, 2005; Hallin and Liska, 2007; Onatski, 2009, 2010, 2012; Doz, Giannone and Reichlin, 2012).

This paper develops a new technique for the estimation of the number of factors in large N, large T approximate factor models by relating the identification and estimation of the number of factors to the asymptotic behavior of sample eigenvalues of large random matrices, thus providing a connection between economics and the new mathematical field of Random Matrix Theory.

Existing approaches to the estimation of the number of factors often have poor finite sample properties, especially in situations where the factors are "weak" and do not strongly dominate the noise component (Ouysse, 2006; Boivin and Ng, 2006; Uhlig, 2008; Bai and Ng, 2008; Onatski, 2012). Factor models continue to be used in many applied settings and the central role they play in finance only serves to strengthen the need for new estimators with improved finite sample performance.

This paper introduces a new approach to the consistent estimation of the number of factors by imposing restrictions on the time and spatial correlation patterns of the error terms. It assumes that the correlation patterns only depend on a low dimension vector of unknown parameters, which can be estimated from the data. We introduce a stopping rule for counting

the number of factors by relating the moments of the sample eigenvalue distribution of the data covariance matrix to the parameters describing the pattern of time and spatial correlations. We show that the moments of the eigenvalue distribution can in fact be derived *analytically* from the model assumptions using properties of noncommutative random variables, thereby greatly simplifying the computational burden. This is the first application of *free probability theory* of noncommutative random variables in econometrics.

Monte Carlo evidence shows that covariance restrictions can greatly improve the finite sample performance in the presence of weak factors, while remaining robust to some degree of misspecification. The method also performs well when applied to the estimation of the number of factors in financial and macroeconomic data.

The rest of the paper is structured as follows. Section 2 introduces the approximate factor model and explains the assumptions underlying our approach. Section 3 presents a consistent procedure for the estimation of the number of factors in high-dimensional models using the moments of the sample eigenvalues of the data covariance matrix. Section 4 shows how free probability theory can be applied to derive the spectral moments analytically from the model assumptions in the presence of both time and spatial correlations. Section 5 compares the finite sample performance of the proposed estimator to that of two competing methods in a series of Monte Carlo designs. Section 6 applies the new estimation procedure to two realistic examples in finance and macroeconomics. Section 7 concludes. Additional background on free probability theory as well as more technical derivations and additional simulation results are found in the online appendix.

2. Approximate Factor Model

Consider the following large (N, T) panel data model with latent factors:

$$R_t = \Lambda F_t + U_t, \tag{1}$$

for $t = 1 \dots T$. R_t is an $N \times 1$ vector of observations, F_t is a $p \times 1$ vector of latent factors, Λ is an $N \times p$ matrix of coefficients (factor loadings) and U_t is an $N \times 1$ vector of idiosyncratic errors. In this model only R_t is observed while Λ, F_t and U_t are unobserved for all t.

The aim of this model is to decompose R_t into a common component ΛF_t and an idiosyncratic component U_t, where the variation in R_t can be explained by a small dimensional set of latent factors F_t. In order to simplify the discussion we shall refer to the cross-sectional dimension N as "individuals", while we let the time-series dimension T denote "periods". Note that the coefficients Λ correspond to individual loadings or weights of the common factors F_t. Let $U = [U_1, U_2, ..., U_T]$ be the $N \times T$ matrix of errors in equation 1. Similarly define the $N \times T$ matrices R and ΛF.

ASSUMPTION 1: *Cross-sectional and Time Dependence: There is a sequence of symmetric positive definite matrices $A_N(\theta_1)$ and $B_T(\theta_2)$, where $\theta = (\theta_1, \theta_2)$ is a parameter vector, such that for all N and T, $U = A_N^{1/2} \epsilon B_T^{1/2}$. Moreover, ϵ has i.i.d. elements, $E(\epsilon_{i,j}) = 0$, $E((\epsilon_{i,j})^2) = 1$, and $E((\epsilon_{i,j})^4) < \infty$.*

Note that $A_N(\theta_1)$ is of dimension $N \times N$ and $B_T(\theta_2)$ is of dimension $T \times T$. We assume that $\dim(\theta)$=S_0, where S_0 is a small finite integer. We can allow $A_N(\theta_1)$ and $B_T(\theta_2)$ to be random or non-random. If they are random, then A_N, B_T and ϵ are independent.

This assumption enables our model to capture a more flexible "approximate factor structure" by allowing for correlations in the idiosyncratic errors (Bai and Ng 2002). Since $\text{Cov}(\text{vec}(U_N)) = B_T \otimes A_N$ for $\text{vec}(U_N) = (U_1' U_2' ... U_T')'$, it restricts the nature of correlations in two important ways. First, it assumes the (N, T) separability of the covariance matrix into an $N \times N$ component A_N and a $T \times T$ component B_T. The matrix A_N captures the cross-sectional dependence between individuals, while B_T captures the form of time dependence and A_N and B_T. It excludes certain forms of dependence such as time varying cross-sectional correlations in the error term. Spatio-temporal covariance models are popular in many areas of statistics (Ma 2003, Stein 2005) and have featured in the recent literature on factor models (Onatski 2010, Onatski 2012). Second, it

assumes that the nature of correlations can be captured by a low dimensional parameter vector θ. This allows the distribution of the error terms to depend on unknown parameters, which are then estimated from the data using a moment based approach.

Discussion: Even though somewhat restrictive this assumption captures the main intuition behind the "approximate factor model". By construction an informative factor model should contain very little structure in the error terms and all important variation should be captured by the factors. Thus, applications of factor models don't typically consider complex structures of the error term, but it is informative to consider departures from a strict diagonal structure for the covariance matrix. It is possible to view this assumption as an approximation, with many desirable properties, to more general covariance structures. In particular it preserves a number of structural features of the original covariance matrix (Van Loan and Pitsianis 1993). Doz, Giannone and Reichlin (2012) consider a quasi-maximum likelihood approach where the true model is an approximate factor which is estimated as an exact factor model. Similarly to our technique, the maximum likelihood approach also has the advantage that it allows for restrictions from economic theory to be imposed.

Few economic models pay close attention to the spatio-temporal structure of the error terms in the context of a factor model. In fact in is typical for models to assume a strict diagonal structure for the covariance of the error terms (Fan, Fan, and Lv 2008). Onatski (2012) provides an example of a factor model for industrial output across industries and time. In this example, the matrix A_N corresponds to the (known) inverse of the input-output matrix for the economy which describes the quantity of input from one industry required to produce one unit of output in a different industry, while B_T captures the auto-regressive nature of production. This assumption does however capture many of the data generating processes discussed in the Econometrics literature on factor models. Chamberlain and Rothschild (1983) discuss an example of an approximate factor model with spatial correlations of lag 1, i.e. $Cov(u_i, u_j) = 0$ if $|i - j| > 1$. In the macroeconometric literature it is common to restrict the errors terms to be AR(1), e.g.

Reis and Watson (2010) or Eickmeier, Lemke and Marcellino (2011). Bai and Ng (2002) and Onatski (2010) use a data generating process where the errors follow an AR(1) process while also exhibiting a small degree of spatial correlation. While in general exclusion restrictions determining the structure of the noise covariance matrix may not be immediately obvious, in some cases they may follow from the structure of the data. For example, when considering a cross-country model similar to that of Diebold et. al. (2008) we may wish to allow for global factors while restricting the idiosyncratic shocks to be country specific.

ASSUMPTION 2: *The number of individuals increases with the sample size. Thus,* $N \to \infty$ *and* $T \to \infty$ *and* $N/T \to c \in (0, \infty)$.

This assumption captures the underlying asymptotic framework and is familiar to the literature on large N, large T panel data (Hahn and Kuersteiner 2002). We depart however from the traditional econometric approach to solving factor models, which operates under the assumption that N is a fixed small number while T is large (Goldberger 1972, Zellner 1970). With the availability of high dimensional panel data where both N and T are large, this alternative asymptotic framework has received renewed attention and is currently an active area of research (Amengual and Watson 2007, Bai and Ng 2002, Onatski 2009, Onatski 2010, Onatski 2012). The use of the limit $N/T \to c \in (0, \infty)$ has recently also received substantial interest in many areas of the statistical literature (Johnstone 2007) and is central to the field of Random Matrix Theory.

Consider the spectral decomposition of an arbitrary symmetric $n \times n$ matrix $C_n = V'DV$. where V is the eigenvector matrix with columns orthogonal to each other and the matrix $D = \text{diag}\{\lambda_1, \lambda_2, ..., \lambda_n\}$ contains the set of eigenvalues of the matrix C_n. We can define the following proper cumulative distribution function $F^{C_n}(\lambda)$ on the spectrum $\lambda_i \in \{\lambda_1, \lambda_2, ..., \lambda_n\}$ of C_n: $F^{C_n}(\lambda) = \frac{1}{n}\{\text{Number of Eigenvalues of } C_n \leq \lambda\} = \frac{1}{n}\sum_{\lambda_i \leq \lambda} 1$.

Note that the spectrum does count multiplicities of eigenvalues. We will label this distribution as the *empirical eigenvalue distribution*. A sequence of $n \times n$ matrices C_n is said to

have an *asymptotic or limiting eigenvalue distribution* if $\lim_{N\to\infty} \mathrm{tr}(C_n^k)$ exists for all $n \in \mathbb{N}$ (Speicher 2005).

ASSUMPTION 3: *Bounded Spectrum: Consider the sequence of matrices A_N and B_T. For any N and T denote by F^{A_N} and F^{B_T} the eigenvalue distributions of A_N and B_T respectively.*

(1) *If $||\,\mathrm{Sp}(A_N)||$ and $||\,\mathrm{Sp}(B_T)||$ denote the spectral norms of A_N and B_T, then both spectral norms are bounded in N and T respectively;*
(2) *As $N \to \infty$ and $T \to \infty$, A_N and B_T have non-random asymptotic eigenvalue distributions (with bounded support) denoted by F^A and F^B.*

This assumption implies that for all N and T the empirical eigenvalue distributions of the matrices A_N and B_T are bounded in N and T. Moreover, the empirical eigenvalue distributions converge *in distribution* to non-random limiting distributions.

This assumption is required in order to guarantee that the eigenvalue distribution of the covariance matrix of the error terms in the factor model, $T^{-1}\sum_{t=1}^{T} U_t U_t'$, converges to a non-random distribution with bounded support (Silverstein 1995, Silverstein and Bai 1995, Bai and Silverstein 1998).[3] This assumption is relatively mild and jointly with Assumption 1 rules out features of the distribution of the error terms which would prevent the identification of latent factors from the noise. Furthermore, it guarantees that the moments of the empirical eigenvalue distribution converge almost surely to non-random quantities, which will prove useful in counting the number of factors..

ASSUMPTION 4: *Pervasive factors: Assume that the factors F_t are independent of U_t. Denote by $\mu_0 = \min\{\mathrm{Sp}(\frac{1}{T}\sum_{t=1}^{T} \Lambda F_t F_t' \Lambda')\}$ the smallest non-zero eigenvalue of the covariance of the*

[3]Note that for the special case where $u_{i,t}$ are i.i.d. random variables with mean 0 and variance 1, finite fourth moment and $N < T$, the asymptotic eigenvalue distribution of the covariance matrix of the error terms corresponds to the celebrated (Marcenko and Pastur 1967) distribution with support bounded on $[(1-\sqrt{c})^2, (1+\sqrt{c})^2]$ as discussed in (Harding 2008).

factors weighted by the factor loadings. Then there is some $M > 0$ *and* $M \to \infty$ *such that* $\mu_0 \geq M$ *as* $N \to \infty$.

This assumption places relatively weak restrictions on the nature of the latent factors and factor loadings. The factors are independent of the noise and the factor loadings are not restricted to being either random or fixed. This assumption clarifies what we typically mean by a "pervasive factor" by requiring that the latent factors impact at least a fraction of the individuals, where the fraction increases with the sample size. Thus, even factors which are relatively weak can be identified in large samples due to their effect on a large number of individuals. For example, we think of pervasive factors in a model of US stocks as supply and demand shocks which impact a large number of firms simultaneously.

3. Counting the Number of Factors

Let $\Sigma = E(RR')$, $\Xi = E(\Lambda FF'\Lambda')$, and $\Omega = E(UU')$ and denote by Σ_N, Ξ_N and Ω_N their sample equivalents: $\Sigma_N = \frac{1}{T}\sum_{t=1}^{T} R_t R_t'$, $\Xi_N = \frac{1}{T}\sum_{t=1}^{T} \Lambda F_t F_t' \Lambda'$, $\Omega_N = \frac{1}{T}\sum_{t=1}^{T} U_t U_t'$. Then by the assumptions of our model we have that $\Sigma = \Xi + \Omega$[4]. The traditional spectral approach to counting the number of factors relies on the observation that by the assumptions of a factor model $\text{rank}(\Xi_N) = p$ and $\text{rank}(\Omega_N) = N$. Hence, the covariance matrix of the observations can be thought of as a p rank perturbation Ξ_N of the idiosyncratic noise covariance Ω_N and we would expect to observe a large number of small eigenvalues and p large eigenvalues. In finite samples, however, it has long been established that accurate estimation of the number of factors using heuristic approaches such as the visual inspection of the scree plot is very difficult, in spite of the popularity of such methods in applied research (Brown, 1989). This is particularly problematic in the case of "weak factors" (Onatski 2012).

[4]Note that we can ignore the cross product terms $\Lambda\left(\frac{1}{T}\sum_{t=1}^{T} F_t U_t'\right)$ and its transpose. We have $\frac{1}{T}\sum_{t=1}^{T} F_t U_t' \to \frac{1}{T}\sum_{t=1}^{T} E(F_t U_t')$ in probability. By Assumption 4 on the independence of the factors and the error term we can ignore this term in large samples.

In this section we introduce a novel approach to counting the number of factors in large factor models, by relying on recent advances in the field of *Random Matrix Theory*[5]. Our approach gives rise to a stopping rule which can be used to consistently count the number of latent factors.

3.1. Preliminaries

Define the Cauchy Transform of an eigenvalue distribution function F^C for some matrix C as: $G_C(\omega) = \int \frac{1}{\omega-\lambda}\, dF^C(\lambda) = \frac{1}{N}\text{tr}\left((\omega I_N - C)^{-1}\right)$, for $w \in \mathbb{C}^+$. This analytic function plays an important role in many random matrix theory results where it serves as an analogue to the Fourier transform in traditional probability theory. For a sample covariance matrix C_N we can define the raw moments of its eigenvalue distribution by: $m^s_{C_N}(\lambda) = N^{-1}\,\text{tr}[(C_N)^s]$. Denote by F^C the limiting distribution of the sample eigenvalue distribution F^{C_N} of C_N. Standard results on bounded moment convergence and continuous mapping (e.g. Billingsley, 1995) imply that if F^{C_N} converges in distribution to F^C, a proper probability distribution with bounded support, then: $\lim_{N\to\infty} m^s_{C_N}(\lambda) = m^s_C(\lambda) = \int \lambda^s dF^C(\lambda) < \infty$.

In order to simplify mathematical notation we restrict our attention to the case where $B_T = I_T$. Furthermore, notice that without loss of generality, we let $0 < c \leq 1$, since the non-zero eigenvalues of CC', for some $N \times T$ dimensional matrix C are the same as the eigenvalues of $C'C$. The remaining $T - N$ eigenvalues of $C'C$ are all zero. First consider the limiting distribution of the eigenvalues of the noise covariance matrix Ω_N defined above:

PROPOSITION 1: *(Silverstein) As $N \to \infty$, $T \to \infty$, and $N/T \to c$, if $F^{A_N} \to F^A$ almost surely in distribution to a pdf on $[0, \infty)$, then the sample eigenvalue distribution $F^{\Omega_N}(\lambda)$ converges to a non-random asymptotic distribution function $F^{\Omega}(\lambda; F^A, c)$ with bounded support.*

[5]See Edelman and Rao (2005) for an introductory treatment.

The proof is given in Silverstein (1995). Assumptions 1, 2 and 3 are sufficient to obtain the result in Silverstein (1995).[6] Let us now focus our attention on the moments $m^s_{\Omega_N}$ of the sample eigenvalue distribution and the moments $m^s_{\Omega}(\theta) = \lim_{N\to\infty}(1/N)\,E\{\mathrm{tr}(\Omega^s_N)\}$ of the limiting eigenvalue distribution.

As a consequence of Proposition 1 the moments of the limiting eigenvalue distribution of the noise covariance matrix exist and are finite. Moreover, it can be shown that the moments satisfy a Central Limit Theorem (Bai and Silverstein, 2004):

PROPOSITION 2: *(CLT) Let* $\overline{g}(w) = -(1-c)/w + cG_{\Omega}(w)$. *Then*

$$N^{-1}\begin{pmatrix} m^1_{\Omega_N} & - & m^1_{\Omega}(\theta) \\ & ... & \\ m^s_{\Omega_N} & - & m^s_{\Omega}(\theta) \end{pmatrix} \sim N(\Delta, V), \tag{2}$$

where for $j = 1 \ldots s$ *and* $k = 1 \ldots s$

$$\Delta_j = -\frac{1}{2\pi i}\int w^j \frac{c\int \overline{g}(w)^3 v^2 (1+v\overline{g}(w))^{-3} dF^A(v)}{\left(1 - c\int \overline{g}(w)^2 v^2 (1+v\overline{g}(w))^{-2} F^A(v)\right)^2} dw$$

$$V_{jk} = -\frac{1}{2\pi^2}\int\int \frac{w_1^j w_2^k}{\left(\overline{g}(w_1) - \overline{g}(w_2)\right)^2}\frac{d\overline{g}(w_1)}{dw_1}\frac{d\overline{g}(w_2)}{dw_2} dw_1 dw_2,$$

where the contours are assumed to be non-overlapping, closed, taken in the positive direction in the complex plane, each enclosing the support of F^{Ω}.

This CLT was introduced by Bai and Siverstein (2004). It is discussed at length in Chapter 9 of Bai and Silversein (2010). The proof relies on Assumptions 1, 2 and 3. Furthermore, Bai and Silverstein (2004, 2010) assume that $E((\epsilon_{i,j})^4) = 3$. Silverstein (1989) has shown that without this condition it is possible that there exist sequences of random matrices for which

[6] Notice that given the assumptions of our model, the limiting eigenvalue distribution of Ω_N depends only on c and θ and can be obtained (implicitly) using the Cauchy transform. As Edelman and Rao (2005) show, obtaining the distribution numerically remains computationally challenging and requires advanced numerical methods.

the first two moments do not converge. Pan and Zhou (2008) show that this condition can be relaxed (Condition 6 in Theorem 1.3) by imposing restrictions on the matrix A_N in relation to the Cauchy transformation of the limiting eigenvalue distribution. This condition is difficult to evaluate but it can be shown that it is satisfied in some cases, e.g. when A_N is diagonal.

EXAMPLE 1: *In general it is not possible to analytically evaluate the integral expressions above over the complex plane. For $A_N = I_N$ the answer however is known and was derived by Jonsson (1982) using a combinatoric proof. In this case the expressions above reduce to:*

$$\Delta_j = \frac{1}{4}\left(\left(1-\sqrt{c}\right)^{2j} + \left(1+\sqrt{c}\right)^{2j}\right) - \frac{1}{2}\sum_{r=0}^{j}\binom{j}{r}^2 c^r$$

$$(V)_{j,k} = 2c^{j+k}\sum_{r_1=0}^{j-1}\sum_{r_2=0}^{k}\binom{j}{r_1}\binom{k}{r_2}\left(\frac{1-c}{c}\right)^{j+k}\sum_{l=1}^{j-k} l\binom{2j-1-(r_1+l)}{j-1}\binom{2k-1-(r_2+l)}{k-1}$$

3.2. Identification and Estimation

In this section we introduce a new approach to estimating the number of latent factors in large dimensional models which simultaneously estimates the unknown parameters θ and also counts the number of latent factors p. Let $\mathrm{Sp}(\Sigma_N)$ denote the spectrum of the covariance matrix Σ_N of the individual units over time, where we have ordered the eigenvalues in decreasing order. That is, $\mathrm{Sp}(\Sigma_N) = \{\lambda_1, \lambda_2, ..., \lambda_N\}$, with λ_1 being the largest eigenvalue and including multiplicities. Now let $\mathrm{Sp}_p(\Sigma_N)$ be the p-truncated spectrum where we have removed the first p largest eigenvalues, $\mathrm{Sp}_p(\Sigma_N) = \mathrm{Sp}(\Sigma_N)\backslash\{\lambda_1, \lambda_2, ..., \lambda_p\}$. Let $\Pi_p(\Sigma_N)$ be the vector of the first K moments of the sample eigenvalue distribution evaluated from the truncated spectrum $\mathrm{Sp}_p(\Sigma_N)$, $\Pi_p(\Sigma_N) = \left[N^{-1}\sum_{\lambda_j\in\mathrm{Sp}_p(\Sigma_N)}\lambda_j^1, N^{-1}\sum_{\lambda_j\in\mathrm{Sp}_p(\Sigma_N)}\lambda_j^2, \ldots, N^{-1}\sum_{\lambda_j\in\mathrm{Sp}_p(\Sigma_N)}\lambda_j^K\right]'$. Note that the above definition allows for $p = 0$, in which case $\Pi_0(\Sigma_N)$ represents the vector corresponding to the first K moments of the sample eigenvalue distribution. Further, denote by $\Pi(\theta) = \left[m_\Omega^1(\theta, c), m_\Omega^2(\theta, c), ..., m_\Omega^s(\theta, c)\right]'$ the vector of moments of the limiting eigenvalue distribution

of Ω_N as $N \to \infty$, $T \to \infty$ and $N/T \to c$. In Section 4 we show how to derive expressions for these limiting moments analytically as functions of θ.

Proposition 2 suggests a minimum distance procedure for estimating the vector of unknown covariance parameters θ from the moments of the sample covariance matrix:

$$\hat{\theta}_p = \text{argmin}_\theta \underbrace{(\Pi(\theta) - \Pi_p(\Sigma_N))' \hat{V}^{-1} (\Pi(\theta) - \Pi_p(\Sigma_N))}_{J(\theta, \text{Sp}_p(\Sigma_N))}, \tag{3}$$

where $(\hat{V})_{j,k}$ is a consistent estimator of the covariance matrix given in equation 2. Two step estimation may be performed by using the estimated parameters from a first step with $V = I$. Note that the p-truncated spectrum is used in the above expression. Thus, a sequence of estimates $\hat{\theta}_0, \hat{\theta}_1, \ldots, \hat{\theta}_p, \ldots$ is obtained as an increasing number of (large) eigenvalues is consecutively removed from the spectrum of the sample covariance matrix.

Now consider the J-statistic corresponding to the minimum distance function defined in 3, $\hat{J}(\hat{\theta}, \text{Sp}_p(\Sigma_N))$, computed for the p-truncated spectrum of the sample covariance matrix Σ_N. We can consistently estimate the number of latent factors from the first K moments of the sample covariance matrix from the following stopping rule:

PROPOSITION 3: *(Stopping Rule). Consider a large dimensional factor model, satisfying Assumptions 1-4 above. Additionally, let* $M = O\left(N^{\frac{2}{K}+\alpha}\right)$, *for some* $\alpha > 0$. *Then,*

$$\widehat{p} = \min\left\{ q : \frac{\widehat{J}(\hat{\theta}, Sp_q(\Sigma_N))}{M^{2K}} \to 0 \right\} \tag{4}$$

provides a consistent estimate of p_0, *the true number of factors.*

Proof: See Appendix. Notice that we need, $N^2/M^K \to 0$ as $N \to \infty$. So, $M = N^{2/K} \log N$ can be a choice of M provided it also satisfies Assumption 4. If the number of moments K is sufficiently large, then in most cases Assumption 4 is sufficient.

Discussion. The intuition behind our stopping rule is that in large samples, we should observe a separation of the spectrum of Σ_N into two parts, a first part with mass $(N-p)/N$

located to the left but bounded by zero from below and a second part with mass p/N to the right which diverges as $N \to \infty$ (Bai and Silverstein, 1999, 2004, 2010; Baik and Silverstein, 2006; Dozier and Silverstein, 2007). Estimating the number of factors thus involves counting the number of eigenvalues to the right of this spectral gap.

In finite samples however this gap may not be easy to locate, especially in models with weak factors and cannot be accomplished by visual inspection of the scree plot. The stopping rule introduced above estimates the number of latent factors consistently as $N \to \infty$. In small samples, when Assumption 4 does not hold, some of the eigenvalues of Σ_N, corresponding to weak factors may undergo a "phase transition", as described in Harding (2008), and converge in probability to the upper bound of the spectrum of Ω_N. In such a case, any eigenvalue based method will produce severely biased results. Harding (2008) shows how this leads to a single factor bias in estimated arbitrage pricing models, when applied to small samples.

4. Using Free Probability to Derive the Spectral Moments

Implementing equation 3 requires us to know the moments $m^s_\Omega(\theta)$ of the limiting eigenvalue distribution of the noise term Ω_N, as functions of the model parameters θ. This could be approached through numerical methods but the computational burden is greatly diminished if these moments were known in analytic form. Deriving these moment expressions using standard combinatorial principles is extremely challenging and has not been attempted in the literature except in very simple cases . In this section we show that, given the assumptions of our model, it is however possible to do so by employing insights from the recent mathematical field of *Free Probability Theory*.[7]

[7]Some elementary results in free probability theory as well as detailed proofs of the theorems in this section are found in the Online Appendix to this paper. More details on free probability theory and its connections to combinatorics can be found in the monograph by Nica and Speicher (2006).

For simplicity we start by limiting our attention to the case where $B_T = I_T$ and discuss the general case at the end of this section. From Assumption 1 it follows that $\Omega_N = (1/T)UU' = (1/T)A_N^{1/2}\epsilon\epsilon' A_N^{1/2}$. Let $\Psi_N = (1/T)\epsilon\epsilon'$ and notice that:

$$m_\Omega^s = \lim_{N\to\infty}\frac{1}{N}E\left\{\text{tr}\left[\frac{1}{T}A_N^{1/2}\epsilon\epsilon' A_N^{1/2}\right]^s\right\} = \lim_{N\to\infty}\frac{1}{N}E\left\{\text{tr}\,\frac{1}{T}[A_N\Psi_N]^s)\right\}. \tag{5}$$

The aim of this section is to introduce a procedure to analytically derive the moments m_Ω^s from on our knowledge of the moments of the limiting eigenvalue distributions of A_N and Ψ_N, which will greatly simplify our computations.

First we need to define the core concepts of a non-commutative probability space and that of *free independence* as an extension of the concept of independence in classical probability theory.[8]

DEFINITION 1: *A non-commutative probability space is a pair $(\mathcal{A}, \phi)$, where $\mathcal{A}$ is an algebra over $\mathbb{C}$ endowed with a unit (1) and ϕ a linear functional on $\mathcal{A}$, $\phi :\to \mathbb{C}$ such that $\phi(1) = 1$.*[9] *An element $X \in \mathcal{A}$ is referred to as a non-commutative random variable in $(\mathcal{A}, \phi)$.*

For the non-commutative probability space $(\mathcal{A}, \phi)$ and a random variable $X \in \mathcal{A}$, we can define the s-th moment of X as $m_X^s = \phi(X^s)$. Computing expectations over random variables in classical probability is often simplified when we can assume independence between the random variables. We now extend the notion of independence in classical probability by employing the concept of freeness or free independence from operator algebras (Voiculescu 1985, Voiculescu 1998, Nica and Speicher 2006).

DEFINITION 2: *(Free Independence):* Let $(\mathcal{A}, \phi)$ be the non-commutative probability space of Definition 1 and $\{\mathcal{A}_1, \mathcal{A}_2, ..., \mathcal{A}_J\} \subset \mathcal{A}$ subalgebras of $\mathcal{A}$ with a unit (1). Then the algebras $\mathcal{A}_1, \mathcal{A}_2, ..., \mathcal{A}_J$ are free with respect to ϕ, if $\phi(a_1 a_2 ... a_K) = 0$ *for* $a_1 \in \mathcal{A}_{j(1)}$, $a_2 \in \mathcal{A}_{j(2)}$, ... ,

[8] This algebra is typically referred to as a von Neumann algebra. See Nica and Speicher (2006) for monograph exposition on free probability theory.

[9] Note that (scalar) classical probability spaces also satisfy the above definition while also assuming commutativeness.

$a_K \in \mathcal{A}_{j(K)}$, where $j(k)$ is an index function on the set $\{1, 2, ..., J\}$ and $j(k) \neq j(k+1)$ for all $k = 1..(K-1)$, and $\phi(a_k) = 0$ for all $k = 1 \ldots K$.

By extension, the random variables $X_1, X_2, ..., X_K \in \mathcal{A}$ are freely independent if the subalgebras generated by them are free with respect to ϕ. Thus, the operator concept of freeness is a particular generalization of the classical probability concept of independence, where freeness with respect to ϕ corresponds to independence of σ-algebras, and free independence of random variables corresponds to the classical notion of independence of random variables.

EXAMPLE 2: *Free independence is a convenient property which amounts to an iterative procedure for computing mixed moments $m^s_{ab} = \phi((ab)^s)$ of products of non-commutative random variables from the moments of the constituent random variables:*

$$m^1_{ab} = \phi(ab) = \phi(a)\phi(b). \tag{6}$$

$$m^2_{ab} = \phi((ab)^2) = \phi(abab) = \phi^2(a)\phi(bb) + \phi^2(b)\phi(aa) - \phi^2(a)\phi^2(b). \tag{7}$$

Notice that while the expression for the first moment is the same as the one we would obtain if a and b were independent random variables in a classical (commutative) probability space, the expression for the second moment does not reduce to the same expression one would obtain if a and b were independent commutative random variables in a classical probability space, since $\phi(abab) \neq \phi(a^2b^2) = \phi(a^2)\phi(b^2)$. The definition of free independence can thus be applied recursively to obtain the mixed higher order moments of ab and other similar products. Additional details are provided in the Online Appendix.

Consider the space of $N \times N$ real matrices $\mathcal{M}_N(\mathbb{R})$ and X a random matrix on this space whose elements $(X)_{i,j}$ are random variables on a classical probability space $(\Theta, \mathfrak{P})$. Define the algebra of functions $\mathcal{A}_N = \bigcap_{1 \leq s < \infty} L^s(X, \mathcal{M}_N)$, for the s-integrable random matrices of dimension $N \times N$ for $1 \leq s < \infty$. Note that this implies that all elements $(X)_{i,j}$ have finite moments since

$(X)_{i,j} \in \mathcal{A}_N$. Furthermore, let $\phi_N : \mathcal{A} \to \mathbb{R}$ be an operator defined as: $\phi_N(Y) = \frac{1}{N} E_X \operatorname{tr}(Y) = \frac{1}{N} \sum_{j=1}^{N} E(Y_{j,j}) = \frac{1}{N} \int_X \operatorname{tr}(Y) d\mathfrak{P}$.

PROPOSITION 4: *$(\mathcal{A}_N, \phi_N)$is a non-commutative probability space.*

This result follows immediately as a particular instance of Definition 1. It implies that we can think of covariance matrices as random variables defined on a non-commutative probability space. This greatly facilitates our ability to compute moments of the type $\phi_N[(A_N \Psi_N)^s]$, if it can be shown that (A_N, Ψ_N) are indeed freely independent.

CLAIM 1: *(A_N, Ψ_N) are freely independent if A_N satisfies Assumption 1 and Ψ_N is "unitarily invariant".*

In the Online Appendix we explore the mathematical concept of unitary invariance, which requires more specialized statistical results, and also discuss general conditions under which two random matrices are asymptotically freely independent . A simple special case of unitary invariance holds when Ψ_N is drawn from a standard Wishart distribution, that is $\Psi_N = (1/T)\epsilon\epsilon'$, where $(\epsilon)_{i,j}$ are distributed iid N(0,1). Then Ψ_N and A_N are freely independent if A_N is a non-random matrix (or a random matrix independent of Ψ_N).

In order to compute the mixed moments $\phi_N[(A_N \Psi_N)^s]$ we require knowledge of the moments of A_N and Ψ_N. From these the mixed moments can be derived by iteration using Definition 2. The moments of A_N follow directly as from any model satisfying Assumption 1 and are known functions of θ. The moments of Ψ_N are given by the following Lemma.

LEMMA 1: *Let ϵ be an $N \times T$ random matrix with elements which are iid with mean 0, variance 1 and finite fourth order moments. Then as $N \to \infty$, $T \to \infty$ and $N/T \to c$, the empirical eigenvalue distribution of $\Psi_N = (1/T)\epsilon\epsilon'$ converges almost surely to the non-random Marcenko-Pastur distribution whose moments are given by:*

$$m_\Psi^s = \lim_{N\to\infty} \frac{1}{N} E \operatorname{tr}\left\{ \left[(1/T)\, \epsilon\epsilon' \right]^s \right] = \sum_{r=1}^{s} \frac{1}{s} \binom{s}{r} \binom{s}{r-1} c^{s-1}.$$

Proof: Jonsson (1982). Note that the moments of $\Psi_N = (1/T)\epsilon\epsilon'$ are the Narayana polynomials in $c = N/T$.

The mixed moments can then be derived from the iterated application of Definition 2, as shown in Example 2. The following proposition provides a general recursive expression for the computation of the mixed moments in terms of the limiting moments of eigenvalue distribution of A_N. Further details are found in the Online Appendix.

PROPOSITION 5: *Let m_Ω^s be the limiting moments of the sample noise covariance Ω and m_A^s be the limiting moments of the matrix A_N. Then for $w \in \mathbb{C}^+$ with $\Im\mathfrak{m}(w) > 0$ we have that*

$$\sum_{s=1}^{\infty} \frac{m_\Omega^s}{w^s} = \sum_{s=1}^{\infty} \frac{m_A^s}{w^s} \left(1 + c \sum_{r=1}^{\infty} \frac{m_\Omega^r}{w^r} \right)^s. \tag{8}$$

Proof: See Online Appendix. Explicit expressions for the moments can easily be computed by expanding this expression in $1/w$ and matching the coefficients on $1/(w^s)$.

EXAMPLE 3: *If $A_N = \sigma I_N$ then:*

$$\begin{aligned} m_\Omega^1 &= \sigma \\ m_\Omega^2 &= (1+c)\,\sigma^2 \\ m_\Omega^3 &= \left(c^2 + 3c + 1\right)\sigma^3 \\ m_\Omega^4 &= (1+c)\left(c^2 + 5c + 1\right)\sigma^4. \end{aligned}$$

4.1. Time-Series Correlations

Let us now consider the case where $A_N = I_N$ and $B_T \neq I_T$. We are interested in $m_\Omega^s = \lim_{N\to\infty} (1/N)\, E\left\{\mathrm{tr}([(1/T)\,\epsilon B_T \epsilon']^s)\right\}$. Notice, however, that the non-zero eigenvalues of $(1/T)\epsilon B_T \epsilon'$ and $(1/T)\epsilon'\epsilon B_T$ are the same. We can now apply Definition 2 to compute the mixed moments

of $(1/T)\epsilon'\epsilon$ and B_T. The general relationship can be summarized by the following recursive expression:

PROPOSITION 6: *Let m^s_Ω be the limiting moments of the spectral distribution of the sample noise covariance Ω and m^s_B be the limiting moments of the spectral distribution of the matrix B_T. Then for $w \in \mathbb{C}^+$ with $\mathfrak{Im}(w) > 0$ we have that*

$$\sum_{s=1}^{\infty} \frac{m^s_\Omega}{w^s} = \frac{1}{c}\left\{\sum_{s=1}^{\infty} \frac{m^s_B}{w^s}\left[c\left(1 + \sum_{r=1}^{\infty} \frac{m^r_\Omega}{w^r}\right)\right]^s\right\}. \tag{9}$$

Proof: See Online Appendix.

EXAMPLE 4: *Assume that the idiosyncratic errors follow an AR(1) process $U_{j,t} = \rho U_{j,t-1} + \epsilon_{j,t}$, for $\epsilon_{j,t}$ white noise such that $E(\epsilon_{j,t}) = 0$ and $E(\epsilon^2_{j,t}) = \sigma^2$. Then,*

$$\begin{aligned}
m^1_\Omega &= m^1_B \\
m^2_\Omega &= cm^2_B + \left(m^1_B\right)^2 \\
m^3_\Omega &= c^2 m^3_B + 3cm^2_B m^1_B + \left(m^1_B\right)^3 \\
m^4_\Omega &= c^3 m^4_B + 4c^2(m^3_B m^1_B + \frac{1}{2}\left(m^2_B\right)^2) + 6cm^2_B\left(m^1_B\right)^2 + \left(m^1_B\right)^4,
\end{aligned}$$

where for any positive integer s,

$$m^s_B = \frac{1}{2\pi}\int_0^{2\pi}\left[\frac{1-\rho^2}{1 - 2\rho\cos(\zeta) + \rho^2}\right]^s d\zeta. \tag{10}$$

It is also possible to consider models where both $B_T \neq I_T$ and $A_N \neq I_T$. In such cases the sample noise covariance matrix is given by: $\Omega_N = \frac{1}{T}A_N^{1/2}\epsilon B_T \epsilon' A_N^{1/2}$. The corresponding moment conditions can be obtained by first computing the moments of the eigenvalue distribution of $\Psi_N = \frac{1}{T}\epsilon B_T \epsilon'$ using Proposition 5 and then computing the moments of the eigenvalue distribution of the product $A_N\Psi_N$ by employing Proposition 6.

5. Monte Carlo

In this section we investigate the finite sample performance of the proposed method and compare it to the two leading competing estimation approaches.[10] We also investigate the robustness of the method to misspecification of the restrictions on the error covariance matrix. We generate Monte Carlo data based on a design similar to that of Bai and Ng (2002) and Onatski (2010): $R_t = \Lambda F_t + U_t,$ where $U = A_N^{1/2} \epsilon B_T^{1/2}$, for $t = 1 \dots T$. R_t is an $N \times 1$ vector of observations, F_t is a $p \times 1$ vector of latent factors, Λ is an $N \times p$ matrix of coefficients (factor loadings) and U_t is an $N \times 1$ vector of idiosyncratic errors. We set the number of latent factors $p_0 = 5$ and draw $\epsilon_{j,t} \sim$ i.i.d. $N(0,1)$ and $F_{j,t} \sim$ i.i.d. $N(0,1)$.

Factor loadings are generated as follows:

$$\Lambda_{j,k} = \sqrt{m_1/\sqrt{m_2}} \text{ , for } j = 1 \dots p_0, \tag{11}$$

$$\Lambda_{j,k} = a\sqrt{m_1/(N-k)} \text{ , for } j = p_0 + 1 \dots N \text{ ,} \tag{12}$$

for $k = 1 \dots p_0$. Furthermore, $a = -1$ if $j = rk$ and $a = +1$ if $j \neq rk$ for $r = 1, 2, 3...$.

We consider four designs for $\sigma^2 = 1$ and $\rho = 0.1$:

Design I : $A_N = \sigma^2 I_N$, $B_T = I_T$, $m_1 = 10$, $m_2 = 1$.
Design II : $A_N = \sigma^2 I_N$, $B_T = I_T$, $m_1 = 3$, $m_2 = N$.
Design III: $A_N = \sigma^2/(1-\rho^2) I_N$, $(B_T)_{i,j} = \rho^{|i-j|}$, $m_1 = 10$, $m_2 = 1$.
Design IV : $A_N = \sigma^2/(1-\rho^2) I_N$, $(B_T)_{i,j} = \rho^{|i-j|}$, $m_1 = 3$, $m_2 = N$.

[10] We limit our attention to alternative approaches from the recent econometric literature. In a recent publication, Passemier and Yao (2012) compare the procedure as introduced in an earlier version of this paper with several other statistical approaches for determining the number of "spikes" in the spectrum of high-dimensional models. They find that the current procedure performs at least as well as the alternatives they consider and in fact has lower MSE.

For each design we vary N from 30 to 450 and T from 100 to 500. The resulting ratio $c = N/T$ varies from 0.3 to 0.9 and we draw 1000 samples for each combination.

We estimate the unknown model parameters p_0, σ^2 and ρ using different estimation techniques. The first estimator (MD) corresponds to the estimation procedure resulting from equations 3 and 4 and setting the moment weighting matrix $\hat{V} = I_K$. Throughout we use $K = 4$ moment conditions. For Designs 1 and 2 the moment weighting matrix V is known and given in Example 1. For these two designs we also report results from a weighted two-step estimator (MDW), which is obtained by first deriving consistent estimates of the model parameters using the unweighted MD estimator, evaluating the estimated moment weighting matrix $\hat{V}$, and then using this weighting matrix in equation 3.

As discussed in the Appendix, our proposed estimator has the feature that the rate of change in the objective function diminishes with the number of factors estimated. In order to guard against the possibility of over-fitting and in order to improve the finite sample performance of our estimator we can add a panel information criterion, which penalizes the objective function in equation 4 if the selected number of factors is too large. Following Bai and Ng (2002) we can consider penalty functions of the form $q\hat{\sigma}^2 g(N,T)$, where q is the number of factors selected, $\hat{\sigma}^2$ is the estimated (average) variance at each step, and $g(N,T)$ a function such that $g(N,T) \to 0$ in large samples. In simulations we have found the following choice to perform very well: $g(N,T) = \left(\frac{N+T}{NT}\right) \log\left(\frac{NT}{N+T}\right)$.

We can augment the estimators MD and MDW defined above with the above penalty function to obtain two additional estimators which we label MD-IC and MDW-IC. We compare our proposed estimator and its variants described above with its leading competitors in the recent econometric literature as introduced by Onatski (2010) and Bai and Ng (2002).

In the tables below we report the mean number of estimated factors and also the percentage of the replications that result in underestimation and overestimation of the number of factors. The later results are reported in the form x/y where $x\%$ of the replications have resulted in

underestimation and $y\%$ of the replications have resulted in overestimation, and $(100-x-y)\%$ of the replications resulted in the correct estimation of the number of latent factors.

Table 1 compares the performance of the different estimators on samples drawn from Designs I and II. Design 1 features factors which are "strong" relative to the idiosyncratic noise. Both the MDW-IC and the Onatski estimator perform very well. Both estimators recover the correct number of factors with near 100% accuracy. This is to be expected as in this design the largest 5 eigenvalues are well separated from the bulk of the remaining eigenvalues of the covariance matrix and thus eigenvalue based methods are very precise. Both estimators outperform the Bai and Ng estimator which tends to underestimate the number of factors severely. As the sample size increases the Bai and Ng estimator underestimates the number of factors in 100% of the samples. Notice that the MD, MDW and MD-IC estimators exhibit a small upward bias and are dominated by the MDW-IC estimator. By inspecting the simulation output we found that even when over-estimating the number of factors these estimators over-estimate by at most 1 factor.

By contrast Design II captures a setting with weak factors. In this design the largest 5 eigenvalues are not strongly separated from the bulk of the remaining eigenvalues and the corresponding scree plot appears to show a continuum of values. Traditional heuristics based on the visual inspection of the scree plot would not be able to discern a clear separation. This scenario thus also presents challenges for spectrum based estimation techniques. All estimators present some degree of downward bias as they occasionally fail to recover all factors from the data. Nevertheless, the estimators introduced in this paper perform substantially better than their competitors. The MD estimators recover between 4 and 5 factors in each simulation. The Onatski estimator finds less than 5 factors in over 90% of the simulations. The Bai and Ng estimator underestimates in 100% of the cases. Its performance actually deteriorates in large samples where it estimates between 0 and 1 factors.[11]

[11]The Online Appendix contains tables showing the mean bias and MSE for the covariance parameters, which are estimated by the MD estimators at the same time as the number of factors. In all cases the

In order to explore the extent to which the finite sample performance of these estimators deteriorates with the amount of noise, we return to Design I and estimate the number of factors in samples where we progressively increase the idiosyncratic variance σ^2 from 1 to 10. In Figure 1 we plot the estimated number of factors for the Bai and Ng, the Onatski and the MD estimators for each level of the variance. The performance of each estimator decreases as the error variance increases. The Bai and Ng estimator in particular fails to recover a positive number of factors as the noise variance increases. The MD estimator outperforms the performance of the Ontaski estimator throughout, even though both of them underestimate the number of factors.

Table 2 explores the performance of the different estimators using Designs III and IV. These designs allow for the error terms to be autocorrelated. We compare the performance of the MD and MD-IC estimators with that of the Onatski and Bai and Ng procedures. We do not compute the MDW estimator since the weighting matrix is difficult to evaluate from expression 2 in Proposition 2 for this example. In Design III the MD-IC and the Onatski estimators perform very well and outperform the Bai and Ng estimator. All procedures underestimate the number of factors in Design IV. The MD and the Onatski procedures are comparable.

The improved finite sample performance of the MD set of procedures does come however at the cost of imposing the correct specification on the unknown error terms. We now investigate the costs of misspecifying the error term in a series of simulations. Table 3 compares the performance of the MDW-IC estimator derived under the assumption of iid errors (corresponding to Designs I and II) but applied to data drawn from Designs III and IV, where the error term is autocorrelated. If the data is generated using Design III the (misspecified) MDW-IC estimator performs fairly well but over-estimates the number of factors in 3% to 32% of the samples. The mean number of factors is very close to 5 indicating that even when it overestimates the factors it does so by no more than 1 factor. Its performance is inferior to that of the Onatski estimator but superior to the procedure of Bai and Ng. When we apply the misspecified MDW-IC to data

covariance parameters are also recovered with small finite sample bias, which diminishes to zero as the sample size increases.

from Design IV with weak factors the MDW-IC estimator performs very similarly to the Onatski estimator. This is due to the fact that the MDW-IC performs better than the Onatski estimator in finite samples with weak factors which compensates for the bias due to misspecification. Note however that these results are driven by the assumptions of Designs III and IV which assume a relatively weak degree of autocorrelation $\rho = 0.1$. In other simulations we have found that our procedure will break down if the degree of misspecification is substantially larger i.e. $\rho > 0.5$.

In order to further explore the impact of misspecification we introduce two variants of Designs III and IV which also exhibit spatial correlations:

Design V : $u_{i,t} = v_{i,t} + \beta v_{i-j,t}$, $m_1 = 10$, $m_2 = 1$.
Design VI: $u_{i,t} = \rho u_{i,t-1} + v_{i,t} + \beta v_{i-j,t}$, $m_1 = 10$, $m_2 = 1$,

where $\rho = 0.3$ and $\beta = 0.1$ and $v_{i,t} \sim N(0,1)$. Note that while it is perfectly feasible to estimate these models using the techniques described in this paper by accounting for both the time and spatial dependence captured by ρ and β, we will generate samples from the designs above but use the MD estimators derived for Designs I and III instead which assume that $\beta = 0$ while accounting for the presence of time correlations. The estimation results are presented in Table 4. Design V is estimated well by all procedures indicating that if the factors are strong a small degree of spatial correlation is of limited concern. The misspecified MD-IC estimator and the Onatski estimator perform very similarly in this situation. Perhaps surprisingly they are dominated by the Bai and Ng procedure.

To conclude we have found that while the MD estimators are robust to a small degree of misspecification they will underperform when applied to models which exhibit a complicated structure of the error terms without properly accounting for it. When correctly specified the MD estimators will however have superior finite sample performance over that of competing estimators, especially in the presence of weak factors.

6. Applications

We apply the proposed method to the estimation of factor models in two common examples of the factor model methodology. First, we estimate the number of pervasive factors underlying US stock returns, and second, we estimate the number of pervasive factors explaining two large panels of macroeconomic measurements. In all cases we assume that $A_N = \sigma^2/(1-\rho^2)I_N$, $(B_T)_{i,j} = \rho^{|i-j|}$.

We construct a sample of daily stock returns over 1298 trading days from January 2003 to December 2007. Data was downloaded from the CRSP database and we consider stocks included in the S&P Index. In order to avoid issues related to missing data, we exclude all companies which are not traded for more than 1 day in any given month and which are not active over the entire period. This results in a sample of 352 stocks. We use the rates on the 90 day Treasury bill to construct excess returns for all stocks.

We apply the proposed estimator of the number of latent factors to a sequence of samples of increasing length, $T \in \{352, 503, 654, 805, 956, 1107, 1258\}$ trading days. Our procedure estimates 3 factors for $T \in \{352, 503\}$ and 4 factors otherwise. By contrast when we apply the Onatski estimator we find the following sequence of estimated factors corresponding to the different sample length: $\{2, 2, 1, 4, 3, 2, 4\}$. We found the Bai and Ng estimator to be numerically very unstable on this data, but Bai and Ng (2002) report the estimated number of factors to be 2 on a similar dataset. Onatski (2010) summarizes a number of similar analyses conducted by other authors on stock returns data, all of which find between 2 and 5 pervasive factors in US stock returns data.

Factor models are also often employed as a data reduction device, whereby a large set of economic indicators is reduced to a small set of statistical factors. This can be done in absence of a structural economic model and the resulting factors, containing the condensed information, can

then be used to predict other outcomes of interest such as consumption, industrial production or inflation (Stock and Watson, 2005).

First, we estimate the number of factors in the classical dataset of Stock and Watson (2005), which is often used as a benchmark in macroeconomic factor models. This database contains 131 economic and financial indicators for the U.S. at monthly frequency for 368 months ending in December 2003. The number of factors estimated by the procedure outlined in this paper is 7. This number is consistent with earlier studies of the same data. Thus, for example, both Stock and Watson (2005) and Bai and Ng (2007), find, using different methods, the number of factors to be 7 in the same dataset. Applying the Onatski estimator we only find 1 factor in this data.

Second, we estimate the number of factors in an alternative dataset of economic indicators for the Euro area which has recently been made available by Angelini et. al. (2011). The data is similar to that of Stock and Watson (2005) and consists of monthly time series collected between 1993 and 2007 on a number of macroeconomic and financial indicators characterizing the Euro area. The number of factors estimated in this dataset is 5. The Onatski estimator applied to the same data does not find any pervasive factors.

It is important to interpret these estimated factors with caution in the absence of an associated economic model as it is difficult to fully understand the economic nature of these estimated factors. From a statistical point of view however, data reduction remains a valuable tool in condensing large datasets and may provide valuable information when used carefully within the confines of economic theory.

7. Conclusion

In this paper we introduce a new econometric approach to the consistent estimation of the number of factors in large factor models. We show that by imposing restrictions on the pattern

of time and spatial correlations of the errors we can derive a consistent spectral estimator for the number of factors with superior finite sample performance in the presence of weak pervasive factors.

The proposed stopping rule for counting the number of factors uses the moments of the sample eigenvalue distribution of the data covariance matrix. Using free probability theory for noncommutative random variables we show that these moments can be derived analytically, thus avoiding the need for a computationally burdensome simulated moments approach. The new approach performs very well in a series on Monte Carlo designs and outperforms in many cases the leading competing approaches currently employed in the econometric literature.

We also applied our estimation procedure to a series of realistic examples in finance and macroeconomics which leads to plausible estimates of the number of factors, consistent with the previous economic literature.

APPENDIX

Proposition 3 proposed the following stopping rule for the consistent estimation of the number of latent factors:

$$\widehat{p} = \min\left\{ q : \frac{\widehat{J}(\hat{\theta}, \mathrm{Sp}_q(\Sigma_N))}{M^{2K}} \to 0 \right\} \tag{A.1}$$

where M is as defined in Assumption 4 and K is the number of moments used. Additionally, we assumed that M depends on N , $M = O\left(N^{\frac{2}{K}+\alpha}\right)$, for some $\alpha > 0$. Actually we need, $N^2/M^K \to 0$ as $N \to \infty$. So, $M = N^{2/K} \log N$ can be a choice of M provided it also satisfies Assumption 4. If the number of moments K is sufficiently large, then in most cases Assumption 4 is sufficient. We will show that the stopping rule given by equation(A.1) will give us a consistent estimate of p, the true number of factors.

Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq M$ be the eigenvalues of the sample factor covariance matrix and $s_1 \geq s_2 \geq \ldots s_N$ be the eigenvalues of the sample noise matrix $T^{-1}\sum_{t=1}^{T} U_t U_t'$. We do not observe λ_i's or s_i's separately, but observe the eigenvalues of their combined covariance matrix Σ_N.

Let us assume initially that the unknown parameters $\hat{\theta}$ were known and did not need to be estimated. In this setting we prove the following Lemma.

LEMMA 2: *For any given parameter θ as $N, T \to \infty$, we have:*

$$\textit{(a)} \quad \frac{\widehat{J}(\theta, Sp_p(\Sigma_N))}{M^{2K}} \to 0.$$

$$\textit{(b)} \quad \frac{\widehat{J}(\theta, Sp_q(\Sigma_N))}{M^{2K}} \geq 1 \quad \textit{for all } q < p.$$

$$\textit{(c)} \quad \frac{\widehat{J}(\theta, Sp_p(\Sigma_N)) - \widehat{J}(\theta, Sp_{p+j}(\Sigma_N))}{M^{2K}} \leq 0 \textit{ for } j \geq 1.$$

$$\textit{(d)} \quad \frac{\widehat{J}(\theta, Sp_{q-1}(\Sigma_N)) - \widehat{J}(\theta, Sp_q(\Sigma_N))}{M^{2K}} \geq 3 \textit{ for all } q < p.$$

where p is the true number of factors in the model and all the above four conditions hold in probability under θ.

The Lemma shows that scaled appropriately, the distance statistics $\widehat{J}(\theta, \mathrm{Sp}_q(\Sigma_N)$ has a detectable decay at p, the true number of factors. From part (a) and (b) of the Lemma, it follows that the stopping rule defined by equation (A.1) is consistent. The lemma also shows that the successive (in factors) difference in the distance statistics (with the same scaling) is significant up to p, as from the first two parts of the lemma we also have: $M^{-2K}[\widehat{J}(\theta, \mathrm{Sp}_{p-1}(\Sigma_N)) - \widehat{J}(\theta, Sp_p(\Sigma_N))] \geq 1$ and becomes insignificant thereafter. These properties reflect the nature of the recovery (detection) of p (true number of factors) and can help in constructing penalty functions. However, any choice of penalty would also depend on the estimates of the unknown parameters θ (or their true values if they were known).

Now we prove Lemma 2.

Proof. (a) We define the sample moments of the noise matrix as $m_S^r = \frac{1}{N}\sum_{i=1}^N s_i^r$ for $r = 1, 2, \ldots, K$. Also, let $I_1 = N^4 \sum_{r=1}^K \left(\frac{m_\Omega^r(\theta) - m_S^r}{N}\right)^2$. Now, note by proposition 2, the difference of the sample and population moments of the noise matrix has an asymptotic multivariate normal distribution

$$\frac{1}{N}\begin{pmatrix} m_\Omega^1(\theta) - m_S^1 \\ \ldots \\ m_\Omega^K(\theta) - m_S^K \end{pmatrix} \to N(\Delta(c), V(c)).$$

And so, $N^{-2}\sum_{r=1}^K (m_\Omega^r(\theta) - m_S^r)^2$ is stochastically bounded.
Also, by our assumptions on M, we have $M^{-2K}N^4 \to 0$. So, $M^{-2K}I_1 \to 0$ in probability.

Now, consider

$$\begin{aligned}
\widehat{J}(\theta, \mathrm{Sp}_p(\Sigma_N)) &= \sum_{r=1}^K \left(N m_\Omega^r(\theta) - \sum_{i=p+1}^N s_i^r \right)^2 \\
&= \sum_{r=1}^K \left(N m_\Omega^r(\theta) - \sum_{i=1}^N s_i^r + \sum_{i=1}^p s_i^r \right)^2 \\
&= I_1 + \sum_{r=1}^K \left(\sum_{i=1}^p s_i^r \right)^2 + 2N \sum_{r=1}^K \left(\sum_{i=1}^p s_i^r \right) \left(m_\Omega^r(\theta) - m_S^r \right).
\end{aligned}$$

Again,

$$\sum_{r=1}^{K}\left(\sum_{i=1}^{p}s_i^r\right)^2 \le \sum_{r=1}^{K}\left(\sum_{i=1}^{p}s_1^r\right)^2$$
$$= p^2\sum_{r=1}^{K}s_1^{2r} = p^2\frac{s_1^{2K+2}-1}{s_1^2-1} = p^2 s_1^{2K}(1+o(1))$$

as s_1 is large as $N\to\infty$.

Also, by Assumption 4, $s_1/M\to 0$ which implies $p^2 s_1^{2K}/M^{2K}\to 0$. And so, $M^{-2K}\sum_{r=1}^{K}\left(\sum_{i=1}^{p}s_i^r\right)^2 = o(1)$. Similarly, for the cross product term we have,

$$M^{-2K}\times N\sum_{r=1}^{K}\left(\sum_{i=1}^{p}s_i^r\right)\left(m_\Omega^r(\theta)-m_S^r\right) \le \frac{N^2}{M^K}\sum_{r=1}^{K}\frac{ps_1^r}{M^K}\left|\frac{m_\Omega^r(\theta)-m_S^r}{N}\right|$$
$$\le p\frac{N^2}{M^K}\sum_{r=1}^{K}\left|\frac{m_\Omega^r(\theta)-m_S^r}{N}\right|$$

which converges to 0 again as $M^{-K}N^2\to 0$ and the rest is stochastically bounded.

Hence we have statement (a) of the Lemma $M^{-2K}\widehat{J}(\theta,\mathrm{Sp}_p(\Sigma_N))\xrightarrow{P_\theta}0$.

(b) Next we have,

$$\widehat{J}(\theta,Sp_{p-1}(\Sigma_N)) = \sum_{r=1}^{K}\left(Nm_\Omega^r(\theta)-\lambda_p^r-s_p^r-\sum_{i=p+1}^{N}s_i^r\right)^2$$
$$= \sum_{r=1}^{K}\left(Nm_\Omega^r(\theta)-\sum_{i=p}^{N}s_i^r\right)^2+\sum_{r=1}^{K}\lambda_p^{2r}-2\sum_{r=1}^{K}\lambda_p^r\left(Nm_\Omega^r(\theta)-\sum_{i=p}^{N}s_i^r\right)$$

In the same way as we showed $M^{-2K}\widehat{J}(\theta,\mathrm{Sp}_p(\Sigma_N))\to 0$ in Probability, we can show:

$$M^{-2K}\sum_{r=1}^{K}\left(Nm_\Omega^r(\theta)-\sum_{i=p}^{N}s_i^r\right)^2\to 0$$

$$\text{and } M^{-K}\left(Nm_\Omega^r(\theta)-\sum_{i=p}^{N}s_i^r\right)=o(1)\text{ for } r=1,\ldots,K.$$

Thus

$$\begin{aligned} M^{-2K}\widehat{J}(\theta, Sp_{p-1}(\Sigma_N)) &= M^{-2K}\left(\sum_{r=1}^{K}\lambda_p^{2r}\right)(1+o(1)) \\ &\geq M^{-2K}\lambda_p^{2K}(1+o(1)) \\ &\geq 1 \text{ as } \lambda_p \geq M. \end{aligned}$$

Similarly we can show that $M^{-2K}\widehat{J}(\theta, Sp_q(\Sigma_N)) \geq 1$ for all $q < p$.

(c) Note that,

$$\begin{aligned} \widehat{J}(\theta, \mathrm{Sp}_{p+1}(\Sigma_N)) &= \sum_{r=1}^{K}\left(Nm_\Omega^r(\theta) - \sum_{i=p+2}^{N} s_i^r\right)^2 \\ &= \sum_{r=1}^{K}\left(Nm_\Omega^r(\theta) - \sum_{i=p+1}^{N} s_i^r - s_{p+2}^r\right)^2 \\ &= \widehat{J}(\theta, \mathrm{Sp}_p(\Sigma_N)) + \sum_{r=1}^{K} s_{p+2}^{2r} - 2N\sum_{r=1}^{K} s_{p+2}^r\left(m_\Omega^r(\theta) - m_S^r + N^{-1}\sum_{i=1}^{p} s_i^r\right). \end{aligned}$$

So,

$$\begin{aligned} \widehat{J}(\theta, \mathrm{Sp}_p(\Sigma_N)) - \widehat{J}(\theta, \mathrm{Sp}_{p+1}(\Sigma_N)) &= -\sum_{r=1}^{K} s_{p+2}^{2r} + 2N\sum_{r=1}^{K} s_{p+2}^r\left(m_\Omega^r(\theta) - m_S^r + N^{-1}\sum_{i=1}^{p} s_i^r\right) \\ &\leq 2N\sum_{r=1}^{K} s_{p+2}^r |m_\Omega^r(\theta) - m_S^r| + 2\sum_{i=1}^{p} s_i^r \\ &\leq 2N^2\sum_{r=1}^{K} s_{p+2}^r \frac{|m_\Omega^r(\theta) - m_S^r|}{N} + 2p\cdot s_1^{2s}. \end{aligned}$$

On dividing by M^{2K} each term on the R.H.S converges to 0 and hence we have,

$$M^{-2K}[\widehat{J}(\theta, \mathrm{Sp}_p(\Sigma_N)) - \widehat{J}(\theta, \mathrm{Sp}_{p+1}(\Sigma_N))] \leq 0.$$

Similarly, the proof can be extended for any $\mathrm{Sp}_{p+j}(\Sigma_N))$ with $j \geq 1$.

(d) Again, on expanding $\widehat{J}(\theta, \mathrm{Sp}_{p-2}(\Sigma_N))$ we have,

$$\widehat{J}(\theta, \mathrm{Sp}_{p-2}(\Sigma_N)) - \widehat{J}(\theta, \mathrm{Sp}_{p-1}(\Sigma_N)) \\ = \sum_{r=1}^{K} (\lambda_{p-1}^r + s_{p-1}^r)^2 - 2\sum_{r=1}^{K} (\lambda_{p-1}^r + s_{p-1}^r) \left(N m_\Omega^r(\theta) - \sum_{i=p}^{N} s_i^r - \lambda_p^r \right)$$

and as before we can dominate the terms without 'λ_i's by those containing λ_i by noting $\lambda_1 \geq \ldots \geq \lambda_p > M$. Hence we get,

$$\widehat{J}(\theta, \mathrm{Sp}_{p-2}(\Sigma_N)) - \widehat{J}(\theta, \mathrm{Sp}_{p-1}(\Sigma_N)) = \left(\sum_{r=1}^{K} \lambda_{p-1}^{2r} + 2\lambda_{p-1}^r \lambda_p^r \right) (1 + o(1)) \geq 3\lambda_p^{2K} (1 + o(1))$$

and the result follows for $q = p - 1$. Following the same line the proof can be generalized for any $q < p$. □

Choice of M: If M is unknown it can be estimated and for our purpose any crude estimate of M would work. Note by our assumptions on M, $\frac{M}{\max(N^{2/K}, \widehat{s}_1)} \to 0$ as $N, T \to \infty$, where $\widehat{s}_1$ is the largest eigenvalue of $\frac{1}{T}\sum_{t=1}^{T} U_t U_t'$. If U_{ij} are Gaussian then the order of $\widehat{s}_1$ is known and thus an estimate of M can be constructed. But, the plot of $\widehat{J}(\theta, Sp_q(\Sigma_N))$ vs q has the following convenient properties, based on which p can be estimated efficiently even without estimation of the order of M:

(1) $\widehat{J}(\theta, Sp_q(\Sigma_N))$ is decreasing for $q \leq p$ and with a huge rate of decrease.

$$\widehat{J}(\theta, Sp_{q-1}(\Sigma_N)) - \widehat{J}(\theta, Sp_q(\Sigma_N)) \geq 3sM^{2s} \text{ for } q \leq p.$$

(2) The rate of change after p is very small compared to the rate of decrease before

$$\frac{|\widehat{J}(\theta, Sp_{p+j}(\Sigma_N)) - \widehat{J}(\theta, Sp_p(\Sigma_N))|}{M^{2s}} \to 0.$$

Unknown θ: The next issue to be resolved is the case when the parameter θ is unknown and needs to be estimated. In this case, we have to be cautious as there can be overfitting of parameters. If there are more than s unknown parameters, then $\widehat{J}(\widehat{\theta}, Sp_q(\Sigma_N)) = 0$ for all $q = 0, 1, \ldots$

and we will not get a consistent estimator based on the method of moments estimation. We assume that the number of unknown parameters (in θ) estimated by minimizing $\widehat{J}(\hat{\theta}, Sp_q(\Sigma_N))$ is small enough so that we do not have overfitting. Furthermore, our stopping rule gives a consistent estimate for p under the following assumption assumption (which can be checked given a particular parametric model): $\frac{\widehat{J}(\widehat{\theta},Sp_q(\Sigma_N))-\widehat{J}(\theta,Sp_q(\Sigma_N))}{M^{2K}} \to 0$ in probability for any given θ and for all q.

References

Amengual, D., and M. W. Watson (2007): "Consistent Estimation of the Number of Dynamic Factors in a Large N and T Panel," *Journal of Business and Economic Statistics*, 25.

Anderson, T. W. (2003): *An Introduction to Multivariate Statistical Analysis*. Wiley, New York.

Angelini, E., G. Camba-Mendez, D. Giannone, L. Reichlin, and G. Runstler (2011): "Short-term forecasts of Euro Area GDP growth," *Econometrics Journal*, 14(1), 25–44.

Bai, J., and S. Ng (2002): "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70(1), 191–221.

Bai, Z., and J. Silverstein (2010): *Spectral Analysis of Large Dimensional Random Matrices*. Springer.

Bai, Z. D., and J. W. Silverstein (1998): "No Eigenvalues Outside the Support of the Limiting Spectral Distribution of Large Dimensional Sample Covariance Matrices," *Annals of Probability*, 26(1), 316–345.

——— (1999): "Exact Separation of Eigenvalues of Large Dimensional Covariance Matrices," *Annals of Probability*, 27(3), 1536–1555.

——— (2004): "CLT for Linear Spectral Statistics of Large-Dimensional Sample Covariance Matrices," *Annals of Probability*, 32(1A), 553–605.

Baik, J., and J. W. Silverstein (2006): "Eigenvalues of Large Sample Covariance Matrices of Spiked Population Models," *Journal of Multivariate Analysis*, J97, 1382–1408.

Billingsley, P. (1995): *Probability and Measure*. Wiley, New York.

Brown, S. J. (1989): "The Number of Factors in Security Returns," *Journal of Finance*, 54, 1247–1262.

Burda, Z., A. Jarosz, and J. Jurkiewicz (2006): "Applying Free Random Variables to Random Matrix Analysis of Financial Data," mimeo.

Chamberlain, G., and M. Rothschild (1983): "Arbitrage, Factor Structure and Mean-Variance Analysis in Large Asset Markets," *Econometrica*, 51(5), 1305–324.

Diebold, F., C. Li, and V. Yue (2008): "Global Yield Curve Dynamics and Interactions: A Dynamic Nelson-Siegel Approach," *Journal of Econometrics*, 146(2), 351–363.

Doz, C., D. Giannone, and L. Reichlin (2012): "A Quasi-Maximum Likelihood Approach for Large Approximate Dynamic Factor Models," forthcoming in the Review of Economics and Statistics.

Dozier, R. B., and J. W. Silverstein (2007): "On the Empirical Distribution of Eigenvalues of Large Dimensional Information-Plus-Noise Type Matrices," *Journal of Multivariate Analysis*, 98(4), 678–694.

Edelman, A., and N. R. Rao (2005): "Random Matrix Theory," *Acta Numerica*, pp. 233–297.

Eickmeier, S., W. Lemke, and M. Marcellino (2011): "Classical Time-Varying FAVAR Models – Estimation, Forecasting, and Structural Analysis," Discussion Paper 04/2011, Deutsche Bundesbank.

Fan, J., Y. Fan, and J. Lv (2008): "High dimensional covariance matrix estimation using a factor model," *Journal of Econometrics*, 147, 186–197.

Goldberger, A. S. (1972): "Maximum-Likelihood Estimation of Regressions Containing Unobservable Independent Variables," *International Economic Review*, 13(1), 1–15.

Grenander, U., and G. Szego (1958): *Toeplitz Forms and their Applications*. University of California Press, Berkeley.

Hahn, J., and G. Kuersteiner (2002): "Asymptotically Unbiased Inference for a Dynamic Panel Model with Fixed Effects when both n and T are Large," *Econometrica*, 70, 1639–1657.

Hallin, M., and R. Liska (2007): "Determining the Number of Factors in the General Dynamic Factor Model," *Journal of the American Statistical Association*, 102(478), 603–617.

Harding, M. (2008): "Explaining the Single Factor Bias of Arbitrage Pricing Models in Finite Samples," *Economics Letters*, 99(1).

Johnstone, I. (2007): "High Dimensional Statistical Inference and Random Matrices," *Proceedings of the European Mathematical Society*.

Jonsson, D. (1982): "Some Limit Theorems for the Eigenvalues of a Sample Covariance Matrix," *Journal of Multivariate Analysis*, 12, 1–38.

Ma, C. (2003): "Spatio-temporal stationary covariance models," *Journal of Multivariate Analysis*, 86, 97–107.

Marcenko, V., and L. Pastur (1967): "Distribution of Eigenvalues for Some Sets of Random Matrices," *Matthematics of the USSR-Sbornik*, 1(4).

Nica, A., and R. Speicher (2006): *Lectures on the Combinatorics of Free Probability*. Cambridge University Press.

Onatski, A. (2009): "Testing Hypotheses about the Number of Factors in Large Factor Models," *Econometrica*, 77(5), 1447–1479.

——— (2010): "Determining the Number of Factors from Empirical Distribution of Eigenvalues," *The Review of Economics and Statistics*, 92(4), 1004–1016.

——— (2012): "Asymptotics of the Principal Components Estimator of Large Factor Models with Weak Factors and i.i.d. Gaussian Noise," *Journal of Econometrics*, 168, 244–258.

Ouysse, R. (2006): "Approximate factor models: finite sample distributions," *Journal of Statistical Computation and Simulation*, 76(4), 279–303.

Passemier, D., and J.-F. Yao (2012): "On Determining the Number of Spikes in a High-Dimensional Spiked Population Model," *Random Matrices: Theory and Applications*, 1(1).

Reis, R., and M. Watson (2010): "Relative Goods' Prices, Pure Inflation, and the Phillips Correlation," *American Economic Journal: Macroeconomics*, 2(3), 128–157.

Silverstein, J. (1995): "Strong Convergence of the Empirical Distribution of Large Dimensional Random Matrices," *Journal of Multivariate Analysis*, 55, 331–339.

Silverstein, J. W., and Z. D. Bai (1995): "On the Empirical Distribution of Eigenvalues of a Class of Large Dimensional Random Matrices," *Journal of Multivariate Analysis*, pp. 175–192.

Speicher, R. (2005): "Free Probability Theory," Notes on Operator Algebras, Ottawa, mimeo.

Stein, M. (2005): "Space-Time Covariance Functions," *Journal of the American Statistical Association*, 100(469).

Stock, J., and M. Watson (2005): "Implications of Dynamic Factor Models for VAR Analysis," mimeo.

Van Loan, C., and N. Pitsianis (1993): "Approximation with Kronecker Products," in *Linear Algebra for Large Scale and Real Time Applications*, ed. by M. Moonen, and G. Golub. Kluwer.

Voiculescu, D. V. (1985): "Symmetries of Some Reduced Free Product C*-Algebras," in *Operator Algebras and their Connections, with Topology and Ergodic Theory, Lecture Notes in Mathematics*, vol. 1132, pp. 556–588. Springer, New York.

——— (1998): "Lectures on Free Probability Theory," in *Lectures on Probability Theory and Statistics*, vol. 1738, pp. 279–349. Springer, Saint Flour 28.

Zellner, A. (1970): "Estimation of Regression Relationships Containing Unobservable Independent Variables," *International Economic Review*, 11(3), 441–454.

FIGURE 1. The effect of increasing the noise-to-signal ratio on the estimated number of factors.

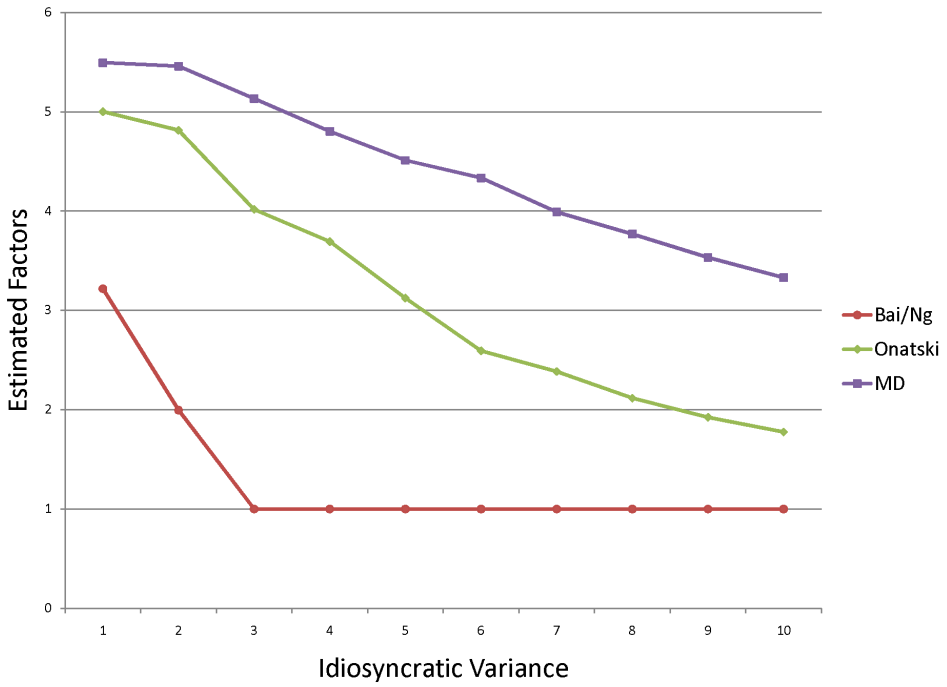

TABLE 1. Monte Carlo simulations for Designs I and II.

Design I

N	T	c	Mean						Under/Over					
			MD	MDW	MD-IC	MDW-IC	Onatski	Bai & Ng	MD	MDW	MD-IC	MDW-IC	Onatski	Bai & Ng
30	100	0.3	5.188	5.078	5.038	5.028	5.004	5.442	0.00/0.13	0.00/0.07	0.00/0.04	0.00/0.03	0.00/0.00	0.00/0.43
50	100	0.5	5.272	5.202	5.112	5.038	4.988	5.000	0.00/0.23	0.00/0.16	0.00/0.11	0.00/0.04	0.01/0.00	0.00/0.00
70	100	0.7	5.490	5.520	5.304	5.052	4.996	4.960	0.00/0.37	0.00/0.37	0.00/0.25	0.00/0.05	0.01/0.00	0.04/0.00
90	100	0.9	5.656	5.890	5.428	5.044	4.978	4.670	0.00/0.45	0.00/0.53	0.00/0.34	0.00/0.04	0.03/0.00	0.33/0.00
90	300	0.3	5.090	5.084	5.016	5.004	5.000	4.932	0.00/0.08	0.00/0.08	0.00/0.02	0.00/0.00	0.00/0.00	0.07/0.00
150	300	0.5	5.226	5.178	5.064	5.000	5.000	4.040	0.00/0.22	0.00/0.17	0.00/0.06	0.00/0.00	0.00/0.00	0.96/0.00
210	300	0.7	5.416	5.394	5.272	5.010	5.000	4.000	0.00/0.34	0.00/0.33	0.00/0.24	0.00/0.01	0.00/0.00	1.00/0.00
270	300	0.9	5.576	5.668	5.428	5.010	5.004	3.868	0.00/0.46	0.00/0.50	0.00/0.38	0.00/0.01	0.00/0.00	1.00/0.00
150	500	0.3	5.130	5.112	5.010	5.008	5.000	4.064	0.00/0.13	0.00/0.11	0.00/0.01	0.00/0.01	0.00/0.00	0.94/0.00
250	500	0.5	5.238	5.184	5.076	5.002	5.020	4.000	0.00/0.22	0.00/0.18	0.00/0.08	0.00/0.00	0.00/0.01	1.00/0.00
350	500	0.7	5.404	5.390	5.220	5.000	5.004	3.942	0.00/0.35	0.00/0.34	0.00/0.21	0.00/0.00	0.00/0.00	1.00/0.00
450	500	0.9	5.494	5.564	5.348	5.000	5.000	3.218	0.00/0.41	0.00/0.47	0.00/0.31	0.00/0.00	0.00/0.00	1.00/0.00

Design II

N	T	c	Mean						Under/Over					
			MD	MDW	MD-IC	MDW-IC	Onatski	Bai & Ng	MD	MDW	MD-IC	MDW-IC	Onatski	Bai & Ng
30	100	0.3	4.912	4.836	4.716	4.694	2.888	5.372	0.20/0.08	0.22/0.05	0.30/0.02	0.32/0.02	0.78/0.00	0.01/0.37
50	100	0.5	4.790	4.700	4.564	4.348	2.768	3.902	0.34/0.11	0.41/0.08	0.46/0.03	0.64/0.01	0.92/0.00	0.96/0.00
70	100	0.7	4.888	4.944	4.650	4.186	2.658	3.156	0.38/0.21	0.37/0.22	0.47/0.12	0.78/0.00	0.98/0.00	1.00/0.00
90	100	0.9	4.936	5.200	4.684	3.960	2.466	2.546	0.39/0.25	0.33/0.31	0.47/0.14	0.89/0.00	0.99/0.00	1.00/0.00
90	300	0.3	4.898	4.980	4.660	4.818	4.124	2.626	0.17/0.06	0.09/0.06	0.34/0.00	0.19/0.01	0.63/0.00	1.00/0.00
150	300	0.5	4.816	4.830	4.578	4.362	3.806	1.942	0.30/0.11	0.27/0.09	0.45/0.03	0.64/0.00	0.88/0.00	1.00/0.00
210	300	0.7	4.882	4.918	4.700	4.162	3.706	1.176	0.34/0.20	0.30/0.19	0.42/0.11	0.84/0.00	0.96/0.00	1.00/0.00
270	300	0.9	4.932	5.064	4.740	4.032	3.478	0.998	0.33/0.22	0.26/0.26	0.42/0.15	0.94/0.00	0.99/0.00	1.00/0.00
150	500	0.3	4.950	5.040	4.692	4.888	4.558	1.998	0.15/0.09	0.04/0.08	0.31/0.00	0.12/0.01	0.42/0.00	1.00/0.00
250	500	0.5	4.858	4.910	4.608	4.366	4.056	1.080	0.27/0.13	0.19/0.10	0.43/0.04	0.64/0.00	0.86/0.00	1.00/0.00
350	500	0.7	4.904	4.890	4.686	4.102	3.926	0.996	0.30/0.19	0.28/0.17	0.40/0.09	0.90/0.00	0.97/0.00	1.00/0.00
450	500	0.9	4.854	4.950	4.704	4.018	3.844	0.638	0.34/0.18	0.29/0.21	0.42/0.13	0.98/0.00	0.99/0.00	1.00/0.00

TABLE 2. Monte Carlo simulations for Designs III and IV.

Design III			Mean				Under/Over			
N	T	c	MD	MD-IC	Onatski	Bai & Ng	MD	MD-IC	Onatski	Bai & Ng
30	100	0.3	5.148	5.000	5.004	5.492	0.00/0.06	0.00/0.00	0.00/0.00	0.00/0.46
50	100	0.5	5.160	4.998	4.996	5.000	0.00/0.12	0.00/0.00	0.00/0.00	0.00/0.00
70	100	0.7	5.350	4.998	4.982	4.960	0.00/0.26	0.00/0.00	0.02/0.00	0.04/0.00
90	100	0.9	5.532	4.994	4.948	4.678	0.00/0.34	0.01/0.00	0.05/0.00	0.32/0.00
90	300	0.3	5.104	5.000	5.010	4.940	0.00/0.10	0.00/0.00	0.00/0.01	0.06/0.00
150	300	0.5	5.456	5.000	5.002	4.044	0.00/0.36	0.00/0.00	0.00/0.00	0.96/0.00
210	300	0.7	5.624	5.000	5.000	3.996	0.00/0.39	0.00/0.00	0.00/0.00	1.00/0.00
270	300	0.9	5.496	5.000	5.002	3.910	0.00/0.34	0.00/0.00	0.00/0.00	1.00/0.00
150	500	0.3	5.238	5.000	5.002	4.068	0.00/0.21	0.00/0.00	0.00/0.00	0.93/0.00
250	500	0.5	5.686	5.000	5.000	4.000	0.00/0.49	0.00/0.00	0.00/0.00	1.00/0.00
350	500	0.7	5.780	5.000	5.006	3.928	0.00/0.46	0.00/0.00	0.00/0.00	1.00/0.00
450	500	0.9	5.512	5.000	5.002	3.252	0.00/0.35	0.00/0.00	0.00/0.00	1.00/0.00

Design IV			Mean				Under/Over			
N	T	c	MD	MD-IC	Onatski	Bai & Ng	MD	MD-IC	Onatski	Bai & Ng
30	100	0.3	4.440	3.762	3.016	5.388	0.57/0.01	0.96/0.00	0.80/0.00	0.01/0.38
50	100	0.5	4.506	3.766	2.616	3.942	0.56/0.07	0.98/0.00	0.94/0.00	0.95/0.00
70	100	0.7	4.532	3.634	2.506	3.142	0.55/0.10	0.99/0.00	0.97/0.00	1.00/0.00
90	100	0.9	4.574	3.438	2.394	2.564	0.59/0.18	0.99/0.00	1.00/0.00	1.00/0.00
90	300	0.3	4.782	3.996	4.046	2.642	0.30/0.08	0.99/0.00	0.66/0.00	1.00/0.00
150	300	0.5	5.024	3.978	3.730	1.954	0.25/0.22	1.00/0.00	0.89/0.00	1.00/0.00
210	300	0.7	5.010	3.862	3.552	1.160	0.36/0.28	1.00/0.00	0.98/0.00	1.00/0.00
270	300	0.9	4.718	3.698	3.348	1.000	0.50/0.19	1.00/0.00	0.99/0.00	1.00/0.00
150	500	0.3	5.022	4.000	4.478	1.996	0.15/0.16	1.00/0.00	0.50/0.00	1.00/0.00
250	500	0.5	5.326	3.978	4.050	1.070	0.17/0.38	1.00/0.00	0.88/0.00	1.00/0.00
350	500	0.7	5.180	3.910	3.914	1.000	0.33/0.32	1.00/0.00	0.96/0.00	1.00/0.00
450	500	0.9	4.772	3.718	3.756	0.626	0.51/0.22	1.00/0.00	0.99/0.00	1.00/0.00

TABLE 3. Estimation of misspecified models. Data is generated according to Designs III and IV but a model with $A_N = \sigma^2 I_N$ and $B_T = I_T$ is assumed.

Design III misspecified

N	T	c	# Factors					
			Mean			Under/Over		
			MDW-IC	Onatski	Bai & Ng	MDW-IC	Onatski	Bai & Ng
30	100	0.3	5.040	5.000	5.522	0.00/0.04	0.00/0.00	0.00/0.50
50	100	0.5	5.060	4.992	5.000	0.00/0.06	0.01/0.01	0.00/0.00
70	100	0.7	5.110	4.994	4.962	0.00/0.11	0.01/0.00	0.04/0.00
90	100	0.9	5.106	4.972	4.678	0.00/0.10	0.03/0.00	0.32/0.00
90	300	0.3	5.032	5.002	4.938	0.00/0.03	0.00/0.00	0.06/0.00
150	300	0.5	5.058	5.000	4.044	0.00/0.06	0.00/0.00	0.96/0.00
210	300	0.7	5.116	5.010	4.000	0.00/0.11	0.00/0.00	1.00/0.00
270	300	0.9	5.146	5.000	3.890	0.00/0.14	0.00/0.00	1.00/0.00
150	500	0.3	5.052	5.000	4.072	0.00/0.05	0.00/0.00	0.93/0.00
250	500	0.5	5.118	5.000	4.000	0.00/0.12	0.00/0.00	1.00/0.00
350	500	0.7	5.256	5.002	3.946	0.00/0.24	0.00/0.00	1.00/0.00
450	500	0.9	5.356	5.000	3.228	0.00/0.32	0.00/0.00	1.00/0.00

Design IV misspecified

N	T	c	# Factors					
			Mean			Under/Over		
			MDW-IC	Onatski	Bai & Ng	MDW-IC	Onatski	Bai & Ng
30	100	0.3	4.718	2.872	5.478	0.30/0.02	0.79/0.00	0.01/0.47
50	100	0.5	4.408	2.682	3.922	0.60/0.02	0.94/0.00	0.95/0.00
70	100	0.7	4.288	2.566	3.198	0.71/0.02	0.98/0.00	1.00/0.00
90	100	0.9	4.070	2.340	2.584	0.81/0.01	0.99/0.00	1.00/0.00
90	300	0.3	4.896	4.060	2.640	0.13/0.03	0.65/0.00	1.00/0.00
150	300	0.5	4.582	3.706	1.942	0.44/0.02	0.90/0.00	1.00/0.00
210	300	0.7	4.458	3.576	1.188	0.57/0.03	0.98/0.00	1.00/0.00
270	300	0.9	4.340	3.412	1.000	0.68/0.02	1.00/0.00	1.00/0.00
150	500	0.3	4.980	4.488	2.002	0.06/0.03	0.48/0.00	1.00/0.00
250	500	0.5	4.784	4.038	1.078	0.26/0.04	0.87/0.00	1.00/0.00
350	500	0.7	4.720	3.896	0.996	0.37/0.09	0.97/0.00	1.00/0.00
450	500	0.9	4.672	3.744	0.648	0.41/0.08	0.99/0.00	1.00/0.00

TABLE 4. Estimation of misspecified models. Data is generated according to Designs V and VI but a model with $A_N = \sigma^2/(1-\rho^2)I_N$, $(B_T)_{i,j} = \rho^{|i-j|}$ is assumed.

Design V			# Factors					
			Mean			Under/Over		
N	T	c	MDW-IC	Onatski	Bai & Ng	MDW-IC	Onatski	Bai & Ng
100	30	0.3	5.000	5.686	4.776	0.02/0.02	0.00/0.63	0.09/0.00
	50	0.5	5.006	5.000	5.010	0.00/0.01	0.00/0.00	0.00/0.01
	70	0.7	5.002	5.000	5.002	0.00/0.00	0.00/0.00	0.00/0.00
	90	0.9	5.002	5.000	5.012	0.00/0.00	0.00/0.00	0.00/0.00
300	90	0.3	5.004	5.000	5.012	0.00/0.00	0.00/0.00	0.00/0.00
	150	0.5	5.000	5.000	5.002	0.00/0.00	0.00/0.00	0.00/0.00
	210	0.7	5.000	5.000	5.002	0.00/0.00	0.00/0.00	0.00/0.00
	270	0.9	5.000	5.000	5.002	0.00/0.00	0.00/0.00	0.00/0.00
500	150	0.3	5.000	5.000	5.000	0.00/0.00	0.00/0.00	0.00/0.00
	250	0.5	5.000	5.000	5.000	0.00/0.00	0.00/0.00	0.00/0.00
	350	0.7	5.000	5.000	5.014	0.00/0.00	0.00/0.00	0.00/0.01
	450	0.9	5.000	5.000	5.002	0.00/0.00	0.00/0.00	0.00/0.00

Design VI			# Factors					
			Mean			Under/Over		
N	T	c	MD-IC	Onatski	Bai & Ng	MD-IC	Onatski	Bai & Ng
100	30	0.3	7.322	7.296	2.762	0.04/0.95	0.00/1.00	0.62/0.35
	50	0.5	6.940	7.760	1.202	0.17/0.82	0.00/1.00	0.87/0.11
	70	0.7	6.404	7.772	1.306	0.20/0.76	0.00/1.00	0.78/0.02
	90	0.9	6.694	7.620	3.084	0.14/0.82	0.00/1.00	0.41/0.01
300	90	0.3	4.502	8.000	4.636	0.46/0.42	0.00/1.00	0.08/0.00
	150	0.5	5.648	7.846	5.002	0.27/0.61	0.00/1.00	0.00/0.00
	210	0.7	6.628	6.824	5.014	0.17/0.79	0.00/0.99	0.00/0.01
	270	0.9	7.490	5.780	5.000	0.06/0.93	0.00/0.69	0.00/0.00
500	150	0.3	4.918	7.994	5.000	0.30/0.40	0.00/1.00	0.00/0.00
	250	0.5	6.810	6.896	5.002	0.05/0.80	0.00/1.00	0.00/0.00
	350	0.7	7.192	5.340	5.000	0.13/0.86	0.00/0.33	0.00/0.00
	450	0.9	7.674	5.008	5.002	0.06/0.94	0.00/0.01	0.00/0.00

ONLINE APPENDIX

B.1. Using Freeness to Derive Mixed Moments of Non-commutative Random Variables

Freeness is a convenient property of random variables since it amounts to an iterative procedure for computing mixed moments of products of random variables from the moments of the constituent random variables. Notice that we can re-write the expression in Definition 2 for the case when $\phi(a_k) \neq 0$ by subtracting the individual means:

$$\phi((a_1 - \phi(a_1)1)(a_2 - \phi(a_2)1)...(a_k - \phi(a_k)1)) = 0. \tag{B.1}$$

Our primary focus is on the computation of mixed moments $m^s_{ab} = \phi((ab)^s)$. We illustrate this process for the first two moments.

If a and b are free then,

$$\phi((a - \phi(a)1)(b - \phi(b)1)) = 0, \tag{B.2}$$

and expanding,

$$\phi(ab - \phi(a)b - a\phi(b) + \phi(a)\phi(b)) = 0, \tag{B.3}$$

$$\phi(ab) - \phi(a)\phi(b) - \phi(a)\phi(b) + \phi(a)\phi(b)) = 0, \tag{B.4}$$

$$m^1_{ab} = \phi(ab) = \phi(a)\phi(b). \tag{B.5}$$

Notice that this expression is the same as the one we would obtain if a and b were independent random variables in a classical probability space. Now consider, $m^2_{ab} = \phi((ab)^2)$. Since a and b are non-commutative, $m^s_{ab} = \phi(abab)$, we start by expanding the expression:

$$\phi((a - \phi(a)1)(b - \phi(b)1)(a - \phi(a)1)(b - \phi(b)1)) = 0. \tag{B.6}$$

If a and b are freely independent then so is the product $abab$. We can apply Definition 2 since all adjacent terms of the product are free and hence:

$$\phi((a-\phi(a)1)\,(b-\phi(b)1)\,(a-\phi(a)1)\,(b-\phi(b)1))=0. \tag{B.7}$$

We can expand this expression to obtain:

$$\begin{aligned}\phi(abab)-\phi(a)\phi(bab)-\phi(b)\phi(aab)+\phi(a)\phi(b)\phi(ab)-\phi(a)\phi(abb)+\\ \phi^2(a)\phi(bb)+\phi(a)\phi(b)\phi(ab)-\phi^2(a)\phi^2(b)-\phi(b)\phi(aba)+\\ \phi(a)\phi(b)\phi(ba)+\phi^2(b)\phi(aa)-\phi^2(a)\phi^2(b)+\phi(a)\phi(b)\phi(ab)-\\ \phi^2(a)\phi(b)\phi(b)-\phi(a)\phi^2(b)\phi(a)+\phi^2(a)\phi^2(b)=0\end{aligned} \tag{B.8}$$

Using the fact that $\phi(ab)=\phi(a)\phi(b)$ and that $\phi(aba)=\phi(aa)\phi(b)$ we can simplify this expression as:

$$\begin{aligned}\phi(abab)-\phi^2(a)\phi(bb)-\phi^2(b)\phi(aa)+\phi^2(a)\phi^2(b)-\phi^2(a)\phi(bb)+\\ \phi^2(a)\phi(bb)+\phi^2(a)\phi^2(b)-\phi^2(a)\phi^2(b)-\\ \phi^2(b)\phi(aa)+\phi^2(a)\phi^2(b)+\phi^2(b)\phi(aa)-\phi^2(a)\phi^2(b)+\\ \phi^2(a)\phi^2(b)-\phi^2(a)\phi^2(b)-\phi^2(a)\phi^2(b)+\phi^2(a)\phi^2(b)=0\end{aligned} \tag{B.9}$$

Since most of the terms in this expression cancel, we obtain

$$\phi(abab)=\phi^2(a)\phi(bb)+\phi^2(b)\phi(aa)-\phi^2(a)\phi^2(b). \tag{B.10}$$

B.2. When are two matrices free?

In general two arbitrary matrices are not freely independent even if the elements of the matrices are independent, since their eigenspaces may satisfy a particular relationship to each other. One of the main insights of random matrix theory is that certain matrices become freely

independent asymptotically as $N \to \infty$ (Voiculescu, 1998). Note that asymptotic freeness for large random matrices Y requires both the convergence of the probability law $\mathfrak{P}$ as $N \to \infty$ and Definition 2 to be satisfied for $\phi_N(Y) = \lim_{N\to\infty} \frac{1}{N} E_X \operatorname{tr}(Y)$. The convergence of the probability law implies the convergence of all moments of the eigenvalue distribution in the large N limit.

Consider the set of matrices S distributed uniformly on the Stiefel manifold (Anderson, 2003, Definition 4.5.1).

DEFINITION 3: *Let S_N be an $N \times N$ matrix satisfying ${S'}_N S_N = I_N$ and $S_N H_N \stackrel{d}{=} S_N$ for all orthogonal matrices H_N. Then S_N is uniformly distributed on the group of square orthogonal matrices $\mathcal{O}(N)$.*

Let μ be the probability measure on the random matrices S_N in Definition 3. Then μ is the unique probability measure on $\mathcal{O}(N)$ such that for some $D \subset \mathcal{O}(N)$, $\mu(\Gamma D) = \mu(D\Gamma) = \mu(D)$ for al $\Gamma \in \mathcal{O}(N)$. The distribution μ is referred to as the Haar (invariant) distribution on $\mathcal{O}(N)$.

LEMMA 3: *Let S_N be an $N \times N$ matrix with the Haar distribution and X_N and Y_N two sequences of random $N \times N$ symmetric matrices such that their empirical eigenvalue distributions converge to proper non-random distributions with bounded support. If S_N is independent of X_N and Y_N then X_N and $S_N Y_N {S'}_N$ are asymptotically free as $N \to \infty$.*

Proof: See Speicher (2005).

Note that by the spectral decomposition of the the matrix Y_N, the effect of the Haar distributed random matrix S_N is to introduce a random rotation in the eigenvectors of Y_N. Asymptotic freeness requires us to identify matrices that are rotationally invariant and thus preserve the information on the eigenvalue distribution independently of the eigenvectors which are now randomly rotated. Hence we are particularly interested in matrices Y_N which are unitarily invariant, that is the spectrum of Y_N and that of $S_N Y_N {S'}_N$ is the same for S_N on the orthogonal group. More formally, we can use the following lemma (Anderson, 2003, Lemma 13.3.2) for normalized matrices.

LEMMA 4: *Let C_N be an arbitrary matrix of order N and define the following normalization:*

$$J(C_N) = \text{diag}\left\{ (C_N)_{1,1} \Big/ |(C_N)_{1,1}|, (C_N)_{2,2} \Big/ |(C_N)_{2,2}|, ..., (C_N)_{N,N} \Big/ |(C_N)_{N,N}| \right\}. \tag{B.11}$$

If the orthogonal matrix S_N of order N has a distribution such that $(S_N)_{i,1} \geq 0$ and if

$$S^*{}_N = J(S_N H_N) S_N H_N$$

has the same distribution for every orthogonal matrix H_N, then S_N has the conditional Haar invariant distribution.

Proof: See Anderson (2003, pp. 542).

The conditional Haar invariant distribution is the conditional distribution of a normalized orthogonal matrix S_N with the Haar distribution, where we let the $(S_N)_{i,1} \geq 0$. It is equal to 2^N times the Haar distribution.

If Y_N is a covariance matrix, unitary invariance requires that the normalized eigenvectors W_N from the spectral decomposition of $Y_N = W_N D_N W'_N$ be distributed conditionally Haar and independent of D_N, the diagonal matrix of eigenvalues. This ensures that further rotations by Haar distributed orthogonal matrices S_N do not change the eigenvalue distribution. Covariance matrices satisfying this requirement include those derived from matrices with iid Normal elements, $Y_N = (1/T)\epsilon\epsilon'$, where $(\epsilon)_{i,j}$ are distributed iid N(0,1). This is captured by the following result:

LEMMA 5: *Let $W_N = (w_1, w_2, ..., w_N)'$ be the matrix of normalized eigenvectors of a covariance matrix Y_N, where $(w)_{1,i} \geq 0$ and where Y_N is distributed according to a Wishart distribution with mean I_N, then W_N has the conditional Haar invariant distribution and W_N is distributed independently of the eigenvalues of Y_N.*

Proof: See Anderson (2003, Theorem 13.3.3).

B.3. Computing Mixed Moments

The mixed moments of $A_N\Psi_N$ can be computed recursively by applying Definition 2 to the moments of A_N and Ψ_N, when these are freely independent. First, notice that the moments of Ψ_N are given by Lemma 1:

$$m_\Psi^1 = 1 \tag{B.12}$$

$$m_\Psi^2 = 1 + c \tag{B.13}$$

$$m_\Psi^3 = 1 + 3c + c^2 \tag{B.14}$$

$$m_\Psi^4 = 1 + 6c + 6c^2 + c^3 \tag{B.15}$$

$$m_\Psi^5 = 1 + 10c + 20c^2 + 10c^3 + c^4 \tag{B.16}$$

$$m_\Psi^6 = 1 + 15c + 50c^2 + 50c^3 + 15c^4 + c^5. \tag{B.17}$$

Free independence implies that $m^1(A\Psi) = m_A^1 m_\Psi^1$. But since $m_\Psi^1 = 1$, we have

$$m^1(A\Psi) = m_A^1. \tag{B.18}$$

Similarly free independence implies that:

$$m^2(A\Psi) = \left(m_A^1\right)^2 m_\Psi^2 + \left(m_\Psi^1\right)^2 m_A^2 - \left(m_A^1\right)^2 \left(m_\Psi^1\right)^2. \tag{B.19}$$

Substituting $m_\Psi^2 = 1 + c$ and $m_\Psi^1 = 1$ in the expression above we obtain:

$$m^2(A\Psi) = \left(m_A^1\right)^2 (1 + c) + m_A^2 - \left(m_A^1\right)^2 \tag{B.20}$$

$$m^2(A\Psi) = m_A^2 + c(m_A^1)^2. \tag{B.21}$$

We can continue this process to obtain:

$$m^3(A\Psi) = m_A^3 + 3cm_A^1 m_A^2 + c^2(m_A^1)^3 \tag{B.22}$$

(B.23) $$m^4(A\Psi) = m_A^4 + 2c\left(\left(m_A^2\right)^2 + 2m_A^1 m_A^3\right) + 6c^2(m_A^1)^2 m_A^2 + c^3(\left(m_A^1\right)^4).$$

The expressions in Example 3 follow immediately by substituting the moments of A_N. This recursive procedure is easy to implement using symbolic software such as Maple or Mathematica in order to derive the analytic expression for any moment of interest.

B.3.1. Proof of Proposition 5. Define the series $\varrho_F(w) = 1/\sum_{s=1}^{\infty} m_F^s w^s$ where m_F^s are the moments of some probability distribution F. Let $S_F(w) = \varrho_F(w)(1+w)/w$. Let X and Y be two free random variables with associated probability measures F and G. Then, $S_F(w)S_G(w) = S_{FG}(w)$ (Voiculescu, 1998).

Now consider the Cauchy Transform G_Ω of Ω_N as $N \to \infty$ and let m_Ω^s be the moments of the asymptotic eigenvalue distribution of Ω. Then $G_\Omega(w) = \sum_{s=0}^{\infty} m_\Omega^s / w^{s+1}$. Following Burda et. al. (2006) we can let $G_\Omega(w) = M_\Omega(w)/w + 1$ such that $M_\Omega(w) = \sum_{s=1}^{\infty} m_\Omega^s / w^s$. Note also that $M_\Omega(\varrho_\Omega(w)) = w$. From the definition of the Cauchy transform we have $N^{-1}\sum_{p=1}^{N} \frac{1}{1-\lambda_p/w} = 1 + w$. Furthermore, $N^{-1}\sum_{p=1}^{N} \frac{1}{1-\lambda_p/\varrho_A(w)} = 1 + w$. If we now multiply both the numerator and denominator by $1/\varrho_\Psi(w)$ we have $N^{-1}\sum_{p=1}^{N} \frac{1/\varrho_\Psi(w)}{1/\varrho_\Psi(w) - \lambda_p/[\varrho_A(w)\varrho_\Psi(w)]} = 1 + w$. Since $S_A(w)S_\Psi(w) = S_{A\Psi}(w)$ we have,

$$N^{-1}\sum_{p=1}^{N} \frac{1/\varrho_\Psi(w)}{1/\varrho_\Psi(w) - z\lambda_p/((1+z)\varrho_{A\Psi}(w))} = 1 + w.$$

Furthermore, we have $N^{-1}\sum_{p=1}^{N} \frac{1}{1-(\lambda_p \varrho_\Psi(w))/((\frac{1+z}{z})\varrho_{A\Psi}(w))} = 1 + w$. If we now substitute $M_\Omega(w)$ for w we have $N^{-1}\sum_{p=1}^{N} \frac{1}{1-\frac{\lambda_p \varrho_\Psi(M(w))}{\frac{1+M_\Omega(w)}{M_\Omega(w)}w}} = 1 + M_\Omega(w)$. Re-writing this expression we obtain $M_\Omega(w) = M_A(\frac{w(1+M_\Omega(w))}{M_\Omega(w)\varrho_\Psi(M(w))})$. Moreover, it ca be shown that $\varrho_\Psi(w) = (1+w)(c+w)/w$. Substituting in the previous expression (and after some further cancelations) we obtain $M_\Omega(w) = M_A(\frac{w}{1+cM_\Omega(w)})$. Re-writing this expression as a series we obtain the expression in Proposition 5: $\sum_{s=1}^{\infty} \frac{m_\Omega^s}{w^s} = \sum_{s=1}^{\infty} \frac{m_A^s}{w^s}(1 + c\sum_{r=1}^{\infty} \frac{m_\Omega^r}{w^r})^s$.

B.3.2. Time-Series Correlations.

Proof of Proposition 6. The proof of equation 9 is very similar to that of given in Proposition 5 with (Ψ, B) substituted for (A, Ψ) and using the fact that $M_{\Psi B}(w) = cM_{\frac{1}{T}\epsilon B\epsilon'}(w)$. We can expand equation 9 as follows:

$$c(\mathcal{M}_\Omega - 1) = c\frac{m_B^1}{w}\mathcal{M} + c^2\frac{m_B^2}{w^2}\mathcal{M}^2 + c^3\frac{m_B^3}{w^3}\mathcal{M}^3 + c^4\frac{m_B^4}{w^4}\mathcal{M}^4 + O(w^{-5}), \tag{B.24}$$

where

$$\mathcal{M} = 1 + \frac{m_\Omega^1}{w} + \frac{m_\Omega^2}{w^2} + \frac{m_\Omega^3}{w^3} + \frac{m_\Omega^4}{w^4} + O(w^{-5}). \tag{B.25}$$

Expanding the RHS of equation B.24 in terms of $1/w$ we obtain:

$$\begin{aligned} c(\mathcal{M} - 1) = {} & \frac{1}{w}cm_B^1 + \frac{1}{w^2}\left(cm_B^1 m_\Omega^1 + c^2 m_B^2\right) + \frac{1}{w^3}\left(cm_B^1 m_\Omega^2 + c^3 m_B^3 + 2c^2 m_B^2 m_\Omega^1\right) + \\ & + \frac{1}{w^4}\left(3c^3 m_B^3 m_\Omega^1 + cm_B^1 m_\Omega^3 + c^2 m_B^2(2m_\Omega^2 + \left(m_\Omega^1\right)^2) + c^3 m_B^4\right) + O(w^{-5}). \end{aligned} \tag{B.26}$$

Dividing both sides by c, equating terms in powers of $1/w$ and substituting recursively for earlier terms we obtain:

$$m_\Omega^1 = m_B^1 \tag{B.27}$$

$$m_\Omega^2 = cm_B^2 + m_B^1 m_\Omega^1 = cm_B^2 + \left(m_B^1\right)^2 \tag{B.28}$$

$$m_\Omega^3 = c^2 m_B^3 + 2cm_B^2 m_\Omega^1 + m_\Omega^2 m_B^1 = c^2 m_B^3 + 2cm_B^2 m_B^1 + \left(cm_B^2 + \left(m_B^1\right)^2\right) m_B^1 \tag{B.29}$$

which can be further simplified as

$$m_\Omega^3 = c^2 m_B^3 + 2cm_B^2 m_\Omega^1 + m_\Omega^2 m_B^1 = c^2 m_B^3 + 2cm_B^2 m_B^1 + \left(cm_B^2 + \left(m_B^1\right)^2\right) m_B^1 \tag{B.30}$$

$$m_\Omega^3 = c^2 m_B^3 + 3cm_B^2 m_B^1 + \left(m_B^1\right)^3 \tag{B.31}$$

Additionally, we have:

$$m_\Omega^4 = m_B^1 m_\Omega^3 + cm_B^2(2m_\Omega^2 + \left(m_\Omega^1\right)^2) + 3c^2 m_B^3 m_\Omega^1 + c^3 m_B^4 \tag{B.32}$$

which after substituting the values of m_Ω^3, m_Ω^2 and m_Ω^1 derived above and simplifying leads to:

$$m_\Omega^4 = c^3 m_B^4 + 4c^2(m_B^3 m_B^1 + \frac{1}{2}\big(m_B^2\big)^2) + 6c m_B^2 \big(m_B^1\big)^2 + \big(m_B^1\big)^4. \tag{B.33}$$

Example 2: Consider a model where the idiosyncratic errors follow an AR(1) process $U_{j,t} = \rho\ U_{j,t-1} + \epsilon_{j,t}$, for $\epsilon_{j,t}$ white noise such that $E(\epsilon_{j,t}) = 0$ and $E(\epsilon_{j,t}^2) = \sigma^2$. Recall that $E(U_{j,t}^2) = \sigma^2/(1-\rho^2)$ and $E(U_{j,t}U_{j,t-k}) = (\sigma^2\rho^k)/(1-\rho^2)$. This implies a separable covariance model with $A_N = (\sigma^2/(1-\rho^2))I_N$ and $(B_T)_{m,n} = \rho^{|m-n|}$. Thus the model for the time series correlations B_T corresponds to a Toeplitz matrix where the first (main) diagonal is 1, the second (upper and lower) diagonals are ρ, the third (upper and lower) diagonals are ρ^2 etc. In order to guarantee that the spectrum of B_T is bounded we need to assume absolute summability, i.e.

$$\sum_{k=0}^{\infty} |\rho|^k = \frac{1}{1-|\rho|} < \infty, \tag{B.34}$$

which implies $|\rho| < 1$. In order to compute the eigenvalue distribution of the matrix B_T as $T \to \infty$ we define the Fourier series $f(\zeta)$ such that

$$f(\zeta) = \lim_{T\to\infty} \sum_{k=-\infty}^{+\infty} \rho^{|k|} \exp(ik\zeta) = \frac{1-\rho^2}{1-2\rho\cos(\zeta)+\rho^2}. \tag{B.35}$$

Let λ_k, $k = 1 \dots T$ be the eigenvalues of the matrix B_T for $T \to \infty$. Then, by a classic theorem of Grenander and Szego (1958), we have that for any positive integer s

$$m_B^s = \lim_{T\to\infty} \frac{1}{T} \sum_{k=1}^{T} \lambda_k^s = \frac{1}{2\pi} \int_0^{2\pi} [f(\zeta)]^s d\zeta. \tag{B.36}$$

TABLE 5. Estimation of the variance parameter σ^2 in Designs I and II.

Design I			Variance parameter								
			Mean Bias				MSE				
N	T	c	MD	MDW	MD-IC	MDW-IC	MD	MDW	MD-IC	MDW-IC	
30	100	0.3	-0.083	-0.062	-0.075	-0.060	0.008	0.005	0.007	0.005	
50	100	0.5	-0.076	-0.064	-0.068	-0.058	0.007	0.005	0.005	0.004	
70	100	0.7	-0.077	-0.074	-0.070	-0.059	0.007	0.006	0.006	0.004	
90	100	0.9	-0.077	-0.082	-0.069	-0.058	0.007	0.008	0.005	0.004	
90	300	0.3	-0.026	-0.020	-0.024	-0.019	0.001	0.000	0.001	0.000	
150	300	0.5	-0.025	-0.021	-0.022	-0.018	0.001	0.001	0.001	0.000	
210	300	0.7	-0.025	-0.023	-0.023	-0.019	0.001	0.001	0.001	0.000	
270	300	0.9	-0.025	-0.025	-0.023	-0.019	0.001	0.001	0.001	0.000	
150	500	0.3	-0.016	-0.013	-0.015	-0.012	0.000	0.000	0.000	0.000	
250	500	0.5	-0.015	-0.013	-0.013	-0.011	0.000	0.000	0.000	0.000	
350	500	0.7	-0.015	-0.014	-0.013	-0.011	0.000	0.000	0.000	0.000	
450	500	0.9	-0.014	-0.015	-0.013	-0.011	0.000	0.000	0.000	0.000	

Design II			Mean Bias				MSE				
N	T	c	MD	MDW	MD-IC	MDW-IC	MD	MDW	MD-IC	MDW-IC	
30	100	0.3	-0.078	-0.063	-0.065	-0.056	0.009	0.006	0.007	0.005	
50	100	0.5	-0.069	-0.059	-0.057	-0.046	0.006	0.005	0.004	0.003	
70	100	0.7	-0.070	-0.069	-0.060	-0.045	0.006	0.006	0.005	0.003	
90	100	0.9	-0.069	-0.076	-0.060	-0.041	0.006	0.007	0.004	0.002	
90	300	0.3	-0.026	-0.023	-0.019	-0.020	0.001	0.001	0.001	0.001	
150	300	0.5	-0.023	-0.021	-0.018	-0.015	0.001	0.001	0.000	0.000	
210	300	0.7	-0.023	-0.023	-0.020	-0.014	0.001	0.001	0.001	0.000	
270	300	0.9	-0.023	-0.024	-0.020	-0.014	0.001	0.001	0.000	0.000	
150	500	0.3	-0.017	-0.015	-0.012	-0.014	0.000	0.000	0.000	0.000	
250	500	0.5	-0.014	-0.013	-0.011	-0.009	0.000	0.000	0.000	0.000	
350	500	0.7	-0.014	-0.013	-0.012	-0.008	0.000	0.000	0.000	0.000	
450	500	0.9	-0.013	-0.014	-0.012	-0.008	0.000	0.000	0.000	0.000	

TABLE 6. Estimation of the variance parameters σ^2 and ρ in Designs III and IV.

Design III			Variance				Rho			
			Mean Bias		MSE		Mean Bias		MSE	
N	T	c	MD	MD-IC	MD	MD-IC	MD	MD-IC	MD	MD-IC
30	100	0.3	-0.080	-0.075	0.008	0.006	-0.065	-0.061	0.009	0.009
50	100	0.5	-0.073	-0.067	0.006	0.005	-0.044	-0.027	0.006	0.007
70	100	0.7	-0.074	-0.063	0.006	0.004	-0.037	-0.006	0.005	0.006
90	100	0.9	-0.075	-0.060	0.006	0.004	-0.019	0.012	0.004	0.005
90	300	0.3	-0.024	-0.023	0.001	0.001	-0.061	-0.050	0.006	0.006
150	300	0.5	-0.026	-0.021	0.001	0.000	-0.045	-0.016	0.004	0.003
210	300	0.7	-0.026	-0.020	0.001	0.000	-0.028	0.000	0.002	0.001
270	300	0.9	-0.024	-0.019	0.001	0.000	-0.011	0.006	0.001	0.001
150	500	0.3	-0.016	-0.013	0.000	0.000	-0.055	-0.038	0.004	0.004
250	500	0.5	-0.017	-0.012	0.000	0.000	-0.038	-0.008	0.002	0.001
350	500	0.7	-0.017	-0.012	0.000	0.000	-0.019	0.003	0.001	0.000
450	500	0.9	-0.015	-0.012	0.000	0.000	-0.005	0.005	0.000	0.000

Design IV			Variance				Rho			
			Mean Bias		MSE		Mean Bias		MSE	
N	T	c	MD	MD-IC	MD	MD-IC	MD	MD-IC	MD	MD-IC
30	100	0.3	-0.052	-0.015	0.005	0.002	0.046	0.216	0.020	0.071
50	100	0.5	-0.060	-0.033	0.005	0.002	0.002	0.122	0.007	0.027
70	100	0.7	-0.061	-0.034	0.005	0.002	0.002	0.099	0.005	0.017
90	100	0.9	-0.061	-0.030	0.005	0.001	0.006	0.092	0.003	0.013
90	300	0.3	-0.023	-0.012	0.001	0.000	-0.022	0.107	0.005	0.015
150	300	0.5	-0.026	-0.014	0.001	0.000	-0.031	0.051	0.003	0.004
210	300	0.7	-0.025	-0.014	0.001	0.000	-0.018	0.040	0.002	0.003
270	300	0.9	-0.022	-0.013	0.001	0.000	-0.004	0.038	0.001	0.002
150	500	0.3	-0.016	-0.008	0.000	0.000	-0.042	0.075	0.004	0.007
250	500	0.5	-0.017	-0.008	0.000	0.000	-0.033	0.035	0.002	0.002
350	500	0.7	-0.016	-0.009	0.000	0.000	-0.014	0.025	0.001	0.001
450	500	0.9	-0.013	-0.008	0.000	0.000	-0.003	0.025	0.000	0.001