

Structural Estimation of High-Dimensional Factor Models*

Matthew C. Harding,[†]

February 2009

Abstract

This paper develops econometric theory for the estimation of large N , large T factor models using structural restrictions from economic models. We employ non-commutative probability theory to derive a new estimator for the number of latent factors based on the moments of the eigenvalue distribution of the sample covariance matrix. Our test combines a minimum distance procedure for the estimation of structural model parameters with a specification test on the empirical eigenvalues to solve the problem of separating the factors from the noise. We also relate the second order unbiased estimation of the factor loadings to instrumental variable methods where the number of instruments is large relative to the sample size, and derive a number of alternatives to Principal Components, which have improved finite sample properties. This procedure is shown to perform very well in applications to asset pricing and data reduction of macroeconomic indicators.

JEL: C33, C32, C46, G11

Keywords: Factor Models, Random Matrix Theory, Principal Components, Free Probability

*I would like to thank Victor Chernozhukov, Alan Edelman, Jerry Hausman and Whitney Newey for their encouragement and support. I am grateful to Jushan Bai, Richard Blundell, Xiaohong Chen, Andrew Chesher, Graham Elliott, Iain Johnstone, Arthur Lewbel, Marcelo Moreira, Alexei Onatski, Raj Rao and James Stock for stimulating comments. Alex Nazarenko provided excellent research support. 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.

[†]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

This paper develops new techniques for the estimation of factor models in large datasets where the number of observations grows with the time dimension. Factor models relate observed data to a small set of unobserved variables which are then estimated. These models underlie many important tools of modern economics and finance, but no definitive econometric theory exists for the case of large panel datasets commonly encountered today.

We relate the identification and estimation of factor models to the asymptotic behavior of estimated eigenvalues of large random matrices, providing a connection between economics and the new mathematical field of Random Matrix Theory. In this paper we take a structural approach and show how the estimation of factor models can be improved by incorporating the economic assumptions of the model into the estimation procedure. In particular, we allow for arbitrary parametric forms of heteroskedasticity and autocorrelation.

We show that the key to identifying the number of latent factors lies in correctly understanding the structure of the noise (idiosyncratic effects) in the data. Once we can separate the estimated eigenvalues of a large factor model into those due to the latent structure and those due to the noise, we can construct a procedure that will consistently estimate the number of factors. Furthermore, we show that Principal Components Analysis (PCA) becomes unreliable for weak factors and relate the second order unbiased estimation of the factor loadings to recent advances in the estimation of models with instrumental variables when the number of instruments is large.

We also apply the methods developed in this paper to two stylized examples using financial and macroeconomic data. In particular, we look at the effect of adding additional observations on the estimation of a pricing model using noisy high-frequency data and at estimating the number of factors present in both US and Euro area macroeconomic data. We find that the methods outlined below perform very well and are ready to be used in more elaborate applications.

While factor models have been used for almost a century, standard 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 theory fails and becomes an unreliable statistical guide to data analysis. New econometric procedures for the estimation of high-dimensional factor models are the subject of active research (Bai and Ng, 2002; Stock and Watson, 2005; Onatski, 2006, 2008).

The application of econometric methods which take into account the special nature of large panel datasets leads us to reconsider stylized facts which we have taken for granted, such as the number of factors explaining most of the variation in financial returns. Harding (2008) shows that in a large panel data setup, the estimated eigenvalues corresponding to strong factors are severely upward biased in finite samples. Since the ratio of

the largest eigenvalue to the sum of all eigenvalues has been traditionally used to measure the effect of the factors, there is a bias towards accepting only a few (3-5) factors as explaining most of the data. In fact, many other factors may exist in the data and contain potentially valuable economic information.

This paper departs from the current literature on factor models insofar as it does not attempt to relax assumptions or attempt non-parametric techniques. Rather, it focuses on an often cited critique of factor models as inherently void of economic meaning and interpretation. This paper addresses this problem by emphasizing an approach which incorporates the information provided by economic models in the form of priors on the functional form or dependence structure between observations as well as the use of exclusion restrictions. The use of these “structural restrictions” adds value by substantially improving the identification of the model and precision of the estimates over leading methods such as Bai and Ng (2002). Furthermore, the use of restrictions derived from economic models improves our ability to interpret the estimated latent factors. While, factors are subject to arbitrary normalizations, the use of restrictions helps in fixing arbitrary rotations and provides a more precise economic understanding of them. For example, Harding (2008a) shows how the use of a financial factor model in conjunction with a New-Keynesian macroeconomic model can help identify supply shocks to the macroeconomy.

This paper further contributes to the literature on factor models by providing a new set of tools for their analysis based on recent developments in the mathematics of random matrices by employing results from free probability theory. These new methods provide a powerful approach to the study of the spectra of stochastic matrices. While it is possible to derive some of the results numerically by simulations, we show how free probability theory can be used to derive exact analytic expressions of the moments of eigenvalue distributions of large random matrices. Additionally, this paper complements and improves the recent application of Random Matrix Theory to factor models (Onatski, 2008). In situations where the structural restrictions are credible, our technique provides an exact determination of the number of factors rather than just a broad inequality constraint on the number of factors by using the information encoded in the bulk of the eigenvalues, without having to rely only on the leading eigenvalues.

The identification of the number of factors is central to the estimation of factor models and in Section 2 we show that it is possible to separate the identification of the number of factors from the estimation of the factor loadings and factor scores, and estimate the number of factors consistently in large factor models. In Section 3 we explain why PCA estimation is inconsistent for weak factors and develop a number of instrumental variable approaches to the estimation of factor loadings with excellent finite sample properties. Section 4 applies the newly developed econometric theory to two realistic examples in finance and macroeconomics.

2. Determining the Number of Factors

We are interested in the following large (N, T) panel data model with latent factors:

$$(2.1) \quad R_t = \Lambda F_t + U_t,$$

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 explain the variation in R_t with reference to a small dimensional set of latent factors F_t by decomposing the observed variation into a common component ΛF_t and an idiosyncratic component U_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 loadings or weights of the common factors F_t for each individual.

This particular statistical model originates in the work of Spearman and Hotelling and has been incorporated in many economic models of interest. The traditional econometric approach to solving this model was derived under the assumption that N is a fixed small number while T is large (Goldberger, 1972; Robinson, 1974; Zellner, 1970). With the availability of large dimensional panel data where both N and T are large, this model has received renewed attention and is currently an active area of research (Amengual and Watson, 2007; Bai and Ng, 2002; Onatski, 2006, 2008).

In finance the factor model of equation 2.1 corresponds to the Arbitrage Pricing Theory (APT) of Ross (1976), which explains the returns $R_{i,t}$ on $i = 1 \dots N$ assets observed over $t = 1 \dots T$ time periods by reference to a small set of risk factors F_t and asset specific shocks $U_{i,t}$. These multi-factor asset pricing models represent a major improvement over simpler single-factor CAPM models in evaluating the risk-return trade-off. While observable proxies have been used for the unobserved factors F_t in many applications, practitioners tend to agree that statistical factors derived from the econometric estimation of the model tend to outperform models evaluated by factor proxies (Miller, 2006).

More recently, the standard factor model above has been incorporated in more complex hybrid models involving both observed and latent factors:

$$(2.2) \quad BY_t + \Gamma Z_t + \Lambda F_t + U_t = 0,$$

where Y_t corresponds to a set of $N \times 1$ dimensional endogenous variable and Z_t is a set of $N \times 1$ dimensional observed exogenous variables with coefficients B and Γ respectively.

In microeconomics, such a model was first used by Gorman (1980) to analyze the characteristics of demand. In particular it is convenient to interpret the term $\Lambda F_t + U_t$ as a multifactor error structure or

interactive fixed effects (Bai, 2005; Pesaran, 2006). This model has recently received considerable attention in labor economics in the study of the relationship between education and earnings (Heckman and Navarro, 2007).

Factor models are also very popular in macroeconomics, where a forecasting model that uses factors constructed from numerous macroeconomic time series can substantially improve forecasting (Stock and Watson, 2006). The recent field of macro-finance has also relied on the estimation of factors from bond yields in order to improve the performance of small-scale structural macroeconomic models (Ang and Piazzesi, 2003).

The primary focus of this paper, however, is determining the number of factors and the estimation of the factor loadings in equation 2.1 for high-dimensional models where both N and T are large. This is captured by the following assumption:

Assumption 1 (Asymptotics): *The number of individuals increases with the sample size. Thus, $N \rightarrow \infty$ and $T \rightarrow \infty$ and $N/T \rightarrow c \in (0, \infty)$.*

This assumption is familiar to the literature on large N and T panel data (Hahn and Kuersteiner, 2002). The constant c , representing the limiting ratio of rows to columns in our panel, will play a very important role in the subsequent discussion.

The traditional statistical approach of Anderson and Rubin (1956) for solving factor models involves the assumption that the errors $U_{i,t}$ are uncorrelated both across individuals and across time. In many economics and finance applications, this assumption has proved to be too restrictive. In particular, Chamberlain and Rothschild (1983) show that if we allow for weak heteroskedasticity and time dependence of the error terms in the APT model, the mean returns are approximately linear in the factor loadings. Thus, the model remains correct under weak departures from the strict factor structure in the large N and T limit.

A major contribution of this paper is the development of an approach that allows us to deal with the approximate factor model where both heteroskedasticity and autocorrelations are possible. In order to do so, we need to describe the precise nature of the weak heteroskedasticity and autocorrelations compatible with the approximate factor model.

Let $U = [U_1, U_2, \dots, U_T]$ be the $N \times T$ matrix of errors in equation 2.1, with elements $U_{i,j}$ for $i = 1 \dots N$ and $t = 1 \dots T$, and where each column t corresponds to a realization of the errors at time t for the N individuals in the sample. Let $\text{vec}(U)$ be the $NT \times 1$ vector obtained by stacking the columns of U , i.e. $\text{vec}(U) = (U_1 U_2 \dots U_T)'$.

Assumption 2 (Cross-sectional and Time Dependence) *There is an $N \times N$ matrix A_N and a $T \times T$ matrix B_T such that $E(\text{vec}(U)) = 0$, $\text{Cov}(\text{vec}(U)) = A_N \otimes B_T$ and $E((U_{i,j})^4) < \infty$ for all N, T and A_N is unrelated to B_T .*

This assumption states that the errors U are mean zero and have finite fourth order moments. The most important assumption lies in 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. Note that this allows for very general forms of cross-sectional and time dependence, such as full correlation matrices. It does, however, limit the number of unknown parameters by assuming that the cross-sectional dependence and the time dependence are unrelated to each other.

In particular, note that one familiar distribution satisfying Assumption 2 is the matrix variate normal distribution (Arnold, 1981), where $\text{vec}(U) \sim \text{Normal}_{(N,T)}(0, A_N \otimes B_T)$, where the distribution function is given by:

$$(2.3) \quad f(U) = (2\pi)^{-\frac{NT}{2}} \det(A_N)^{-T/2} \det(B_T)^{-N/2} \exp \left\{ \text{tr} \left[-\frac{1}{2} U' A_N^{-1} U B_T^{-1} \right] \right\}.$$

In deriving the results in this paper, Assumption 2 above is sufficient for almost all proofs. Strictly speaking, the most general case of our estimator, as discussed in Section 2.4 which can handle arbitrary forms of cross-sectional and temporal dependence requires the stronger Normality assumption of equation 2.3. This is however, less a limitation of our procedure but rather a consequence of the fact that we use very recent theoretical results from free probability theory, which is still a very young discipline where many results remain unproven under more general assumptions. We have no reasons to believe it would not work under the more general assumptions above without explicitly requiring Normality.

The framework of approximate factor structure derives identification by requiring that the covariance matrix of the observations R_t has p unbounded eigenvalues corresponding to the latent factors and $N - p$ bounded eigenvalues corresponding to the idiosyncratic noise component. In order to account for these additional constraints, we first require some additional definitions. Consider the spectral decomposition of an arbitrary symmetric $n \times n$ matrix C_n . Since the matrix C_n is symmetric, we can find a matrix V with columns that are orthogonal to each other such that $A_n = V' D V$. The matrix $D = \text{diag}\{\lambda_1, \lambda_2, \dots, \lambda_n\}$ contains the set of eigenvalues of the matrix C_n , while the columns of V are the eigenvectors of C_n . We can now define the following proper cumulative distribution function $F^{C_n}(\lambda)$ on the spectrum $\lambda_i \in \{\lambda_1, \lambda_2, \dots, \lambda_n\}$ of C_n :

$$(2.4) \quad F^{C_n}(\lambda) = \frac{1}{n} \{\text{Number of Eigenvalues of } C_n \leq \lambda\} = \frac{1}{n} \sum_{\lambda_i \leq \lambda} 1.$$

Note that the spectrum does count multiplicities of eigenvalues.

Assumption 3 (Bounded Spectrum): *Denote by F^{A_N} and F^{B_T} the eigenvalue distribution of the matrix A_N and B_T respectively. Then as $N \rightarrow \infty$ and $T \rightarrow \infty$, $F^{A_N} \rightarrow F^A$ and $F^{B_T} \rightarrow F^B$, the eigenvalue distributions converge to non-random limiting distributions F^A and F^B . Moreover, let $\|\text{Sp}(A_N)\|$ and*

$\|\text{Sp}(B_T)\|$ denote the spectral norms of A_N and B_T . Assume that both spectral norms are bounded in N and T respectively.

This assumption is required in order to guarantee that the (scaled) eigenvalue distribution of $N^{-1}UU'$ converges to a non-random distribution as $N \rightarrow \infty$, and moreover that the spectrum of $N^{-1}UU'$ is also bounded (Silverstein and Bai, 1995; Bai and Silverstein, 1998). If $A_N = I_N$ and $B_T = I_T$ then the limiting distribution of $N^{-1}UU'$ converges to the limiting distribution of Marcenko and Pastur (1967) with support bounded on $[(1 - \sqrt{c})^2, (1 + \sqrt{c})^2]$ as discussed in Harding (2008b).

Assumption 2 and 3 rule out certain types of explosive behavior of the error terms and require that the variances converge for each time series to a finite value. Additionally, this assumption ensures that the moments of the empirical eigenvalue distribution of $N^{-1}UU'$ converge almost surely since the limiting distribution has bounded support. These moments will play an essential role in our procedure to estimate the number of factors.

Notice that the assumptions on the idiosyncratic error terms imposed above allow for a large range of empirically relevant models. Many forms of heteroskedasticity and autocorrelation are consistent with the assumptions on weak dependence. While some stochastic processes in the time-series literature are naturally excluded, such as unit roots, others such as structural breaks are allowed. In particular, notice that the assumptions above are consistent with structural breaks in the variance occurring at some unknown change points. It seems that these models have not been discussed so far in the literature on factor models, but they are undoubtedly important in the context of factor models based on large N, T panel data where we have the temptation to include data going back for many years and thus potentially covering more than one variance regime. To illustrate this point let us assume that we are interested in constructing a model using data before and after a financial crisis such as the Asian crisis of the late 1990s, but we are unsure as to the exact change point in the time series of the variances. If we assume τ periods to have been in the variance regime σ_1 and $T - \tau$ periods in the variance regime σ_2 , a simple model constructed along the lines of Assumptions 2 and 3 would be $A_N = I$ and $B_T = [\sigma_1 \otimes I_\tau] \oplus [\sigma_2 \otimes I_{T-\tau}]$, where \oplus stands for the direct sum of the two matrix spaces. Hence the spectrum of A_N consists of $\lambda_i = 1$ for $i = 1 \dots N$, while the spectrum of B_T consists of $\lambda_i = \sigma_1$ for $i = 1 \dots \tau$ and $\lambda_i = \sigma_2$ for $i = \tau + 1 \dots T$. These two spectra satisfy the boundedness condition of Assumption 2 and are thus admissible. Moreover, τ does not have to be determined a priori and can be a model parameter that is estimated at the same time as the number of factors. This illustrates the flexibility of our approach in the estimation of factor models by using the empirical eigenvalue distribution.

Assumption 4 (Pervasive factors): Assume that the factors F_t are independent of U_t . Denote by $\mu_0 = \min\{\text{Sp}(\frac{1}{T} \sum_{t=1}^T \Lambda F_t F_t' \Lambda')\}$, the smallest eigenvalue of the covariance of the factors weighted by the factor loadings. Then for all $N \rightarrow \infty$, there is some $M > 0$ and $M \rightarrow \infty$ such that $\mu_0 \geq M$.

Notice that this is consistent with both random and fixed factor loadings Λ . This assumption requires 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 will be revealed in large samples due to their effect on a large number of individuals. The pervasive factors reflect the structural part of an economic model. We can think of them as the systemic component of our model to be distinguished from the idiosyncratic noise perturbations. For example, we think of pervasive factors as global supply and demand shocks which impact a large number of firms simultaneously (Harding, 2008a). Note, however, that pervasive factors are not the only kind of factors one might be interested in. While not the subject of this paper, we could adjust the current methodology to estimate small scale factors that are related to a small number of firms but where the number of firms affected by them does not increase with the sample size. For example, several firms might rely on the same supply network and thus have correlated fluctuations. In certain circumstances, we might be able to distinguish these factors from the background idiosyncratic noise. These small scale factors are, however, excluded by Assumption 4 since for them $\mu_0 \rightarrow 0$ as $N \rightarrow \infty$.

Small scale factors might be particularly interesting if we wish to detect unusual correlations in large samples or individual idiosyncracies or may correspond to multicollinearity in the data, a topic that has received some attention in the statistics literature and may also have potential applications to portfolio choice. Assumption 4, however, corresponds to the current economic practice and mirrors the assumptions in Chamberlain and Rothschild (1983).

2.1. Factor Identification Strategy

The classical statistics literature on factor models recognizes that we can write the covariance matrix of the observations as:

$$(2.5) \quad \Sigma_N = \frac{1}{T} \sum_{t=1}^T R_t R_t' = \frac{1}{T} \sum_{t=1}^T \Lambda F_t F_t' \Lambda' + \frac{1}{T} \sum_{t=1}^T U_t U_t' = \Xi_N + \Omega_N,$$

where $\text{rank}(\Xi_N) = p$ and $N \rightarrow \infty$. Thus, the covariance matrix of the observations can be thought of as a finite (p) rank perturbation Ξ_N of the idiosyncratic noise covariance Ω_N . If we were to observe the population equivalents of our matrices Ξ_N and Ω_N , which we denote by Ξ_0 and Ω_0 , under the assumptions of our factor model, we would observe an infinite number of small bounded eigenvalues for Ω_0 and a small number of infinite eigenvalues for Ξ_0 . In finite samples, however, it has been noted that even for the simplest case of the strict factor model with $A_N = I_N$ and $B_T = I_T$ both the eigenvalues of Ξ_N and the eigenvalues of Ω_N increase with N (Brown, 1989). Until very recently this has prompted economists and statisticians to believe that factor identification based on the empirical distribution of eigenvalues is impossible (Bai and Ng, 2002). In this section we show how factor identification based on the empirical eigenvalue distribution is in fact possible and provides a very powerful new approach to factor analysis in a large class of models.

In order to identify the number of factors we rely on recent advances in the field of random matrix theory which provides mathematical tools that enable us to characterize the empirical eigenvalue distribution for many symmetric matrices (Edelman and Rao, 2005). Our approach is structural in that it relies on explicit assumptions about the form of cross-sectional and time dependence. In particular, Assumption 2 states that the covariance of the idiosyncratic terms is separable between a cross-sectional correlation matrix A_N and a time dependence matrix B_T . For many cases of interest it is sufficient to impose a specific structural form on these two matrices and parameterize these matrices as $A_N(\theta_A)$ and $B_T(\theta_B)$, where $\theta = (\theta_A, \theta_B)$ is a low dimensional vector of unknown structural covariance parameters. For example, the model with two variance regimes discussed above depends on $\theta = (\sigma_1, \sigma_2, \tau)$, where σ_1 and σ_2 are the two variances and τ/T is the probability of being in the first regime. Our procedure estimates the unknown parameter vector θ at the same time as the number of factors p . The main question in identifying the finite rank p perturbation due to the pervasive factors is how does the empirical eigenvalue distribution of Ω_N depend on (A_N, B_T) ? While in general we cannot analytically characterize the empirical distribution of Ω_N , in Section 2.2 we show that we can compute the moments of the empirical eigenvalue distribution of Ω_N in terms of A_N and B_T , which gives rise to a minimum distance estimation procedure and a downward testing procedure of the moment conditions that correctly identifies the number of factors.

In order to simplify mathematical notation we restrict our attention to the case where $B_T = I_T$ and discuss the remaining case in Section 2.4. 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.

Define the Cauchy Transform of an eigenvalue distribution function F^C for some matrix C as:

$$(2.6) \quad G_C(w) = \frac{1}{N} \lim_{N \rightarrow \infty} E \left\{ \text{tr} \left[\frac{1}{wI_N - C} \right] \right\} = \int \frac{1}{w - \lambda} F^C(\lambda)$$

for $w \in \mathbb{C}^+$ with $\Im(w) > 0$. 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. In particular, it allows us to recover the eigenvalue probability density function from the Stieltjes-Perron Inversion:

$$(2.7) \quad \frac{dF^C(\lambda)}{d\lambda} = -\frac{1}{\pi} \lim_{\xi \rightarrow 0} \Im[G_C(w + i\xi)].$$

First consider the limit distribution of the eigenvalues of the noise covariance matrix Ω_N .

Proposition 1: As $N \rightarrow \infty$, $T \rightarrow \infty$, $N/T \rightarrow c$, $A_N \rightarrow A$ and $F^{A_N} \rightarrow F^A$, the empirical eigenvalue distribution $F^{\Omega_N}(\lambda)$ converges to a non-random asymptotic distribution function $F^\Omega(\lambda; F^A, c)$ with bounded support.

Proof: See Silverstein (1995), Rao and Edelman (2006).

Notice that the asymptotic distribution depends on c and also on the asymptotic eigenvalue distributions of the matrices A (and B in the more general case) corresponding to the population values of the cross-sectional and time-series correlations. The resulting asymptotic distribution function can be derived implicitly for certain types of matrices A in terms of its Cauchy transformation, but requires numerical methods to evaluate (Rao and Edelman, 2006). Therefore, we focus our attention on a set of linear spectral statistics corresponding to the moments of the eigenvalue distribution, which have more convenient properties. For an arbitrary covariance matrix C we define

$$(2.8) \quad m_C(\lambda) = \int g(\lambda) dF^C(\lambda) = \frac{1}{N} \sum_{j=1}^N f(\lambda_j),$$

where $\lambda_j \in \text{Sp}(C)$. We are especially interested in the monomials, $g(\lambda) \in \{\lambda, \lambda^2, \dots, \lambda^s\}$ and denote the corresponding moments by $\{m_C^1, m_C^2, \dots, m_C^s\}$. Notice that these monomials define the raw moments of the eigenvalue distribution F^C .

If C is some empirical covariance matrix C_N , then

$$(2.9) \quad m_{C_N}^s(\lambda) = N^{-1} \text{tr}[(C_N)^s].$$

Moreover, standard results on bounded moment convergence and continuous mapping (e.g. Billingsley, 1995) imply that if $C_N \rightarrow C$ and $F^{C_N} \rightarrow F^C$, a proper probability distribution with bounded support, then:

$$(2.10) \quad \lim_{N \rightarrow \infty} m_{C_N}^s(\lambda) = m_C^s(\lambda) = \int \lambda^s dF^C(\lambda) < \infty.$$

In particular note that for the covariance model introduced in Assumption 2, the moments of the eigenvalue distribution of the error term in our factor model exist and are finite as a consequence of Proposition 1. The challenge consists of being able to compute the limiting moments of the eigenvalue distribution. Below we introduce a procedure based on free probability theory that relates the moments of the limiting eigenvalue distribution to the moments of the eigenvalue distribution of the cross-sectional and time-series correlation matrices.

Moreover, it can be shown that the moments of the eigenvalue distribution of a random covariance matrix C_N satisfy a Central Limit Theorem (Bai and Silverstein, 2004):

Proposition 2 (CLT): *Let $\bar{g}(w) = -(1-c)/w + cG_C(w)$. Then*

$$(2.11) \quad N^{-1} \begin{pmatrix} m_{C_N}^1 & - & m_C^1(\lambda) \\ & \dots & \\ m_{C_N}^s & - & m_C^s(\lambda) \end{pmatrix} \sim N(\Delta, V),$$

where for $j = 1 \dots s$ and $k = 1 \dots s$

$$(2.12) \quad \Delta_j = -\frac{1}{2\pi i} \int w^j \frac{c \int \bar{g}(w)^3 v^2 (1 + v\bar{g}(w))^{-3} dF^A}{\left(1 - c \int \bar{g}(w)^2 v^2 (1 + v\bar{g}(w))^{-2} dF^A\right)^2} dw$$

$$(2.13) \quad V_{jk} = -\frac{1}{2\pi^2} \iint \frac{w_1^j w_2^k}{(\bar{g}(w_1) - \bar{g}(w_2))^2} \frac{d\bar{g}(w_1)}{dw_1} \frac{d\bar{g}(w_2)}{dw_2} dw_1 dw_2,$$

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^C .

In general however it is not possible to compute these integrals over the complex plane analytically. For $A = I_N$ the answer is known and was derived by Jonsson (1982) using a combinatoric proof. In this case the expressions above reduce to:

$$(2.14) \quad \Delta_j = \frac{1}{4} \left((1 - \sqrt{c})^{2j} + (1 + \sqrt{c})^{2j} \right) - \frac{1}{2} \sum_{r=0}^j \binom{j}{r}^2 c^r$$

$$(2.15) \quad (V)_{j,k} = 2c^{j+k} \sum_{r_1=0}^{j-1} \sum_{r_2=0}^k \binom{j}{r_1} \binom{k}{r_2} \left(\frac{1-c}{c}\right)^{j+k} \sum_{l=1}^{j-k} l \binom{2j-1-(r_1+l)}{j-1} \binom{2k-1-(r_2+l)}{k-1}$$

Recall that our strategy for identifying the number of latent factors involves performing an eigenvalue decomposition of the covariance matrix between the observed time series, Σ_N . Moreover, we have assumed that the covariance matrix between the unobserved error terms is Ω_N , where the error terms were drawn from a Normal matrix variate distribution with separable cross-sectional and time series correlation (Assumption 2). Note that by Proposition 1, the empirical eigenvalue distribution of Ω_N converges to some non-random proper distribution function $F^\Omega(\lambda; \theta, c)$ as $N \rightarrow \infty$, $T \rightarrow \infty$ and $N/T \rightarrow c$ where θ is the unknown vector of covariance parameters.

Let us assume for the moment that our data does not contain unobserved latent factors, that is $\Xi_N = 0$. In this case, Proposition 2 suggests a minimum distance procedure for estimating the vector of unknown covariance parameters θ using the moments of the empirical eigenvalue distribution. Let

$$(2.16) \quad \Pi(\hat{\Omega}_N) = [N^{-1} \text{tr}(\Omega_N^1), N^{-1} \text{tr}(\Omega_N^2), \dots, N^{-1} \text{tr}(\Omega_N^s)]',$$

be the vector of the first s moments of the empirical eigenvalue distribution of Ω_N and denote by

$$(2.17) \quad \Pi(\theta) = [m_\Omega^1(\theta, c), m_\Omega^2(\theta, c), \dots, m_\Omega^s(\theta, c)]',$$

the corresponding vector of limiting moments as $N \rightarrow \infty$, $T \rightarrow \infty$ and $N/T \rightarrow c$. In Section 2.2 we show how to derive expressions for these limiting moments analytically. In order to estimate the vector of

unknown parameters, we could apply the following minimum distance procedure:

$$(2.18) \quad \hat{\theta} = \operatorname{argmin}_{\theta} \left(\Pi(\theta) - \Pi(\hat{\Omega}_N) \right)' \hat{V}^{-1} \left(\Pi(\theta) - \Pi(\hat{\Omega}_N) \right),$$

where $(\hat{V})_{j,k}$ is a consistent estimate of equation 2.13.

Our focus however, is on estimating the rank of the matrix Ξ_N when $\operatorname{rank}(\Xi_N > 0)$ subject to the constraint that we only observe $\Sigma_N = \Xi_N + \Omega_N$. This implies that we have to use the spectral decomposition of the covariance matrix Σ_N to estimate both the rank of Ξ_N and any additional covariance parameters θ that Ω_N depends on.

By the assumptions of our factor model, $\operatorname{rank}(\Xi_N) = p \ll N$. Moreover, we know that the p eigenvalues capturing the effect of the latent factors F_t diverge to infinity as $N \rightarrow \infty$ while the N eigenvalues corresponding to the noise term U_t remain bounded. In large enough samples, this produces 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 $N \rightarrow \infty$ (Bai and Silverstein, 1998, 2004; Baik and Silverstein, 2006; Dozier and Silverstein, 2007). In particular note that the p eigenvalues do not “pull” the remaining $N - p$ eigenvalues to the right. Identifying the number of latent factors thus requires us to estimate the number of eigenvalues to the right of this spectral gap which separates the eigenvalues due to the noise term U_t from those due to the latent factors. In finite samples however this gap is not evident due to the presence of weak factors for which the corresponding eigenvalues are close to the upper bound of the distribution of eigenvalues due to the idiosyncratic terms. Thus, we require statistical techniques in order to separate the eigenvalues due to the factors from those due to the noise.

If we were to compare the asymptotic moment expressions of equation 2.17 for $\Pi(\theta)$, with the moments of the empirical eigenvalue distribution of the observed covariance matrix Σ_N ,

$$(2.19) \quad \Pi(\Sigma_N) = \left[N^{-1} \operatorname{tr}(\Sigma_N^1), N^{-1} \operatorname{tr}(\Sigma_N^2), \dots, N^{-1} \operatorname{tr}(\Sigma_N^s) \right]',$$

we would find a poor match. The asymptotic moment expressions are correct for the unobserved covariance matrix Ω_N but not for the observed covariance matrix $\Sigma_N = \Xi_N + \Omega_N$. We can exploit this inconsistency in the moment conditions of the eigenvalue distribution which occurs in the presence of latent factors to specify a downward testing moment selection procedure (Andrews, 1999).

Let $\operatorname{Sp}(\Sigma_N)$ denote the spectrum of the covariance matrix of the observations, Σ_N , where we have ordered the eigenvalues in decreasing order. That is, $\operatorname{Sp}(\Sigma_N) = \{\lambda_1, \lambda_2, \dots, \lambda_N\}$, with λ_1 being the largest eigenvalue and including multiplicities. Note that expression 2.19 can be re-written as:

$$(2.20) \quad \Pi(\Sigma_N) = \left[N^{-1} \sum_{\lambda_j \in \operatorname{Sp}(\Sigma_N)} \lambda_j^1, N^{-1} \sum_{\lambda_j \in \operatorname{Sp}(\Sigma_N)} \lambda_j^2, \dots, N^{-1} \sum_{\lambda_j \in \operatorname{Sp}(\Sigma_N)} \lambda_j^s \right]'$$

Now let $\text{Sp}_p(\Sigma_N)$ be the truncated spectrum where we have removed the first p largest eigenvalues, $\text{Sp}_p(\Sigma_N) = \text{Sp}(\Sigma_N) \setminus \{\lambda_1, \lambda_2, \dots, \lambda_p\}$. Let $\Pi_p(\Sigma_N)$ be the vector of the first s empirical moment conditions evaluated using the truncated spectrum $\text{Sp}_p(\Sigma_N)$,

$$(2.21) \quad \Pi_p(\Sigma_N) = \left[N^{-1} \sum_{\lambda_j \in \text{Sp}_p(\Sigma_N)} \lambda_j^1, N^{-1} \sum_{\lambda_j \in \text{Sp}_p(\Sigma_N)} \lambda_j^2, \dots, N^{-1} \sum_{\lambda_j \in \text{Sp}_p(\Sigma_N)} \lambda_j^s \right]'$$

Notice that if p is the true number of latent factors then the moment conditions $\Pi_p(\Sigma_N)$ will match with the asymptotic moments for the covariance of error terms $\Pi(\theta)$ from expression 2.20 above. By evaluating the moment conditions on the truncated spectrum, we have removed the effect of the latent factors and we expect the distance between $(\Pi_p(\Sigma_N) - \Pi(\theta))$ to be small if the correct number of factors p has been identified and large otherwise. This suggests a testing procedure based on the minimized objective function, commonly referred to as the J -test (Hansen, 1982). The number of unobserved factors, \hat{p} is estimated by:

$$(2.22) \quad \hat{p} = \underset{p=0,1,2,\dots}{\text{argmin}} \hat{J}(\hat{\theta}; \text{Sp}_p(\Sigma_N)),$$

where the vector of unknown covariance parameters is estimated using the moment conditions computed from the truncated spectrum $\text{Sp}_p(\Sigma_N)$. The procedure is applied recursively by truncating the spectrum of the observed covariance matrix Σ_N from the right and re-estimating the vector of parameters θ until the corresponding J -test is minimized. In Section 2.3 we show that the J -test is minimized after the spectrum of the observed covariance matrix Σ_N has been truncated by the true number of factors, thereby estimating \hat{p} consistently.

Notice that the true number of factors is only revealed asymptotically as $N \rightarrow \infty$. Identifying the true number of pervasive factors requires the sample to be large enough such that the spectrum separates between a set of $N-p$ eigenvalues corresponding to the error terms U_t and a set of p eigenvalues corresponding to the factors F_t . Unfortunately, in small samples the eigenvalues corresponding to the p factors may not always separate and exhibit a phase transition phenomenon where eigenvalues corresponding to weak factors do not detach from the spectrum of the error terms and converge in probability to the upper bound of the spectrum of Ω_N rather than to their true asymptotic limits which diverge with N . Harding (2008b) shows how this leads to a single factor bias in estimated arbitrage pricing models commonly used in finance. Note that this is not a feature of the estimation procedure but rather of the data; in small samples, some weak factors are obfuscated by the error terms and cannot be identified. Our procedure, however, does guarantee that if a factor is strong enough to overcome the phase transition phenomenon and emerge from the shadow of the error terms, it will be picked up by our algorithm and will be correctly identified as a latent factor. This allows us to estimate all the weak factors which can be identified from the data in addition to the strong factors which have traditionally been estimated.

2.2. Free Probability Derivation of Moments

In order to implement the identification strategy outlined above, we need to compute the limiting moments of the eigenvalue distribution of the covariance of the error terms U_t as a function of the covariance model parameters θ ,

$$(2.23) \quad m_{\Omega}^s = \lim_{N \rightarrow \infty} (1/N) E \{ \text{tr}(\Omega_N^s) \}.$$

Ignoring time series correlations for the moment, one of the implications of Assumption 2 is that we can write $\Omega_N = (1/T)UU' = (1/T)A_N^{1/2}\epsilon\epsilon'A_N^{1/2}$, where $U = A_N^{1/2}\epsilon$ and $(\epsilon)_{i,j}$ is iid mean zero with finite fourth order moments. Moreover, we assume that A_N and ϵ are independent. Denote the covariance of the iid terms $(\epsilon)_{i,j}$ by $\Psi_N = (1/T)\epsilon\epsilon'$ and notice that:

$$(2.24) \quad \begin{aligned} m_{\Omega}^s &= \lim_{N \rightarrow \infty} (1/N) E \left\{ \text{tr} \left(\left[(1/T) A_N^{1/2} \epsilon \epsilon' A_N^{1/2} \right]^s \right) \right\} \\ &= \lim_{N \rightarrow \infty} (1/N) E \{ \text{tr}((1/T) [A_N \Psi_N]^s) \}. \end{aligned}$$

The focus of this section is to introduce a procedure to analytically derive the large N limiting eigenvalue distribution moments m_{Ω}^s based on our knowledge of the limiting eigenvalue distribution of A_N and Ψ_N . Note, however, that even though by assumption A_N and Ψ_N have limiting eigenvalue distributions with bounded support and A_N is independent of Ψ_N this is not sufficient to guarantee that the eigenvalue distribution of $(A_N \Psi_N)^s$ will depend only on the underlying limiting distributions. The free probability approach developed below will provide additional conditions under which it is possible to relate the moments of the limiting eigenvalue distribution of mixed moments of products of random matrices to the limiting moments of the eigenvalues of their constituent matrices.

The computation of expressions such as those of the moments m_{Ω}^s can be prohibitive analytically if we start from the individual elements of the random matrix due to the combinatoric complexity of the resulting traces of powers of mixed products between matrices. Instead we prefer to think of the whole random matrix as a random variable on a non-commutative probability space defined below.

Consider a probability space Θ and the Hilbert space $L^\infty(\Theta)$ of bounded measurable functions h defined on Θ , corresponding to the random variables on Θ . The functions h are allowed to be complex valued but for the purpose of estimating a factor model we can restrict our attention to real random variables on Θ . Furthermore, there exists a probability law \mathfrak{P} on h which measures the probability that the value of h lies in a certain sub-interval in the (real or complex) image of h . The space of bounded linear functionals ϕ on the Hilbert space $L^\infty(\Theta)$ given by $\phi(h) = \int_{\Theta} h d\mathfrak{P}$ defines a von Neumann algebra, \mathcal{A} , on Θ .

Definition 1: *A non-commutative probability space is a pair (\mathcal{A}, ϕ) , where \mathcal{A} is an algebra endowed with a unit (1) and ϕ a linear functional on \mathcal{A} such that $\phi(1) = 1$.*

Note that classical probability spaces also satisfy the above definition but that we are interested in relaxing the commutativity assumption imposed by classical probability on scalar random variables. Furthermore, it is convenient to also assume that \mathcal{A} is a von Neumann algebra as discussed above.

For the non-commutative probability space (\mathcal{A}, ϕ) 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 from operator algebras (Voiculescu, 1985; Speicher, 2005).

Definition 2 (Freeness): *Let (\mathcal{A}, ϕ) be the non-commutative probability space of Definition 1 and $\{\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_J\} \subset \mathcal{A}$ subalgebras of \mathcal{A} with the a unit (1). Then the algebras $\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_J$ are free with respect to ϕ if*

$$(2.25) \quad \phi(a_1 a_2 \dots a_K) = 0$$

if $a_1 \in \mathcal{A}_{j(1)}, a_2 \in \mathcal{A}_{j(2)}, \dots, a_K \in \mathcal{A}_{j(K)}$, where $j(k)$ is an index function on the set $\{1, 2, \dots, J\}$ and $j(k) \neq j(k+1)$ for all $k = 1 \dots (K-1)$, and $\phi(a_k) = 0$ for all $k = 1 \dots K$.

By extension, the random variables $X_1, X_2, \dots, X_K \in \mathcal{A}$ 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. Note, however, that this particular extension of independence to non-commutative probability spaces is but one of the possible non-commutative extensions with convenient properties for the analysis of covariance matrices.

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 equation 2.25 for the case when $\phi(a_k) \neq 0$ by subtracting the individual means:

$$(2.26) \quad \phi((a_1 - \phi(a_1)1)(a_2 - \phi(a_2)1) \dots (a_k - \phi(a_k)1)) = 0.$$

Our primary focus is on the computation of mixed moments $m_{ab}^s = \phi((ab)^s)$. If a and b are free then,

$$(2.27) \quad \phi((a - \phi(a)1)(b - \phi(b)1)) = 0,$$

and expanding,

$$(2.28) \quad \phi(ab - \phi(a)b - a\phi(b) + \phi(a)\phi(b)) = 0,$$

$$(2.29) \quad \phi(ab) - \phi(a)\phi(b) - \phi(a)\phi(b) + \phi(a)\phi(b) = 0,$$

$$(2.30) \quad m_{ab}^1 = \phi(ab) = \phi(a)\phi(b).$$

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, $m_{ab}^2 = \phi((ab)^2)$. Since a and b are non-commutative, $m_{ab}^s = \phi(abab)$, we start by expanding the expression:

$$(2.31) \quad \phi((a - \phi(a)1)(b - \phi(b)1)(a - \phi(a)1)(b - \phi(b)1)) = 0.$$

In the Appendix we show that this leads to the following expression for the mixed second moment in a and b :

$$(2.32) \quad \phi(abab) = \phi^2(a)\phi(bb) + \phi^2(b)\phi(aa) - \phi^2(a)\phi^2(b).$$

This expression however 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

$$(2.33) \quad \phi(abab) \neq \phi(a^2b^2) = \phi(a^2)\phi(b^2).$$

The definition of free independence can thus be applied recursively to obtain the mixed higher order moments of ab and other similar products.

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 (Θ, \mathfrak{P}) . Define the algebra of functions

$$(2.34) \quad \mathcal{A}_N = \bigcap_{1 \leq s < \infty} L^s(X, \mathcal{M}_N)$$

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 $\phi_N : \mathcal{A} \rightarrow \mathbb{R}$ be an operator defined as:

$$(2.35) \quad \phi_N(Y) = \frac{1}{N} E_X \operatorname{tr}(Y) = \frac{1}{N} \sum_{j=1}^N E(Y_{j,j}) = \frac{1}{N} \int_X \operatorname{tr}(Y) d\mathfrak{P}.$$

Corollary 1: (\mathcal{A}_N, ϕ_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 random matrices both in terms of the usual commutative probability spaces on which each element of the matrix is defined but also in terms of the whole matrix as a random variable defined on a non-commutative probability space. In particular given the connection between the trace and the eigenvalues of the matrix, it turns out to be more convenient to think of the covariance matrices in our factor model in terms of the non-commutative probability space.

Recall that our interest in using non-commutative probability is mainly due to the necessity of computing moments of the type $\phi_N[(A_N \Psi_N)^s]$. In order to employ the moment expansion procedure above, we would first need to show that (A_N, Ψ_N) are freely independent. In general, however, two arbitrary matrices are not freely independent since their eigenspaces may satisfy a particular relationship to each other, even if the elements of the matrices are independent.

One of the main insights of random matrix theory is that certain matrices become freely independent asymptotically as $N \rightarrow \infty$ (Voiculescu, 1998). Note that asymptotic freeness for large random matrices Y requires both the convergence of the probability law \mathfrak{P} as $N \rightarrow \infty$ and Definition 2 to be satisfied for $\phi_N(Y) = \lim_{N \rightarrow \infty} \frac{1}{N} E_X \text{tr}(Y)$. 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 all $\Gamma \in \mathcal{O}(N)$. The distribution μ is referred to as the Haar (invariant) distribution on $\mathcal{O}(N)$.

Proposition 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 $N \rightarrow \infty$.

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 3.1: Let C_N be an arbitrary matrix of order N and define the following normalization:

$$(2.36) \quad J(C_N) = \text{diag} \left\{ (C_N)_{1,1} / |(C_N)_{1,1}|, (C_N)_{2,2} / |(C_N)_{2,2}|, \dots, (C_N)_{N,N} / |(C_N)_{N,N}| \right\}.$$

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:

Proposition 4: *Let $W_N = (w_1, w_2, \dots, w_N)'$ 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).

Returning to the moment expressions 2.23 and 2.24, we see that A_N and Ψ_N are asymptotically free if Ψ_N is unitarily invariant. An important special case is given by Proposition 4 where Ψ_N is drawn from a standard Wishart distribution, that is $\Psi_N = (1/T)\epsilon\epsilon'$, where $(\epsilon)_{i,j}$ are distributed iid $N(0,1)$. This implies that we can apply Definition 2 and compute mixed moments of $(A_N \Psi_N)^s$ using the recursive procedure outlined above. Since the sequence of matrices A_N is given by our parametric model (or estimated by some alternative procedure), it has known moments $m_{A_N}^s$ for all N as $N \rightarrow \infty$. The moments of Ψ_N are known to converge to the moments of the Marcenko-Pastur distribution under a general set of assumptions.

Proposition 5: *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 $N \rightarrow \infty$, $T \rightarrow \infty$ and $N/T \rightarrow c$, 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:*

$$(2.37) \quad m_{\Psi}^s = \lim_{N \rightarrow \infty} \frac{1}{N} E \text{tr} \left\{ [(1/T)\epsilon\epsilon']^s \right\} = \sum_{r=1}^s \frac{1}{s} \binom{s}{r} \binom{s}{r-1} c^{s-1}.$$

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

$$(2.38) \quad m_{\Psi}^1 = 1$$

$$(2.39) \quad m_{\Psi}^2 = 1 + c$$

$$(2.40) \quad m_{\Psi}^3 = 1 + 3c + c^2$$

$$(2.41) \quad m_{\Psi}^4 = 1 + 6c + 6c^2 + c^3$$

$$(2.42) \quad m_{\Psi}^5 = 1 + 10c + 20c^2 + 10c^3 + c^4$$

$$(2.43) \quad m_{\Psi}^6 = 1 + 15c + 50c^2 + 50c^3 + 15c^4 + c^5.$$

We can now use the moments given by equation 2.37 and the expressions in equation 2.30 and 2.32 to compute the mixed moments of $A_N\Psi_N$ in the large N limit using the property of free independence between A_N and Ψ_N . From equation 2.30 we know that $m^1(A\Psi) = m_A^1 m_{\Psi}^1$. But since $m_{\Psi}^1 = 1$, we have

$$(2.44) \quad m^1(A\Psi) = m_A^1.$$

Similarly by equation 2.32 we know that

$$(2.45) \quad m^2(A\Psi) = (m_A^1)^2 m_{\Psi}^2 + (m_{\Psi}^1)^2 m_A^2 - (m_A^1)^2 (m_{\Psi}^1)^2.$$

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

$$(2.46) \quad m^2(A\Psi) = (m_A^1)^2 (1 + c) + m_A^2 - (m_A^1)^2$$

$$(2.47) \quad m^2(A\Psi) = m_A^2 + c(m_A^1)^2.$$

We can continue this process to obtain:

$$(2.48) \quad m^3(A\Psi) = m_A^3 + 3cm_A^1 m_A^2 + c^2(m_A^1)^3$$

$$(2.49) \quad m^4(A\Psi) = m_A^4 + 2c \left((m_A^2)^2 + 2m_A^1 m_A^3 \right) + 6c^2 (m_A^1)^2 m_A^2 + c^3 ((m_A^1)^4).$$

Thus, we have shown how to compute the moments of the eigenvalue distribution of the covariance of the error terms U_t , $\Omega_N = (1/T)UU' = (1/T)A_N^{1/2}\epsilon\epsilon'A_N^{1/2}$ in terms of the moments of the eigenvalue distribution of A_N in the large N limit. Since these moments will be functions of the unknown parameters θ , the expressions derived above will also be functions of θ once we substitute a precise covariance model for A_N . To illustrate, consider the model with $A_N = \sigma I_N$. We have only one unknown parameter, the variance

scale coefficient σ , $\theta = \{\sigma\}$. Substituting the moments of A in equations 2.44, 2.47, 2.48 and 2.49, we obtain the first four spectral moments of the white noise covariance matrix to be:

$$(2.50) \quad m_{\Omega}^1 = \sigma$$

$$(2.51) \quad m_{\Omega}^2 = (1 + c) \sigma^2$$

$$(2.52) \quad m_{\Omega}^3 = (c^2 + 3c + 1) \sigma^3$$

$$(2.53) \quad m_{\Omega}^4 = (1 + c) (c^2 + 5c + 1) \sigma^4.$$

The free probability framework introduced above allows us to compute the moments of the empirical eigenvalue distribution of the noise covariance matrix in terms of the population covariance matrix assumed by our factor model by a number of algebraic operations on free moments. While these computations are relatively straightforward and only involve basic algebra, higher order moments may involve a substantial number of terms and thus it may be more convenient to use a mathematical software package such as Maple or Mathematica to derive the moment expressions. It is also possible to derive the moment expressions using the Cauchy transform defined in equation 2.6. In the Appendix we show that the moment expressions derived above are also more generally given by the following implicit relationship:

Proposition 6: *Let m_{Ω}^s be the limiting moments of the empirical noise covariance Ω and m_A^s be the limiting moments of the correlation matrix A . Then for $w \in \mathbb{C}^+$ with $\Im(w) > 0$ we have that*

$$(2.54) \quad \sum_{s=1}^{\infty} \frac{m_{\Omega}^s}{w^s} = \sum_{s=1}^{\infty} \frac{m_A^s}{w^s} \left(1 + c \sum_{r=1}^{\infty} \frac{m_{\Omega}^r}{w^r} \right)^s.$$

The relationship between the first four moments is given by:

$$(2.55) \quad m_{\Omega}^1 = m_A^1$$

$$(2.56) \quad m_{\Omega}^2 = m_A^2 + c(m_A^1)^2$$

$$(2.57) \quad m_{\Omega}^3 = m_A^3 + 3cm_A^1m_A^2 + c^2(m_A^1)^3$$

$$(2.58) \quad m_{\Omega}^4 = m_A^4 + 2c \left((m_A^2)^2 + 2m_A^1m_A^3 \right) + 6c^2(m_A^1)^2m_A^2 + c^3((m_A^1)^4).$$

Proof: See Appendix. In order to derive the expressions for the moments in equations 2.44, 2.47, 2.48 and 2.49 we can expand this expression in $1/w$ and match the coefficients on $1/(w^s)$.

In the next section we revisit our estimator for the number of factors based on the identification strategy outlined in Section 2.1 and the procedure for deriving the limiting moments of the eigenvalue distribution as described in this section and analyze its statistical properties.

2.3. Implementation and Finite Sample Performance

Recall the basic model setup of our factor model, $R_t = \Lambda F_t + U_t$, for $t = 1 \dots T$ where 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. The identification strategy outlined above implies a computational procedure that leads to the consistent estimation of the number of factors p in the factor model with $N \rightarrow \infty$, $T \rightarrow \infty$ and $N/T \rightarrow c$. The advantage of this procedure consists in that it does not require the estimation of the unknown factor loadings Λ and factor scores F_t first and is therefore unaffected by complications resulting from the estimation of weak factor scores, which will be discussed in Section 3.2.

We can now summarize the steps required for the implementation of our estimator for the number of factors. For simplicity we continue to assume that $B_T = I_T$. This assumption will be relaxed in the next section.

First, we choose a parametric model for the idiosyncratic error terms U_t in terms of the population covariance matrix $A_N(\theta)$, where the correlations are described in terms of the low dimensional parameter vector θ . Second, we compute the moments of the eigenvalue distribution of $A_N(\theta)$ for large N , $\{m_A^1, m_A^2, m_A^3, \dots\}$. Third, we apply the free probability result of Proposition 6 to compute the moments of the asymptotic eigenvalue distribution of the covariance matrix Ω of U_t for a large (N, T) sample drawn from a distribution with covariance matrix A_N . We label these moments as $\Pi(\theta) = [m_\Omega^1, m_\Omega^2, m_\Omega^3, \dots]'$. Fourth, we use a minimum distance approach to estimate the unknown covariance parameters θ by minimizing a weighted distance between $\Pi(\theta)$ and its sample equivalent $\Pi(\Sigma_N) = [N^{-1} \text{tr}(\Sigma_N^1), N^{-1} \text{tr}(\Sigma_N^2), \dots]$ applied to the covariance matrix of observations R_t denoted by Σ_N . Fifth, we remove the largest eigenvalue of the spectrum of Σ_N and re-estimate the parameters θ using the minimum distance procedure. We repeat step 5 by progressively removing large eigenvalues until an (arbitrary) upper bound on the number of factors has been reached. Sixth, we compare the minimized objective functions, $\hat{J}(\hat{\theta})$ obtained by removing large eigenvalues and choose the one which is smallest within the set of minimized objective functions. The number of eigenvalues which had been removed for the computation of that objective function is our consistent estimate of the number of factors.

Proposition 7: *Let Σ_N be the covariance matrix of observations R_t in a large N, T factor model with $N \rightarrow \infty$, $T \rightarrow \infty$ and $N/T \rightarrow c$. Let $\text{Sp}_p(\Sigma_N)$ be the spectrum of the matrix Σ_N where we removed the largest p eigenvalues and $\hat{J}(\hat{\theta}; \text{Sp}_p(\Sigma_N))$ the (scaled) minimized objective function of the minimum distance procedure for the estimation of θ outlined above. Then, for*

$$(2.59) \quad \hat{p} = \underset{p=0,1,2,\dots}{\text{argmin}} \hat{J}(\hat{\theta}; \text{Sp}_p(\Sigma_N)),$$

is a consistent estimate of the number of factors p_0 of the factor model $R_t = \Lambda F_t + U_t$.

Proof: See Appendix.

In order to implement the estimation procedure described above we compute the minimum distance estimates of θ using the Nelder-Mead algorithm available in most common software packages such as Matlab or Gauss. The optimal weighting matrix for the moment conditions Π is difficult to compute analytically for the approximate factor model with arbitrary A_N . By Proposition 2, however, it may be possible to use the bootstrap or the jackknife to estimate the weighting matrix, since the moments are asymptotically Normal. The properties of such a procedure are not known.

For the strict factor model with iid errors, the optimal weighting matrix is easy to implement and we can use an efficient two-step procedure which uses the estimated $\hat{\theta}$ from a first step estimation that employs equal weighting of the moment conditions. We label the one-step minimum distance procedure MD and the two-step weighted estimator MDW.

In simulations we have also found that the performance of our estimator can be improved by adding a panel information criterion which penalizes the objective function in equation 3.13 if the selected number of factors is too large. The intuition is that in some cases the difference between the estimated \hat{J} at p_0 and at $p_0 + 1$ may be very small. In such cases it is beneficial to augment equation 3.13 with a penalty function of the form $p\hat{\sigma}^2g(N, T)$, where p is the number of excluded eigenvalues, $\hat{\sigma}^2$ is the estimated (average) variance at step p and $g(N, T)$ a function such that $g(N, T) \rightarrow 0$ in large samples. In simulations we have found the following choice due to Bai and Ng (2002) to perform very well:

$$(2.60) \quad g(N, T) = \left(\frac{N+T}{NT} \right) \log \left(\frac{NT}{N+T} \right).$$

We can augment the estimators MD and MDW defined above with the additional penalty function $p\hat{\sigma}^2g(N, T)$ to obtain two alternative estimators which we label MD-IC and MDW-IC.

In Figure 1 we plot the objective function given by equation 2.59 for a particular simulation of the exact factor model with 5 factors using the design given in the Appendix. The objective function is minimized for all four choices of estimators of the number of factors (MD, MDW, MD-IC, MDW-IC) at the correct number of factors. In Table 1 and 2 we explore the finite sample properties of our estimators for different choices of N and T such that $c \in \{0.3, 0.5, 0.7, 0.9\}$. We use two simulation designs, one with strong factors and the other one with weak factors and a strict factor model. We report the mean number of chosen factors over 5000 simulations. While both the use of optimal weighting and of the panel information criterion improve the performance of the estimator our estimators appear to work well in all cases. Furthermore, we can estimate the unknown variance parameter $\theta = \sigma^2$ accurately and with low MSE without having to estimate the unobserved factors first. A particular advantage of our approach is that it works very well irrespective of whether the factors are weak or strong. This makes it especially useful when trying to estimate the weak factors in a model and not just the few strong ones.

We also implemented simulations for the leading alternative estimator of Bai and Ng (2002) using code made available to us by the authors.³ The Bai and Ng (2002) method is sometimes criticized by practitioners as being extremely sensitive on the maximum number of factors considered by the algorithm. Sensitivity to “starting values” is an frequent problem in numerical analysis. We have conducted extensive simulations using different designs and also by varying the maximum number of factors parameter, but have only found small differences in outcomes under variations of this parameter. For the simulation design with strong factors we have found the Bai and Ng (2002) method to perform well for smaller samples but to exhibit a small negative bias in large N and large T settings. For the simulation design with weak factors we have found the Bai and Ng (2002) method to both underestimate and overestimate the number of factors depending on the different combinations of N and T used in the simulations. The results in Table 2 show that in particular for large N and large T samples, where the ratio of noise-to-signal is high, the Bai and Ng (2002) estimator substantially underestimates the number of factors relative to the minimum distance estimator proposed above. Furthermore, the minimum distance estimator exhibits lower mean squared errors than the Bai and Ng (2002) estimator in all samples.

To conclude, in situations where the structural assumptions used by the minimum distance estimator are warranted, we find the estimator to have both lower bias and lower mean squared error than the estimator of Bai and Ng (2002). The minimum distance estimator performs substantially better in situations where the factors are relatively weak relative to the noise by using structural information about the nature of the noise to reveal the factors.

2.4. Time Series Correlations

In some applications we may wish to allow for weak time series correlations of the idiosyncratic errors. If the true model is such that the idiosyncratic errors are correlated over time but we wrongly assume that they are independent over time, the number of factors that is estimated using a misspecified model will be biased. We investigate this further in Table 3 using the simulation design given in the Appendix with five strong factors and autocorrelated idiosyncratic errors. We use the moments of the eigenvalue distribution for the misspecified model to construct the moment conditions and employ the minimum distance procedure described above.

We notice that the number of estimated factors is upward-biased. This is due to the fact that if the true model has autocorrelated idiosyncratic errors the resulting eigenvalue distribution will have a larger spectral radius than the distribution for uncorrelated idiosyncratic errors. This leads to eigenvalues in the right tail of the distribution of eigenvalues due to the noise in the factor model to be falsely categorized as factors. The results in Table 3 are for a relatively weak degree of autocorrelation (0.1). We have found that, for the case

³Code is not currently available for the implementation of the Onatski (2008) estimator, which presents some computational challenges. We were not able to implement this estimator successfully in our simulations.

where the true model has autocorrelations in excess of 0.3 and where the misspecified model is estimated, our estimator will fail to converge. Table 3 also shows that if the degree of misspecification is small our estimators will have fairly small bias. This suggests that our approach is robust to minor deviations from the assumed parametric model but will fail if the degree of misspecification is large. If the true model is one where the idiosyncratic errors are autocorrelated, we can construct the correct estimator if we impose the correct assumptions on the parametric form of the autocorrelations.

Recall that by Assumption 2 the case of time-correlated idiosyncratic errors implies a model where $B_T \neq I_T$ such that the spectral density of B_T converges to a non-random distribution with bounded spectrum. Consider, for example, 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. Note that in asset return factor models such as APT the degree of autocorrelation would typically be small.

In order to guarantee that the spectrum of B_T is bounded we need to assume absolute summability, i.e.

$$(2.61) \quad \sum_{k=0}^{\infty} |\rho|^k = \frac{1}{1 - |\rho|} < \infty,$$

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

$$(2.62) \quad f(\zeta) = \lim_{T \rightarrow \infty} \sum_{k=-\infty}^{+\infty} \rho^{|k|} \exp(ik\zeta) = \frac{1 - \rho^2}{1 - 2\rho \cos(\zeta) + \rho^2}.$$

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

$$(2.63) \quad m_B^s = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{k=1}^T \lambda_k^s = \frac{1}{2\pi} \int_0^{2\pi} [f(\zeta)]^s d\zeta.$$

This expression gives the moments of the eigenvalue distribution of the population time-covariance matrix B_T for large T . The univariate integral above can easily be evaluated using numerical integration techniques. The above approach can be employed to compute the moments of the population spectrum F^B for other choices of B_T in the large T limit.

In order to apply the procedure outlined in the section above we have to compute the moments of the asymptotic eigenvalue distribution of the covariance matrix of observations for $N \rightarrow \infty$ and $T \rightarrow \infty$ and

$N/T \rightarrow c$:

$$(2.64) \quad m_{\Omega}^s = \lim_{N \rightarrow \infty} (1/N) E \{ \text{tr}(\Omega_N^s) \}.$$

If we assume that cross-sectional correlations are given by a scale factor times I_N then we can write $\Omega_N = (1/T)UU' = (1/T)\epsilon B_T \epsilon'$, where $U = \epsilon B_T^{1/2}$ and $(\epsilon)_{i,j}$ is iid mean zero with finite fourth order moments. Then,

$$(2.65) \quad m_{\Omega}^s = \lim_{N \rightarrow \infty} (1/N) E \{ \text{tr}([(1/T) \epsilon B_T \epsilon']^s) \}.$$

Notice, however, that the non-zero eigenvalues of $(1/T)\epsilon B_T \epsilon'$ and $(1/T)\epsilon' \epsilon B_T$ are the same. Hence we can apply the free probability procedure presented in Section 2.2 to compute the mixed moments of $(1/T)\epsilon' \epsilon$ and B_T . For the case where B_T is a Toeplitz matrix corresponding to an AR(1) process, the moments of the eigenvalue distribution of B_T were given above. The relationship between the moments of the eigenvalue distribution of the covariance matrix of the observations and the moments of the eigenvalue distribution of B_T can be summarized by the following result:

Proposition 8: *Let m_{Ω}^s be the limiting moments of the empirical noise covariance Ω and m_B^s be the limiting moments of the correlation matrix B . Then for $w \in \mathbb{C}^+$ with $\Im(w) > 0$ we have that*

$$(2.66) \quad \sum_{s=1}^{\infty} \frac{m_{\Omega}^s}{w^s} = \frac{1}{c} \left\{ \sum_{s=1}^{\infty} \frac{m_B^s}{w^s} \left[c \left(1 + \sum_{r=1}^{\infty} \frac{m_{\Omega}^r}{w^r} \right) \right]^s \right\}.$$

The relationship between the first four moments is given by:

$$(2.67) \quad m_{\Omega}^1 = m_B^1$$

$$(2.68) \quad m_{\Omega}^2 = c m_B^2 + (m_B^1)^2$$

$$(2.69) \quad m_{\Omega}^3 = c^2 m_B^3 + 3c m_B^2 m_B^1 + (m_B^1)^3$$

$$(2.70) \quad m_{\Omega}^4 = c^3 m_B^4 + 4c^2 (m_B^3 m_B^1 + \frac{1}{2} (m_B^2)^2) + 6c m_B^2 (m_B^1)^2 + (m_B^1)^4.$$

Proof: See Appendix.

In Table 4 we illustrate the performance of our estimator by Monte-Carlo simulations for a model with 5 factors constructed according to the design given in the Appendix and with the additional requirement that the idiosyncratic errors are autocorrelated with coefficient $\rho = 0.3$. We apply two procedures to estimate the unknown parameters (p, σ^2, ρ) , corresponding to the number of factors, the variance scale and the degree of autocorrelation respectively. The first procedure is the unweighted minimum distance method, while the second procedure augments the minimum distance objective function with the panel information criterion in order to estimate p . Both methods were described in the previous section. We notice that the estimator

of p based on the minimum distance procedure augmented with the panel information criterion performs extremely well in choosing the correct number of factors. Similarly the unknown covariance parameters σ^2 and ρ are also estimated precisely with low MSE.

Additionally, we compare the performance of the minimum distance estimator with that of the Bai and Ng (2002) estimator using the same Monte Carlo samples. We implement two version of the Bai and Ng (2002) estimator using either 5 or 15 as the maximum number of factors over which the algorithm searches. Both implementations appear to generate a downward bias in the estimated number of factors relative to the minimum distance estimator. The estimator does not appear to be sensitive to the maximum number of factors parameter. The minimum distance estimator also has substantially lower mean squared error.

The expressions given in Propositions 6 and 8 above cover the cases where either B_T or A_N are known to be the identity matrix up to a scaling factor. It is possible however to estimate models where both $B_T \neq I_T$ and $A_N \neq I_T$. In such cases the covariance matrix of the residuals is given by:

$$(2.71) \quad \Omega_N = \frac{1}{T} A_N^{1/2} \epsilon B_T \epsilon' A_N^{1/2}.$$

The corresponding moment conditions can be applied by first computing the moments of the eigenvalue distribution of $\Psi_N = \frac{1}{T} \epsilon B_T \epsilon'$ using Proposition 8 and then computing the moments of the eigenvalue distribution of the product $A_N \Psi_N$ by employing Proposition 6.

In some cases we may not be able to specify an exact parametric model for A_N or B_T . It may, however, be possible to derive estimates of A_N and B_T using a consistent method for estimating the residuals $U_{i,t}$ without estimating the number of factors or the factor loadings and factor scores. Such a procedure was recently suggested by Pesaran (2006) and involves augmenting the factor model with observed factor proxies constructed from the cross-sectional averages of the model. In this situation we may be able to estimate the number of factors by a two-step procedure which first derives consistent estimates of A_N or B_T and then uses the estimated moments to extract the correct number of factors. Note that however in general, it will not be possible to estimate both A_N and B_T from a single sample.

3. Second Order Unbiased Estimation of the Factor Model

3.1. Inconsistency of Principal Components for Weak Factors

Consider the classical factor model of equation 2.1 $R_t = \Lambda F_t + U_t$, for $t = 1 \dots T$. Recall that 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 . Estimation of the classical factor model requires estimation of the $N \times p$ matrix of factor loadings and predicting the $p \times T$ values of the latent factors F_t (factor scores). In this section we

will restrict our attention to the exact factor model and assume that $U_{i,t}$ is iid Normal with mean 0 and variance σ^2 .

Traditionally, for small values of N , factor models are estimated by maximum likelihood methods. High dimensional factor models however require the estimation of Np parameters which proves to be computationally infeasible if N is larger than 25. Thus, practitioners often employ Principal Components Analysis (PCA) applied to the covariance matrix of observations:

$$(3.1) \quad \Sigma_N = \frac{1}{T} \sum_{t=1}^T R_t R_t' = \Lambda \left(\frac{1}{T} \sum_{t=1}^T F_t F_t' \right) \Lambda' + \frac{1}{T} \sum_{t=1}^T U_t U_t' = \Xi_N + \Omega_N,$$

in order to estimate the factor loadings (Jolliffe, 2002). The equation above does not allow for the separate identification of both factor loadings and factor scores and we have to impose additional identifying restrictions. In particular it is common to assume that the factors F_t are orthogonal to each other and have unit variance. Moreover, since Ξ_N is invariant to orthogonal rotations S of the factor loadings, $\Xi_N = \Lambda \Lambda' = (\Lambda S') (S \Lambda')'$ for $S' S = I$, we require additional normalizations. Principal Components normalizes the Euclidean distance of the estimated factor loadings, $\|\tilde{\Lambda}_j\|_2 = 1$ for $j = 1 \dots p$. The statistics literature often refers to this as “fixing the rotation” of the factors, a process that consists of reporting the normalized factor loadings $\tilde{\Lambda}_j = \hat{\Lambda}_j / \|\hat{\Lambda}_j\|_2$ for $j = 1 \dots p$ and estimated factors that are orthogonal to each other.

The PCA estimator $\tilde{\Lambda}$ is the $N \times p$ matrix of the first p (normalized) eigenvectors from the spectral decomposition $\Sigma_N = V D V'$, associated with the p largest diagonal elements of D , where the columns of V are orthonormal by construction. The factor loadings are defined only up to a change in sign. Given estimates of the factor loadings, we can estimate factor scores by Generalized Least Squares regressions on the cross-section. Alternatively, the factor scores can be approximated by the first p normalized eigenvectors of the $T \times T$ matrix $(1/N) R' R$, where R is the $N \times T$ matrix with columns R_1, R_2, \dots, R_T (Connor and Korajczyk, 1986).

Define

$$(3.2) \quad \tilde{\mu} = \frac{\min(\text{Sp}(\lim_{N \rightarrow \infty} (\Lambda \Lambda')))}{\max(\text{Sp}(\lim_{N \rightarrow \infty} (\Omega_N)))}$$

which corresponds to the ratio of the minimum eigenvalue of the spectrum due to the factors over the maximum eigenvalue due to the noise term. It can be thought of as a measure of the spectral gap discussed in Section 2.1. Thus, it also measures the “strength” of the factors. A strong factor corresponds to a factor for which $\tilde{\mu} \gg 0$, while a weak factor leads to a corresponding eigenvalue for which $\tilde{\mu}$ is close to zero. By Assumptions 3 and 4 we have $\tilde{\mu} > 0$ as $N \rightarrow \infty$, if the factors are identified. In finite samples, however, we expect to have both weak and strong factors.

Under the asymptotic framework of Assumption 1, $N \rightarrow \infty$ and $T \rightarrow \infty$ and $N/T \rightarrow c \in (0, \infty)$ it has recently been noticed that the estimated sample eigenvectors are inconsistent estimators of the corresponding population eigenvectors (Hoyle and Rattray, 2004; Paul, 2007; Onatsky, 2006). Let us assume that for each factor $j = 1 \dots p$ the corresponding vector of true factor loadings Λ_j and its estimate $\hat{\Lambda}_j$ have been normalized such that $\|\Lambda_j\|_2 = \|\hat{\Lambda}_j\|_2 = 1$. Denote by $\varnothing(x, y)$ the cosine of the angle between two arbitrary vectors x and y , where $\varnothing(x, y) = x'y/(\|x\|_2 \cdot \|y\|_2)$. The proposition below states conditions under which consistency continues to hold for the classical factor model estimated by PCA even though the sample eigenvectors are inconsistent.

Proposition 9 (Consistency of PCA): *The degree of inconsistency in the estimates of the factor loadings Λ as $N \rightarrow \infty$, $T \rightarrow \infty$ and $N/T \rightarrow c \in (0, \infty)$ is given by*

$$(3.3) \quad \sqrt{\frac{1 - \frac{c}{\tilde{\mu}^2}}{1 + \frac{c}{\tilde{\mu}}}} \leq \varnothing(\hat{\Lambda}_j, \Lambda_j) \leq 1.$$

If $\tilde{\mu} \rightarrow \infty$ as $N \rightarrow \infty$ then the PCA estimate of Λ_j is consistent, i.e. $\varnothing(\hat{\Lambda}_j, \Lambda_j) \rightarrow 1$.

Proof: See Appendix.

For random factor loadings Λ we can think of \varnothing as a measure of the correlation between the two vectors. It is perhaps surprising that the PCA estimator of a factor model is consistent, but it is important to stress that it follows as a result of the more specialized assumptions imposed by an economic factor model and it does not hold true for an arbitrary application of PCA. Notice that in a factor model the inconsistency of the estimated factor loadings (and factor scores) depends on the ratio between c and $\tilde{\mu}$ only and, under the asymptotic framework of Assumption 1, c converges to a constant, while under Assumption 4 on pervasiveness of economic factors, $\tilde{\mu}$ diverges as $N \rightarrow \infty$. Since the ratio tends to zero the inconsistency disappears.

In finite samples, however, it might be the case that the measure of the spectral gap, $\tilde{\mu}$, is close to zero for the weak factors. In such cases the estimation of the factor loadings may suffer substantial biases, presenting challenges for the estimation of weak economic factors. In order to investigate this effect, we use Monte Carlo to simulate a factor model with 5 weak factors using the simulation design described in the Appendix. We calibrate the simulations such that $\tilde{\mu} < 2$. In Table 5 we present the results for the estimated coefficients on the factor loadings using PCA. We notice that PCA performs very poorly in this case. It is interesting to note that, by construction, the model has both weak and strong factors. The PCA estimates, however, are poor for all factor loadings, not just the ones on the weak factors. In the section below we discuss alternative estimation procedures employing instrumental variables. Table 5 shows that, by contrast, an instrumental variables approach continues to provide satisfactory estimates even though PCA fails. This is due to the very different approach to estimation of the two methods and will be discussed in more detail below. Note that the results of Table 5 also seem to indicate that the PC estimator is likely not to have moments for the

case of weak factors. This is similar to the problems encountered in the estimation of equations with weak instruments (Hahn, Kuersteiner and Hausman, 2004).

PCA does not only offer a poor approach to estimation in the presence of weak factors, it also suffers two more serious short-comings which are easily corrected by alternative instrumental variables based procedures. First, PCA estimation of the factor loadings does not allow us to impose economic restrictions on the estimated coefficients. Such restrictions are common in the macroeconomics literature and are particularly important in Factor Augmented VAR models (Stock and Watson, 2005). Even small departures from the standard framework, such as imposing exclusion restrictions on the factors in some equations, present major challenges. Restrictions severely limit the use of standard eigenvector techniques in the estimations.

Second, performing inference on the estimated factor loadings is very difficult due to the complicated distributions of eigenvectors (Bai, 2003; Paul, 2007; Onatski, 2007). In particular, the asymptotic distributions depend on the large eigenvalues but the eigenvalues themselves are only observed with bias in the sample (Paul, 2007; Onatski, 2008; Harding, 2008b).

3.2. Estimation by Instrumental Variables

Factor analysis can also be thought of as a generalization of multivariate linear regression analysis with measurement error (Madansky, 1964). Although this fact has been recognized for a very long time, the application of instrumental variables (IV) procedures to the estimation of factor models is generally regarded as inferior to estimation by PCA. In this section we show that high-dimensional factor models estimated by IV suffer from a finite sample bias problem similar to that encountered in the recent econometrics literature on estimation with many instruments (Hansen, Hausman and Newey, 2006). While this explains the practitioners' reluctance to apply IV methods in large datasets, it also provides a solution to the second order unbiased estimation by using IV estimators with improved finite sample performance.

Above we have seen how PCA imposes a normalization on the factor loadings $\tilde{\Lambda}_j = \hat{\Lambda}_j / \|\hat{\Lambda}_j\|_2$. Other normalizations are also possible without loss of generality. In particular partition the matrix of factor coefficients as follows:

$$(3.4) \quad \Lambda = \begin{pmatrix} \hat{\Lambda}_1 \\ \hat{\Lambda}_2 \end{pmatrix},$$

where $\hat{\Lambda}_1$ is a $p \times p$ submatrix of $\hat{\Lambda}$. We can now define the normalized factor loadings to be:

$$(3.5) \quad \tilde{\Lambda} = \begin{pmatrix} \hat{\Lambda}_1 \\ \hat{\Lambda}_2 \end{pmatrix} \hat{\Lambda}_1^{-1} = \begin{pmatrix} I_p \\ \hat{\Lambda}_2 \hat{\Lambda}_1^{-1} \end{pmatrix} = \begin{pmatrix} I_p \\ \tilde{\Lambda}_2 \end{pmatrix}.$$

Under this normalization we have, $R_{j,t} = F_{j,t} + u_{j,t}$, for $j = 1 \dots p$. This means that, without loss of generality, we can choose the first p observations to act as proxies measured with error of the underlying

latent factors. Now choose any observation $R_{p+k,t}$, for $k \geq 1$. Substituting the first p vectors of observations for the p factors in the equation for observation $p+k$ we obtain:

$$(3.6) \quad R_{p+k,t} = \sum_{j=1}^p \tilde{\Lambda}_{j,2}^k R_{j,t} + u_{p+k,t} - \sum_{j=1}^p \tilde{\Lambda}_{j,2}^k u_{j,t},$$

where $\tilde{\Lambda}_{j,2}^k$ corresponds to the row k and column j entry of the normalized loadings matrix $\tilde{\Lambda}_2$.

If we assume an exact factor structure with $u_{i,t}$ iid then we have

$$(3.7) \quad E \left[R_{p+m,t} \left(R_{p+k,t} - \sum_{j=1}^p \tilde{\Lambda}_2^{p+k} R_{j,t} \right) \right] = 0,$$

for $m \geq 1$ and $m \neq k$. Therefore we can use all observations other than the first p observations and observation $p+k$ as instruments for the p factor proxies used in the equation for R_{p+k} .

Notice that for the exact factor model with $A_N = I_N$ and $B_T = I_T$ we have $N-p-1$ moment conditions and the observations on the remaining $N-p-1$ variables ($R_{p+1}, \dots, R_{p+k-1}, R_{p+k+1}, \dots, R_N$) are valid instruments which can be used in order to estimate the coefficients $\tilde{\Lambda}_2^j$ for $j = 1..p$.

For other choices of A_N and B_T not all moment conditions will be available. The identification of equation 3.6 under such conditions will follow the usual procedures for the identification of equations with endogenous right hand side variables. This framework, however, is sufficiently flexible and allows us to complement the set of equations with additional linear restrictions or moment conditions in order to guarantee identification.

Notice that for the exact factor model, each $N-p$ equation $R_{p+k,t}$ can be written as in equation 3.6. Each equation is identified with a degree of over-identification equal to $N-2p-1$. Therefore we can estimate all the factor loadings $\tilde{\Lambda}_2$ using instrumental variables by repeatedly estimating each set of factor loadings using all other observations except for the first p observations as instruments.

For simplicity let us focus on a model with only one factor:

$$(3.8) \quad \begin{aligned} R_{1,t} &= f_t + u_{1,t} \\ R_{2,t} &= \lambda_2 f_t + u_{2,t} \\ &\dots \\ R_{N,t} &= \lambda_N f_t + u_{N,t}, \end{aligned}$$

where $\lambda_2, \lambda_3, \dots, \lambda_N$ are the factor loadings that we are interested in estimating and f is the unobserved latent factor. Let us assume that $u_{i,t}$ are distributed jointly Normal with $E(u_{i,t}) = 0$ and $E(u_{i,t}u_{j,t}) = 0$ and $E(u_{i,t}^2) = \sigma_i^2$. These assumptions allow us to investigate the case where the maximum number of possible instruments, $N-2$, is allowed. More general correlations restrict the number of instruments that can be employed but will not affect the intuition behind the next result. Identification in such settings follows the usual rules for the identification of systems of equations.

Consider the estimation of λ_k for $k = 2 \dots N$. Using $R_{1,t}$ as a proxy for the latent factor we obtain:

$$(3.9) \quad R_{k,t} = \lambda_k R_{1,t} + \epsilon_{k,t}$$

where $\epsilon_{k,t} = u_{k,t} - \lambda_2 u_{1,t}$. Furthermore, notice that for each equation $j > 1, j \neq k$ we have $f_t = (1/\lambda_j)(R_{j,t} - u_{j,t})$. Hence we can write the reduced form equation for $R_{1,t}$ as

$$(3.10) \quad R_{1,t} = z_t \pi + v_t,$$

where

$$(3.11) \quad z_t = (R_{2,t}, \dots, R_{k-1,t}, R_{k+1,t}, \dots, R_{N,t})$$

$$(3.12) \quad v_t = u_{1,t} - U_t \pi$$

$$(3.13) \quad U_t = (u_{2,t}, \dots, u_{k-1,t}, u_{k+1,t}, \dots, u_{N,t}).$$

Notice that the dimension of z_t is $N - 2$ which is potentially large relative to the sample size T . Let us first assume that the number of potential instruments is large relative to the sample size but less than the sample size, i.e. $c = N/T < 1$. Let $E(v_t^2) = \sigma_v^2$ and let $\chi^2 = \pi' z' z \pi / \sigma_v^2$ be the concentration parameter.

We wish to investigate the bias in the estimation of λ_k using the most common IV estimation procedure, Two-Stage Least Squares (2SLS):

$$(3.14) \quad \hat{\lambda}_k^{2SLS} = \frac{R_1 P_z R_k}{R_1 P_z R_1},$$

where $P_z = z(z'z)^{-1}z'$, and $R_j = (R_{j,1}, \dots, R_{j,T})'$.

The next proposition gives the second-order bias of the 2SLS estimate.

Proposition 10 (Second Order Bias of 2SLS in Factor Models): *We expect the bias in estimating the factor loadings by 2SLS to be approximately:*

$$(3.15) \quad E(\hat{\lambda}_k^{2SLS}) - \lambda_k \cong \frac{E(\epsilon v)}{\sigma_v^2} \frac{K - 2}{\chi^2} = -\lambda_k \frac{\sigma_1^2}{\sigma_v^2} (N - 4) \frac{(1 - R^2)}{(T - N - 2) R^2},$$

where R^2 is the theoretical R^2 of the first stage regression.

This expression reveals that the 2SLS estimate of the factor loadings is biased downward. Moreover, the amount of bias is proportional to the degree of over-identification and monotonically increasing in N . This explains why estimation by 2SLS suffers relative to estimation by PCA even for strong factors as the 2SLS bias increases with N .

Note that the expression given above is similar to the expressions found by Hahn and Hausman (2002 a,b) in their investigation of estimation bias in simultaneous equations. The relationship is due to the asymptotic

framework used in our analysis of high-dimensional factor models. We allow the number of cross-sectional observations to grow with the sample size at the same rate. Estimation of the factor loadings by 2SLS uses some of these observations as instruments, thereby generating a model with many instruments very similar to that of Bekker (1994). Note however, that there are also differences due to the very specific form which the correlation between the structural equation and the reduced form takes in the factor model. In particular, the bias expression depends on the variances in each equation.

In order to explore the behavior of different IV methods in the estimation of factor models we focus in Table 6 on a simple model with one factor constructed using the design outlined in the Appendix for different choices of N and T such that $c \in \{0.3, 0.5, 0.7, 0.9\}$. Our aim is to compare the behavior of IV estimators and understand the nature of the “many instruments” bias exhibited by 2SLS in this case. As such, we do not choose an example with very weak factors where PCA will fail so as not to give the impression that IV methods have a very substantial advantage by design. Rather our aim is to show that IV methods perform at least as well as a PCA approach, and in some cases, as discussed in Table 5, IV methods prove to be more robust to the presence of very weak factors.

In Table 6a we report the mean bias, the median bias and the means squared error (MSE) of the factor loadings estimates for the case with $c < 1$. In the second column we report the results for the standard 2SLS procedure. As predicted by the bias expression in equation 3.15 above we observe a negative bias which is increasing in $c = N/T$.

In order to correct for the bias due to the large number of instruments employed in the estimation of the factor loadings, we evaluate the performance of additional estimators familiar from the literature on many instruments as possible solutions to the many instruments problem. We consider the Fuller (1977) estimator, which is a modification of the Limited Information Maximum Likelihood (LIML) estimator with parameter $a = 1$. Additionally, we compute a Bias Corrected 2SLS derived by solving for λ_k in equation 3.15 of Proposition 10. The exact expression for these estimators are given in the Appendix. Furthermore, we also estimate the Continuously Updating Estimator (CUE) (Newey and Windmeijer, 2005), a particular choice of a generalized empirical likelihood estimator. In order to overcome the computational issues associated with the CUE estimator we employ a bounded Nelder-Mead algorithm.

Table 6a shows that all the estimators considered above perform well and can be used to obtain second order unbiased estimates of the factor loadings. Fuller seems to have comparable MSE to CUE, but the mean bias and MSE performance of CUE suffers for $c = 0.9$. This may indicate that CUE has a moment problem when the number of instruments is close to the sample size. Additionally, the estimators considered appear to be median unbiased. CUE appears to dominate the performance of the other estimators in terms of MSE except for the case with $c = 0.9$, while Fuller dominates in terms of mean bias.

So far we have only considered cases where $c < 1$, i.e. where the number of possible instruments is less than the sample size. In many applications, however, the number of individuals may be larger than the number of time periods over which the sample is observed. In this case the standard IV estimation methods fail since they employ the $P_z = z(z'z)^{-1}z'$ projection and $z'z$ becomes singular when $N > T$. At first glance this appears to be a limitation of IV procedures since they restrict the number of instruments to be less than the sample size.

In Table 6a and Table 6b we also investigate a proposal of Theil (1973) which advocates the use of an incomplete projection which avoids the singularity of $z'z$ by using $P_{D,z} = z(D)^{-1}z'$ for some positive definite matrix D . The resulting 2SLS estimator which we call T-2SLS is given by $\hat{\lambda}_k^{2SLS} = (R_1 P_{D,z} R_k) / (R_1 P_{D,z} R_1)$. A similar modification can be applied to the CUE estimator by choosing a moment weighting matrix which depends on D rather than on $z'z$. The exact expression is given in the Appendix and we label the resulting estimator T-CUE. We use Monte-Carlo to investigate the behavior of the estimate of the factor loadings for a simple choice of $D = I$ for both $c < 1$ and $c > 1$.

For $c < 1$ the bias of T-2SLS is only slightly higher than that of BC2SLS while delivering a lower MSE than Fuller, BC2SLS and CUE. For $c > 1$ the bias of T-2SLS becomes substantial. By contrast T-CUE performs extremely well for both $c < 1$ and $c > 1$ in terms of both bias and MSE. For $c < 1$ T-CUE seems to avoid the moment problem of CUE for $c = 0.9$ and seems to outperform Fuller and BC2SLS. We plan to explore the performance of estimators employing Theil's modification in future work to establish their distributional properties and the optimal choice of the matrix D . The simulations presented in this section, however, indicate that Fuller, BC2SLS, CUE and T-CUE can be used as alternatives for the accurate estimation of the factor loadings. Moreover, T-CUE can be used even when $N > T$, and the number of instruments exceeds the sample size. Given the difficulties inherent in estimating weak factors using PCA or imposing structural restrictions on the factor loadings, the methods explored in this section provide much needed accuracy and flexibility for the estimation of factor models.

4. Applications

In order to explore the useability of the proposed estimation procedure for large factor models to real world data, we estimate two stylized models. First, we look at the effect of adding more data on the estimation of a classical APT model of stock returns. Second, we re-consider the classical problem of choosing the number of factors in a large dataset of macroeconomic indicators.

4.1. The Number of Factors in Large and Noisy Asset Covariance Matrices

The use of factor models in economics is historically linked to the development of APT models of systemic risk in finance (Ross, 1976; Chamberlain and Rothschild, 1983). Over the years, numerous empirical studies

have been conducted estimating the “true” number of factors from financial data. There are good reasons to believe that this question is not as well posed as it may seem at first glance. Indeed, no agreement exists in the literature as to the number of risk factors. It depends on the market and type of assets under consideration, the sampling frequency and time-frame of the data. *Ceteris paribus*, Random Matrix Theory makes precise predictions of the effect of adding more time periods of data to a sample used in the estimation of an APT model. Our ability to extract factors from financial data depends on the ratio of signal to noise in the data since noisy data leads to imprecise estimates of the empirical covariance matrix. In high-dimensional settings it can be shown that the key to identification is the ratio of individual units (stocks) to the number of time periods, $c = N/T$. Exact analytic results are given in Harding (2008b) and are not repeated here. The underlying intuition is however very simple. While the eigenvalues corresponding to the strong factors diverge easily in large samples, the eigenvalues of weak factors remain bounded and their behavior may be indistinguishable from that of eigenvalues due to noise. The extent to which the noise may obfuscate the presence of latent factors depends on a large extent on the ratio of individual units to time periods, c . Since we are averaging over time-periods, in the absence of nonstationarity, having more time-periods leads to a more precise estimation of the empirical covariance matrix and a more condensed distribution of eigenvalues resulting from the noise part of the model. This makes it easier for weaker factors to manifest themselves in the data. Therefore, we would expect that in the absence of structural breaks, estimating the APT model would reveal more latent risk factors when using longer time series.

Notice that this prediction is contrary to and distinct from the often cited *Epps* effect, where the correlation among price changes decreases substantially as we increase the sampling frequency (Epps, 1979; Huang and Jo, 1995; Toth and Kertesz, 2008). We are interested in understanding the behavior of APT models as we add more time periods to the model, rather than subdivide the existing time-frame and sample at higher frequencies. The Epps effect shows that at very high frequencies, market microstructure noise may dominate and limit our ability to make suitable inferences.

In order to investigate our prediction that the estimated number of latent factors in a standard APT model increases as we add more time periods of observations, i.e. is a decreasing function of $c = N/T$, we construct a sample of daily stock returns over a 10 year period, starting in January 1996. Data was downloaded from the CRSP database and we initially consider all companies that are traded on three main US stock exchanges AMEX, NYSE and NASDAQ. Given the nature of this stylized exercise, we do not construct any particular type of portfolio and all firms on which data is available are admissible. Given the infrequent trading in some stocks, we only include those companies for which data is consistently recorded. In order to avoid the obfuscation of our results by missing data issues, we exclude all companies which are not traded for more than 1 day in any given month. Using this admittedly arbitrary criterion we only keep 2,336 stocks from a total of 17,517 stocks on which data is available over the period of interest. 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. Thus, the first sample contains data on the first year of the data, the second sample covers the first two years of the data and so on. Notice that a year is measured as the number of trading days which equals 252 calendar days in most years. The last sample which includes the data for the entire 10 year period covers 2,539 trading days.

In Figure 2 we plot the estimated number of factors from each of the 10 samples as a function of the number of trading days used in the estimation while keeping the number of stocks equal against the ratio $c = N/T$. As we use more trading days in the data, the estimated number of factors increases from 5 in the one year sample to 40 in the 10 year sample. It is important to remember that in this simple example it is not possible to evaluate the extent to which any one of the 40 factors estimated over a period of 10 years did in fact persist over the entire period. The assumption of factor pervasiveness only states that in order for a factor to be identified it has to be strong enough in the sense of affecting the majority of stocks over a significant period of time. It may well be that some of these factors were not active over the entire sampling period.

4.2. Data Reduction of Macroeconomic Measurements

Factor models are 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 abstraction of an 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). It may also be possible to combine a factor model with a structural economic model, whereby the data reduction is performed conditional on the structure of the model, thus in effect imposing additional identification restrictions which fix the rotation of the factors along preset dimensions thereby improving our ability to interpret the estimated factors from an economic perspective. This idea can be implemented in a DSGE framework to account for the imperfect measurement of key macroeconomic quantities (Boivin and Giannoni, 2006) or in a New-Keynesian model where financial data can be used to determine the stochastic dimension of the macroeconomy and accurately measure supply shocks (Harding, 2008a).

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 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. We have found this number to be robust to a number of alternative specifications, such as varying the penalty function, normalizing the covariance matrix to unit variances or letting the constant c change over the estimation as eigenvalues are removed. Notice that 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. Onatski (2008) concludes that the number of factors is greater than 2, but the proposed estimator is not sensitive enough to determine exactly how many factors there are in this dataset.

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. (2008). The data is similar to that of Stock and Watson (2005) and consists of 100 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 6. We also find this number to be robust to a number of variations of our estimator.

It is important to interpret the number of estimated factors with caution. As we have seen in our first example the estimates may be sensitive to the dimensions of the available data and potentially valuable information may be obscured by the noise inherent in such measurements. Furthermore, in the absence of an associated economic model it is not clear how to give these factors an economic interpretation. 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.

5. Conclusion

In this paper we introduce new econometric theory for the estimation of large panel data models with unobserved latent variables. We show that it is possible to estimate the number of factors consistently for both the exact and approximate factor model without having to estimate first the factor loadings or factor scores. Our procedure allows for arbitrary models of heteroskedasticity and autocorrelation.

This paper also contributes to the theoretical understanding of the behavior of large random matrices by providing a series of mathematical tools based on free probability theory which are employed to characterize the spectra of random matrices.

We show that our procedure performs better in a number of Monte Carlo methods than the leading method of Bai and Ng (2002). The improvement in precision comes from the use of parametric restrictions which are motivated by economic theory. The use of economic restrictions also improves our ability to associate economic meaning to estimated latent factors.

Additionally, we have shown that in factor models with weak factors, the estimation of factor loadings by PCA is inconsistent. To solve this problem we develop alternative IV based procedures with excellent finite sample properties. We relate the IV estimation of the factor model to current research on many and weak instruments.

We also consider a series of realistic examples in finance and economic which employ factor models and show that our methods perform very well on real data and provide useful insights.

6. Appendix

Simulation design

We simulate a models $R_t = \Lambda F_t + U_t$, where $U = A_N^{1/2} \epsilon B_T^{1/2}$ for $\epsilon_{j,t}$ iid $N(0, 1)$ and $F_{j,t}$ iid $N(0, 1)$. A_N and B_T are as discussed in the text. Factor loadings are generated as follows: let $\Lambda_{j,k} = \sqrt{m_1/\sqrt{m_2}}$ for $j = 1 \dots p_0$ and $k = 1 \dots p_0$, where p_0 is the number of factors. For $j = p_0 + 1 \dots N$ and $k = 1 \dots p_0$ we have $\Lambda_{j,k} = a\sqrt{m_1/(N-k)}$ where $a = -1$ if $j = rk$ and $a = +1$ if $j \neq rk$ for $r = 1, 2, 3, \dots$. We can generate weak factors by setting $m_1 = 3$ and $m_2 = N$ and strong factors by setting $m_1 = 10$ and $m_2 = 1$.

Proof of Equation 2.37: 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:

$$(6.1) \quad \phi((a - \phi(a)1)(b - \phi(b)1)(a - \phi(a)1)(b - \phi(b)1)) = 0.$$

We can expand this expression to obtain:

$$(6.2) \quad \begin{aligned} & \phi(abab) - \phi(a)\phi(bab) - \phi(b)\phi(aab) + \phi(a)\phi(b)\phi(ab) - \phi(a)\phi(abb) + \\ & \quad \phi^2(a)\phi(bb) + \phi(a)\phi(b)\phi(ab) - \phi^2(a)\phi^2(b) - \phi(b)\phi(aba) + \\ & \quad \phi(a)\phi(b)\phi(ba) + \phi^2(b)\phi(aa) - \phi^2(a)\phi^2(b) + \phi(a)\phi(b)\phi(ab) - \\ & \quad \phi^2(a)\phi(b)\phi(b) - \phi(a)\phi^2(b)\phi(a) + \phi^2(a)\phi^2(b) = 0 \end{aligned}$$

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

$$(6.3) \quad \begin{aligned} & \phi(abab) - \phi^2(a)\phi(bb) - \phi^2(b)\phi(aa) + \phi^2(a)\phi^2(b) - \phi^2(a)\phi(bb) + \\ & \quad \phi^2(a)\phi(bb) + \phi^2(a)\phi^2(b) - \phi^2(a)\phi^2(b) - \\ & \quad \phi^2(b)\phi(aa) + \phi^2(a)\phi^2(b) + \phi^2(b)\phi(aa) - \phi^2(a)\phi^2(b) + \\ & \quad \phi^2(a)\phi^2(b) - \phi^2(a)\phi^2(b) - \phi^2(a)\phi^2(b) + \phi^2(a)\phi^2(b) = 0 \end{aligned}$$

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

$$(6.4) \quad \phi(abab) = \phi^2(a)\phi(bb) + \phi^2(b)\phi(aa) - \phi^2(a)\phi^2(b).$$

Proof of Proposition 6: 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 $N \rightarrow \infty$ 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 $M_\Omega(\varrho_\Omega(w)) = w$. If we now apply equation 2.6 to the heteroskedasticity matrix A 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 can 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 equation 2.54: $\sum_{s=1}^{\infty} \frac{m_\Omega^s}{w^s} = \sum_{s=1}^{\infty} \frac{m_A^s}{w^s} (1+c \sum_{r=1}^s \frac{m_\Omega^r}{w^r})$.

Proof of Proposition 7: We need to show that $\lim_{N,T \rightarrow \infty} \Pr(\hat{J}(\hat{\theta}, \text{Sp}_p(\Sigma_N)) > \hat{J}(\hat{\theta}, \text{Sp}_{p_0}(\Sigma_N)))$ for all $p \neq p_0$ and $p \leq p_{\max}$. Let $\tilde{\Pi} = [Nm_\Omega^1, Nm_\Omega^2, \dots, Nm_\Omega^s]$ and

$$\tilde{\Pi}_p = [\sum_{\lambda_j \in \text{Sp}_p(\Sigma_N)} \lambda_j^1, \sum_{\lambda_j \in \text{Sp}_p(\Sigma_N)} \lambda_j^2, \dots, \sum_{\lambda_j \in \text{Sp}_p(\Sigma_N)} \lambda_j^s].$$

For simplicity we consider the case of the minimum distance estimator with equal weighting. First consider the case where $p < p_0$. Then,

$$\begin{aligned} \hat{J}(\hat{\theta}, \text{Sp}_p(\Sigma_N)) &= \sum_{r=1}^s \left(Nm_\Omega^r - \sum_{\lambda_j \in \text{Sp}_p(\Sigma_N)} \lambda_j^r \right)^2 \\ (6.5) \quad &= \sum_{r=1}^s \left(Nm_\Omega^r - \sum_{\lambda_j \in \text{Sp}_{p_0}(\Sigma_N)} \lambda_j^r \right)^2 + \mathcal{J}(\lambda_{p_0+1}, \dots, \lambda_p), \end{aligned}$$

$$(6.6) \quad \hat{J}(\hat{\theta}, \text{Sp}_p(\Sigma_N)) = \hat{J}(\hat{\theta}, \text{Sp}_{p_0}(\Sigma_N)) + \mathcal{J}(\lambda_{p_0+1}, \dots, \lambda_p).$$

The term $\mathcal{J}(\lambda_{p_0+1}, \dots, \lambda_p)$ consists of polynomial terms which depend on eigenvalues resulting from the latent factors. Hence by the pervasiveness of the factors (Assumption 4) $\mathcal{J}(\lambda_{p_0+1}, \dots, \lambda_p) \rightarrow \infty$ as $N \rightarrow \infty$ and $T \rightarrow \infty$. A fortiori, $\hat{J}(\hat{\theta}, \text{Sp}_p(\Sigma_N)) > \hat{J}(\hat{\theta}, \text{Sp}_{p_0}(\Sigma_N))$ with probability 1.

Now consider the case with $p > p_0$. Repeating the steps above we have $\hat{J}(\hat{\theta}, \text{Sp}_p(\Sigma_N)) > \hat{J}(\hat{\theta}, \text{Sp}_{p_0}(\Sigma_N))$ with probability 1. In this case however, it is no longer true that $\mathcal{J}(\lambda_{p_0+1}, \dots, \lambda_p) \rightarrow \infty$. Rather, we have

$\mathcal{J}(\lambda_{p_0+1}, \dots, \lambda_p) > 0$. This is a consequence of the fact that for $p \neq p_0$ the moment conditions on the eigenvalue distribution are going to be misspecified. The objective function is however asymmetric since it diverges for $p < p_0$ but is increasing for $p > p_0$. Adding an information criterion does not change affect the case where $p < p_0$. For the case where $p > p_0$ however adding an information criterion penalized the objective function proportionally to $(p - p_0)$ which is advantageous for values of p close to p_0 such that $\mathcal{J}(\lambda_{p_0+1}, \dots, \lambda_p)$ is small.

It is important to understand the intuition behind this result. While the case with $p < p_0$ is immediate due to the fact that the eigenvalues due to the factors diverge, the case with $p > p_0$ is perhaps less so. If we mistakenly throw away $p - p_0$ eigenvalues, why would this mistake not be asymptotically negligible since we are throwing away only a finite number of them. It is important to realize that the we are not removing these $p - p_0$ eigenvalues at random but rather we are removing the $p - p_0$ *largest eigenvalues*. Since the eigenvalue distribution due to the noise has a finite support this is equivalent to a placing a restriction on the domain of the pdf. This restriction becomes easily apparent in simulations where we show that the finite moments are very sensitive to this form of restriction. It may of course happen that in very large samples the distinction between the minimized objective function at p_0 and the minimized objective function at p where, $p - p_0$ is small and $p > p_0$ is very small, thereby having an objective function which is essentially flat to the right of p_0 over the interval $[p_0, p]$. This explains the small upward bias of the estimator in some simulation designs. In such cases we recommend choosing the smallest number of factors at which the objective function is minimized. In all simulations however, we have found this not to be an important issue. The objective function was sensitive enough to detect the correct number of factors in all attempted Monte Carlo designs. But we do recommend checking the plot of the minimized objective function in order to eliminate such a possibility in some applications to real world data.

Proof of Proposition 8: The proof of equation 2.66 is very similar to that of given in Proposition 6 with (Ψ, B) substituted for (A, Ψ) and using the fact that $M_{\Psi B}(w) = cM_{\frac{1}{T}\epsilon B\epsilon'}(w)$. In order to derive equations 2.67, 2.68, 2.69 and 2.70 we can expand equation 2.66 as follows:

$$(6.7) \quad c(\mathcal{M}_\Omega - 1) = c\frac{m_B^1}{w}\mathcal{M} + c^2\frac{m_B^2}{w^2}\mathcal{M}^2 + c^3\frac{m_B^3}{w^3}\mathcal{M}^3 + c^4\frac{m_B^4}{w^4}\mathcal{M}^4 + O(w^{-5}),$$

where

$$(6.8) \quad \mathcal{M} = 1 + \frac{m_\Omega^1}{w} + \frac{m_\Omega^2}{w^2} + \frac{m_\Omega^3}{w^3} + \frac{m_\Omega^4}{w^4} + O(w^{-5}).$$

Expanding the RHS of equation 5.7 in terms of $1/w$ we obtain:

$$(6.9) \quad c(\mathcal{M} - 1) = \frac{1}{w}cm_B^1 + \frac{1}{w^2}(cm_B^1m_\Omega^1 + c^2m_B^2) + \frac{1}{w^3}(cm_B^1m_\Omega^2 + c^3m_B^3 + 2c^2m_B^2m_\Omega^1) + \\ + \frac{1}{w^4}\left(3c^3m_B^3m_\Omega^1 + cm_B^1m_\Omega^3 + c^2m_B^2(2m_\Omega^2 + (m_\Omega^1)^2) + c^3m_B^4\right) + O(w^{-5}).$$

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

$$(6.10) \quad m_\Omega^1 = m_B^1$$

$$(6.11) \quad m_\Omega^2 = cm_B^2 + m_B^1 m_\Omega^1 = cm_B^2 + (m_B^1)^2$$

$$(6.12) \quad m_\Omega^3 = c^2 m_B^3 + 2cm_B^2 m_\Omega^1 + m_\Omega^2 m_B^1 = c^2 m_B^3 + 2cm_B^2 m_B^1 + (cm_B^2 + (m_B^1)^2) m_B^1$$

which can be further simplified as

$$(6.13) \quad m_\Omega^3 = c^2 m_B^3 + 2cm_B^2 m_\Omega^1 + m_\Omega^2 m_B^1 = c^2 m_B^3 + 2cm_B^2 m_B^1 + (cm_B^2 + (m_B^1)^2) m_B^1$$

$$(6.14) \quad m_\Omega^3 = c^2 m_B^3 + 3cm_B^2 m_B^1 + (m_B^1)^3$$

Additionally, we have:

$$(6.15) \quad m_\Omega^4 = m_B^1 m_\Omega^3 + cm_B^2 (2m_\Omega^2 + (m_\Omega^1)^2) + 3c^2 m_B^3 m_\Omega^1 + c^3 m_B^4$$

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

$$(6.16) \quad m_\Omega^4 = c^3 m_B^4 + 4c^2 (m_B^3 m_B^1 + \frac{1}{2} (m_B^2)^2) + 6cm_B^2 (m_B^1)^2 + (m_B^1)^4.$$

Proof of Proposition 9: A detailed discussion of the inconsistency of PCA in large N , T models can be found in Paul (2007) and Onatski (2006, 2007). Here we show that the maximum degree of inconsistency depends on our measure $\tilde{\mu}$ of the spectral gap. Let $d_j = \lim_{N \rightarrow \infty} \hat{\Lambda}' \hat{\Lambda}$. Since the non-zero eigenvalues of $\hat{\Lambda}' \hat{\Lambda}$ are the same as the eigenvalues of $\hat{\Lambda} \hat{\Lambda}'$ we have that $\text{Sp}(\lim_{N \rightarrow \infty} (\hat{\Lambda} \hat{\Lambda}')) = \{d_1, d_2, \dots, d_{p_0}\}$, where p_0 is the number of factors in our factor model. Let $\tilde{d} = \min \text{Sp}(\lim_{N \rightarrow \infty} \hat{\Lambda} \hat{\Lambda}')$. If we restrict our attention to the exact factor model we have $\sigma^2 = \max(\text{Sp}(\lim_{N \rightarrow \infty} (\Omega_N)))$. By Theorem 1 of Onatski (2006) we have that the degree of inconsistency for the j -th Principal Component is given by

$$(6.17) \quad \varnothing(\hat{\Lambda}_j, \Lambda_j) = \sqrt{\frac{d_j^2 - \sigma^4 c}{d_j(d_j + \sigma^2 c)}}$$

Note however that $\varnothing(\hat{\Lambda}_j, \Lambda_j)$ is monotonically increasing in the eigenvalues d_j due to the factors:

$$(6.18) \quad 0 < \frac{\partial \varnothing(\hat{\Lambda}_j, \Lambda_j)}{\partial d_j} = \frac{c\sigma^2(d_j^2 + 2\sigma^2 d_j + \sigma^4 c)}{2\sqrt{\frac{d_j^2 - \sigma^4 c}{d_j(d_j + \sigma^2 c)}}(d_j + c\sigma^2)^2 d_j^2},$$

since $d_j > 0$. Hence,

$$(6.19) \quad \sqrt{\frac{\tilde{d}^2 - \sigma^4 c}{\tilde{d}^2 + \sigma^2 \tilde{d} c}} < \varnothing(\hat{\Lambda}_j, \Lambda_j).$$

We can re-write the expression on the left as:

$$(6.20) \quad \sqrt{\frac{1 - \left(\frac{\sigma^2}{\tilde{d}}\right)^2 c}{1 + \frac{\sigma^2}{\tilde{d}} c}} < \varnothing(\hat{\Lambda}_j, \Lambda_j).$$

If we define $\tilde{\mu} = \tilde{d}/\sigma^2$ we obtain the expression for the inconsistency of PCA as given in equation 3.3.

Proof of Proposition 10: Consider the estimation of the factor loadings in a strict factor model with 1 latent factor given by 3.8. We can sequentially estimate the factor loadings by estimating λ_k for $k = 2..N$ using $R_{k,t} = \lambda_k R_{1,t} + \epsilon_{k,t}$ for $k = 2..N$ and $\epsilon_{k,t} = u_{k,t} - \lambda_k u_{1,t}$. Focusing on the estimation of equation k we can re-write the estimating equation as a system of equations with two endogenous variables y_1 and y_2 in a form familiar to the IV literature:

$$(6.21) \quad y_1 = \beta y_2 + \epsilon = \beta z \pi + w$$

$$(6.22) \quad y_2 = z \pi + v$$

where $y_1 = R_k$, $y_2 = R_1$, $z = (R_2, R_3, \dots, R_{k-1}, R_{k+1}, \dots, R_N)$ and $v = u_{1,t} - U_t \pi$. Let $K = \dim(\pi)$. For the strict factor model $K = N - \#\text{factors} - 1$. In our case $K = N - 2$.

Furthermore, assume that the reduced form errors (w, v) are i.i.d Normal distributed as

$$(6.23) \quad \begin{pmatrix} w \\ v \end{pmatrix} \sim N(0, \Omega) = N\left(0, \begin{pmatrix} \sigma_w^2 & \sigma_{wv} \\ \sigma_{wv} & \sigma_v^2 \end{pmatrix}\right),$$

and let the covariance between the structural equation (1) and the reduced form equation (2) for the endogenous variable y_2 , be $\sigma_{\epsilon v}$. Note that for the exact factor model with 1 factor $\sigma_{\epsilon v} = -\lambda_k \sigma_1^2$ where $\sigma_1^2 = \text{Var}(u_1)$. Let the concentration parameter be given by $\chi^2 = \pi' z' z \pi / \sigma_v^2$. We are interested in the bias of the 2SLS estimator as a function of the concentration parameter. The finite sample bias of the 2SLS estimator $\hat{\beta}_{2\text{SLS}}$ was derived by Richardson (1968) and is given by the following expression:

$$(6.24) \quad E(\hat{\beta}_{2\text{SLS}}) - \beta = \frac{\sigma_{\epsilon v}}{\sigma_{vv}} \exp(-\chi^2/2) {}_1F_1(K/2 - 1; K/2; \chi^2/2),$$

where ${}_1F_1(a; b; c)$ denotes the confluent hypergeometric function ${}_1F_1(a; b; c)$ given by the following expansion

$$(6.25) \quad {}_1F_1(a; b; c) = \sum_{j=1}^{\infty} \frac{(a)_j c^j}{(b)_j j!},$$

Note that the confluent hypergeometric function is defined in terms of Pochhammer's symbol $(a)_j$ which corresponds to the ascending factorial:

$$(6.26) \quad (a)_j = \prod_{k=0}^{j-1} (a+k) = a(a+1)(a+2)\dots(a+j-1) \quad \text{for } (a)_0 = 1.$$

Let $A = \exp(-\chi^2/2) {}_1F_1(K/2 - 1; K/2; \chi^2/2)$ and consider an expansion of A for large values of χ^2 :

$$(6.27) \quad A = \exp(-\chi^2/2) \frac{\Gamma(K/2)}{\Gamma(K/2 - 1)} \exp(\chi^2/2) \left(\frac{\chi^2}{2}\right)^{(K/2-1-K/2)} + O\left[\left(\frac{\chi^2}{2}\right)^{-2}\right]$$

$$(6.28) \quad A = 2 \frac{\Gamma(K/2)}{\Gamma(K/2 - 1)} \left(\frac{1}{\chi^2}\right) + O\left[\left(\frac{\chi^2}{2}\right)^{-2}\right]$$

But since $\Gamma(K/2) = (K/2 - 1)\Gamma(K/2 - 1)$ we have

$$(6.29) \quad A = \frac{K-2}{\chi^2} + O\left[\left(\frac{\chi^2}{2}\right)^{-2}\right]$$

If we now substitute the first term of A in our bias expression we obtain

$$(6.30) \quad E(\hat{\beta}_{2SLS}) - \beta \cong \frac{\sigma_{ev}}{\sigma_v^2} \frac{K-2}{\chi^2} = \frac{\sigma_{ev}}{\sigma_v^2} (K-2) \frac{(1-R^2)}{(N-K)R^2},$$

where R^2 corresponds to the R^2 of the first stage regression. This expression corresponds to the approximate bias expression given in Hahn and Hausman (2002 a, b). Recall that $K = N - 2$ and $\sigma_{ev} = -\lambda_k \sigma_1^2$. Hence we obtain the expression in Proposition 10:

$$(6.31) \quad E(\hat{\lambda}_k^{2SLS}) - \lambda_k \cong -\lambda_k \frac{\sigma_1^2}{\sigma_v^2} (N-4) \frac{(1-R^2)}{(T-N-2)R^2}.$$

IV Estimators used in Section 3.2:

The estimation of the loadings in a factor model using IV methods requires applying IV estimators recursively to determine the set of loadings λ_k for each cross-sectional unit. The complete set of loadings is obtained by adding the normalizations on the first p loadings. For simplicity we give the IV estimators under the setup of equations 5.21 and 5.22. At each step however a different set of observations are substituted for y_1 and z . Let $P_z = z(z'z)^{-1}z'$ and $Q_z = I - P_z$.

$$2SLS: \hat{\beta} = (y_2' P_z y_1) / (y_2' P_z y_2)$$

Fuller: $\hat{\beta} = (y_2' P_z y_1 - \kappa y_2' Q_z y_1) / (y_2' P_z y_2 - \kappa y_2' Q_z y_1)$ for $\kappa = \phi - 1 / (T - N - 2)$, where $\phi = \min \text{Sp}\{W' P_z W (W' Q_z W)^{-1}\}$ and $W = (y_1, y_2)$.

BC2SLS: Uses the idea in Hahn and Hausman (2002b) to solve for the population coefficient from the second-order bias expression.

$$(6.32) \quad \hat{\beta}_{\text{BC}} = \hat{\beta}_{\text{2SLS}} / \left(1 - \frac{\sigma_1^2}{\sigma_v^2} (N - 4) \frac{(1 - R^2)}{(T - N - 2) R^2} \right)$$

where $\hat{\beta}_{\text{2SLS}}$ is the 2SLS estimator. We can make this estimator feasible by substituting R^2 with the estimated first stage R^2 , σ_v^2 with the estimated first stage variance and estimate $\sigma_1^2 = \text{Var}(y_1) - 1$, if we normalize the variance of the factors to 1.

CUE: $\hat{\beta} = \text{argmin}_{\beta \in B} \hat{g}(\beta)' \hat{\Omega}(\beta) \hat{g}(\beta) / 2$ for $\hat{g}(\beta) = z'(y_1 - \beta y_2)$ and $\hat{\Omega}(\beta) = E(\hat{\sigma}_t^2 z_t' z_t)$, where $\hat{\sigma}_t^2 = E(\epsilon_t^2)$. In order to optimize this objective function we applied a bounded version of the Nelder-Mead algorithm which allows us to impose restrictions on the parameter space B in order to avoid the multiple minima occasionally found in the simulations.

T-2SLS: The idea behind Theil's modification is to replace the term $(z'z)^{-1}$ by a matrix D . Here we take $D = I$ and therefore we have $\hat{\beta} = (y_2' z z' y_1) / (y_2' z z' y_2)$. Note that this estimator is consistent since $\hat{\beta} = \beta_0 + (y_2' z z' \epsilon) / (y_2' z z' y_2)$ and $\text{plim } T^{-1} z' \epsilon = 0$.

T-CUE: The estimator is the same as CUE but uses a different weighting matrix $\hat{\Omega}(\beta) = E(\hat{\sigma}_t^2)$.

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.
- ANDERSON, T. W., AND H. RUBIN (1956): “Statistical Inference in Factor Analysis,” Cowles Foundation Working Paper.
- ANDREWS, D. W. K. (1999): “Consistent Moment Selection Procedures for Generalized Method of Moments Estimation,” *Econometrica*, 67(3), 543–564.
- ANG, A., AND M. PIAZZESI (2003): “A No-Arbitrage Vector Autoregression of Term Structure Dynamics with Macroeconomic and Latent Variables,” *Journal of Monetary Economics*, 50, 745–787.
- ARNOLD, S. F. (1981): *The Theory of Linear Models and Multivariate Analysis*. Wiley, New York.
- BAI, J. (2003): “Inferential Theory for Factor Models of Large Dimensions,” *Econometrica*, 71, 135–171.
- (2005): “Panel Data Models with Interactive Fixed Effects,” mimeo.
- BAI, J., AND S. NG (2002): “Determining the Number of Factors in Approximate Factor Models,” *Econometrica*, 70(1), 191–221.
- (2007): “Determining the Number of Primitive Shocks in Factor Models,” *Journal of Business and Economic Statistics*, 25(1).
- 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.
- (2004): “CLT for Linear Spectral Statistics of Large-Dimensional Sample Covariance Matrices,” *Annals of Probability*, 32(1A), 553–605.
- (2005): “Exact Separation of Eigenvalues of Large Dimensional Covariance Matrices,” mimeo.
- 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.
- BURMEISTER, E., AND M. B. MCELROY (1988): “Joint Estimation of Factor Sensitivities and Risk Premia for the Arbitrage Pricing Theory,” *Journal of Finance*, 43(3), 721–733.
- CHAMBERLAIN, G., AND M. ROTHSCILD (1983): “Arbitrage, Factor Structure and Mean-Variance Analysis in Large Asset Markets,” *Econometrica*, 51(5), 1305–324.
- CONNOR, G., AND R. A. KORAJCZYK (1986): “Performance Measurement with the Arbitrage Pricing Theory: A New Framework for Analysis,” *Journal of Financial Economics*, 15, 373–394.
- 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.
- EPPS, T. W. (1979): “Comovements in Stock Prices in the Very Short Run,” *Journal of the American Statistical Association*, 74(366).
- FULLER, W. A. (1977): “Some Properties of a Modification of the Limited Information Estimator,” *Econometrica*, 45(4), 939–953.
- GOLDBERGER, A. S. (1972): “Maximum-Likelihood Estimation of Regressions Containing Unobservable Independent Variables,” *International Economic Review*, 13(1), 1–15.
- GORMAN, W. M. (1980): “A Possible Procedure for Analysing Quality Differentials in the Egg Market,” *Review of Economic Studies*, 47, 843–856.
- GRENNANDER, U., AND G. SZEGO (1958): *Toeplitz Forms and their Applications*. University of California Press, Berkeley.
- HAHN, J., AND J. HAUSMAN (2002a): “A New Specification Test for the Validity of Instrumental Variables,” *Econometrica*, 70(1), 163–189.
- (2002b): “Notes on Bias in Estimators for Simultaneous Equation Models,” *Economics Letters*, 75, 237–241.

- 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.
- HAHN, J., G. KUERSTEINER, AND J. HAUSMAN (2004): “Estimation with Weak Instruments: Accuracy of Higher Order Bias and MSE Approximations,” *Econometrics Journal*, 7(1), 272–306.
- HANSEN, C., J. HAUSMAN, AND W. NEWEY (2006): “Estimation with Many Instrumental Variables,” mimeo.
- HANSEN, L. P. (1982): “Large Sample Properties of Generalized Method of Moments Estimators,” *Econometrica*, 50, 1029–1054.
- HARDING, M. (2008a): “Do Global Factors Affect the US Economy?,” mimeo.
- (2008b): “Explaining the Single Factor Bias of Arbitrage Pricing Models in Finite Samples,” *Economics Letters*, 99(1).
- HECKMAN, J. J., AND S. NAVARRO (2007): “Dynamic Discrete Choice and Dynamic Treatment Effects,” *Journal of Econometrics*, 136(2), 341–396.
- HOYLE, D., AND M. RATTRAY (2004): “Principal Component Analysis Eigenvalue Spectra From Data with Symmetry Breaking Structure,” *Physical Review E*, 69.
- HUANG, R. D., AND H. JO (1995): “Data Frequency and the Number of Factors in Stock Returns,” *Journal of Banking and Finance*, 19, 987–1003.
- JOLLIFFE, I. T. (2002): *Principal Component Analysis*. Springer, New York.
- JONSSON, D. (1982): “Some Limit Theorems for the Eigenvalues of a Sample Covariance Matrix,” *Journal of Multivariate Analysis*, 12, 1–38.
- MADANSKY, A. (1964): “Instrumental Variables in Factor Analysis,” *Psychometrika*, 29(2), 105–113.
- MILLER, G. (2006): “Factor Models and Fundamentalism,” Morgan Stanley Capital International - BARRA, Publication 181.
- NEWEY, W., AND F. WINDMEIJER (2005): “GMM with Many Weak Moment Conditions,” mimeo.
- ONATSKI, A. (2006): “Determining the Number of Factors from Empirical Distribution of Eigenvalues,” mimeo.
- (2007): “Asymptotics of the Principal Components Estimator of Large Factor Models with Weak Factors and i.i.d. Gaussian Noise,” mimeo.
- (2008): “Testing Hypotheses about the Number of Factors in Large Factor Models,” mimeo.
- PAUL, D. (2007): “Asymptotics of the Leading Sample Eigenvalues of a Spiked Covariance Model,” *Statistica Sinica*.
- PESARAN, M. H. (2006): “Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure,” *Econometrica*, 74(4), 967–1013.
- RICHARDSON, D. H. (1968): “The Exact Distribution of a Structural Coefficient Estimator,” *American Statistical Association Journal*, December, 1214–1226.
- ROBINSON, P. M. (1974): “Identification, Estimation and Large-Sample Theory for Regressions Containing Unobservable Variables,” *International Economic Review*, 15(3), 680–692.
- ROSS, S. (1976): “The Arbitrage Theory of Capital Asset Pricing,” *Journal of Finance*, 13, 341–360.
- 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 (54): “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.
- STOCK, J., AND M. WATSON (2003): “Forecasting Output and Inflation: The Role of Asset Prices,” *Journal of Economic Literature*, 41, 788–829.
- (2005): “Implications of Dynamic Factor Models for VAR Analysis,” mimeo.
- (2006): “Forecasting with Many Predictors,” in *Handbook of Economic Forecasting*, ed. by G. Elliott, C. J. Granger, and A. Timmermann. Elsevier, Amsterdam.
- THEIL, H. (1973): “A Simple Modification of the Two-Stage Least Squares Procedure for Undersized Samples,” in *Structural Equation Models in the Social Sciences*, ed. by A. Goldberger, and O. Duncan. Seminar Press, New York.
- TOTH, B., AND J. KERTESZ (2008): “The Epps Effect Revisited,” mimeo.
- 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.
- WACHTER, K. W. (1978): “The Strong Central Limits of Random Matrix Spectra for Sample Matrices of Independent Elements,” *Annals of Probability*, 6(1), 1–18.
- ZELLNER, A. (1970): “Estimation of Regression Relationships Containing Unobservable Independent Variables,” *International Economic Review*, 11(3), 441–454.

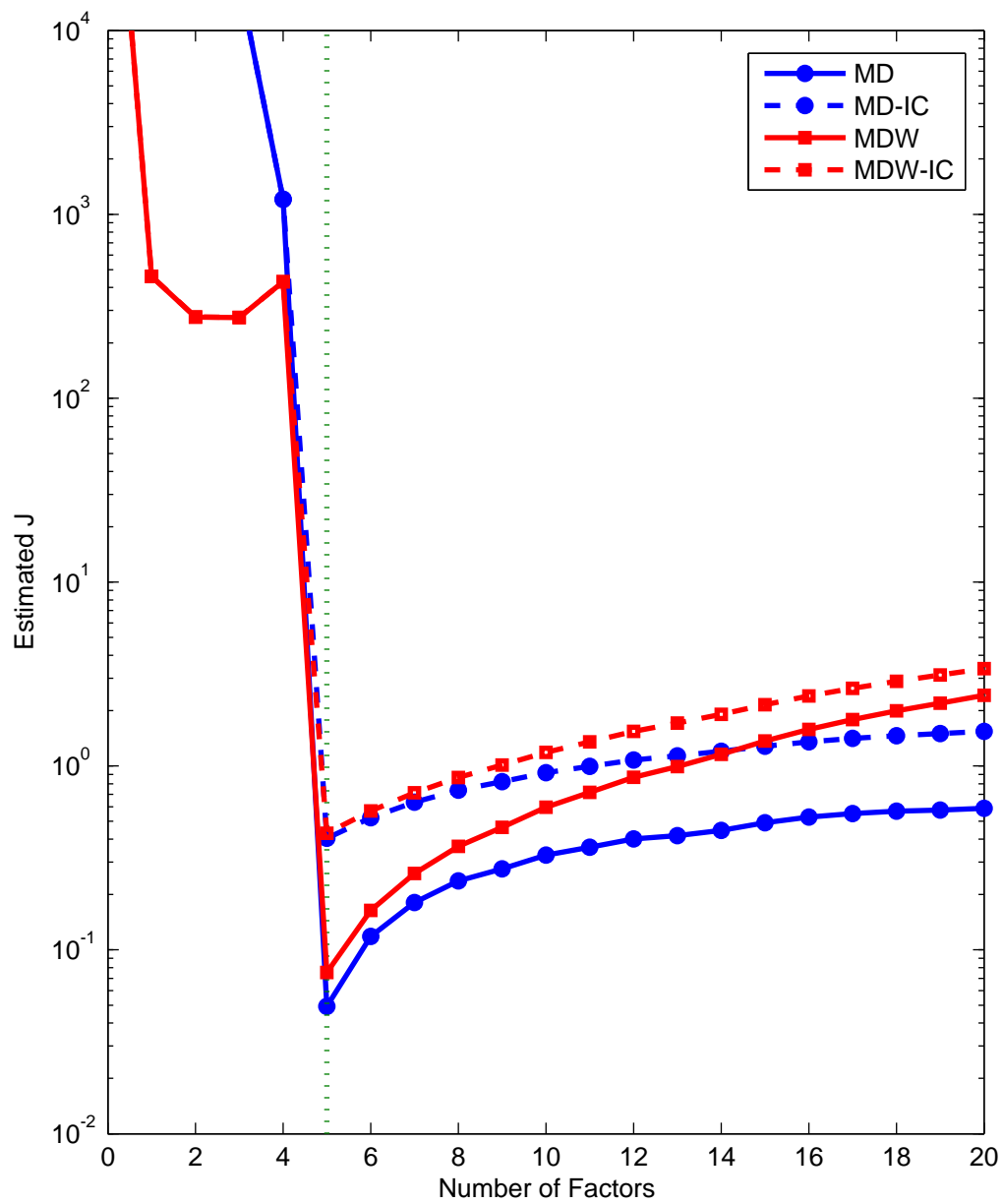


FIGURE 1. Objective function used to estimated the number of factors.

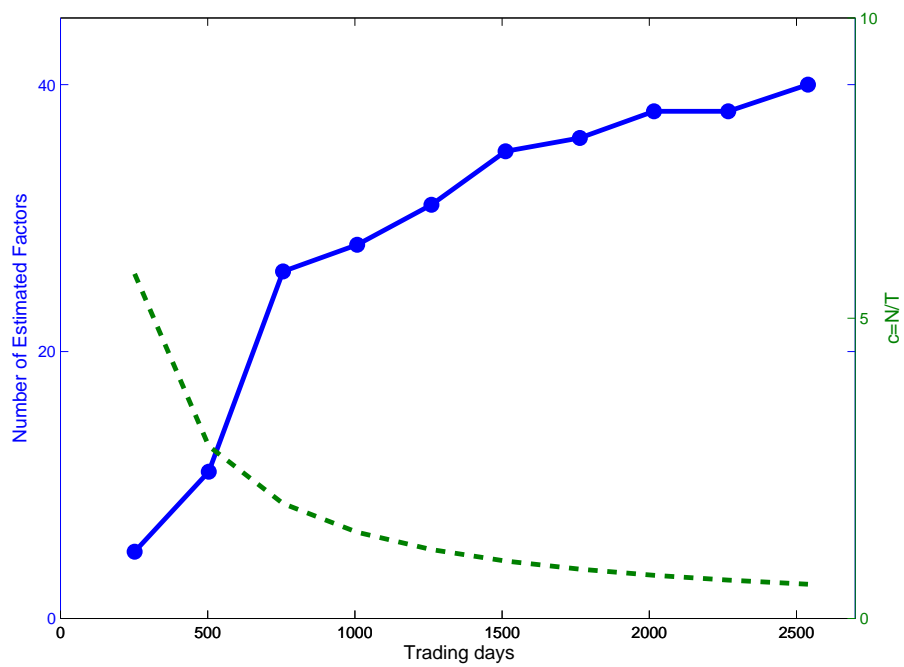


FIGURE 2. Estimation of the number of factors for $N = 300$ stock returns and $T = 252 \dots 2539$ (1 to 10 trading years).

Table 1: Estimating the Number of Strong Factors. # Factors = 5

Number of Factors														
N	T	c	Mean						MSE					
			MD	MDW	MD-IC	MDW-IC	BN5	BN15	MD	MDW	MD-IC	MDW-IC	BN5	BN15
30	100	0.3	5.198	5.074	5.020	5.028	4.976	4.984	0.586	0.082	0.020	0.028	0.024	0.016
50	100	0.5	5.344	5.246	5.136	5.042	4.422	4.422	0.660	0.394	0.156	0.046	0.582	0.582
70	100	0.7	5.488	5.532	5.312	5.036	3.934	3.944	0.716	0.940	0.416	0.044	1.238	1.208
90	100	0.9	5.712	6.054	5.484	5.040	3.568	3.568	1.288	2.666	0.704	0.048	2.296	2.300
90	300	0.3	5.118	5.100	5.006	5.002	4.004	3.998	0.122	0.104	0.006	0.002	0.996	1.006
150	300	0.5	5.248	5.190	5.096	5.012	3.472	3.442	0.308	0.234	0.112	0.012	2.584	2.674
210	300	0.7	5.428	5.434	5.244	5.004	3.004	3.006	0.612	0.578	0.300	0.004	3.988	3.982
270	300	0.9	5.586	5.662	5.446	5.006	2.984	2.978	0.890	1.050	0.626	0.006	4.080	4.110
150	500	0.3	5.102	5.072	5.002	5.002	3.880	3.912	0.106	0.076	0.002	0.002	1.360	1.264
250	500	0.5	5.256	5.212	5.090	5.002	3.002	3.002	0.308	0.248	0.102	0.002	3.994	3.994
350	500	0.7	5.446	5.412	5.252	5.000	2.998	3.000	0.586	0.496	0.284	0.000	4.010	4.000
450	500	0.9	5.566	5.624	5.418	5.006	2.806	2.806	0.846	0.948	0.570	0.006	4.970	4.970

Idiosyncratic Variance													
N	T	c	Mean Bias				MSE						
			MD	MDW	MD-IC	MDW-IC	MD	MDW	MD-IC	MDW-IC			
30	100	0.3	-0.081	-0.060	-0.072	-0.058	0.008	0.004	0.006	0.004			
50	100	0.5	-0.078	-0.066	-0.069	-0.058	0.007	0.005	0.005	0.004			
70	100	0.7	-0.074	-0.071	-0.067	-0.056	0.006	0.006	0.005	0.003			
90	100	0.9	-0.077	-0.085	-0.069	-0.056	0.007	0.009	0.005	0.003			
90	300	0.3	-0.025	-0.020	-0.022	-0.018	0.001	0.000	0.001	0.000			
150	300	0.5	-0.024	-0.020	-0.022	-0.018	0.001	0.000	0.001	0.000			
210	300	0.7	-0.024	-0.023	-0.022	-0.018	0.001	0.001	0.001	0.000			
270	300	0.9	-0.024	-0.024	-0.022	-0.018	0.001	0.001	0.001	0.000			
150	500	0.3	-0.015	-0.011	-0.014	-0.011	0.000	0.000	0.000	0.000			
250	500	0.5	-0.015	-0.012	-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			

Estimators used

- a) MD Minimum Distance Parameter Estimation and J-test Objective Function
- b) MDW Two-Step Minimum Distance Parameter Estimation with Optimal Covariance Matrix and J-test
- c) MD-IC Augmentation of J-test in Estimator a) with Panel Information Criterion
- d) MDW-IC Augmentation of J-test in Estimator b) with Panel Information Criterion
- e) BN5 Bai and Ng (2002) Estimator with kmax=5
- f) BN15 Bai and Ng (2002) Estimator with kmax=15

Table 2: Estimating the Number of Weak Factors. # Factors = 5

Number of Factors														
N	T	c	Mean						MSE					
			MD	MDW	MD-IC	MDW-IC	BN5	BN15	MD	MDW	MD-IC	MDW-IC	BN5	BN15
30	100	0.3	5.176	5.064	5.028	5.016	3.798	12.904	0.444	0.076	0.040	0.024	1.606	63.016
50	100	0.5	5.256	5.178	5.088	4.992	2.774	6.100	0.528	0.306	0.180	0.072	5.134	1.700
70	100	0.7	5.304	5.314	5.110	4.812	2.144	3.984	0.696	0.754	0.354	0.260	8.308	1.232
90	100	0.9	5.444	5.738	5.294	4.774	1.740	3.102	0.816	2.010	0.586	0.274	10.840	3.742
90	300	0.3	5.082	5.084	5.008	5.008	2.026	2.976	0.086	0.088	0.008	0.008	8.870	4.120
150	300	0.5	5.196	5.146	5.062	4.998	1.258	1.986	0.264	0.162	0.110	0.014	14.194	9.102
210	300	0.7	5.302	5.296	5.152	4.890	0.998	1.222	0.494	0.424	0.272	0.114	16.018	14.446
270	300	0.9	5.378	5.452	5.238	4.676	0.810	1.000	0.706	0.768	0.514	0.328	17.710	16.000
150	500	0.3	5.090	5.078	5.008	5.008	1.672	2.000	0.102	0.086	0.008	0.008	11.296	9.000
250	500	0.5	5.190	5.126	5.034	4.992	1.000	1.076	0.234	0.146	0.058	0.008	16.000	15.468
350	500	0.7	5.274	5.232	5.120	4.890	0.844	0.996	0.426	0.316	0.272	0.114	17.404	16.036
450	500	0.9	5.308	5.414	5.182	4.644	0.106	0.598	0.572	0.646	0.410	0.356	24.046	19.618

Idiosyncratic Variance													
N	T	c	Mean Bias				MSE						
			MD	MDW	MD-IC	MDW-IC	MD	MDW	MD-IC	MDW-IC			
30	100	0.3	-0.088	-0.067	-0.080	-0.065	0.009	0.005	0.008	0.005			
50	100	0.5	-0.081	-0.070	-0.073	-0.063	0.008	0.006	0.006	0.005			
70	100	0.7	-0.076	-0.073	-0.068	-0.057	0.007	0.006	0.005	0.004			
90	100	0.9	-0.077	-0.083	-0.071	-0.056	0.007	0.008	0.006	0.004			
90	300	0.3	-0.028	-0.022	-0.026	-0.021	0.001	0.001	0.001	0.001			
150	300	0.5	-0.026	-0.022	-0.024	-0.020	0.001	0.001	0.001	0.000			
210	300	0.7	-0.026	-0.025	-0.024	-0.020	0.001	0.001	0.001	0.000			
270	300	0.9	-0.025	-0.026	-0.023	-0.018	0.001	0.001	0.001	0.000			
150	500	0.3	-0.017	-0.013	-0.016	-0.013	0.000	0.000	0.000	0.000			
250	500	0.5	-0.016	-0.013	-0.014	-0.012	0.000	0.000	0.000	0.000			
350	500	0.7	-0.015	-0.014	-0.014	-0.012	0.000	0.000	0.000	0.000			
450	500	0.9	-0.015	-0.016	-0.014	-0.011	0.000	0.000	0.000	0.000			

Estimators used

- a) MD Minimum Distance Parameter Estimation and J-test Objective Function
- b) MDW Two-Step Minimum Distance Parameter Estimation with Optimal Covariance Matrix and J-test
- c) MD-IC Augmentation of J-test in Estimator a) with Panel Information Criterion
- d) MDW-IC Augmentation of J-test in Estimator b) with Panel Information Criterion
- e) BN5 Bai and Ng (2002) Estimator with kmax=5
- f) BN15 Bai and Ng (2002) Estimator with kmax=15

Table 3: Estimating the Number of Strong Factors. # Factors = 5. Misspecified model with AR1(0.1) Idiosyncratic Errors.

Number of Factors										
N	T	c	Mean				MSE			
			MD	MDW	MD-IC	MDW-IC	MD	MDW	MD-IC	MDW-IC
30	100	0.3	5.226	5.120	5.062	5.034	0.394	0.144	0.066	0.034
50	100	0.5	5.460	5.384	5.212	5.044	0.664	0.572	0.244	0.048
70	100	0.7	5.792	5.878	5.480	5.092	1.500	1.846	0.696	0.108
90	100	0.9	6.272	6.792	5.914	5.152	2.964	5.328	1.762	0.172
90	300	0.3	5.344	5.292	5.040	5.028	0.448	0.368	0.044	0.028
150	300	0.5	5.886	5.804	5.476	5.052	1.358	1.228	0.592	0.056
210	300	0.7	6.538	6.670	6.146	5.066	3.262	3.802	2.034	0.070
270	300	0.9	7.338	7.884	6.978	5.146	6.606	9.864	4.902	0.150
150	500	0.3	5.596	5.536	5.088	5.028	0.796	0.680	0.092	0.028
250	500	0.5	6.500	6.450	5.916	5.112	2.956	2.866	1.344	0.116
350	500	0.7	7.574	7.784	7.064	5.268	7.598	8.944	5.044	0.320
450	500	0.9	8.486	9.028	8.126	5.366	13.118	17.192	10.758	0.426

Idiosyncratic Variance										
N	T	c	Mean Bias				MSE			
			MD	MDW	MD-IC	MDW-IC	MD	MDW	MD-IC	MDW-IC
30	100	0.3	-0.084	-0.065	-0.075	-0.061	0.009	0.005	0.007	0.005
50	100	0.5	-0.079	-0.070	-0.067	-0.058	0.007	0.006	0.005	0.004
70	100	0.7	-0.087	-0.087	-0.074	-0.062	0.009	0.009	0.006	0.004
90	100	0.9	-0.093	-0.109	-0.081	-0.063	0.010	0.014	0.008	0.004
90	300	0.3	-0.028	-0.024	-0.021	-0.020	0.001	0.001	0.001	0.001
150	300	0.5	-0.031	-0.029	-0.023	-0.020	0.001	0.001	0.001	0.000
210	300	0.7	-0.035	-0.037	-0.029	-0.020	0.001	0.002	0.001	0.000
270	300	0.9	-0.040	-0.047	-0.035	-0.021	0.002	0.002	0.001	0.000
150	500	0.3	-0.017	-0.016	-0.010	-0.011	0.000	0.000	0.000	0.000
250	500	0.5	-0.022	-0.022	-0.016	-0.012	0.001	0.001	0.000	0.000
350	500	0.7	-0.027	-0.030	-0.023	-0.013	0.001	0.001	0.001	0.000
450	500	0.9	-0.030	-0.036	-0.027	-0.014	0.001	0.001	0.001	0.000

Estimators used

- a) MD Minimum Distance Parameter Estimation and J-test Objective Function
- b) MDW Two-Step Minimum Distance Parameter Estimation with Optimal Covariance Matrix and J-test
- c) MD-IC Augmentation of J-test in Estimator a) with Panel Information Criterion
- d) MDW-IC Augmentation of J-test in Estimator b) with Panel Information Criterion

Table 4: Estimating the Number of Factors with Autocorrelated Idiosyncratic Errors

		Factors									
N	T	c	Mean				MSE				
			MD	MD-IC	BN5	BN15	MD	MD-IC	BN5	BN15	
30	100	0.3	5.460	5.000	4.984	4.986	0.612	0.000	0.016	0.018	
50	100	0.5	6.640	5.004	4.466	4.466	4.312	0.004	0.538	0.538	
70	100	0.7	7.358	5.022	3.954	3.954	9.306	0.030	1.194	1.194	
90	100	0.9	6.730	5.014	3.608	3.608	6.406	0.014	2.176	2.176	
90	300	0.3	6.770	5.000	4.006	4.006	6.134	0.000	0.994	0.994	
150	300	0.5	6.098	5.000	3.526	3.526	3.150	0.000	2.422	2.422	
210	300	0.7	5.720	5.000	3.008	3.008	1.548	0.000	3.976	3.976	
270	300	0.9	5.560	5.002	2.988	2.988	1.096	0.002	4.060	4.060	
150	500	0.3	6.284	5.000	3.904	3.904	3.968	0.000	1.288	1.288	
250	500	0.5	5.840	5.000	3.002	3.002	1.928	0.000	3.994	3.994	
350	500	0.7	5.710	5.000	2.994	2.994	1.338	0.000	4.030	4.030	
450	500	0.9	5.580	5.002	2.834	2.834	1.008	0.002	4.830	4.830	

		Idiosyncratic Variance					Time Series Correlation				
N	T	c	Mean Bias			MSE		Mean Bias		MSE	
			MD	MD-IC	MD	MD-IC	MD	MD-IC	MD	MD-IC	
30	100	0.3	-0.092	-0.068	0.010	0.006	-0.168	-0.104	0.039	0.025	
50	100	0.5	-0.130	-0.067	0.020	0.005	-0.148	-0.028	0.029	0.005	
70	100	0.7	-0.144	-0.068	0.025	0.005	-0.106	-0.015	0.017	0.003	
90	100	0.9	-0.117	-0.064	0.017	0.004	-0.046	-0.001	0.005	0.001	
90	300	0.3	-0.049	-0.021	0.003	0.001	-0.094	-0.021	0.013	0.001	
150	300	0.5	-0.036	-0.022	0.002	0.001	-0.031	-0.005	0.002	0.000	
210	300	0.7	-0.031	-0.023	0.001	0.001	-0.015	-0.002	0.000	0.000	
270	300	0.9	-0.027	-0.022	0.001	0.001	-0.008	0.001	0.000	0.000	
150	500	0.3	-0.026	-0.014	0.001	0.000	-0.042	-0.012	0.003	0.000	
250	500	0.5	-0.020	-0.013	0.001	0.000	-0.016	-0.004	0.000	0.000	
350	500	0.7	-0.018	-0.013	0.000	0.000	-0.008	0.000	0.000	0.000	
450	500	0.9	-0.016	-0.013	0.000	0.000	-0.004	0.001	0.000	0.000	

Estimators used

- a) MD Minimum Distance Parameter Estimation and J-test Objective Function
- b) MD-IC Augmentation of J-test in Estimator a) with Panel Information Criterion
- c) BN5 Bai and Ng (2002) Estimator with kmax=5
- d) BN15 Bai and Ng (2002) Estimator with kmax=15

Table 5: Inconsistency of PCA for Weak Factors

Mean Bias												
N	T	c	PC1	PC2	PC3	PC4	PC5	IV1	IV2	IV3	IV4	IV5
70	100	0.7	0.202	0.067	-0.186	-0.016	-0.140	-0.094	0.073	-0.090	-0.084	0.094
90	100	0.9	-1.189	0.409	-1.678	-0.425	3.726	-0.128	0.096	0.122	-0.072	0.126
210	300	0.7	39.957	32.094	6.431	-52.497	-8.939	-0.014	0.023	0.006	-0.054	0.003
350	500	0.7	0.091	0.059	-0.022	-0.069	-0.058	-0.001	-0.011	0.009	-0.009	0.018
450	500	0.9	0.006	0.002	0.005	-0.014	-0.001	-0.002	0.045	0.019	-0.024	0.044

Median Bias												
N	T	c	PC1	PC2	PC3	PC4	PC5	IV1	IV2	IV3	IV4	IV5
70	100	0.7	0.000	-0.003	-0.013	0.010	-0.003	-0.105	0.068	-0.065	-0.081	0.113
90	100	0.9	0.007	0.003	0.010	-0.007	0.013	-0.102	0.087	0.112	-0.074	0.098
210	300	0.7	0.002	-0.003	-0.005	-0.005	-0.002	-0.017	0.019	0.016	-0.031	0.023
350	500	0.7	-0.001	-0.004	0.003	0.006	0.003	0.001	0.016	0.014	-0.003	0.013
450	500	0.9	0.006	-0.001	0.004	0.003	-0.004	0.006	0.037	0.017	-0.031	0.048

MSE												
N	T	c	PC1	PC2	PC3	PC4	PC5	IV1	IV2	IV3	IV4	IV5
70	100	0.7	19.102	4.802	12.596	5.541	13.308	0.373	0.329	0.297	0.301	0.351
90	100	0.9	949.050	173.110	2414.200	606.960	12735	0.490	0.472	0.393	0.462	0.455
210	300	0.7	1337400	895190	50012	2451500	75091	0.423	0.293	0.268	0.348	0.297
350	500	0.7	4.771	2.633	0.421	3.836	2.689	0.352	0.289	0.192	0.346	0.239
450	500	0.9	0.627	0.168	0.076	0.480	0.165	0.848	0.689	0.548	0.738	0.764

Table 6(a): Estimation of Factor Loadings ($c < 1$)

			Mean Bias							
N	T	c	PCA	2SLS	Fuller	BC2SLS	CUE	T-2SLS	T-CUE	
30	100	0.3	0.016	-0.230	-0.003	0.028	0.010	-0.014	0.017	
50	100	0.5	0.010	-0.338	0.009	0.047	0.024	-0.040	0.010	
70	100	0.7	0.013	-0.413	0.022	0.105	0.036	-0.054	0.014	
90	100	0.9	0.009	-0.470	0.020	0.157	-0.004	-0.072	0.010	
90	300	0.3	0.003	-0.237	0.000	0.012	0.004	-0.026	0.003	
150	300	0.5	0.006	-0.341	-0.001	0.021	0.005	-0.041	0.005	
210	300	0.7	0.008	-0.416	0.005	0.022	0.013	-0.055	0.008	
270	300	0.9	0.005	-0.474	0.038	0.036	0.111	-0.074	0.005	
150	500	0.3	0.001	-0.240	-0.001	0.005	0.002	-0.028	0.001	
250	500	0.5	-0.002	-0.342	-0.001	0.011	0.004	-0.047	-0.002	
350	500	0.7	0.004	-0.417	0.000	0.016	0.006	-0.059	0.004	
450	500	0.9	0.001	-0.473	0.016	0.022	0.102	-0.076	0.001	

			MSE							
N	T	c	PCA	2SLS	Fuller	BC2SLS	CUE	T-2SLS	T-CUE	
30	100	0.3	0.024	0.066	0.028	0.065	0.030	0.022	0.025	
50	100	0.5	0.023	0.125	0.039	0.127	0.045	0.022	0.025	
70	100	0.7	0.023	0.180	0.078	0.334	0.094	0.021	0.024	
90	100	0.9	0.023	0.229	0.394	0.459	0.289	0.022	0.023	
90	300	0.3	0.007	0.061	0.009	0.019	0.009	0.008	0.008	
150	300	0.5	0.008	0.120	0.013	0.033	0.013	0.008	0.007	
210	300	0.7	0.008	0.176	0.018	0.047	0.019	0.009	0.007	
270	300	0.9	0.007	0.228	0.117	0.065	0.205	0.012	0.009	
150	500	0.3	0.004	0.060	0.005	0.010	0.005	0.005	0.004	
250	500	0.5	0.004	0.119	0.007	0.017	0.007	0.006	0.005	
350	500	0.7	0.004	0.175	0.011	0.025	0.012	0.007	0.005	
450	500	0.9	0.004	0.225	0.036	0.033	0.143	0.009	0.005	

Table 6(b): Estimation of Factor Loadings ($c > 1$)

			Mean Bias			MSE		
N	T	c	PCA	T-2SLS	T-CUE	PCA	T-2SLS	T-CUE
150	100	1.5	0.011	-0.116	0.009	0.025	0.029	0.025
300	100	3	0.003	-0.194	0.005	0.024	0.050	0.025
500	100	5	0.022	-0.251	0.023	0.026	0.074	0.027
450	300	1.5	0.004	-0.115	0.004	0.008	0.019	0.008
900	300	3	-0.001	-0.191	-0.001	0.007	0.040	0.008
1500	300	5	0.005	-0.250	0.005	0.008	0.066	0.009
750	500	1.5	0.002	-0.116	0.002	0.004	0.017	0.005
1500	500	3	0.000	-0.189	0.000	0.004	0.038	0.005
2500	500	5	0.004	-0.250	0.004	0.004	0.063	0.005