LINEAR ESTIMATION OF THE NUMBER OF FACTORS

Seung C. Ahn
Arizona State University

Marcos F. Perez
Wilfrid Laurier University

This version: March 2008

Abstract

This paper provides a linear method to estimate the number of factors in linear factor models. We generalize the non-linear methodology proposed by Cragg and Donald (1997). The method is based on the estimation of the rank of a covariance matrix using a Minimum Chi-Squared statistic (MINCHI2). MINCHI2 method is general in the sense that it does not make distributional assumptions about the response vector and allows for heteroskedasticity and time series autocorrelation. Nevertheless, MINCHI2 requires a non-linear optimization that often fails to locate solutions. We reexamine the properties of the MINCHI2 estimator using a GMM approach and find that the new estimator involves linear and non-linear moment conditions. Based on this finding, we propose an alternative estimation procedure that utilizes only the linear conditions. Asymptotically, the alternative estimator is less efficient than the MINCHI2 estimator, but it is much easier to compute. Monte Carlo simulations show that the new method has good small sample properties.

LINEAR ESTIMATION OF THE NUMBER OF FACTORS

1. Introduction

The use of factor models becomes increasingly popular in the economic literature. A well known-example is the Arbitrage Price Theory (APT, Ross, 1976), where asset returns are generated by a factor structure. In the finance literature, the APT model has been extensively used to analyze the prices of the systematic risks in the stock, money, or fixed income securities markets. An excellent summary of factor models used in the literature can be found in Campbell, Lo and Mackinlay (1997). There are many other examples. Analyzing the data from G7 countries, Gregory and Head (1999) found that cross-country variations in productivity and investment have common components. Gorman (1981) and Lewbel (1991) found that if consumers are utility maximizers, their budget shares for individual goods or services purchased should be driven by at most three factors. Stock and Watson (2005) showed that many macroeconomic variables in US are driven by a smaller number of common factors. Ahn, Lee and Schmidt (2006) showed that the time pattern of the fluctuations in individual firms' technical productivities can be estimated based on a factor model. More examples can be found in Bai (2003). For these studies, the estimation of the common factors is an important research task.

To identify and estimate the factors, it is important to estimate the correct number of factors (L). This is particularly true for testing the APT hypothesis, which predicts that a vector of cross-section expected returns on risky asset should be a linear combination of a vector of ones and the matrix of factor loadings. The APT hypothesis is often tested by conducting cross-section regression of average returns on estimated factor loadings (*e.g.* Black, Jensen and Scholes, 1972; Fama and MacBeth, 1973; Fama and French, 1993). The coefficients of the factor loadings are factor risk prices. If too few factors are used in the estimation of an APT model, the estimated factor prices will be inconsistent. In contrast, if too many factors are used, that is, if some of the used factors are not correlated with returns, the factor prices are not identified (Kan and Zhang, 1999).

In this paper we present a simple GMM method that can be use to consistently estimate the number of factors in a linear factor model under general distributional assumptions. For the case *exact* factor models the method is appropriate if either the number of cross-sectionals or the number time-series observations is large. It can also be used for approximate factor models in which both the numbers of cross-section and time series observations are large. The advantages of the new method are as follows. First, the method is computationally easy to implement. All necessary procedures are based on closed-form solutions. Any software that can implement three stage least squares (3SLS) can be used. Second, the method can be applied as long as the error components are either cross-sectionally independent or zero-autocorrelated. Approximate models can be also analyzed if we are able to construct portfolios with the characteristic that number of individual assets in each portfolio is sufficiently large.

The earlier empirical studies of the APT factor models (*e.g.*, Roll and Ross, 1980; and Brown and Weinstein, 1983) were based on the maximum likelihood method of Jöreskog (1967). Using this method, a researcher estimates factor loadings and variances of idiosyncratic errors of asset returns concurrently, and test for the number of latent factors using a likelihood-ratio test. However, the maximum-likelihood method requires quite restrictive assumptions to hold, such as the assumption that the idiosyncratic error terms in a model are independently distributed over different cross-section units and over time. More general approaches have been developed that allow cross-sectional correlations (approximate factor models) and non-normal data. A common approach is to construct candidate factors, repeat the estimation and testing of the model for different number of factors (L), and observe if the tests are sensitive to increasing L. Lehman & Modest (1988) and Connor & Korajczyk (1988) used this technique to analyze the US stock returns. Success of this method would depend on the quality of the chosen candidate factors. Another approach is to use the estimators of the ranks of matrices (*e.g.*, Gill and Lewbel, 1992; Cragg and Donald, 1996, 1997). If the idiosyncratic error components of the response variables analyzed are cross-sectionally independent (exact factor model), the variance matrix of the response variables (*e.g.*, returns) is decomposed into a diagonal matrix and a matrix with a rank equal to L. Thus, the number of the common factors (L) can be found by estimating the rank of the difference between the estimates of the variance and the diagonal matrices. A pitfall of this approach is that it is computationally burdensome, especially if the number of response variables analyzed is large.[1]

[1] Rank of a matrix can be estimated by the Lower-Diagonal-Upper triangular decomposition test (LDU)

More recently, Bai and Ng (2002) have developed a general estimation method for the number of factors. Their least squares (LS) estimation method is designed for data with large number of response variables (*N*) over a large number of time series observations (*T*). The idiosyncratic errors of the response variables are allowed to be cross-sectionally correlated and autocorrelated. Chamberlain and Rothschild (1983) show that the APT hypothesis holds even if idiosyncratic component of the asset returns are cross-sectionally correlated (approximate factor models). Thus, the LS method of Bai and Ng (2002) would be a useful tool for testing the APT hypothesis. However, it is important to note that the method results in inconsistent estimators if either *N* or *T* is small (Bai, 2003). The simulation results reported in Bai and Ng (2002) also indicate that the number of factors is not accurately estimated if one of *N* and *T* is less than 40. Thus, the LS method would be inappropriate for the studies using smaller sets of response variables.

The rest of the paper is organized as follows. Section 2 is devoted to the estimation procedure. Section 3 exhibits our Monte Carlo simulation results and finite-sample properties of our test. Section 4 applies our method to the US stock market. Concluding remarks are provided in Section 5.

2. **Model and Assumptions**

The object of our study is a linear factor model, where a response vector from a population linearly depends of finite number of underlying common factors (L_0). Specifically:

$$r_{it} = \alpha_i + \beta_i' f_t + \varepsilon_{it}\,, \tag{1}$$

---

developed by Gill and Lewbel (1992) and Cragg and Donald (1996). This method requires a Gaussian elimination procedure and division of the response variables into two non-overlapping groups. The Gaussian elimination procedure is complicated if too big matrices are analyzed. Alternatively, Cragg and Donald (1997) propose a Minimum Chi-Squared statistic (MINCHI2). This method is general in the sense that it requires no distributional assumption about the response variables and allows for heteroskedasticity and autocorrelation. But the principal problem of MINCHI2 is that some nonlinear optimization procedures are required and the procedures often fail to locate solutions. Donald, Fortuna and Pipiras (2005) also suggest some theoretical problems related to the methods.

where r_{it} is the value of the response variable i (= 1, 2, …, N) at the time t (= 1, 2, …, T), α_i is an intercept, f_t is an $L \times 1$ vector of unobservable common factors, β_i is an $1 \times L$ vector of the factor loadings for the response variable i, and the ε_{it} are the idiosyncratic components of response variables which are cross-sectionally uncorrelated. Thus, the response variables r_{it} are cross-sectionally correlated only through the common factors f_t. Usual factor analysis typically applies to demeaned data with $E(r_{it}) = 0$ for all i and t. But we do not impose such restrictions. An exact factor model implies that the response variables x_{it} are correlated only through common factors f_t, this implies that the idiosyncratic components ε_{it} are cross-sectionally uncorrelated. Fist we assume that N is relatively small and T is large, so the asymptotic theory will work for $N \to \infty$ as T is fixed. Modifications for in which T is large and N is small are straight forward and will be considered later.

For convenience, we adopt the following notation. We use $r_{\bullet t}$ to denote the vector that includes all the cross-sectional observations of the response variable r_{it} at time t. Similarly, $r_{i\bullet}$ denotes the vector including all of the time series observations of r_{it} for the response variable i. The vectors $\varepsilon_{i\bullet}$ and $\varepsilon_{\bullet t}$ are similarly defined. Using this notation, we can stack the equations in (1) for given t by:

$$r_{\bullet t} = \alpha + \mathrm{B} f_t + \varepsilon_{\bullet t}, \tag{2}$$

where $\alpha = (\alpha_1, \alpha_2, ..., \alpha_N)'$ and $\mathrm{B} = (\beta_1, \beta_2, ..., \beta_N)'$. Including the non-zero vector of response-variable-specific intercepts into the model, we can assume that $E(f_t) = 0$ without loss of generality.

Since our method to estimate the number of factors (L) is an application of GMM, we require a set of sufficient conditions under which usual GMM theories apply and the number of factors can be identified. For asymptotics, we use "$\to_p$" and "$\to_d$" to denote "converges in probability" and "converges in distribution," respectively. The basic assumptions are the following:

Assumption A: The factors in f_t are non-constant variables with finite moments up to the fourth order, $E(f_t) = 0_{L_o \times 1}$ and $E(f_t f_t') = \Omega_f$ for all *t*, and $T^{-1}\Sigma_{t=1}^{T} f_t f_t' \to_p \Omega_f$ as $T \to \infty$, where Ω_f is a $L_o \times L_o$ finite and positive definite matrix.

Assumption B: $rank(\mathrm{B}) = L_o$.

Assumption C: There exists a constant $m \in (0, \infty)$, such that for all *T* (with fixed *N*), (C1) the errors ε_{it} have finite moments up to the eighth order with $E(\varepsilon_{it} \mid f_1, f_2, ..., f_t) = 0$ for all *i* and *t*; (C2) $E(\varepsilon_{it}\varepsilon_{i's} \mid f_1, f_2, ..., f_t) = 0$ for all $i \neq i'$, $s \geq t$; (C3) $\left| T^{-1}\Sigma_{t=1}^{T}\Sigma_{s=1}^{T} E(\varepsilon_{is}\varepsilon_{it}) \right| \leq m$ for all *i* and *t*; (C4) $T^{-1/2}\Sigma_{t=1}^{T} w_t \to_d N\left(0_{(N+L_o)\times 1}, \Lambda\right)$, as $T \to \infty$, where $w_t = (h_t \otimes \varepsilon_{\bullet t}) - E(h_t \otimes \varepsilon_{\bullet t})$, $h_t = (1, f_t', \varepsilon_{\bullet t}')'$, and $\Lambda = p\lim_{T\to\infty} T^{-1}\Sigma_{s=1}^{T}\Sigma_{t=1}^{T} E(w_t w_s')$.

In Assumption A, we assume that the factors are covariance stationary; that is, the variance matrix of f_t, $Var(f_t) = \Omega_f$, is same for all t. We adopt this assumption for expository convenience and it can be relaxed without altering our results. The required assumption is that $T^{-1}\Sigma_{t=1}^{T} f_t f_t' \to_p \Omega_f$ as $T \to \infty$. Most of the general mixing processes satisfy this condition (White, 1999).

Assumption B implies that the true number of factors is L_o. Under Assumption (C1), the factors are weakly exogenous to the idiosyncratic errors. Assumption (C2) restricts the error terms to be cross-sectionally uncorrelated[2]. Thus, with (C2), the model (2) is an exact factor model. If some instrumental variables correlated with the factors

---

[2] Alternatively, when the errors are cross-sectionally correlated, but not autocorrelated over time, an exact model can be obtained by rewriting the model (2) as $r_{i\bullet} = F\beta_i + \varepsilon_{i\bullet}$, where $F = (f_1, f_2, ..., f_T)'$. If the errors are serially uncorrelated, the variance matrix of $\varepsilon_{i\bullet}$ becomes diagonal. When *T* is small, we can estimate L_o by applying the method we discuss below to this alternative model.

are observable, we could use them to estimate the number of factors, even allowing the errors to be cross-sectionally correlated. Such cases will be discussed in section 4.2. Also, even if the errors are cross sectionally correlated, the method can be used to estimate the number of factors by grouping the response variables appropriately (e.g., portfolios), as explained in section 4.1

Assumption (C3) indicates that the autocovariances of the error terms are absolutely summable, while (C4) is nothing but a central limit theorem. When factors and errors follow general mixing processes, both Assumptions (C3) and (C4) hold.

3. Estimation of the Number of Factors

. For notational convenience let assume that $E[r_{\bullet t}] = \alpha = 0$, this assumption is made without lost of generality as it will be explained later. Given this assumption we can write equation (2) as:

$$r_{\bullet t} = \mathrm{B} f_t + \varepsilon_{\bullet t} \tag{3}$$

Under Assumptions A-C, the covariance matrix of the response variable can be written as:

$$E[r_{\bullet t} r_{\bullet t}'] = \Sigma = \mathrm{B}\Omega_f \mathrm{B}' + \Psi , \tag{4}$$

where Ψ is the $N \times N$ diagonal matrix of the variances of ε_{it} and Σ the covariance matrix of the response variable. Many of the methods popularly used for factor analysis are based in this variance decomposition. The ML estimation of Jöreskog (1967) and the Minimum Chi-Squared statistic (MINCHI2) of Cragg and Donald (1997) estimate L_o based on estimates of B and Ψ. But use of these methods is somewhat limited. The legitimacy of the ML method requires some strong distributional assumptions on data such as normality. Use of MINCHI2 does not require such strong distributional assumptions, but it often suffers from the computational difficulty of estimating Ψ, since non linear moment conditions are involved in the estimation.

Our moment conditions are based in the existence a matrix $H_o = (H_{1,o}, H_{2,o}\ )'$ which is a $N \times (N - L_o)$ full column matrix such that $H_o{}'\beta = 0_{(N-L_0)\times 1}$. Clearly, the matrix H_o is not unique since for any nonsingular matrix A, $(H_o A)'\beta = 0_{(N-L_o)\times 1}$. To avoid this under-identification problem, the restriction $H_{o,1} = -I_{N-L_o}$ is imposed. Then if we multiply equation (3) by the matrix H_o the following equation holds:

$$E[H_o{}' r_{\bullet t} r_{\bullet t}{}' - H_o{}'\Psi] = H_o{}'\Sigma - H_o{}'\Psi = 0_{(N-L_o)\times N} \tag{4}$$

Note that equation (4) is defined at the true number of factors L_o and at the true parameter values. Based on equation (4) we construct the moment conditions that will be used in the estimation. Let us denote by L the number of factors we use for estimation which could be different from L_o. The corresponding matrix H_L will have the form: $H_L = (H_{1,L}, H_{2,L}\ )'$ which is a $N \times (N - L)$ full column matrix and $H_{1,L} = -I_{N-L}$. Given the structure of H_L and Ψ we have to estimate $(N - L\ ) \times L$ parameters of the matrix $H_{2,L}$, and *N* parameters of the matrix Ψ. Let's call θ_L the vector of these $L \times (N - L) + N$ parameters. We define the moment function:

$$m_t(\theta_L / L) = vec(H_L{}' r_{\bullet t} r_{\bullet t}{}' - H_L{}'\varepsilon_{\bullet t}\varepsilon_{\bullet t}{}') \tag{5}$$
$$E[m_t(\theta_L \mid L\ )] = vec[MC(\theta_L \mid L)] = vec(H_L{}'\Sigma - H_L{}'\Psi)$$

Thus, the moment condition (5) implies that under Assumptions A-C, when $L = L_o$, $E[m_t(\theta_L \mid L\ )] = 0$ if and only if $\theta_L = \theta$. That is, our moment conditions will hold just at the true value of the parameters, and if and only if the true number of factors (L_o) was used in the estimation. Let analyze in detail the set of moment conditions in (5). For any value of *L,* the matrices Σ and Ψ can be written as partitioned matrices as follows:

$$\Psi = \begin{pmatrix} \underset{(N-L)\times(N-L)}{\Psi^{11}} & \underset{(N-L)\times L}{0} \\ \underset{L\times(N-L)}{0} & \underset{L\times L}{\Psi^{22}} \end{pmatrix}_{N\times N} \qquad \Sigma = \begin{pmatrix} \underset{(N-L)\times(N-L)}{\Sigma^{11}} & \underset{(N-L)\times L}{\Sigma^{12}} \\ \underset{L\times(N-L)}{\Sigma^{12}} & \underset{L\times L}{\Sigma^{22}} \end{pmatrix}_{N\times N}$$

Given the form of $H = (-I_{N-L}, H_2)'$ the moment conditions on (5) can be written as:

$$MC(\theta_L / L) = \left(-I_{N-L}, H_2\right) \begin{pmatrix} \underset{(N-L)\times(N-L)}{\Sigma^{11}} & \underset{(N-L)\times L}{\Sigma^{12}} \\ \underset{L\times(N-L)}{\Sigma^{12}} & \underset{L\times L}{\Sigma^{22}} \end{pmatrix}_{N\times N} - \left(-I_{N-L}, H_2\right) \begin{pmatrix} \underset{(N-L)\times(N-L)}{\Psi^{11}} & \underset{(N-L)\times L}{0} \\ \underset{L\times(N-L)}{0} & \underset{L\times L}{\Psi^{22}} \end{pmatrix}_{N\times N} = 0$$

this implies that:

$$MC(\theta_L / L) = \begin{bmatrix} \Sigma^{11} - \Psi^{11} - H_2{}'\Sigma^{12} \\ \Sigma^{12} - H_2{}'(\Psi^{22} + \Sigma^{22}) \end{bmatrix} = 0`$$

We can write these moments conditions as two subsets of moment conditions:

$$MC1(\theta_L / L) = \left[\Sigma^{11} - \Psi^{11} - H_2{}'\Sigma^{12}\right] = 0$$

$$MC2(\theta_L / L) = \left[\Sigma_t^{12} - H_2{}'(\Sigma_t^{22} + \Psi^{22})\right] = 0$$

MC1 involves $(N-L)^2$ moment conditions, and all of them are linear in θ_L. MC2 includes $(N-L)\times L$ moment conditions and all of them are non linear in θ_L. MC1 involves $(N-L)\times L$ parameters of the matrix $H_{2,L}$ and $(N-L)$ parameters of the matrix Ψ^{11}. MC2 includes the $(N-L)\times L$ parameters of the matrix $H_{2,L}$ and L parameters of the matrix Ψ^{22}. Then we can define each subset of the moment conditions as function of their parameters as follows:

$$MC\left(\theta_{H_{2,L}}, \theta_{\Psi^{11}}, \theta_{\Psi^{22}} / L\right) = \begin{bmatrix} MC1\ (\theta_{H_{2,L}}, \theta_{\Psi^{11}} / L) \\ MC2\ (\theta_{H_{2,L}}, \theta_{\Psi^{22}} / L) \end{bmatrix}$$

Given that Ψ^{11} is diagonal, the $(N-L)$ moment conditions in the diagonal of MC1 are the only ones that involve the $(N-L)$ parameters of Ψ^{11}. Define the off diagonal elements of MC1 as MC1o and the diagonal elements of MC1 as MC1d as follows:

$$MC1(\theta_{H_{2,L}}, \theta_{\Psi^{11}} / L) = \begin{bmatrix} \text{MC1o}\ (\theta_{H_{2,L}} / L) \\ \text{MC1d}\ (\theta_{H_{2,L}}, \theta_{\Psi^{11}} / L) \end{bmatrix} \tag{6}$$

where MC1d $(\theta_{H_{2,L}}, \theta_{\Psi^{11}} / L)$ involves $(N-L)$ moment conditions.

Also $MC2\ (\theta_{H_{2,L}}, \theta_{\Psi^{22}} / L)$ can be redefined as two subsets of moment conditions. The first subset will include the $(N-L)\times L$ parameters of the matrix $H_{2,L}$. The second subset will have L moment conditions that will depend of L parameters of the matrix Ψ^{22} and $(N-L)\times L$ parameters of the matrix $H_{2,L}$.

__Lemma 1:__ *There exists a linear combination of* $MC2\ (\theta_{H_{2,L}}, \theta_{\Psi^{22}} / L)$ *such as we can divide the moment conditions in two subsets such as:*

$$MC2(\theta_{H_{2,L}}, \theta_{\Psi^{22}} / L) = \begin{bmatrix} MC2a\ (\theta_{H_{2,L}} / L) \\ MC2b\ (\theta_{H_{2,L}}, \theta_{\Psi^{22}} / L) \end{bmatrix}$$

where $MC2b\ (\theta_{H_{2,L}}, \theta_{\Psi^{22}} / L)$ *includes L moment conditions.*

Using equation (6) and *Lemma 1* the moment condition in (5) can be written as follows:

$$MC\left(\theta_{H_{2,L}},\theta_{\Psi^{11}},\theta_{\Psi^{22}}/L\right)=\begin{bmatrix} MC1o\ (\theta_{H_{2,L}}/L) \\ MC1d\ (\theta_{H_{2,L}},\theta_{\Psi^{11}}/L) \\ MC2a\ (\theta_{H_{2,L}}/L) \\ MC2b\ (\theta_{H_{2,L}},\theta_{\Psi^{22}}/L) \end{bmatrix} \quad (7)$$

Given the moment conditions (7) such as $E[m_t(\theta_L \mid L\ )]=vec[MC(\theta_L \mid L)]$ we now define the GMM minimization problem for any given L:

$$\min_{\theta_L} C_T(\theta_L \mid W_T(L),L)=TM_T(\theta_L)'[W_T(L)]^{-1}M_T(\theta_L), \quad (8)$$

where $\theta_L=(\theta_{H_{2,L}},\theta_{\Psi^{11}},\theta_{\Psi^{22}})$, $M_T(\theta_L \mid L)=T^{-1}\Sigma_{t=1}^{T}m_t(\theta_L \mid L)$ is the sample mean of the moment functions $m_t(\theta_L \mid L)$, and the weighting matrix $W_T(L)$ is a positive-definite matrix with a non-stochastic and finite probability limit, say $W(L)$. As explained before, the set of moment conditions in (7) not just linear but also non linear moment conditions. Then, nonlinear optimization procedures are required and the procedures often fail to locate solutions as documented by Donald, Fortuna and Pipiras (2005) .

Also several moment conditions in (7) are redundant in the sense they are linear combination of each other. Using these facts, we propose an alternative estimation procedure that utilizes only the linear conditions. The following propositions are devoted reduce the number of non-linear moment conditions by replacing them with linear ones.

<u>*Proposition 1:*</u> *Let* $(\hat{\theta}_{H_{2,L}},\hat{\theta}_{\Psi^{11}},\hat{\theta}_{\Psi^{22}})$ *be the solution of the GMM minimization problem defined in (8) that involves all the conditions in* $MC\left(\theta_{H_{2,L}},\theta_{\Psi^{11}},\theta_{\Psi^{22}}/L\right)$ equation *(7). Let* $(\tilde{\theta}_{H_{2,L}})$ *be the solutions of (8) but including only the following moment conditions MCA:*

$$MCA\left(\theta_{H_2,L}/L\right)=\begin{bmatrix} MC1o\ (\theta_{H_{2,L}}/L) \\ MC2a\ (\theta_{H_2,L}/L) \end{bmatrix}$$

Then: $\hat{\theta}_{H_{2,L}} = \tilde{\theta}_{H_{2,L}}$ *and the asymptotic covariance matrix of the estimators is also the same.*

The proof is in the appendix and follows directly from Ahn, Schmidt (1992) separability result. The intuition behind the proof is that the GMM estimators of $\hat{\theta}_{H_{2,L}}$ does not change is we add equal number of parameters and moment conditions. The reason for this result is that the moment condition just define $\theta_{\Psi^{11}}$ *and* $\theta_{\Psi^{22}}$ in terms of $\hat{\theta}_{H_{2,L}}$. Proposition 2 let us to estimate the parameters $\hat{\theta}_{H_2}$ using an smaller set of moment conditions. The original set of moment condition in (7) involves $(N-L)^2$ linear and $(N-L)\times L$ non linear moment conditions and a total of $(N-L)\times L+N$ parameters. Now, $MCA\left(\theta_{H_{2,L}}/L\right)$ involves only $(N-L)(N-L-1)$ linear moment conditions in $MC1o\ (\theta_{H_2,L}/L)$ and $L\times(N-L-1)$ non linear moment conditions in $MC2a\ (\theta_{H_{2,L}}/L)$, and the estimation of $(N-L)\times L$ parameters in $H_{2,L}$.

<u>*Proposition 2*</u> *A least* $L\times(L-1)/2$ *moment conditions from* $MC2a\ (\theta_{H_{2,L}}/L)$ *are not a linear combinations of the moment conditions on* $MC1o\ (\theta_{H_2,L}/L)$.

Proposition 2 implies that we can not replace $L\times(L-1)/2$ non linear moment conditions from $MC2a\ (\theta_{H_{2,L}}/L)$ with any linear combinations of the linear moment conditions in $MC1o\ (\theta_{H_2,L}/L)$. This implies that some moment conditions are not redundant. We argue that if use only the linear moment conditions we will loosing some asymptotic efficiency in the estimation of the parameters $H_{2,L}$, but the model will be parsimonious and computable. In section five we evaluate by simulations, the effect that this efficiency lost in parameter estimation can have in the estimation of the number o factors.

Now, we explain how to use the moment function to consistently estimate the factors. Using just the moment conditions in $MC1o\ (\theta_{H_{2,L}} / L)$, the moment function can be defined as $E[m_t(\theta_L \mid L\ )] = vec[MC1o\ (\theta_{H_{2,L}} / L)]$ for given L. Let us rewrite the GMM minimization problem described in equation (8):

$$\min_{\theta_L} C_T(\theta_L \mid W_T(L), L) = TM_T(\theta_L)'[W_T(L)]^{-1} M_T(\theta_L), \tag{8}$$

Let $\hat{\theta}_L$ denote the GMM estimator minimizing $C_T(\theta_L \mid W_T(L), L)$; and use $\hat{\theta}_L^o$ to denote the GMM estimator minimizing $C_T(\theta_L \mid W_T(L_o), L_o)$ (i.e. at the true number of factors). We can estimate the model assuming different values of L (number of factors). Then we propose to estimate the number of factors by selecting the best specified model. The model section criterion method has been used extensively in determining the order of ARMA processes in time series analysis, specifically by Hannan and Quinn (1979), Hannan (1980,1981), Atkinson (1981), and Nishii (1988). Cragg and Donald (1997) use this method to estimate the ranks of matrices. Following these studies, we define the following criterion function:

$$MS_T(L) = C_T(\hat{\theta}_L \mid W_T(L), L) f(T)^{-1} - g(L) \tag{9}$$

where $f(T)$ and $g(L)$ are predefined functions of T (the number of observations) and L (the number of factors), respectively. Note that our approach is based $C_T(\hat{\theta}_L \mid W_T(L), L)$ statistic, which is simply the overidentifying restriction test statistic (Hansen, 1982). With appropriate choices of $f(T)$ and $g(L)$, a consistent estimate of L can be obtained by minimizing the criterion function $MS_T(L)$. There are many possible choices of $f(T)$ and $g(L)$. One commonly used criterion is:

Schwarz Criterion (BIC): $f(T) = \ln(T)$, and $g(T) = df$.

In BIC, $g(L)$ is simply the degrees of overidentifying restrictions in the moment conditions. With (9) and BIC, the following result establishes that the moment conditions on (7) can be used to estimate the number of factors.

Proposition 3: Let $\hat{L}$ be the minimizer of $MS_T(L)$ with BIC. Then, $\hat{L} \to_p L_o$.

The proof of Proposition 3 is the same as Proposition 2 in Ahn-Perez (2008) with different moment conditions. Observe that Proposition 3 holds even if the optimal GMM estimator is not used. This is a very important property since in the GMM literature, many studies have shown that optimal GMM estimators often have poor finite-sample properties, especially when data are autocorrelated or/and too many moment functions are used (see, for example, Altongi and Segal, 1996; Andersen and Sørensen, 1996; and Christiano and den Haan, 1996). One of the main reasons for this problem is that for such cases, the optimal weighting matrix, $[\tilde{W}_T(L_o)]^{-1}$ is poorly estimated. Note also that our approach is based $c_T(\tilde{b}_L^o \mid \tilde{W}_T(L_o), L_o)$ statistic, which is simply the overidentifying restriction test statistic (Hansen, 1982). In our Monte Carlo simulations (section 5) we compare the performance of the BIC criterion with the following criterions :

Akaike Information (AIC): $f(T) = 1$, and $g(T) = 2(P-L)(Q-L)$

Schwarz Criterion 2 (BIC2): $f(T) = \log(T)$, and $g(T) = (P-L)(Q-L)$.

Schwarz Criterion 3 (BIC3): $f(T) = \ln(T)$, and $g(T) = (P-L)(Q+1)$.

It can be shown that Proposition 2 holds using the modifications to the Schwarz Criterion labeled BIC2 and BIC3. AIC criterion is commonly used, but it leads to inconsistent estimation of the number of factors L_o

4. Close Form solution

In this section we explain how to compute the GMM minimization problem of equation (7) in close form. The moment conditions in $MC1o\ (\theta_{H_{2,L}} / L)$ are defined as:

$$MC1o\ (\theta_{H_{2,L}} / L) = offdiag\left[\Sigma_t^{11} - H_{2,L}'\Sigma_t^{12}\right] = 0$$

where *offdiag[Z]* implies the off-diagonal element of a matrix Z. The covariance restrictions implied by $MC1o\ (\theta_{H_{2,L}} / L)$ can be viewed as follows:

$$offdiag\begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1(N-L)} \\ \sigma_{21} & \ddots & & \vdots \\ \vdots & & \ddots & \\ \sigma_{(N-L)1} & \cdots & & \sigma_{(N-L)(N-L)} \end{bmatrix} -$$

-

$$offdiag\left(\begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1L} \\ h_{21} & & & h_{2L} \\ \vdots & & & \\ h_{(N-L)1} & \cdots & & h_{(N-L)L} \end{bmatrix} \times \begin{bmatrix} \sigma_{(N-L+1)1} & \sigma_{(N-L+1)2} & \cdots & \sigma_{(N-L+1)(N-L)} \\ \sigma_{(N-L+2)1} & \sigma_{(N-L+2)2} & \cdots & \sigma_{(N-L+2)(N-L)} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{N1} & & \cdots & \sigma_{N(N-L)} \end{bmatrix}\right) = 0$$

Where σ_{ij} and h_{ij} are the ((i,j) elements of the matrices Σ and $H_{2,L}$. From here is easy to see that these moment conditions are the same implied by a system of $(N-L)$ of equations with instruments defined as follows:

$$\begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_{(N-L)} \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1L} \\ h_{21} & & & h_{2L} \\ \vdots & & & \\ h_{(N-L)1} & \cdots & & h_{(N-L)L} \end{bmatrix} \times \begin{bmatrix} r_{(N-L+1)} \\ r_{(N-L+2)} \\ \vdots \\ r_N \end{bmatrix}$$

(10)

$$Instruments = \begin{bmatrix} r_2 & r_3 & \cdots & r_{(N-L)} \\ r_1 & r_3 & r_4 & r_{(N-L)} \\ \vdots & & \ddots & \\ r_1 & r_2 & & r_{(N-L)} \end{bmatrix}$$

Given Proposition 3, the system can be estimated using any non-optimal weighting matrix. This includes a block diagonal weighting matrix, or equivalently the estimation of system equation by equation[3]. We had assume that for notational simplicity that $E[r_{\bullet t}] = \alpha = 0$. From the system in equation (10) it is clear that this assumption can be removed by just including and intercept term for every equation in the system. At the same time a vector of ones can be included as instrument for every equation. This will not alter the results since we will add as many parameters as moment conditions.

Summarizing, the system of equations (10) should be estimated by instrumental variable estimation, equation by equation. The Hansen overidentifying restrictions test statistic $C_T(\hat{\theta}_L | W_T(L), L)$ can be computed for each equation and then added up. This procedure should be repeated for several values of L. The estimated number of factors will be the value of L that minimizes any consistent model selection criterion (BIC).

5. Small Sample Properties

In this section we report results of our Monte Carlo experiments developed to check the small sample properties of the test

5.1. Data Generation

The foundation of our Monte Carlo exercises is the following the three-factor model:

$$r_{it} = \alpha_i + \beta_{i1} f_{1t} + \beta_{i2} f_{2t} + \beta_{i3} f_{3t} + \varepsilon_{it} = \alpha_i + c_{1,it} + c_{2,it} + c_{3,it} + \varepsilon_{it}, \quad (15)$$

[3] The moment conditions $MC1o\ (\theta_{H_{2,L}} / L)$ include some redundant moment conditions. If the system is estimated equation by equation the estimation can be done without drooping them.

where the f_{kt} (k = 1, 2, 3) are the common factors of the model. Our benchmark model is the three-factor model of Fama and French (1993): EMR (excess market return), SMB, and HML.[4] We generate randomly β_{ik} and f_{kt} to match the moments of the Fama-French data. That is, we generate data such that the moments of $c_{k,it}$ match the counterparts from the data that Fama and French (1993) used. At the sample means of the estimated betas ($\bar{\beta}_1$, $\bar{\beta}_2$, and $\bar{\beta}_3$) for the 25 size and book-to-market portfolios, the estimated variances of the Fama-French common components are the following:

$$var(\bar{\beta}_1 \times EMR) = 21.72; var(\bar{\beta}_2 \times SMB) = 4.50; var(\bar{\beta}_3 \times HML) = 1.29.$$

Two types of idiosyncratic error components are used. First, we generate the errors which are cross-sectionally heteroskedastic, but not autocorrelated. Specifically, the errors are drawn from $N(0, \sigma_i^{FF})$, where the σ_i^{FF} are the variances of the residuals from the time-series regressions of (15) for each i. The values of σ_i^{FF} are between 1.21 and 3.78, with the average of 2.016. Thus, the variances of the first and second common components at the means of betas are more than twice as great as the average variance of the idiosyncratic components, while the variance of the third common component (1.29) is smaller. We define the signal to noise ratio (SNR) of a common component ($c_{k,it}$) as the ratio of the variances of the common component and the idiosyncratic error component. In our simulation, the SNRs are approximately 10.8 , 2.2, and 0.65 for common components 1, 2, and 3, respectively.

Second, we generate the error terms from a simple AR (1) process: $\varepsilon_{it} = \rho_i \varepsilon_{i,t-1} + v_{it}$. Using the residuals from the time-series regressions of (15), we estimate the parameters ρ_i and estimate $\mathrm{var}(v_{it})$ such that $\mathrm{var}(\varepsilon_{it}) = \sigma_i^{FF}$. The errors generated by this way are cross-sectional heteroskedastic and serially correlated over time.

---

[4] The Fama-French factors are constructed using the 6 value-weight portfolios formed on size and book-to-market. SMB (Small Minus Big) is the average return on the three small portfolios minus the average return on the three big portfolios. HML (High Minus Low) is the average return on the two value portfolios minus the average return on the two growth portfolios. EMR is the excess return on the market: the value-weight returns on all NYSE, AMEX, and NASDAQ stocks (from CRSP) minus the one-month Treasury bill rate (from Ibbotson Associates). See Fama and French (1993) for a complete description of the factor returns.

5.2. Estimation of the number of factors

Using data generated using three factors as defined in section 5.1, we estimate the number of factors using the model selection criterion methods. We perform our simulations with six different combinations of T and N: T = 500 and 1000*;* and N = 12, 15, and 25. The values of N and T are chosen to be close to the sample sizes most often used in the finance literature. For each combination, we consider two cases: the cases with autocorrelated (AR(1)) and serially uncorrelated idiosyncratic errors.

Given the generated data we estimate the system defined in equation (10) equation by equation, using the White (1980) estimator for the weighting matrix of each equation[5]. We report in Table 1 results for the estimation of the number of factors using the model selection criterion method. We use the four different penalty functions defined in section 3: *BIC1 BIC2 BIC3* and *AIC*. Results show that the methodology has good small sample properties. For the case of no autocorrelated errors (Panel A), *BIC1* criterion is the most accurate for $N= 12$ and *15,* estimating correctly three factors 90-95% of the time. As expected results improve as T increases. For the case of $N=25$ we estimate 3 factors around 80% of the time, which suggest that bigger sample size is required for data with $N>25$. The probability of over-estimation increases as N increases, from 5-8% when $N=12$ to 17-18% when $N=25$.

Panel B of Table 1 includes estimation results when errors are autocorrelated. Results are very similar to Panel A, so we can conclude that the method is robust to time series autocorrelation.

---

[5] We also experimented with other optimal and non optimal weighting matrixes. Results do not show significant differences. Equation by equation estimation is computationally faster and easier.

As explained in section 3 in order to avoid under-identification, our methodology requires restrictions on the parameter matrix, specifically $H_L = (-I_{N-L}\ , H_{2,L}\ )'$. The effect of this restriction in our system of equation (10) is the definition of which response variables are use as regressors in the estimation. To check the robustness of our results, we now investigate if estimation results differ if we change the parameter restriction. That is we want to check if result vary when we use different sub-set of response variables as regressors in the system of equations (10). To accomplish this objective we generate one set of portfolio returns, and then, we randomly sort them in 100 different ways. For each sorting, we estimate the number of factors. Results from this experiment with $T = 500$ are presented in Table 3. [6]

Results are similar to the ones reported in Tables 1 and 2. BIC1 criterion estimates the number of factors more accurately in all cases. This experiment confirms that the GMM estimation results could change depending on the parameter restriction used.. Changing in the parameter restrictions seems to be more important as N increases. The true number of factors (three) is always estimated with the highest frequency. Given these results we suggest that the estimation should be repeated using many randomly sorted data. Our experiments suggest that the number of factors can be more accurately estimated if 100 different sorting are used for estimation.[7] The estimator we propose is the "highest frequency" (HS) estimator, which is simply the number of factors most often estimated from 100 different sorts. In order to investigate the finite-sample properties of this estimator, we perform the following experiment: We generate 1,000 different sets of portfolio returns. For each data set, we estimate the number of factors sorting the data in 100 random ways and compute the HS estimate. The results (not reported) show that the HS estimates 3 factors 100% of the time, for N = 12, 15, and 25, whether or not idiosyncratic errors are autocorrelated or not.

The last part of our simulation exercises tries to evaluate the performance of our methods when a factor explains a very small proportion of the total variation of the

---

[6] In the following tables we do not report results for T=1000 in order to save space and since results improve as T increases. The main conclusions are not altered when T=1000. Results are available upon request.

[7] We also performed the same experiment using all possible sorts of portfolio returns into two groups. Results do differ significantly with the ones presented just using 100 random specifications of the groups.

response variable. We will call such factor a *weak* factor. In our simulation, the variance of the common component ($c_{k,it} = \beta_{ik} f_{kt}$) associated to a weak factor will be small compared with the variance of the idiosyncratic component. In order words, a weak factor is a factor with a low signal to noise ratio (SNR). As described in section 5.1, our data was generated using as a benchmark the three-factor model of Fama and French (1992), where the SNRs of three factors are 10.8, 2.2, and 0.65, respectively. To generate the data with one weak factor, we reduce the SNR of the second common component (SMB) and increase the one of the first common component (EMR). We do so because we wish to generate data such that the total variations in the response variables explained by the three factors and the variations in idiosyncratic errors remain constant. In this experiment, we reduce the SNR of the second common component to four different values: 1.0, 0.50, and 0.25. Note that SNR of the third common component is keep constant to a value smaller than 1. As we have done before, we generate one set of portfolio returns, and then, we random sort data in 100 different ways. For each sorting, we estimate the number of factors. Results are presented in Table 3.

For all SNR examined the highest estimated frequency is always tree factors, independently of the value of *N* and the penalty criterion used. As expected as SNR decreases we estimate 3 factors with smaller frequency. Also as *N* increases results deteriorate. With these results HS estimator appears to be a reliable estimator even in the presence of weak factors. In order to confirm this fact, we carry out the same experiment we conduct before: we generate one set of portfolio returns, and then, we randomly sort them in 100 different ways. For each sorting, we estimate the number of factors by the HS method. We repeat this experiment for 1,000 generated samples with three different SNRs of the second common component: 1.0, 0.5 and 0.25. Results are presented in Table 4. For SNR equal to 1 (Panel A) and equal to 0.5 (Panel B) results confirm our conjecture. HS estimates 3 factors with very high probability (97-100%) independently of the value *N* and the penalty criterion. For the case of SNR equal 0.25 *BIC2* seems to estimate more accurately three factors with probabilities of 96-99%. *BIC1* criterion, estimates three factors 67% of the time and underestimates the number of factors 33% of the time.

6. Conclusions

We propose in this paper a linear methodology to estimate the number of factors when one dimension of the data is small. The test is independent of the factors, since it is assumed that they are unobservable. Since GMM is used it is feasible to allow for time series autocorrelation and heteroskedasticity and cross sectional heteroskedasticity of disturbances. Monte Carlo simulations show that the test performs better when N is small and T is large. We prove that the methodology is computationally simple, since just requires the estimation of a system of equations by instrumental variables. An important feature is that the use of optimal weighting matrix ins not require , so the system can be estimated equation by equation. Several penalty criterions have been used

Our simulations show that the method is more precise using the *BIC1* penalty criterion. Since the method requires identification restrictions in the estimated parameters, we recommend to estimate the number of factors for different sorting of the respose variable. (100 or less random sorting appear to be enough). Simulation show that the number of factors most often estimated from different sorts is a very reliable estimator (i.e. the one with the highest frequency). *BIC1* penalty criterion is the most accurate in finite sample detecting factors with high signal to noise ration (strong factors). *BIC2* tends to capture weaker (SNR smaller than 0.25) factors, so results using both criterions should be compared.

APPENDIX

Lemma 1: *There exists a linear combination of* $MC2\,(\theta_{H_{2,L}}, \theta_{\Psi^{22}} / L)$ *such as we can divide the moment conditions in two subsets such as:*

$$MC2(\theta_{H_{2,L}}, \theta_{\Psi^{22}} / L) = \begin{bmatrix} MC2a\ (\theta_{H_{2,L}} / L) \\ MC2b\ (\theta_{H_{2,L}}, \theta_{\Psi^{22}} / L) \end{bmatrix}$$

where $MC2b\ (\theta_{H_{2,L}}, \theta_{\Psi^{22}} / L)$ *includes L moment conditions.*

Proof:

The moment conditions are $MC2(\theta_{H_{2,L}}, \theta_{\Psi^{22}} / L) = \left[\Sigma^{12} - H_{2,L}\,'(\Psi^{22} + \Sigma^{22}) \right] = 0$, that are written are:

$$\begin{bmatrix} \sigma_{1,(N-L+1)} & \sigma_{1,(N-L+2)} & \cdots & \sigma_{1,N} \\ \sigma_{2,(N-L+1)} & \ddots & & \vdots \\ \vdots & & \ddots & \\ \sigma_{(N-L),(N-L+1)} & \cdots & & \sigma_{(N-L),N} \end{bmatrix} -$$

$$\left(\begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1L} \\ h_{21} & & & h_{2L} \\ \vdots & & & \\ h_{(N-L),1} & h_{(N-L),2} & \cdots & h_{(N-L),L} \end{bmatrix} \times \begin{bmatrix} \sigma_{(N-L+1),(N-L+1)} - \psi^{22}_{N-L+1} & \sigma_{(N-L+1),(N-L+2)} & \cdots & \sigma_{(N-L+1),N} \\ \sigma_{(N-L+2),(N-L+1)} & \sigma_{(N-L+2),(N-L+2)} - \psi^{22}_{N-L+2} & \cdots & \sigma_{(N-L+2),N} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{N,(N-L+1)} & \cdots & & \sigma_{N,N} - \psi^{22}_{N} \end{bmatrix} \right) = 0$$

Let analyze the first column of moment conditions given by:

$$\begin{bmatrix} \sigma_{1,(N-L+1)} \\ \sigma_{2,(N-L+1)} \\ \vdots \\ \sigma_{(N-L),(N-L+1)} \end{bmatrix} - \begin{bmatrix} h_{1,1} \times (\sigma_{(N-L+1),(N-L+1)} - \psi^{22}_{N-L+1}) + h_{1,2} \times \sigma_{(N-L+2),(N-L+1)} + \cdots + h_{1,L} \times \sigma_{N,(N-L+1)} \\ h_{2,1} \times (\sigma_{(N-L+1),(N-L+1)} - \psi^{22}_{N-L+1}) + h_{2,2} \times \sigma_{(N-L+2),(N-L+1)} + \cdots + h_{2,L} \times \sigma_{N,(N-L+1)} \\ \vdots \\ h_{(N-L),1} \times (\sigma_{(N-L+1),(N-L+1)} - \psi^{22}_{N-L+1}) + h_{(N-L),2} \times \sigma_{(N-L+2),(N-L+1)} + \cdots + h_{(N-L),L} \times \sigma_{N,(N-L+1)} \end{bmatrix} = 0$$

Let $h_{1,1}, h_{2,1} \cdots h_{(N-L),1} > 0$ [8] and solve for the terms involving the parameters $\psi^{22}_{i,j}$ as follows:

$$\begin{bmatrix} \dfrac{\sigma_{1,(N-L+1)} - h_{1,2} \times \sigma_{(N-L+2),(N-L+1)} - \cdots - h_{1,L} \times \sigma_{N,(N-L+1)}}{h_{1,1}} \\ \dfrac{\sigma_{2,(N-L+1)} - h_{2,2} \times \sigma_{(N-L+2),(N-L+1)} - \cdots - h_{2,L} \times \sigma_{N,(N-L+1)}}{h_{2,1}} \\ \vdots \\ \dfrac{\sigma_{(N-L),(N-L+1)} - h_{(N-L),2} \times \sigma_{(N-L+2),(N-L+1)} - \cdots - h_{(N-L),L} \times \sigma_{N,(N-L+1)}}{h_{(N-L),1}} \end{bmatrix} - \begin{bmatrix} \sigma_{(N-L+1),(N-L+1)} - \psi^{22}_{N-L+1} \\ \sigma_{(N-L+1),(N-L+1)} - \psi^{22}_{N-L+1} \\ \vdots \\ \sigma_{(N-L+1),(N-L+1)} - \psi^{22}_{N-L+1} \end{bmatrix} = 0$$

Clearly the $(N-L)$ previous moment conditions can be written as: one moment condition involving $\sigma_{(N-L+1),(N-L+1)} - \psi^{22}_{N-L+1}$ and $(N-L-1)$ moment conditions free of the parameter ψ^{22}_{N-L+1}. Similar procedure can be performed for all L columns of the matrix $MC2b\ (\theta_{H_{2,L}}, \theta_{\Psi^{22}} / L)$ and this will lead to the result.

Proposition 1: Now, $MCA\left(\theta_{H_{2,L}} / L\right)$ involves $(N-L)(N-L-1)$ linear moment conditions in $MC1o\ (\theta_{H_{2,L}} / L)$ and $L \times (N-L-1)$ non linear moment conditions in $MC2a\ (\theta_{H_{2,L}} / L)$, and the estimation of $(N-L) \times L$ parameters in $H_{2,L}$. Then the all parameters in $H_{2,L}$ can be identified by $MCA\left(\theta_{H_{2,L}} / L\right)$. Also $MC1d\ (\theta_{H_{2,L}}, \theta_{\Psi^{11}} / L)$ includes $(N-L)$ moment condition and $(N-L)$ parameters is $\theta_{\Psi^{11}}$. Finally

---

[8] We assume this without lost of generality, since if these coefficients are equal to zero the non linear moment conditions will disappear. If no moment conditions are non-linear Proposition 2 is not needed neither Lemma 1.

$MC2b\ (\theta_{H_{2,L}}, \theta_{\Psi^{22}} / L)$ includes L moment conditions and L parameters. Then direct application of Theorem 1 of Ahn, Schmidt (1992) gives the result.

<u>*Proposition 2*</u> *Let analyze the form of* $MC1o\ (\theta_{H_2,L} / L)$.

$$MC1o\ (\theta_{H_2,L} / L) = offdiag \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1(N-L)} \\ \sigma_{21} & \ddots & & \vdots \\ \vdots & & \ddots & \\ \sigma_{(N-L)1} & \cdots & & \sigma_{(N-L)(N-L)} \end{bmatrix} -$$

$$offdiag\left(\begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1L} \\ h_{21} & & & h_{2L} \\ \vdots & & & \\ h_{(N-L)1} & \cdots & & h_{(N-L)L} \end{bmatrix} \times \begin{bmatrix} \sigma_{(N-L+1)1} & \sigma_{(N-L+1)2} & \cdots & \sigma_{(N-L+1)(N-L)} \\ \sigma_{(N-L+2)1} & \sigma_{(N-L+2)2} & \cdots & \sigma_{(N-L+2)(N-L)} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{N1} & \cdots & & \sigma_{N(N-L)} \end{bmatrix} \right) = 0$$

By observing these moment conditions ity is clear that they do not include the $(L^2 - L)/2$ off diagonal elements of the square matrix Σ^{22}:

$$\Sigma^{22} = \begin{bmatrix} \sigma_{(N-L),(N-L)} & \sigma_{(N-L),(N-L+1)} & \cdots & \sigma_{(N-L),N} \\ \sigma_{(N-L+1),(N-L)} & \ddots & & \sigma_{(N-L+1),N} \\ \vdots & & & \\ \sigma_{N,(N-L)} & & \sigma_{N,(N-1)} & \sigma_{N,N} \end{bmatrix}$$

Since $MC2a\ (\theta_{H_{2,L}} / L)$ involves the off-diagonal elements in Σ^{22}, not linear combination of $MC1o\ (\theta_{H_2,L} / L)$ can lead to the moment conditions in $MC2a\ (\theta_{H_{2,L}} / L)$.

REFERENCES

Ahn, Seung C., Young H. Lee, and Peter Schmidt, 2007a, Stochastic frontier models with multiple time-varying individual effects, *Journal of Productivity Analysis* 27 (1), 1-12.

Ahn, Seung C., Young H. Lee, and Peter Schmidt, 2007b, Panel data models with multiple time-varying individual effects, mimeo, Arizona State University.

Ahn, Seung C., Stephan Dieckmann and Marcos F. Perez, 2007, Estimating the Common Factor in Credit Spreads, mimeo, Arizona State University.

Andersen, Torben G., and Bent E. Sørensen, 1996, GMM estimation of a stochastic volatility model: A Monte Carlo study, *Journal of Business & Economic Statistics* 14, 328-352

Andrews, Donald W.K., 1991, Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* 59, 817-858.

Atkinson, Anthony .C., 1981 Likelihood ratios, posterior odds and information criteria, *Journal of Econometrics* 16, 15-20

Bai, Jusan and Serena Ng, 2002, Determining the number of factors in approximate factor models, *Econometrica* 70, 191-221.

Bai, Jusan, 2003, Inferential theory for factor models of large dimensions, *Econometrica* 71, 135-171.

Bekaert, Geert , Robert J. Hodrick and Xiaoyan Zhang 2005, International Stock Return Comovements. *Johnson School Research paper series No. 35-06*

Black, Fisher, Michael C. Jensen, and Myron S. Scholes, 1972, The capital asset pricing model: Some empirical tests, in Michael C. Jensen, ed.: *Studies in the Theory of Capital Markets*, Praeger, New York.

Brown, Stephen J. and Mark I. Weinstein, 1983, A new approach to testing asset pricing models: The bilinear paradigm, *Journal of Finance* 38, 711-743.

Brown, Stephan J., 1989, The number of factors in security returns, *Journal of Finance* 44(5), 1247-1262.

Campbell, John Y., Andrew W. Lo, and A. Craig MacKinlay, 1997, *The Econometrics of Financial Markets*, Princeton University Press, Princeton, New Jersey.

Chamberlain, Gary and Michael Rothschild, 1983, Arbitrage, factor Structure, and mean variance analysis on large asset markets, *Econometrica* 51, 1281-1304.

Chen, Nai Fu, Richard Roll and Stephen Ross, 1986, Economic forces and the stock market, *Journal of Business* 59, 368-403.

Connor, Gregory, 1984, A unified beta pricing theory, *Journal of Economic Theory* 34, 13-31.

Connor, Gregory and Robert A. Korajczyk, 1988, Risk and return in an equilibrium APT: Application of a new test methodology, *Journal of Financial Economics* 21, 255-289.

Connor, Gregory and Robert A. Korajczyk, 1993, A test for the number of factors in an approximate factor model. *Journal of Finance* 48, 1263-1291.

Cragg John and Stephen Donald, 1996, On the asymptotic properties of LDU based test of the rank of a matrix, *Journal of the American Statistical Association* 91, 1301-1309.

Cragg John and Stephen Donald, 1997, Inferring the rank of a matrix, *Journal of Econometrics* 76, 223-250

Donald Stephen G., Fortuna Natercia, and Vladas Pipiras, 2005, On the rank estimation in symmetric matrices: the case of indefinite estimators, CEMPRE Working Paper

Fama, Eugene F., and James D. MacBeth, 1973, Risk, return, and equilibrium: empirical tests, *Journal of Political Economy* 71, 607-636.

Fama, Eugene F., and Kenneth R. French, 1993, Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics* 33, 3-56.

Ferson, Wayne, and Stephen R. Foerster, 1994, Finite sample properties of the generalized methods of moments tests of conditional asset pricing models, *Journal of Financial Economics* 36, 29-56.

Gill, Len, and Arthur Lewbel, 1992, Testing the rank and definiteness of estimated matrices with application to factor, state-space and ARMA models, *Journal of the American Statistical Association*, 87, 766-776.

Grinblatt, Mark and Sheridan Titman, 1985, Approximate factor structures: Interpretations and implications for empirical tests, *Journal of Finance* 40, 1367-1373.

Gorman, William M., 1981, Some Engel curves, in *Essay in the Theory and Measurement of Consumer Behavior in Honor of Sir Richard Stone*, ed. by A. Deaton, New York: Cambridge University Press.

Gregory, Allan W., and Allen C. Head, 1999, Common and country-specific fluctuations in productivity, investment, and the current account, *Journal of Monetary Economics* 44, 423.

Hannan, Edward J., 1980, The estimation of the order of an ARMA process, *Annals of Statistics* 8, 1071-1081.

Hannan, Edward. J., 1981, Estimating the dimension of linear system, *Journal of Multivariate analysis* 11, 459-473.

Hannan, Edward J., and B. G. Quinn, 1979, The determinants of the order of an autoregression, *Journal of the Royal Statistical Society* 41, 190-195.

Hansen, Lars P., 1982, Large sample properties of generalized method of moments estimators, *Econometrica* 50, 1029-1054.

Hinkley, David V., 1977, Jackknifing in unbalance situations, *Technometrics* 19, 285-292.

Jones, Christopher S., 2001, Extracting factors from heteroskedastic returns, *Journal of Financial Economics* 62, 293-325.

Jöreskog, Karl G., 1967, Some contributions to maximum likelihood factor analysis, *Psychometrika* 34, 183-202.

Kan, Raymond, and Chu Zhang, 1999, Two-pass tests of asset pricing models with useless factors, *Journal of Finance* 54, 204-235.

Lehmann, Bruce N., and David M. Modest, 1988, The empirical foundations of the arbitrage pricing theory, *Journal of Financial Economics* 21, 213-254.

Lewbel, Arthur, 1991, The rank of demand system: Theory and nonparametric estimation, *Econometrica* 59, 711-730.

Newey, Whitney K., and Kenneth D. West, 1987, A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix, *Econometrica* 55 (3), 703-708

Newey, Whitney K., and Kenneth D. West, 1994. Automatic Lag Selection in Covariance Matrix Estimation *The Review of Economic Studies*, 61, 631-653

Nishii, Ryuie, 1988, Maximum likelihood principle and model selection when the true model is unspecified, *Journal of Multivariate Analysis*, 27, 392-403

Roll, Richard W. and Stephan A. Ross, 1980, An empirical investigation of the arbitrage pricing theory, *Journal of Finance* 35, 1073-1103.

Ross, Stephen A., 1976, The arbitrage theory of capital asset pricing, *Journal of Economic Theory* 13, 341-360.

Stock, James H., and Mark Watson, 2005, Implication of dynamic factor models for VAR analysis, NBER working paper 11467.

White, Halbert L., 1980, A heteroskedasticity-consistent covariance matrix estimator and direct test for heteroskedasticity, *Econometrica* 48, 817-838

White, Halbert L., 1999, *Asymptotic Theory for Econometricians*: Academic Press, San Diego, California.

TABLE 1

Estimating the number of Factors by the Model Selection Criterion Method

The model selection criterion method is used to estimate the number of factors for the data generated with three factors (*Lo = 3*). The total number of simulations is 1,000. The system is estimated equation by equation using White estimator as weighting matrix. We report results for the information criterions: BIC1 BIC2 BIC3 and AIC as described in the paper

Panel A: NO AUTOCORRELATION

N	# Factors	T=500				T=1000			
		BIC1	BIC2	BIC3	AIC	BIC1	BIC2	BIC3	AIC
12	≤ 2	0.3	0.1	0.1	0.1	0.2	0.2	0.2	0.0
	3	91.5	83.0	89.5	73.7	94.3	86.6	93.3	75.2
	4	8.0	16.3	10.1	23.9	5.5	12.7	6.5	23.0
	≥ 5	0.2	0.6	0.3	2.3	0.0	0.5	0.0	1.8
	Average	3.08	3.17	3.11	3.28	3.05	3.14	3.06	3.27
15	≤ 2	0.6	0.0	0.2	0.0	0.2	0.0	0.2	0.0
	3	91.2	76.8	88.5	66.1	91.8	80.8	88.1	70.0
	4	8.1	20.6	10.9	28.3	7.3	17.3	11.0	25.1
	≥ 5	0.1	2.6	0.4	5.6	0.7	1.9	0.7	4.9
	Average	3.08	3.26	3.12	3.40	3.09	3.21	3.12	3.35
25	≤ 2	2.6	0.0	0.5	0.0	2.9	2.9	0.0	0.0
	3	79.3	53.6	73.0	42.5	81.0	63.3	74.8	47.9
	4	17.0	34.6	23.0	37.8	18.1	31.5	23.1	37.7
	≥ 5	1.1	11.8	3.5	19.7	0.9	5.2	2.1	14.4
	Average	3.17	3.59	3.30	3.82	3.20	3.42	3.27	3.70

Panel B: AUTOCORRELATION

N	# Factors	T=500				T=1000			
		BIC1	BIC2	BIC3	AIC	BIC1	BIC2	BIC3	AIC
12	≤ 2	0.3	0.1	0.1	0.1	0.2	0.2	0.2	0.0
	3	91.1	83.3	89.5	73.6	94.4	86.5	93.5	74.8
	4	8.4	16.1	10.1	24.2	5.4	12.9	6.3	23.3
	≥ 5	0.2	0.5	0.3	2.1	0.0	0.4	0.0	1.9
	Average	3.09	3.17	3.11	3.28	3.05	3.14	3.06	3.27
15	≤ 2	0.6	0.0	0.2	0.0	0.2	0.0	0.2	0.0
	3	91.4	77.0	88.4	66.1	90.9	80.8	88.0	69.5
	4	7.9	20.6	11.0	28.1	8.2	17.3	11.1	25.7
	≥ 5	0.1	2.4	0.4	5.8	0.7	1.9	0.7	4.8
	Average	3.08	3.25	3.12	3.40	3.09	3.21	3.12	3.35
25	≤ 2	2.2	0.0	0.6	0.0	0.0	0.0	0.0	0.0
	3	79.3	55.2	72.0	42.4	81.3	63.3	75.3	48.5
	4	17.2	33.3	24.0	37.7	17.8	32.0	22.9	37.0
	≥ 5	1.3	11.5	3.4	19.5	0.9	4.2	1.8	14.1
	Average	3.18	3.56	3.30	3.76	3.20	3.42	3.27	3.70

TABLE 2

Effects of Random Sorting

A *single* data set is generated with T=500 from a three-factor model. The number of factors is estimated by model selection criterion method. This estimation is conducted for 100 randomly chosen sorts of the response variables. The system is estimated equation by equation using White estimator as weighting matrix.

N	# Factors	NO AUTOCORRELATION				AUTOCORRELATION			
		BIC1	BIC2	BIC3	AIC	BIC1	BIC2	BIC3	AIC
12	≤ 2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	3	97.7	95.3	97.2	91.4	93.0	87.0	91.0	85.0
	4	2.3	4.7	2.8	8.6	7.0	13.0	9.0	14.0
	≥ 5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
	Average	3.02	3.05	3.03	3.09	3.07	3.13	3.09	3.16
15	2	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0
	3	90.3	81.6	89.3	71.4	91.0	79.0	91.0	74.0
	4	9.5	17.5	10.3	25.5	7.0	18.0	7.0	21.0
	≥ 5	0.2	0.9	0.4	3.1	0.0	3.0	2.0	5.0
	Average	3.10	3.19	3.11	3.32	3.05	3.24	3.11	3.31
25	≤ 2	2.9	2.7	0.0	0.0	0.0	0.0	0.0	0.0
	3	84.0	68.0	76.0	56.0	81.0	65.0	76.0	50.0
	4	15.0	24.0	21.0	33.0	18.0	25.0	20.0	35.0
	≥ 5	1.0	8.0	3.0	11.0	1.0	9.0	4.0	15.0
	Average	3.17	3.41	3.27	3.56	3.20	3.40	3.28	3.65

TABLE 3

Effects of Weak Factor in One Random Sample

A single data set with T = 500 is generated from a three-factor model using different signal to noise ratios for the second common component. The variances of the response variables explained by three factors are held constant. The number of factors is estimated by the model selection criterion method. The method is applied to 100 randomly chosen sorts of response variables.

PANEL A: SIGNAL TO NOISE FOR COMMON COMPONENT 2 = 1.00

# Factors	N=12				N=15				N=25			
	BIC1	BIC2	BIC3	AIC	BIC1	BIC2	BIC3	AIC	BIC1	BIC2	BIC3	AIC
≤ 1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	93.0	87.0	93.0	85.0	91.0	80.0	91.0	73.0	81.0	64.0	76.0	50.0
4	7.0	13.0	7.0	14.0	6.0	17.0	7.0	21.0	18.0	26.0	20.0	35.0
≥ 5	0.0	0.0	0.0	1.0	1.0	3.0	2.0	6.0	1.0	10.0	4.0	15.0
Average	3.07	3.13	3.07	3.16	3.06	3.23	3.11	3.33	3.20	3.47	3.28	3.67

PANEL B: SIGNAL TO NOISE FOR COMMON COMPONENT 2 = 0.5

# Factors	N=12				N=15				N=25			
	BIC1	BIC2	BIC3	AIC	BIC1	BIC2	BIC3	AIC	BIC1	BIC2	BIC3	AIC
≤ 1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	1.0	0.0	0.0	0.0	3.0	0.0	2.0	0.0	2.0	0.0	1.0	0.0
3	91.0	86.0	90.0	81.0	91.0	77.0	90.0	66.0	80.0	61.0	76.0	47.0
4	8.0	12.0	10.0	17.0	6.0	18.0	7.0	28.0	17.0	29.0	19.0	35.0
≥ 5	0.0	2.0	0.0	2.0	0.0	5.0	1.0	6.0	1.0	10.0	4.0	18.0
Average	3.07	3.16	3.10	3.21	3.03	3.28	3.07	3.40	3.17	3.50	3.26	3.76

PANEL C: SIGNAL TO NOISE FOR COMMON COMPONENT 2 = 0.25

# Factors	N=12				N=15				N=25			
	BIC1	BIC2	BIC3	AIC	BIC1	BIC2	BIC3	AIC	BIC1	BIC2	BIC3	AIC
≤ 1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	25.0	1.0	12.0	0.0	13.0	1.0	7.0	1.0	31.0	1.0	10.0	0.0
3	72.0	80.0	81.0	77.0	85.0	74.0	86.0	60.0	59.0	56.0	70.0	40.0
4	3.0	18.0	7.0	21.0	2.0	20.0	7.0	31.0	9.0	28.0	17.0	35.0
≥ 5	0.0	1.0	0.0	2.0	0.0	5.0	0.0	8.0	1.0	15.0	3.0	25.0
Average	2.78	3.19	2.95	3.25	2.89	3.29	3.00	3.46	2.80	3.59	3.13	3.96

TABLE 4

Effects of Weak Factors in 1,000 Random Samples

1,000 random samples are generated from a three-factor model with different signal to noise ratios of the second common component. Each sample contains 500 time series observations. For each data set, the variances of the response variables explained by three factors are held constant. The number of factors is estimated by applying the model selection criterion to 100 randomly chosen sorts of response variables. The estimated number of factors for each sample is the number estimated the most frequently.

PANEL A: SIGNAL TO NOISE FOR COMMON COMPONENT 2 = 1.00

# Factors	N=12				N=15				N=25			
	BIC1	BIC2	BIC3	AIC	BIC1	BIC2	BIC3	AIC	BIC1	BIC2	BIC3	AIC
≤ 1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	100.0	100.0	100.0	92.0	100.0	100.0	100.0	97.0	100.0	100.0	100.0	89.0
4	0.0	0.0	0.0	8.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	11.0
more	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Average	3.0	3.0	3.0	3.1	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.1

PANEL B: SIGNAL TO NOISE FOR COMMON COMPONENT 2 = 0.5

# Factors	N=12				N=15				N=25			
	BIC1	BIC2	BIC3	AIC	BIC1	BIC2	BIC3	AIC	BIC1	BIC2	BIC3	AIC
≤ 1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	3.0	0.0	2.0	0.0	3.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
3	97.0	100.0	98.0	87.0	97.0	100.0	99.0	94.0	100.0	98.0	100.0	80.0
4	0.0	0.0	0.0	13.0	0.0	0.0	0.0	6.0	0.0	2.0	0.0	20.0
more	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Average	3.0	3.0	3.0	3.1	3.0	3.0	3.0	3.1	3.0	3.0	3.0	3.2

PANEL C: SIGNAL TO NOISE FOR COMMON COMPONENT 2 = 0.25

# Factors	N=12				N=15				N=25			
	BIC1	BIC2	BIC3	AIC	BIC1	BIC2	BIC3	AIC	BIC1	BIC2	BIC3	AIC
≤ 1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	33.0	4.0	13.0	2.0	31.0	1.0	10.0	0.0	30.0	0.0	0.0	0.0
3	67.0	96.0	87.0	86.0	69.0	99.0	90.0	91.0	70.0	98.0	100.0	68.0
4	0.0	0.0	0.0	12.0	0.0	0.0	0.0	9.0	0.0	2.0	0.0	32.0
more	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Average	2.7	3.0	2.9	3.1	2.7	3.0	2.9	3.1	2.7	3.0	3.0	3.3