Work in Progress
1st June 2003

ASYMPTOTIC PRINCIPAL COMPONENTS ESTIMATION OF LARGE FACTOR MODELS

CHRIS HEATON† AND VICTOR SOLO‡

†*Department of Economics, Macquarie University and School of Economics, University of New South Wales, E-mail: cheaton@efs.mq.edu.au*
‡*School of Electrical Engineering and Telecommunications, University of New South Wales, E-mail: vsolo@unsw.edu.au*

ABSTRACT: There has been much recent interest in forecasting based on factor analysis models for large numbers of observable variables (p) and large numbers of observations (T). Some nice asymptotic results have been produced showing that under certain conditions, as $(p,T) \to (\infty, \infty)$ principal components analysis can be used to carry out the forecasting, thereby avoiding the need to fit a full factor analysis model. However, the question of how large p needs to be in order for the asymptotic theory to provide an adequate approximation in practice is open. In this paper we develop probability bounds for the forecast accuracy of principal component forecasts for stationary processes in terms of an empirically determinable noise to signal ratio. We develop a hypothesis test for this bound for which asymptotics in T hold even with p large. We apply this test to US macrodata.

Correspondence to Chris Heaton, Department of Economics, Macquarie University, North Ryde NSW 2109, AUSTRALIA; cheaton@efs.mq.edu.au.

Introduction

In this paper we consider forecasting with a factor model of the form

(1) $x_t = Bf_t + n_t$

(2) $y_t = \beta' f_t + v_t \qquad t=1,..,T$

where x_t is a $p\times1$ vector of observable variables which will be used as predictors of y_t, a scalar observable variable. f_t is a $k\times1$ vector of zero-mean unobservable factors, n_t is a $p\times1$ vector of zero-mean unobservable 'noise' terms and v_t is a zero-mean scalar noise term. B and β are a $p\times k$ coefficient matrix and a $k\times1$ coefficient vector respectively, which are both unobserved. We will assume throughout that $E(f_t f_t') = I_k$. A factor model of this type might be considered when the number of predictor variables is relatively large, making more traditional forecasting models such as VAR and regression less attractive or infeasible.

It is well known that when f_t and n_t in equation (1) are serially uncorrelated and Gaussian, and $\Psi = E(n_t n_t')$ is diagonal, that B is identified only up to an orthogonal transformation under some rank conditions. However, this lack of complete identifiability is not a problem in forecasting applications. Heaton and Solo (2002) show that when f_t and n_t are stationary Gaussian processes and n_t has a diagonal spectral density matrix, that B is identified up to sign changes and reordering of the factors under fairly general conditions. When the covariance matrix of n_t, Ψ is not necessarily diagonal, identification is more problematic. Following Chamberlain (1983) and Chamberlain and Rothchild (1983) such models are referred to as approximate factor models[1]. For any approximate factor model given by equation (1) if we define $B^* = \mu B$ and $n_t^* = n_t + (1-\mu)Bf_t$, then $x_t = B^* f_t + n_t^*$ is an observationally equivalent approximate factor representation of x_t. Forni, Hallin, Lippi and Reichlin (2000) have shown that under certain conditions which force only the first k eigenvalues of Ω to diverge as $p \to \infty$, the common component Bf_t is asymptotically identified. However, for finite p, identification of the parameters is clearly a problem.

There are at least three approaches that one might take to parameter estimation in the above model. Firstly, one might assume a strict factor structure and use traditional static factor analysis maximum likelihood estimation, as described by Lawley and Maxwell (1971) and Jöreskog (1967), to estimate B and Ψ under Gaussian assumptions. Factor estimates could then be produced and used to estimate β. In practice, MLE procedures for static factor analysis are often plagued by improper solutions,[2] and a degree of finessing on the part of the user is often required to find a set of estimates. However, using combinations of algorithms, the estimation of models with well over one hundred variables is computationally feasible.

[1] The more traditional factor models with a diagonal error covariance are referred to as strict factor models.

[2] 'Improper solutions' refers to iterations in which the estimate of an element of Ψ converges (slowly) towards zero. Improper solutions in factor analysis are often referred to as Heywood cases.

A more satisfactory model for forecasting might be constructed by specifying the factors f_t and the noise n_t as finite-order autoregressions, maintaining the assumption of strict factor structure. An estimation procedure for such a model, based on the Kalman Filter, was detailed by Watson and Engle (1983) and has been used widely. Camba-Mendez, Kapetanios and Smith and Weale (2001) used such a procedure to estimate factors from sets of leading indicators. These were then incorporated in VAR models to forecast European GDP growth. Unfortunately, the computational cost of the Watson and Engle algorithm is high and increases in p^3 making the estimation of models with a large number of variables infeasible. However, Heaton and Solo (2002) have derived an accelerated EM algorithm which is much more efficient and can handle much larger data sets. Also, Kapetanios (2002) has derived a non-iterative subspace algorithm which provides consistent estimates of factor model parameters. Accordingly, the somewhat severe constraints on model size imposed by the Watson and Engle algorithm are no longer binding.

The third approach to estimation is to replace the unobservable factors in equation (2) with the sample principal components of x_t and to estimate β by ordinary least squares. The idea that principal components and factors may be closely related is an old one but only recently has some substance been put behind this suggestion. Schneeweiss and Mathes (1995) and Schneeweiss (1997) give easily interpretable conditions under which population factors in a static exact factor model are close to population principal components in some sense. Stock and Watson (1998) prove under quite general conditions that the forecasts derived from regression on sample principal components converge in probability to the mean-square efficient forecasts that could be computed if f_t was observable as $p \to \infty$ and $T \to \infty$ jointly, with a restriction on the joint growth rates of p and T. In particular, their result holds for approximate factor models. This may be important since the assumption that the elements of n_t are mutually independent becomes less believable as p gets large. Bai and Ng (2002) prove consistent estimation of a transformation of the factor scores under a less restrictive joint growth rate, and derive model selection criteria for choosing the number of factors. Forni, Hallin, Lippi and Reichlin (2000) provide a consistency theorem for a dynamic version of the model for which the factors are estimated by the principal components of the estimated spectral density matrix. Stock and Watson (1998) also carried out an extensive forecasting simulation using US macrodata, which suggested that the principal components forecasting technique performs well relative to small dimension regression and VAR.

The obvious attractions of the principal components technique to applied forecasters are the relative ease with which the forecasts can be produced, the fact that an optimality result exists to justify their use, and the fact that this optimality result applies to the more general approximate factor model. All that is required to compute forecasts is the computation of eigenvectors and the execution of a least squares regression. Since such procedures are standard in most econometric and statistical courses and software packages, specialist software and expertise are not required. However, one must be concerned whether the asymptotic principal components theory holds in any particular case. Of concern to us is the interpretation of the theoretical requirement that $p \to \infty$. Of course, this is a purely abstract concept. In practice, p is a variable over which the analyst has limited control, and one must wonder whether p is 'large enough' for the asymptotic principal components theories to provide a satisfactory result in any particular application. Our aim then is to try to

answer the question of when the 'infinite-p' results will provide a satisfactory approximation in practice. It turns out that what seems to matter is not so much that p be large but rather that the magnitude of the noise to signal ratio be small, where the noise to signal ratio is defined as $\rho = \frac{\sigma^2}{\lambda_k}$ where λ_k is the k^{th} eigenvalue of Ω, and σ^2 is the largest eigenvalue of Ψ. Schneeweiss (1997) has shown that the population principal components of $BB' + \Psi$ converge to B as a slightly different noise to signal ratio approaches zero[3]. We note that Bai and Ng (2002), Forni, Hallin, Lippi and Reichlin (2000) and Stock and Watson (1998) all make assumptions that bound the eigenvalues of Ψ but require the eigenvalues of BB' to diverge, causing ρ to converge to zero as $p \to \infty$. However, we can demonstrate with simple examples that there is no finite value of p which guarantees a small noise to signal ratio, and that a finite increase in a finite value of p does not guarantee a reduction in the noise to signal ratio. Consider first a factor model in which $\Psi = \sigma^2 I_p$ and denote the j^{th} eigenvalue of BB' as d_j. Then $\rho = \frac{\sigma^2}{d_k + \sigma^2}$. Thus, we can write down the noise to signal ratio without needing to know p. In other words, it is always possible to write down a (somewhat contrived) factor model with any finite value of p and a noise to signal ratio close to unity[4]. The second example is a numerical one. Consider the following two models.

Model 1:

$$B = \begin{pmatrix} -0.5975 \\ -1.2106 \\ -0.7027 \end{pmatrix}, \ \Psi = \begin{pmatrix} 0.8176 & 0 & 0 \\ 0 & 0.74558 & 0 \\ 0 & 0 & 0.4622 \end{pmatrix}$$

Model 2:

$$B = \begin{pmatrix} -0.5975 \\ -1.2106 \\ -0.7027 \\ 0.3564 \end{pmatrix}, \Psi = \begin{pmatrix} 0.8176 & 0 & 0 & 0 \\ 0 & 0.74558 & 0 & 0 \\ 0 & 0 & 0.4622 & 0 \\ 0 & 0 & 0 & 0.9514 \end{pmatrix}$$

Note that Model 2 is simply Model 1 with an extra variable added. It is easy to compute the noise to signal ratio for Model 1 to be 0.2708 and for Model 2 to be 0.3011. Thus, in this case, increasing the number of variables in the model has raised the noise to signal ratio, rather than reducing it as we might have hoped.

The objective of our research project is to derive a theory for factor model forecasting for finite p and to understand when the asymptotic principal components approximation is likely to be useful in practice. In what follows we define a

[3] More precisely the population principal components multiplied by a sign matrix converge to B scaled by the eigenvalues of BB'.
[4] It will become clear in the next section that a noise to signal ratio of 1 is very poor.

theoretical 'ideal' forecast based on unidentified factors, and derive a probability bound which links the ideal forecast to an identified forecast computed from principal components. The size of this bound is related to the magnitude of the noise to signal ratio. We discuss estimation of the noise to signal ratio and testing hypotheses about its magnitude, and conclude with an empirical example using the Stock and Watson (1998) data.

A Small Noise Theory of Factor Analysis

For the model given by equations (1) and (2) let $E(x_t x_t') = \Omega = BB' + \Psi$ and $E(v_t^2) = \sigma_v^2$. Let D be the k×k diagonal matrix of ordered eigenvalues of BB' and U be the corresponding p×k matrix of eigenvectors; let Λ be the k×k diagonal matrix of the first k ordered eigenvalues of Ω and Q be the corresponding p×k matrix of eigenvectors. $\Lambda_\perp$ and $Q_\perp$ contain the remaining eigenvalues and eigenvectors. We therefore have the spectral decompositions

$$BB' = UDU' \text{ and } \Omega = Q\Lambda Q' + Q_\perp \Lambda_\perp Q_\perp'.$$

We define σ^2 to be the maximum eigenvalue of Ψ. We then define the noise to signal ratio to be

$$\rho = \frac{\sigma^2}{\lambda_k}$$

We define the ideal estimator of the factor to be

$$f_{Qt} = \Lambda^{-\frac{1}{2}} Q' x_t,$$

the ideal regression estimator to be

$$\beta_Q = \frac{1}{T}\sum_{t=1}^{T} f_{Qt} y_t$$

and the ideal forecast to be

$$s_{Qh} = \beta_Q' f_{Qh}.$$

Replacing population parameters with sample estimates, we define the empirical factor estimate, the empirical regression estimator and the empirical forecast to be

$$\hat{f}_{Qt} = \hat{\Lambda}^{-\frac{1}{2}} \hat{Q}' x_t, \qquad \hat{\beta}_Q = \frac{1}{T}\sum_{t=1}^{T} \hat{f}_{Qt} y_t, \qquad \text{and} \qquad \hat{s}_{Qh} = \hat{\beta}_Q' \hat{f}_{Qh}$$

We wish to investigate the properties of the ideal and empirical forecasts for finite p and a given noise to signal ratio. It is also of interest to investigate the asymptotics of these quantities as:

(I) $T \to \infty$, ρ fixed;

(II) $\left(\frac{1}{T}, \rho\right) \to (0,0)$.

By the latter of these we mean the following.

Definition: We say $Z_{T,\rho} \xrightarrow{p} 0$ as $\left(\frac{1}{T}, \rho\right) \to (0,0)$ if $\forall \varepsilon > 0,\ 0 < \delta \le 1,$

$\exists T_0 = T(\delta, \varepsilon) \in \mathbb{N},\ \rho_0 = \rho(\delta, \varepsilon) \in \mathbb{N}$ such that $T > T_0$ and $\rho < \rho_0$

$\Rightarrow P\left\{\left\|Z_{T,\rho}\right\| > \varepsilon\right\} < \delta$.

The development of these asymptotics would help to link the small-p theory to the infinite-p theory that already exists. This complete theory is still under development.

By the Markov inequality, $P\left\{\left\|Z_{T,\rho}\right\| > \varepsilon\right\} < \frac{E\left\|Z_{T,\rho}\right\|}{\varepsilon}$. We will use this inequality to obtain the desired results. Our approach proceeds in two stages; first we consider the behaviour of an ideal forecast for which we obtain tight bounds and secondly we consider the behaviour of an empirical deviation from that ideal. Consider first the ideal forecast. We define the forecast deviation as

$$e_h = s_{Qh} - s_h = \beta_Q' f_{Qh} - \beta' f_h$$

It seems natural to measure the quality of the ideal forecast by a non-dimensional ratio

$$\frac{\mathrm{var}(e_h)}{\mathrm{var}(\beta' f_h)} = \frac{\mathrm{var}(e_h)}{\beta'\beta}$$

however this proves to be somewhat tedious and is under development. For the present purposes a simpler more rapid approach, but one which preserves the idea of measuring the relative magnitude of forecast error variances is to consider instead the first moment measure $\frac{E|e_h|}{\sqrt{\beta'\beta}}$. We make the following assumptions.

Assumptions:

1. $\sum_{j=0}^{T-1} \left|E(y_t y_{t-j})\right| < \infty$, $\sum_{j=0}^{T-1} \left|E(y_t x_{t-j})\right| < \infty_p$ $\Omega = E(x_t x_t') < \infty_{p\times p}$;
2. $E(f_t f_t') = I_k$, $E(v_t^2) = \sigma_v^2$, and $E(n_t n_t') = \Psi$ a positive definite matrix;
3. $E(f_t n_t') = 0$, $E(f_t v_t) = 0$, and $E(n_t v_t) = 0$.
4. f_t, v_t and n_t are Gaussian.

Thus, our results apply to stationary Gaussian processes. Assumptions 2 and 3 are technically convenient but, since the model is not identified, do not restrict the covariance of x_t and y_t. The assumption of Gaussianity may be easily relaxed, but at the cost of some ease of interpretability.

We define $r^2 = \frac{\|\beta\|^2}{\|\beta\|^2 + \sigma_v^2}$. Note that r^2 is the proportion of the variance of y_t that is explained by the factors. Thus, it may be interpreted as the population analogue of the R^2 statistic from regression analysis. We also denote

$$\gamma = \sum_{j=1}^{T-1} \left| \frac{E(y_t y_{t-j})}{E(y_t^2)} \right| + \sup_i \sum_{j=1}^{T-1} \left| \frac{E(y_t s_{i,t-j})}{\sqrt{E(y_t^2)E(s_{it}^2)}} \right|$$

We then have the following.

Theorem 1: Under the above assumptions

$$\frac{E|e_h|}{\sqrt{\beta'\beta}} \leq \sqrt{k\left(\rho^2 + \frac{2k\gamma}{r^2 T}\right)} + \sqrt{k\rho\left(1 + \frac{2\gamma}{r^2 T}\right)}$$

Proof: See appendix.

In conjunction with the Markov inequality, Theorem 1 allows us to place a probability bound on the difference between the principal component forecast and the ideal forecast. For example,

$$P\left\{\left|\frac{s_{Qh} - s_h}{\sqrt{VAR(s_h)}}\right| > 1\right\} \leq \sqrt{k\left(\rho^2 + \frac{2k\gamma}{r^2 T}\right)} + \sqrt{k\rho\left(1 + \frac{2\gamma}{r^2 T}\right)}$$

where the left hand side of the inequality is the probability that the principal component forecast will deviate from the ideal forecast by more than one standard deviation of the ideal forecast. In order for the ideal forecast to be reasonably close to the theoretically optimal forecast, we need the noise to signal ratio to be fairly small and the sample size to be reasonably large. Precisely how large the sample size needs to be will depend on the magnitude of the autocovariances of the data and the proportion of the variance of the forecast variable that is determined by the factors. Furthermore, we can demonstrate that

$$e_h = s_{Qh} - s_h \xrightarrow{p} 0 \text{ as } \left(\frac{1}{T}, \rho\right) \to (0,0).$$

Of additional interest is the asymptotic behaviour of the empirical forecast. This is far more complicated and a complete theory is under development. However, the following result is straightforward. We define the empirical forecast deviation to be

$\hat{e}_h = \hat{\beta}'_Q \hat{f}_{Qh} - \beta'_Q f_{Qh}$. We then have

Theorem 2: If the elements of Λ are distinct then, for fixed ρ and fixed p, $\hat{e}_h \xrightarrow{p} 0$ as $T \to \infty$.

Proof: See appendix.

Thus, as T grows, the empirical forecast converges in probability to a forecast which differs from the mean square efficient forecast by a known probability bound which is related to the noise to signal ratio.

Measurement of the Noise-to-Signal Ratio

Given the above theory, measurement of the noise to signal ratio is a concern of some practical importance. Since Ψ is not identified, the eigenvalue σ^2 and accordingly the noise to signal ratio ρ are not identified. Thus, direct estimation of the noise to signal ratio is not possible. However, it is possible to consistently estimate a lower bound on the noise to signal ratio. Let $\Phi = \sigma^2 I_p - \Psi$. Then $\Phi + \Omega = BB' + \sigma^2 I_p$. Note that $eig_j(\Phi) = \sigma^2 - \sigma_j^2$, where $eig_j(.)$ denotes the j^{th} ordered eigenvalue of its matrix argument, so $eig_j(\Phi) \geq 0 \ \forall j = 1,..,p$. Thus, Φ is positive semi-definite. It follows from Magnus and Neudecker (1991, p.208, Theorem 9) that $eig_j(BB' + \sigma^2 I_p) \geq eig_j(\Omega)$, i.e. $d_j + \sigma^2 \geq \lambda_j \ \forall j = 1,..,p$. Since $d_{k+1} = 0$, $\lambda_{k+1} \leq \sigma^2$. We therefore have that

$$\frac{\lambda_{k+1}}{\lambda_k} \leq \rho$$

This expression makes it clear that in order for the noise to signal ratio to be small, implying that the k-principal component forecast is close to the theoretical ideal forecast, there must exist a large gap between the k^{th} and $(k+1)^{th}$ eigenvalues of the covariance matrix of predictor variables. This links asymptotic principal component techniques to the traditional principal component literature. Obviously, this ratio of eigenvalues may be consistently estimated by the corresponding ratio of sample eigenvalues.

A hypothesis test of the magnitude of the lower bound of the noise to signal ratio may be based on

$$f(\hat{\lambda}) = \frac{\hat{\lambda}_{k+1}}{\hat{\lambda}_k} - \frac{\lambda_{k+1}}{\lambda_k}$$

where $\lambda = (\lambda_{k+1} \quad \lambda_k)'$ and $\hat{w}$ denotes a sample estimate of a population parameter w. We assume that λ_{k+1} and λ_k are distinct. It is well known that $z = \sqrt{T}(\hat{\lambda} - \lambda)$ is asymptotically Gaussian. Since $\sqrt{T}f(\hat{\lambda}) = \frac{\partial f}{\partial \lambda'} z + O(T^{-\frac{1}{2}})$ it follows that $\sqrt{T}f(\hat{\lambda})$ is asymptotically Gaussian. Furthermore, $VAR\left[\sqrt{T}f(\hat{\lambda})\right] = \left(\frac{2\lambda_{k+1}}{\lambda_k}\right)^2 + O(T^{-1})$ so

$$f(\hat{\lambda}) = \frac{\frac{\sqrt{T}}{2}\left(\frac{\hat{\lambda}_{k+1}}{\hat{\lambda}_k} - \frac{\lambda_{k+1}}{\lambda_k}\right)}{\frac{\lambda_{k+1}}{\lambda_k}} \Rightarrow N(0,1) \text{ as } T \to \infty \text{ with p fixed, where } \Rightarrow \text{ denotes}$$

convergence in distribution.

From the previous section, as $T \to \infty$ $P\left\{\left|\frac{s_{Qh} - s_h}{\sqrt{VAR(s_h)}}\right| > 1\right\} \le \sqrt{k\rho}\left(\sqrt{\rho} + 1\right)$

If we can choose a desired numerical bound for the above probability, which we denote as α, then we may solve the equation $\alpha = \sqrt{k\rho}\left(\sqrt{\rho} + 1\right)$ to find a corresponding bound for ρ, which we denote ρ_α. Since $\frac{\lambda_{k+1}}{\lambda_k} \le \rho$, we may test H_0: $\frac{\lambda_{k+1}}{\lambda_k} = \rho_\alpha$; H_1: $\frac{\lambda_{k+1}}{\lambda_k} > \rho_\alpha$. Note that rejection of the null implies $\rho > \rho_\alpha$. Some values of ρ, k and α which solve the above equation $\alpha = \sqrt{k\rho}\left(\sqrt{\rho} + 1\right)$ are given in the table below.

Table 1: Values of ρ_α

α	k					
	1	**2**	**3**	**4**	**5**	**6**
0.01	9.8E-05	4.93E-05	3.3E-05	2.48E-05	1.98E-05	1.65E-05
0.05	0.002277	0.001169	0.000788	0.000596	0.000479	0.0004
0.1	0.008392	0.004397	0.002996	0.002277	0.001839	0.001839
0.25	0.042893	0.023495	0.016375	0.012628	0.010302	0.008714
0.5	0.133975	0.076666	0.05473	0.042893	0.035416	0.030237

The table and histogram below show the results of a Monte Carlo simulation of this test. We initially choose to set the number of observations to 100, the number of variables to 5 and the number of factors to 2. We then raise the number of variables to 50, then to 200. We choose the first 3 population eigenvalues to be 100, 75 and 5. The

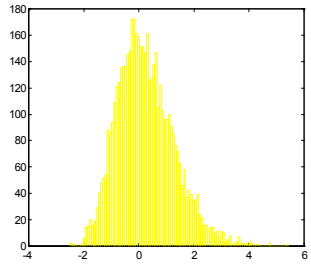
remaining eigenvalues decay linearly to 0.001. Thus, $\frac{\lambda_{k+1}}{\lambda_k} = 0.0667$. We conduct 5000 simulations of the test statistic for each model.

Table 2: Empirical and Theoretical Distributions of the Test Statistic (k=2, T=100, 5000 simulations)

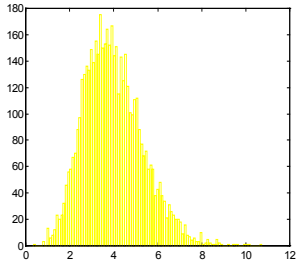
Theoretical Percentiles (α)	**Empirical Percentiles**			
		p = 5	p = 50	p = 200
	10%	0.155	0.9934	1
	5%	0.095	0.98	1
	1%	0.0334	0.9102	1

The elements in the table are the proportions of the empirical probability mass that lie above the critical value corresponding to α. Thus, for example, the 5% critical value for a standard Gaussian distribution is 1.645, and the table shows that, for a model with 5 observable variables, 0.095 of the empirical probability mass lies above 1.645.

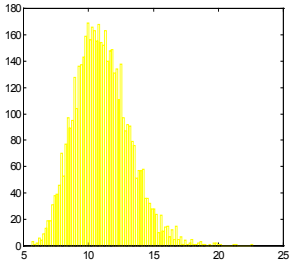
p = 5



p = 50



p = 200



It is clear in this example that the first order asymptotic approximation does not work well and has a strong bias that increases in p, the number of observable variables. A similar phenomenon has been previously noted in the literature. When testing for the number of factors underlying returns of securities traded on the New York and American stock exchanges, Dhrymes, Friend and Gultekin (1984) find that the number of factors chosen by a standard test statistic based on eigenvalues increases with the number of securities analysed in the model. Bai and Ng (2002, p.195) state that it can be shown that all non-zero eigenvalues of a sample covariance matrix increase with p and they claim that a test based on sample eigenvalues is not feasible. However, neither of these papers offers a satisfactory explanation of this phenomenon. In order to explain the deterioration in the test as p gets large, we need to investigate the higher order asymptotics of the test statistic. Fujikoshi (1980) has derived a general expression for the Edgeworth expansion of a function of the eigenvalues of a sample covariance matrix which may be applied here. Not surprisingly, the resulting expression is complex. However, we can present a simpler

partial explanation of the failure of the test by restricting our attention to the first moment of the test statistic. Taking a Taylor Series expansion of $f(\hat{\lambda})$ about λ yields

$$\sqrt{T}f(\hat{\lambda}) = \frac{\lambda_{k+1}}{\lambda_k^2}z_k - \frac{1}{\lambda_k}z_{k+1} - \frac{2\lambda_{k+1}}{\sqrt{T}\lambda_k^3}z_k^2 + \frac{1}{\sqrt{T}\lambda_k^2}z_k z_{k+1} + O(T^{-\frac{3}{2}})$$

where z_I is the i^{th} element of z. Denoting $\varpi = \text{vec}(S)$ and $\zeta = \text{vec}(\Omega)$, a Taylor series expansion of the j^{th} sample eigenvalue is

$$\hat{\lambda}_j - \lambda_j = \frac{1}{\sqrt{T}}\left(q_j' \otimes q_j'\right)h + \frac{1}{2T}h'\left[\left(I_p \otimes q_j\right) + \left(q_j \otimes I_p\right)\right]\left(q_j' \otimes \sum_{i\neq j}\frac{q_j q_j'}{\lambda_j - \lambda_i}\right)h + O(T^{-\frac{3}{2}})$$

where $h = \sqrt{T}\left(\varpi - \zeta\right)$. Since pre- and post-multiplying a matrix by an orthogonal matrix does not change its eigenvalues we can, without loss of generality, restrict attention to the case where $\Omega = \begin{pmatrix} \Lambda & 0 \\ 0 & \Lambda_\perp \end{pmatrix}$. This greatly simplifies the following expressions. Using the above we have

$$E(z_j) = \sqrt{T}E\left(\hat{\lambda}_j - \lambda_j\right) = \frac{1}{\sqrt{T}}\sum_{i\neq j}\frac{M_{ijij}}{\lambda_j - \lambda_i} + O(T^{-1})$$

$$E(z_j^2) = TE\left(\hat{\lambda}_j - \lambda_j\right)^2 = M_{jjjj} + O(T^{-\frac{1}{2}})$$

$$E(z_j z_l) = TE\left(\hat{\lambda}_j - \lambda_j\right)\left(\hat{\lambda}_l - \lambda_l\right) = M_{jjll} + O(T^{-\frac{1}{2}})$$

where $M_{ijrs} = E\left\{\left[\frac{1}{T}\sum_{t=1}^{T}v_{it}v_{jt} - E\left(\frac{1}{T}\sum_{t=1}^{T}v_{it}v_{jt}\right)\right]\left[\frac{1}{T}\sum_{t=1}^{T}v_{rt}v_{st} - E\left(\frac{1}{T}\sum_{t=1}^{T}v_{rt}v_{st}\right)\right]\right\}$ and v_{wt} is the t^{th} observation of the w^{th} element of the p×1 vector of observations which has been transformed to have the diagonal covariance explained above. We therefore have

$$\sqrt{T}E\left[f(\hat{\lambda})\right] = \frac{1}{\lambda_k\sqrt{T}}\left(\frac{\lambda_{k+1}}{\lambda_k}\sum_{i\neq k}\frac{M_{ikik}}{\lambda_k - \lambda_i} - \sum_{i\neq k+1}\frac{M_{i(k+1)i(k+1)}}{\lambda_{k+1} - \lambda_i} - \frac{2\lambda_{k+1}}{\lambda_k^2}M_{kkkk} + \frac{1}{\lambda_k}M_{(k+1)(k+1)kk}\right) + O(T^{-1})$$

It is likely that the second term in the brackets in this expression would tend to be large for large p. Since the smallest p-k eigenvalues are bounded above by λ_{k+1} and below by zero, if p is quite large then the smallest p-k eigenvalues may be quite close

to each other. Consequently, the second summed term in the brackets would be the sum of a large number of fairly large terms.

There are at least three approaches that might be tried to circumvent this problem: a Barlett-type adjustment that corrects for the $T^{-1/2}$ term to give a statistic which is better approximated by the standard normal density in finite samples; using an Edgeworth expansion to give a better approximation of the density of the unadjusted statistic; or judiciously redesigning the test to avoid the problem. We have had by far the most success with the third of these approaches. It was shown previously that $\lambda_j = \sigma^2 + d_j$ for j = 1,..,p. Since $d_j = 0$ for $j \geq k+1$ it follows that $\frac{1}{p-k}\sum_{j=k+1}^{p} \lambda_j \leq \lambda_{k+1} \leq \sigma^2$. We therefore construct a test of H_0: $\frac{\frac{1}{p-k}\sum_{j=k+1}^{p} \lambda_j}{\lambda_k} = \rho_\alpha$; H_1: $\frac{\frac{1}{p-k}\sum_{j=k+1}^{p} \lambda_j}{\lambda_k} > \rho_\alpha$. Note that rejection of the null implies $\rho > \rho_\alpha$.

Consider the statistic $\hat{\theta} = \sqrt{T}\sum_{j=k+1}^{p} \hat{\lambda}_j - \rho_\alpha (p-k)\hat{\lambda}_k$. A Taylor series approximation of the expected value of $\hat{\theta}$ is

$$E(\hat{\theta}) = -\frac{1}{\sqrt{T}}\left\{[1-\rho_\alpha(p-k)]\sum_{i\neq k}\frac{M_{ikik}}{\lambda_k - \lambda_i} + \sum_{j=1}^{k-1}\sum_{i\neq k}\frac{M_{ijij}}{\lambda_j - \lambda_i}\right\} + O(T^{-\frac{3}{2}})$$

Note that the denominators in the summed terms in this expression involve distances from the first k eigenvalues only. Since under the null the distance between eigenvalues k and k+1 is large, the summed terms are likely to be small even if p is quite large. Thus, a test statistic based on $\hat{\theta}$ is less likely to suffer from bias problems as p gets large. It is straightforward to show that

$$VAR(\hat{\theta}) = 2\left[\rho_\alpha^2 (p-k)^2 \lambda_k^2 + \sum_{j=k+1}^{p} \lambda_j^2\right] + O(T^{-1}).$$

We therefore propose the test statistic

$$\phi = \sqrt{\frac{T}{2}}\frac{\left(\sum_{j=k+1}^{p} \hat{\lambda}_j - \rho_\alpha (p-k)\hat{\lambda}_k\right)}{\sqrt{\rho_\alpha^2 (p-k)^2 \hat{\lambda}_k^2 + \sum_{j=k+1}^{p} \hat{\lambda}_j^2}}$$

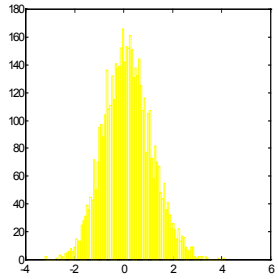
By an argument similar to that presented above, $\phi \Rightarrow N(0,1)$ as $T \to \infty$ with p fixed. We now conduct Monte Carlo simulations for this test that are similar to those conducted for the previous statistic.

Table 3: Empirical and Theoretical Distributions of the Alternative Test Statistic (k=2, T=100, 5000 simulations)

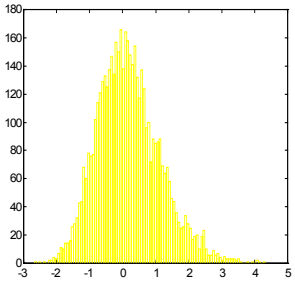
Theoretical Percentiles (α)	**Empirical Percentiles**		
	p = 5	p = 50	p = 200
10%	0.1282	0.1264	0.0566
5%	0.0696	0.0728	0.0286
1%	0.0178	0.0256	0.0076

The elements in the table are the proportions of the empirical probability mass that lie above the critical value corresponding to α. Thus, for example, the 5% critical value for a standard Gaussian distribution is 1.645, and the table shows that, for a model with 5 observable variables, 0.095 of the empirical probability mass lies above 1.645.

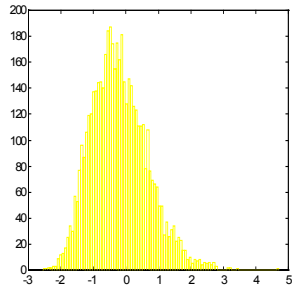
p = 5



p = 50



p = 200

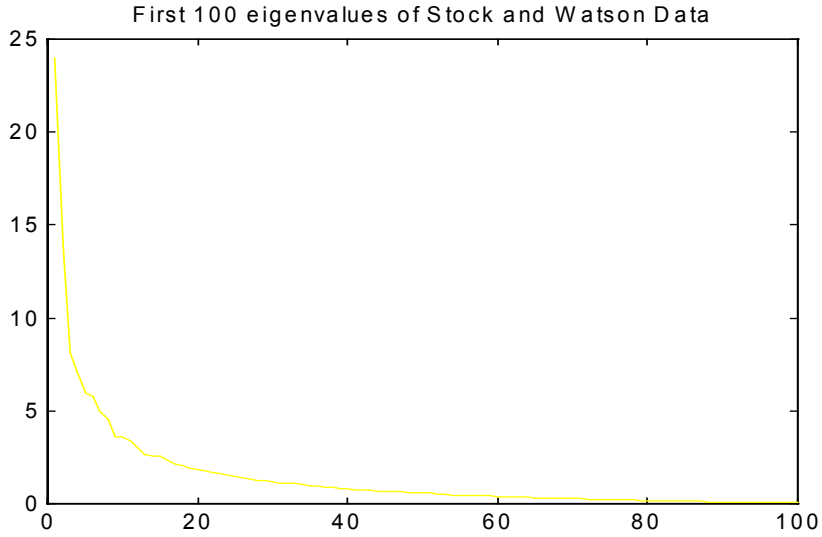


While there is some evidence of skewness, the extreme bias in p that exists in the original test statistic is not present in the alternative test statistic, even when p is very large.

An Empirical Example

Stock and Watson (1998) have collected a large data set of variables describing the US macroeconomy which they employ in a forecasting experiment using a factor model. The interested reader is directed to their paper for a description of the data, and to Mark Watson's web site (http://www.wws.Princeton.EDU/~mwatson/publi.html) for the data set. We follow Stock and Watson in taking logs and/or differences or double-differences for some variables. Following appropriate transformations the balanced panel contains 149 variables measured monthly from March 1959 to December 1998. These variables are rescaled to a zero mean and unit variance.

The plot below shows the first 100 eigenvalues of the Stock and Watson data.



Note that the first few sample eigenvalues drop sharply but the plot levels out after that. It should also be noted that, when compared to the values in Table 1, the ratios $\hat{\lambda}_{k+1} / \hat{\lambda}_k$ are not of the appropriate order of magnitude to support an argument that principal component forecasts are close to the mean-square efficient forecast. However, it should be remembered that it is possible that these ratios are being estimated with a substantial bias. In order to determine whether a near-efficiency argument can be supported for this data set, we conduct the hypothesis test of the previous section for factor models of orders 1 to 6. We choose values for ρ_α corresponding to probability bounds of 5%, 10%, 25% and 50%. The results of the tests are presented in the following table.

Table 4: Results of Test for Stock and Watson Data

	K					
ρ_α	1	2	3	4	5	6
0.05	72.9153	88.1466	92.2046	94.7054	96.3465	98.5772
0.1	38.8245	75.5352	88.07	92.0174	94.5692	96.5609
0.25	-2.7385	19.1625	49.8191	63.2659	74.4605	79.3342
0.5	-11.3836	-4.3701	8.6955	17.1004	27.5315	32.3147

Note that for a 1-factor model we cannot reject a probability bound of 0.25 and for a 2-factor model we cannot reject a probability bound of 0.5, using significance levels of 5 per cent. For all other factor models and probability bounds we can strongly reject the hypothesis about the probability bound. Based on these results we cannot argue convincingly that principal component forecasts using the Stock and Watson data are likely to be close to the theoretical mean-square efficient forecast.

Conclusions

It is clearly the case that regressions on principal components of large numbers of predictor variables has something to offer macroeconomic forecasters. What is less clear is whether these forecasts are optimal, or whether a full factor analysis needs to be carried out in order to generate efficient forecasts from large dimensional data sets. We have produced a probability bound based on a noise to signal ratio which indicates when principal component forecasts will be close to a mean-square efficient forecast. We have also produced a test for the magnitude of this noise to signal ratio. Our empirical results using the Stock and Watson data set apparently indicate that our probability bound does not suggest that the principal component forecasts are near optimal in this case. This raises the possibility that a full factor analysis might produce superior forecasts.

Appendix

We define $R = U'Q$ and denote the largest eigenvalue of a matrix X by $\lambda_{max}(X)$. The proof of Theorem 1 will make use of the following lemmas.

Lemma 1: If $x \sim N(0,\Gamma)$ and α and β are vectors of conformable dimension then $E(\alpha'x)^2(\beta'x)^2 = \alpha'\Gamma\alpha\beta'\Gamma\beta + 2(\alpha'\Gamma\beta)^2$. The proof is elementary.

Corollary: $\text{var}(\alpha'x\beta'x) = \alpha'\Gamma\alpha\beta'\Gamma\beta + (\alpha'\Gamma\beta)^2$. Again, the proof is elementary.

Lemma 2: If $z_t = \begin{pmatrix} w_t \\ u_t \end{pmatrix}$ is Gaussian and $E(z_t z_{t-j}') = \Gamma^{(j)} = \begin{pmatrix} \Gamma_w^{(j)} & \Gamma_{wu}^{(j)} \\ \Gamma_{uw}^{(j)} & \Gamma_u^{(j)} \end{pmatrix}$, then using Lemma 1, and the Cauchy-Schwarz inequality,

$$\text{var}(a'S_{wu}b) = \text{var}\left(\frac{1}{T}\sum_{t=1}^{T} a'w_t b'u_t\right) \le \frac{2}{T}\left(a'\Gamma_w^{(0)}ab'\Gamma_u^{(0)}b + \sum_{j=1}^{T-1} a'\Gamma_w^{(j)}ab'\Gamma_u^{(j)}b + \sum_{j=1}^{T-1} a'\Gamma_{wu}^{(j)}ab'\Gamma_{wu}^{(-j)}b\right)$$

where a and b are vectors of conformable dimension.

Corollary: $E(\alpha'u)^2(\beta'v)^2 = \alpha'\Gamma_u\alpha\beta'\Gamma_v\beta + 2(\alpha'\Gamma_{uv}\beta)^2 \le 3\alpha'\Gamma_u\alpha\beta'\Gamma_v\beta$

Lemma 3: If Z is a random vector and e_i is a k×1 vector of zeros but with a 1 in position i, and M is a k×k constant, then

$$E(Z'M'MZ) = E(Z'M'\sum_{i=1}^{k} e_i e_i'MZ) = \sum_{i=1}^{k} E(Z'M'e_i)(e_i'MZ) = \sum_{i=1}^{k} E(e_i'MZ)^2$$

Lemma 4: $\Lambda = Q'(UDU' + \Psi)Q$

Proof: $\Omega Q = Q\Lambda \Rightarrow (UDU' + \Psi)Q = Q\Lambda$. Premultiplying by Q' gives the result.

Lemma 5: If $M = D^{\frac{1}{2}}R\Lambda^{-1}R'D^{\frac{1}{2}}$, then the eigenvalues of $I - M$ are equal to the eigenvalues of $\Lambda^{-\frac{1}{2}}Q'\Psi Q\Lambda^{-\frac{1}{2}}$.

Proof: The eigenvalues of $I - M$ are the solutions of

$$0 = \left|\lambda I - (I - M)\right| = \left|(\lambda - 1)I + M\right| = \left|(\lambda - 1)I + D^{\frac{1}{2}}R\Lambda^{-1}R'D^{\frac{1}{2}}\right|$$

$$= \left|(\lambda - 1)I + R\Lambda^{-1}R'D\right| = \left|(\lambda - 1)I + I - Q'\Psi Q\Lambda^{-1}\right| \text{ from Lemma 4}$$

$$= \left|\lambda I + Q'\Psi Q\Lambda^{-1}\right| = \left|\lambda I + \Lambda^{-\frac{1}{2}}Q'\Psi Q\Lambda^{-\frac{1}{2}}\right|.$$

Lemma 6: The eigenvalues of $D^{\frac{1}{2}}R\Lambda^{-2}R'D^{\frac{1}{2}}$ are equal to the eigenvalues of $\Lambda^{-1}R'DR\Lambda^{-1}$.

Proof: The eigenvalues of $D^{\frac{1}{2}}R\Lambda^{-2}R'D^{\frac{1}{2}}$ are the solutions for λ of

$$0 = \left|\lambda I - D^{\frac{1}{2}}R\Lambda^{-2}R'D^{\frac{1}{2}}\right| = \left|\lambda I - R\Lambda^{-2}R'D\right| = \left|\lambda I - \Lambda^{-1}R'DR\Lambda^{-1}\right|$$

Proof of Theorem 1: Defining $S_{xy} = \frac{1}{T}\sum_{t=1}^{T} x_t y_t$, the forecast deviation is

$$e_h = \beta_Q' f_{Qh} - \beta' f_h = \left(\Lambda^{-\frac{1}{2}}Q'S_{xy}\right)'\Lambda^{-\frac{1}{2}}Q'x_h - \beta' f_h = S_{xy}'Q\Lambda^{-1}Q(Bf_h + n_h) - \beta' f_h$$

$$= e_a + e_b \text{ where}$$

$e_a = \left(S_{xy}'Q\Lambda^{-1}Q'B - \beta'\right)f_h$ and $e_b = S_{xy}'Q\Lambda^{-1}Q'n_h$.

First consider e_b. From the Cauchy-Schwarz inequality we have

$$\left|e_b\right| \le \left\|S_{xy}'Q\Lambda^{-1}\right\|\left\|Q'n_h\right\| \text{ and } \left(E\left|e_b\right|\right)^2 \le E\left\|S_{xy}'Q\Lambda^{-1}\right\|^2 E\left\|Q'n_h\right\|^2 \qquad \text{(A1)}$$

We have that $E\left\|Q'n_h\right\|^2 = \text{tr}(Q'\Psi Q) \le \sigma^2 k$. (A2)

Also

$$E\left\|S_{xy}'Q\Lambda^{-1}\right\|^2 = E\left(S_{xy}'Q\Lambda^{-2}Q'S_{xy}\right) = \sum_{i=1}^{k} E\left(e_i'\Lambda^{-1}Q'S_{xy}\right)^2 \text{ from Lemma 3}$$

$$= \sum_{i=1}^{k}\left[\operatorname{var}\left(e_i'\Lambda^{-1}Q'S_{xy}\right) + E\left(e_i'\Lambda^{-1}Q'S_{xy}\right)^2 \right] = \sum_{i=1}^{k}\left[\operatorname{var}\left(e_i'\Lambda^{-1}Q'S_{xy}\right) + E\left(e_i'\Lambda^{-1}Q'B\beta\right)^2 \right] \tag{A3}$$

Now $\sum_{i=1}^{k}\left(e_i'\Lambda^{-1}Q'B\beta\right)^2 = \beta BQ\Lambda^{-2}Q'B\beta = \beta D^{\frac{1}{2}}U'Q\Lambda^{-2}Q'UD^{\frac{1}{2}}\beta$

$$= \beta D^{\frac{1}{2}}R\Lambda^{-2}R'D^{\frac{1}{2}}\beta \le \beta'\beta\lambda_{max}\left(\Lambda^{-1}R'DR\Lambda^{-1}\right) \text{ by Lemma 6}$$
$$\le \beta'\beta\lambda_{max}\left(\Lambda^{-1}\Lambda\Lambda^{-1}\right) \text{ by Lemma 4}$$
$$\le \beta'\beta\lambda_k^{-1} \tag{A4}$$

Also, from Lemma 2

$\sum_{i=1}^{k}\operatorname{var}\left(e_i'\Lambda^{-1}Q'S_{xy}\right) \le \Upsilon_1 + \Upsilon_2 + \Upsilon_3$ where

$$\Upsilon_1 = \frac{2}{T}\sum_{i=1}^{k}e_i'\Lambda^{-1}Q'\Omega Q\Lambda^{-1}e_i\sigma_y^{(0)2}$$

$$\Upsilon_2 = \frac{2}{T}\sum_{i=1}^{k}\sum_{j=1}^{T-1}\left|e_i'\Lambda^{-1}Q'E(x_t x_{t-j}')Q\Lambda^{-1}e_i\sigma_y^{(j)2}\right|$$

$$\Upsilon_3 = \frac{2}{T}\sum_{i=1}^{k}\sum_{j=1}^{T-1}\left|e_i'\Lambda^{-1}Q'E(x_t y_{t-j})E(y_t x_{t-j}')Q\Lambda^{-1}e_i\right|$$

where $\sigma_y^{(j)2} = E(y_t y_{t-j})$. We have that

$$\Upsilon_1 = \frac{2}{T}\sigma_y^{(0)2}\operatorname{tr}\left(\Lambda^{-1}\Lambda\Lambda^{-1}\right) \le \frac{2}{T}\sigma_y^{(0)2}\lambda_k^{-1}$$

$\Upsilon_2 = \frac{2}{T}\sum_{i=1}^{k}\sum_{j=1}^{T-1}\left|e_i'\Lambda^{-\frac{1}{2}}E(s_t s_{t-j}')\Lambda^{-\frac{1}{2}}e_i\sigma_y^{(j)2}\right|$ where $s_t = \Lambda^{-\frac{1}{2}}Q'x_t$ is the principal component vector of x_t. So

$$\Upsilon_2 = \frac{2}{T}\sum_{i=1}^{k}\sum_{j=1}^{T-1}\left|\frac{E(s_t s_{t-j}')}{\lambda_i}\sigma_y^{(j)2}\right| \le \frac{2}{T}\sum_{i=1}^{k}\sum_{j=1}^{T-1}\frac{1}{\lambda_i}\left|E(s_t s_{t-j}')\right|\left|\sigma_y^{(j)2}\right| \le \frac{2}{T}\lambda_k^{-1}\sum_{j=1}^{T-1}\left|\sigma_y^{(j)2}\right|$$

by Cauchy-Schwarz, and

$$\Upsilon_3 = \frac{2}{T}\sum_{i=1}^{k}\sum_{j=1}^{T-1}\frac{1}{\lambda_i}\left|E(s_{it}y_{t-j})E(y_t s'_{i,t-j})\right| \le \frac{2}{T}\sum_{i=1}^{k}\sum_{j=1}^{T-1}\frac{1}{\lambda_i}\left|E(s_{it}y_{t-j})\right|\left|E(y_t s'_{i,t-j})\right| \le \frac{2\sigma_y^{(0)2}}{T}\sum_{i=1}^{k}\sum_{j=1}^{T-1}\frac{1}{\lambda_i}\left|E(y_t s'_{i,t-j})\right|$$

$$\le \frac{2\sigma_y^{(0)2}}{T}\lambda_k^{-1}\sup_i\sum_{j=1}^{T-1}\left|E(y_t s'_{i,t-j})\right|$$

$$\text{so } \sum_{i=1}^{k}\text{var}\left(e_i'\Lambda^{-1}Q'S_{xy}\right) \le \frac{2}{T\lambda_k}\left(\sum_{j=0}^{T-1}\left|\sigma_y^{(j)2}\right| + \sigma_y^{(0)}\sup_i\sum_{j=1}^{T-1}\left|E(y_t s'_{i,t-j})\right|\right) = \frac{2}{T\lambda_k}\sigma_y^{(0)2}\gamma$$

(A5)

$$\text{where } \gamma = \sum_{j=1}^{T-1}\left|\frac{E(y_t y_{t-j})}{E(y_t^2)}\right| + \sup_i\sum_{j=1}^{T-1}\left|\frac{E(y_t s_{i,t-j})}{\sqrt{E(y_t^2)E(s_{it}^2)}}\right|$$

Equations A3, A4, and A5 yield

$$E\left\|S'_{xy}Q\Lambda^{-1}\right\|^2 \le \lambda_k^{-1}\left(\frac{2\sigma_y^{(0)2}\gamma}{T} + \beta'\beta\right)$$

which when combined with A1 and A2 yield

$$\frac{\left(E|e_b|\right)^2}{\|\beta\|^2} \le \frac{\sigma^2 k}{\lambda_k}\left(\frac{2}{T}\frac{\sigma_y^{(0)2}\gamma}{\|\beta\|^2} + 1\right).$$

We now consider e_a. By the Cauchy-Schwarz inequality we have

$$\left(E|e_a|\right)^2 \le E\left\|S'_{xy}Q\Lambda^{-1}QB - \beta'\right\|^2 E\|f_h\|^2 = kE\left\|S'_{xy}Q\Lambda^{-1}QB - \beta'\right\|^2 \qquad \text{(A6)}$$

$$\text{Now } E\left\|S'_{xy}Q\Lambda^{-1}QB - \beta'\right\|^2 = \sum_{i=1}^{k}E\left[e_i'\left(B'Q\Lambda^{-1}Q'S_{xy} - \beta\right)^2\right] \text{ from Lemma 3}$$

$$= \sum_{i=1}^{k}\text{var}\left[e_i'\left(B'Q\Lambda^{-1}Q'S_{xy} - \beta\right)\right] + \sum_{i=1}^{k}E\left[e_i'\left(B'Q\Lambda^{-1}Q'S_{xy} - \beta\right)\right]^2 \qquad \text{(A7)}$$

but from Lemma 2

$$= \sum_{i=1}^{k} \mathrm{var}\left[e_i' \left(B'Q\Lambda^{-1}Q'S_{xy} - \beta \right) \right] \leq \Delta_1 + \Delta_2 + \Delta_3 \quad \text{where}$$

$$\Delta_1 = \sum_{i=1}^{k} \frac{2}{T} e_i' B'Q\Lambda^{-1}Q'\Omega Q\Lambda^{-1}Q'Be_i \sigma_y^{(0)2}$$

$$\Delta_2 = \frac{2}{T} \sum_{i=1}^{k} \sum_{j=1}^{T-1} \left| e_i' B'Q\Lambda^{-1}Q'E(x_t x_{t-j})Q\Lambda^{-1}Q'Be_i \sigma_y^{(j)2} \right|$$

$$\Delta_3 = \frac{2}{T} \sum_{i=1}^{k} \sum_{j=1}^{T-1} \left| e_i' B'Q\Lambda^{-1}Q'E(x_t y_{t-j}) e_i' B'Q\Lambda^{-1}Q'E(x_t y_{t-j}) \right|$$

We have that

$$\Delta_1 = \frac{2\sigma_y^{(0)2}}{T} \mathrm{tr}\left(B'Q\Lambda^{-1}Q'B \right) \leq \frac{2\sigma_y^{(0)2}}{T} \mathrm{tr}\left(B'\Omega^{-1}B \right) \leq \frac{2k\sigma_y^{(0)2}}{T}$$

$$\Delta_2 = \frac{2}{T} \sum_{i=1}^{k} \sum_{j=1}^{T-1} \left| e_i' B'Q\Lambda^{-\frac{1}{2}} E(s_t s_{t-j}) \Lambda^{-\frac{1}{2}} Q'Be_i \sigma_y^{(j)2} \right|$$

let $v_i = e_i' B'Q\Lambda^{-\frac{1}{2}} = \begin{pmatrix} 0 & \dots & 0 & v_i & 0 & \dots & 0 \end{pmatrix}$ then

$$\Delta_2 = \frac{2}{T} \sum_{i=1}^{k} \sum_{j=1}^{T-1} \left| v_i E(s_t s_{t-j}) v_i' \sigma_y^{(j)2} \right| = \frac{2}{T} \sum_{i=1}^{k} \sum_{j=1}^{T-1} \left| v_i^2 E(s_t s_{t-j})_i \sigma_y^{(j)2} \right| \leq \frac{2}{T} \sum_{i=1}^{k} \sum_{j=1}^{T-1} v_i^2 \left| \sigma_y^{(j)2} \right|$$

$$\leq \frac{2}{T} \sum_{i=1}^{k} e_i' BQ\Lambda^{-1}Q'Be_i \sum_{j=1}^{T-1} \left| \sigma_y^{(j)2} \right| \leq \frac{2}{T} \sum_{i=1}^{k} \mathrm{tr}(B'\Omega^{-1}B) \sum_{j=1}^{T-1} \left| \sigma_y^{(j)2} \right| \leq \frac{2k}{T} \sum_{j=1}^{T-1} \left| \sigma_y^{(j)2} \right|$$

$$\Delta_3 = \frac{2}{T} \sum_{i=1}^{k} \sum_{j=1}^{T-1} \left| e_i' B'Q\Lambda^{-\frac{1}{2}} E(s_t y_{t-j}) E(y_t s_{t-j}') \Lambda^{-\frac{1}{2}} Q'Be_i \right| = \frac{2}{T} \sum_{i=1}^{k} e_i' B'Q\Lambda^{-1}Q'Be_i \sum_{j=1}^{T-1} \left| E(s_t y_{t-j}) E(y_t s_{t-j}') \right|$$

$$\leq \frac{2}{T} \mathrm{tr}(B'\Omega^{-1}B) \sigma_y^{(0)} \sup_i \sum_{j=1}^{T-1} \left| E(y_t s_{t-j}') \right|$$

So

$$\sum_{i=1}^{k} \mathrm{var}\left[e_i' \left(B'Q\Lambda^{-1}Q'S_{xy} - \beta \right) \right] \leq \frac{2k}{T} \sum_{j=1}^{T-1} \left| \sigma_y^{(j)2} \right| + \sigma_y^{(0)} \sup_i \sum_{j=1}^{T-1} \left| E(y_t s_{t-j}') \right| = \frac{2k}{T} \sigma_y^{(0)} \gamma \quad \text{(A8)}$$

Also $\sum_{i=1}^{k} E\left[e_i' \left(B'Q\Lambda^{-1}Q'S_{xy} - \beta \right) \right] = e_i'(M - I)\beta$ where $M = D^{\frac{1}{2}} R \Lambda^{-1} R' D^{\frac{1}{2}}$

So $\sum_{i=1}^{k} E\left[e_i' \left(B'Q\Lambda^{-1}Q'S_{xy} - \beta \right) \right]^2 = \beta'(M - I)^2 \beta$.

But $M^2 \leq B'Q\Lambda^{-1}Q'\Omega Q\Lambda^{-1}Q'B = M$ so $(M - I)^2 \leq (I - M) \Rightarrow M \leq I_k$

Therefore $\sum_{i=1}^{k} E\left[e_i' \left(B'Q\Lambda^{-1}Q'S_{xy} - \beta \right) \right]^2 \leq \beta'(M - I)^2 \beta \leq \beta'\beta \left[\lambda_{max}(I - M) \right]^2$

$$= \beta'\beta \left[\lambda_{max} \left(\Lambda^{-\frac{1}{2}} Q'\Psi Q \Lambda^{-\frac{1}{2}} \right) \right]^2 \text{ from Lemma 5}$$

$$\leq \beta'\beta \left(\frac{\sigma^2}{\lambda_k} \right)^2 \qquad \text{(A9)}$$

Combining (A6)-(A9) yields

$$\frac{\left(E|e_a| \right)^2}{\|\beta\|^2} \leq k \left(\left(\frac{\sigma^2}{\lambda_k} \right)^2 + \frac{2k}{T} \frac{\sigma_y^{(0)2} \gamma}{\|\beta\|^2} \right)$$

Noting that $\rho = \frac{\sigma^2}{\lambda_k}$ and $\frac{1}{r^2} = \frac{\sigma_y^2}{\|\beta\|^2}$, combining the above results yields the result of the theorem.

Proof of Theorem 2: Since the elements of diag(Λ) are distinct, it follows from the implicit function theorem that there exist differentiable functions $h : N(\Omega) \longrightarrow \mathbb{R}^{k \times l}$ and $g : N(\Omega) \longrightarrow \mathbb{R}^{p \times k}$ where $N(\Omega)$ is a neighborhood of Ω, such that $\Lambda = h(\Omega)$ and $Q = g(\Omega)$. Since h and g are differentiable, they are continuous. Assume that Q is subject to an appropriate normalization and consider the sequence $\{S_T\}_{T=p}^{\infty}$ where $S_T = \frac{1}{T} \sum_{t=1}^{T} x_t x_t'$. Since x_t is Gaussian and g and h are continuous, the Slutsky Theorem gives $\hat{\Lambda} = h(S_T) \xrightarrow{p} \Lambda$ and $\hat{Q} = g(S_T) \xrightarrow{p} Q$ as $T \longrightarrow \infty$. Thus $\hat{e}_h = \hat{\beta}_Q' \hat{f}_{Qh} - \beta_Q' f_{Qh} = S_{yx}' \left(\hat{Q}\hat{\Lambda}^{-1}\hat{Q}' - Q\Lambda^{-1}Q' \right) x_h \xrightarrow{p} 0$ as $T \longrightarrow \infty$.

References

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

Breusch, T. S., (1986) "Hypothesis Testing in Unidentified Models" *Review of Economic Studies* p.635-651.

Camba-Mendez, G., Kapetanios, G., Smith, R. J., and Weale, M. R., (2001) "An Automatic Leading Indicator of Economic Activity: Forecasting GDP Growth for European Countries" *Econometrics Journal* 4(1) p.S56-S90.

Chamberlain, G. (1983) "Funds, factors, and diversification in arbitrage pricing models" *Econometrica* 51, p.1281-1304.

Chamberlain, G. and Rothchild, M. (1983) "Arbitrage, factor structure, and mean-variance analysis in large asset markets" *Econometrica* 51, p.1305-1324.

Dhrymes, P.J., Friend, I, and Gultekin, N.B. (1984) "A critical reexamination of the empirical evidence on the arbitrage pricing theory" The Journal of Finance, 34(2), June, p.323-350.

Forni, M., Hallin, M., Lippi, M. and Reichlin, L. (2000) "The Generalized Dynamic Factor Model: Identification and Estimation" *The Review of Economics and Statistics* 82(4), p.540-554.

Heaton, C. and Solo, V. (2002) "Identification and Estimation of Causal Factor Models of Stationary Time Series" *Macquarie Economics Research Papers* Macquarie University: Sydney, March.

Jöreskog, K. G. (1967). "Some Contributions to Maximum Likelihood Factor Analysis." *Psychometrica* 32: 443-482.

Kapetanios, G., (2002) "Factor Analysis Using Subspace Factor Models: Some Theoretical Results and an Application to UK Inflation Forecasting" Working Paper No.466, Department of Economics, Queen Mary, University of London.

Lawley, D. N. and A. E. Maxwell (1971). *Factor Analysis as a Statistical Method*. Elsevier: New York.

Magnus, J. R., and Neudecker, H. (1991) *Matrix Differential Calculus with Applications in Statistics and Econometrics* John Wiley and Sons: New York.

Poskitt, D. S. and Tremayne, A. R., (1981) "An Approach to Testing Linear Time Series Models" *Annals of Statistics* 9(5), p.974-986.

Schneeweiss, H. (1997) "Factors and Principal Components in the Near Spherical Case" *Multivariate Behavioural Research* 32(4), p.375-401.

Schneeweiss, H. and Mathes, H. (1995) "Factor Analysis and Principal Components" *Journal of Multivariate Analysis* 55(1), p.105-124.

Stock J.H. and Watson M.W. (1998) "Diffusion Indexes" *NBER Working Paper W6702*

Watson, M. W. and R. F. Engle (1983). "Alternative Algorithms for Estimation of Dynamic MIMIC, Factor, and Time Varying Coefficient Regression Models." *Journal of Econometrics* 23: 385-400.