Principal Components and Factor Analysis. A Comparative Study.

Travaglini, Guido

Università degli Studi di Roma "La Sapienza"

13 October 2011

Principal Components and Factor Analysis. A Comparative Study.

Guido Travaglini, Università "La Sapienza", Roma, Italy.

Second Draft, October 13, 2011.

JEL classification: C01, C02, C52.

Keywords: Principal Components, Factor Analysis, Matrix Norm.

Abstract

A comparison between Principal Component Analysis (PCA) and Factor Analysis (FA) is performed both theoretically and empirically for a random matrix $X:(n\times p)$, where n is the number of observations and both coordinates may be very large. The comparison surveys the asymptotic properties of the factor scores, of the singular values and of all other elements involved, as well as the characteristics of the methods utilized for detecting the true dimension of *X*. In particular, the norms of the FA scores, whichever their number, and the norms of their covariance matrix are shown to be always smaller and to decay faster as $n\to\infty$. This causes the FA scores, when utilized as regressors and/or instruments, to produce more efficient slope estimators in instrumental variable estimation. Moreover, as compared to PCA, the FA scores and factors exhibit a higher degree of consistency because the difference between the estimated and their true counterparts is smaller, and so is also the corresponding variance. Finally, FA usually selects a much less encumbering number of scores than PCA, greatly facilitating the search and identification of the common components of *X*.

1. Introduction

Principal Component Analysis (PCA) is a widely used method for dimension reduction of large databases in the presence of collinearity. It dates back to the works of Pearson (1901) and has been extensively analyzed along several decades (Anderson, 1958, 1963; Cattell, 1966; Jolliffe, 1982, 2002). Factor Analysis (FA) is a construct that extends the features of PCA to a stochastic environment, and stems from the contribution of several authors especially in more recent times (Anderson and Rubin, 1956; Stock and Watson, 2002, 2005; Bai 2003; Bai and Ng, 2002, 2006, 2007, 2008, 2010).

More recent approaches to either method, e.g. Supervised PCA initiated by Bair et al. (2006), and Dynamic FA pioneered by Sargent and Sims (1977), for ease of treatment are not considered here, but both authorships supply considerable amounts of proofs as to efficiency and consistency issues (e.g. Forni et al., 2004; Kapeitanos and Marcellino, 2007).

PCA and FA share the search for a common structure characterized by few common components, usually known as "scores" that determine the observed variables contained in matrix *X*. However, the two methods differ on the characterization of the scores as well as on the technique adopted for selecting their true number.

In PCA the scores are the orthogonalized principal components obtained through rotation, while in FA the scores are latent variables determined by unobserved factors and loadings which involve idiosyncratic error terms. The dimension reduction of *X* implemented by each method produces a set of fewer homogenous variables – the true scores – which contain most of the model's information.

In addition, the true number of scores in FA is determined by means of formal statistical procedures different from the classical methods applied in PCA, like screeplot inspection or eigenvalue relative magnitudes. The FA procedures consist of information criteria addressed at detecting, within a prespecified range, the minimum variance of the error terms plus a penalty for overfitting.

The main features of the FA and of the PCA models are examined in Sect. 2, together with some proofs regarding the relative magnitude of their respective factor scores. The asymptotic properties of all of the components of the models and of the singular values are examined in Sect. 3, while Sect. 4 offers a comparative overview of the methods utilized for detecting the true number of scores. Sect. 5 concludes by showing the results achieved in the previous Sections.

Briefly, the conclusions report of the advantages in adopting FA as opposed to PCA for component evaluation and/or instrumental variable estimation purposes. Under FA, the scores are in fact shown to produce more efficient slope estimators when utilized as regressors and/or instruments. Together with the factors they also exhibit a higher degree of consistency even for large sample dimensions. Finally under FA, dimension reduction is definitely more stringent, greatly facilitating the search and identification of the common components of the available dataset.

2. The FA and PCA models and factor scores

In this Section we explain the FA and the PCA models together with the determination of the matrix of their scores. From the Abstract, the reader is reminded that the time and space domains of the data matrix X are respectively represented by the fields of integers: $1,..., n < \infty$, and $1,..., p < \infty$, where $n <=> p$. In addition, for any n and p, the Euclidean norm of a random vector or matrix $A:(n \times p)$ is denoted as $\|A\|$, which corresponds to its maximum singular value $s_{\max}$. Similarly for the covariance matrix $AA':(n \times n)$, whose Euclidean norm is denoted as $\|AA'\|$. It is worth remembering that the singular values of A correspond to the square-rooted eigenvalues (or singular values) of AA'.

2.1. The FA model.

The standard FA model may be expressed as follows

1) $$X = F\Lambda + e$$

where $X:(n \times p)$, $F:(n \times n)$, $\Lambda:(n \times p)$ and $e:(n \times p)$. Matrix X is a user-supplied dataset henceforth posited to be I(0), and F are the estimated factors, Λ are the estimated factor loadings and e the idiosyncratic error matrix. Matrices F and Λ are generically known as the components of eq. (1).

Assumption 1. The major assumptions regarding eq. (1) are:
i) total independence between factors and loadings on the one hand and the idiosyncratic error on the other, i.e. $\mathrm{E}(F'e) = 0$ and $\mathrm{E}(e\Lambda) = 0$;

ii) the error matrix is homoscedastic, $e \sim IID(0, \Sigma_e)$, and is also not autocorrelated, namely, $\mathrm{E}(e_j e_s') = 0$, $\forall j, s \in n; j \neq s$;

iii) the errors are stationary, i.e. the error matrix covariance is bounded from above, namely, $\|\mathrm{E}(e_j e_j')\| \ll \infty$, $\forall j \in n$;

iv) the statistical distribution of *F*, for any *n* and *p* and for $X \sim NID(\mathbf{0,1})$, is $F \sim NID(\mathbf{0,1})$, where **0, 1** are $n \times p$ matrices of zeros and ones.
Assumptions 1. (*i*)-(*iii*) are typical of classical FA (Anderson, 1984) where in general *p* is fixed and much smaller than *n* and which estimates the factor loadings by Maximum Likelihood of the equation system

$$X_{ji} = \lambda_i F_j + \varepsilon_{ji};\ j \in n,\ i \in p\,.$$

Recently, the efficiency and consistency of this method has been seriously questioned for panels characterized by a large *p* (e.g. Bai, 2003).

Assumption 1. (*iv*) regarding the factors follows from the distributional properties assumed to hold for the dataset *X*. A similar distribution regards the factor loadings Λ, as will be shown shortly.
For $n < p$, which is the data dimensional setting utilized henceforth, let the covariance matrix of the non-error part of eq. (1) be

2) $$\Sigma = XX'/np : (n \times n)$$

whose spectral decomposition representation obtained by applying SVD is $\Sigma = FSF'$, where $F'F/n = I_n$ an identity matrix of size $n \times n$, while $S:(n \times n)$ is the diagonal matrix of real and positive singular values or eigenvalues $s_j \ll 1,\ \forall j \in n$ in descending order (Bai and Ng, 2002, 2010), which are bounded from above by virtue of the Marčenko-Pastur distribution law (Sect. 3.1). In particular, it should be noted that $\sum_{j=1}^{n} s_i = 1$ (Onatski, 2009).

By construction, given $r < p$ the size of the true FA scores or equivalently of the factors, obtained by a method to be later described (Sect. 4), we have

3) $$\hat{C} = \hat{F}\hat{\Lambda} \equiv \hat{F}\hat{F}'\hat{X}/n$$

where $\hat{C}:(n \times r)$ is a linear transformation of $\hat{X}:(n \times r) \subset X$, and where $\hat{F}:(n \times r) = n^{1/2}F,\ \hat{F} \subset F$ are the true factors and $\hat{\Lambda}:(r \times r),\ \hat{\Lambda} \subset \Lambda$ are the corresponding true factor loadings. By construction, we also have $\hat{S} \subset S$ where $\hat{S}:(r \times r) = \hat{\Lambda}\hat{\Lambda}'/p$ is a diagonal matrix and $\left\|\hat{S}\right\| = s_{\max} < 1$. Given Assumption 1. (*iv*) and eq. (3), there follows that $\hat{F}_j \sim NID(0,1)$, $\bar{\Lambda} \sim IID(0,\bar{S})$ which means that $(rn)^{-1}\sum_{j=1}^{n}\sum_{i=1}^{r}(\lambda_{ij} - \bar{\lambda}_j) = r^{-1}\sum_{i=1}^{r} s_j$, and finally that $\hat{C} \sim IID(0,\hat{\Delta})$, where the covariance matrix $\hat{\Delta}$ is such that $1 > \left\|\hat{\Delta}\right\| > \left\|\hat{S}\right\| > 0$.

From eq. (1) there follows that $X - \hat{C} = \hat{e}$, where $\hat{e}$ is the empirical estimate of *e* with given variance and $\hat{C}:(n \times r)$ is the true factor score (or common component) matrix. For $r < p$ we have $X \neq \hat{C}$, but for $r \to p$ we have $\hat{C} \to X$ and such that for fixed *n* at $r = p$ the reported matrices are identical. In practice, the linear transformation process fades away as the true number of factors equals the column number of *X*.

For $n > p$, $X'X:(p \times p) = \breve{F}\breve{S}\breve{F}'$ where $\breve{S}:(p \times p)$ is a diagonal matrix of singular values with the same properties as those of S defined above, and $\breve{F}:(p \times p) \neq F$. The corresponding factor score matrix is $\breve{C}:(n \times r) \equiv \hat{C}$ and, identical to above, for $r \to p$, $\hat{C} \to X$.

In both cases, and within the context of a stochastic setting, the true factors and factor loadings $\hat{F}$, $\hat{\Lambda}$ cannot be separately identified without previous knowledge of their consistently estimated counterparts F, Λ, and a specific transformation matrix $H:(r\times r)$ is required to produce the appropriate estimator *F* of $\hat{F}H$ and Λ of $\hat{\Lambda}(H)^{-1}$, as described in Theorem 1 of Bai and Ng (2002), in Bai (2003) and in Bai and Ng (2008), as well as in Stock and Watson (2005). A succinct explanation of this identification problem will be provided in Sect. 3.3. However, the true scores $\hat{C}$ are uniquely identified and consistently estimated.

2.2. The PCA model

The classical PCA true scores, denoted as $\tilde{C}$, are obtained by a SVD procedure different from that exhibited in eq. (2). In fact we have the following spectral decomposition of *X* expressed as:

4) $$X = FDB'$$

where *F* is provided above, while $D:(n\times p)$ is a diagonal matrix of eigenvalues or singular values, and $B:(p\times p)$ is a unitary matrix such that $B'B = I_n$.

If we let eq. (2) utilize the definition of *X* in eq. (4) for comparative purposes, we have

$$\Sigma = FDB'BD'F'/np$$

whereby ultimately $\tilde{D}\tilde{D}'/np = \hat{S}$ and $\tilde{D} = \left(np\hat{S}\right)^{1/2}$ such that $\left\|\tilde{D}\right\| > \left\|\hat{S}\right\|,\ \forall n, p$. In other words, the elements of $\tilde{D}$ denoted as $d_j,\ j\in n$, by construction supersede all of the elements of $\hat{S}$ because $1 > d_j = s_j^{1/2},\ \forall j\in n$ by virtue of the relationship between the singular values of $A:(n\times p)$ to the eigenvalues of its covariance matrix $AA':(n\times n)$.

For $r < p$ under the assumption that the number of both true scores is identical, we have $\tilde{D} = D:(r\times r)$ and thus

5) $$\tilde{C} = \hat{F}\tilde{D}$$

where $\tilde{C}:(n\times r)$ is the matrix of the true PCA scores, under the assumption, later on to be relaxed, that *r* is optimally detected and is the same as that of the FA model.

There are significant differences between the true factor scores $\hat{C}$ and the true PCA scores $\tilde{C}$. By consequence, alternate use among them for empirical research is not innocuous with regard to end results (e.g. Bernanke and Boivin, 2003, Bernanke et al., 2005; Bai and Ng, 2006; Kapeitanos and Marcellino, 2007; Forni and Gambetti, 2008). In practice, use of either as regressors or instrumental variables may lead to sizably different parameter estimations. In addition, serious consistency issues concerning PCA in the presence of large *n* and *p* have been raised by some authors (Bai and Ng, 2002).

By appealing to eqs. (3) and (5), the following theorems illustrate the mean and covariance features of $\hat{C}$ as compared to those of $\tilde{C}$, after assuming they are both stationary processes originated from X an I(0) time series. Then we have:

Theorem 1. For a given dataset $X = \mathrm{I}(0)$ and for any n and p, the true PCA scores $\tilde{C}$ are greater than the true FA scores $\hat{C}$, namely, $\|\hat{C}\| < \|\tilde{C}\|$. $\|\hat{\Delta}\| < \|\tilde{\Delta}\|$

Proof: given that $\hat{C} = \hat{F}\hat{F}'\hat{X}/n$ and $\tilde{C} = \hat{F}\tilde{D}$, there follows by the Cauchy-Schwarz Inequality of the inner product of two independent sequences that $\|\hat{C}\| \le \|\hat{F}\hat{F}'/n\| \cdot \|\hat{X}\|$ and that $\|\tilde{C}\| \le \|n^{-1/2}\hat{F}\| \cdot \|\tilde{D}\|$. From eq. (3) $XX'/np = F\tilde{D}B'B\tilde{D}'F'/np$, whereby $\|XX'\| = \|D'D/np\|$, since $\|F\| = \|B\| = 1$. By consequence we have $\|\tilde{D}\| = \|X\| > \|\hat{X}\|$, and since $\|\hat{F}\hat{F}'/n\| = 1$, there immediately follows that $\|\hat{C}\| < \|\tilde{C}\|$. This proves that the true PCA scores are always larger than the true factor scores.

Theorem 2. In addition to what expressed in Theorem 1, $\tilde{C}$ are orthogonal by construction while $\hat{C}$ are not orthogonal. Moreover, for any n and p the covariance of the former, $\tilde{\Delta} = \left(\tilde{C}'\tilde{C}\right)/np$, is larger than the covariance of the latter, $\hat{\Delta} = \hat{C}'\hat{C}/np$, as shown in Sect. 2.1.

Proof: let the $r \times r$ covariance matrices be rewritten as follows

$$\tilde{\Delta} = \tilde{D}'\hat{F}'\hat{F}\tilde{D}/np = \tilde{D}'\tilde{D}/n\,p$$

and

$$\hat{\Delta} = \hat{X}'\hat{F}\hat{F}'\hat{F}\hat{F}'\hat{X}/np = \hat{X}'\hat{F}\hat{F}'\hat{X}/n\,p.$$

Since $\tilde{D}$ is a diagonal matrix by virtue of eq. (4), so is $\tilde{D}'\tilde{D}$ and any linear transform thereof when operated by a scalar like *np*, including $\tilde{\Delta}$. This proves the orthogonality of $\tilde{C}$.

This orthogonality may also be easily proven by the fact that $\|\hat{\Delta}\| < \|\tilde{\Delta}\|$ where $\tilde{\Delta}_{jj} > 0,\ \Delta_{ij} = 0;\ i, j \in r$, namely, the off-diagonals of the PCA covariance are all zero, which proves by construction the orthogonality of its scores.

After recalling from eqs. (1) and (4) that the FA and the PCA models share exactly the same factors, in order to prove that the first covariance matrix is larger than the second we observe that their respective norms are $\|\tilde{D}'\tilde{D}/np\| = \|\tilde{D}\tilde{D}'/np\|$ and $\|\hat{X}'\hat{F}\hat{F}'\hat{X}/np\|$. Then we need to prove that $\|\tilde{D}\| > \|\hat{X}'\hat{F}\|$. From eq. (4), we have $\|F\| = \|\hat{F}\| = \|B\| = 1$ and $\|X\| = \|D\|$, but $\|D\| = \|\tilde{D}\| > \|\hat{X}\|$ then $\|\tilde{D}\| > \|\hat{X}'\hat{F}/n\|$ whereby there immediately follows that $\tilde{\Delta} > \hat{\Delta}$.

Since the FA scores are proven to be asymptotically smaller in magnitude than their PCA counterparts, they are definitely preferable in applied work when utilized as regressors and/or instruments in large-sample instrumental-variable estimation procedures such as the Generalized Method of Moments (GMM) (Hansen, 1982; Newey and Windmeijer, 2009), because they generate smaller Jacobians and produce by consequence more efficient slope parameters (Kapeitanos and Marcellino, 2007).

From the inequality $\left\|\hat{C}\right\| < \left\|\tilde{C}\right\|$ there automatically follows that $\left\|\hat{C}u\right\| < \left\|\tilde{C}u\right\|$, where u is any nonzero-element stochastic vector or matrix of the same length as that of the scores.

Let the standard single-equation model of length n be $y = X\beta + u$ where y and X are the appropriately stacked arguments and z the selected instrument matrix. Then $\beta:(r\times 1)$ is a (consistent) OLS estimator and u is defined as the first-stage disturbance vector. After defining β_0 as the population parameter value, and $z = \hat{C}$ or $z = \tilde{C}$, the Jacobian matrix is

$$G:(r\times r) = -z'X_j,\ j\in r$$

and the asymptotic covariance of the GMM-estimated parameter vector is

$$n^{\frac{1}{2}}(\beta_{GMM} - \beta_0) = \left(G'\Omega^{-1}G\right)^{-1} \qquad (6)$$

where

$$\beta_{GMM} = \left(G'\Omega^{-1}G\right)^{-1} G'\Omega^{-1}z'y$$

(Hansen, 1982; Newey and Windmeijer, 2009) and $\Omega:(r\times r) = w'w/n$ is the weight matrix, and $w:(n\times r) = \hat{C}u$ or $\tilde{C}u$ are the sample moments. Simple algebra of eq. (6) implies that for given G the following is true:

$$\underset{\|\Omega\|\to 0}{Lim}\left\|\left(G'\Omega^{-1}G\right)^{-1}\right\| = 0$$

namely, the smaller the norm of the weight matrix the smaller is the norm of the covariance matrix of the GMM estimated parameter vector. By consequence:

$$\underset{\|\Omega\|\to 0}{Lim}\ n^{\frac{1}{2}}(\beta_{GMM} - \beta_0) = 0 \qquad (7)$$

which holds insofar as $\left\|\hat{C}u\right\| < \left\|\tilde{C}u\right\|$, as assumed above.

3. Asymptotic properties of errors, factors, factor loadings and factor scores

Knowledge of the asymptotic behavior of the true PCA and FA scores, respectively $\tilde{C}$ and $\hat{C}$, and of the components of the latter, namely $\hat{F}$ and $\hat{\Lambda}$, is important especially in large model settings. Equally important is knowledge of the distribution properties of the matrix e of the idiosyncratic errors, inextricably tied to the components and to the dataset. The limit distribution of most of these variables is provided by Bai (2003) and of Bai and Ng (2002, 2008). Other properties remain to be analyzed with particular emphasis on convergence/divergence issues in an asymptotic setting. Illustration of such properties requires at first a few definitions and some basics in random matrix theory (RMT).

3.1. Theoretical aspects of RMT: matrix norms and singular values

For the data dimensional feature here utilized, which is $n < p$, the aspect ratio $y = p/n$ is such that $1 < y < \infty$ and $\underset{n\to\infty}{Lim}\, y = 1$ with p fixed, and $\underset{p\to\infty}{Lim}\, y = \infty$ with n fixed. Unless the specifications $n, p \to \infty$ are required for clarity purposes, y will be henceforth utilized.

Consider now the random vector or matrix A introduced in the previous Section, whose elements are in particular assumed to be N.I.D. real numbers with zero mean and finite fourth moments. For $n = p$ the matrix is real asymmetric. Consider also the matrix $AA':(n\times n)$ which in turn is always real symmetric.

Proposition 1. For the sequence $y = 1,...,\infty$, the matrix norm of A and of AA', for any $y \in [1,\infty]$, is respectively denoted as $\|A\|_y$ and $\|AA'\|_y$. Then the following apply:

i) $\|A\|_y$ and $\|AA'\|_y$ form each a Cauchy sequence respectively denoted as $(a_y) = \{\|A\|_y\}$ or $(a_y) = \{\|AA'\|_y\}$, whose elements $\{a_i\}$ for $i \in [1,\infty]$ belong to the Banach space of real numbers;

ii) both sequences are monotone and may be stationary (if $a_i \lesseqgtr a_{i+1},\ \forall i$) or increasing (if $a_i \le a_{i+1},\ \forall i$) or decreasing ($a_i \ge a_{i+1},\ \forall i$). In the last two cases, (a_y) is asymptotically weakly convergent in L^2 to a right-endpoint upper or lower bound, respectively;

iii) the lower bound is $lb = \inf_{1\le y\le\infty}(a_y)$, such that $\{a_i\} \ge lb,\ \forall i$, and the sequence exhibits negative-valued first derivatives for $a_i - a_{i+1} \approx 0,\ \forall i$;

iv) the upper bound is $ub = \sup_{1\le y\le\infty}(a_y)$, such that $\{a_i\} \le ub,\ \forall i$, and the sequence exhibits positive-valued first derivatives for $a_i - a_{i+1} \approx 0,\ \forall i$;

v) by the uniform boundedness principle, points (*iii*) and (*iv*) respectively imply

$$\underset{y\to 1,\, y\to\infty}{P\,\mathrm{Lim}}\left(\left(a_y\right) = lb\right) = 1 \quad \text{and} \quad \underset{y\to 1,\, y\to\infty}{P\,\mathrm{Lim}}\left(\left(a_y\right) = ub\right) = 1;$$

vi) the rate of convergence in probability of (a_y) for p fixed and $n \to \infty$ is $O_p(n^{\alpha})$, where α is an integer or a fraction, and $\alpha < 0$ if $\exists\, lb$ while $\alpha > 0$ if $\exists\, ub$. Finally, $(a_y) = O_p(1)$ if $\alpha = 0$. The same reasoning can be extended to the case of n fixed and $p \rightarrow \infty$ whereby we have $\left(a_y\right) = O_p\left(p^{\alpha}\right)$ and related values of α.

Proposition 1 implies that, for changing *y*, the norms of the two matrices change and form a sequence of real numbers that may stay stationary or rise toward an upper bound or decrease toward a lower bound. The rates of convergence in probability vary depending on the existence and on the magnitude of the bounds. The sequence is monotone and continuous and obeys the fundamental laws of such functions (see e.g. Sokal, 2010 and the references cited therein). The upper- or lower-bound norms are supplied by recent research in RMT which establishes the limit distribution of maximum singular values $s_{\max}$ of the reported random matrices.

If we let $s_{\max} = \left\| A/n \right\|$ be the Euclidean norm of A/n, for $n \rightarrow \infty$ (or $y \rightarrow 1$), $s_{\max}$ exhibits an asymptotic behavior that obeys the limit Marčenko-Pastur (MP) distribution law whereby $s_{\max} \sim \left(\sqrt{n} + \sqrt{p}\right)/n$ (Johnstone, 2001; Vershynin, 2011). If instead we let $s_{\max} = \left\| AA'/n \right\|$ be the Euclidean norm of the covariance matrix AA'/n, then under mild conditions regarding its distribution and for $n \rightarrow \infty$ we have $s_{\max} \sim \left(\sqrt{n} + \sqrt{p}\right)^2 / n = \left(1 + \sqrt{y}\right)^2$ (Geman, 1980). Needless to say, if the denominator of *A* or of AA' is *np* rather than *n*, the values of $s_{\max}$ are computed accordingly. This happens for example when computing the upper bound of the singular values of matrix Σ in eq. (2), which is $s_{\max} \sim \left(\sqrt{n} + \sqrt{p}\right)^2 / np$.

In addition, for $n \rightarrow \infty$, the rate of convergence in probability of the empirical spectral distribution (ESD) of $s_{\max}$ toward its corresponding limit spectral distribution (LSD) is proven to be $O_p\left(n^{-1/4}\right)$ (Bai,1993; Bai et al., 2000; Rudelson and Vershynin, 2010).

Proposition 2. Any data-based sequence $\left(a_y\right)$ may be empirically estimated over the interval $1 < y < \infty$ and tested for the parameter α implied by the function $\left(n^{\alpha}\right)$ shown in Proposition 1. (*vi*) and by consequence for their convergence in probability implied by Proposition 1.(*v*). If $\left(a_y\right)$ is a sequence of matrix norms, the estimation process may better be defined as Empirical Norm Estimation (ENE).

The estimation may be carried through by nonlinear least squares applied to the following equation

$$\left(a_j\right) = j^{\alpha} + \upsilon_j, \; j \in n$$

where $\mathrm{E}\left(\upsilon_j\right) = 0$ with finite variance. If the function $\left(p^{\alpha}\right)$ is utilized, the same equation may be estimated with the index $i \in p$ replacing the index *j*.

One important application of computing the ENE regards the series included in eqs. (1) and (4). If X is a n-sized random *NID* vector, for $y \to 1$, $\left\| X \right\|_y = O_p\left(n^{\alpha}\right)$ where $1/3 \le \alpha \le 2/3$. At the other extreme, for a very large fixed p and for $n \to \infty$ we have $\left\| X \right\|_y = O_p\left(1\right)$, since

$$\left\| \hat{F}\hat{F}'/n \right\|_y = \left\| \hat{F}/\sqrt{n} \right\|_y \equiv 1,\ \forall y \in [1,\infty].$$

Other applications regard the scores and their components, as shown in the following Sections.

3.2. Asymptotic properties of errors

Under conditions of normality about the dataset X and for n fixed and $p \to \infty$ $\left(\text{or } y \to \infty\right)$, the two following Assumptions hold:

Assumption 2. The matrix e of the errors exhibits a (close-to) diagonal covariance matrix $\Omega : \left(p \times p\right) = \mathrm{E}\left(ee'\right)$ under the hypothesis of asymptotic orthogonality;

Assumption 3. The covariance matrix, diagonal or not, exhibits over $y \in \left[1,\infty\right]$ the normed sequence $\left\| \Omega \right\|_y = O_p\left(p^{\alpha}\right)$, $\alpha \le 0$, which is a decaying process or at least $O_p\left(1\right)$ under the hypothesis of asymptotic homoscedasticity.

In other words: $\mathrm{E}\left(e_{ji}e_{jk}\right) = 0;\ i,k \in p;\ \forall j \in n$ and $n^{-1}\sum_{j=1}^{n} e_{ji}^2 = \sigma_i^2;\ \forall i \in p,\ \forall j \in n$, where σ_i^2 is the i.th column mean for any given $n \ge 2$ (Bai, 2003). In other words, for fixed n and $p \to \infty$, the off-diagonals of matrix Ω tend to zero while the mean variance of the errors converges to a constant unit value under normality, namely, $p^{-1}\sum_{i=1}^{p} \sigma_i^2 = \sigma^2$, or also $\sigma^2 = \left(np\right)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{p} \hat{e}_{ji}^2$ where $\sigma^2 \le 1$.

3.3. Asymptotic properties of factors and factor loadings

Two more assumptions hold in probability regarding the true factors and the true factor loadings (Assumptions A and B in Bai, 2003; Bai and Ng, 2002, 2008, 2010):

Assumption 4. $\hat{F}$ is stochastic such that $\mathrm{E}\left(\left\| \hat{F} \right\|^4\right) \ll \infty$ and the true factor covariance matrix is $n^{-1}\sum_{i=1}^{n} \hat{F}_i\hat{F}_i' \xrightarrow{p} \Sigma_{\hat{F}}$, where $\Sigma_{\hat{F}} : \left(r \times r\right)$ is symmetric p.d.;

Assumption 5. $\hat{\Lambda}$ is deterministic such that $\mathrm{E}\left(\left\| \hat{\Lambda} \right\|\right) \ll \infty$ or stochastic such that $\mathrm{E}\left(\left\| \hat{\Lambda} \right\|^4\right) \ll \infty$, and the true factor loading covariance matrix is $\hat{\Lambda}\hat{\Lambda}'/p \xrightarrow{p} \Sigma_{\hat{\Lambda}}$, where $\Sigma_{\hat{\Lambda}} : \left(r \times r\right)$ is symmetric p.d.

Assumption 6. Following on from the two previous Assumptions, we have $\hat{\Sigma} = \Sigma_{\hat{\Lambda}}^{1/2}\Sigma_{\hat{F}}\Sigma_{\hat{\Lambda}}^{1/2}$ which is symmetric p.d. of size $r \times r$ and whose eigenvalue diagonal matrix, defined as V, is the asymptotic counterpart of $\hat{S}$.

The two matrices $\Sigma_{\hat{F}}$ and $\Sigma_{\hat{\Lambda}}$ provide consistent estimates of the factors and of the factor loadings and are instrumental to the analysis of the asymptotic differences between the estimated and the true factors, i.e. the asymptotic approximation of F toward $\hat{F}$ introduced in Section 2, and fully described by Bai and Bai and Ng in the works cited herein.

The authors show that $\hat{F}$ and $\hat{\Lambda}$ cannot be identified without previous knowledge of their estimated counterparts F and Λ, and that the trivial solution $F - n^{-1/2}\hat{F} \equiv 0$, where $\hat{F} = n^{1/2}F$, must be avoided. However, as noted in Sect. 2, the true scores are always uniquely identified and consistently estimated.

Theorem 3. Let $H:(r \times r)$ be the transformation matrix of $\hat{F}$ into F such that $\hat{F} = FH$ and of $\hat{\Lambda}$ into Λ such that $\hat{\Lambda} = \Lambda H^{-1}$, as shown in Bai (2003), Bai and Ng(2002, 2008, 2010) and Stock and Watson (2002, 2005). Let also $\hat{F} = n^{1/2}F$ as defined in Sect.2.1. Under these circumstances, neither F nor Λ can be identified if H is trivially taken to be the diagonal matrix $n^{1/2}I_r$ with $I_r:(r \times r)$ an identity matrix. By consequence, for identification to be achieved H must be sparse and invertible and, in its most generic form, made up of a deterministic and of a stochastic component.

Proof: Assume that

8) $$H = n^{1/2}I_r + \Theta_r,\ \Theta_r \sim IID(0, \Xi)$$

where I_r and $\Theta_r:(r \times r)$ respectively are the deterministic and the stochastic component, whose covariance is Ξ. Then we may have $\hat{F} = FH$ where the deterministic component is not sufficient for identification, but the stochastic component is crucial. Then $\hat{F}$ is a linear stochastic rotation of F where identification is achieved by means of $H \neq n^{1/2}I_r$.

There are several different possibilities of computing matrix H, either theoretically or empirically. Only two, however, are considered here. The first is

9) $$H = V^{-1}\left(\hat{F}\hat{F}'/n\right)\left(\hat{\Lambda}\hat{\Lambda}'/p\right)$$

where V is defined in Assumption 6 and the first bracketed expression originates from appropriate manipulation of the matrix Σ, as shown by Bai (Appendix A, 2003). Since the limits of the first and of the second bracketed expressions in eq. (9) respectively are $\Sigma_{\hat{F}}$ and $\Sigma_{\hat{\Lambda}}$, while the limit of V is $\Sigma_{\hat{\Lambda}}$, we have $H = \Sigma_{\hat{F}}$. However, since $\underset{n\to\infty}{Lim}\ \Sigma_{\hat{F}} = I_r$, then $\underset{n\to\infty}{Lim}\, H = I_r$ (Bai and Ng, 2010).

The second way of computing H has been proposed by Stock and Watson (2002, 2005) and stems from the OLS model $X = \hat{C}\hat{H} + v$, where $\hat{H}:(r \times r)$ and X, $\hat{C}$ and v are all of size $n \times r$ and the last is a disturbance matrix. The factors and the factor loadings utilized to produce $\hat{C}$ are computed by reduced rank regression. The ensuing matrix $\hat{H} = I_r + \Theta_r$, where Θ_r is defined above and

$\mathrm{E}\left(\hat{H}_{ik}\right)=0;\ i,k\in r;\ i\neq k$. This means that the *i*.th true factor is determined not only by the *i*.th estimated factor with unit slope on the main diagonal of $\hat{H}$ but also by (at least some) other *k* estimated factors.

Whichever the method adopted to determine the matrix *H*, identification of $\hat{F}$ and $\hat{\Lambda}$ directly follows through given knowledge of the estimated values of F and Λ. Their asymptotic properties relative to the corresponding true values can now be treated.

For any *p* fixed and given *r*, the following norm regarding the relationship between estimated and true factors holds (Bai, 2003):

10) $$n^{1/2}\left\|F_j-\hat{F}_jH\right\|^2=O_p(1),\ \forall j\in n$$

where $F_j:(1\times r)$ and $\hat{F}_j:(1\times r)$. Eq. (10) implies that, asymptotically, the squared deviation between $\hat{F}$ and *F* is a stationary process that does not change over the given time span *n*. In other words, the true factors are always sufficiently close to the *r*-sized matrix of the estimated factors, irrespective of the magnitudes of *n* and *p*. In addition, the asymptotic distribution of the factor differences is (Bai, 2003, Theorem 1):

11) $$p^{1/2}\left(F_j-\hat{F}_jH\right)\sim NID\left(0,V^{-1}Q\hat{\Gamma}_jQV^{-1}\right),\ \forall j\in n$$

where $\hat{\Gamma}_j:(1\times n)$ is the estimated mean covariance of $\hat{e}$ and Λ, obtained from the matrix $\hat{e}\Lambda':(n\times r)$ and the invertible matrix $Q:(r\times r)=V^{1/2}\Upsilon\Sigma_\Lambda^{1/2}$, where Υ are the eigenvectors associated to the diagonal matrix V. According to some authors (Connor and Korajczik, 1986), if there is heteroskedasticity and autocorrelation, consistent estimation of $\hat{F}$ for large *n* is unfeasible and thus eq. (10) does not apply. According to other authors (Bai, 2003, Theorem 5) consistent estimation of $\hat{F}$ is still possible via a modification of eqs. (8) and (9), whereby we have

10) $$p^{1/2}\left(F_j-\hat{F}_j\bar{H}'\right)\sim NID\left(0,V^{-1}Q\Gamma QV^{-1}\right),\ \forall j\in n$$

where $\bar{H}=HV\left(V-\sigma^2\right)^{-1}$ and $\Gamma=\mathrm{plim}\left(\hat{\Lambda}'\Omega\hat{\Lambda}\right)$. Obviously, unless $\sigma^2=0$, $\left\|\bar{H}\right\|<\left\|H\right\|$. In practice, the factor matrix deviations for any *n* are zero-mean and have asymptotic variance conditional on the covariance matrices of the factors, of the factor loadings, and of the idiosyncratic error, whichever its characterization.

The asymptotic distribution of the difference between the estimated and the true factor loadings is (Bai, 2003, Theorem 2):

13) $$n^{1/2}\left(\Lambda_i-\hat{\Lambda}_iH^{-1}\right)\sim NID\left(0,(Q')^{-1}\Phi_iQ^{-1}\right),\ \forall i\in p$$

where $\hat{\Lambda}_i$ is the true factor loading matrix for each p and $\Phi_i:(1\times p)$ is defined as the estimated covariance of the mixture of $\hat{e}$ and of the true factors, obtained from the matrix $\hat{e}'\hat{F}:(p\times r)$ and precisely is

$$\Phi_i=(np)^{-1}\sum_{i=1}^{p}\sum_{j=1}^{n}\hat{F}_{ji}^2\hat{e}_{ji}^2 .$$

Several major issues emerge from the reported distributions. As the reader will notice at first, eqs. (11) and (12) are a function of the covariance of the factor loadings and the idiosyncratic error, while eq. (13) is a function of the covariance of the latter and of the factors. In essence, the distribution of each component depends on the mixture of the errors and of the other component. Secondly, The expected covariance of the mixing processes involved is zero everywhere, namely, $\mathrm{E}(\hat{\Gamma}_j)=0,\ \forall j\in n$ and $\mathrm{E}(\Lambda_i)=0,\ \forall i\in p$. Thirdly, the two norms $\left\|F_j-\hat{F}_jH\right\|,\ \forall j\in n$ and $\left\|\Lambda_i-\hat{\Lambda}_iH^{-1}\right\|,\ \forall i\in p$ are computed by ENE to be $O_p(1)$, as demonstrated elsewhere (Bai 2003, Bai and Ng, 2002, 2008). Finally, the estimated asymptotic distributions of eqs. (11) or (12) and (13) respectively, both derived for the entire time and space domain *n* and *p*, are

14) $$F_{\mathrm{H}}=p^{1/2}(F_n-F_nH)\sim NID(0,FF_n/p)$$

and

15) $$\Lambda_{\mathrm{H}}=n^{1/2}\left(\Lambda_n-\hat{\Lambda}_nH^{-1}\right)\sim NID\left(0,\Lambda\Lambda_p/n\right)$$

where $F_H:(n\times r)$, $\Lambda:(p\times r)$, $FF_n=F_H'F_H$ and $\Lambda\Lambda_p=\Lambda_{\mathrm{H}}'\Lambda_{\mathrm{H}}$. The last two are $r\times r$ p.d. covariance matrices – of which the first will be extensively treated shortly – and both share an upper bound as $n\to\infty$ (Chamberlain and Rothschild, 1983; Bai and Ng, 2002) determined by the MP rule.

While eq. (13) does not apply in the PCA model provided by eq. (4) since neither the idiosyncratic error nor the loadings exist, the PCA counterparts of eqs. (11) and (12), if of any interest, may be obtained by similarity reasoning. In fact, eq. (9) may be rewritten as

9.1) $$\tilde{H}=\tilde{V}^{-1}\left(\hat{F}\hat{F}'/n\right)\tilde{D}$$

where $\tilde{V}$ is the diagonal matrix of the eigenvalues of $\tilde{D}^{1/2}\Sigma_{\hat{F}}\tilde{D}^{1/2}$. The asymptotic difference between the estimated and the true PCA factors is then

16) $$F_{\tilde{H}}=p^{1/2}\left(F_n-\hat{F}_n\tilde{H}'\right)\sim NID\left(0,\tilde{F}\tilde{F}_n/p\right)$$

whose $r\times r$ p.d. covariance matrix is $\tilde{F}\tilde{F}_n$, which is comparable with FF_n. Eqs. (14) and (16) are thus respectively the FA and the PCA asymptotic factor estimator distribution functions. It will be shown that significant differences exist between the two and that, as with scores, also the PCA factors asymptotically fare worse than the FA factors.

Theorem 4. Let the Eqs. (14) and (16) be the FA and the PCA asymptotic factor estimator distributions. Then we have $\left\|\tilde{H}\right\| > \left\|H\right\|$ and $\left\|\tilde{F}\tilde{F}_n\right\| > \left\|FF_n\right\|$. This means that under PCA not only the difference between the estimated factors and their true counterparts is larger in absolute terms, but also that its asymptotic variance is larger.

Proof: it is a well-established fact that both $\left\|\tilde{D}\right\| > \left\|\Lambda\Lambda'/p\right\|$ and $\left\|\tilde{V}\right\| > \left\|V\right\|$ whereby, comparing eq. (9) to eq. (9.1), there immediately follows that $\left\|\tilde{H}\right\| > \left\|H\right\|$. The expanded version of each covariance matrix, independent of scaling, respectively is

$$\left(F_n - \hat{F}_n\tilde{H}'\right)'\left(F_n - \hat{F}_n\tilde{H}'\right) = I_r + \tilde{H}\hat{F}_n^{'}\hat{F}_n\tilde{H}' - 2F_n^{'}\hat{F}_n\tilde{H}'$$

and

$$\left(F_n - \hat{F}_n\hat{H}'\right)'\left(F_n - \hat{F}_n\hat{H}'\right) = I_r + \hat{H}\hat{F}_n^{'}\hat{F}_n\hat{H}' - 2F_n^{'}\hat{F}_n\hat{H}'$$

Since $\left\|\left(\tilde{H}\hat{F}'\hat{F} - 2F'\hat{F}\right)\tilde{H}'\right\| > \left\|\left(\hat{H}\hat{F}'\hat{F} - 2F'\hat{F}\right)\hat{H}'\right\|$, there follows immediately that $\left\|\tilde{F}\tilde{F}_n\right\| > \left\|FF_n\right\|$.

In practice, the covariance matrix of the asymptotic distribution of the differences between the estimated and the true factors under the PCA model is always larger (in norm) than the corresponding covariance of the FA model.

3.4. Asymptotic properties of factor scores and of singular values

From Theorem 1 and for $r < p$, it is demonstrated that the true PCA scores are larger in norm than the true FA scores. This is valid for both fixed p, as $y \to 1$, and for n fixed and $y \to \infty$. Formally, eqs. (3) and (5) respectively produce

$$\left\|\hat{C}\right\|_y \le \left\|\hat{F}\hat{F}'/n\right\|_y \cdot \left\|\hat{X}\right\|_y$$

and

$$\left\|\tilde{C}\right\|_y \le \left\|n^{-1/2}\hat{F}\right\|_y \cdot \left\|\tilde{D}\right\|_y$$

such that by Theorem 1 we have $\left\|\hat{C}\right\|_y < \left\|\tilde{C}\right\|_y$. In addition, both $\left\|\hat{C}\right\|_y$ and $\left\|\tilde{C}\right\|_y$ are estimated by ENE to be $O_p(1)$.

For changing y, also the estimated norms of the covariance of the true FA scores and of the true PCA scores differ in convergence probability. The covariance matrices respectively are $\hat{\Delta}$ and $\tilde{\Delta}$ (Sect. 2) and their norms are both decreasing as a function of the sample size, and precisely are $\left\|\hat{\Delta}\right\|_y = O_p\left(n^{-1/2}\right)$ and $\left\|\tilde{\Delta}\right\|_y = O_p\left(n^{-1/3}\right)$. Because the former is steeper than the latter, we can comfortably conclude that the score covariance under FA asymptotically approaches, as $n \to \infty$, its lower bound at a faster rate than that of the score covariance under PCA. For fixed *n* and $p \to \infty$, instead, both covariance norms are $O_p(1)$.

Asymptotically, the singular value matrix $\hat{S}$ associated to the true FA factor scores behaves differently from the singular value matrix $\tilde{D}$ associated to the true PCA scores. In fact, for sufficiently large *p* and $n \to \infty$, we have $\left\|\hat{S}\right\|_y = O_p\left(n^{-1/4}\right)$ while $\left\|\tilde{D}\right\|_y = O_p(1)$. The latter corresponds to the convergence in probability of the norm of *X*, as implied by Theorem 1, while the former corresponds to the reported findings on random covariance matrices pertaining to RMT (Bai, 1993; Bai et al., 2000; Rudelson and Vershynin, 2010).

The implications of these findings are far reaching: the asymptotic behavior for fixed *p* and $n \to \infty$ of the singular values under FA exhibits a rate of decay that supports by virtue of eq. (3) the conclusions achieved in Theorem 1 about the magnitude of the scores $\hat{C}$. In other words, the longer the sample size *n* the smaller the factor scores while the PCA scores remain unprejudiced.

Finally, the difference between the estimated and the true factor scores, for $y \to 1$, is characterized by the following asymptotic distribution

17) $$n^{1/2}\left(C_{ji} - \hat{C}_{ji}\right) \sim NID\left(0, W_{ji}\right);\ \forall j \in n,\ \forall i \in p$$

where $W_{ji} : (n \times p) = \hat{F}\Sigma_{\hat{F}}^{-1}\Phi_i\Sigma_{\hat{F}}^{-1}\hat{F}'/n$ as provided in Theorem 1 by Bai and Ng (2002) and Theorem 3 in Bai (2003). The mean value of this matrix, for all elements enclosed in *n* and *p*, is given by $p^{-1}\sum_{i=1}^{p}\left(n^{-1}\sum_{j=1}^{n}W_{ji}\right) = m_j,\ \forall j \in n$, which decreases for $n \to \infty$ and is $O_p\left(n^{-1/4}\right)$. In practice, the difference between the estimated and the true factor scores exhibits an asymptotically decaying variance.

The asymptotic distribution of the factor scores that matches eqs. (14) and (15) for both $n, p \to \infty$ is an extension of eq. (14), namely,

18) $$n^{1/2}\left(C - \hat{C}\right) \sim NID(0, W)$$

where $W : (r \times r)$. In PCA, eqs. (17) and (18) do not apply because there is no measurable distance between estimated and true scores. Probably the difference between the scores computed by standard PCA methods and the *r* scores obtained under FA, defined as true, may be the best approximation to reproduce eqs. (17) and (18). To save space, while recalling that $\hat{F}$ is common to both PCA and FA, we utilize only the second, namely

19) $$n^{1/2}\left(\tilde{C} - \hat{C}\right) \sim NID\left(0, \tilde{W}\right)$$

where $\tilde{W} : (r \times r)$. It is worth noting that for *p* fixed and $n \to \infty$ the two covariance matrices are found by ENE to be $W,\ \tilde{W} = O_p(1)$.

4. Detecting the true number of scores in the FA and PCA models

Thus far, for expository purposes, the true number of scores $r < p$ has been assumed to be the same for FA and PCA. The assumption is now relaxed since they traditionally differ as to the method utilized for detecting *r*. While modern FA predominantly utilizes formal statistical procedures, PCA sticks to classical methods based on screeplot testing (Cattell, 1966) or on other eigenvalue- or variance-based methods (Jolliffe, 1982, 2002; Draper and Smith 1981; Myers 2000; Loehlin, 2004).

In FA modeling there are two major strands of analysis for the detection of the number of true scores. One method is essentially probabilistic and is based on the search of the minimal estimated variance stemming from the difference between the observed dataset and the estimated factor scores across different options. The other method is a modernization of Cattell's method.

More precisely, the first method is introduced by Bai and Ng (2002) and includes eight formal statistical procedures to compute the true number of scores. There are three so-called Panel Criteria (PC) and three Information Criteria (IC), to which the standard Akaike and Bayesian Information Criteria (AIC and BIC) are added.

The first method extends to the Eigenvalue Ratio Estimator (ERE) by Ahn and Horenstein (2009), while the second method includes the Edge Distribution Estimator (EDE) by Onatski (2009), and the Eigenvalue Ratio Test (ERT) by Ahn and Horenstein (2009).

As for the first method, let $k_{\max}$ be the maximum number of scores admitted, so as to produce the sequence $k = 1,...,k_{\max}$. Take also eq. (1) solved for the disturbances: $e = X - F\Lambda$, which for each $j \in n,\ i \in k_{\max} \ll p$ may be written as follows

20) $$\hat{e}_{ji} = x_{ji} - f_{ji} \cdot \hat{\lambda}_i$$

and define

21) $$V_k = \min\left((np)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{p} \hat{e}_{ji}^2 \right)$$

which is the sum of the squared residuals averaged over both *n* and *p* where the bracketed expression equals σ^2 defined in Assumption 3.

All of the ICs cited by the authors are based on detecting the minimum variance V_k plus a penalty for overfitting within the *k* sequence. Only one IC for ease of space is reported here as it is recommended by most users (Bai and Ng, 2007; Alessi et al., 2009):

22) $$IC_k = \ln\left(V_k\right) + k\left(\frac{n+p}{np}\right)\ln\left(n\right);\ \forall k \in k_{\max}.$$

According to Bai and Ng (Theorem 1, 2002), we have for r^* the population-true number of scores

23) $$r = \arg\min_{1 \le k \le k_{\max}}\left(IC_k\right)$$

and $$\underset{n,p\to\infty}{P\lim}\left(r = r^*\right) = 1$$

and also

$$\underset{n,p\to\infty}{L\mathrm{im}}\left(\left(\frac{n+p}{np}\right)\ln\left(n\right)\right) = \infty$$

An extension of eq. (21) is provided by the following ratio

24) $$ERE_k = \frac{V_{k-1} - V_k}{V_k - V_{k+1}};$$

while the eigenvalue-based method is

25) $$ERT1_k = s_k / s_{k+1};\ ERT2_k = \log\left(s_k\right)/\log\left(s_{k+1}\right)$$

which is a formal version of the screeplot test, wherein s_k is the *k*.th singular value of matrix *S* derived from eq. (2). The true number of scores $r < p$ is detected whenever a maximum is achieved in either of them, namely,

26) $$r = \arg\max_{1 \le k \le k_{\max}}\left(ERE_k\right)$$

and

27) $$r = \arg\max_{1 \le k \le k_{\max}}\left(ERTj_k\right);\ j\text{=}1,\ 2$$

whereby

$$\underset{n,p\to\infty}{P\lim}\left(r = r^*\right) = 1.$$

Eqs. (23), (24) and (25) are shown to produce identical values of r in most Montecarlo experimentations when *n* and the select value of $k_{\max}$ respectively are sufficiently large (Ahn and Horenstein, 2009).

Somewhat different results are obtained with Onatski's EDE which strictly requires nonstationary series, i.e. $X = I(1)$ and is based on the asymptotic distribution of the eigenvalues of its covariance matrix. Some of the eigenvalues are shown by the author to diverge toward infinity, while others – the true eigenvalues $s_j, j \in k_{\max}$ – converge toward a finite value. Therefore their respective first differences tend to infinity and to zero. Let δ be some prespecified threshold value obtained from the distribution of s_j such that $\infty >> \delta > 0$ (Onatski, 2009), then

(28) $$EDE_k = \left|s_k - s_{k+1}\right| - \delta,\ \forall k \in k_{\max}$$

where

$$r = \arg\max\left\{k \le k_{\max} :\ \left|s_k - s_{k+1}\right| \ge \delta\right\},\ \forall k \in k_{\max}$$

namely, the number of scores corresponds to the largest of the absolute differences between contiguous eigenvalues must and the parameter δ. In such case, $r \ge 1$, elsewise $r = 0$.

In PCA the criteria for the detection of the true number of scores r proposed in the literature are the outdated screeplot visual inspection and eigenvalue- or variance-based methods, amongst which Cross Validation is very popular (Krzanowski, 1987). Briefly, they concentrate on the computation of the *n* singular values and/or of their corresponding shares, fixing the value of r at some stopping time dictated, when using correlation matrices, by the Kaiser-Guttman rule (Draper and Smith, 1998; Loehlin, 2004) which states that

29) $$r = \arg_{1 \le j \le n}\left(s_j \ge 1\right),\ j \in n$$

or more generally by

30) $$r = \arg_{1 \le j \le n}\left(\sum_{i=1}^{r} s_i \Big/ \sum_{j=1}^{n} s_j \ge c_r\right)$$

where c_r is some arbitrary share of the covariance matrix of *X* such that the cumulative sum $\sum_{j=1}^{r} c_j \approx 1$.

Both criteria exhibited in eqs. (29) and (30), and many more with similar constructs (e.g. Myers, 2000), may select a very large r, definitely larger than that envisaged by the FA authors (Bai and Ng, 2002, 2008; Stock and Watson, 2002, 2005). For instance, in the case of eq. (30) approximately one half of the singular values of the correlation matrix exceed unity, while over $2n/3$ is needed for the cumulative sum of shares to exceed 95%.

Both these occurrences put the stopping time for r to a rather impressive number for $n \to \infty$, although a recent addition to PCA modeling, Supervised PCA (Bair et al., 2006) by appealing to correlations of *X* with the endogenous variable in a pre-established OLS context, may sizably reduce this number. By any means, however, the FA methods suggested in eqs. (22), (24), (25) and (28) fix r to very small numbers even in the presence of very large-sized datasets, also in its modified dynamic versions (Bernanke et al., 2005; Forni et al., 2004; Stock and Watson, 2002, 2005). In fact Montecarlo simulations conducted over stationary X matrices (n=50, p=60; n=150, p=350) on average respectively put r between 2 and 3 and between 1 and 3.

5. Conclusion

The main distinctive features of the FA and of the PCA models have been examined in this paper. By exploiting some important results in the literature (Bai 2003; Bai and Ng, 2002, 2006, 2007, 2008, 2010), factors, factor loadings and scores, as well as singular values and the FA errors have undergone careful scrutiny especially with regard to their asymptotic properties in the presence of large datasets and to the magnitude of their norms.

In sum, FA is shown to largely outperform PCA on several grounds: *i*) when utilized as regressors and/or instruments, the FA scores produce more efficient slope estimators in instrumental variable estimation, notably in GMM; *ii*) both FA scores and factors exhibit substantial consistency, because the asymptotic difference between the estimated and their true counterparts and the corresponding variances are smaller; *iii*) the FA procedure of dimension reduction produces a much more limited amount of true scores, thereby greatly facilitating the search and identification of the common components of large datasets.

References

Ahn S.C. and Horenstein A.R. (2009) *Eigenvalue Ratio Test for the Number of Factors*, mimeo, Arizona State University and Instituto Autónomo Tecnológico de México.

Alessi L., Barigozzi M. and Capasso M. (2009) *A Robust Criterion for Determining the Number of Factors in Approximate Factor Models*, ECORE Discussion Paper 97, European Central Bank, Frankfurt am Main, Germany.

Anderson T. W. (1958) *An Introduction to Multivariate Statistical Analysis*, Wiley, New York, N.Y.

Anderson T. W. (1963) *Asymptotic Theory for Principal Component Analysis*, 34, 122-148.

Anderson T. W. (1984) *An Introduction to Multivariate Statistical Analysis*, 2^{nd}. Ed., Wiley Series in Probability and Statistics, New York, N.Y.

Anderson T.W. and Rubin H. (1956) *Statistical Inference in Factor Analysis*, Cowles Foundation Paper No. 103.

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

Bai J. (2003) *Inferential Theory for Factor Models of Large Dimensions*, Econometrica, 71, 135-171.

Bai J. and Ng S. (2006) *Instrumental Variable Estimation in a Data Rich Environment*, NYU mimeo.

Bai J. and Ng S. (2007) *Determining the Number of Primitive Shocks in Factor Models*, Journal of Business and Economic Statistics, 26, 52- 60.

Bai J. and Ng S. (2008) *Selecting Instrumental Variable Estimation in a Data Rich Environment*, Journal of Time Series Econometrics, 1, 1-32.

Bai J. and Ng S. (2010) *Principal Components, Estimation and Identification of the Factors*, mimeo, Department of Economics, Columbia University.

Bai Z.D. (1993) *Convergence Rate of Expected Spectral Distribution of Large Random Matrices. Part II. Sample Covariance Matrices*, The Annals of Probability, 21, 649-672.

Bai Z.D., Miao B. and Yao Samos-Matisse J., (2000), *Convergence Rates of Spectral Distributions of Large Sample Covariance Matrices*, mimeo.

Bair E., Hastie T., Paul D. and Tibshirani R. (2006) *Prediction by Supervised Principal Components*, Journal of the American Statistical Association, 101, 119-137.

Bernanke B. and Boivin J. (2003) *Monetary Policy in a Data-Rich Environment*, Journal of Monetary Economics, L, 525-546.

Bernanke B., Boivin J. and Eliasz P. (2005), *Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach*, mimeo.

Cattell, R.B. (1966) *The Scree Test for the Number of Factors,* Multivariate Behavioral Research, 1, 245-276.

Chamberlain G. and Rothschild M. (1983) *Arbitrage, factor structure, and mean-variance analysis on large asset markets*, Econometrica, 51, 1281-1304.

Connor G. and Korajczyk R. (1986) *Performance Measurement with the Arbitrage Pricing Theory: A New Framework for Analysis*, Journal of Financial Economics, 15, 373-394.

Draper N.R. and Smith H. (1998) *Applied Regression Analysis*, 3^{d} Ed.**,** Wiley-Interscience, New York, N.Y.

Forni, M., Hallin M., Lippi M. and Reichlin L. (2004) *The Generalized Dynamic Factor Model: Consistency and Rates*, Journal of Econometrics 119, 231-255.

Forni M. and Gambetti L. (2008) *The Dynamic Effects of Monetary Policy: A Structural Factor Model Approach*, Centre for Economic Policy Research, CEPR, DP7098.

Geman S. (1980) *A Limit Theorem for the Norm of Random Matrices*, The Annals of Probability, 2, 252-261.

Hansen L.P. (1982) *Large Sample Properties of Generalized Method of Moments Estimator*, Econometrica, 50, 1029-1054.

Johnstone I.M. (2001) *On the Distribution of the Largest Eigenvalue in Principal Component Analysis*, The Annals of Statistics, 29, 295-327.

Jolliffe I. (1982) *A Note on the use of Principal Components in Regression*, Applied Statistics, 31, 300-303.

Jolliffe I. (2002) *Principal Component Analysis*, 2nd. Ed., Springer-Verlag, New York, N.Y.

Kapeitanos G. and Marcellino M. (2007) *Factor-GMM Estimation with Large Sets of Possibly Weak Instruments*, Working Papers 577, Queen Mary University of London, Department of Economics.

Krzanowski W.J. (1987) *Cross-Validation in Principal Component Analysis*, Biometrics, 43, 575-584.

Loehlin, J.C. (2004) *Latent Variable Models: an Introduction to Factor, Path, and Structural Analysis*, 4th Ed., Lawrence Erlbaum Associates, N.J.

Myers R.H. (2000) *Classical and Modern Regression with Applications*, 2nd Ed., Duxbury Press.

Newey W. and Windmeijer F. (2009) *GMM with Many Weak Moment Conditions*, Econometrica, 77, 687-719.

Onatski A. (2009) *Determining the Number of Factors from Empirical Distribution of Eigenvalues*, manuscript, Economics Department, Columbia University.

Pearson K. (1901) *On Lines and Planes of Closest Fit to Systems of Points in Space*, Philosophical Magazine 2, 559–572.

Rudelson M. and Vershynin R. (2010) *Non-asymptotic Theory of Random Matrices: Extreme Singular Values*, Proceedings of the International Congress of Mathematicians, Hyderabad, India.

Sargent, T.J. and Sims C.A. (1977) *Business Cycle Modeling without Pretending to have too much a priori Economic Theory*, In C.A. Sims, Ed., *New Methods in Business Research*, Federal Reserve Bank of Minneapolis, Minneapolis.

Sokal A.D. (2010) *A Really Simple Elementary Proof of the Uniform Boundedness Theorem*, mimeo, Dept. of Physics, New York University.

Stock J.H. and Watson M.W. (2002) *Forecasting Using Principal Components from a Large Number of Predictors,* Journal of the American Statistical Association, 97, 1167–1179.

Stock J.H. and Watson M.W. (2005) *Implications of Dynamic Factor Models for VAR Analysis,* NBER Working Paper No. 11467.

Vershynin R. (2011) *Introduction to the Non-asymptotic Analysis of Random Matrices*, University of Michigan, mimeo.

Computations of empirical norms and Montecarlo simulations were performed with @Matlab R2011a. Codes are available upon request from the author.