View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Research Papers in Economics

CCSO Centre for Economic Research

University of Groningen

CCSO Working Papers **July, 2006**

CCSO Working paper 2006/05

On information in static and dynamic factor models

Pieter W. Otter
Faculty of Economics, University of Groningen

Jan P.A.M. Jacobs
Faculty of Economics, University of Groningen

CCSO Working Papers
http://www.eco.rug.nl/ccso

On information in static and dynamic factor models

Pieter W. Otter[*] and Jan P.A.M. Jacobs

Faculty of Economics, University of Groningen

July 2006

Presented at
The Far Eastern Meeting of the Econometric Society,
July 9–12, 2006, Beijing, China

Abstract

This paper employs concepts from information theory in factor models. We show that in the exact factor model the whole distribution of eigenvalues of the covariance matrix contributes to the information and not only the largest ones. In addition, we derive the condition that the first q say eigenvalues diverge whereas the rest remain bounded in the static model rather than having to assume it. Finally, we calculate information in static and dynamic factor models, which can be used to find the dimensions of the factor space. We illustrate the concepts with simulation experiments.

Keywords factor analysis, model selection, information

JEL-code C32, C52, C82

[*] Corresponding author: Pieter W. Otter, Department of Econometrics, University of Groningen, P.O. Box 800, 9700 AV Groningen, The Netherlands. Tel.: +31 50 363 3782. Fax: +31 50 363 3720. Email: p.w.otter@eco.rug.nl

1 Introduction

With the proliferation of huge data sets, factor models are becoming more and more popular. Approximate factor models exploit the intuitively appealing idea that variations in a large number of economic variables can be adequately modelled by a small number of reference variables, in other words that movements in a large number of series are driven by a limited number of common 'factors'. A more recent development is the introduction of dynamic factor models, in which factors can affect variables with lead and lags. Recent examples are Stock and Watson (2002a, 2002b), Camba-Mendez, Kapetanios, Smith, and Weale (2001), and the Generalized Dynamic Factor Model of Forni, Hallin, Lippi and Reichlin (2000).

A natural question to ask is how much information is in the data set. This question has two dimensions: (i) how many factors are sufficient to adequately capture the information in the data set? and (ii) is there an 'optimal' size of the data set, i.e. does an additional variable add information? To start with the latter, the size of the data set does not need to be very large to obtain reasonable precise factor estimates. Boivin and Ng (2003) and Inklaar, Jacobs, and Romp (2005) find that some 40 variables are sufficient using Monte Carlo simulations and a comparison to conventional NBER-type business cycle indicators, respectively. Bai and Ng (2002) come to the same conclusion.

The determination of the optimal number of factors is a topic of ongoing research. Two main approaches can be distinguished. Forni et al. (2000) advocate heuristic inspection of eigenvalues against the number N of the series.

Each factor should explain at least a prespecified percentage of total variance. The average over the first q empirical eigenvalues diverges, whereas the average of the $(N - q)$ smallest eigenvalues is relatively stable. An alternative route is taken by Bai and Ng (2002), who propose the use of information criteria to determine the optimal number of static factors r as a trade-off between goodness-of-fit and overfitting. The Bai and Ng information criteria give an upper bound for the number of dynamic factors q, since the number of static factors $r = q(p+1)$ is the maximum combination of dynamic factors and their lag p. Recently, Kapetanios (2004, 2005) provides an alternative to information criteria based on a sequence of tests on the largest eigenvalue of the sample covariance matrix.

This paper exploits concepts from information theory in static and dynamic factor models, in particular Kullback-Leibler numbers. We link entropy and information (negative entropy) from data to factor models, and derive the distribution of eigenvalues in relation to information. We show that the whole distribution of eigenvalues of the covariance matrix in the data and the exact factor model contributes to the information and not only the largest ones. In addition, we analyse a strict factor model. By this we derive the condition that the first q say eigenvalues diverge whereas the rest remain bounded in the static model rather than having to assume it as for example Forni et al. (2002) do. Finally, we calculate information in static and dynamic factor models, which can be used to find the dimensions of the factor space. Simulation experiments illustrate our methods.

The paper is structured as follows. Section 2 discusses information in the data set. Section 3 looks at static factor models focusing at modelling crosssectional correlations, whereas Section 4 considers dynamic factor models and autocovariances. Section 5 reports some simulation experiments. Section 6 concludes.

2 Information in data

Let $\boldsymbol{x}_t$ be an N-dimensional vector of observed data at time t, $t = 1, \ldots, T$. The data is demeaned and normalized, and normally distributed with mean zero and variance $\mathrm{E}(\boldsymbol{x}_t \boldsymbol{x}_t') = \boldsymbol{\Gamma}_0$, i.e. $\boldsymbol{x}_t \sim \mathbb{N}(\mathbf{0}, \boldsymbol{\Gamma}_0)$, where $\mathrm{diag}(\boldsymbol{\Gamma}_0) = (1, 1 \ldots, 1)$, $\mathrm{tr}(\boldsymbol{\Gamma}_0) = N$ and $\boldsymbol{\Gamma}_i = \mathrm{E}(\boldsymbol{x}_t \boldsymbol{x}_{t-i}')$ are the autocovariances of $\boldsymbol{x}_t$.

The entropy, denoted by H, as measure of disorder is for a stationary, normally distributed vector given by

$$2H_x = cN + \mathrm{logdet}(\boldsymbol{\Gamma}_0),$$

where $c \equiv \log(2\pi) + 1 \approx 2.84$, with $2H_{x,max} = cN$ in case $\boldsymbol{\Gamma}_0 = \boldsymbol{I}_N$, see e.g. Goodwin and Payne (1977). The information or negentropy is defined as

$$I_x \equiv 2H_{x,max} - 2H_x = -\mathrm{logdet}(\boldsymbol{\Gamma}_0) \geq 0, \qquad (1)$$

which is zero in case $\boldsymbol{\Gamma}_0 = \boldsymbol{I}_N$.

Assuming that the autocovariance matrix $\boldsymbol{\Gamma}_0$ has full rank, we can apply the decomposition

$$\boldsymbol{\Gamma}_0 = \boldsymbol{C}\boldsymbol{\Lambda}\boldsymbol{C}', \text{ with } \boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_N);\ \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N > 0. \qquad (2)$$

Therefore, we have $\mathrm{tr}(\boldsymbol{\Gamma}_0) = \mathrm{tr}(\boldsymbol{\Lambda}) = N$.

Let $||\boldsymbol{A}||_E = \left(\sum_{i,j} |a_{i,j}|^2\right)^{1/2} = \mathrm{tr}(\boldsymbol{A}'\boldsymbol{A})^{1/2}$ be the Euclidean (Schur) norm of the matrix $\boldsymbol{A}$. So, $||\boldsymbol{\Gamma}_0||_E^2 = \sum_{i,j} |\gamma_{i,j}|^2$ measures the magnitude of correlation between $\boldsymbol{x}_{i,t}$ and $\boldsymbol{x}_{j,t}$ with $\gamma_{i,i} = 1$. From Equation (2) we have

$$||\boldsymbol{\Gamma}_0||_E^2 = \mathrm{tr}(\boldsymbol{C}\boldsymbol{\Lambda}^2\boldsymbol{C}') = \mathrm{tr}(\boldsymbol{\Lambda}^2) = \sum_{i=1}^{N} \lambda_i^2$$

Let $\boldsymbol{\lambda} = (\lambda_1 \dots \lambda_N)'$, then $||\boldsymbol{\Gamma}_0]]_E^2 = \boldsymbol{\lambda}'\boldsymbol{\lambda}$ which under the restriction $\boldsymbol{\iota}'\boldsymbol{\lambda} = \sum \lambda_i = N$ attains its maximum when $\lambda_1 = N$ and $\lambda_j = 0, j = 2, \dots, N$.

Absolute information can be rewritten as

$$\begin{aligned} I_x &= -\operatorname{logdet}(\boldsymbol{\Gamma}_0) = -\sum_{j=1}^{N} \log \lambda_j \geq 0 \\ &= -\sum_{j=q+1}^{N} \log \tilde{\lambda}_j - \sum_{j=1}^{q} \log \lambda_j \geq 0, \quad \tilde{\lambda}_j \leq 1, \lambda_j > 1 \\ &= \sum_{j=q+1}^{N} \log \left(1/\tilde{\lambda}_j\right) - \sum_{j=1}^{q} \log \lambda_j \geq 0 \\ &= \Big(\log(1/\tilde{\lambda}_N) - \log(\lambda_1)\Big) + \Big(\log(1/\tilde{\lambda}_{N-1}) - \log(\lambda_2)\Big) + \dots \geq 0. \end{aligned}$$

So, absolute information I_x is positive if $\left(1/\tilde{\lambda}_N\right) > \lambda_1$, $\left(1/\tilde{\lambda}_{N-1}\right) > \lambda_2, \dots$ or $\mathrm{tr}(\boldsymbol{\Lambda}^{-1}) > \mathrm{tr}(\boldsymbol{\Lambda}) = N$. The information I_x is determined by the magnitude of the $(N-q)$ eigenvalues $\tilde{\lambda}_j^2 < 1$ and the magnitude of the q largest eigenvalues λ_j. The larger $\mathrm{tr}(\boldsymbol{\Lambda}^{-1}) - N$, the more information. This can be seen by using a divergence information measure based on the Kullback-Leibler numbers and the entropy of the eigenvalues.

Kullback-Leibler information and divergence

Let $f_1(\tilde{\boldsymbol{x}}) : \tilde{\boldsymbol{x}} \sim \mathcal{N}_N\left(\mathbf{0}, \boldsymbol{\Gamma}_0 = \boldsymbol{C}\boldsymbol{\Lambda}\boldsymbol{C}'\right)$ be the density function of $\boldsymbol{x}$ (time index suppressed), then $f_1(\boldsymbol{x}) : \boldsymbol{x} \sim \mathcal{N}_N\left(\mathbf{0}, \boldsymbol{\Lambda}\right)$ where $\boldsymbol{x} = \boldsymbol{C}'\tilde{\boldsymbol{x}}$. Let $f_2(\tilde{\boldsymbol{x}}) : \tilde{\boldsymbol{x}} \sim \mathcal{N}_N\left(\mathbf{0}, \boldsymbol{I}_N\right)$. Then $f_2(\boldsymbol{x}) : \boldsymbol{x} \sim \mathcal{N}_N\left(\mathbf{0}, \boldsymbol{I}_N\right)$ with $\boldsymbol{x} = \boldsymbol{C}'\tilde{\boldsymbol{x}}$. The so-called *Kullback-Leibler* numbers are defined as

$$G_1 = \mathrm{E}_{f_1}\left(\log\left(\frac{f_1(\boldsymbol{x}}{f_2(\boldsymbol{x})}\right)\right) \text{ and } G_2 = \mathrm{E}_{f_2}\left(\log\left(\frac{f_2(\boldsymbol{x}}{f_1(\boldsymbol{x})}\right)\right) \tag{3}$$

and $G(\boldsymbol{x}) = G_1(\boldsymbol{x}) + G_2(\boldsymbol{x})$ is the measure of information for discriminating between the two density functions with $G(\boldsymbol{x}) = 0$ in case $f_1(\boldsymbol{x}) = f_2(\boldsymbol{x})$ and $G = \infty$ in case of perfect discrimination, see Young and Calvert (1974, pp 245–245). For a general background see Burnham and Anderson (2002).

For $\mathrm{tr}\left(\boldsymbol{\Gamma}_0\right) = \mathrm{tr}(\boldsymbol{\Lambda}) = N$ we have $G_1(\boldsymbol{x}) = -\mathrm{logdet}(\boldsymbol{\Lambda})$ and $G_2(\boldsymbol{x}) = \mathrm{logdet}(\boldsymbol{\Lambda}) + \frac{1}{2}\left(\mathrm{tr}(\boldsymbol{\Lambda}^{-1}) - N\right)$. Therefore

$$2G(\boldsymbol{x}) = \mathrm{tr}(\boldsymbol{\Lambda}^{-1}) - N = \mathrm{tr}(\boldsymbol{\Lambda}^{-1}) - \mathrm{tr}(\boldsymbol{\Lambda}) = \sum_{j=1}^{N} \frac{(1-\lambda_j^2)}{\lambda_j}, \tag{4}$$

from which it can be seen that $G(\boldsymbol{x})$ is not discriminating if $\lambda_j \approx 1$ but is discriminating for "small" $\lambda_j < 1$.

Consider the special case of the transformation with $\boldsymbol{x}_t \sim \mathcal{N}_N\left(\mathbf{0}, \boldsymbol{\Gamma}_0 = \boldsymbol{C}\boldsymbol{\Lambda}\boldsymbol{C}'\right)$ with $\boldsymbol{C} = (c_1\ c_2\ \ldots\ c_N)$ and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N \geq 0$, i.e.

$$\boldsymbol{y}_t = \boldsymbol{C}_1'\boldsymbol{x}_t \text{ or } \boldsymbol{y}_t' = \boldsymbol{x}_t'\boldsymbol{C}_1 = \boldsymbol{x}_t'(c_1\ c_2\ \ldots\ c_k)$$

where $\{c_j, j = 1, \dots k\}$ are the first k eigenvectors, then $\mathrm{E}(\boldsymbol{y}_t) = 0$ and $\mathrm{var}(\boldsymbol{y}_t) = \boldsymbol{\Lambda}_1 = \mathrm{diag}(\lambda_1 \dots \lambda_k)$.

This transformation, in which $\boldsymbol{y}_t$ is called the *feature*-vector, is known as the *Karhunen-Loève* expansion. Young and Calvert (1974) show that the Karhunen-Loève expansion is an optimal minimax entropy feature extractor in case $f(\boldsymbol{x})$ is not Gaussian (e.g. a mixture of density functions $\{f(\boldsymbol{x}) : \mathrm{E}(\boldsymbol{x}_t) = 0, \mathrm{E}(\boldsymbol{x}_t \boldsymbol{x}_t') = \boldsymbol{R}\}$. Estimates of the feature vector $\boldsymbol{y}_t$ through time are given below, relating the feature vector to principal components.

The foregoing shows that the distribution of the eigenvalues is important, which can be measured by the entropy of the eigenvalues. Because $\mathrm{tr}(\boldsymbol{\Lambda}) = N$ we have $\bar{\lambda}_j = \lambda_j / N$ with $0 \leq \bar{\lambda}_j \leq 1$ and

$$H_{\bar{\lambda}} = -\sum_j \bar{\lambda}_j \log \bar{\lambda}_j \tag{5}$$

with $H_{\bar{\lambda}}^{max} = \log(N)$ for $\bar{\lambda}_j = 1/N$ for all j. As mentioned above in the ideal case we have $\lambda_1 = N$ $(\bar{\lambda}_1 = 1)$ and $\lambda_j = 0, j = 2, \dots, N$ and $H_{\bar{\lambda}} = 0$ (with the usual convention $\bar{\lambda}_j \log \bar{\lambda}_j = 0$ for $\bar{\lambda}_j = 0$). The information contained in the eigenvalues is $I_{\bar{\lambda}} = \log(N) - H_{\bar{\lambda}}$ or the relative information

$$I_{\bar{\lambda}}^R = 1 - \frac{H_{\bar{\lambda}}}{\log(N)},$$

with $0 \leq I_{\bar{\lambda}}^R \leq 1$.

3 Static factor model

Up to now we did not exploit the fact that the data is driven by a number of factors. Now let $\boldsymbol{x}_t$ be driven by k factors

$$\boldsymbol{x}_t = \boldsymbol{B}\boldsymbol{F}_t + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{x}_t \in \mathbb{R}^n, \boldsymbol{F}_t \sim \mathbb{N}_k\left(\boldsymbol{0}, \boldsymbol{I}_k\right), \boldsymbol{\varepsilon}_t \sim \mathbb{N}_N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_N), \tag{6}$$

where $\boldsymbol{B} \in \mathbb{R}^{N\times k}$ are the matrix of factor loadings, with Euclidean norm $||\boldsymbol{B}||_E^2 = \sum_{i,j} |b_{i,j}|^2 = \mathrm{tr}(\boldsymbol{B}'\boldsymbol{B})$. We can apply the Singular Value Decomposition (SVD) to the matrix of factor loadings

$$\boldsymbol{B} = \boldsymbol{U}_N \boldsymbol{S} \boldsymbol{V}_k' \text{ with } \boldsymbol{S} = \begin{pmatrix} \boldsymbol{S}_k \\ \boldsymbol{0} \end{pmatrix}, \tag{7}$$

where the columns of $\boldsymbol{U}_N$ and $\boldsymbol{V}_k$ are orthonormal, i.e. $\boldsymbol{U}_N\boldsymbol{U}_N' = \boldsymbol{U}_N'\boldsymbol{U}_N = \boldsymbol{I}_N$ and $\boldsymbol{V}_k\boldsymbol{V}_k' = \boldsymbol{V}_k'\boldsymbol{V}_k = \boldsymbol{I}_k$, $\boldsymbol{S}_k = \mathrm{diag}(s_1, \ldots s_k), s_1 \geq s_2 \geq s_k > 0$ (s_i's are the singular values). So, $\boldsymbol{B}$ has Euclidean norm $||\boldsymbol{B}||_E^2 = \mathrm{tr}(\boldsymbol{B}'\boldsymbol{B}) = \sum_{i=1}^k s_i^2$. Let $\boldsymbol{b}_j \equiv (b_{j1} \ldots b_{jk})$ be the j-th row of $\boldsymbol{B}$, i.e. the vector of factor loadings of the j-th component of $\boldsymbol{x}_t$. Then using $\mathrm{tr}(\boldsymbol{B}'\boldsymbol{B}) = \mathrm{tr}(\boldsymbol{B}\boldsymbol{B}')$

$$\mathrm{tr}(\boldsymbol{B}\boldsymbol{B}') = \sum_{j=1}^{N} ||b_j||_2^2 \text{ with } ||b_j||_2^2 = b_j b_j'$$

and

$$S_N \equiv \sum_{i=1}^{k} s_i^2(N) = \sum_{i=1}^{N} ||b_j||_2^2,$$

i.e. the singular values function of N are proportionally increasing with N provided $||b_j||_2^2 \neq 0$ for all j, so $S_N = \mathcal{O}(N)$.

Rewrite the factor model (6) as

$$\boldsymbol{x}_t = \boldsymbol{U}_N \boldsymbol{S} \boldsymbol{V}_k' \boldsymbol{F}_t + \tilde{\boldsymbol{\varepsilon}}_t = \boldsymbol{U}_N \boldsymbol{S} \tilde{\boldsymbol{F}}_t + \tilde{\boldsymbol{\varepsilon}}_t,$$

where $\tilde{\boldsymbol{F}}_t = \boldsymbol{V}_k' \boldsymbol{F}_t$ with variance $\operatorname{var}(\tilde{\boldsymbol{F}}_t) = \boldsymbol{V}_k' \boldsymbol{V}_k = \boldsymbol{I}_k$ and $\tilde{\boldsymbol{\varepsilon}}_t = \boldsymbol{U}_N \boldsymbol{\varepsilon}_t$ with variance $\operatorname{var}(\tilde{\boldsymbol{\varepsilon}}_t) = \sigma^2 \boldsymbol{U}_N \boldsymbol{U}_N' = \sigma^2 \boldsymbol{I}_N$. The factor model becomes

$$\boldsymbol{x}_t = \boldsymbol{U}_N \left[\boldsymbol{S} \tilde{\boldsymbol{F}}_t + \boldsymbol{\varepsilon}_t \right],$$

and has autocovariance

$$\boldsymbol{\Gamma}_0 = \mathrm{E}(\boldsymbol{x}_t \boldsymbol{x}_t') = \boldsymbol{U}_N \left(\begin{pmatrix} \boldsymbol{S}_k & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} + \sigma^2 \boldsymbol{I}_N \right) \boldsymbol{U}_N' = \boldsymbol{U}_N \boldsymbol{A} \boldsymbol{U}_N'.$$

Matrix $\boldsymbol{A}$ is equal to $\boldsymbol{A} = \operatorname{diag}(s_1^2 + \sigma^2, s_2^2 + \sigma^2, \ldots, s_k^2 + \sigma^2, \sigma^2, \ldots, \sigma^2) \in \mathbb{R}^{N \times N}$. Normalisation using $\operatorname{tr}(\boldsymbol{\Gamma}_0) = \operatorname{tr}(\boldsymbol{A}) = \sum_{j=1}^k s_j^2 + N\sigma^2 = N$ yields

$$\tilde{\boldsymbol{A}} = \operatorname{diag}\left(\frac{s_1^2 + \sigma^2}{\bar{s}_N + \sigma^2}, \frac{s_2^2 + \sigma^2}{\bar{s}_N + \sigma^2}, \ldots, \frac{s_k^2 + \sigma^2}{\bar{s}_N + \sigma^2}, \frac{\sigma^2}{\bar{s}_N + \sigma^2}, \ldots, \frac{\sigma^2}{\bar{s}_N + \sigma^2} \right),$$

where $\bar{s}_N \equiv \sum_{j=1}^k s_j^2 / N$. The first element in $\tilde{\boldsymbol{A}}$ is larger than one, because $\bar{s}_N \leq \frac{k}{N} s_1^2$.

Let $c(\boldsymbol{B}'\boldsymbol{B}) = s_1 / s_k$ be the condition number of $\boldsymbol{B}'\boldsymbol{B}$. Then $s_1^2 = c^2(\boldsymbol{B}'\boldsymbol{B}) s_k^2$ and $\bar{s}_N \leq \frac{k}{N} c^2(\boldsymbol{B}'\boldsymbol{B}) s_k^2$ from which follows that $\frac{s_j^2 + \sigma^2}{\bar{s}_N + \sigma^2} > 1$ for $j = 1, \ldots, k$ if $c(\boldsymbol{B}'\boldsymbol{B}) < \sqrt{\frac{N}{k}}$. The larger the average magnitude of $\boldsymbol{B}$ measured by $\bar{s}_N$ the smaller the $(N - k)$ elements $\frac{\sigma^2}{\bar{s}_N + \sigma^2} < 1$

Adding a variable

What is the effect of adding a new variable $\boldsymbol{x}_{N+1,t}$ to the data set $\boldsymbol{x}_t$? The matrix of factor loading becomes $\boldsymbol{B}_{N+1} = \begin{pmatrix} \boldsymbol{B} \\ \boldsymbol{b}_{N+1} \end{pmatrix}$ with

$$\begin{aligned} \operatorname{tr}(\boldsymbol{B}'_{N+1}\boldsymbol{B}_{N+1}) =& ||\boldsymbol{B}_{N+1}||_E^2 = \operatorname{tr}(\boldsymbol{B}'\boldsymbol{B}) + \boldsymbol{b}_{N+1}\boldsymbol{b}'_{N+1} \\ =& S_N + ||\boldsymbol{b}_{N+1}||_2^2 = \sum_{j=1}^{k} s_j^2(N+1) \equiv S_{N+1}, \end{aligned}$$

i.e. $S_{N+1} = S_N + ||\boldsymbol{b}_{N+1}||_2^2$. Therefore, if $||\boldsymbol{b}_{N+1}||_2^2 \neq 0$, the sum of squared singular values is proportionally increasing with N, $s_N = \mathcal{O}(N)$, $\bar{s}_N \equiv S_N/N$ is bounded, and $s_j^2, j = 1, \ldots, k$ diverge whereas the last $(n-k)$ elements of $\tilde{\boldsymbol{A}}$ are bounded.

The information in the autocovariance matrix $\boldsymbol{\Gamma}_0$ is equal to $2I_x = -\text{logdet}(\boldsymbol{\Gamma}_0) = -\sum_{j=1}^{N} \log \lambda_j = -\sum_{j=1}^{N} \log \tilde{a}_{jj}$, where $\tilde{a}_{jj}$ is the j-th element of $\tilde{\boldsymbol{A}}$, or

$$2I_x = (N-k)\log\left(\frac{\bar{s}_N + \sigma^2}{\sigma^2}\right) - \sum_{j=1}^{k} \log\left(\frac{s_j^2 + \sigma^2}{\bar{s}_N + \sigma^2}\right).$$

Do variables add information? Recall that for $\boldsymbol{x}_t(N) \in \mathbb{R}^N$ with autocovariance $\mathrm{E}(\boldsymbol{x}_t(N)\boldsymbol{x}'_t(N)) = \boldsymbol{\Gamma}_0(N)$ the entropy is defined as $2H_{x_t(N)} = cN + \text{logdet}\left(\boldsymbol{\Gamma}_0(N)\right)$ and the information as $2I_{x_t(N)} = 2H_{x,max} - 2H_x = -\text{logdet}\left(\boldsymbol{\Gamma}_0(N)\right) \equiv \boldsymbol{I}_N$. Define the relative information (per component of $\boldsymbol{x}_t(N)$) as:

$$2I_N^R = \frac{2H_{max} - 2H_x(N)}{2H_{max}} = \frac{I_N}{2H_{max}} = \frac{I_N}{cN}.$$

If $H_{x(N)}$ is equal to H_{max} then $2I_N^R = 0$; if $H_{x(N)} = 0$ then $2I_N^R = 1$. So, an additional variable $\boldsymbol{x}_{N+1,t}$ adds information if

$$2I_{N+1}^R > 2I_N^R \text{ or} \frac{I_{N+1}}{c(N+1)} > \frac{I_N}{cN}, \text{ i.e.}(I_{N+1} - I_N) > I_N/N.$$

The $(N+1)$-th variable need to add more information than the average contribution of the N variables already included in the data set.

Least squares estimation and feature extraction

The static factor model with "true" number of factors equal to $\bar{k}$ is given by

$$\boldsymbol{x}_t = \boldsymbol{B}\boldsymbol{F}_t + \boldsymbol{\varepsilon}_t, \qquad t = 1, \ldots,$$

where the factor loading matrix $\boldsymbol{B} \in \mathbb{R}^{N \times \bar{k}}$ with $\text{rank}(\boldsymbol{B}) = \bar{k}$, the factors are orthogonal $\mathrm{E}\left(\boldsymbol{F}_t\boldsymbol{F}_t'\right) = \boldsymbol{I}_{\bar{k}}$, and the errors have mean zero $\mathrm{E}(\boldsymbol{\varepsilon}_t) = 0$ and variance $\mathrm{E}(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t') = \boldsymbol{V}_{\boldsymbol{\varepsilon}}$. Note that we do not assume a strict factor mapping here.

Consider T demeaned observations of the j-th variable, $j = 1, \ldots, N$, collected in the vector $\tilde{\boldsymbol{x}}_j \in \mathbb{R}^T$. Let $\boldsymbol{x}_j = \tilde{\boldsymbol{x}}_j / s_{\boldsymbol{x}_j}$ with $s_{\boldsymbol{x}_j}^2 = ||\tilde{\boldsymbol{x}}_j||_2^2/T$ then $||\boldsymbol{x}_j||_2^2 = T$. Write the $(T \times N)$ matrix $\boldsymbol{X}$ as $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1' \\ \vdots \\ \boldsymbol{x}_T' \end{pmatrix}$, $\boldsymbol{x}_t \in \mathbb{R}^N, \boldsymbol{x}_t' = (x_{1,t} \ldots x_{N,t})$. For $T \geq N$ we have $\frac{1}{T}\boldsymbol{X}'\boldsymbol{X} = \hat{\boldsymbol{\Gamma}}_0 = \frac{1}{T}\sum_{t=1}^T \boldsymbol{x}_t\boldsymbol{x}_t'$, with $\text{tr}(\hat{\boldsymbol{\Gamma}}_0) = N$.

To apply singular value compositions we distinguish two cases: (i) $T \geq N$ and (ii) $T < N$. Ad (i) If $T \geq N$ the SVD becomes

$$\boldsymbol{X} = \boldsymbol{U}_T \begin{pmatrix} \boldsymbol{S}_N \\ \boldsymbol{0} \end{pmatrix} \boldsymbol{V}_N' = \boldsymbol{U}_N \boldsymbol{S}_N \boldsymbol{V}_N',$$

with $\boldsymbol{U}_T = (\boldsymbol{u}_1 \ldots \boldsymbol{u}_T)$ orthonormal $(\boldsymbol{U}_T'\boldsymbol{U}_T = \boldsymbol{U}_T\boldsymbol{U}_T' = \boldsymbol{I}_T)$, $\boldsymbol{V}_N = (\boldsymbol{v}_1 \ldots \boldsymbol{v}_N)$ orthonormal $(\boldsymbol{V}_N'\boldsymbol{V}_N = \boldsymbol{V}_N\boldsymbol{V}_N' = \boldsymbol{I}_N)$, and $\boldsymbol{S}_N = \operatorname{diag}(s_1 \ldots s_N), s_1 \geq s_2 \geq \ldots \geq s_N \geq 0$. Ad (ii) For $T < N$ we get

$$\boldsymbol{X} = \boldsymbol{U}_T(\boldsymbol{S}_T \ \boldsymbol{0})\boldsymbol{V}_N' = \boldsymbol{U}_T\boldsymbol{S}_T\boldsymbol{V}_T',$$

with $\boldsymbol{V}_T = (\boldsymbol{v}_1 \ldots \boldsymbol{v}_T)$ and $\boldsymbol{S}_T = \operatorname{diag}(s_1 \ldots s_T)$.

Defining $kmax = \min(N, T)$ the two SVDs can be written as

$$\boldsymbol{X} = \boldsymbol{U}_{kmax}\boldsymbol{S}_{kmax}\boldsymbol{V}_{kmax}'. \tag{8}$$

If $kmax = N$, the autocovariance $\hat{\boldsymbol{\Gamma}}_0 = \frac{1}{T}\boldsymbol{X}'\boldsymbol{X} = \frac{1}{T}\boldsymbol{V}_N\boldsymbol{S}_N^2\boldsymbol{V}_N' = \boldsymbol{V}_N\hat{\boldsymbol{\Lambda}}_N\boldsymbol{V}_N'$, or $\hat{\boldsymbol{\Lambda}}_N = \frac{1}{T}\boldsymbol{S}_2^2$, which implies the $\hat{\lambda}_j = s_j^2/T,\ j = 1, \ldots, N$, are eigenvalues of $\hat{\boldsymbol{\Gamma}}_0$.

Select $k < kmax$ and apply a *least squares decomposition*

$$\boldsymbol{X} = \boldsymbol{U}_k\boldsymbol{S}_k\boldsymbol{V}_k' + \boldsymbol{U}_2\boldsymbol{S}_2\boldsymbol{V}_2' \equiv \hat{\boldsymbol{X}} + \boldsymbol{E} \tag{9}$$

where $\boldsymbol{U}_k = (\boldsymbol{u}_1 \ldots \boldsymbol{u}_k),\ \boldsymbol{U}_2 = (\boldsymbol{u}_{k+1} \ldots \boldsymbol{u}_{kmax}),\ \boldsymbol{V}_k = (\boldsymbol{v}_1 \ldots \boldsymbol{v}_k),\ \boldsymbol{V}_2 = (\boldsymbol{v}_{k+1} \ldots \boldsymbol{v}_{kmax}),\ \boldsymbol{S}_k = \operatorname{diag}(s_1 \ldots s_k)$ and $\boldsymbol{S}_2 = \operatorname{diag}(s_{k+1} \ldots s_{kmax})$ with

$s_1 \geq s_2 \geq \ldots \geq s_{kmax} \geq 0$. The Euclidean norm of the errors $\boldsymbol{E}$ is equal to $||\boldsymbol{E}||_E^2 = ||\boldsymbol{X} - \hat{\boldsymbol{X}}||_E^2 = \mathrm{tr}(\boldsymbol{E}'\boldsymbol{E}) = \sum_{j=k+1}^{kmax} s_j^2$, which is the minimum. Note that the least squares properties $\boldsymbol{E}'\hat{\boldsymbol{X}} = \boldsymbol{V}_2\boldsymbol{S}_2\boldsymbol{U}_2'\boldsymbol{U}_k\boldsymbol{S}_k\boldsymbol{V}_k' = 0$ and $\hat{\boldsymbol{X}}'\boldsymbol{E} = 0$ are satisfied because of the orthogonality of $\boldsymbol{U}$.

The first component of the least squares decomposition (9) becomes

$$\hat{\boldsymbol{X}} = \begin{pmatrix} \hat{\boldsymbol{x}}_1' \\ \vdots \\ \hat{\boldsymbol{x}}_T' \end{pmatrix} = \begin{pmatrix} \hat{\boldsymbol{F}}_{1,1}' \\ \vdots \\ \hat{\boldsymbol{F}}_{1,T}' \end{pmatrix} \boldsymbol{S}_k\boldsymbol{V}_k',$$

where $\boldsymbol{U}_k' \equiv (\hat{\boldsymbol{F}}_{1,1} \ldots \hat{\boldsymbol{F}}_{1,T})$ are realizations of the k factors. So we have

$$\hat{\boldsymbol{x}}_t = \boldsymbol{V}_k\boldsymbol{S}_k\boldsymbol{F}_{1,t} = \hat{\boldsymbol{B}}_1\boldsymbol{F}_{1,t}, \tag{10}$$

where $\hat{\boldsymbol{B}}_1 = (s_1\boldsymbol{v}_1 \ \ldots \ s_k\boldsymbol{v}_k)$, with condition number $c(\hat{\boldsymbol{B}}_1'\hat{\boldsymbol{B}}_1) = s_1/s_k$ and $||\frac{1}{\sqrt{T}}\hat{\boldsymbol{B}}_1||_E^2 = \frac{1}{T}\sum_{j=1}^k s_j^2 = \sum_{j=1}^k \hat{\lambda}_j$. The sample covariance matrix of the factors is $\sum_{t=1}^T \boldsymbol{F}_{1,t}'\boldsymbol{F}_{1,t} = \boldsymbol{U}_k'\boldsymbol{U}_k = \boldsymbol{I}_k$.

The residuals are equal to

$$E = \begin{pmatrix} \boldsymbol{e}_1' \\ \vdots \\ \boldsymbol{e}_T' \end{pmatrix} = \begin{pmatrix} \boldsymbol{F}_{2,1} \\ \vdots \\ \boldsymbol{F}_{2,T} \end{pmatrix} \boldsymbol{S}_2\boldsymbol{V}_2', \tag{11}$$

with $\boldsymbol{U}_2' = (\boldsymbol{F}_{2,1} \ldots \boldsymbol{F}_{2,T})$. Therefore

$$\boldsymbol{e}_t = \boldsymbol{V}_2\boldsymbol{S}_2\boldsymbol{F}_{2,t} = \hat{\boldsymbol{B}}_2\boldsymbol{F}_{2,t}$$

with $\hat{\boldsymbol{B}}_2 = (s_{k+1}\boldsymbol{v}_{k+1} \ \ldots \ s_{kmax}\boldsymbol{v}_{kmax})$ and $\sum_{t=1}^{T} \boldsymbol{F}_{2,t}' \boldsymbol{F}_{2,t} = \boldsymbol{U}_2'\boldsymbol{U}_2 = \boldsymbol{I}_{kmax-k}$. The residuals are generated by $(kmax - k)$ independent factors. The condition number of $c(\hat{\boldsymbol{B}}_2'\hat{\boldsymbol{B}}_2)$ is equal to $c(\hat{\boldsymbol{B}}_2'\hat{\boldsymbol{B}}_2) = s_{k+1}/s_{kmax}$. In addition we have

$$||\frac{1}{\sqrt{T}}\hat{\boldsymbol{B}}_2||_E^2 = \frac{1}{T}\sum_{j=k+11}^{kmax} s_j^2 = \sum_{j=k+1}^{kmax} \hat{\lambda}_j.$$

Combining (10) and (11) we get

$$\boldsymbol{x}_t = \hat{\boldsymbol{x}}_t + \boldsymbol{e}_t = \hat{\boldsymbol{B}}_1\boldsymbol{F}_{1,t} + \hat{\boldsymbol{B}}_2\boldsymbol{F}_{2,t} \tag{12}$$

with $\hat{\boldsymbol{x}}_t'\boldsymbol{e}_t = \boldsymbol{F}_{1,t}'\boldsymbol{S}_k\boldsymbol{V}_k'\boldsymbol{V}_2\boldsymbol{S}_2\boldsymbol{F}_{2,t} = 0$ because of the orthogonality of $\boldsymbol{V}$, and autocovariance

$$\hat{\boldsymbol{\Gamma}}_0 = \frac{1}{T}\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}} + \frac{1}{T}\hat{\boldsymbol{E}}'\hat{\boldsymbol{E}} = \frac{1}{T}\hat{\boldsymbol{B}}_1\hat{\boldsymbol{B}}_1' + \hat{\boldsymbol{V}}_\varepsilon,$$

with $\hat{\boldsymbol{V}}_\varepsilon = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{e}_t\boldsymbol{e}_t' = \frac{1}{T}\hat{\boldsymbol{B}}_2\hat{\boldsymbol{B}}_2'$ and $\mathrm{tr}(\hat{\boldsymbol{V}}_\varepsilon) = \frac{1}{T}\sum_{j=k+1}^{kmax} s_j^2 = \sum_{j=k+1}^{kmax}\hat{\lambda}_j$ is minimum.

Feature extraction and Karhunen-Loève expansion

In Equation (8) we observed that

$$\boldsymbol{X} = \boldsymbol{U}_{kmax}\boldsymbol{S}_{kmax}\boldsymbol{V}_{kmax}'.$$

Let $\boldsymbol{V}_k = (\boldsymbol{v}_1 \ \ldots \ \boldsymbol{v}_k)$ then $\boldsymbol{X}\boldsymbol{V}_k = \boldsymbol{U}_k\boldsymbol{S}_k$ or

$$\begin{pmatrix} \boldsymbol{x}'_1 \\ \vdots \\ \boldsymbol{x}'_T \end{pmatrix} (\boldsymbol{v}_1 \ \ldots \ \boldsymbol{v}_k) \equiv \begin{pmatrix} \boldsymbol{y}'_1 \\ \vdots \\ \boldsymbol{y}'_T \end{pmatrix} = \begin{pmatrix} \boldsymbol{F}'_{1,1} \\ \vdots \\ \boldsymbol{F}'_{1,T} \end{pmatrix} \boldsymbol{S}_k,$$

where $\boldsymbol{y}_t = \boldsymbol{S}_k\boldsymbol{F}_{1,t}, \ t = 1, \ldots, T$ is the feature vector with covariance matrix $\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{y}_t\boldsymbol{y}'_t = \boldsymbol{S}_k^2/T = \hat{\Lambda}_k$ so the components of $\boldsymbol{y}_t$ are linear combinations of the factors. Because $\begin{pmatrix} \boldsymbol{y}'_1 \\ \vdots \\ \boldsymbol{y}'_T \end{pmatrix} = (s_1\boldsymbol{u}_1 \ \ldots \ s_k\boldsymbol{u}_k)$ the j-th component of the feature vector through time $\{\boldsymbol{y}_t, t = 1, \ldots, T\}$ is $s_j\boldsymbol{u}_j$, that is the j-th singular value times the j-th principal component $\boldsymbol{u}_j \in \mathbb{R}^T$.

4 Dynamic factor models[1]

Let $\boldsymbol{x}_t$ be an N-dimensional vector of observed data at time t, $t = 1, \ldots, T$, which is driven by q dynamic factors $\boldsymbol{u}_t$ with loadings $\boldsymbol{B}_j$ up to lag p, i.e. $j = 1, \ldots, p$, and idiosyncratic components $\boldsymbol{\varepsilon}_t$

$$\boldsymbol{x}_t = \boldsymbol{B}_0\boldsymbol{u}_t + \boldsymbol{B}_1\boldsymbol{u}_{t-1} + \ldots + \boldsymbol{B}_p\boldsymbol{u}_{t-p} + \boldsymbol{\varepsilon}_t. \tag{13}$$

Equation (13) is the (dynamic) factor representation of the data. Note that factors, loadings and idiosyncratic components are not observable. In vector

[1] This section draws upon Jacobs and Otter (2006).

notation the model becomes

$$\boldsymbol{x}_t = \begin{pmatrix} \boldsymbol{B}_0 & \boldsymbol{B}_1 & \dots & \boldsymbol{B}_p \end{pmatrix} \begin{pmatrix} \boldsymbol{u}_t \\ \boldsymbol{u}_{t-1} \\ \vdots \\ \boldsymbol{u}_{t-p} \end{pmatrix} + \boldsymbol{\varepsilon}_t \equiv \boldsymbol{B}\boldsymbol{F}_t + \boldsymbol{\varepsilon}_t. \tag{14}$$

We make the following assumptions. First, the q-dimensional vector of factors is Gaussian White Noise (GWN) with $E(\boldsymbol{u}_t) = 0$ and $\text{var}(\boldsymbol{u}_t) = I_q$, the q-dimensional identity matrix. Secondly, the idiosyncratic components $\boldsymbol{\varepsilon}_t$ is GWN with $E(\boldsymbol{\varepsilon}_t) = 0$ and $\text{var}(\boldsymbol{\varepsilon}_t) = \boldsymbol{V}$ and factors $\boldsymbol{u}_t$ and idiosyncratic components $\boldsymbol{\varepsilon}_t$ are independent. This assumptions imply that the generalized dynamic factor model of Forni, Hallin, Lippi and Reichlin (2000a) and Forni and Lippi (2001), which allows some correlation among idiosyncratic components, can be dealt with too. Thirdly, the matrix of loadings $\boldsymbol{B}$ has full (column-)rank, i.e. $\text{rank}(\boldsymbol{B}) = (p+1)q$ with $(p+1)q < N$.

In the sequel we give a procedure based on canonical correlation to determine the information content in the set of autocovariances $\{\boldsymbol{\Gamma}_i,\ i = 0, 1, 2\dots\}$ with $\boldsymbol{\Gamma}_i = \mathrm{E}(\boldsymbol{x}_t\boldsymbol{x}'_{t-i})$ together with the dimension of the dynamic factors (q) and the lag order (p). The basic idea is the following. Let

$$\begin{pmatrix} \boldsymbol{x}_t \\ \boldsymbol{x}_{t-i} \end{pmatrix} \sim \mathcal{N}_{2N}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Gamma}_0 & \boldsymbol{\Gamma}_i \\ \boldsymbol{\Gamma}'_i & \boldsymbol{\Gamma}_0 \end{pmatrix} \right),$$

where $\{\boldsymbol{x}_t,\ t = 1, 2, \dots\}$ is assumed to be stationary. The canonical correlation procedure linearly transforms the $2N$-dimensional vector into the

$2N$-dimensional vector

$$\begin{pmatrix} \boldsymbol{y}_t \\ \boldsymbol{y}_{t-i} \end{pmatrix} \sim \mathcal{N}_{2N}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{I}_N & \boldsymbol{S}_i \\ \boldsymbol{S}_i & \boldsymbol{I}_N \end{pmatrix} \right),$$

where $\boldsymbol{S}_i$ is a diagonal matrix, i.e. $\boldsymbol{S}_i = \mathrm{diag}(s_{i,1} \ \ldots \ s_{i,N})$, where $\{s_{i,j}, \ j = 1, \ldots, N\}$ are the canonical correlation coefficients with $0 \leq s_{i,j} \leq 1$. Using Bartlett's test statistic a subset of the canonical correlation coefficients are tested against zero. In case $\boldsymbol{S}_i = \mathbf{0}$ is accepted, the conditional entropy of $\boldsymbol{y}_t$ given $\boldsymbol{y}_{t-i}$ is $2H_y^{max} = cN$, whereas if $\boldsymbol{S}_i \neq \mathbf{0}$ the conditional entropy of $\boldsymbol{y}_t$ equals $2H_y = cN + \mathrm{logdet}(\boldsymbol{I} - \boldsymbol{S}_i^2)$ and hence

$$2I(\boldsymbol{y}_t|\boldsymbol{y}_{t-i}) = -\mathrm{logdet}(\boldsymbol{I} - \boldsymbol{S}_i^2) = -\sum_{j=1}^{\bar{N} \leq N} \ln(1 - s_{i,j}^2), \qquad (15)$$

where $\bar{N}$ is the number of significant canonical correlation coefficients as outcomes of the testing procedure for p and q given below.

Estimation of p and q

First we demean the N components of our data matrix $\boldsymbol{x}_t$ and take unit variances, obtaining $\tilde{\boldsymbol{x}}_t$. Let

$$\hat{\boldsymbol{\Gamma}}_i = \frac{1}{T-i} \sum_{t=i}^{T} \tilde{\boldsymbol{x}}_t \tilde{\boldsymbol{x}}'_{t-i}, \quad i = 0, 1, 2, \ldots$$

be a consistent estimate of the autocovariances $\boldsymbol{\Gamma}_i$. Assuming that rank($\hat{\boldsymbol{\Gamma}}_0$) has full rank N, we can apply the spectral decomposition

$$\hat{\boldsymbol{\Gamma}}_0 = \boldsymbol{C}\boldsymbol{\Lambda}^{1/2}\boldsymbol{\Lambda}^{1/2}\boldsymbol{C}' = \hat{\boldsymbol{\Gamma}}_0^{1/2}(\hat{\boldsymbol{\Gamma}}_0^{1/2})', \tag{16}$$

with $\hat{\boldsymbol{\Gamma}}_0^{-1/2} = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{C}'$ and the Singular Value Decomposition (SVD)

$$\hat{\boldsymbol{\Gamma}}_0^{-1/2}\hat{\boldsymbol{\Gamma}}_i(\hat{\boldsymbol{\Gamma}}_0^{-1/2})' = \boldsymbol{H}_i\boldsymbol{S}_i\boldsymbol{Q}_i \tag{17}$$

where the columns of $\boldsymbol{H}$ and $\boldsymbol{Q}$ are orthogonal, i.e. $\boldsymbol{H}_i'\boldsymbol{H}_i = \boldsymbol{H}_i\boldsymbol{H}_i' = \boldsymbol{I}_N$ and $\boldsymbol{Q}_i'\boldsymbol{Q}_i = \boldsymbol{Q}_i\boldsymbol{Q}_i' = \boldsymbol{I}_N$, and $\boldsymbol{S}_i = \text{diag}(s_{i,1}, s_{i,2}, \ldots, s_{i,N})$, an $N \times N$ diagonal matrix with singular values $s_{i,j} \in [0,1]$ with $s_{i,1} > s_{i,2} > \ldots > s_{i,N} \geq 0$. The canonical correlation coefficients (singular values) $s_{i,j}$ are estimates of the equivalent population canonical coefficients $\rho_{i,j}$, see Otter (1990, 1991).

To test the null hypothesis that the $N-k$ smallest population canonical coefficients for the ith autocovariance are equal to zero

$$H_0 : \rho_{i,k+1} = \ldots = \rho_{i,N} = 0,$$

we calculate Bartlett's test statistic

$$\chi^2 = -\left[T - \frac{1}{2}(2N+1)\right]\sum_{j=k+1}^{N}\ln(1-s_{i,j}^2), \tag{18}$$

for all values of $k = 0, 1, 2, \ldots$. This statistic follows a χ^2 distribution under the null with degrees of freedom $df = (N-k)^2$.

Test procedure

The procedure essentially comes down to the linear transformation of $\tilde{\boldsymbol{x}}_t$ and $\tilde{\boldsymbol{x}}_{t-i}$ into canonical vectors $\boldsymbol{y}_t = \boldsymbol{A}_i\tilde{\boldsymbol{x}}_t$ and $\boldsymbol{y}_{t-i} = \boldsymbol{G}_i\tilde{\boldsymbol{x}}_{t-i}$ with $\boldsymbol{A}_i = \boldsymbol{H}_i'\hat{\boldsymbol{\Gamma}}_0^{-1/2}$ and $\boldsymbol{G}_i = \boldsymbol{Q}_i\hat{\boldsymbol{\Gamma}}_0^{-1/2}$ with $E(\boldsymbol{y}_t\boldsymbol{y}_{t-i}') = \boldsymbol{S}_i$ and unit variance matrices. The conditional variance of $\boldsymbol{y}_t$ given $\boldsymbol{y}_{t-1}$ equals $(I_N - \boldsymbol{S}_i^2)$ and hence the estimated information is given by Equation (15) above. From Equation (14) and the normalisation we have $\boldsymbol{\Gamma}_i = \boldsymbol{D}^{-1/2}\boldsymbol{B}^{(i)}\boldsymbol{B'}\boldsymbol{D}^{-1/2}$ where $\boldsymbol{D} = \mathrm{diag}(\sigma_1^2, ..., \sigma_N^2)$ with the variances of the components of $\boldsymbol{x}_t$ as elements. The $N \times (p+1)q$ dimensional matrix $\boldsymbol{B}^{(i)} = (\boldsymbol{B}_i \ \boldsymbol{B}_{i+1} \ \ldots \ \boldsymbol{B}_p \ \boldsymbol{0} \ \ldots \ \boldsymbol{0})$ has rank $(p+1-i)q$ and hence the rank of $\boldsymbol{\Gamma}_i$ is $(p+1-i)q$ for lags $i = 1, .., p$ and zero for lags greater than p. The ranks of $\hat{\boldsymbol{\Gamma}}_i$ are estimated by the number of significant singular values using Bartlett's test statistic as follows.

If for a given significance level the hypothesis that all population canonical coefficients for the $(p+i)$th autocovariance $(i > 0)$ are equal to zero, i.e. $H_0 : \rho_{p+i,1} = \rho_{p+i,2} = \ldots = \rho_{p+i,N} = 0$, is accepted whereas the hypothesis that all population canonical coefficients for the (p)th autocovariance are equal to zero, $H_0 : \rho_{p,1} = \rho_{p,2} = \ldots = \rho_{p,N} = 0$ is rejected, but the hypothesis that the $(N-q)$ smallest canonical coefficients are equal to zero, $H_0 : \rho_{p,q+1} = \rho_{p,q+2} = \ldots = \rho_{p,N} = 0$, is accepted, then the estimated lag order is p and the estimated number of factors or $\dim(\boldsymbol{u}_t)$ equals q.

5 Simulation

To illustrate our procedures, we run some simple simulation experiments in MATLAB along the lines of Bai and Ng (2002). We simulate data from the dynamic factor model of Equation (13)

$$x_{it} = \sum_{l=1}^{p} \sum_{j=1}^{q} b_{ilj} u_{jt-l} + \sigma \varepsilon_{it}$$

for different maximum lag orders p, common factors q, and noise standard deviation σ. The static factor model that corresponds to this dynamic one has $k = (p+1)q$ factors. The elements of the loadings b_{ilj} are independent drawings from a uniform distribution in the interval $[-2, +2]$. All time-series are standardised, i.e. demeaned and scaled to unit variance. We assume $N < T$, in particular $T = N^{\alpha},\ \alpha > 1$ in line with Forni et al. (2004).

Table 1 lists the first seven eigenvalues and the last eigenvalue of the autocovariance matrix $\boldsymbol{\Gamma}(0)$ for $k = 4,\ 6$, $\sigma = 1,\ 5$ and different values of the number of variables N and the number of observations $T \equiv N^{1.1}$, together with the information measures KL the discrimination function of Equation (4), information I_x as defined in Equation (1), maximum entropy H_x^{max}, the entropy of the eigenvalues $H_{\bar{\lambda}}$ as defined in Equation (5), and the maximum entropy of the eigenvalues $H_{\bar{\lambda}}^{max}$. The larger the number of variables and observations, the more information is in the factor model as can be seen from higher values of the eigenvalues larger than one and higher values of the information measures in the last five columns. By construction the eigenvalues are decreasing. For $\sigma = 1$ the fifth (seventh) eigenvalue becomes equal to or smaller than one for $k = 4,\ 6$. With high noise $\sigma = 5$, this pattern is less clear. The fifth (and seventh) eigenvalue jumps to a smaller value, but does not become smaller than one.

Table 1: Eigenvalues and information in static factor model

k	σ	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_N	KL	I_x	H_x^{max}	$H_{\bar{\lambda}}$	$H_{\bar{\lambda}}^{max}$
		$N = 50,\ T = 74$												
4	1	13.7	11.7	8.1	7.1	0.71	0.63	0.58	0.006	901	94	142	2.19	3.91
	5	4.8	4.3	3.2	2.5	2.2	2.1	1.9	0.027	152	30	142	3.42	3.91
6	1	12.1	10.1	8.8	5.6	4.4	3.4	0.45	0.0046	1430	111	142	2.18	3.91
	5	4.6	3.7	3.4	2.8	2.5	2.2	2.0	0.035	138	29	142	3.42	3.91
		$N = 100,\ T = 158$												
4	1	26.3	22.1	18.6	15.4	0.70	0.66	0.59	0.0072	1693	205	283	2.28	4.60
	5	6.9	5.9	5.4	4.4	2.4	2.3	2.2	0.0443	226	55	283	4.11	4.60
6	1	19.9	17.2	16.1	14.4	11.5	9.2	0.4	0.0037	2394	232	283	2.35	4.60
	5	7.2	6.0	5.6	4.7	4.2	3.5	2.0	0.0392	254	60	283	4.05	4.60
		$N = 200,\ T = 340$												
4	1	48.4	45.0	39.3	31.6	0.96	0.86	0.71	0.0091	2976	417	567	2.44	5.29
	5	12.2	11.0	10.6	9.0	2.4	2.3	2.3	0.0497	386	108	567	4.71	5.29
6	1	35.6	33.4	29.9	29.1	26.3	21.1	0.5	0.0065	4078	472	567	2.51	5.29
	5	12.3	10.8	8.7	8.3	7.8	6.8	2.3	0.0424	429	117	567	4.67	5.29
		$N = 500,\ T = 931$												
4	1	110.7	102.5	96.3	95.3	1.0	0.8	0.8	0.0109	6124	*	1419	2.71	6.21
	5	24.3	24.1	22.2	18.3	2.4	2.3	2.3	0.0637	794	248	1419	5.58	6.21
6	1	85.1	79.9	77.8	71.9	66.2	58.9	0.71	0.0079	9374	*	1419	2.63	6.21
	5	23.6	22.0	21.0	20.1	18.8	16.4	2.25	0.0558	902	282	1419	5.46	6.21

Notes: $T = N^{1.1}$; $k \equiv (p+1)q$ is the number of factors in the static factor model, where p is the maximum lag order, and q is the number of common factors; σ is the variance of the noise; $\lambda_1, \ldots, \lambda_7$ are the first seven eigenvalues of the autocovariance matrix $\boldsymbol{\Gamma}(0)$, λ_N is the last eigenvalue; KL is the discrimination function of Equation (4); I_x is information as defined in Equation (1); H_x^{max} is maximum entropy; $H_{\bar{\lambda}}$ is entropy of the eigenvalues, Equation (5); and $H_{\bar{\lambda}}^{max}$ is maximum entropy of the eigenvalues.
An $*$ indicates that I_x is not defined because $\boldsymbol{\Gamma}(0)$ is singular.

Table 2 provides information on the dynamic factor model. Again we simulate data from the dynamic factor model of Equation (13). For $p = 2$, $q = 2$ and different combinations of N and T, we calculate maximum entropy H_{max}, the number of significant singular values $nsig$, the information in the autocovariance matrix $I(\boldsymbol{y}_t|\boldsymbol{y}_{t-l})$ at the first four lags as defined in Equation (15), and I_x, the information in $\boldsymbol{\Gamma}_0$. The table allows the following observations. First, the information in $\boldsymbol{\Gamma}_0$ is significantly larger than the information in the autocovariances at the first four lags. Secondly, the information in $\boldsymbol{\Gamma}_0$ increases with the number of variables and observations. At $N = 50$, $T = 354$ and $\sigma = 1$ information in the autocovariance matrix $\boldsymbol{\Gamma}_0$ (I_x) is equal to 85.37, whereas I_x rises to 193.5 for $N = 100$, $T = 1000$. However, the rather limited additional information suggests that the number of variables need not be very large to get reasonable precise factor estimates, as mentioned in the Introduction. Thirdly, the information in $\boldsymbol{\Gamma}_0$ decreases with noise; for example with N=100 and T=1000, I_x goes from 193.5 if σ equals one, to 54.3 and 23.2 for σ equal to three and five. Finally, the number of significant singular values of the autocovariance matrix becomes zero after two lags. Exceptions occur because the outcomes in each row are based on a single simulation run.

Table 2: Information in dynamic factor model ($q = 2,\ p = 2,\ k = 6$)

	H_{max}	$nsig$	$I(\boldsymbol{y}_t \mid \boldsymbol{y}_{t-l})$	$nsig$	$I(\boldsymbol{y}_t \mid \boldsymbol{y}_{t-l})$	$nsig$	$I(\boldsymbol{y}_t \mid \boldsymbol{y}_{t-l})$
$N = 50,\ T = 354$							
		$\sigma = 1$		$\sigma = 3$		$\sigma = 5$	
I_x			85.37		21.34		10.74
$l = 1$	141.9	5	16.41	4	7.10	4	4.57
2	141.9	2	7.89	2	3.44	2	2.14
3	141.9	0	0	1	0.66	1	0.63
4	141.9	0	0	0	0	0	0
$N = 100,\ T = 1000$							
		$\sigma = 1$		$\sigma = 3$		$\sigma = 5$	
I_x			193.5		54.3		23.2
$l = 1$	283.8	4	17.9	4	9.3	4	5.8
2	283.8	2	9.2	2	4.9	2	2.9
3	283.8	0	0	0	0	0	0
4	283.8	0	0	0	0	0	0
$N = 38,\ T = 180$							
		$\sigma = 1$		$\sigma = 3$		$\sigma = 5$	
I_x			62.75		17.77		7.61
$l = 1$	107.84	5	15.77	4	6.65	2	2.23
2	107.84	3	9.27	2	3.61	3	3.32
3	107.84	1	1.05	0	0	1	1.13
4	107.84	0	0	1	1.15	1	1.03

Notes: H_{max} is maximum entropy; $nsig$ is the number of significant singular values of the autocovariance at lag l ; $I(\boldsymbol{y}_t|\boldsymbol{y}_{t-l})$ is information in the autocovariance matrix with lag l as defined in Equation (15); I_x is the information in the autocovariance matrix $\boldsymbol{\Gamma}_0$ as defined in Equation (1).

The bottom panel of Table 2 reports information outcomes for the data set dimensions of Inklaar et al. (2005), i.e. $N = 38$ and $T = 180$. To illustrate the procedure for obtaining estimates of the number of dynamic factors q and lags p, we run 1000 replications of the dynamic factor model of Equation (13) for different combinations of (p, q, σ) for the same data set dimensions. Table 3 lists the average number of significant canonical correlation coefficients (at the 5% significance level) for autocovariances up to and including the fifth lag. Due to the large number of degrees of freedom the Bartlett test statistic of Equation (18) has been replaced by the standardised z-statistic. In addition, we print the signal-to-noise ratio (SN) for each component $x_{i,t}$, which is defined as the ratio of the common variance due to the factors and the variance due to the noise, i.e, $SN = 4(p+1)q/3\sigma^2$ and information I_x in the autocovariance matrix $\boldsymbol{\Gamma}_0$ as defined in Equation (1).

Table 3 shows that for most of the combinations of p and q with low noise the procedure performs well especially in the estimation of the lag order (a drop after lag p) with for lag $p+1$ an average of significant singular values less than one. For example, the average number of singular values for $(p, q, \sigma) = (1, 1, 1)$ drops from 1.31 for autocovariances at one lag to 0.41 at lag two.

As explained in the test procedure, one expects a drop in the number of significant singular values from $((p - i)q + q)$ to $(p - i)q$ which equals q, the number of dynamic factors, if the lag is increased from from i to $(i + 1)$ (for i not greater than p). So, subtracting the number of significant singular values at lag $(i + 1)$ from the number at lag i provides a method to estimate q. Consider for example the combination ($q = 2$, $p = 3$, $\sigma = 1$

Table 3: Dynamic factor model: Monte Carlo simulations ($N = 38,\ T = 180$, 1000 replications)

q	p	σ	avg. of sign. sing. vals at autocov. with lag					SN	I_x
			1	2	3	4	5		
1	1	1	1.31	0.41	0.55	0.65	0.74	2.67	38.76
2	1	1	2.25	0.45	0.55	0.66	0.78	5.33	55.82
3	1	1	3.29	0.41	0.52	0.62	0.76	8.00	58.98
4	1	1	4.26	0.44	0.51	0.63	0.78	10.67	64.8
5	1	1	5.22	0.41	0.53	0.62	0.77	13.33	67.12
6	1	1	6.22	0.43	0.54	0.63	0.74	16.00	67.70
2	1	1	2.25	0.45	0.55	0.66	0.78	5.33	55.82
2	2	1	4.27	2.40	0.52	0.65	0.78	8.00	64.92
2	3	1	6.20	4.33	2.46	0.61	0.80	10.67	70.36
2	4	1	8.19	6.29	4.41	2.57	0.76	13.33	65.30
2	4	3	7.76	5.94	4.20	2.39	0.77	1.48	18.18
2	4	5	5.90	4.52	3.37	2.08	0.79	0.53	10.90
2	4	7	3.69	2.90	2.21	1.53	0.76	0.27	6.64

Notes: p is the maximum lag order; q is number of common factors; σ is the variance of the noise; SN is the signal-to-noise ratio of the variance due to the factors and the variance due to the noise, $SN = 4(p+1)q/3\sigma^2$: and I is information as defined in Equation (1).

in Table 3. Subtracting the number of singular values at lag i from those at lag $(i+1)$ results in (1.87 1.87 1.85 -0.19), producing an estimate of two dynamic factors. The estimation of the number of factors shows a slight overestimation in case of low noise and a serious systematic underestimation in case of increased noise. Take for example (p, q) equal to (2,4). A noise level of 1 leads on average to 8.19 singular values which is close to the expected value $pq = 8$, whereas a higher noise level of 7 gives 3.69 singular values on average, as expected because of the very low signal-to-noise ratio.

6 Conclusion

This paper has shown that concepts from information theory can fruitfully be applied in the analysis of factor models. The information in the data set can be obtained from the autocovariance matrix. Using Kullback-Leibler numbers we demonstrated that the whole distribution of the eigenvalues of the autocovariance matrix contributes to the information and not only the largest ones. In addition we calculated information in static and dynamic factor models, which enabled us to work out whether an additional variable adds information and to estimate the optimal number of dynamic factors q and lag p. To illustrate the concepts we run simulation experiments with static and dynamic factor models for some ad hoc data set dimensions.

Kullback-Leibler numbers are related to the Akaike information criterion (AIC). Future research will look into the relation between our methods and the information criteria of Bai and Ng (2002). Besides we plan to address asymptotics, i.e. N and T going to infinity, and put our methods to the test with ‘real’ data sets like for example the Morkmon (Den Reijer, 2005) or Eurocoin (Altissimo et al., 2001) data sets. A practical complication of both data sets is that the number of variables exceeds the number of observations. This can in principle be handled by the least squares procedure given in Section 3. Another option is to order the variables according to correlation, doing the analysis on the first forty, say, variables, and checking whether additional variables contain information using the relative information formulas mentioned in the same section.

References

Altissimo, F., A. Bassanetti, R. Cristadoro, M. Forni, M. Lippi, L. Reichlin, and G. Veronese (2001), “Eurocoin: A real time coincident indicator of the euro area business cycle”, Discussion Paper 3108, Centre for Economic Policy Research.

Bai, J. and S. Ng (2002), “Determining the number of factors in approximate factor models”, *Econometrica*, **70**, 191–221.

Boivin, J. and S. Ng (2003), “Are more data always better for factor analysis?”, Working Paper 9829, National Bureau of Economic Research [forthcoming in *Journal of Econometrics*].

Burnham, K.P. and D.R. Anderson (2002), *Model selection and multimodel inference : a practical information-theoretic approach*, 2nd edition, Springer, New York.

Camba-Mendez, G., G. Kapetanios, R.J. Smith, and M.R. Weale (2001), “An automatic leading indicator of economic activity: forecasting GDP growth for European countries”, *Econometrics Journal*, **4**, S56–S90.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2000a), “The generalized dynamic-factor model: Identification and estimation”, *The Review of Economics and Statistics*, **82**, 540–554.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2004), “The generalized dynamic factor model: consistency and rates”, *Journal of Econometrics*, **119**, 231–255.

Forni, M. and M. Lippi (2001), "The generalized factor model: Representation theory", *Econometric Theory*, **17**, 1113–1141.

Goodwin, G.C. and R.L. Payne (1977), *Dynamic system identification : experiment design and data analysis*, Academic Press, New York, London.

Inklaar, R.C., Jan J.P.A.M. Jacobs, and W.E. Romp (2005), "Business cycle indexes: Does a heap of data help?", *Journal of Business Cycle Measurement and Analysis*, **1**, 309–336.

Jacobs, J.P.A.M. and P.W. Otter (2006), "Determining the number of factors and lag order in dynamic factor models: A minimum entropy approach", *Econometric Reviews* [forthcoming].

Kapetanios, G. (2004), "A new method for determining the number of factors in factor models with large dimensions", Working Paper 525, Queen Mary University of London.

Kapetanios, G. (2005), "A testing procedure for determining the number of factors in approximate factor models with large dimensions", Working Paper 551, Queen Mary University of London.

Otter, P.W. (1990), "Canonical correlation in multivariate time series analysis with an application to one-year-ahead and multiyear-ahead macroeconomic forecasting", *Journal of Business and Economic Statistics*, **8**, 453–457.

Otter, P.W. (1991), "On Wiener-Granger causality, information and canonical correlation", *Economics Letters*, **35**, 187–191.

Reijer, A.H.J den (2004), “Forecasting dutch GDP using large scale factor models”, Working Paper 28/2005, De Nederlandsche Bank.

Stock, J.H. and M.W. Watson (2002a), “Diffusion indexes”, *Journal of the American Statistical Association*, **97**(460), 147–162.

Stock, J.H. and M.W. Watson (2002b), “Macroeconomic forecasting using diffusion indexes”, *Journal of Business and Economic Statistics*, **20**(2), 147–162.

Young, T.Y. and T.W. Calvert (1974), *Classification, estimation and pattern recognition*, Elsevier, New York; London.