

Information, data dimension and factor structure

Jan P.A.M. Jacobs

University of Groningen

Pieter W. Otter

University of Groningen

Ard H.J. den Reijer*

Sveriges Riksbank

Revised January 2010

Abstract

This paper employs concepts from information theory to choosing the dimension of a data set. We propose a relative information measure connected to Kullback-Leibler numbers. By ordering the series of the data set according to the measure, we are able to obtain a subset of the data set that is most informative to model a variable of interest. The method can be used as a first step in the construction of a dynamic factor model or a leading index, as illustrated with the U.S. macroeconomic data set of Stock and Watson (2005).

Keywords: Kullback-Leibler numbers, information, factor structure, data set dimension, dynamic factor models, leading index

JEL-code: C32, C52, C82

*Corresponding author: Ard H.J. den Reijer, Sveriges Riksbank, Modelling Division, 103 37 Stockholm, Sweden. Tel.: +46 8 787 0149. Email: ard.den.reijer@riksbank.se

1 Introduction

With the proliferation of huge data sets a natural question to ask is how much information there is in a data set. Is there an ‘optimal’ size of the data set in relation to some variable(s) of interest, in other words can we confine attention to a subset of the series instead of having to monitor all series in a data set? The question seems especially relevant for factor models, which exploit the idea that movements in a large number of series are driven by a limited number of common ‘factors’. For a recent overview see Bai and Ng (2008a).

Although convergence of factor estimates requires large cross-sections and large time dimensions, see e.g. Forni and Lippi (2001) and Bai (2003), the data set need not be very large to obtain reasonably precise factor estimates. Boivin and Ng (2006) 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) also conclude that the number of series need not be very large to get precise factor estimates. The question whether we can confine attention to a subset of the variables is also relevant for the construction of leading indexes, which aims at selecting indicators with predictive power out of a large number of candidates too.

Our paper is related to Bai and Ng (2008b), who use ‘hard’ and ‘soft’ thresholding to reduce the influence of uninformative predictors for a variable. Hard thresholding involves some pretest procedure, while under soft

thresholding the top ranked predictors according to some soft-thresholding rule are kept.

Building upon Otter and Jacobs (2006), the paper exploits concepts from information theory, in particular Kullback-Leibler numbers, to analyse information in the data.¹ We propose a relative information measure based on Gaussian distributed data with a clear link to Kullback-Leibler numbers. The measure is discussed in more detail assuming an approximate factor structure in the data; a test procedure is given whether an additional variable adds information. Ordering the series of the data set according to the measure enables us to identify a subset of the data set that is most informative to modelling a variable of interest. The method can be used as a first step in the construction of a dynamic factor model or a leading index.

We illustrate the concepts with the macroeconomic data set of Stock and Watson (2005), which consists of 132 monthly U.S. variables and runs from 1959–2003. We find that relative information is maximized for 40–50 series if we are interested in modelling industrial production and CPI inflation. Allowing for pure leads only or both leads and lags does not affect the size of the subset to a great extent.

The paper is structured as follows. Section 2 discusses our relative information measure, how it works out assuming an approximate factor structure in the data, and presents a test procedure. Section 3 applies our method to the U.S. data set. Section 4 concludes.

¹Jacobs and Otter (2008) apply similar information concepts to derive a formal test for the number of common factors and the lag order in a dynamic factor model.

2 Information in data

2.1 Kullback-Leibler numbers and information

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

$$G_1 = E_{f_1} \left(\log \left(\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \right) \right) \text{ and } G_2 = E_{f_2} \left(\log \left(\frac{f_2(\mathbf{x})}{f_1(\mathbf{x})} \right) \right), \quad (1)$$

and $G = G_1 + G_2$ is the measure of information for discriminating between the two density functions with $G = 0$ in case $f_1(\mathbf{x}) = f_2(\mathbf{x})$ and $G = \infty$ in case of perfect discrimination, see Young and Calvert (1974, p245). For a general background see Burnham and Anderson (2002).

For $\text{tr}(\mathbf{\Gamma}) = \text{tr}(\mathbf{A}) = N$ we have $G_1 = -\log\det(\mathbf{A})$ and $G_2 = \log\det(\mathbf{A}) + \frac{1}{2}(\text{tr}(\mathbf{A}^{-1}) - N)$. Therefore

$$2G = \text{tr}(\mathbf{A}^{-1}) - N = \text{tr}(\mathbf{A}^{-1}) - \text{tr}(\mathbf{A}) = \sum_{j=1}^N \frac{(1 - \lambda_j^2)}{\lambda_j}, \quad (2)$$

from which it can be seen that G is small (not discriminating) if the eigenvalues λ_j are close to 1, but becomes large (discriminating) for “small” eigenvalues.

For the Gaussian case we can use alternative measures of entropy and information. Let \mathbf{x}_t again be an N -dimensional vector of observed data at time t , $t = 1, \dots, T$. The data is demeaned and normalized, and normally

distributed with mean zero and variance $\mathbb{E}(\mathbf{x}_t \mathbf{x}_t') = \mathbf{\Gamma}$, i.e. $\mathbf{x}_t \sim \mathbb{N}(\mathbf{0}, \mathbf{\Gamma})$, where $\text{diag}(\mathbf{\Gamma}) = (1, 1, \dots, 1)$ and $\text{tr}(\mathbf{\Gamma}) = N$. Here we make the additional assumption that all eigenvalues are positive. The entropy as measure of disorder for a stationary, normally distributed vector is given by

$$2H_x = cN + \log\det(\mathbf{\Gamma}),$$

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

$$2\text{Inf}_x \equiv 2(H_{x,max} - H_x) = -\log\det(\mathbf{\Gamma}) \geq 0, \quad (3)$$

which is zero in case $\mathbf{\Gamma} = \mathbf{I}_N$. We define the *relative information* as

$$\text{Inf}_N^R = \frac{2H_{max} - 2H_{x(N)}}{2H_{max}} = \frac{2\text{Inf}_N}{2H_{max}} = \frac{2\text{Inf}_N}{cN}. \quad (4)$$

If $H_{x(N)}$ is equal to H_{max} then $\text{Inf}_N^R = 0$; if $H_{x(N)} = 0$ then $\text{Inf}_N^R = 1$.

2.2 Relative information measure Inf_N^R in the approximate factor model

In this section we consider the relative information measure Inf_N^R in more detail assuming an approximate factor structure in the data. Let \mathbf{x}_t be driven by k factors

$$\mathbf{x}_t = \mathbf{B}_N \mathbf{F}_t + \boldsymbol{\varepsilon}_t, \quad \mathbf{x}_t \in \mathbb{R}^N, \quad \mathbf{F}_t \sim \mathcal{N}_k(\mathbf{0}, \mathbf{I}_k), \quad \boldsymbol{\varepsilon}_t \sim \mathcal{N}_N(\mathbf{0}, \boldsymbol{\Psi}_{11}), \quad (5)$$

where $\mathbf{B}_N \in \mathbb{R}^{N \times k}$ is the matrix of factor loadings. Note that this approximate factor model is sufficiently general to cover the static and the dynamic case. The variance between the first N elements of \mathbf{x}_t is equal to $\mathbf{\Gamma}(N) = \mathbf{B}_N \mathbf{B}'_N + \mathbf{\Psi}_{11}$.

Adding a variable $x_{N+1,t}$ we have

$$\begin{pmatrix} \mathbf{x}_t \\ x_{N+1,t} \end{pmatrix} = \begin{pmatrix} \mathbf{B}_N \\ \mathbf{b}_{N+1} \end{pmatrix} \mathbf{F}_t + \begin{pmatrix} \boldsymbol{\varepsilon}_t \\ \varepsilon_{N+1,t} \end{pmatrix}, \quad (6)$$

with covariance $\mathbf{\Gamma}(N+1) = \begin{pmatrix} \mathbf{\Gamma}(N) & \mathbf{\Gamma}_{12} \\ \mathbf{\Gamma}_{21} & 1 \end{pmatrix}$, where $\mathbf{\Gamma}_{12} = \mathbf{B}_N \mathbf{b}'_{N+1} + \mathbf{\Psi}_{12}$ with $\mathbf{\Psi}_{12} = \text{E}(\boldsymbol{\varepsilon}_t \varepsilon_{N+1,t})$. Because of the normalisation we have $\mathbf{b}_{N+1} \mathbf{b}'_{N+1} + \sigma_{N+1}^2 = 1$, where $\sigma_{N+1}^2 = \text{E}(\varepsilon_{N+1,t}^2)$. Using the rule of determinants for partitioned matrices we get

$$\det(\mathbf{\Gamma}(N+1)) = \det(\mathbf{\Gamma}(N))(1 - a_{N+1}),$$

with $a_{N+1} \equiv (\mathbf{b}_{N+1} \mathbf{B}'_N + \mathbf{\Psi}'_{12}) \mathbf{\Gamma}^{-1}(N) (\mathbf{B}_N \mathbf{b}'_{N+1} + \mathbf{\Psi}_{12})$ and $0 \leq (1 - a_{N+1}) \leq 1$.

After some calculations the following relation between the relative information measures Inf_{N+1}^R and Inf_N^R can be established:

$$\text{Inf}_{N+1}^R = \text{Inf}_N^R - \frac{1}{N+1} \left(\frac{\log(1 - a_{N+1})}{c} + \text{Inf}_N^R \right). \quad (7)$$

Therefore a variable $x_{N+1,t}$ adds relative information, i.e. $\text{Inf}_{N+1}^R > \text{Inf}_N^R$, if $-\log(1 - a_{N+1}) > c \text{Inf}_N^R$, provided $\text{E}(x_{N+1,t} \mathbf{x}'_t) = (\mathbf{b}_{N+1} \mathbf{B}'_N + \mathbf{\Psi}'_{12}) = \mathbf{\Gamma}'_{12} \neq 0$. The latter condition can be tested by means of the procedure described in Sec-

tion 2.3 below. Rewriting a_{N+1} as $a_{N+1} = \mathbb{E}(x_{N+1,t}\mathbf{x}'_t)\mathbf{C}\mathbf{\Lambda}^{-1}\mathbf{C}'\mathbb{E}(x_{N+1,t}\mathbf{x}'_t)'$ it is seen that the covariance between $x_{N+1,t}$ and \mathbf{x}_t is amplified by the inverse of the eigenvalues consistent with the criterion based on the KL-numbers, see Equation (2).

The second term on the right-hand side of Equation (7) serves as a threshold which can be seen as follows. From the equation we have $\text{Inf}_{N+1}^R = \text{Inf}_N^R$ if $a_{N+1} = 1 - \exp(-c\text{Inf}_N^R)$. Whenever Inf_N^R is close to zero, Inf_{N+1}^R increases for relative small values of a_{N+1} whereas if Inf_N^R is close to one, a_{N+1} should be close to one to allow $x_{N+1,t}$ to add information. Note that the threshold in Equation (7) is determined by the covariance between $x_{N+1,t}$ and \mathbf{x}_t , $\mathbf{\Lambda}^{-1}$ which measures the degree of correlation between the components of \mathbf{x}_t consistent with the KL-measure G , the magnitude of Inf_N^R , and the dimension of the data vector.

Equation (7) can be simplified by the following procedure. Let $\mathbf{\Gamma}(N) = \mathbf{C}\mathbf{\Lambda}\mathbf{C}'$ with $\mathbf{\Gamma}(N)$ regular and consider the linear transforms $\tilde{\mathbf{x}}_t = \mathbf{U}'\mathbf{\Lambda}^{-1/2}\mathbf{C}'\mathbf{x}_t$ and $\tilde{x}_{N+1,t} = v^{-1}x_{N+1,t}$ with \mathbf{U} orthogonal, i.e. $\mathbf{U}'\mathbf{U} = \mathbf{U}\mathbf{U}' = \mathbf{I}_N$, and $v^2 = 1$ obtained by the singular value decomposition (SVD) $\mathbf{\Lambda}^{-1/2}\mathbf{C}'\mathbf{\Gamma}_{12} = \mathbf{U}\mathbf{\Sigma}v$ with $\mathbf{\Sigma} = (\phi_1 \ 0 \ \dots \ 0)'$. From this SVD it can be seen that $\mathbf{\Gamma}_{12} = 0$ implies $\mathbf{\Sigma} = 0$. Then

$$\begin{pmatrix} \tilde{\mathbf{x}}_t \\ \tilde{x}_{N+1,t} \end{pmatrix} \sim \mathcal{N}_{N+1} \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tilde{\mathbf{\Gamma}}(N+1) \equiv \begin{pmatrix} \mathbf{I}_N & \mathbf{\Sigma} \\ \mathbf{\Sigma}' & 1 \end{pmatrix} \right),$$

and hence $\det(\tilde{\mathbf{\Gamma}}(N+1)) = \det(\mathbf{I}_N)(1 - \phi_1^2)$. So the information of the transformed vector $(\tilde{\mathbf{x}}'_t \ \tilde{x}_{N+1,t})'$ becomes $\tilde{\text{Inf}}_{N+1} = -\log(1 - \phi_1^2)/2$, where

$\phi_1 \in [0, 1]$ is the canonical correlation coefficient. Because $\tilde{\text{Inf}}_N = 0$, Equation (7) becomes $\tilde{\text{Inf}}_{N+1}^R = -\log(1 - \phi_1^2)/c(N + 1)$.

2.3 A test procedure

Replacing $\tilde{\mathbf{T}}(N + 1)$ by a consistent estimate $\hat{\mathbf{T}}(N + 1)$ and applying the same procedure yields $\hat{\text{Inf}}_{N+1} = -\log(1 - \hat{\phi}_1^2)/2$. Under $H_0 : \phi_1 = 0$, the Bartlett test statistic

$$-[T - 1/2(N + 2)] \log(1 - \hat{\phi}_1^2) = [T - 1/2(N + 2)] 2\hat{\text{Inf}}_{N+1}$$

follows asymptotically a χ^2 -distribution with N degrees of freedom, see e.g. Muirhead (1982). Testing the hypothesis $\phi_1 = 0$ is basically testing whether the transformed vector $(\tilde{\mathbf{x}}'_t \tilde{x}_{N+1,t})'$ has maximum entropy, i.e. no correlation at all. If the null hypothesis is rejected, the estimated relative information of the transformed variables equals

$$\hat{\text{Inf}}_{N+1}^R = -\log(1 - \hat{\phi}_1^2)/c(N + 1).$$

Under the null hypothesis the expected value of $2\hat{\text{Inf}}_{N+1}$ is $N/[T - 1/2(N + 2)]$ and hence the expected value of the relative information of the transformed vector $(\tilde{\mathbf{x}}'_t \tilde{x}_{N+1,t})'$ is

$$\text{E} \left\{ \hat{\text{Inf}}_{N+1}^R \right\} \approx 1/c[T - 1/2(N + 2)] \approx 1/[cT - H_{max}]$$

where $H_{max} = c(N + 1)/2$ is the maximum entropy of $(\tilde{\mathbf{x}}'_t \tilde{x}_{N+1,t})'$.

2.4 MSE-prediction

From the foregoing we have $\tilde{\mathbf{x}}_t = \mathbf{A}\mathbf{x}_t$ with $\mathbf{A} = \mathbf{U}'\mathbf{\Lambda}^{-1/2}\mathbf{C}'$. Given a realization $\tilde{x}_{N+1,t} = v^{-1}x_{N+1,t}$ the conditional mean (predictor) of $\tilde{\mathbf{x}}_t$ is $\tilde{\mathbf{x}}_t^P = \mathbf{\Sigma}\tilde{x}_{N+1,t}$ with conditional variance $\text{var}\{\tilde{\mathbf{x}}_t^P\} = \mathbf{I} - \mathbf{\Sigma}\mathbf{\Sigma}' = \text{diag}((1 - \phi_1^2), 1, \dots, 1)$ and information $-\log(1 - \phi_1^2)/2$. Hence if $\phi_1 = 0$ implying $\mathbf{\Sigma} = 0$ the vector $\tilde{\mathbf{x}}_t^P$ has maximum entropy and no information.

The conditional MSE-predictor of \mathbf{x}_t itself is

$$\mathbf{x}_t^P = \mathbf{A}^{-1}\tilde{\mathbf{x}}_t^P = \phi_1\mathbf{C}\mathbf{\Lambda}^{1/2}\mathbf{u}_1\tilde{x}_{N+1,t},$$

where \mathbf{u}_1 is the first column of the orthonormal matrix \mathbf{U} . The conditional variance of \mathbf{x}_t^P is

$$\text{var}\{\mathbf{x}_t^P\} = \mathbf{A}^{-1}(\mathbf{A}^{-1})' - \mathbf{A}^{-1}\mathbf{\Sigma}\mathbf{\Sigma}'(\mathbf{A}^{-1})^{-1} = \mathbf{\Gamma}(N) - \mathbf{A}^{-1}\mathbf{\Sigma}\mathbf{\Sigma}'(\mathbf{A}^{-1})^{-1},$$

from which it can be seen that $\mathbf{\Gamma}(N)$ exceeds $\text{var}\{\mathbf{x}_t^P\}$ by a positive definite matrix if $\phi_1 \neq 0$.

3 Application

In the application below, we use the relative information measure introduced above to order a macroeconomic data set. Plots of the relative information measures against the number of variables indicate which subset is most informative for modelling a variable of interest.

3.1 The data set

In this section we evaluate the performance of the suggested approach on the Stock and Watson (2005) U.S. macroeconomic data set, which consists of monthly observations on 132 macroeconomic time series from 1959M1 up to and including 2003M12. The series cover 14 categories: real output and income; employment and hours; real retail, manufacturing and trade sales; consumption; housing starts and sales; real inventories; orders; stock prices; exchange rates; interest rates and spreads; money and credit quantity aggregates; price indexes; average hourly earnings; and miscellaneous. The series are transformed by taking logarithms and/or differencing when necessary to assure approximate stationarity. In general, first differences of logarithms (growth rates) are used for real quantity variables, first differences are used for nominal interest rates, and second differences of logarithms for price series (changes in inflation). Moreover, the series are adjusted for outliers by replacing the observations of the transformed variables with absolute median deviations larger than 6 times the interquartile range with the median value of the preceding 5 observations. The specific transformations and the list of series are given in Appendix A of Stock and Watson (2005).

3.2 Information in the data set

We order the data set according to the relative information measure with respect to two target variables: the first difference of the log of total industrial production (IP hereafter) and the second difference of the log of the consumer price index (CPI hereafter) using the following procedure: (i) the initial

variable of the ordered data set is the target variable; (ii) the variable that maximizes the respective relative information from the remaining data is added to the ordered data set, and so on. Let $\mathcal{D}(n)$ be the ordered data set that consists of n variables. Then variable x_i , $i = 1, \dots, N - n$ of the remaining data set is chosen for which holds that $i = \operatorname{argmax} I_{N(\mathcal{D}(n+1))}^R$ with $\mathcal{D}(n+1) = \{\mathcal{D}(n), x_i\}$. The full data set consists of $N = 132$ time series variables, with $T = 540$ observations covering the sample 1959M1–2003M12. Since the number of observations T is much larger than the number of series N , all eigenvalues of the covariance matrix of \mathbf{x}_t differ from zero and our relative information measure is computationally stable.

Table 1 presents the orders of the first 50 variables according to the two relative information criteria for both target variables. The table allows the following observations. The first ten series that are included in the subset for IP belong to the group of Industrial Production; the first ten series for CPI are price indices. Second, price indices are generally speaking not informative for IP (the exception is series # 114: NAPM commodity price index), while production series do not appear in the first fifty variables of the ordered data subset for CPI (with one exception series # 19: NAPM production). Finally, variables enter the ordered data sets in clusters. For IP, the relative information measure first selects a group of industrial production variables, followed by employment series, interest rates and spreads, and housing starts and sales. With CPI as target variable, the relative information measure starts with picking price indices, followed by employment, orders, interest rates and spreads, housing starts and sales, and employment.

Table 1: Ranking of series according to relative information

order	IP series #	CPI series #
1	6	115
2	16	124
3	20	123
4	7	119
5	8	125
6	13	127
7	14	122
8	9	117
9	12	128
10	11	121
11	19	39
12	62	37
13	61	38
14	50	34
15	64	33
16	37	40
17	38	41
18	34	43
19	33	50
20	40	61
21	41	19
22	43	62
23	42	64
24	63	42
25	114	63
26	39	114
27	102	102
28	101	101
29	100	100
30	99	99
31	97	98
32	96	97
33	98	96
34	95	95
35	59	59
36	54	54
37	56	56
38	51	51
39	60	60
40	55	55
41	58	58
42	53	53
43	57	57
44	52	52
45	49	49
46	47	47
47	44	44
48	36	36
49	74	74
50	68	68

Notes. See the table in the appendix for the description of the variables.

Figure 1: Relative information of ordered data set

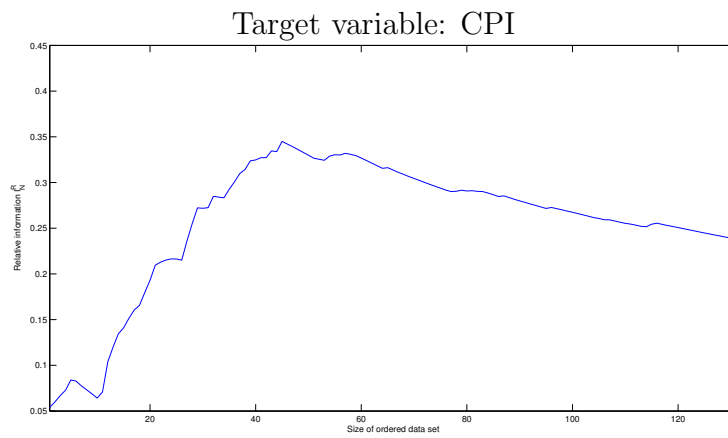
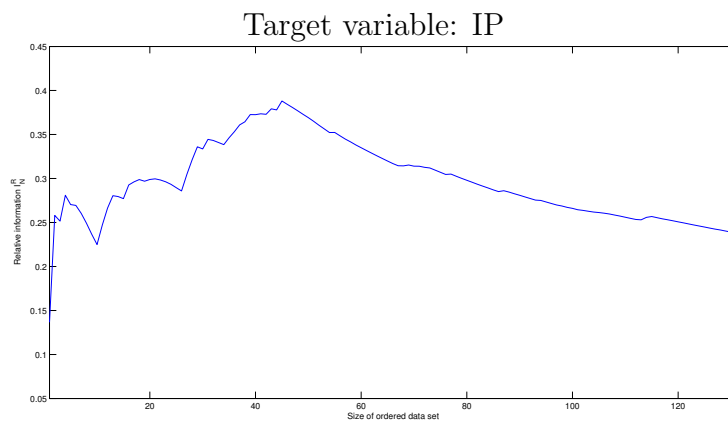
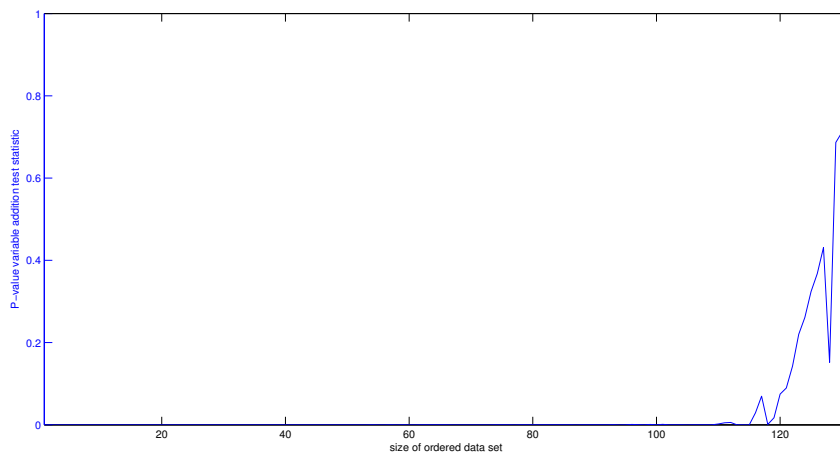


Figure 1 shows the evolution in relative information if we order the data set according to the target variables IP (top panel) and CPI (bottom panel). The figure reveals that sometimes relative information decreases with the addition of a single series, but increases if a batch of variables is added. For both target variables relative information attains a global maximum if we take between 40 and 50 series in line with the findings of Boivin and Ng (2006) and Inklaar et al. (2005).

Figure 2 shows outcomes (p-values) of the test described in Section 2.3 whether an additional variable adds information. The null hypothesis is that an additional variable is not correlated with the variables already included in the set. Hence, low p-values indicate that an additional variable adds information. We note that the outcomes of the test are not sensitive to the initial condition, i.e. the choice of the target variable. The figure suggest that some 120 series are informative. This finding does not contradict our conclusion that relative information, measured by the ratio of information, Inf_N , and maximum entropy cN , is maximized for 40–50 series. More than this number of series add information to the ordered data set, i.e. $\text{Inf}_{N+1} > \text{Inf}_N$ for $40 < N < 120$, but apparently the additional information does not exceed the increase in entropy in these series, $\text{Inf}_{N+1} - \text{Inf}_N < c(N+1) - cN = c$, and therefore $\text{Inf}_{n+1}^R < \text{Inf}_n^R$.

Figure 2: Does an additional variable add information?



3.3 Allowing for pure leads and leads and lags

Our methods can be quite useful to reduce the size of a data set as a first step in the construction of dynamic factor models or leading indexes. To that purpose we calculate the relative information within the data set allowing for both leads and lags and pure leads only. If leads and lags are allowed, individual variables can be selected with a ‘lead’ j between $-k$ and k periods, and a pure lead of $j = 0, \dots, k$ periods. In either case, variable $x_{i,t-j}$, $i = 0, \dots, N - n$ and $j = 0, \dots, k$ of the remaining data set is chosen with a ‘lead’ of j periods for which holds that $\{i, j\} = \arg \max \text{Inf}_{N(D(n+1,j))}^R$ with $D(n+1, j) = \{D(n), x_{i,t-j}\}$. Here we take a three-year horizon and set $k = 36$. This implies that the data is shortened to around 500 observations for the pure lead case, and to around 460 observations if we allow leads and lags.

Table 2 shows the order of the first 50 variables on the basis of relative information I_N^R when we allow for pure leads only, and both leads and lags. The overlap between the two cases for the first 50 series is close to 100% for both target variables. If we allow only pure leads, the first 30 series enter the ordered data set for IP without a lead. Note that this sequence may differ from the static case, because the reduction in the number of observations affects the eigenvalues of the covariance matrix of the data matrix \mathbf{x}_t . Housing starts, sales variables and hours enter the ordered data set with a lead of more than two years. If series may enter with leads and lags, the majority of the first fifty variables is selected with no lead/lag or a small lag. Strikingly, most housing starts and sales variables now get a lag of five months, whereas hours enter the ordered data set with a lag of over one year.

If leads and lags are allowed with CPI as target variable, the first ten variables—all price indices—enter the ordered data set without a lead or a lag. All other variables enter the ordered data set with considerable lags, with the exception of two housing variables which get a twenty months lead. The maximum lag is two and a half years for interest rates and spreads. Allowing pure leads only yields a maximum lead of 31 months for hours worked, and a lead of 14 months for housing starts and sales variables. The majority of the series however enters the ordered data set with no lead or a small lead.

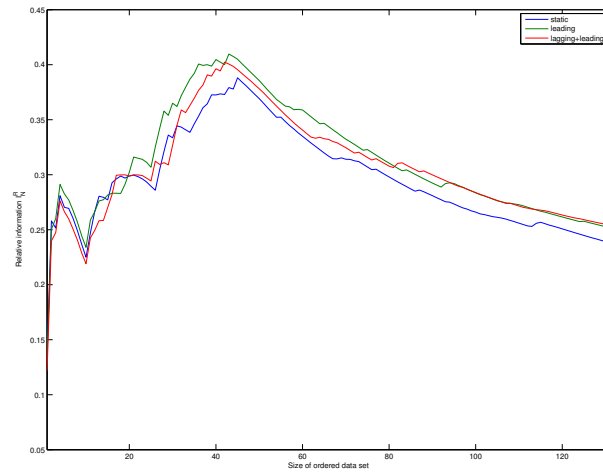
Table 2: Ranking of series according to relative information: pure leads, and leads and lags

order	pure leads		IP leads and lags		pure leads		CPI leads and lags	
	series #	lead	series #	lead(+)/lag(-)	series #	lead	series #	lead(+)/lag(-)
1	6		6		115		115	
2	16	0	16	0	124	0	124	0
3	20	0	20	0	122	0	125	0
4	7	0	7	0	123	0	127	0
5	8	0	8	0	119	0	119	0
6	13	0	13	0	125	0	123	0
7	14	0	14	0	127	0	122	0
8	9	0	9	0	117	0	117	0
9	12	0	12	0	128	0	128	0
10	11	0	11	0	121	0	121	0
11	38	0	38	0	56	14	44	-20
12	37	0	37	0	59	14	40	-20
13	34	0	34	0	54	14	33	-20
14	33	0	33	0	51	14	34	-20
15	40	0	50	-1	60	14	37	-20
16	41	0	61	-1	55	14	38	-20
17	43	0	19	-1	57	14	41	-20
18	42	0	62	-1	52	14	43	-20
19	50	0	63	-2	58	14	50	-21
20	61	0	64	-2	53	14	61	-21
21	19	0	40	0	47	31	19	-21
22	62	0	41	0	49	31	62	-21
23	64	0	43	0	64	5	63	-22
24	63	0	42	0	50	6	64	-22
25	114	0	114	-3	61	6	42	-20
26	102	0	49	-26	19	6	114	-22
27	101	0	47	-26	62	6	102	-30
28	100	0	53	9	63	6	101	-30
29	99	0	58	9	102	0	100	-30
30	97	0	100	0	101	0	99	-30
31	96	0	101	0	100	0	49	-25
32	59	26	102	0	99	0	47	-25
33	56	26	99	0	97	0	52	20
34	54	26	59	-5	96	0	57	20
35	51	26	56	-5	98	0	59	-20
36	60	26	54	-5	38	6	56	-20
37	55	26	51	-5	37	6	54	-20
38	58	26	60	-5	34	6	51	-20
39	53	26	55	-5	33	6	60	-20
40	57	26	57	-5	40	6	55	-20
41	52	26	52	-5	41	6	58	-20
42	98	0	97	0	43	6	53	-20
43	49	27	96	0	42	6	97	-30
44	47	27	98	0	114	3	96	-30
45	95	0	95	23	95	0	98	-30
46	39	0	44	-34	39	6	95	-30
47	44	0	39	0	36	6	36	-20
48	36	0	74	0	44	19	39	-20
49	68	3	36	0	68	8	74	-15
50	74	28	45	0	74	19	68	-20

Notes. See the table in the appendix for the description of the variables.

Figure 3: Comparison of relative information

Target variable: IP



Target variable: CPI

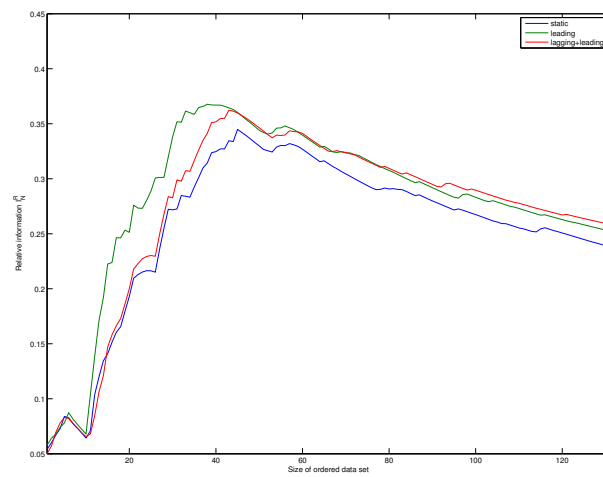


Figure 3 compares relative information in the Stock and Watson data set with respect to IP (top panel) and CPI (bottom panel) in the static case, pure leads only, and both leads and lags. The relative information patterns for the three cases are similar. However relative information is higher if the series are allowed to enter the ordered data set with pure leads only or both leads and lags than in the static case. Hence, there is scope in the data for constructing a dynamic factor model and a leading index for GDP. The maximum is attained at around 40 series for both target variables.

4 Conclusion

This paper fruitfully applied concepts from information theory in the analysis of large data sets. We defined a relative information measure linked to Kullback-Leibler numbers. The application of the measures enabled us to order a data set and to identify a subset of the data that is most informative to modelling a variable of interest.

We illustrated our methods with the Stock and Watson (2005) U.S. macroeconomic data set consisting of 132 times series variables with 540 observations. With around 40–50 series relative information is maximized for industrial production and inflation. Approximately the same number of series enter the ordered data set if leads and lags or pure leads are allowed. We conclude that our method can indeed produce a considerable reduction in the dimension of a data set.

Our relative information measure is based on the eigenvalues of the covariance matrix of the data, which is only defined if the number of observations

T exceeds the number of series N . Future research will deal with the mirror situation of $N > T$.

Appendix A: The Stock and Watson (2005)

U.S. macroeconomic data set

Table A.1 lists the 132 series of the Stock and Watson (2005) U.S. data set, with number, mnemonic, and description of the variable. For details like the transformation applied to the series and sources see Stock and Watson (2005) Appendix A. As is required for factor estimation, the variables are standardized by subtracting their mean and then dividing by their standard deviation. This standardization is necessary to avoid overweighting of large variance series in the factor estimation.

Table A.1: Description of the Stock and Watson data set

#	Short name	Mnemonic	Description
1	PI	A0M052	Personal income (AR, bil. chain 2000 \$)
2	PI less transfers	A0M051	Personal income less transfer payments (AR, bil. chain 2000 \$)
3	Consumption	A0M224_R	Real Consumption (AC) A0m224/gmde
4	M&T sales	A0M057	Manufacturing and trade sales (mil. Chain 1996 \$)
5	Retail sales	A0M059	Sales of retail stores (mil. Chain 2000 \$)
6	IP: total	IPS10	INDUSTRIAL PRODUCTION INDEX - TOTAL INDEX
7	IP: products	IPS11	INDUSTRIAL PRODUCTION INDEX - PRODUCTS, TOTAL
8	IP: final prod	IPS299	INDUSTRIAL PRODUCTION INDEX - FINAL PRODUCTS
9	IP: cons gds	IPS12	INDUSTRIAL PRODUCTION INDEX - CONSUMER GOODS
10	IP: cons dble	IPS13	INDUSTRIAL PRODUCTION INDEX - DURABLE CONSUMER GOODS
11	iIP:cons nondble	IPS18	INDUSTRIAL PRODUCTION INDEX - NONDURABLE CONSUMER GOODS
12	IP:bus eqpt	IPS25	INDUSTRIAL PRODUCTION INDEX - BUSINESS EQUIPMENT
13	IP: matls	IPS32	INDUSTRIAL PRODUCTION INDEX - MATERIALS
14	IP: dble mats	IPS34	INDUSTRIAL PRODUCTION INDEX - DURABLE GOODS MATERIALS
15	IP:nondble mats	IPS38	INDUSTRIAL PRODUCTION INDEX - NONDURABLE GOODS MATERIALS
16	IP: mfg	IPS43	INDUSTRIAL PRODUCTION INDEX - MANUFACTURING (SIC)
17	IP: res util	IPS307	INDUSTRIAL PRODUCTION INDEX - RESIDENTIAL UTILITIES
18	IP: fuels	IPS306	INDUSTRIAL PRODUCTION INDEX - FUELS
19	NAPM prodrn	PMP	NAPM PRODUCTION INDEX (PERCENT)
20	Cap util	A0M082	Capacity Utilization (Mfg)
21	Help wanted indx	LHEL	INDEX OF HELP-WANTED ADVERTISING IN NEWSPAPERS (1967=100;SA)
22	Help wanted/emp	LHELX	EMPLOYMENT: RATIO; HELP-WANTED ADS:NO. UNEMPLOYED CLF
23	Emp CPS total	LHEM	CIVILIAN LABOR FORCE: EMPLOYED, TOTAL (THOUS.,SA)
24	Emp CPS nonag	LHNAG	CIVILIAN LABOR FORCE: EMPLOYED, NONAGRIC.INDUSTRIES (THOUS.,SA)
25	U: all	LHUR	UNEMPLOYMENT RATE: ALL WORKERS, 16 YEARS & OVER (%SA)
26	U: mean duration	LHU680	UNEMPLOY.BY DURATION: AVERAGE(MEAN)DURATION IN WEEKS (SA)
27	U 5 wks	LHU5	UNEMPLOY.BY DURATION: PERSONS UNEMPL.LESS THAN 5 WKS (THOUS.,SA)
28	U 5-14 wks	LHU14	UNEMPLOY.BY DURATION: PERSONS UNEMPL.5 TO 14 WKS (THOUS.,SA)
29	U 15+ wks	LHU15	UNEMPLOY.BY DURATION: PERSONS UNEMPL.15 WKS + (THOUS.,SA)
30	U 15-26 wks	LHU26	UNEMPLOY.BY DURATION: PERSONS UNEMPL.15 TO 26 WKS (THOUS.,SA)
31	U 27+ wks	LHU27	UNEMPLOY.BY DURATION: PERSONS UNEMPL.27 WKS + (THOUS.,SA)
32	UI claims	A0M005	Average weekly initial claims, unemploy. insurance (thous.)
33	Emp: total	CES002	EMPLOYEES ON NONFARM PAYROLLS - TOTAL PRIVATE
34	Emp: gds prod	CES003	EMPLOYEES ON NONFARM PAYROLLS - GOODS-PRODUCING
35	Emp: mining	CES006	EMPLOYEES ON NONFARM PAYROLLS - MINING
36	Emp: const	CES011	EMPLOYEES ON NONFARM PAYROLLS - CONSTRUCTION
37	Emp: mfg	CES015	EMPLOYEES ON NONFARM PAYROLLS - MANUFACTURING
38	Emp: dble gds	CES017	EMPLOYEES ON NONFARM PAYROLLS - DURABLE GOODS
39	Emp: nondbles	CES033	EMPLOYEES ON NONFARM PAYROLLS - NONDURABLE GOODS
40	Emp: services	CES046	EMPLOYEES ON NONFARM PAYROLLS - SERVICE-PROVIDING
41	Emp: TTU	CES048	EMPLOYEES ON NONFARM PAYROLLS - TRADE, TRANSPORTATION, AND UTILITIES
42	Emp: wholesale	CES049	EMPLOYEES ON NONFARM PAYROLLS - WHOLESALE TRADE
43	Emp: retail	CES053	EMPLOYEES ON NONFARM PAYROLLS - RETAIL TRADE
44	Emp: FIRE	CES088	EMPLOYEES ON NONFARM PAYROLLS - FINANCIAL ACTIVITIES
45	Emp: Govt	CES140	EMPLOYEES ON NONFARM PAYROLLS - GOVERNMENT
46	Emp-hrs nonag	A0M048	Employee hours in nonag. establishments (AR, bil. hours)
47	Avg hrs	CES151	AVERAGE WEEKLY HOURS OF PRODUCTION OR NONSUPERVISORY WORKERS ON PRIVATE NONFARM PAYROLLS - GOODS-PRODUCING
48	Overtime: mfg	CES155	AVERAGE WEEKLY HOURS OF PRODUCTION OR NONSUPERVISORY WORKERS ON PRIVATE NONFARM PAYROLLS - MFG OVERTIME HOURS
49	Avg hrs: mfg	A0M001	Average weekly hours, mfg. (hours)
50	NAPM empl	PMEMP	NAPM EMPLOYMENT INDEX (PERCENT)
51	HStarts: Total	HSFR	HOUSING STARTS:NONFARM(1947-58);TOTAL FARM&NONFARM(1959-)(THOUS.,SAAR)
52	HStarts: NE	HSNE	HOUSING STARTS:NORTHEAST (THOUS.U.)S.A.
53	HStarts: MW	HSMW	HOUSING STARTS:MIDWEST(THOUS.U.)S.A.
54	HStarts: South	HSSOU	HOUSING STARTS:SOUTH (THOUS.U.)S.A.
55	HStarts: West	HSWST	HOUSING STARTS:WEST (THOUS.U.)S.A.
56	BP: total	HSBR	HOUSING AUTHORIZED: TOTAL NEW PRIV HOUSING UNITS (THOUS.,SAAR)
57	BP: NE	HSBNE	HOUSES AUTHORIZED BY BUILD. PERMITS:NORTHEAST(THOU.U.)S.A
58	BP: MW	HSBMW	HOUSES AUTHORIZED BY BUILD. PERMITS:MIDWEST(THOU.U.)S.A.
59	BP: South	HSBSOU	HOUSES AUTHORIZED BY BUILD. PERMITS:SOUTH(THOU.U.)S.A.
60	BP: West	HSBWST	HOUSES AUTHORIZED BY BUILD. PERMITS:WEST(THOU.U.)S.A.
61	PMI	PMI	PURCHASING MANAGERS' INDEX (SA)
62	NAPM new ordrs	PMNO	NAPM NEW ORDERS INDEX (PERCENT)
63	NAPM vendor del	PMDEL	NAPM VENDOR DELIVERIES INDEX (PERCENT)
64	NAPM Invent	PMNV	NAPM INVENTORIES INDEX (PERCENT)

#	Short name	Mnemonic	Description
65	Orders: cons gds	A0M008	Mfrs' new orders, consumer goods and materials (bil. chain 1982 \$)
66	Orders: dble gds	A0M007	Mfrs' new orders, durable goods industries (bil. chain 2000 \$)
67	Orders: cap gds	A0M027	Mfrs' new orders, nondefense capital goods (mil. chain 1982 \$)
68	Unf orders: dble	A1M092	Mfrs' unfilled orders, durable goods indus. (bil. chain 2000 \$)
69	M&T invent	A0M070	Manufacturing and trade inventories (bil. chain 2000 \$)
70	M&T invent/sales	A0M077	Ratio, mfg. and trade inventories to sales (based on chain 2000 \$)
71	M1	FM1	MONEY STOCK: M1(CURR.TRAV.CKS,DEM DEP,OTHER CK'ABLE DEP)(BIL\$,SA)
72	M2	FM2	MONEY STOCK:M2(M1+O'NITE RPS,EURO\$,G/P&B/D MMMFS&SAV&SM TIME DEP)(BIL\$,SA)
73	M3	FM3	MONEY STOCK: M3(M2+LG TIME DEP,TERM RP'S&INST ONLY MMMFS)(BIL\$,SA)
74	M2 (real)	FM2DQ	MONEY SUPPLY - M2 IN 1996 DOLLARS (BCI)
75	MB	FMFBA	MONETARY BASE, ADJ FOR RESERVE REQUIREMENT CHANGES(MIL\$,SA)
76	Reserves tot	FMRRRA	DEPOSITORY INST RESERVES:TOTAL,ADJ FOR RESERVE REQ CHGS(MIL\$,SA)
77	Reserves nonbor	FMRNBA	DEPOSITORY INST RESERVES:NONBORROWED,ADJ RES REQ CHGS(MIL\$,SA)
78	C&I loans	FCLNQ	COMMERCIAL & INDUSTRIAL LOANS OUSTANDING IN 1996 DOLLARS (BCI)
79	C&I loans	FCLBMC	WKLY RP LG COM'L BANKS:NET CHANGE COM'L & INDUS LOANS(BIL\$,SAAR)
80	Cons credit	CCINRV	CONSUMER CREDIT OUTSTANDING - NONREVOLVING(G19)
81	Inst cred/PI	A0M095	Ratio, consumer installment credit to personal income (pct.)
82	S&P 500	FSPCOM	S&P'S COMMON STOCK PRICE INDEX: COMPOSITE (1941-43=10)
83	S&P: indust	FSPIN	S&P'S COMMON STOCK PRICE INDEX: INDUSTRIALS (1941-43=10)
84	S&P div yield	FSDXP	S&P'S COMPOSITE COMMON STOCK: DIVIDEND YIELD (% PER ANNUM)
85	S&P PE ratio	FSPXE	S&P'S COMPOSITE COMMON STOCK: PRICE-EARNINGS RATIO (%NSA)
86	FedFunds	FYFF	INTEREST RATE: FEDERAL FUNDS (EFFECTIVE) (% PER ANNUM,NSA)
87	Commpaper	CP90	Commercial Paper Rate (AC)
88	3 mo T-bill	FYGM3	INTEREST RATE: U.S.TREASURY BILLS,SEC MKT,3-MO.(% PER ANN,NSA)
89	6 mo T-bill	FYGM6	INTEREST RATE: U.S.TREASURY BILLS,SEC MKT,6-MO.(% PER ANN,NSA)
90	1 yr T-bond	FYGT1	INTEREST RATE: U.S.TREASURY CONST MATURITIES,1-YR.(% PER ANN,NSA)
91	5 yr T-bond	FYGT5	INTEREST RATE: U.S.TREASURY CONST MATURITIES,5-YR.(% PER ANN,NSA)
92	10 yr T-bond	FYGT10	INTEREST RATE: U.S.TREASURY CONST MATURITIES,10-YR.(% PER ANN,NSA)
93	Aaabond	FYAAAC	BOND YIELD: MOODY'S AAA CORPORATE (% PER ANNUM)
94	Baa bond	FYBAAC	BOND YIELD: MOODY'S BAA CORPORATE (% PER ANNUM)
95	CP-FF spread	SCP90	cp90-fyff
96	3 mo-FF spread	SFYGM3	fygm3-fyff
97	6 mo-FF spread	SFYGM6	fygm6-fyff
98	1 yr-FF spread	SFYGT1	fygt1-fyff
99	5 yr-FFspread	SFYGT5	fygt5-fyff
100	10yr-FF spread	SFYGT10	fygt10-fyff
101	Aaa-FF spread	SFYAAAC	fyaaac-fyff
102	Baa-FF spread	SFYBAAC	fybaac-fyff
103	Ex rate: avg	EXRUS	UNITED STATES;EFFECTIVE EXCHANGE RATE(MERM)(INDEX NO.)
104	Ex rate: Switz	EXRSW	FOREIGN EXCHANGE RATE: SWITZERLAND (SWISS FRANC PER U.S.\$)
105	Ex rate: Japan	EXRJAN	FOREIGN EXCHANGE RATE: JAPAN (YEN PER U.S.\$)
106	Ex rate: UK	EXRUK	FOREIGN EXCHANGE RATE: UNITED KINGDOM (CENTS PER POUND)
107	EX rate: Canada	EXRCAN	FOREIGN EXCHANGE RATE: CANADA (CANADIAN \$ PER U.S.\$)
108	PPI: fin gds	PWFSA	PRODUCER PRICE INDEX: FINISHED GOODS (82=100,SA)
109	PPI: cons gds	PWFCSA	PRODUCER PRICE INDEX:FINISHED CONSUMER GOODS (82=100,SA)
110	PPI: int matls	PWMSA	PRODUCER PRICE INDEX:INTERMED MAT,SUPPLIES & COMPONENTS(82=100,SA)
111	PPI: crude matls	PWCMSA	PRODUCER PRICE INDEX:CRUDE MATERIALS (82=100,SA)
112	Commod: spot price	PSCCOM	SPOT MARKET PRICE INDEX:BLS & CRB: ALL COMMODITIES(1967=100)
113	Sens matls price	PSM99Q	INDEX OF SENSITIVE MATERIALS PRICES (1990=100)(BCI-99A)
114	NAPM com price	PMCP	NAPM COMMODITY PRICES INDEX (PERCENT)
115	CPI-U: all	PUNEW	CPI-U: ALL ITEMS (82-84=100,SA)
116	CPI-U: apparel	PU83	CPI-U: APPAREL & UPKEEP (82-84=100,SA)
117	CPI-U: transp	PU84	CPI-U: TRANSPORTATION (82-84=100,SA)
118	CPI-U: medical	PU85	CPI-U: MEDICAL CARE (82-84=100,SA)
119	CPI-U: comm.	PUC	CPI-U: COMMODITIES (82-84=100,SA)
120	CPI-U: dbles	PUCD	CPI-U: DURABLES (82-84=100,SA)
121	CPI-U: services	PUS	CPI-U: SERVICES (82-84=100,SA)
122	CPI-U: ex food	PUXF	CPI-U: ALL ITEMS LESS FOOD (82-84=100,SA)
123	CPI-U: ex shelter	PUXHS	CPI-U: ALL ITEMS LESS SHELTER (82-84=100,SA)
124	CPI-U: ex med	PUXM	CPI-U: ALL ITEMS LESS MEDICAL CARE (82-84=100,SA)
125	PCE defl	GMDC	PCE,IMPL PR DEFL:PCE (1987=100)
126	PCE defl: dlbes	GMDCD	PCE,IMPL PR DEFL:PCE; DURABLES (1987=100)
127	PCE defl: nondble	GMDCN	PCE,IMPL PR DEFL:PCE; NONDURABLES (1996=100)
128	PCE defl: services	GMDCS	PCE,IMPL PR DEFL:PCE; SERVICES (1987=100)
129	AHE: goods	CES275	AVERAGE HOURLY EARNINGS OF PRODUCTION OR NONSUPERVISORY WORKERS ON PRIVATE NONFARM PAYROLLS - GOODS PRODUCING
130	AHE: const	CES277	AVERAGE HOURLY EARNINGS OF PRODUCTION OR NONSUPERVISORY WORKERS ON PRIVATE NONFARM PAYROLLS - CONSTRUCTION
131	AHE: mfg	CES278	AVERAGE HOURLY EARNINGS OF PRODUCTION OR NONSUPERVISORY WORKERS ON PRIVATE NONFARM PAYROLLS - MANUFACTURING
132	Consumer expect	HHSNTN	U. OF MICH. INDEX OF CONSUMER EXPECTATIONS(BCD-83)

Acknowledgements

Views expressed are those of the individual authors and do not necessarily reflect official positions of the Riksbank. The paper has benefited from comments received following presentations of previous versions at the Far Eastern Meeting of the Econometric Society, Beijing, China, July 2006, the 13th International Conference on Computing in Economics and Finance, Montréal, Canada, June 2007, the Research Forum: New Developments in Dynamic Factor Modelling, Centre for Central Banking Studies of the Bank of England, London, October 2007, the Far Eastern Meeting of the Econometric Society, Singapore, July 2008, the Conference on Factor Structures for Panel and Multivariate Time Series Data, Maastricht, September 2008, and seminars at De Nederlandsche Bank and the Sveriges Riksbank.

References

- Bai, J. (2003), “Inferential theory for factor models of large dimensions”, *Econometrica*, **71**, 135–171.
- Bai, J. and S. Ng (2002), “Determining the number of factors in approximate factor models”, *Econometrica*, **70**, 191–221.
- Bai, J. and S. Ng (2008a), “Large dimensional factor analysis”, *Foundations and Trends in Econometrics*, **3**, 89–163.
- Bai, J. and S. Ng (2008b), “Forecasting economic time series using targeted predictors”, *Journal of Econometrics*, **146**, 304–317.
- Boivin, J. and S. Ng (2006), “Are more data always better for factor analysis?”, *Journal of Econometrics*, **132**, 169–194.
- Burnham, K.P. and D.R. Anderson (2002), *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd edition, Springer, New York.
- 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., J.P.A.M. Jacobs, and W.E. Romp (2004), “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 (2008), “Determining the number of factors and lag order in dynamic factor models: A minimum entropy approach”, *Econometric Reviews*, **27**, 385–397.
- Muirhead, R.J. (1982), *Aspects of Multivariate Statistical Theory*, Wiley, New York.
- Otter, P.W. and J.P.A.M. Jacobs (2006), “On information in static and dynamic factor models”, Working Paper #2006/5, CCSO [presented at the Far Eastern Meeting of the Econometric Society, July 2006, Beijing, China].
- Stock, J.H. and M.W. Watson (2005), “Implications of dynamic factor models for VAR analysis”, Working paper 11467, National Bureau of Economic Research.
- Young, T.Y. and T.W. Calvert (1974), *Classification, Estimation and Pattern Recognition*, Elsevier, New York; London.