



# Information, data dimension, and factor structure

Jan Jacobs, Pieter W. Otter and Ard den Reijer

University of Groningen

De Nederlandsche Bank



Research Forum: New developments in dynamic factor  
modelling, 10 October 2007, London



## Abstract



This paper employs concepts from information theory as described in Otter and Jacobs (2006) to choosing the dimension of a data set. We calculate relative measures of information in the data in terms of eigenvalues and derive criteria to determine the 'optimal' size of the data set, in particular whether an extra variable adds information. The methods can be used as a first step in the construction of a dynamic factor model or a leading index, as illustrated with a macroeconomic data set on The Netherlands.



## Structure



1. Introduction

2. Information in data

- Kullback-Leibler
- relative information  $I_N^R$
- relative eigenvalues information  $I_{\bar{\lambda}}^R$

3. Application

4. Conclusion



## Motivation



**Factor models:** variations in a large number of economic variables can be modelled by a small number of variables, or movements in a large number of series are driven by a limited number of common 'factors'

### Issues

- number of factors  
Correct specification of the number of factors is central to both the theoretical and the empirical validity of factor models  
Jacobs and Otter (*Econometric Reviews* forthcoming)



## Motivation



**Factor models:** variations in a large number of economic variables can be modelled by a small number of variables, or movements in a large number of series are driven by a limited number of common ‘factors’

### Issues

- number of factors  
Correct specification of the number of factors is central to both the theoretical and the empirical validity of factor models  
Jacobs and Otter (*Econometric Reviews* forthcoming)
- size of data set, or number of variables  
Need not be very large to get precise factor estimates; some 40 series are sufficient (Bai and Ng, 2002; Boivin and Ng, 2006; Inklaar, Jacobs, and Romp, 2005)



## Information



Let  $\mathbf{x}_t$  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  $E(\mathbf{x}_t \mathbf{x}_t') = \mathbf{\Gamma}_0$ , i.e.  $\mathbf{x}_t \sim \mathbb{N}(\mathbf{0}, \mathbf{\Gamma}_0)$ , where  $\text{diag}(\mathbf{\Gamma}_0) = (1, 1, \dots, 1)$ ,  $\text{tr}(\mathbf{\Gamma}_0) = N$  and  $\mathbf{\Gamma}_i = E(\mathbf{x}_t \mathbf{x}_{t-i}') are the autocovariances of  $\mathbf{x}_t$ .$



## Information



Let  $\mathbf{x}_t$  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  $E(\mathbf{x}_t \mathbf{x}_t') = \mathbf{\Gamma}_0$ , i.e.  $\mathbf{x}_t \sim \mathbb{N}(\mathbf{0}, \mathbf{\Gamma}_0)$ , where  $\text{diag}(\mathbf{\Gamma}_0) = (1, 1, \dots, 1)$ ,  $\text{tr}(\mathbf{\Gamma}_0) = N$  and  $\mathbf{\Gamma}_i = E(\mathbf{x}_t \mathbf{x}_{t-i}') are the autocovariances of  $\mathbf{x}_t$ .$

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

$$2H_x = cN + \log \det(\mathbf{\Gamma}_0),$$

where  $c \equiv \log(2\pi) + 1 \approx 2.84$ , with  $2H_{x,max} = cN$  in case  $\mathbf{\Gamma}_0 = \mathbf{I}_N$ .

The information or negentropy is defined as

$$I_x \equiv 2H_{x,max} - 2H_x = -\log \det(\mathbf{\Gamma}_0) \geq 0, \quad (1)$$

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



## Kullback-Leibler

Let  $f_1(\tilde{\mathbf{x}}) : \tilde{\mathbf{x}} \sim \mathcal{N}_N(\mathbf{0}, \mathbf{\Gamma}_0 = \mathbf{C}\mathbf{\Lambda}\mathbf{C}')$  be the density function of  $\mathbf{x}$ , then  $f_1(\mathbf{x}) : \mathbf{x} \sim \mathcal{N}_N(\mathbf{0}, \mathbf{\Lambda})$  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

$$G_1 = E_{f_1} \left( \log \left( \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \right) \right) \text{ and } G_2 = E_{f_2} \left( \log \left( \frac{f_2(\mathbf{x})}{f_1(\mathbf{x})} \right) \right) \quad (2)$$

and  $G(\mathbf{x}) = G_1(\mathbf{x}) + G_2(\mathbf{x})$  is the measure of information for discriminating between the two density functions with  $G(\mathbf{x}) = 0$  in case  $f_1(\mathbf{x}) = f_2(\mathbf{x})$  and  $G = \infty$  in case of perfect discrimination.





For  $\text{tr}(\mathbf{I}_0) = \text{tr}(\mathbf{A}) = N$  we have  $G_1(\mathbf{x}) = -\log\det(\mathbf{A})$  and  
De Nederlandsche Bank  
 $G_2(\mathbf{x}) = \log\det(\mathbf{A}) + \frac{1}{2} (\text{tr}(\mathbf{A}^{-1}) - N)$ . Therefore

$$2G(\mathbf{x}) = \text{tr}(\mathbf{A}^{-1}) - N = \text{tr}(\mathbf{A}^{-1}) - \text{tr}(\mathbf{A}) = \sum_{j=1}^N \frac{(1 - \lambda_j^2)}{\lambda_j}, \quad (3)$$

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



De Nederlandsche Bank

For  $\text{tr}(\mathbf{I}_0) = \text{tr}(\mathbf{A}) = N$  we have  $G_1(\mathbf{x}) = -\log\det(\mathbf{A})$  and  
 $G_2(\mathbf{x}) = \log\det(\mathbf{A}) + \frac{1}{2} (\text{tr}(\mathbf{A}^{-1}) - N)$ . Therefore

$$2G(\mathbf{x}) = \text{tr}(\mathbf{A}^{-1}) - N = \text{tr}(\mathbf{A}^{-1}) - \text{tr}(\mathbf{A}) = \sum_{j=1}^N \frac{(1 - \lambda_j^2)}{\lambda_j}, \quad (3)$$

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

The distribution of the eigenvalues can be measured by their entropy. Because  $\text{tr}(\mathbf{A}) = N$  we have  $\bar{\lambda}_j = \lambda_j/N$  with  $0 \leq \bar{\lambda}_j \leq 1$  and

$$H_{\bar{\lambda}} = - \sum_j \bar{\lambda}_j \log \bar{\lambda}_j \quad (4)$$

with  $H_{\bar{\lambda}}^{\max} = \log(N)$  for  $\bar{\lambda}_j = 1/N$  for all  $j$ .



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



## Do variables add information?

Recall that for  $\mathbf{x}_t(N) \in \mathbb{R}^N$  with autocovariance  $E(\mathbf{x}_t(N)\mathbf{x}_t'(N)) = \mathbf{\Gamma}_0(N)$  the entropy is defined as  $2H_{\mathbf{x}_t(N)} = cN + \log \det(\mathbf{\Gamma}_0(N))$  and the information as  $2I_{\mathbf{x}_t(N)} = 2H_{\mathbf{x},max} - 2H_{\mathbf{x}} = -\log \det(\mathbf{\Gamma}_0(N)) \equiv I_N$ .

Define the *relative information* (per component of  $\mathbf{x}_t(N)$ ) as:

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

If  $H_{\mathbf{x}(N)}$  is equal to  $H_{max}$  then  $2I_N^R = 0$ ; if  $H_{\mathbf{x}(N)} = 0$  then  $2I_N^R = 1$ .

The information contained in the eigenvalues is  $I_{\bar{\lambda}} = \log(N) - H_{\bar{\lambda}}$  or the *relative eigenvalues information*

$$I_{\bar{\lambda}}^R = 1 - \frac{H_{\bar{\lambda}}}{\log(N)}, \quad (5)$$

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



## Do variables add information?

So, an additional variable  $x_{N+1,t}$  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.



## Do variables add information?

So, an additional variable  $x_{N+1,t}$  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.

The rank of the autocovariance matrix  $\mathbf{I}_0(N)$  cannot exceed  $T$ , and so both criteria—which are determined by the eigenvalues of the covariance matrix—become infeasible in case  $N > T$ .



## Application



Information in data set (Den Reijer 2005), which consists of the data underlying MORKMON supplemented with variables possessing leading indicator characteristics:  $N = 124$ ;  $T = 92$

Six groups of variables:

1. GDP, gross value added and productivity
2. Industrial Production and capacity utilization
3. Prices
4. Financial
5. External sector
6. Surveys



## Procedure



(i) initial variable is target variable (GDP growth and change of CPI-inflation);

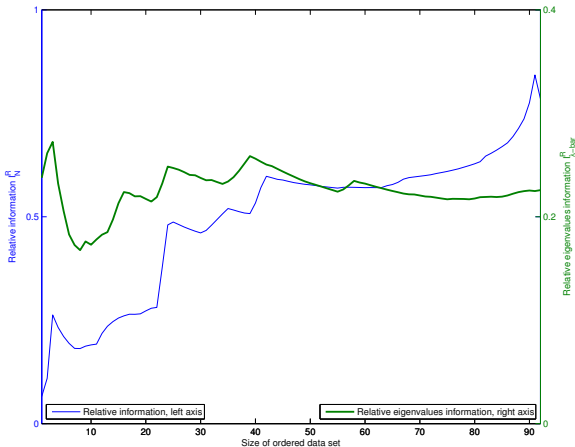
(ii) from the remaining data the variable is added with highest information  $(I_N^R, I_{\bar{\lambda}}^R)$ .

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  $i = \operatorname{argmax}_{\bar{\lambda}(\mathcal{D}(n+1))} I_{\bar{\lambda}(\mathcal{D}(n+1))}^R$ , respectively,  $i = \operatorname{argmax}_{\mathcal{D}(n+1)} I_{\mathcal{D}(n+1)}^R$  with  $\mathcal{D}(n + 1) = \{\mathcal{D}(n), X_i\}$ .



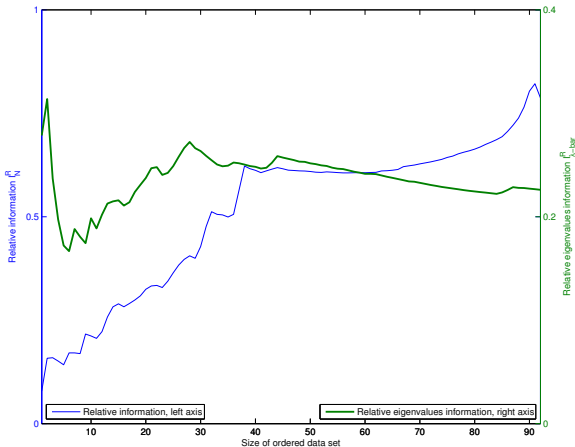


## Relative information with respect to GDP





## Relative information with respect to CPI





## Ranking of series



order	GDP				CPI			
	eigenvalues	info	series #	group	eigenvalues	info	series #	group
1	33	2	33	2	59	3	59	3
2	34	2	34	2	46	3	46	3
3	32	2	32	2	58	3	58	3
4	4	1	4	1	44	3	44	3
5	13	1	13	1	56	3	54	3
6	86	5	2	1	48	3	48	3
7	90	5	88	5	54	3	56	3
8	42	2	7	1	47	3	43	3
9	41	2	24	2	92	5	92	5
10	7	1	40	2	43	3	47	3
11	24	2	122	6	51	3	49	3
12	40	2	108	6	53	3	51	3
13	112	6	112	6	49	3	50	3
14	122	6	123	6	50	3	53	3
15	108	6	21	1	52	3	52	3
16	21	1	124	6	80	4	78	4
17	123	6	110	6	84	4	83	4
18	107	6	27	2	79	4	81	4
19	116	6	29	2	81	4	84	4
20	88	5	30	2	78	4	79	4



## Results




The table presents the index numbers of the variables of the data set and the corresponding groups to which they belong.

- The first three series that are included in the data set for GDP belong to the group of Industrial Production. The first seven series that are included in the data set for CPI belong to the group Prices. 'Multi-group'-optima emerge around 40 series for GDP and around 30 series for CPI.
- The overlap between the two ordered data sets is about 80% for both target variables.
- Price variables (group 3) are not informative for GDP, and real variables (group 1) only enter the ordered data set for CPI after the optimum number is reached. Moreover, variables from all five remaining categories are chosen.



## Allowing for leads and lags: Procedure

 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 leads and lags and pure leads.

- (i) initial variable is target variable (GDP growth and change of CPI-inflation);
- (ii) from the remaining data the variable is added with highest information ( $I_N^R, I_\lambda^R$ ). 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  $j = 1 \dots, k$  periods.



## Allowing for leads and lags: Procedure



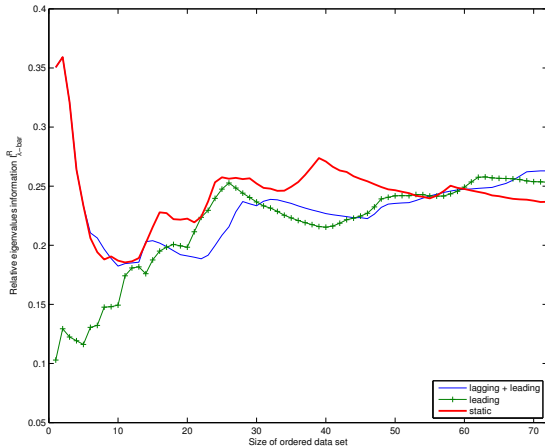
In either case, variable  $X_{i,t-j}$ ,  $i = 1, \dots, N - n$  and  $j = 0, \dots, k$  of the remaining data set is chosen with a 'lead' of  $j$  periods for which holds that  $\{i, j\} = \arg \max_{\lambda(D(n+1,j))} I_{\lambda(D(n+1,j))}^R$  or  $\{i, j\} = \arg \max_{N(D(n+1,j))} I_{N(D(n+1,j))}^R$  with

$$D(n+1, j) = \{D(n), X_{i,t-j}\}.$$

In our calculations we set  $k = 10$ , which shortens the effective sample size to  $(T - 2k =)72$  periods when allowing for leads and lags, and  $(T - k =)82$  in case of pure leads.

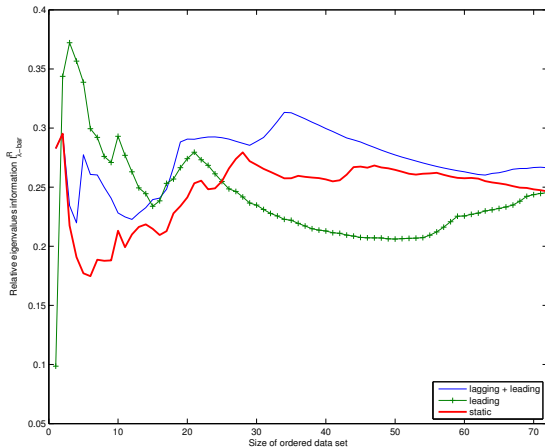


## Allowing for leads and lags: GDP





## Allowing for leads and lags: CPI







## Ranking of series



order	GDP				CPI							
	series #	group	leads and lags lead(+)/lag(-)	pure leads series # group lead	series #	group	leads and lags lead(+)/lag(-)	pure leads series # group lead				
1	1	1	0	1	1	0	57	3	0	57	3	0
2	33	2	0	12	1	3	59	3	0	51	3	9
3	34	2	0	3	1	3	46	3	0	53	3	9
4	32	2	0	73	4	7	44	3	0	49	3	9
5	4	1	0	114	6	9	92	5	-1	50	3	9
6	13	1	0	102	5	5	43	3	-1	52	3	9
7	113	6	-8	104	5	5	48	3	-1	54	3	7
8	109	6	-8	54	3	10	54	3	-1	48	3	7
9	105	5	4	48	3	10	56	3	-1	44	3	8
10	104	5	3	56	3	10	47	3	-1	43	3	7
11	48	3	3	92	5	8	110	6	-2	92	5	7
12	54	3	3	43	3	8	124	6	-2	47	3	7
13	47	3	3	44	3	9	108	6	-2	102	5	2
14	92	5	3	101	5	3	122	6	-2	56	3	8
15	43	3	3	51	3	10	112	6	-2	104	5	2
16	44	3	4	53	3	10	21	1	-2	80	4	7
17	56	3	3	49	3	10	36	2	-3	84	4	7
18	102	5	3	50	3	10	23	2	-3	79	4	7
19	60	3	-9	52	3	10	117	6	-3	81	4	7
20	57	3	4	17	1	6	35	2	-3	83	4	7



## Results



With GDP as the first variable, leading indicators are typically preferred over lagging indicators. As a result, the overlap with the optimal data set of the static case is around 65%. In case of CPI, however, variables are more often selected with a lag which implies that CPI itself is a leading indicator. If variables are restricted to only enter the ordered data set with a lead, the overlap of the optimal data set with the corresponding static data set is 60%.



## Results



With GDP as the first variable, leading indicators are typically preferred over lagging indicators. As a result, the overlap with the optimal data set of the static case is around 65%. In case of CPI, however, variables are more often selected with a lag which implies that CPI itself is a leading indicator. If variables are restricted to only enter the ordered data set with a lead, the overlap of the optimal data set with the corresponding static data set is 60%.

- Relative information is highest if we allow for leads and lags, as it nests both the static and pure leading case. However, the lines in the figures cross, since sets of variables differ. When more and more variables are selected and the data sets converge, relative information allowing for leads and lags and pure leads exceeds the information of the static case.




## Results



In case of GDP, only focussing on leading indicators suggests an optimum of around 60 leading indicators. In case of CPI, the picture is less clear. Since inflation leads most variables in the data set, the pure leading case has more relative information for the first 15 variables in the ordered data set, which almost all belong to the prices category (i.e. energy and commodity prices). The optimal size of the data set becomes around 35 variables, if we allow leading and lagging variables.




## Conclusion

 Concepts from information theory can be fruitfully applied in the analysis of large data sets.

Application of the proposed information measures on the Dutch data set: around 40 series provide maximum amount of information with respect to GDP and CPI.



## Conclusion

 Concepts from information theory can be fruitfully applied in the analysis of large data sets.

Application of the proposed information measures on the Dutch data set: around 40 series provide maximum amount of information with respect to GDP and CPI.

Twice as much leading series, around 60, is the optimal size when GDP is the first variable and allowing for leads and lags, around 35 variables are optimal for CPI. We conclude that our methods can indeed produce a considerable reduction in the dimension of a data set