Journal of Multivariate Analysis 106 (2012) 80–91

Contents lists available at SciVerse ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Information, data dimension and factor structure

Jan P.A.M. Jacobs [a,*], Pieter W. Otter [a], Ard H.J. den Reijer [b]

[a] University of Groningen, The Netherlands

[b] Sveriges Riksbank, Sweden

ARTICLE INFO

Article history:
Received 21 September 2010
Available online 22 November 2011

JEL classification:
C32
C52
C82

AMS subject classifications:
62-07
62H25
62P20
94A17

Keywords:
Kullback–Leibler numbers
Information
Factor structure
Data set dimension
Dynamic factor models
Leading index

ABSTRACT

This paper employs concepts from information theory for 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 a data set that is most informative. The method can be used as a first step in the construction of a dynamic factor model or a leading index, as illustrated with a Monte Carlo study and with the US macroeconomic data set of Stock and Watson [20].

1. Introduction

With the proliferation of huge data sets a natural question to ask is how much information is there 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 [2].

Although convergence of factor estimates requires large cross-sections and large time dimensions, see e.g. Forni and Lippi [9] and Bai [1], the data set need not be very large to obtain reasonably precise factor estimates. Boivin and Ng [6] and Inklaar et al. [13] 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 [3] 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.[1]

* Correspondence to: Faculty of Economics and Business, University of Groningen, P.O. Box 800, 9700 AV Groningen, The Netherlands.
E-mail address: j.p.a.m.jacobs@rug.nl (J.P.A.M. Jacobs).

[1] Another issue in the construction of (dynamic) factor models is the determination of the number of factors. For a discussion of the literature and a criterion for the determination of the number of factors, see [19].

0047-259X/$ – see front matter
doi:10.1016/j.jmva.2011.11.003

Building upon Otter and Jacobs [18], the paper exploits concepts from information theory, in particular Kullback–Leibler numbers, to analyse information in the data.[2] 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 recursive procedure including a test as to whether an additional variable adds information is given. Ordering the series of the data set according to the measure enables us to identify a subset of a data set that is most informative. The method can be used as a first step in the construction of a dynamic factor model or a leading index.

Our paper is related to Bai and Ng [4], who study 'hard' and 'soft' thresholding to reduce the influence of uninformative predictors for a variable from the point of view of factor forecasting. Hard thresholding involves some pretest procedure, while under soft thresholding the top ranked predictors according to some soft-thresholding rule are kept. Our paper fits into the category of soft thresholding; we also seek to identify a subset of a larger data set that is most informative. However, in contrast with the penalized regression models studied by Bai and Ng [4], the Least Absolute Shrinkage Selection Operator (LASSO) model of Tibshirani [22] and the elastic net rule of Zou and Hastie [24], our method is based on a quantitative measure of information adopting a factor model framework and does not rely on an external regression method.

We illustrate the concepts with a Monte Carlo simulation and with the macroeconomic data set of Stock and Watson [20], which consists of 132 monthly US variables and runs from 1959–2003. We find that relative information is indeed maximized for a limited number of series. In the Stock and Watson data set relative information is maximized for 40–50 series, if we are interested in modelling industrial production and CPI inflation.

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. After a Monte Carlo study in Section 3, we apply our method to the US data set of Stock and Watson [20] in Section 4. Section 5 concludes.

2. Information in data

2.1. Kullback–Leibler numbers and information

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

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

and $G = G_1 + G_2$ is the measure of information for discriminating between the two density functions with $G = 0$ in the case of $f_1(\boldsymbol{x}) = f_2(\boldsymbol{x})$ and $G = \infty$ in case of perfect discrimination; see [23, p. 245]. For a general background, see [7].

For $\mathrm{tr}(\boldsymbol{\Gamma}) = \mathrm{tr}(\boldsymbol{\Lambda}) = N$ we have $G_1 = -\mathrm{logdet}(\boldsymbol{\Lambda})$, where G_1 is the mean information in $\boldsymbol{x}$ for discriminating between $f_1(\boldsymbol{x})$ and $f_2(\boldsymbol{x})$, see [15], and $G_2 = \mathrm{logdet}(\boldsymbol{\Lambda}) + \frac{1}{2}\left(\mathrm{tr}(\boldsymbol{\Lambda}^{-1}) - N\right)$. Therefore

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

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.

We can also use the entropy measure. Let $\boldsymbol{x}_t$ again 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 $E(\boldsymbol{x}_t\boldsymbol{x}_t') = \boldsymbol{\Gamma}$, i.e. $\boldsymbol{x}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Gamma})$, where $\mathrm{diag}(\boldsymbol{\Gamma}) = (1, 1, \ldots, 1)$ and $\mathrm{tr}(\boldsymbol{\Gamma}) = N$. 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

$$2H_x = -2E_x[\log f(\boldsymbol{x})] = cN + \mathrm{logdet}(\boldsymbol{\Gamma}),$$

where $c \equiv \log(2\pi) + 1 \approx 2.84$, with $2H_{x,\max} = cN$ in the case of $\boldsymbol{\Gamma} = \boldsymbol{I}_N$; see e.g. Goodwin and Payne (1977) [10]. The information or negentropy is defined as

$$2\mathrm{Inf}_x \equiv 2(H_{x,\max} - H_x) = -\mathrm{logdet}(\boldsymbol{\Gamma}) \geq 0, \tag{3}$$

which is zero in the case of $\boldsymbol{\Gamma} = \boldsymbol{I}_N$. This measure coincides with Kullback–Leibler information G_1. We define the *relative information* as

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

If $H_{x(N)}$ is equal to $H_{\max}$ then $\mathrm{Inf}_N^R = 0$; if $H_{x(N)} = 0$ then $\mathrm{Inf}_N^R = 1$. The relative information equals the weighted mean information per variable in the data vector $\boldsymbol{x}_t$, where the weight is $1/c$.

[2] Jacobs and Otter [14] apply similar information concepts to derive a formal test for the number of common factors and the lag order in a dynamic factor model.