

# A criterion for the number of factors in a data-rich environment

Pieter W. Otter  
University of Groningen

Jan P.A.M. Jacobs  
University of Groningen, CAMA and CIRANO

Ard H.J. den Reijer\*  
Sveriges Riksbank

This version: February 2011

## Abstract

Recently, Onatski (2009) provided a formal justification for the use of the scree test in determining the number of factors in large factor models. We derive a criterion for the determination of the number of factors, that is associated with Onatski's test statistic but does not rely on frequency domain analysis. Monte Carlo experiments compare the finite-sample properties of our criterion to Bai and Ng (2007), Hallin and Liška (2007) and Onatski (2009), and show that our criterion is superior. An application with the U.S. macroeconomic data set of Stock and Watson (2005) illustrates our procedure.

*Keywords:* factor model, number of factors, test

*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](mailto:ard.den.reijer@riksbank.se)

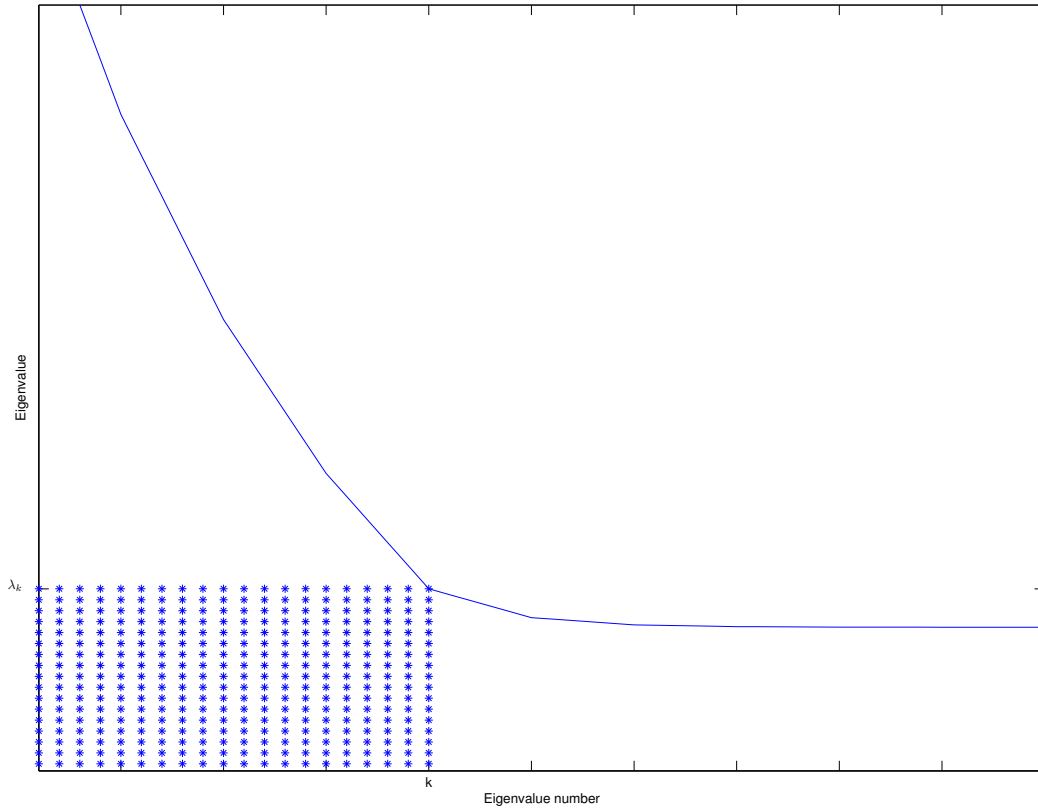
# 1 Introduction

A widely used method to analyse large quantities of data in the social sciences is factor analysis, in which the variation in a large number of observed variables is described in fewer unobserved variables, or movements in a large number of series are driven by a limited set of common ‘factors’. One of the issues in factor analysis is the determination of the number of unobserved variables to retain, i.e. the number of factors. Various methods are in use: (i) heuristic methods like the Kaiser criterion in which only factors with eigenvalues greater than 1 are retained, or the scree test of Cattell (1966), which will be explained in more detail below; (ii) stopping rules, see e.g. Peres-Neto, Jackson and Somers (2005); or (iii) principal components analysis, see e.g. Jolliffe (2002, Chapter 6) or Coste, Ecosse, Leplège, and Pouchot (2005).

In recent years, large dimensional factor models have become more and more popular in econometric research too. For an overview of recent developments see Bai and Ng (2008). The determination of the number of factor is still high on the research agenda despite the fact that many studies have proposed solutions and consistent estimators using different factor model and distributional assumptions. See e.g. Connor and Korajczyk (1993); Bai and Ng (2002, 2007), Amengual and Watson (2007), Kapetanios (2010), Hallin, and Liška (2007), Harding (2009), Jacobs and Otter (2008), and Onatski (2009).

Recently, Onatski (2009) provided a formal justification for the use of the scree test in determining the number of factors. The scree test, a graphical technique, consists of plotting the eigenvalues  $\lambda_k$  against its component num-

Figure 1: Stylized scree plot



ber  $k$  as in Figure 1, and deciding at which value of  $k$  the slopes of the plotted points are ‘steep’ to the left of  $k$  and ‘not steep’ at the right of  $k$ . This value of  $k$ , which defines an ‘elbow’ in the graph, is then taken to be the number of factors to be retained. Onatski proposes to look at the evolution in the fraction of subsequent differences between eigenvalues,  $(\gamma_k - \gamma_{k+1})/(\gamma_{k+1} - \gamma_{k+2})$ , where  $\gamma_i$  is the  $i$ -th largest eigenvalue of the smoothed periodogram estimate of the spectral density matrix of data at a pre-specified frequency. He describes the asymptotic distribution of the statistic as the number of variables

$N$  and the number of observations  $T$  rise as a function of the Tracy-Widow distribution, and tabulates the critical values of the test.

This paper provides an alternative criterion associated to Onatski's test statistic, which in our set-up runs in singular values rather than frequency domain eigenvalues. Our approach is more general. Whereas Onatski assumes that the data have the generalized dynamic structure introduced by Forni, Hallin, Lippi and Reichlin (2000), our approach does not start by assuming an explicit factor model of the data, but investigates whether a factor structure is appropriate. Monte Carlo show the superiority of our criterion compared to Bai and Ng (2007), Hallin and Liška (2007) and Onatski (2009) for different combinations of  $N$  and  $T$ .

The rest of the paper is structured as follows. Section 2 describes our criterion, whereas Section 3 presents some Monte Carlo simulations experiments. Section 4 applies our test to the Stock and Watson (2005) data set. Section 5 concludes.

## 2 Method

### 2.1 Population model

Consider a  $N$ -dimensional stationary normalized random vector  $\mathbf{x}_t$  with zero mean and covariance matrix  $\mathbf{\Gamma}$ , with  $\text{tr}(\mathbf{\Gamma}) = N$ . The vector  $\mathbf{x}_t$  can be expressed as  $\mathbf{x}_t = \mathbf{B}\tilde{\mathbf{F}}_t$ , with  $\tilde{\mathbf{F}}_t \sim \mathcal{N}_N(\mathbf{0}, \mathbf{I}_N)$ ,  $\mathbf{B}\mathbf{B}' = \mathbf{\Gamma}$ , while  $\mathbf{B}$  is an  $N \times N$  matrix. Applying the singular value decomposition (SVD)  $\mathbf{B} = \mathbf{C}\mathbf{\Sigma}\mathbf{W}'$ , where  $\mathbf{C}'\mathbf{C} = \mathbf{C}\mathbf{C}' = \mathbf{I}_N$ ,  $\mathbf{W}'\mathbf{W} = \mathbf{W}\mathbf{W}' = \mathbf{I}_N$ ,  $\mathbf{\Sigma} = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_N)$

with singular values  $\sigma_1 > \sigma_2 > \dots > \sigma_N$ , we get

$$\mathbf{x}_t = \mathbf{C}\boldsymbol{\Sigma}\mathbf{W}'\tilde{\mathbf{F}}_t \equiv \mathbf{C}\boldsymbol{\Sigma}\mathbf{F}_t, \quad \mathbf{F}_t \sim \mathcal{N}_N(\mathbf{0}, \mathbf{I}_N).$$

The variance-covariance matrix  $\boldsymbol{\Gamma}$  becomes  $\boldsymbol{\Gamma} = \mathbf{C}\boldsymbol{\Sigma}^2\mathbf{C}' \equiv \mathbf{C}\boldsymbol{\Lambda}\mathbf{C}'$ , so  $\sigma_j^2 = \lambda_j$ ,  $j = 1, \dots, N$  and  $\text{tr}(\boldsymbol{\Sigma}^2) = \text{tr}(\boldsymbol{\Lambda}) = N$ . For any  $1 < k < N$  we have the following orthogonal least squares decomposition

$$\mathbf{x}_t = \mathbf{C}\boldsymbol{\Sigma}\mathbf{F}_t = \mathbf{C}_1\boldsymbol{\Sigma}_1\mathbf{F}_{1,t} + \mathbf{C}_2\boldsymbol{\Sigma}_2\mathbf{F}_{2,t} \equiv \tilde{\mathbf{x}}_t + \boldsymbol{\varepsilon}_t, \quad (1)$$

where

$$\begin{aligned} \mathbf{C} &= \begin{bmatrix} \mathbf{C}_1 & \mathbf{C}_2 \end{bmatrix}, \quad \mathbf{C}_1 \in \mathbb{R}^{N \times k}, \mathbf{C}_2 \in \mathbb{R}^{N \times (N-k)}, \\ \boldsymbol{\Sigma} &= \begin{bmatrix} \boldsymbol{\Sigma}_1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_2 \end{bmatrix}, \quad \boldsymbol{\Sigma}_1 = \text{diag}(\sigma_1, \dots, \sigma_k), \boldsymbol{\Sigma}_2 = \text{diag}(\sigma_{k+1}, \dots, \sigma_N), \\ \mathbf{F}_t &= \begin{bmatrix} \mathbf{F}'_{1,t} & \mathbf{F}'_{2,t} \end{bmatrix}', \quad \dim(\mathbf{F}'_{1,t}) = k, \dim(\mathbf{F}'_{2,t}) = N - k. \end{aligned}$$

We have  $\text{E}\{\tilde{\mathbf{x}}_t\boldsymbol{\varepsilon}'_t\} = \mathbf{C}_1\boldsymbol{\Sigma}_1\text{E}\{\mathbf{F}_{1,t}\mathbf{F}'_{2,t}\}\boldsymbol{\Sigma}'_2\mathbf{C}'_2 = 0$  and  $\text{E}\{\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}'_t\} = \text{tr}(\boldsymbol{\Sigma}_2^2) = \sum_{j=k+1}^N \sigma_j^2$  is the minimum.

Let  $\|\mathbf{A}\|_E = \left(\sum_{i,j} |a_{i,j}|^2\right)^{1/2} = \text{tr}(\mathbf{A}'\mathbf{A})^{1/2}$  be the Euclidean (Schur) norm of the matrix  $\mathbf{A}$ . The expected Euclidean norm of  $\mathbf{x}_t$ ,  $\|\mathbf{x}_t\| = \sqrt{\mathbf{x}'_t\mathbf{x}_t}$  can be written as

$$\text{E}\{\|\mathbf{x}_t\|^2\} = \underbrace{\text{E}\{\|\tilde{\mathbf{x}}_t\|^2\}}_{\text{'explained' variance}} + \underbrace{\text{E}\{\|\boldsymbol{\varepsilon}_t\|^2\}}_{\text{'error' variance}} = \sum_{j=1}^k \lambda_j + \sum_{j=k+1}^N \lambda_j = N, \quad (2)$$

with  $\text{tr}(\mathbf{\Gamma}) = N$ . Hence  $\text{E}\{||\mathbf{x}_t||^2/\sqrt{N}\} = \sum_{j=1}^k \lambda_j/N + \sum_{j=k+1}^N \lambda_j/N = 1$ . From Equation (1) the ‘explained’ variance can be approximated by using the factor basis  $\mathbf{C}_1 \mathbf{\Sigma}_1$

$$\text{E}\{||\tilde{\mathbf{x}}_t||^2\} = \sum_{j=1}^k \lambda_j/N = \frac{k}{N} \lambda_k + \sum_{j=1}^k \delta_j(k)/N, \quad (3)$$

with  $\delta_j(k)/N \equiv (\lambda_j - \lambda_k)/N > 0$ ,  $j = 1, \dots, k-1$ . So, the ‘explained’ variance  $\text{E}\{||\tilde{\mathbf{x}}_t||_E^2\}$  has a lower bound  $J(k)$  equal to  $J(k) \equiv \frac{k}{N} \lambda_k$ . Because the eigenvalues are ordered,  $\lambda_1 > \lambda_2 > \dots > \lambda_N$ ,  $J(k)$  is a trade-off between  $k$  and  $\lambda_k/N$ : if  $k$  increases,  $\lambda_k/N$  becomes smaller. Therefore, we can use  $J(k)$  as a criterion to maximize the lower bound of the ‘explained’ variance in the data. But from Equation (2) we have

$$\begin{aligned} \text{E}\{||\mathbf{x}_t/\sqrt{N}||^2\} &= \sum_{j=1}^k \lambda_j/N + \sum_{j=k+1}^N \lambda_j/N \equiv \frac{k}{N} \lambda_k + \sum_{j=1}^k \delta_j(k)/N + S_\varepsilon(k) \\ &= J(k) + J^c(k) = 1, \end{aligned}$$

with  $J^c(k) \equiv \sum_{j=1}^k \delta_j(k)/N + S_\varepsilon(k)$  and  $S_\varepsilon(k) \equiv \sum_{j=k+1}^N \lambda_j/N$  is the ‘error’ variance. Therefore, maximizing the criterion  $J(k)$  is equivalent to minimizing  $J^c(k)$ , where  $\sum_{j=1}^k \delta_j(k)/N$  is the penalty function when increasing  $k$ .

Here we introduce a possible factor structure in the data vector  $\mathbf{x}_t$ .

**Definition 1**  $\mathbf{x}_t$  has a factor structure if there exists a  $\kappa < N$  such that

$$\begin{aligned} \lim_{N \rightarrow \infty} \lambda_j/N = \tilde{\lambda}_j > 0 &\Rightarrow \lambda_j = O(N), \quad j = 1, 2, \dots, \kappa \\ \lim_{N \rightarrow \infty} \lambda_j/N = 0 &\Rightarrow \lambda_j = o(N), \quad j = \kappa + 1, \dots, N. \end{aligned}$$

Note that it might not be meaningful to impose a factor structure when  $\kappa$  is ‘large’ and  $\tilde{\lambda}_\kappa$  is ‘small’.

If  $\mathbf{x}_t$  has a factor structure, then

$$\bar{J}(k) = \lim_{N \rightarrow \infty} J(k) = \begin{cases} k\tilde{\lambda}_k > 0 & \text{for } k = 1, 2, \dots, \kappa, \\ 0 & \text{for } k = \kappa + 1, \dots, N. \end{cases}$$

If  $J(k)$  for  $k = 1, 2, \dots$  is increasing and reaches its maximum at some  $k$ , it is to be expected that the derivative  $DJ(k) \equiv \Delta J(k)/\Delta k = (J(k+1) - J(k))/(k+1 - k) = J(k+1) - J(k) = -\delta_k k + \lambda_{k+1}/N$ , with  $\delta_k \equiv (\lambda_k - \lambda_{k+1})/N > 0$  is positive and decreasing whenever  $J(k)$  has a maximum.

Assuming that a factor structure exists, we look at  $\lim_{N \rightarrow \infty} DJ(\kappa) = \overline{DJ}(\kappa)$ . Since  $\lim_{N \rightarrow \infty} \delta_\kappa = \lim_{N \rightarrow \infty} \frac{\lambda_\kappa}{N} - \lim_{N \rightarrow \infty} \frac{\lambda_{\kappa+1}}{N+1} = \tilde{\lambda}_\kappa > 0$ ,  $\overline{DJ}(\kappa) = -\tilde{\lambda}_\kappa \kappa < 0$ , while  $\overline{DJ}(\kappa + j) = 0$  for  $j = 1, 2, \dots$ . In addition  $\overline{DJ}(k) = -\tilde{\delta}_k k + \tilde{\lambda}_k$  for  $k < \kappa$ , with  $\tilde{\delta}_k = \lim_{N \rightarrow \infty} \lambda_k/N - \lim_{N \rightarrow \infty} \lambda_{k+1}/(N+1) = (\tilde{\lambda}_k - \tilde{\lambda}_{k+1}) > 0$ . So  $\overline{DJ}(k)$  is positive, decreasing for  $k < \kappa$  with  $\overline{DJ}(\kappa) = -\tilde{\lambda}_\kappa \kappa < 0$  and zero for  $k > \kappa$ , which implies  $\kappa = \arg \min \overline{DJ}(k)$ ,  $k = 1, 2, \dots$ . Note that the magnitude of the negative value  $\overline{DJ}(\kappa) = -\tilde{\lambda}_\kappa \kappa < 0$  gives an indication for the appropriateness of a factor structure. If  $\kappa$  is large and  $\overline{DJ}(\kappa)$  is small, then a factor structure is not meaningful.

The correspondence to the test statistic of Onatski (2009) can be seen as follows. Since  $DJ(k)$  is decreasing,  $DJ(k+1) < DJ(k)$  which is equivalent to  $J(k+2) - J(k+1) < J(k+1) - J(k)$ , from which follows that  $J(k) - J(k+1) < J(k+1) - J(k+2)$ . Using the definition of  $\bar{J}(k) \equiv k\tilde{\lambda}_k$  we obtain

$$\frac{\tilde{\lambda}_k - \tilde{\lambda}_{k+1}}{\tilde{\lambda}_{k+1} - \tilde{\lambda}_{k+2}} = \begin{cases} 1 + \frac{2}{k}, & k = 1, 2, \dots, \kappa - 1, \\ \infty, & k = \kappa, \dots \end{cases}$$

So with normalized data, the Onatski (2009) test statistic has an upper bound equal to  $1 + 2/k$ .

## 2.2 Data

Let the  $T \times N$  data matrix  $\mathbf{X}$  consist of normalized realizations of  $\mathbf{x}_t \in$

$$\mathbb{R}^N, t = 1, 2, \dots, T: \mathbf{X} = \begin{bmatrix} \mathbf{x}'_1 \\ \mathbf{x}'_2 \\ \vdots \\ \mathbf{x}'_T \end{bmatrix}. \text{ The singular value decomposition to } \mathbf{X}$$

yields  $\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{C}'$ . We distinguish two cases: (i)  $T \geq N$  and (ii)  $T < N$ .

(i) If  $T > N$  the SVD becomes

$$\mathbf{X} = \mathbf{U} \begin{bmatrix} \mathbf{S}_1 \\ \mathbf{0} \end{bmatrix} \mathbf{C}' = \mathbf{U}_1 \mathbf{S}_1 \mathbf{C}',$$

where  $\mathbf{U}'\mathbf{U} = \mathbf{U}\mathbf{U}' = \mathbf{I}_T$ ,  $\mathbf{U} = \begin{bmatrix} \mathbf{U}_1 & \mathbf{U}_2 \end{bmatrix}$ ,  $\mathbf{U}_1 \in \mathbb{R}^{T \times N}$  and  $\mathbf{S} = \text{diag}(s_1, s_2, \dots, s_N)$ ,  $s_1 > s_2 > \dots > s_N$ ;

(ii) For  $T < N$  we get

$$\mathbf{X} = \mathbf{U} \begin{bmatrix} \mathbf{S}_1 & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{C}'_1 \\ \mathbf{C}'_2 \end{bmatrix} = \mathbf{U} \mathbf{S}_1 \mathbf{C}',$$

with  $\mathbf{C} = \begin{bmatrix} \mathbf{C}_1 & \mathbf{C}_2 \end{bmatrix}$ ,  $\mathbf{C}_1 \in \mathbb{R}^{N \times T}$ ,  $\mathbf{U} \in \mathbb{R}^{T \times T}$  and  $\mathbf{S} = \text{diag}(s_1, s_2, \dots, s_T)$ ,  
 $s_1 > s_2 > \dots > s_T$ .

Defining  $\bar{k} = \min(T, N)$  the two SVDs can be written as

$$\mathbf{X} = \mathbf{U}_{\bar{k}} \mathbf{S}_{\bar{k}} \mathbf{C}'_{\bar{k}},$$

with  $\mathbf{U}_{\bar{k}} = \mathbf{U}$  and  $\mathbf{C}_1 \in \mathbb{R}^{N \times T}$  when  $\bar{k} = T$ , and  $\mathbf{U}_{\bar{k}} = \mathbf{U}_1 \in \mathbb{R}^{T \times N}$  and  $\mathbf{C}_{\bar{k}} = \mathbf{C}$  when  $\bar{k} = N$ . The Euclidean norm of matrix  $\mathbf{X}$  is equal to  $\|\mathbf{X}\|_E^2 = \sum_{i,j} |x_{ij}|^2 = \text{tr}(\mathbf{X}'\mathbf{X}) = \text{tr}(\mathbf{S}_{\bar{k}}) = \sum_{j=1}^{\bar{k}} s_j^2$ . Let  $\hat{\mathbf{\Gamma}} = \mathbf{X}'\mathbf{X}/T$  be an estimate of the covariance matrix  $\mathbf{\Gamma}$ . Then  $\|\frac{1}{\sqrt{T}}\mathbf{X}\|_E^2 = \text{tr}(\mathbf{X}'\mathbf{X}/T) = \text{tr}(\mathbf{\Gamma}) = \sum_{j=1}^{\bar{k}} s_j^2/T = N$ , because of the normalization of the data, and hence  $\|\frac{1}{\sqrt{NT}}\mathbf{X}\|_E^2 = \sum_{j=1}^{\bar{k}} s_j^2/NT = 1$ . And letting  $s_j^2/T \equiv \hat{\lambda}_j$ , we get  $\sum_{j=1}^{\bar{k}} \hat{\lambda}_j/N = 1$ .

Using the SVD  $\frac{\mathbf{X}}{\sqrt{NT}} = \mathbf{U}_{\bar{k}} \frac{\mathbf{S}_{\bar{k}}}{\sqrt{NT}} \mathbf{C}'_{\bar{k}}$  we get

$$\left\| \frac{\mathbf{X}}{\sqrt{NT}} \right\|_E^2 = \text{tr} \left( \mathbf{U}_{\bar{k}} \frac{\mathbf{S}_{\bar{k}}}{\sqrt{NT}} \mathbf{C}'_{\bar{k}} \mathbf{C}_{\bar{k}} \frac{\mathbf{S}_{\bar{k}}}{\sqrt{NT}} \mathbf{U}'_{\bar{k}} \right) = \sum_{j=1}^{\bar{k}} \frac{s_j^2}{NT} = 1.$$

Define  $\bar{s}_j^2 \equiv \frac{s_j^2}{NT}$ ,  $j = 1, \dots, \bar{k}$ , so  $\{\bar{s}_j\}$  are the singular values of  $\mathbf{X}/\sqrt{NT}$ .

## 2.3 Estimation

As seen above, our criterion is equal to  $\widehat{DJ}(k) = (k+1)\bar{s}_{k+1}^2 - k\bar{s}_k^2$  for  $k = 1, \dots, \bar{k}$  and we propose to  $\min_{1 \leq k \leq \bar{k}} \widehat{DJ}(k)$ , with  $\hat{k} \equiv \arg \min(\widehat{DJ}(k))$ .

The scaled data matrix  $\mathbf{X}/\sqrt{NT}$  can be decomposed as

$$\frac{\mathbf{X}}{\sqrt{NT}} = \mathbf{U}_1 \bar{\mathbf{S}}_1 \mathbf{C}'_1 + \mathbf{U}_2 \bar{\mathbf{S}}_2 \mathbf{C}'_2 = \hat{\mathbf{X}} + \mathbf{E},$$

with  $\mathbf{U}_1 \in \mathbb{R}^{T \times \hat{k}}$ ,  $\bar{\mathbf{S}}_1 = \text{diag}(\bar{s}_1, \dots, \bar{s}_{\hat{k}})$ ,  $\mathbf{C}_1 \in \mathbb{R}^N \times \hat{k}$ ,  $\mathbf{U}_2 \in \mathbb{R}^{T \times (\bar{k} - \hat{k})}$ ,  $\bar{\mathbf{S}}_1 = \text{diag}(\bar{s}_{\hat{k}+1}, \dots, \bar{s}_{\bar{k}})$ , and  $\mathbf{C}_2 \in \mathbb{R}^{N \times (\bar{k} - \hat{k})}$ . Because  $\hat{\mathbf{X}} = \begin{bmatrix} \hat{\mathbf{x}}'_1 \\ \vdots \\ \hat{\mathbf{x}}'_T \end{bmatrix} = \mathbf{U}_1 \bar{\mathbf{S}}_1 \mathbf{C}'_1$ , we

have  $\hat{\mathbf{x}}_t = \mathbf{C}_1 \bar{\mathbf{S}}_1 \hat{\mathbf{F}}_{1,t}$ , where  $\mathbf{U}'_1 = \begin{bmatrix} \hat{\mathbf{F}}_{1,1} & \dots & \hat{\mathbf{F}}_{1,T} \end{bmatrix}$  with  $\hat{\mathbf{F}}_{1,t} \in \mathbb{R}^{\hat{k} \times 1}$ , and the estimated variance of the factors equals  $\mathbf{U}'_1 \mathbf{U}_1 = \sum_{t=1}^T \hat{\mathbf{F}}_{1,t} \hat{\mathbf{F}}'_{1,t} = \mathbf{I}_{\hat{k}}$ . From

$\mathbf{E} = \begin{bmatrix} e_1 \\ \vdots \\ e_T \end{bmatrix} = \mathbf{U}_2 \bar{\mathbf{S}}_2 \mathbf{C}'_2$ , we get  $\mathbf{e}_t = \mathbf{C}_2 \bar{\mathbf{S}}_2 \hat{\mathbf{F}}_{2,t}$ , where  $\mathbf{U}'_2 = \begin{bmatrix} \hat{\mathbf{F}}_{2,1} & \dots & \hat{\mathbf{F}}_{2,T} \end{bmatrix}$ ,

with variance  $\mathbf{U}'_2 \mathbf{U}_2 = \sum_{t=1}^T \hat{\mathbf{F}}_{2,t} \hat{\mathbf{F}}'_{2,t} = \mathbf{I}_{\bar{k} - \hat{k}}$ . So the estimated model is

$$\mathbf{x}_t = \hat{\mathbf{x}}_t + \mathbf{e}_t,$$

where  $\hat{\mathbf{x}}_t = \mathbf{C}_1 \bar{\mathbf{S}}_1 \hat{\mathbf{F}}_{1,t}$  and  $\mathbf{e}_t = \mathbf{C}_2 \bar{\mathbf{S}}_2 \hat{\mathbf{F}}_{2,t}$ , with  $\mathbf{x}'_t \mathbf{e}_t = 0$  for all  $t$  because  $\mathbf{C}'_1 \mathbf{C}_2 = 0$ .

## 2.4 Asymptotics

To be done.

### 3 Monte Carlo experiments

To assess the finite-sample properties of our criterion, we compare it to the methods of Bai and Ng (2007) (BN henceforth), Hallin and Liška (2007) (HL henceforth) and Onatski (2009).<sup>1</sup> We consider the generalized dynamic factor structure

$$x_{it} = \Lambda_{i1}(L) F_{1t} + \dots + \Lambda_{ik1}(L) F_{kt} + e_{it}, \quad (4)$$

where  $\Lambda_{i1}(L) = \sum_{u=0}^{\infty} \Lambda_{ij}^{(u)} L^u$  with lag operator  $L$ , factor loadings  $\Lambda_{ij}^{(u)}$ , factors  $F_{jt}$  and idiosyncratic term  $e_{it}$ .

We replicate Onatski's modification of Hallin and Liška's (2007) Monte Carlo experiment and generate data from model (4) as follows:

1. the  $k$ -dimensional factor vectors  $F_{jt}$  are i.i.d.  $N(0, I_k)$ .
2. the filters  $\Lambda_{ik}(L)$ , ( $i = 1, \dots, n$ ;  $k = 1, \dots, q$ ) are randomly generated independently from the  $F_{jt}$ 's by one of the following two devices:

**MA loadings:**  $\Lambda_{ik}(L) = b_{ij}^{(0)} \left(1 + b_{ij}^{(1)} L\right) \left(1 + b_{ij}^{(2)} L\right)$  with i.i.d. and mutually independent coefficients  $b_{ij}^{(0)} \sim N(0, 1)$ ,  $b_{ij}^{(1)} \sim U[0, 1]$  and  $b_{ij}^{(2)} \sim U[0, 1]$

**AR loadings:**  $\Lambda_{ik}(L) = b_{ij}^{(0)} \left(1 - b_{ij}^{(1)} L\right)^{-1} \left(1 - b_{ij}^{(2)} L\right)^{-1}$  with i.i.d. and mutually independent coefficients  $b_{ij}^{(0)} \sim N(0, 1)$ ,  $b_{ij}^{(1)} \sim U[.8, .9]$  and  $b_{ij}^{(2)} \sim U[.5, .6]$

---

<sup>1</sup>We thank Alexei Onatski and Roman Liška for making available their Matlab programs. The Bai-Ng programs are downloaded from Serena Ng's homepage.

3. the idiosyncratic components  $e_{it}$  follow  $AR(1)$ -processes both cross-sectionally and over time:  $e_{it} = \rho_i e_{it-1} + v_{it}$  and  $v_{it} = \rho v_{i-1t} + u_{it}$ , with i.i.d coefficients  $\rho_i \sim U[-.8, .8]$ ,  $\rho = 0.2$  and  $u_{it} \sim N(0, 1)$  i.i.d. and independently generated from  $\Lambda_{ik}(L)$  and  $F_{jt}$ , cf. Onatski (2009). The support  $[-.8, .8]$  of the uniform distribution has been chosen to match the range of the first-order autocorrelations of the estimated idiosyncratic components of the Stock and Watson (2005) dataset.
4. For each  $i$ , the variance of  $e_{it}$  and that of the common components  $\sum_{j=1}^k \Lambda_{ij}(L) F_{jt}$  are normalized such that their variances equal  $0.4 + 0.05k$  and  $1 - (0.4 + 0.05k)$ , respectively. Hence, a 2-factor model explains 50% of the data variation and a 7-factor model 75% for  $\sigma = 1$ . As a final step, the idiosyncratic part is magnified by  $\sigma \geq 1$ .

Then the different test procedures are employed to determine the number of factors in the simulated data sets. For the Onatski-procedure, the parameter  $\alpha$  equals the maximum of 0.01 and the p-value of the test of  $H_0 : k = 0$  vs.  $H_1 : 0 < k \leq 4$ . So,  $\alpha$  is calibrated such that the test has enough power to reject the false null hypothesis of no factors. Then the algorithm proceeds to test  $H_0 : k = k_1$  vs.  $H_1 : k_1 < k \leq 4$ . If  $H_0$  is not rejected, stop. Otherwise, test  $H_0 : k = k_1 + 1$  vs.  $H_1 : k_1 + 1 < k \leq 4$ . Repeat the procedure until  $H_0$  is not rejected. The Onatski-test requires the parameter  $m$  for grid size of approximating frequencies and is set at  $m = 30, 40, 65$  for  $T = 70, 120, 500$  respectively. Denoted in the original notation of the corresponding paper, for the Bai-Ng estimator, we use the  $\widehat{D}_{1,k}$  statistic for the residuals of a VAR(4), set the maximum number of static factors at 10 and consider

$\delta = 0.1$  and  $m = 2$ . For the Hallin-Liška estimator, we use the information criterion  $IC_{2;n}^T$  with penalty  $p_1(n, T)$ , set the truncation parameter  $M_T$  at  $\lceil 0.7\sqrt{T} \rceil$  and consider subsample sizes  $(n_j, T_j) = (n - 10j, T - 10j)$  with  $j = 0, 1, \dots, 3$ . We chose the penalty multiplier  $c$  on a grid  $0.01 : 0.01 : 3$  using Hallin-Liška’s second “stability interval” procedure.<sup>2</sup> Finally, we note that our proposed procedure does not require auxiliary parameters and is therefore straightforward to implement.

Table 1 reports the percentages of 500 simulation that deliver 1, 2, 3 and 4 estimated number of factor  $\hat{k}$  for Onatski’s (2009, Table IV) choices of  $n, T$  and  $\sigma^2$ . Some minor differences occur. The Bai-Ng application in case  $\sigma^2 = 1$  for AR-loadings shows better results in our application, while Onatski obtains better results for the Hallin-Liška-estimator. The table shows that our criterion procedure clearly outperforms the other procedures.<sup>3</sup> Table 2 reports the results of the extended simulation analysis with the true number of factors being  $k = 3$ , an extended  $(n, T)$ –grid and estimators being constrained to lie in the range from 1 to 10.

---

<sup>2</sup>In case the algorithm does not produce a second “stability interval”, we refine the increments of the grid to 0.001 instead. If the algorithm still fails to produce a second “stability interval”, then the algorithm determines the number that prevails just after the end of the original first “stability interval”.

<sup>3</sup>The only two exceptions are  $n = 70, T = 70, \sigma^2 = 1$  and  $n = 100, T = 120, \sigma^2 = 1$  with AR-loadings.

Table 1: Monte Carlo replications of the dynamic factor model

$N$	$T$	$\hat{k} =$	Onatski			Hallin-Liška			Bai-Ng			Jacobs-Otter-Reijer						
			1	2	3	4	1	2	3	4	1	2	3	4				
		$\sigma^2$	MA loadings															
70	70	1	0	100	0	0	7.8	92.2	0	0	0.4	99.6	0	0	0	100	0	0
70	70	2	1.6	97	1.2	0.2	5.6	93.8	0.6	0	0.6	99.4	0	0	0.6	99.4	0	0
70	70	4	13.4	77.6	6.2	2.8	24.4	75.2	0.4	0	56.6	43.4	0	0	2.4	97.2	0.4	0
100	120	1	0	100	0	0	0	100	0	0	0	100	0	0	0	100	0	0
100	120	3	0	100	0	0	18.4	81.6	0	0	0	100	0	0	0	100	0	0
100	120	6	3.2	94	1.8	1	58.6	41.4	0	0	56.6	43.4	0	0	0.4	99.6	0	0
150	500	1	0	100	0	0	0	100	0	0	0	100	0	0	0	100	0	0
150	500	8	0	100	0	0	0.4	99.6	0	0	0	100	0	0	0	100	0	0
150	500	16	0.2	96.2	2.6	1	79.6	20.4	0	0	99.8	0.2	0	0	0	100	0	0
$n$	$T$	$\sigma^2$	AR loadings															
70	70	1	14.4	83.2	2	0.4	37.6	61.4	1	0	2.6	97.4	0	0	14.6	85.4	0	0
70	70	2	37.4	54.8	6.2	1.6	23.8	75.4	0.8	0	27.6	72.4	0	0	24.2	75.6	0	0.2
70	70	4	57.2	31.2	7	4.6	49.4	49.8	0.8	0	79	21	0	0	33	59.6	3	4.4
100	120	1	0.2	98.6	1	0.2	0.8	99.2	0	0	0	100	0	0	1.6	98.4	0	0
100	120	3	11.2	80.4	5.6	2.8	42.2	57.8	0	0	9.8	90.2	0	0	8	92	0	0
100	120	6	31	51.2	12	5.8	64.2	35.8	0	0	71.4	28.6	0	0	20.4	78.8	0.8	0
150	500	1	0	99.4	0.4	0.2	0	100	0	0	0	100	0	0	0	100	0	0
150	500	8	0	97.6	1	1.4	4	96	0	0	3.4	96.6	0	0	0	100	0	0
150	500	16	5	86.8	4.8	3.4	21.4	78.6	0	0	96.6	3.4	0	0	0.4	99.6	0	0

**Notes.**

Percentage of 500 Monte Carlo replications resulting in estimates 1, 2, 3 or 4 of the true number of factors  $k = 2$  according to the different estimators. The estimators are constrained to be in the range from 1 to 4.

Table 2: Monte Carlo replications of the dynamic factor model

$N$	$T$	$\hat{k}$	$\sigma^2$	Onatski				Hallin-Liska				Bai-Ng				Jacobs-Otter-Reijer			
				1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
				MA loadings															
70	70	1	0.4	0	99	0.6	3.8	5.2	83.6	6.2	0	29.2	70.8	0	0	0.4	99.6	0	
70	70	4	23.4	20.8	34.2	6.8	31.4	48.6	19	1	68.2	28.6	3.2	0	0	11.8	82.8	1	
70	70	8	41	20.8	9.4	5	96.6	3.2	0.2	0	100	0	0	0	0.4	6.2	13.4	4	
70	70	16	43.8	16.2	6.4	5.6	100	0	0	0	100	0	0	0	0	0	0.2	1.6	
100	120	1	0	0	100	0	0	0	100	0	0	0.8	99.2	0	0	0	100	0	
100	120	4	3	1	93.4	1.4	10.8	50.2	39	0	1	37.2	61.8	0	0	2	98	0	
100	120	8	24.2	11.8	39.8	6	98.4	1.6	0	0	99.8	0.2	0	0	0.2	8.2	86.2	0.6	
100	120	16	36.6	15.8	11.8	5.8	100	0	0	0	100	0	0	0	0.2	2	5.4	5.4	
120	100	1	0	0	99.6	0.4	0.2	0	99.8	0	0	1	99	0	0	0	100	0	
120	100	4	1.8	2.6	88	6.2	9.2	48.4	42.4	0	3.6	39.4	57	0	0	1.6	98.4	0	
120	100	8	16	16.4	38.6	12.8	97.6	2.4	0	0	99.8	0.2	0	0	0.4	9.8	77.6	1.2	
120	100	16	35.2	20.8	12	7.6	100	0	0	0	100	0	0	0	0.2	2.2	5.8	1.4	
150	500	1	0	0	100	0	0	0	100	0	0	0	100	0	0	0	100	0	
150	500	4	0	0	100	0	0	0	100	0	0	0	100	0	0	0	100	0	
150	500	8	0.2	0	99.6	0.2	0.4	23	76.6	0	0	7	93	0	0	0	100	0	
150	500	16	9	6.6	65.6	3	100	0	0	0	100	0	0	0	0	0	100	0	
500	150	1	0	0	100	0	0	0	100	0	0	0	100	0	0	0	100	0	
500	150	4	0	0	100	0	35.8	44	20.2	0	0	0	100	0	0	0	100	0	
500	150	8	0.4	0	99	0.6	96.6	3.4	0	0	1.6	25.6	72.8	0	0	0	100	0	
500	150	8	0.4	0	99	0.6	96.6	3.4	0	0	1.6	25.6	72.8	0	0	0	100	0	
				AR loadings															
70	70	1	18.6	18	53.8	3.2	24.2	12.6	60.8	1.6	0.2	37.6	62.2	0	7.6	27.2	65.2	0	
70	70	4	56	19	7	3.4	44.8	44.8	10.4	0	69.6	29.6	0.8	0	19.8	36.4	16.2	2.6	
70	70	8	59	15	3.8	4.2	90.2	9.4	0.4	0	98.2	1.8	0	0	9.6	8.8	2	0.8	
70	70	16	48.2	14.4	6.8	2.4	100	0	0	0	100	0	0	0	1.6	0.2	0.2	2.2	
100	120	1	1.4	0.8	94.8	1.2	1.2	9.4	89.4	0	0	0.8	99.2	0	0	7.6	92.4	0	
100	120	4	18.6	17.8	36.4	7.2	36.2	47.6	16.2	0	16.6	66.2	17.2	0	4.4	26	69.4	0.2	
100	120	8	38.6	22.4	13.8	6.6	88.6	11.4	0	0	90.6	9.4	0	0	11.4	29	35.2	1.8	
100	120	16	42	15.2	7.2	8.4	100	0	0	0	100	0	0	0	3.6	6.8	2.2	1.8	
120	100	1	0.2	0.8	98.8	0.2	8	27.4	64.6	0	0	1.6	98.4	0	0.8	11.4	87.8	0	
120	100	4	9.2	15.2	68.2	5.2	33.2	46.6	20.2	0	22.4	60.8	16.8	0	7.8	29	62.2	0.4	
120	100	8	28.8	35.2	27	3.6	87.2	12.4	0.4	0	91.2	8.8	0	0	15.6	32.6	24.4	3.2	
120	100	16	41.2	30	12	4.4	99.8	0.2	0	0	100	0	0	0	6.4	9.4	2.8	1.8	
150	500	1	0	0	98.8	0.4	0	0.2	99.8	0	0	0	100	0	0	0	100	0	
150	500	4	0	0	99	0.2	2.8	2	95.2	0	0	0	100	0	0	0	100	0	
150	500	8	0.2	0	93	4.4	8.8	9	82.2	0	1.8	37.4	60.8	0	0	0.6	99.4	0	
150	500	16	9.6	4.8	54	7.4	64.6	18.8	16.6	0	99.2	0.8	0	0	3.2	96.8	0		
500	150	1	0	0	99.6	0	0	0.4	99.6	0	0	0	100	0	0	0.4	99.6	0	
500	150	4	2.8	2.8	93.2	0.6	45.8	37.4	16.8	0	0	5.2	94.8	0	0.4	7.4	92.2	0	
500	150	8	29.2	15.6	45	3.4	81.4	17.8	0.8	0	11.6	64	24.4	0	2.6	12.6	84.8	0	
500	150	8	29.2	15.6	45	3.4	81.4	17.8	0.8	0	11.6	64	24.4	0	2.6	12.6	84.8	0	

**Notes.**

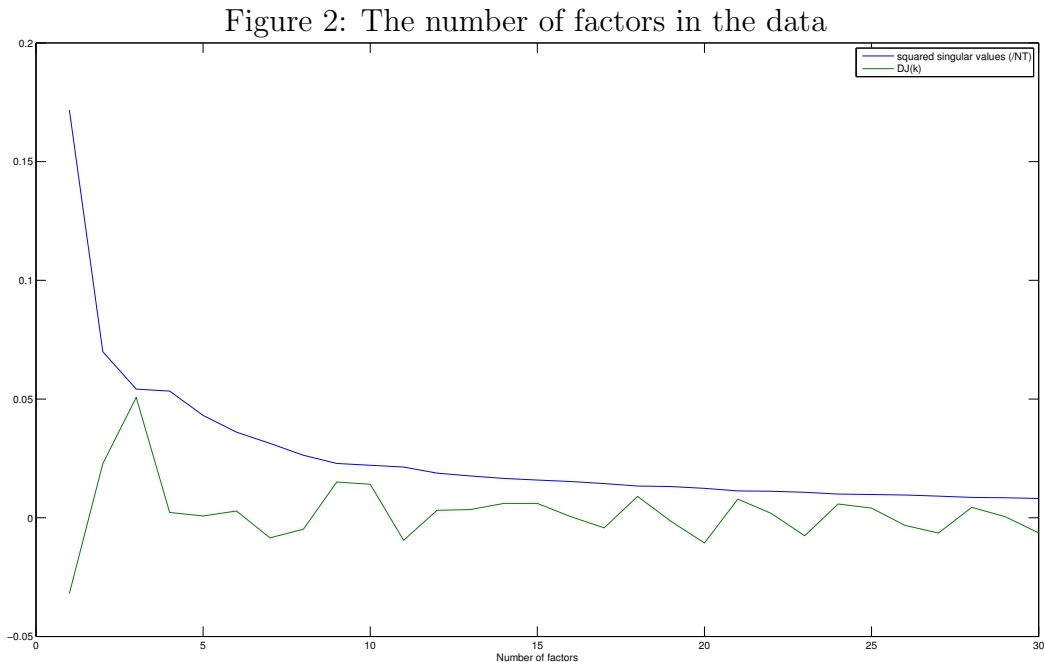
Percentage of 500 Monte Carlo replications resulting in estimates 1, 2, 3 or 4 of the true number of factors  $k = 3$  according to the different estimators. The estimators are constrained to be in the range from 1 to 10.

Three general observations emerge from the tables: (i) all procedures have a tendency to underestimate rather than overestimate the true number of factors; (ii) the Bai-Ng and the Hallin-Liška estimators do not capture the true number of factors even in the large dimension case  $(n, T) = (150, 500)$ , if the data is noisy, i.e. for  $\sigma^2 = 16$ ; (iii) though the Onatski-procedure is also based on the scree test, our proposed procedure shows generally a better performance. Our procedure does not involve auxiliary calculations related to the spectral decomposition, which might explain its relative efficiency.

## 4 Application

In this section we evaluate the performance of the suggested approach to the Stock and Watson (2005) macroeconomic data set, which consists of monthly observations on  $N = 132$  macroeconomic time series running from 1960M1 through 2003M12 ( $T = 528$ ). The authors describe how to transform the series by taking logarithms and/or differencing when necessary to assure approximate stationarity and correcting for outliers. This data set is considered a representative summary of the U.S. economy. Hallin and Liška (2007) propose to split the sample into two parts to account for a structural change in the U.S. economy that occurred around 1982–1983. Based on the subperiod 1960–1982 ( $T=276$ ), they find  $\hat{q} = 3$  factors and  $\hat{q} = 1$  factor for the second subperiod 1983-2003. For the full sample, they find  $\hat{q} = 1$  factor though its identification is less clear. Onatski (2009) restricts the analysis to business cycle frequencies and explicitly excludes cycle longer than 10 years. Employing his test procedure within the algorithm that continues increasing

the number of factors as long as the null hypothesis keeps on being rejected results in  $\hat{q} = 1$  factor for the full sample period. The alternative hypothesis of only 2 factors cannot be rejected at a significance level higher than 4%. Bai and Ng (2007) estimate  $\hat{q} = 4$  factors employing the full sample, but point out that there is substantial variation over the sample.

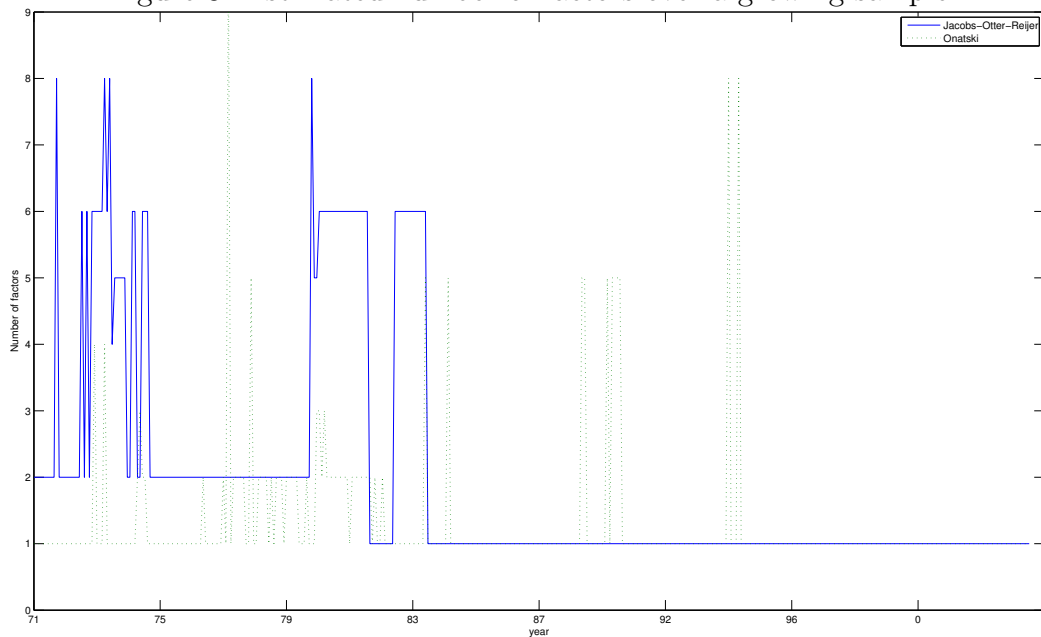


Notes.  $\bar{s}_j^2$  are the scaled (by  $NT$ ) squared singular values for  $j = 1, \dots, 30$ ;  $DJ(j)$  is our proposed criterion to estimate the number of factors.

Figure 2 shows the scaled (by  $NT$ ) squared singular values  $\bar{s}_j^2$  together with our proposed criterion  $DJ(j)$  to estimate the number of factors for  $j = 1, \dots, 30$ . The estimator for the number of factors  $\hat{k} = \arg \min_j \widehat{DJ}(j)$  results in  $\hat{k} = 1$  for this data set. We repeated this exercise and Figure 3 shows the resulting number of factors over the samples with ending dates running from 1970M12 (sample size  $T=133$ ) to 2003M12 (sample size  $T=528$ ) for both our proposed procedure and Onatski's (2009) algorithm (with  $m=65$ ). The first

part of the sample up to 1983 provides quite varying results suggesting mainly 2, but also possibly 6 factors for our proposed procedure, while Onatski tends to find 1 factor. For the larger samples ending later than 1983, both procedures result in  $\hat{k} = 1$  factor, while the irregularities remain for Onatski's procedure.

Figure 3: Estimated number of factors over a growing sample



Notes. The estimated number of factors is based on samples with end dates running from 1970M12 to 2003M12 (sample sizes from  $T = 133$  to  $T = 528$ ). The number of factors are based on Onatski (2009) with  $m = 65$  and our proposed procedure.

## 5 Conclusion

This paper presented a simple criterion to determine the number of factors in a data-rich environment. Our procedure is straightforward to implement and does neither require the specification of several auxiliary parameters as

in Bai and Ng (2007), nor the specification of an automated search procedure as in Hallin and Liška (2007). Our procedure is closely related to Onatski (2009), but is more straightforward (and therefore more efficient) as it does not involve cumbersome numerical transformations related to spectral decompositions and Fourier transformations. Monte Carlo simulations indicate that our criterion outperforms the procedures of Bai and Ng (2007), Hallin and Liška (2007) and Onatski (2009). Like Onatski, we find strong evidence of one dynamic factor  $\hat{k} = 1$  in the Stock and Watson (2005) data set.

## **Acknowledgements**

The paper has benefited from comments received following a presentation of a previous version at a Workshop on ‘Dynamic Factor Modelling’, Queen Mary College, London, October 2007. Views expressed are those of the individual authors and do not necessarily reflect official positions of the Riksbank.

## References

- Amengual, D. and M. Watson (2007), “Consistent estimation of the number of factors in a large  $N$  and  $T$  panel”, *Journal of Business & Economic Statistics*, **25**, 91–96.
- Bai, J. and S. Ng (2002), “Determining the number of factors in approximate factor models”, *Econometrica*, **70**, 191–221.
- Bai, J. and S. Ng (2007), “Determining the number of primitive shocks in factor models”, *Journal of Business and Economic Statistics*, **25**, 52–60.
- Bai, J. and S. Ng (2008), “Large dimensional factor analysis”, *Foundations and Trends in Econometrics*, **3**, 89–163.
- Cattell, R.B. (1966), “The scree test for the number of factors”, *Multivariate Behavioral Research*, **1**, 245–276.
- Connor, G. and R. Korajczyk (1993), “A test for the number of factors in an approximate factor model”, *The Journal of Finance*, **58**, 1263–1291.
- Coste, J., S. Bouée, E. Ecosse, A. Leplège, and J. Pouchot (2005), “Methodological issues in determining the dimensionality of composite health measures using principal component analysis: Case illustration and suggestions for practice”, *Quality of Life Research*, **14**, 641–654.
- Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2000), “The generalized dynamic-factor model: Identification and estimation”, *The Review of Economics and Statistics*, **82**, 540–554.

- Hallin, M. and R. Liška (2007), “Determining the number of factors in the general dynamic factor model”, *Journal of the American Statistical Association*, **102**, 603–617.
- Harding, Matthew C. (2009), “Structural estimation of high-dimensional factor models”, mimeo, Stanford University.
- 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.
- Jolliffe, I.T. (2002), *Principal Component Analysis*, Springer Series in Statistics, 2nd edition, Springer, New York.
- Kapetanios, G. (2010), “A testing procedure for determining the number of factors in approximate factor models with large datasets”, *Journal of Business & Economic Statistics*, **28**, 397–409.
- Onatski, A. (2009), “Testing hypotheses about the number of factors in large factor models”, *Econometrica*, **77**, 1447–1479.
- Perez-Neto, P.R., D.A. Jackson, and K.M. Somers (2005), “How many principal components? Stopping rules for determining the number of non-trivial axes revisited”, *Computational Statistics & Data Analysis*, **49**, 974–997.
- Stock, J.H. and M.W. Watson (2005), “Implications of dynamic factor models for VAR analysis”, Working paper 11467, National Bureau of Economic Research.