# 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, University of Tasmania, CAMA and CIRANO

Ard H.J. den Reijer
Sveriges Riksbank

This version: April 2015

**Abstract**

This paper derives a new criterion for the determination of the number of factors in static approximate factor models, that is strongly associated with the scree test. Our criterion looks for the number of eigenvalues for which the difference between adjacent eigenvalue – eigenvalue component number blocks is maximized. Monte Carlo experiments compare the properties of our criterion to the Edge Distribution (ED) estimator of Onatski (2010) and the two eigenvalue ratio estimators of Ahn and Horenstein (2013). Our criterion outperforms the latter two for all sample sizes and the ED estimator of Onatski (2010) for samples up to 300 variables/observations.

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

*JEL-code*: C32, C52, C82

*Correspondence to Jan P.A.M. Jacobs, Faculty of Economics and Business, University of Groningen, PO Box 800, 9700 AV GRONINGEN, the Netherlands. Tel.: +31 50 363 3681. Email: j.p.a.m.jacobs@rug.nl

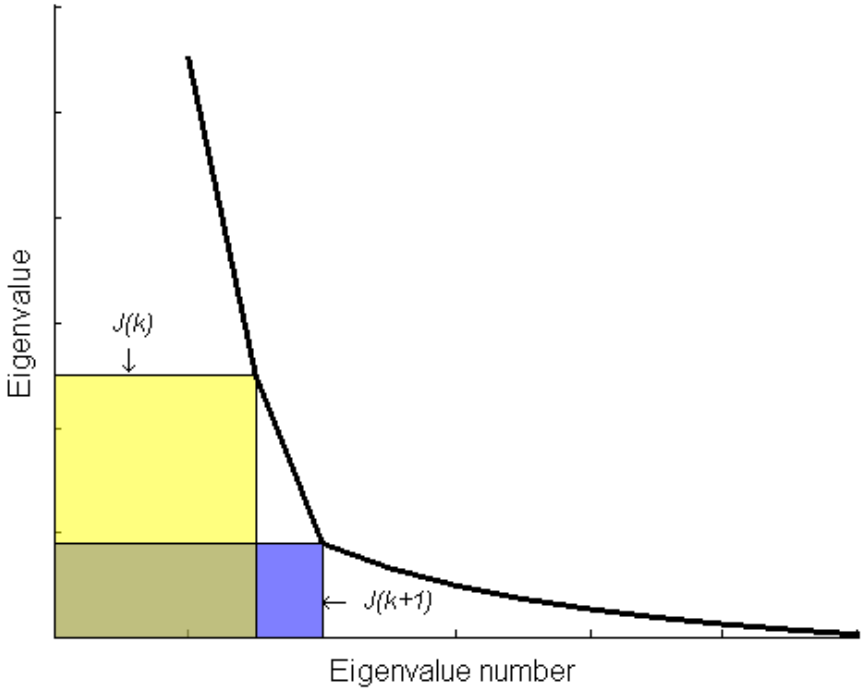
# 1 Introduction

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 *et al.*, (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) or Stock and Watson (2011). The determination of the number of factors 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. Connor and Korajczyk (1993), Bai and Ng (2002), Onatski (2009), Onatski (2010), Ahn and Horenstein (2013), Harding (2013) and Caner and Han (2014) develop estimation methods for the number of factors in static factor models. Recent examples for the number of dynamic factors are Amengual and Watson (2007), Hallin and Liška (2007), Bai and Ng (2007), Jacobs and Otter (2008), Kapetanios (2010) and Breitung and Pigorsch (2013).

Figure 1: Graphical illustration of our criterion in a scree plot

Find the value of $k$ for which the difference between adjacent eigenvalue – eigenvalue number blocks $DJ(k) \equiv J(k) - J(k+1) \equiv k\lambda_k - (k+1)\lambda_{k+1}$ is maximized




We derive a criterion for the determination of the number of factors in approximate static factor models, that is strongly associated to the scree test. This is a graphical technique, which consists of plotting the eigenvalues $\lambda_k$ against its eigenvalue number, and deciding at which value of $k$ the slopes of the plotted points are 'steep' to the left of $k$ and 'not steep' to 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. Our criterion is based on the comparison of surfaces under the scree plot, as illustrated in Figure 1. We look for the value of $k$ for which the difference between the adjacent products of the

eigenvalue numbers times the corresponding eigenvalues, in other words the difference between adjacent eigenvalue-eigenvalue number blocks ($DJ(k) \equiv J(k) - J(k+1) \equiv k\lambda_k - (k+1)\lambda_{k+1}$) is maximized.

In simulation experiments we compare our criterion to a couple of other estimators based on eigenvalues and also associated to the scree test. The Edge Distribution (ED) estimator of Onatski (2010) is based on the fact that any finite number of the largest of the bounded eigenvalues of the sample covariance matrix cluster around a single point. His estimator consistently separates the diverging eigenvalues from the cluster and counts the number of the separated eigenvalues, which is his estimate of the number of factors. Ahn and Horenstein (2013) propose the Eigenvalue Ratio (ER) and the Growth Ratio (GR) estimators. The ER estimator is obtained by maximizing the ratio of two adjacent eigenvalues arranged in descending order

$$\frac{\lambda_k}{\lambda_{k+1}} = \frac{V\left(k-1\right) - V\left(k\right)}{V\left(k\right) - V\left(k+1\right)},$$

while the GR estimator maximizes

$$\frac{\ln\left(V\left(k-1\right)\right) - \ln\left(V\left(k\right)\right)}{\ln\left(V\left(k\right)\right) - \ln\left(V\left(k+1\right)\right)} = \frac{\ln\left(1+\lambda_k^*\right)}{\ln\left(1+\lambda_{k+1}^*\right)},$$

where $V(k) = \sum_{j=k+1}^{m} \lambda_j$, $\lambda_k^* \equiv \frac{\lambda_k}{\sum_{j=k+1}^{m} \lambda_j}$ and $m = \min(n, T)$ for the number of variables $n$ and the number of observations $T$. Our criterion outperforms the two eigenvalue test ratios of Ahn and Horenstein (2013) for all sample sizes, except for their base scenario. It also outperforms the ED estimator of Onatski (2010) for samples up to 300 variables/observations. This conclusion

is robust for variation in the signal to noise ratio and situations where "weak" factors are present, which may have a huge impact (Onatski 2012).

The rest of the paper is structured as follows. Section 2 derives our criterion and shows its consistency using the set-up of Onatski (2010). Section 3 presents Monte Carlo simulation experiments. Section 4 concludes.

# 2 Method

## Our criterion

To introduce our method we consider the basis specification of Onatski (2010), which is discussed below, in a slightly different notation. Consider the factor model

$$\boldsymbol{x}_t = \boldsymbol{B}\boldsymbol{f}_t + \boldsymbol{\varepsilon}_t, \tag{1}$$

where $\boldsymbol{x}_t$ is an $n$-dimensional vector of stochastic normalized variables, i.e. $\mathrm{E}\{x_{it}\} = 0$ and $\mathrm{var}\{x_{it}\} = 1$, for $i = 1, \ldots, n$ for all $t$; $\boldsymbol{B} = (b_1 \ldots b_k)$ with $b_j \in \mathbb{R}^n$, is a matrix of factor loadings; $\boldsymbol{f}_t \in \mathbb{R}^k$ a vector of factors independently distributed from the idiosyncratic component vector $\boldsymbol{\varepsilon}_t$, which is further specified below.

Let $\mathrm{E}\{\boldsymbol{x}_t\boldsymbol{x}_t'\} = \boldsymbol{V} = \boldsymbol{C}\boldsymbol{\Lambda}\boldsymbol{C}'$ with ordered eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$. Assuming $\mathrm{rank}(\boldsymbol{B}\boldsymbol{B}') = k$

$$\mathrm{tr}(\mathrm{E}\{\boldsymbol{x}_t\boldsymbol{x}_t'\}) = \sum_{j=1}^{n} \lambda_j = \sum_{j=1}^{k} \lambda_j + \sum_{j=k+1}^{n} \lambda_j = n, \tag{2}$$

where $\sum_{j=1}^{k} \lambda_j$ is the explained variance which can be rewritten as $\sum_{j=1}^{k} \lambda_j = k\lambda_k + \sum_{j=1}^{k}(\lambda_j - \lambda_k)$.

By defining $J(k) \equiv k\lambda_k$ to be the minimum explained variance we have $J(k+1) = J(k) - DJ(k)$ for $k = 1, 2, \ldots$, where $DJ(k)$ is the change in the minimum explained variance if the factor space is increased by one, with $DJ(k) \leq 0$ or $DJ(k) > 0$. Our criterion becomes now to maximize $DJ(k)$ as illustrated in Figure 1.

In the simulations in Section 3 below, we employ the DJS estimator which consists of the constrained optimization $\widehat{k} = \underset{k}{\operatorname{argmax}}\, \widehat{DJ}\,(k)$ with the constraint that $\widehat{DJ}\left(\widehat{k}-1\right) < 0$. Since this constraint does not affect the asymptotic properties of our criterion, we will investigate the large sample properties of $DJ(k)$ using the framework of Onatski (2010).

*Remark 1.* To measure the sensitivity of the criterion $DJ(k)$ with respect to a small change in the eigenvalues, the total derivative can be used

$$\partial DJ(k) = k\partial\lambda_k - (k+1)\partial\lambda_{k+1},$$

with $\partial\lambda_k$ and $\partial\lambda_{k+1}$ small positive changes. For Ahn and Horenstein (2013)'s eigenvalue ratio $ER(k) \equiv \lambda_k/\lambda_{k+1}$ we have

$$\partial ER(k) = \lambda_{k+1}^{-1}(\partial\lambda_k - ER(k)\partial\lambda_{k+1})$$

with is sensitive to small changes in the eigenvalues in case of $\lambda_{k+1} << 1$, whereas our criterion is stable.

## Asymptotics

Onatski (2010) considers the approximate factor model

$$\boldsymbol{X}^{(n,T)} = \boldsymbol{B}^{(n,T)}\boldsymbol{F}^{(n,T)} + \boldsymbol{e}^{(n,T)}, \tag{3}$$

where $\boldsymbol{X}^{(n,T)}$ is an $n \times T$ matrix of data on $n$ cross-sectional units observed over $T$ time periods, $\boldsymbol{B}^{(n,T)}$ is an $n \times r$ matrix whose $(i,j)$th element is interpreted as the loading of the $j$th factor on the $i$th cross-sectional unit, $\boldsymbol{F}^{(n,T)}$ is an $r \times T$ matrix whose $(j,t)$th element is interpreted as the value of the $j$th factor at time $t$, and $\boldsymbol{e}^{(n,T)} = \boldsymbol{A\varepsilon G}$ is an $n \times T$ matrix of the idiosyncratic components of the data, the $n \times n$ matrix $\boldsymbol{A}$ and the $T \times T$ matrix $\boldsymbol{G}$ are two largely unrestricted deterministic matrices, and $\boldsymbol{\varepsilon}$ is an $n \times T$ matrix with i.i.d. Gaussian entries, so that both cross-sectional and temporal correlation of the idiosyncratic terms is allowed.

Let the ordered eigenvalues of the sample matrix $\left(\boldsymbol{X}^{(n,T)\prime}\boldsymbol{X}^{(n,T)}\right)/T(n)$ be $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_m$ with $m = \min(n,T)$ and $\sum_{i=1}^{m}\hat{\lambda}_i = n$. We assume that Assumption 1 and 2, Lemma 1–3, and Theorem 1 of Onatski (2010) hold.

Let $\{T(n),\ n \in \mathbb{N}\}$ be a sequence of positive integers such that $n/T(n) \to c > 0$ as $n \to \infty$. Let $k_{max}/n \to 0$ as $n \to \infty$ with the maximum possible number of factors $k_{max}$ assumed a priori given sample size $n$, $T(n)$. Let $\widehat{DJ}(k) = k\left(\widehat{\lambda}_k - \widehat{\lambda}_{k+1}\right) - \widehat{\lambda}_{k+1}$.

Onatski (2010, p1007) writes that his Theorem 1 suggests a way to estimate the true number of factors $r$. For $k(n) > r$ and large enough $n$, $\hat{\lambda}_{k+1}$ is finite with probability one as $n \to \infty$. For any $k > r$ the difference $\left(\hat{\lambda}_k - \hat{\lambda}_{k+1}\right)$

converges to zero with probability one, while the difference $\left(\hat{\lambda}_r - \hat{\lambda}_{r+1}\right)$ diverges to infinity. It follows that the estimator $\widehat{DJ}(k) = k\widehat{\lambda}_k - (k+1)\widehat{\lambda}_{k+1}$ converges as $n, T(n) \to \infty$ for $k > r$ while it diverges to infinity for $k = r$. Hence, our estimator $\widehat{DJ}(k)$ is consistent.

*Remark 2.* By considering $\widehat{DJ}(k)/k$ and threshold $\hat{\lambda}_{k+1}/k$, our estimator fits in the family of estimators of Onatski (2010, Equation (10))

$$\hat{r}(\hat{\delta}) = \max\left\{k \leq k_{max} : \ \hat{\lambda}_k - \hat{\lambda}_{k+1} \geq \hat{\delta}\right\}, \tag{4}$$

with $\hat{\delta} = \frac{1}{k}\hat{\lambda}_{k+1}$. So, whereas Onatski has to estimate the threshold in an ad-hoc way, our threshold is simply $\hat{\lambda}_{k+1}/k$.

*Remark 3.* Onatski's family of estimators in Equation (4) is consistent even for weak factors, which are defined as factors whose explanatory power for response variables grows slower than the rate of $n$. If the $k$-th factor is weak, then $\text{plim}_{m\to\infty}\hat{\lambda}_k = 0$, but $\text{plim}_{m\to\infty} m\hat{\lambda}_k = \infty$.

*Remark 4.* The eigenvalue ratio estimators of Ahn and Horenstein (2013) require more strict assumptions to prevent the denominator of the ratios from becoming equal to zero.

# 3 Monte Carlo experiments

We compare finite-sample simulations of our DJS estimator with the two alternatives proposed by Ahn and Horenstein (2013), the eigenvalue ratios ER and GR, and the ED estimator proposed by Onatski (2010). For all the

estimators considered, the argument search is performed over $k = 1, \ldots, k_{max}$ with $k_{max} = 8$.

We employ the data generating process as specified in Ahn and Horenstein (2013), which is also used by Bai and Ng (2002) and Onatski (2010). The foundation of the simulation exercise is the following approximate factor model:

$$x_{it} = \sum_{j=1}^{r} b_{ij} f_{jt} + \sqrt{\theta} u_{it}; \qquad u_{it} = \sqrt{\frac{1-\rho^2}{1+2J\beta^2}} e_{it}, \tag{5}$$

where $e_{it} = \rho e_{i,t-1} + (1-\beta)\,\nu_{it} + \beta \sum_{h=\max(i-J,1)}^{\min(i+J,n)} \nu_{ht}$ and the $\nu_{ht}$ and $b_{ij}$ are all drawn from $N(0,1)$. The idiosyncratic components $u_{it}$ are normalized such that their variances are equal to one for most of the cross-section units $J$.[1] The control parameter $\theta$ is the inverse of the signal to noise ratio (SNR) for the individual factors because $\text{var}(f_{jt}) / \text{var}\left(\sqrt{\theta} u_{it}\right) = 1/\theta$. When it is necessary to change the SNRs of all factors, we adjust parameter $\theta$. However, we also simulate a single weak factor by drawing from $N(0, \theta_{wf})$ for the weak factor and from $N(0,1)$ for the other factors, so $\theta_{wf}$ represents the relative dominance (or weakness) of the single factor. The magnitude of the time series correlation in the idiosyncratic component is controlled by parameter $\rho$. Note that Equation (5) describes an approximate static factor model, so no autocorrelation for the factors is assumed. Parameter $\beta$ governs the magnitude of cross-sectional correlation. We will focus on the specification with both serially and cross-sectionally correlated errors, $\rho = 0.5$, $\beta = 0.2$, $J = \max(10, n/20)$. Despite the fact that the means of the factors, the factor loadings and the idiosyncratic component are all zero in the data generating

[1] More specifically for units $J + 1 \leq i \leq n - j$.

process (5), we use double demeaned data, i.e. $x_{it} - T^{-1}\sum_t x_{it} - n^{-1}\sum_i x_{it} + (nT)^{-1}\sum_{i,t} x_{it}$, in order to avoid the one-factor bias problem as identified by Brown (1989).[2]

**Base scenario**

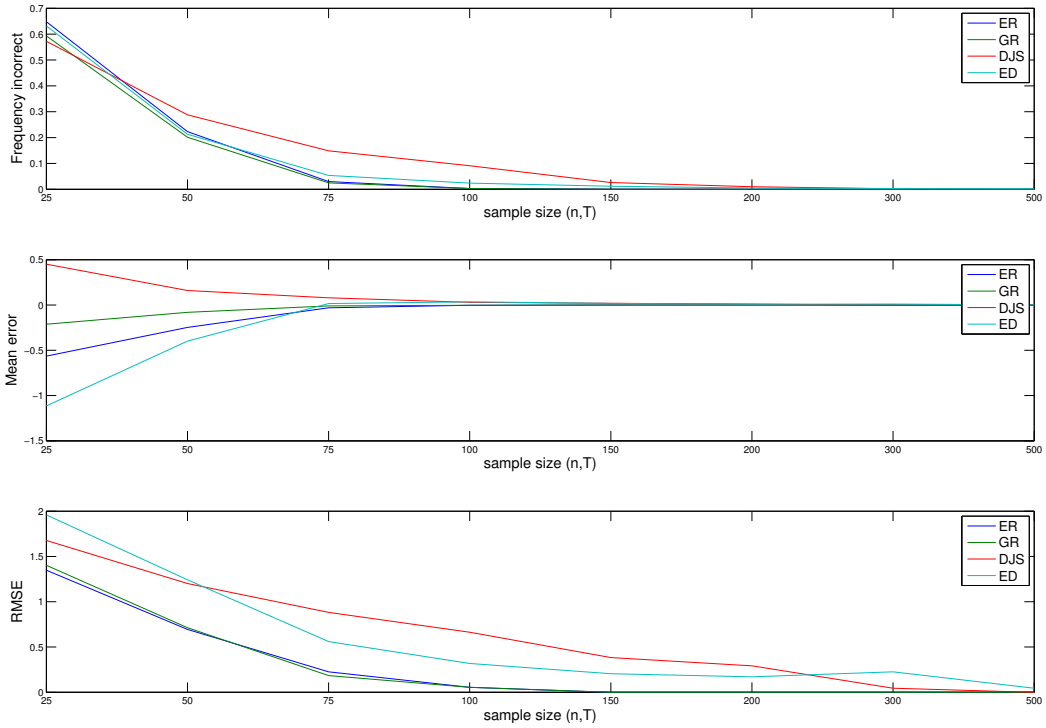
We focus on the model with $r = 3$ factors and configurations of the sample size over the grid $(n, T) = 25, 50, 75, 100, 150, 200, 300, 500$, inverse signal to noise parameter $\theta$ and the relative weakness of one of the three factors $\theta_{wf}$. Based on 1000 simulations for each configuration, we compute for each of the four different estimators DJS, ER, GR and ED the estimated number of factors $\widehat{k}$, i.e. the mode, and three performance statistics, the mean error, the root mean squared error (RMSE) and the frequency of incorrect estimated number of factors. To illustrate the measures, suppose 1000 simulations produce 700 correct outcomes of $\hat{k} = 3$, 200 outcomes of $\hat{k} = 2$ and 100 outcomes of $\hat{k} = 4$, the latter two both incorrect. Then the mean error equals 0.1, the RMSE $\sqrt{(}0.3)$, and the frequency of incorrect estimated number of factors is 0.3, which hence does not pass a 10% threshold level.

In the simulations based on Ahn and Horenstein's (2013) baseline specification consisting of a three-factor model with $\theta_{wf} = \theta = 1$, all estimators produce a mode equal to three factors. Figure 2 shows the results for the performance statistics. The figure shows that our proposed estimator DJS compares reasonably well with the other ones for this base scenario, but performs less well than the others. The figure also shows that the AH estimators

[2]We employ double demeaned data for the estimators of Ahn and Horenstein (2013) ER, GR and our proposed estimator DJS, but not for Onatski's (2010) ED estimator.

come out well in this base scenario with $\theta = \theta_{wf} = 1$. Below we will see that this outcome is not robust.

Figure 2: Performance of different estimators




Note. Simulations are based on $\theta = \theta_{wf} = 1$.

## Robustness

To check the robustness of the performance of the different estimators, we extend the grid for the inverse signal to noise parameter $\theta$ to 18 points: $\theta = [\frac{1}{2}, \frac{3}{4}, 1, \frac{5}{4}, \frac{6}{4}, \frac{7}{4}, 2, \frac{9}{4}, \frac{10}{4}, \frac{11}{4}, 3, \frac{13}{4}, \frac{14}{4}, \frac{15}{4}, 4, 5, 7, 12]$. In addition, we extend the grid for the weak factor parameter $\theta_{wf}$ to 15 points: $\theta_{wf} = \left[\frac{3}{32}, \frac{1}{8}, \frac{3}{16}, \frac{1}{4}, \frac{3}{8}, \frac{1}{2}, \frac{3}{4}, 1, \frac{3}{2}, 2, 3, 4, 6, 8, 12\right]$. The combined grid consists then of $18 \times 15 = 270$ different configurations, each consisting of 1000 simulations. The perfor-

mance of the different estimators for each configuration is summarized by the estimated number of factors, i.e. the mode of the simulations, and the frequency of the incorrect estimated number of factors. For each of the 270 configurations these performance statistics are further summarized by the percentages in Table 1, representing the fraction of the different configurations for which the mode consists of the true number of factors, i.e. $\widehat{k} = 3$. The percentages presented in Table 2 show the fraction of the different configurations for which the frequency of incorrect estimated number of factors is larger than 10%. To keep the tables and outcomes readable, we only report outcomes for sample sizes where the number of variables $n$ is equal to the number of observations $T$, as in Ahn and Horenstein (2013).

The first block in the two tables summarize the results of all 270 different configurations. The estimators DJS and ED clearly outperform the other alternatives for large sample sizes. Since this is not apparent in the standard configuration $\theta = \theta_{wf} = 1$ shown in Figure 2, the outperformance must by construction be due to variation in $\theta$ and $\theta_{wf}$.

Table 1: Correct number of factors according to the different estimators

| $(n,T)$ | 25 | 50 | 75 | 100 | 150 | 200 | 300 | 500 |
|---|---|---|---|---|---|---|---|---|
| | $\forall\theta,\forall\theta_{wf}$, grid = 270 | | | | | | | |
| ER | 7 | 13 | 17 | 20 | 20 | 21 | 31 | 42 |
| GR | 5 | 10 | 14 | 16 | 17 | 19 | 28 | 38 |
| DJS | 27 | 33 | 37 | 40 | 37 | 39 | 50 | 56 |
| ED | 11 | 19 | 24 | 29 | 29 | 33 | 55 | 70 |
| | $\forall\theta,\theta_{wf}=1$, grid = 18 | | | | | | | |
| ER | 22 | 44 | 50 | 61 | 67 | 72 | 89 | 94 |
| GR | 17 | 39 | 50 | 61 | 67 | 72 | 89 | 94 |
| DJS | 39 | 44 | 50 | 61 | 56 | 61 | 83 | 89 |
| ED | 17 | 39 | 44 | 56 | 56 | 61 | 89 | 94 |
| | $\forall\theta,\theta_{wf}=[4,6,8,12]$, grid = 72 | | | | | | | |
| ER | 3 | 6 | 10 | 11 | 11 | 13 | 19 | 33 |
| GR | 0 | 1 | 4 | 4 | 4 | 4 | 10 | 21 |
| DJS | 51 | 60 | 68 | 71 | 68 | 71 | 85 | 94 |
| ED | 17 | 26 | 33 | 39 | 40 | 46 | 78 | 94 |
| | $\forall\theta_{wf},\theta=1$, grid = 15 | | | | | | | |
| ER | 27 | 33 | 47 | 47 | 47 | 47 | 60 | 73 |
| GR | 27 | 27 | 40 | 40 | 40 | 40 | 47 | 67 |
| DJS | 60 | 60 | 60 | 60 | 60 | 60 | 60 | 60 |
| ED | 60 | 60 | 67 | 67 | 67 | 73 | 80 | 87 |
| | $\forall\theta_{wf},\forall\theta>1$, grid = 225 | | | | | | | |
| ER | 1 | 6 | 9 | 12 | 12 | 13 | 24 | 36 |
| GR | 0 | 4 | 8 | 10 | 11 | 12 | 22 | 32 |
| DJS | 22 | 28 | 33 | 36 | 33 | 35 | 48 | 55 |
| ED | 0 | 9 | 15 | 20 | 20 | 24 | 49 | 65 |

**Notes**.
The results are based on the mode of 1000 Monte Carlo replications. The presented percentage is the fraction of the grid for which the mode equals the true number of factors $r = 3$ according to the different estimators. In case $\forall\theta,\forall\theta_{wf}$, the grid consists of $18\times15 = 270$ different simulations. The grid sizes for each case are reported on the first line.

Table 2: Frequency of incorrect number of factors according to the different estimators

| $(n,T)$ | 25 | 50 | 75 | 100 | 150 | 200 | 300 | 500 |
|---|---|---|---|---|---|---|---|---|
| | $\forall\theta,\forall\theta_{wf},\text{grid}=270$ | | | | | | | |
| ER | 100 | 97 | 93 | 90 | 89 | 87 | 77 | 62 |
| GR | 100 | 98 | 96 | 93 | 91 | 90 | 80 | 67 |
| DJS | 100 | 100 | 100 | 98 | 86 | 74 | 56 | 47 |
| ED | 100 | 96 | 89 | 83 | 81 | 77 | 56 | 35 |
| | $\forall\theta,\theta_{wf}=1,\text{grid}=18$ | | | | | | | |
| ER | 100 | 89 | 78 | 72 | 67 | 61 | 28 | 11 |
| GR | 100 | 89 | 83 | 72 | 67 | 61 | 28 | 11 |
| DJS | 100 | 100 | 100 | 72 | 67 | 61 | 28 | 11 |
| ED | 100 | 94 | 78 | 72 | 61 | 56 | 17 | 6 |
| | $\forall\theta,\theta_{wf}=[4,6,8,12],\text{grid}=72$ | | | | | | | |
| ER | 100 | 99 | 97 | 94 | 94 | 92 | 86 | 74 |
| GR | 100 | 100 | 100 | 99 | 99 | 99 | 93 | 83 |
| DJS | 100 | 100 | 100 | 100 | 83 | 53 | 21 | 11 |
| ED | 100 | 92 | 83 | 75 | 72 | 67 | 36 | 11 |
| | $\forall\theta_{wf},\theta=1,\text{grid}=15$ | | | | | | | |
| ER | 100 | 100 | 73 | 67 | 67 | 67 | 47 | 27 |
| GR | 100 | 100 | 87 | 73 | 73 | 73 | 53 | 40 |
| DJS | 100 | 100 | 100 | 93 | 60 | 40 | 40 | 40 |
| ED | 100 | 100 | 40 | 40 | 40 | 33 | 27 | 13 |
| | $\forall\theta_{wf},\forall\theta>1,\text{grid}=225$ | | | | | | | |
| ER | 100 | 100 | 99 | 97 | 96 | 95 | 85 | 70 |
| GR | 100 | 100 | 100 | 98 | 97 | 96 | 87 | 73 |
| DJS | 100 | 100 | 100 | 99 | 90 | 81 | 59 | 48 |
| ED | 100 | 100 | 99 | 93 | 91 | 87 | 63 | 40 |

**Notes**.
The results are based on the frequency of incorrect estimated number of factors for 1000 Monte Carlo replications. The presented percentage is the fraction of the grid for which the frequency of incorrectly estimated number of factors is larger than 10%. In the case $\forall\theta,\forall\theta_{wf}$, the grid consists of $18\times 15=270$ different simulations. Grid sizes are denoted by 'grid'.

The second block in the two tables only considers the configurations without weak factors, i.e. $\theta_{wf} = 1$, and summarizes the results based on the 18 different configurations for $\theta$. In the absence of weak factors no obvious differences exist in the performance of the proposed estimators, except perhaps for small sample sizes. All the estimators are equally robust to variations in the signal to noise ratio.

The third panel in Table 1 and Table 2, which summarizes the results over the grid $\theta_{wf} = [4, 6, 8, 12]$, clearly shows that the estimators ER and GR are not robust to the presence of weak factors. Moreover, the outcomes reveal that for smaller sample sizes, DJS more often correctly estimates the true number of factors than ED, while ED shows a lower frequency of incorrect estimated number of factors.

The fourth panel in Table 1 and Table 2 considers the baseline case for the signal to noise ratio, i.e. $\theta = 1$, and summarizes the results based on the 15 different weak factor configurations $\theta_{wf}$. In these configurations, the ER and GR estimators perform quite well comparably to the alternatives, especially at the larger sample sizes. However, the robustness for these two estimators breaks down in case of a weak overall factor structure, in which the idiosyncratic component explains a larger part of the variability than the common factors together, i.e. in case the inverse signal to noise ratio $\theta > 1$. The fifth panel in Table 1 and Table 2 considers configurations of an overall weak factor structure. The results confirm the better performance of the ED and DJS estimators.

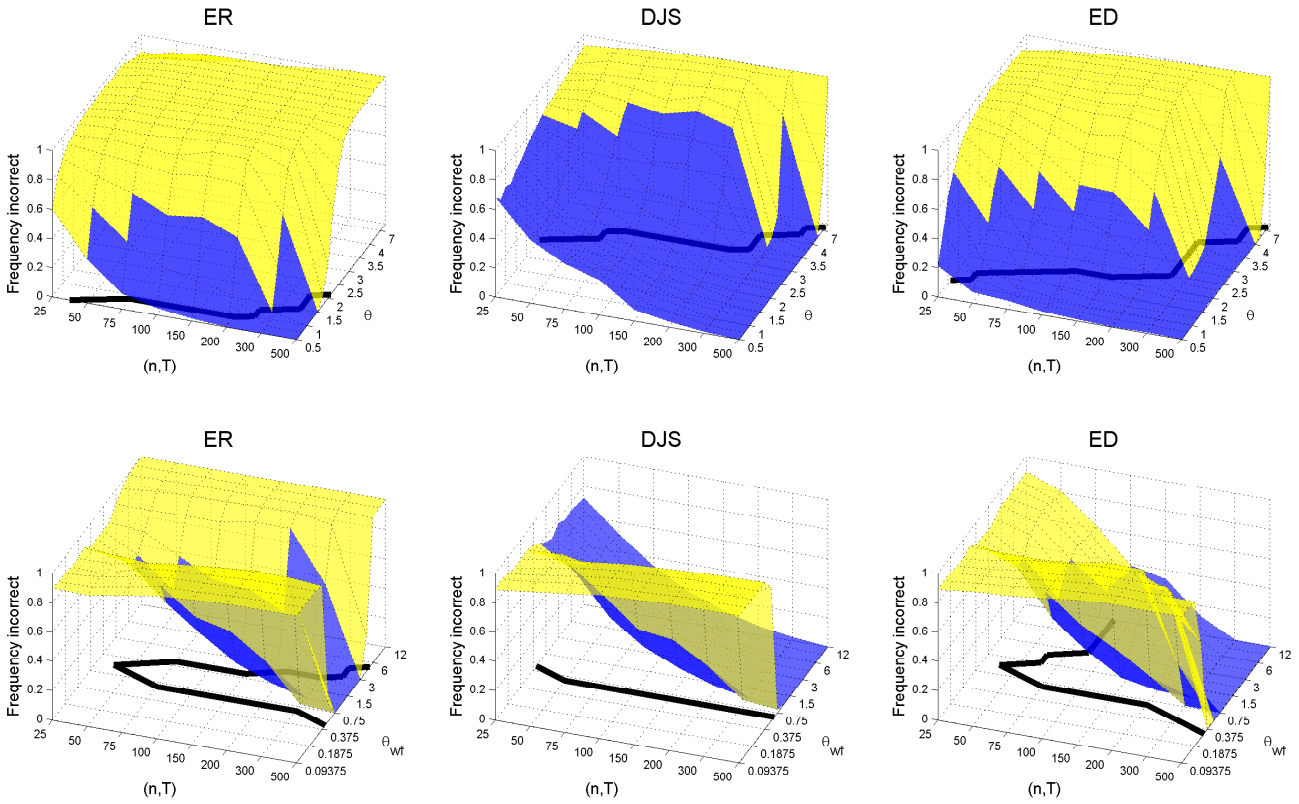
**Impact of weak factors**

The top three graphs in Figure 3 further illustrate the impact of the introduction of a weak factor. We show the frequency of incorrect number of estimated factors for different values of the inverse signal to noise ratio $\theta$ for the estimators ER, DJS and ED for $\theta_{wf} = 4$. While the upper panel of Figure 2 plots this statistic for $\theta_{wf} = \theta = 1$, the graphs in the first row of Figure 3 plot this statistic for all values of $\theta$ and $\theta_{wf} = 4$. The mode of the estimated number of factors being correct, i.e. $\widehat{k} = 3$, is represented by the blue colored surface, while the opposite is presented by the yellow colored surface. The demarcation between the yellow and blue surfaces projected on the $((n, T), \theta)$-plane is represented by the thick black line. The three graphs clearly show that in the weak factor case, our proposed DJS outperformes the ED estimator, while the ER estimator only performs well for high signal to noise ratios, i.e. low values of $\theta$.[3]

The three graphs in the bottom row of Figure 3 show the results for the specific case of $\theta = 2$. While the upper panel of Figure 2 plots the statistic for $\theta_{wf} = \theta = 1$, the second row of Figure 3 plots this statistic for all values of $\theta_{wf}$ and $\theta = 2$. The graphs clearly shows that the ER estimator performs poorly in case one factor is relatively very weak, i.e. $\theta_{wf} < \frac{3}{8}$, or relatively strong, i.e. $\theta_{wf} > 3$, even at large sample sizes. In contrast, the ED and DJS estimators converge for large sample sizes in case $\theta_{wf} > 0.75$, while the DJS estimator shows some outperformance for the correct number of factors even for small sample sizes.

[3]A similar conclusion holds for the GR estimator, not shown here.

Figure 3: Performance of estimators for different factor structures

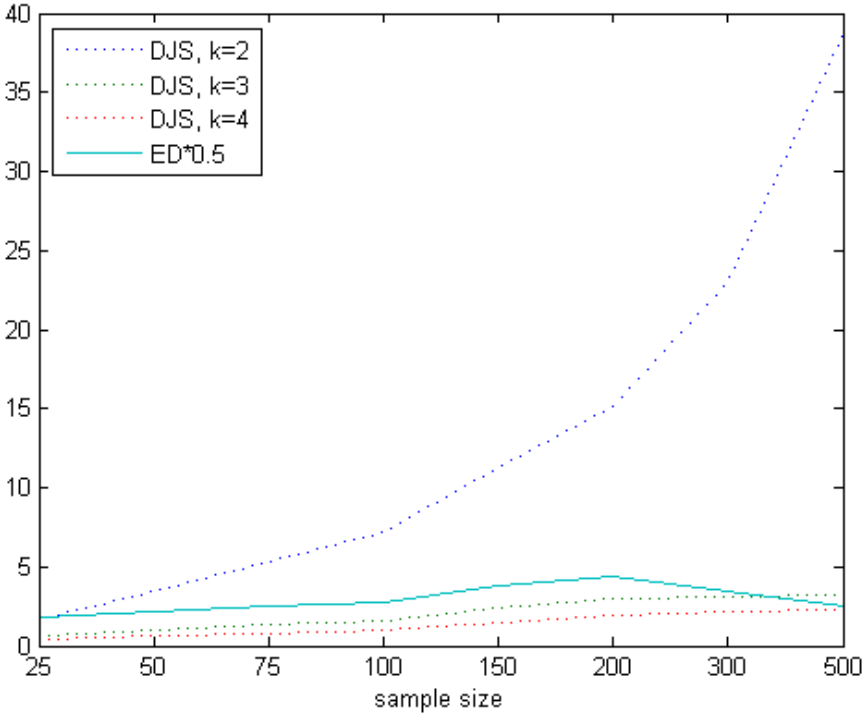



**Notes**.
The graphs in the top row show the frequency of incorrect estimated number of factors for all values of $\theta$ and $\theta_{wf} = 4$. The graphs in the bottom row show the frequency of incorrect estimated number of factors for $\theta = 2$. The blue colored surfaces show the parameter configurations where the mode of the estimated number of factors is correct, i.e. $r = 3$, while the yellow colored surfaces show the opposite case. The demarcation between the yellow and blue surfaces projected on the $((n, T), \theta)$-plane is represented by the thick black line.

### Comparison of ED and DJS: size of threshold

We have seen above in Section 2 that our $DJ(k)/k$ estimator belongs to the family of estimators of Onatski (2010) with threshold value $\hat{\delta}$ in Equation (4). Figure 4 shows the difference between the thresholds of the two estimators for different sample sizes for simulations with $r = 3$ factors and $\theta = \theta_{wf} = 1$.

Figure 4: Comparison of ED and DJS thresholds ($\theta = 1$ and $\theta_{wf} = 1$, and $r = 3$ factors)




The graph shows the ED threshold multiplied by .5, since Onatski determines the threshold with a regression and multiplies the outcome ad hoc by 2. Our DJS threshold is equal to $\hat{\lambda}_{k+1}/k$ and shown for $k = 2,\ 3$ and 4 where $\hat{\lambda}$s are the eigenvalues of $\boldsymbol{X}'\boldsymbol{X}/T$. The figure shows that the threshold $\hat{\delta}_3$ which corresponds to $k = 2$ diverges for large sample sizes as expected whereas $\hat{\delta}_4$

and $\hat{\delta}_5$ which correspond to $k = 3$ and $k = 4$ respectively, converge and are close to the ED threshold.

# 4 Conclusion

This paper presents a simple criterion to determine the number of factors in a data-rich environment, based on the comparison of surfaces under the scree plot. Our criterion is intuitive appealing and straightforward to implement. Our procedure is closely related to Onatski (2010). Monte Carlo simulations taking into account weak factors reveal that our criterion outperforms the two eigenvalue test ratios of Ahn and Horenstein (2013) for all sample sizes, and Onatski's (2010) edge distribution estimator, except for large samples.

# Acknowledgements

We would like to thank Paul Bekker, Kees Bouwman, Tom Wansbeek and Mark Watson for helpful suggestions. The paper has benefited from comments received following presentations at the 7th Netherlands Econometric Study Group Meeting, Groningen, June 2012, the European Meeting of the Econometric Society, Gothenburg, August 2013, and the 7th International Conference on Computational and Financial Econometrics (CFE 2013), London, December 2013. Views expressed are those of the individual authors and do not necessarily reflect official positions of Sveriges Riksbank.

# References


Ahn, S.C. and A.R. Horenstein (2013), "Eigenvalue ratio test for the number of factors", *Econometrica*, **81**, 1203–1227.

Amengual, D. and M.W. 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 & Economic Statistics*, **25**, 52–60.

Bai, J. and S. Ng (2008), "Large dimensional factor analysis", *Foundations and Trends in Econometrics*, **3**, 89–163.

Breitung, J. and U. Pigorsch (2013), "A canonical correlation approach for selecting the number of dynamic factors", *Oxford Bulletin of Economics and Statistics*, **75**, 23–36.

Brown, S.J. (1989), "The number of factors in security returns", *The Journal of Finance*, **44**, 1247–1262.

Caner, Mehmet and Xu Han (2014), "Selecting the correct number of factors in approximate factor models: The large panel case with group bridge estimators", *Journal of Business & Economic Statistics*, **32**, 359–374.

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.

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. (2013), "Estimating the number of factors in large 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.

Onatski, A. (2010), "Determining the number of factors from empirical distribution of eigenvalues", *The Review of Economics and Statistics*, **92**, 1004–1016.

Onatski, A. (2012), "Asymptotics of the principal components estimator of large factor models with weakly influential factors", *Journal of Econometrics*, **168**, 244–258.

Peres-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 (2011), "Dynamic factor models", in M.P. Clements and D.F. Hendry, editors, *Oxford Handbook of Forecasting*, Oxford University Press, Oxford.