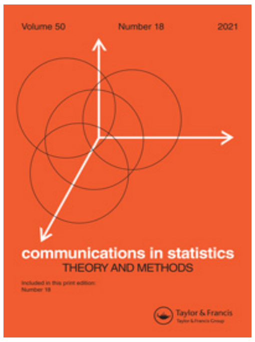

Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Theory and Methods

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/lsta20

A criterion for the number of factors

Ard H. J. den Reijer, Jan P. A. M. Jacobs & Pieter W. Otter

To cite this article: Ard H. J. den Reijer, Jan P. A. M. Jacobs & Pieter W. Otter (2021) A criterion for the number of factors, Communications in Statistics - Theory and Methods, 50:18, 4293-4299, DOI: 10.1080/03610926.2020.1713376

To link to this article: https://doi.org/10.1080/03610926.2020.1713376

Published online: 06 Mar 2020.

COMMUNICATIONS IN STATISTICS—THEORY AND METHODS
2021, VOL. 50, NO. 18, 4293–4299
https://doi.org/10.1080/03610926.2020.1713376

OPEN ACCESS

A criterion for the number of factors*

Ard H. J. den Reijer[a], Jan P. A. M. Jacobs[b,c,d,e], and Pieter W. Otter[b]

[a]Monetary Policy Department, Sveriges Riksbank, Stockholm, Sweden; [b]Faculty of Economics and Business, University of Groningen, Groningen, The Netherlands; [c]University of Tasmania, Hobart, Tasmania, Australia; [d]CAMA, Canberra, Australia; [e]CIRANO, Montréal, Canada

ABSTRACT

This note proposes a new criterion for the determination of the number of factors in an approximate static factor model. The criterion is strongly associated with the scree test and compares the differences between consecutive eigenvalues to a threshold. The size of the threshold is derived from a hyperbola and depends only on the sample size and the number of factors k. Monte Carlo simulations compare its properties with well-established estimators from the literature. Our criterion shows similar results as the standard implementations of these estimators, but is not prone to a lack of robustness against a too large *a priori* determined maximum number of factors $k_{\max}$.

ARTICLE HISTORY
Received 15 February 2019
Accepted 28 December 2019

KEYWORDS
Static factor model; number of factors; scree test

JEL-CODE
C32 C52 C82

1. Introduction

A wide range of methods has been proposed to determine the number of common factors for static approximate factor models concerning a data set with a large number of cross-section units (n) and time series observations (T). Bai and Ng (2002) propose to estimate the number of factors (r) by minimizing information criterion functions employing a penalty that depends on both n and T. Onatski (2010) develops data-dependent methods for a threshold value, which ideally should be slightly larger than the magnitude of the $(r+1)^{th}$ eigenvalue. Both methods require a pre-specified maximum possible number of factors. Ahn and Horenstein (2013) propose to look at ratios of eigenvalues thereby circumventing the need to specify a threshold.[1]

Similar to the latter two methods, our criterion for the determination of the number of factors is strongly associated with the scree test of Cattell (1966), which consists of plotting the eigenvalues λ_k of the scaled sample covariance matrix in descending order of magnitude against their corresponding ordinal eigenvalue numbers k, and deciding at which r they level off. The break between the 'steep' slope to the left of r and the leveling off to the right indicates an 'elbow' in the graph.

CONTACT Jan P. A. M. Jacobs j.p.a.m.jacobs@rug.nl Faculty of Economics and Business, University of Groningen, PO Box 800, 9700 AV Groningen, the Netherlands.

*The present version of this note has benefited from suggestions of anonymous referees, Paul Bekker, Kees Bouwman, Tom Wansbeek and Mark Watson, and from comments received following several conferences, workshops and seminars. Moreover, we thank Alexei Onatski for providing his simulation code. Views expressed are those of the individual authors and do not necessarily reflect official positions of Sveriges Riksbank.

Our proposed criterion is based on the comparison of surfaces under the scree plot. Like Onatski (2010), we look for the maximum k for which the difference between adjacent eigenvalues, *i.e.*, $\lambda_k - \lambda_{k+1}$ is larger than its corresponding threshold, *i.e.*, $\bar{\lambda}_{k+1}$. Based on a no-factor structure benchmark, the threshold $\bar{\lambda}_{k+1}$ is derived as the reciprocal function of $k+1$, horizontally scaled by an harmonic number. Hence, the corresponding benchmark scree plot $\{\bar{\lambda}_{k+1}, k+1\}$, for all k is an hyperbola, which does not show an 'elbow'. In accordance with Bai and Ng (2002), our proposed threshold $\bar{\lambda}_k$ is a function only of sample size n and T and thereby, unlike Onatski (2010), not data-dependent. Moreover, as our proposed threshold $\bar{\lambda}_{k+1}$ varies with k, there is no need to pre-specify a maximum number of factors $k_{\max}$.

The rest of the note is structured as follows. Section 2 derives our criterion as an application of Onatski (2010). Section 3 compares our criterion with the ones of Bai and Ng (2002), Onatski (2010) and Ahn and Horenstein (2013) in a Monte Carlo simulation. Section 4 concludes.

2. Method

Let the approximate factor model with the number of unobserved factors r be given by

$$\boldsymbol{X} = \boldsymbol{\Lambda}\boldsymbol{F}' + \boldsymbol{\xi} \tag{1}$$

where $\boldsymbol{X}$ is an $n \times T$ matrix with observations, $\boldsymbol{\xi}$ an $n \times T$ matrix with idiosyncratic components. The common components are determined by the matrix of factor loadings $\boldsymbol{\Lambda}$ and the matrix of factors $\boldsymbol{F}$ with rank r. The scaled sample covariance matrix $\boldsymbol{XX}'/nT$ has eigenvalues in descending order of their magnitude $\lambda_1 \geq \ldots \geq \lambda_n$.[2]

Let $\sum_{j=1}^{k} \lambda_j$ be the cumulative explanatory power of the first k factors, which can be rewritten as $\sum_{j=1}^{k} \lambda_j = k\lambda_k + \sum_{j=1}^{k}(\lambda_j - \lambda_k)$. Define $J(k) \equiv k\lambda_k$, which can be interpreted as the minimum possible explanatory power of the k factors. Define the no-factor structure benchmark as the condition that $J(k) = J(l), \forall k, l$. For the corresponding eigenvalues $\bar{\lambda}_k$, it then holds that $\bar{\lambda}_1 = k\bar{\lambda}_k$. Moreover, the unity sum of scaled eigenvalues $1 = \sum_{j=1}^{n} \bar{\lambda}_j = \bar{\lambda}_1 H_n$, with harmonic number $H_n = \sum_{j=1}^{n} \frac{1}{j}$ enables to quantify $\bar{\lambda}_k = \frac{1}{kH_n}$.

Figure 1 shows the hyperbola $\bar{\lambda}$ together with the empirical scree plot λ obtained from a simulated factor-model with $r=3$. Decomposing $\lambda_k = \bar{\lambda}_k + \delta_k$, the figure shows that the first r diverging eigenvalues explain by assumption more than their no-factor benchmark equivalents, *i.e.*, $\delta_k \geq 0$ for $k \leq r$. As by definition $\sum_{j=1}^{r} \delta_j = -\sum_{j=r+1}^{n} \delta_j$, the empirical scree plot λ must cross the hyperbola $\bar{\lambda}$. As it holds that $\lambda_r - \lambda_{r+1} = \delta_r - \delta_{r+1} + w(r)$, a lower bound $w(r) = \frac{1}{r(r+1)H_n}$ can be obtained for the empirical scree plot between the points of crossing, *i.e.*, between $k=r$ and $k = r+1$. However, we propose a tighter threshold as $r \times w(r) = \bar{\lambda}_{r+1}$, thereby requiring that the difference between λ_r and λ_{r+1} meets the cumulative minimum of the r preceding eigenvalues.

The approach fits within Onatski's (2010, Equation (10)) family of estimators:

$$\hat{r}(\hat{\alpha}(n\lambda), k_{\max}) = \max\{k \leq k_{\max} : \lambda_k - \lambda_{k+1} \geq \hat{\alpha}(n\lambda)\} \tag{2}$$

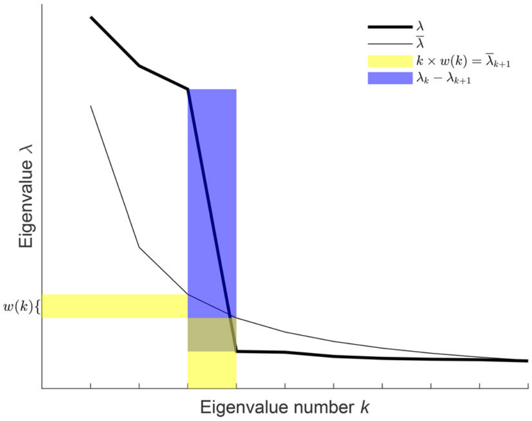

Figure 1. Graphical illustration of our criterion in a scree plot. Find the maximum k for which the difference between adjacent eigenvalues, *i.e.*, $\lambda_k - \lambda_{k+1}$ (blue plus yellow-blue) is larger than its corresponding threshold, *i.e.*, $\bar{\lambda}_{k+1}$ (yellow plus yellow-blue).

with constant $\hat{\alpha}(n\lambda)$ obtained by a regression involving $n\lambda$.[3] Onatski (2010, p1007) writes in his Theorem 1 that for $k > r$, $n\lambda_k$ is finite and that the difference $n(\lambda_k - \lambda_{k+1})$ converges to zero, while the difference $n(\lambda_r - \lambda_{r+1})$ diverges to infinity with probability one as $n, T \to \infty$.

We propose $\tilde{r}(\tilde{\alpha}(\bar{\lambda}, k))$, which deviates from $\hat{r}(\hat{\alpha}(n\lambda), k_{\max})$ in three ways: i) the varying threshold $\tilde{\alpha}(\bar{\lambda}, k) = \bar{\lambda}_{k+1} = \frac{1}{(k+1)H_n}$ is a function of the ordered eigenvalue number k that converges to zero for either $k, \min\{n, T\} \to \infty$, while Onatski's (2010) threshold is constant $\forall k$; ii) the threshold $\tilde{\alpha}(\bar{\lambda}, k)$ is a function of H_n and can thereby *a priori* be determined as a function of sample size $\{n, T\}$, while Onatski's (2010) threshold $\hat{\alpha}(n\lambda)$ is a function of the empirical λ and can thereby only be determined *a posteriori*; and iii) as $\lambda_k - \lambda_{k+1} \leq \lambda_k \leq \bar{\lambda}_k$, the varying threshold cannot be passed (apart from random error) for $k > r$. So, there is no need to specify a $k_{\max}$ parameter even though $\tilde{\alpha}(\bar{\lambda}, k) \to 0$ for $k \to \infty$.

3. Monte Carlo simulation

We compare finite-sample simulations of our proposed criterion with the estimators proposed by Bai and Ng (2002) (BN),[4] Onatski (2010) (ON) and the two alternatives proposed by Ahn and Horenstein (2013), the Eigenvalue Ratio (ER) and the Growth Ratio (GR). The ER estimator of k is obtained by maximizing the ratio of two adjacent eigenvalues arranged in descending order.

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

$$x_{it} = \sum_{j=1}^{r} b_{ij} f_{jt} + \sqrt{\theta} u_{it}; \quad u_{it} = \sqrt{\frac{1-\rho^2}{1+2J\beta^2}} e_{it} \tag{3}$$

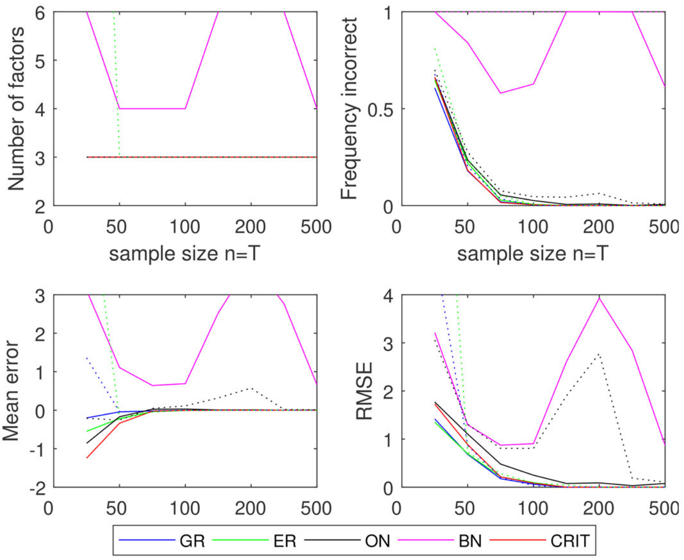

Figure 2. Performance of different estimators. *Note.* The different estimators consist of our proposed criterion (CRIT), Ahn and Horenstein's (2013) Eigenvalue Ratio (ER) and Growth Ratio (GR) and Onatski's (2010) estimator (ON) and Bai and Ng (2002)'s BIC3 estimator (BN). The number of factors is determined by an argument search up to a maximum of $k_{\max} = 8$ factors (straight lines), alternatively $k_{\max} = 20$ factors (dotted lines). Note that the dotted lines for BN lie outside the graph.

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 ν_{ht}, b_{ij} and f_{jt} 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.[5] The control parameter θ is the inverse of the signal to noise ratio (SNR) for the individual factors because $\text{var}(f_{jt})/\text{var}\left(\sqrt{\theta}u_{it}\right) = 1/\theta$. The magnitude of the time series correlation in the idiosyncratic component is controlled by parameter ρ. Note that Equation (3) describes an approximate static factor model and assumes no autocorrelation for the factors. Parameter β governs the magnitude of cross-sectional correlation and parameter J the number of correlated units. We will focus on the specification with $r = 3$ factors, $\theta = 1$ and 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 process (3), we use double demeaned data, i.e., $x_{it} - T^{-1}\sum x_{it} - n^{-1}\sum x_{it} + (nT)^{-1}\sum x_{it}$, in order to avoid the one-factor bias problem as identified by Brown (1989).[6]

3.1. Simulation results

Based on 1000 simulations for each of the sample sizes in the grid $n = T = 25, 50, 75, 100, 150, 200, 300, 500$,[7] we compute the estimated number of factors $\hat{k}$, 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 $\hat{k} =$

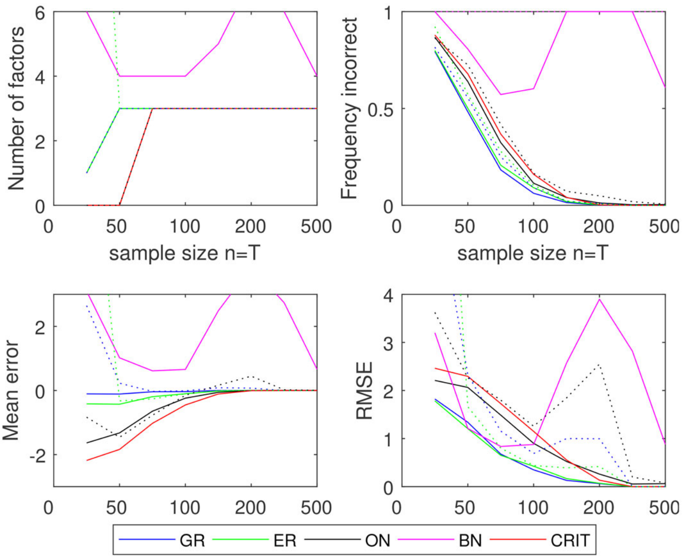

Figure 3. Performance of different estimators (cont.). *Note.* Similar to Figure 3 though for simulation with a lower signal to noise ratio of $\theta = 2$.

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 is the square root of 0.3 and the frequency of incorrect estimated number of factors is 0.3.

Figure 2 shows the performance statistics for the five estimators considered, where the argument search is performed over $k = 1, ..., k_{\max}$ with the standard specification of $k_{\max} = 8$. As a robustness check, the three dotted lines show the equivalent statistics for the case $k_{\max} = 20$. The figure shows that our proposed criterion compares well to the alternatives in the standard simulation. First, as documented by Ahn and Horenstein (2013) the BN alternative does not perform so well in case the idiosyncratic component exhibits cross-sectional correlation. Second, the other alternatives show not to be robust against the case $k_{\max} = 20$. Especially the ER and GR alternatives reveal small sample sensitivity. As ER and GR consist of fractions with eigenvalues in the denominator, both are sensitive to small random changes in case $\lambda_k \ll 1$, *i.e.*, for large $k_{\max}$. Onatski's (2010) estimator of the threshold $\hat{\alpha}(n\lambda)$ involves a regression on the empirical λ and hence, incorporates random instabilities in case of a large $k_{\max}$.

Figure 3 shows the results of the simulation with a lower signal to noise ratio of $\theta = 2$. For this edge case, all the estimators exhibit poor small-sample performance. For medium to large sample sizes, the performance of the different alternatives is more similar with exception of the BN-estimator. The ER and GR estimators with the argument search up to $k_{\max} = 8$ show some outperformance, but still exhibit a lack of robustness against this parameter.

Finally, Figure 4 shows the results of the simulation with a higher number of factors $r = 5$. Here again, the ER and GR estimators show some outperformance apart from the case with small samples and a high $k_{\max} = 20$. Note moreover that our proposed criterion shows a similar performance as compared to Onatski's (2010) estimator.

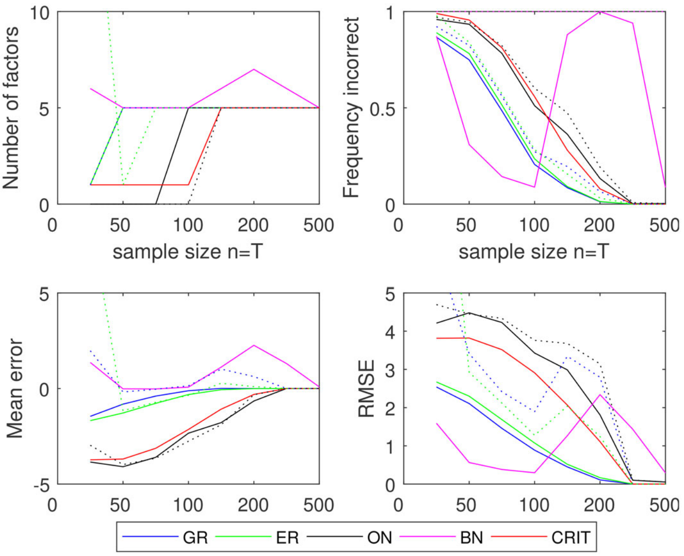

Figure 4. Performance of different estimators (cont.). *Note.* Similar to Figure 3 though for simulation with a higher number of factors $r = 5$.

As an empirical application, we employed the different estimators on the latest vintage of FRED-MD, see McCracken and Ng (2016). This large macroeconomic database is sampled at a monthly frequency, updated monthly using the Federal Reserve Data (FRED) database and thereby publicly accessible.[8] Based on this database consisting of $n = 128$ series with $T = 725$ months of observations, the estimated number of static factors vary between one for CRIT, two for ER and GR, five for ON and finally BN says eight, all estimated with $k_{\max} = 20$. The difference in results might be due to stochastics, *i.e.*, n is relatively small, while T relatively large, possibly a dynamic factor structure[9] or non linearities in the data.

4. Conclusion

This note presents a simple criterion to select the number of factors in an approximate static factor model, based on the comparison of surfaces under the scree plot. The criterion is an application of Onatski (2010), but with a varying threshold that is not data-dependent and only related to the sample size. In contrast to the alternatives, our proposed criterion does not require a pre-specified maximum number of factors $k_{\max}$.

Standard Monte Carlo simulations reveal a performance in line with the alternatives proposed by Onatski (2010) and the two alternatives of Ahn and Horenstein (2013). However, the alternatives show a lack of robustness against larger values of $k_{\max}$.

Notes

1. Recent contributions include Wu (2018) and Choi and Jeong (2019).
2. In case $n > T$, then $\lambda_i = 0$ for $i > T$. Without loss of generality, we assume $n \leq T$ for ease of notation.

3. Note that Onatski (2010) employs eigenvalues of the non scaled sample covariance matrix $\boldsymbol{XX}'/T$, *i.e.,* $n\lambda$ in our notation.
4. Like Ahn and Horenstein (2013), we only report the BIC_3 estimator being the best-performing one of the proposed estimators in this simulation set-up.
5. More specifically for units $J+1 \leq i \leq n-j$.
6. Ahn and Horenstein (2013) employ double demeaned data for ER and GR, while Onatski (2010) does not for ON. Our simulation results show no substantive performance differences between plain simulation data and double-demeaned simulation data for all five estimators.
7. For reasons of space, we take n equal to T in the simulations. Results in which n and T differ from each other lead to qualitatively similar conclusions and are available upon request.
8. See https://research.stlouisfed.org/econ/mccracken/fred-databases/.
9. However, the static factor representation of a dynamic factor model is possible in case the lenghts of the lags are finite.

References

Ahn, S. C., and A. R. Horenstein. 2013. Eigenvalue ratio test for the number of factors. *Econometrica* 81:1203–27. doi:10.3982/ECTA8968.

Bai, J., and S. Ng. 2002. Determining the number of factors in approximate factor models. *Econometrica* 70 (1):191–221. doi:10.1111/1468-0262.00273.

Brown, S. J. 1989. The number of factors in security returns. *The Journal of Finance* 44 (5):1247–62. doi:10.1111/j.1540-6261.1989.tb02652.x.

Cattell, R. B. 1966. The scree test for the number of factors. *Multivariate Behavioral Research* 1 (2):245–76. doi:10.1207/s15327906mbr0102_10.

Choi, I., and H. Jeong. 2019. Model selection for factor analysis: Some new criteria and performance comparisons. *Econometric Reviews* 38 (6):577–96. doi:10.1080/07474938.2017.1382763.

McCracken, M. W., and S. Ng. 2016. FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics* 34 (4):574–89. doi:10.1080/07350015.2015.1086655.

Onatski, A. 2010. Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics* 92 (4):1004–16. doi:10.1162/REST_a_00043.

Wu, J. 2018. Eigenvalue difference test for the number of common factors in the approximate factor models. *Economics Letters* 169:63–9. doi:10.1016/j.econlet.2018.05.009.