Computational Statistics and Data Analysis 112 (2017) 235–241

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

# Transformed contribution ratio test for the number of factors in static approximate factor models

Qiang Xia [a], Rubing Liang [a,*], Jianhong Wu [b]

[a] *College of Mathematics and Informatics, South China Agricultural University, Guangzhou 510642, China*
[b] *School of Mathematics and Science, Shanghai Normal University, Shanghai 200234, China*

ARTICLE INFO

*Article history:*
Received 25 November 2015
Received in revised form 7 January 2017
Accepted 2 March 2017
Available online 30 March 2017

*Keywords:*
Approximate factor model
Eigenvalues
Number of factors
Transformed contribution ratio

ABSTRACT

Determining the number of factors ($r$) is of importance in static approximate factor models. Under some mild conditions, the $r$ largest eigenvalues of the variance matrix of $N$ response variables go to infinity as $N$ increases, while the rest are bounded. Then, "Eigenvalue Ratio" (ER) and "Growth Ratio" (GR) estimators have been well exploited by maximizing the ratio of two adjacent eigenvalues in the literature. As a modification of ER and GR estimators, the new estimator named as "Transformed Contribution Ratio" (TCR) is obtained by maximizing the ratio of two adjacent transformed contribution of the eigenvalues. Under the same conditions of ER and GR estimators, the resulted estimator can be proved to be consistent. It can be further shown that, comparing with the competitors in the existing literature, the new method has desired performance on truly selecting the value of the number of latent common factors, especially when both strong and weak factors or some dominant factors are in static approximate factor models. Monte Carlo simulation experiments and one real data application are carried out for illustration.


## 1. Introduction

In the analysis of large dimensional factor models, one of fundamental issues is the consistent determination of the number of common factors. Recently, attempts in this direction have been focused on this issue in economic and financial data with both large cross-section dimension ($N$) and time series observations ($T$). Many examples include but not limited to Bai and Ng (2002), Stock and Watson (2002), Onatski (2006, 2010), Alessi et al. (2010), Fan et al. (2013) and Ahn and Horenstein (2013) for static approximate factor models of Chamberlain and Rothschild (1983). Pan and Yao (2008), Lam and Yao (2012) and Xia et al. (2015) focused on the determination of the number of factors for high-dimensional time series. Much of the other related econometric research, such as Forni et al. (2000), Hallin and Liska (2007), Amengual and Watson (2007), Bai and Ng (2007), and Onatski (2009), among others, were papers on dynamic factor models.

These estimators, however, are not directly comparable as they are based on their different sets of hypothesis. Meanwhile, under the same predetermined conditions, various methods can lead to different estimation results for the number of factors. In this paper, we will improve eigenvalue-based ratio-type method for the number of common factors suggested by Ahn and Horenstein (2013) in static approximate factor models, the improved method can be expected to have desired performance comparing with the existing methods in the literature. Up to our knowledge, Bai and Ng (2002) proposed to estimate the number of factors in static approximate factor models by minimizing the information criteria of model selection, named PC

* Corresponding author. Fax: +86 85285393.
*E-mail address:* stat2003@163.com (R. Liang).

http://dx.doi.org/10.1016/j.csda.2017.03.005
0167-9473/

and IC. But the criteria of Bai and Ng (2002) are known to overestimate in practice. Onatski (2006) obtained a consistent estimator of the $(r+1)$th largest eigenvalue and then the corresponding threshold value can be found easily only if it is slightly larger than the $(r+1)$th largest eigenvalue, where $r$ is the true number of common factors. However, the method is available only under the assumption that the idiosyncratic errors are either autocorrelated or cross-sectionally correlated, but not both (Ahn and Horenstein, 2013). Onatski (2010) considered the so-called "Edge Distribution" estimator by using the differenced eigenvalues. Alessi et al. (2010) introduced a tuning multiplicative constant in the penalty to improve the method of Bai and Ng (2002), an idea that was proposed by Hallin and Liska (2007) in the context of dynamic factor models. As reported by Ahn and Horenstein (2013), the methods proposed by Bai and Ng (2002) and Onatski (2010) have worse finite sample properties in the case with cross-sectional dependency although they do perform well in the case with independent idiosyncratic errors. Two eigenvalues-based ratio-type estimators of Ahn and Horenstein (2013) were proposed and shown to perform well even when the idiosyncratic errors are cross-sectionally dependent or serially correlated. Lam and Yao (2012) used the similar idea to deal with the factor modeling for high-dimensional time series based on the dimension-reduction viewpoint. As argued by Ahn and Horenstein (2013), almost all the above estimators need to predetermine the possible maximum of the number of factors, which may lead to overestimation or underestimation of the number when the maximum (hereafter, denoted by *kmax*) is too large or too small. But, the ER and GR estimators are not sensitive to *kmax* by the simulation experiments of Ahn and Horenstein (2013).

This paper proposes a determination approach by improving eigenvalue-based ratio-type method for the number of common factors in static approximate factor models. According to Ahn and Horenstein (2013), the ER estimator performs well under the cases of strong factors, while the GR estimator outperforms the ER estimator in the case when both strong and weak factors or some dominant factors are in static approximate factor models. The new method named as "Transformed Contribution Ratio" (hereafter TCR) is based on the ratio values of adjacent transformed contribution of eigenvalues arranged in descending order, and expected to have both advantages of the ER and GR estimators. Under some mild conditions, the resulted estimator can be proved to be consistent. It can be further shown that, comparing with the competitors in the existing literature, the new estimator has desired performance on truly selecting the value of the number of latent common factors, especially when both strong and weak factors or some dominant factors are in static approximate factor models. Monte Carlo simulation experiments and one real data application are carried out for illustration.

The rest of this paper is organized as follows. In Section 2, we introduce the new method of determining the number of common factors in the approximate factor models and state its asymptotic consistency. Section 3 carries out some simulation experiments to examine the finite sample performance of the new estimation. One real data analysis for illustration are given in Section 4. Section 5 provides some concluding comments. Technical details on the proofs of theorems are described in Appendix.

Throughout this paper, $A'$ means the transpose of the matrix $A$, $\|A\| = [trace(A'A)]^{1/2}$ denotes the norm of the matrix $A$, and $\psi_i(\Omega)$ denotes the $i$th largest eigenvalue of a positive semi-definite matrix $\Omega$.

## 2. Estimation of the number of factors

Consider the following static approximate factor model of Chamberlain and Rothschild (1983),

$$y_{\cdot t} = \Lambda F_t + u_{\cdot t}, \quad t = 1, 2, \ldots, T, \tag{1}$$

where $y_{\cdot t} = (y_{1t}, \ldots, y_{Nt})'$, $F_t$ is an $r$-dimensional vector of common factors, $\Lambda = (\lambda_1, \lambda_2, \ldots, \lambda_N)'$, $\lambda_i$ is an $r$-dimensional vector of factor loadings, and $u_{\cdot t} = (u_{1t}, \ldots, u_{Nt})'$ is the vector of the idiosyncratic error. The number of factors $(r)$ is unknown and needs to be estimated. For the sake of statement, we denote that $Y = (y_{\cdot 1}, \ldots, y_{\cdot T})'$ is the $T \times N$ observation matrix with the element $y_{it}$, $F = (F_1, F_2, \ldots, F_T)'$, and $U = (u_{\cdot 1}, \ldots, u_{\cdot T})'$ is the $T \times N$ idiosyncratic error matrix with the element $u_{it}$, $i = 1, \ldots, N$, $t = 1, \ldots, T$. And then, model (1) can be rewritten as the following matrix form

$$Y = F\Lambda' + U. \tag{2}$$

Let

$$\tilde{\mu}_{NT,i} =: \psi_i\big(Y'Y/(NT)\big) = \psi_i\big(YY'/(NT)\big) \tag{3}$$

denote the $i$th largest eigenvalue of $Y'Y/(NT)$. Ahn and Horenstein (2013) found that the first $r$ eigenvalues are $O_p(1)$ and the rest are at most $O_p\left(\frac{1}{m}\right)$, and then proposed the eigenvalue-based ratio-type estimators, i.e., ER and GR, to determine the number of common factors as follows

$$\hat{r}_{ER} = \arg\max_{1\le i\le kmax} \frac{\tilde{\mu}_{NT,i}}{\tilde{\mu}_{NT,i+1}}, \qquad \hat{r}_{GR} = \arg\max_{1\le i\le kmax} \frac{\ln(V_{i-1}) - \ln(V_i)}{\ln(V_i) - \ln(V_{i+1})}, \tag{4}$$

where $m = \min\{N, T\}$ and $M = \max\{N, T\}$, $V_i = \sum_{j=i+1}^{m} \tilde{\mu}_{NT,j}$, and *kmax* is the predetermined possible maximum value of the number of factors.

Ahn and Horenstein (2013) argued that the two objective functions in (4) at the true value point of the number of factors are $O_p(m)$ and the others $O_p(1)$, and then the eigenvalue ratio-based estimators can be proven to be consistent when $m$ is sufficiently large. In practice, however, when not all factors are strong, the number of common factors is not detected