Economics Letters 144 (2016) 102–106

Contents lists available at ScienceDirect

Economics Letters

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

Robust determination for the number of common factors in the approximate factor models

Jianhong Wu

College of Mathematics and Science, Shanghai Normal University, Shanghai 200234, China

ARTICLE INFO

Article history:
Received 17 February 2016
Accepted 20 April 2016
Available online 4 May 2016

JEL classification:
C1
C13

Keywords:
Approximate factor model
Determining the number of factors
Dominant factor
Eigenvalues
Transformation function

ABSTRACT

This paper proposes a robust determination method for the number of common factors in the approximate factor models. The new method is based on the ratio values of some transformation function of adjacent eigenvalues arranged in descending order. 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 method has desired performance on truly selecting the value of the number of latent common factors whether there are dominant factors or not. Monte Carlo simulation is carried out for illustration.

1. Introduction

In the analysis of large dimensional factor models (Chamberlain and Rothschild, 1983), one of fundamental issues is how to consistently determine the number of common factors. In the last decade, a lot of work focus on the issue for the approximate factor models with large individual number n and time length T. The reader may refer to, e.g., Bai and Ng (2002), Bai and Ng (2007), Onatski (2006, 2010), Hallin and Liska (2007), Pan and Yao (2008), Bathia et al. (2010), Lam and Yao (2012), Ahn and Horenstein (2013), Caner and Han (2014) and Xia et al. (2015) for more details. Clearly, various methods can lead to different estimation results for the number of factors due to their different predetermined conditions and different applicability, although they are all proven to be consistent. In this paper we will propose a robust determination method for the number of factors in approximate factor models, which can be expected to have desired performance. For the sake of simplicity and without loss of generality, we mainly focus on the static factor models here.

In the following, we give some simple reviews for the main determination methods in the static factor models in the literature. Up to our knowledge, Bai and Ng (2002) should be the first one to study theoretically the determination method of the number of factors in approximate factor models with large individual number and large time length. They obtained the estimators of the number of factors by minimizing one of the two model selection criterion functions, names PC and IC. One serious problem is that, when the penalty term is replaced by its limit multiple, the resulted estimator can be different based on the finite sample in practice although its asymptotic consistency still hold. Onatski (2006, 2010) 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 only available for the case in which the idiosyncratic errors are either autocorrelated or cross-sectionally correlated, but not both (Ahn and Horenstein, 2013). As reported in 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. Caner and Han (2014) followed the similar idea of Bai and Ng (2002) and then proposed a group bridge estimator, which is also dependent on the choices of some parameters in finite sample size. This is similar to that of Bai and Ng (2002). All the above estimators need to predetermine the possible maximum

E-mail address: wjhstat1@163.com.

http://dx.doi.org/10.1016/j.econlet.2016.04.026

of the number of factors, which may lead to overestimation or underestimation of the number when the maximum (hereafter, denoted by $r_{\max}$) is too large or too small, as argued by Ahn and Horenstein (2013). Xia et al. (2015) proposed a ridge-type estimator and a BIC-type estimator for the number of factors in factor modeling for volatility of multivariate time series. However, all the eigenvalue ratio-type estimators in the existing literature have worse performance when some factors are dominant, which can be partly verified in the simulation study in this paper.

This paper proposes a robust determination method for the number of common factors in the approximate factor models. The new method is based on the ratio values of some transformation function of adjacent eigenvalues arranged in descending order and the use of the so-called ridge-type parameter (e.g., Xia et al., 2015). 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 whether there are dominant factors or not. Monte Carlo simulation is carried out for illustration.

The rest of this paper is organized as follows. In the next section, we propose the new determination method of the number of 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. Technical details on the proofs of Theorem 1 are described in the Appendix.

2. Estimation of the number of factors

Consider the following static approximate factor model

$$y_{it} = \lambda_i' F_t + u_{it}, \quad i = 1, 2, \ldots, n,\ t = 1, 2, \ldots, T, \tag{1}$$

where F_t is an r-dimensional vector of common factors, λ_i is an r-dimensional vector of factor loadings, and u_{it} is the idiosyncratic error. The number r of factors is unknown and needs to be estimated. Denote $F = (F_1, F_2, \ldots, F_T)'$, $\Lambda = (\lambda_1, \ldots, \lambda_n)'$, Y is the $n \times T$ observation matrix with the element y_{it}, and U is the $n \times T$ idiosyncratic error matrix with the element u_{it}, $i = 1, \ldots, n$, $t = 1, \ldots, T$. Model (1) can be rewritten as the following matrix form

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

For the sake of notations, let $m = \min\{n, T\}$ and $M = \max\{n, T\}$. Denote the eigenvalues of $\frac{Y'Y}{nT}$ arranged in descending order $\tilde{\mu}_{nT,i}$, $i = 1, \ldots, T$. That is, $\tilde{\mu}_{nT,i} =: \Psi_i(\frac{Y'Y}{nT})$ means the ith largest eigenvalue of $\frac{Y'Y}{nT}$ (hereafter $\Psi_i(\cdot)$ means the ith largest eigenvalue of a positive semi-definite matrix $\cdot$). Ahn and Horenstein (2013) found that the first r eigenvalues are $O_p(1)$ and the rest are at most $O_p(\frac{1}{m})$ under some conditions, and then proposed the eigenvalue-based ratio-type estimators as follows

$$\begin{aligned} \hat{r}_{ER} &= \arg \max_{1 \le i \le r_{\max}} \frac{\tilde{\mu}_{nT,i}}{\tilde{\mu}_{nT,i+1}}, \\ \hat{r}_{GR} &= \arg \max_{1 \le i \le r_{\max}} \frac{\ln(V_{i-1}) - \ln(V_i)}{\ln(V_i) - \ln(V_{i+1})}, \end{aligned} \tag{3}$$

where $V_i = \sum_{j=i+1}^{m} \tilde{\mu}_{nT,j}$, $r_{\max}$ is the predetermined possible maximum value of the number of factors. This eigenvalue-based ratio-type estimation method has also been used by Wang (2012) and Lam and Yao (2012) for different models and scenarios. Ahn and Horenstein (2013) argued that the two objective functions in (3) 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 some factors are dominant, the other factors cannot be detected accurately and then the number of factors may be underestimated. In order to avoid the underestimation of the number especially in the case with the existence of dominant factors, we can use the modified objective function

$$\frac{2\Phi(\tilde{\mu}_{nT,i}) - 1}{2\Phi(\tilde{\mu}_{nT,i+1}) - 1} \tag{4}$$

to replace those of (3) suggested by Ahn and Horenstein (2013), where $\Phi(\cdot)$ is the cumulated distributed function (Abbreviation: cdf) of the standard normal distribution. Note that the transformation function $2\Phi(\tilde{\mu}_{nT,i}) - 1 = Pr(|\xi| \le \tilde{\mu}_{nT,i})$, where ξ is a standardly normally distributed random variable. Clearly, this transformation function's value is close to 0 (or 1) when the positive eigenvalue $\tilde{\mu}_{nT,i}$ is very small (or very large). One of main roles of the transformation is the standardization or shrinkage to the large eigenvalues, and then avoid the underestimation of the number of factors when there exist some dominant factors. By the similar method in Ahn and Horenstein (2013), the above objective function in (4) can be proven to be $O_p(1)$ as $i < r$, $O_p(m)$ as $i = r$ and $O_p(1)$ as $r_{\max} \ge i > r$, which can be expected to guarantee the consistency of the estimation.

Although (Ahn and Horenstein, 2013) argued that the adjacent eigenvalue ratio estimator is not sensitive to the choice of $r_{\max}$ unless it is too large or too small, the choice is always a problem due to its arbitrariness in practice. Moreover, the ratio value of two adjacent eigenvalues may be not stable when the two eigenvalues are very close to zero or even likely to be practical zero (see, e.g., Lam and Yao, 2012, Xia et al., 2015). So, in this paper we consider the ridge-type estimator as follows,

$$\hat{r} = \arg \max_{1 \le i \le m-1} \frac{2\Phi(\tilde{\mu}_{nT,i} + c) - 1}{2\Phi(\tilde{\mu}_{nT,i+1} + c) - 1}. \tag{5}$$

The ridge-type estimation method is also used by Xia et al. (2015) to avoid the instability of the $\frac{0}{0}$-type ratio values. As Xia et al. (2015) argued, this estimate is a modification through adding a positive value in the eigenvalues, which can be seen as an eigenvalues decomposition of the target matrix plus the diagonal matrix cI_T (hereafter I_k means the $k \times k$ identity matrix). Clearly, the ridge-type estimator can avoid the instability of the $\frac{0}{0}$-type ratio values and then need not to preset the possible maximum values of the number of factors ($r_{\max}$).

Remark 1. As Xia et al. (2015) argued, two principles are necessary for the choice of c. Firstly, it should not deteriorate the effectiveness for estimating the number r, and then the absolute value of c should be as small as possible. Secondly, in order to guarantee the consistency of the estimate, the convergence rate of c to zero should be slower than $O(\frac{1}{m})$ which is the convergence rate of the corresponding estimates of those zero eigenvalues. So, it can be suggested $c = \frac{\log(m)}{10m}$ in practice, which is shown to perform well in the simulation study in the next section. If one thinks that the choice of the parameter c is also arbitrary and similar to that of the possible maximum number $r_{\max}$, and then the following alterative is suggested,

$$\hat{r}_0 = \arg \max_{1 \le i \le r_{\max}} \frac{2\Phi(\tilde{\mu}_{nT,i}) - 1}{2\Phi(\tilde{\mu}_{nT,i+1}) - 1}.$$

Let d be a large positive constant. Our theoretical results are derived based on the assumptions as follows.

Assumption A. Let $\mu_{nT,i} = \Psi_i(\frac{\Lambda'\Lambda}{n}\frac{F'F}{T})$ for $i = 1, \ldots, r$. Then, for each $i = 1, \ldots, r$, $plim_{m \to \infty} \mu_{nT,i} = \mu_i \in (0, +\infty)$. Moreover, the number r of common factors is assumed to be finite.