Thailand Statistician
April 2025; 23(2): 393-406
http://statassoc.or.th
Contributed paper

# Comparison of Methods Used for Estimation of Number of Factors for High Dimensional Factor-Augmented Functional-Coefficient Forecasting Models

**Jiraroj Tosasukul*[a, b]**
[a] Department of Mathematics, Faculty of Science, Naresuan University, Phitsanulok, Thailand.
[b] Research Center for Academic Excellence in Mathematics, Naresuan University, Phitsanulok, Thailand.
*Corresponding author; e-mail: jirarojt@nu.ac.th

Received: 26 December 2023
Revised: 16 September 2024
Accepted: 21 October 2024

**Abstract**

Estimating the number of common factors is challenging particularly in nonparametric high dimensional time series model where the dimensions are bigger than the sample size being considered. In this article, we study the different techniques for determining the number of common factors which are considered in the context of reducing dimensions in the factor-augmented functional-coefficient model with high dimensions. The concept behind all the estimate techniques that are used in this research is based on the evaluation of eigenvalues derived from the correlation and covariance matrix. The performance of the estimators is compared using percentages and the average of the estimated factor numbers. The results of the simulation experiments showed that the growth rates of residual variances technique exhibits outstanding performance compared to other methods in situations when the dimension is greater than or equal to the large sample size. Nevertheless, both modified eigenvalues thresholding and the eigenvalue difference criterion techniques provide better results in comparison to other methods in cases when the dimension is smaller than the small sample size. This study employed an empirical approach, using an actual dataset of Australia's quarterly consumer price index (CPI) to provide evidence for the estimating techniques used for determining the factor number.

**Keywords:** Factor models, functional-coefficient models, number of common factors, eigenvalues, principal component analysis.

## 1. Introduction

In several fields, including finance and economics, an enormous amount of data has been collected over the last few years. The fast advancement of technology is the cause. An important concern related to enormous data sets is that the number of variables often gets extensive and often surpasses the sample size. As a result, traditional statistical methods for analyzing data become impractical. For instance, the challenges associated with inferential statistics within the framework

of population means testing (see Chongcharoen et al. 2022) and multiple regression techniques (see Choosawat et al. 2020). The dimension reduction approach is widely employed. The focus of dimensional reduction processes is to get rid of unnecessary features and find significant ones within extensive datasets. In addition, dimension reduction is a machine learning method used to decrease the number of attributes or features in a set of big data.

In general, the high-dimensional factor model is suggested for dealing with the datasets in fields of finance and business economics rather than the screening and shrinking methods that have been developed to choose relevant variables (see Bühlmann and Van de Geer 2011, Fan and Lv 2008, and Hastie et al. 2015). For this reason, both reduction procedures may not perform optimally when used in the study of significantly correlated enormous datasets (see Chamberlain and Rothschild 1983, Fama and French 1992, and Bai and Ng 2002). This issue can become problematic in the context of time series analysis since the predictors often include lagged values and there is a significant correlation between the regressors. This issue may become even more severe when applied to the time series models.

Many approaches have been recommended for solving this issue, with the goal of dimension reduction in clouding the factor-augmented functional-coefficient model (FA-FCM) (see Li et al. 2020). The key idea behind the FA-FCM is that the model extends the functional-coefficient model to high-dimensional. The key idea behind the FA-FCM is that the model extends the functional-coefficient model to high-dimensions in datasets of time series. The model combines the extremely dimensional factor model and the time series model, wherein the coefficient values might be changed via a stochastic index variable. Thus, in order to construct the full model, it is necessary to determine both the factor regressors and the forecasting coefficient functions. It is evident that the latent factor number is generally much less than the total number of exogenous variables included in the factor model. Hence, the aim of reducing dimensions has been succeeded.

One of the most important topics to consider while estimating the enhanced prediction model is the factor number selection procedure, since the number of factors remains unknown. Many approaches have been provided for determining an optimum suitable number of factors. Considering that the principal component analysis is applied in the process of estimating the factor models, it is usual practice to employ certain naive procedures to figure out the total number of components. Jolliffe (2003) developed the concept of cumulative percentage of total variance, Kaiser (1960) presented Kaiser's rule, and Cattell (1966) established the scree plot.

Through the use of the principal component analysis, additional techniques for determining the optimal number of factors have been developed in three separate domains, with the aim of enhancing efficiency. Firstly, some information criteria are employed in order to optimize the reasonable factor number, (see Bai and Ng 2002, Hallin and Liška 2007). Secondly, the factor number can be determined by applying the statistical inference techniques of confidence intervals and hypothesis testing, (see for example, Pan and Yao 2008, Onatski 2009, Kapetanios 2010, and Cai et al. 2020). The third domain is through estimation that principally that are mainly relied on the eigenvalue analysis of both the covariance matrix and correlation matrix, (see Onatski 2010, Lam and Yao 2012, Ahn and Horenstein 2013, and Fan et al. 2022).

Inspired and motivated by the above problem, we are interested in the aforementioned methods for determining the common factor number in the factor model component of the FA-FCM. By implementing the principal component analysis based on the eigenvalue analysis, we aim to compare some estimation methods to optimize the factor number with extremely high dimensional exogenous variables. In the following order, the remaining parts of this work are presented. The FA-FCM is described in Section 2. Estimation method using the principal component technique and methods for

estimating the factor number in factor model are presented in Section 3. The simulation studies and their results are detailed in Section 4. The paper is concluded in Section 5

## 2. The FA-FCM

In this section, we discuss the model with the factor-augmented functional-coefficient that was presented by Li et al. (2020). Let $y_{t+1}$ denotes a response variable. The FA-FCM takes the form:

$$y_{t+1} = \mathbf{F}_t'\boldsymbol{\beta}_1(u_t) + \mathbf{Y}_t'\boldsymbol{\beta}_2(u_t) + \epsilon_{t+1}, \tag{1}$$

and

$$\mathbf{Z}_t = \mathrm{B}\mathbf{F}_t + V_t, \tag{2}$$

where $\mathbf{F}_t = (F_{t1}, \cdots, F_{tk_0})'$ is called latent common factor with $k_0$ dimension in the first term. The following term, $\mathbf{Y}_t = (y_t, \cdots, y_{t-d_0+1})'$ represents $d_0$ the lags vector and $\epsilon_{t+1}$ is the error term. Both coefficient function vectors are defined by $\boldsymbol{\beta}_1(\cdot) = \left[\beta_{11}(\cdot), \cdots, \beta_{1k_0}(\cdot)\right]'$ and $\boldsymbol{\beta}_2(\cdot) = \left[\beta_{21}(\cdot), \cdots, \beta_{2d_0}(\cdot)\right]'$ with the index variable $(u_t)$ varies on time $t$. Furthermore, model $(2)$ is referred to as a factor model, where B and $V_t$ are factor loadings matrix and idiosyncratic errors vector, respectively. It is important to highlight that dimensionality reduction methods are often employed in the construction of the unobservable $\mathbf{F}_t$. These methods reduce the number of observable exogenous features or regressors, $\mathbf{Z}_t$, by their height. Normally, the factor number $(k_0)$ is significantly lower than the dimension regressors $(p)$.

## 3. Methodology

### 3.1. Estimation method of principal component analysis

In the last ten years, there has been significant growth in the methods of data collecting and information technology. As a result, it is now possible to collect a significant amount of data, which tends to be defined by a large number of explanatory features, over a long period of time. When the sample size for a given data, analysis is weak in comparison to the number of variables, many classical statistical analysis methods can become ineffective. Consequently, it is critical to design effective statistical techniques for the analysis of data with high dimensions. For addressing this concern dimension reduction is frequently used. The principal component analysis, which is one of the techniques used for dimension reduction, is frequently employed in practice to reduce originally observed regressors with high dimensions. To use traditional principal components analysis in a multivariate context, one must first convert the original set of variables into a new set of variables. This new set of variables must be arranged so that the most significant variables maintain almost all the variance observed in the original set of variables.

According to the factor model (2) in the previous section, the first thing that we need to do is estimate the unobserved factors which can be used as surrogates for the original regressors. Let $\mathbb{Z} = (\mathbf{Z}_1, \cdots, \mathbf{Z}_n)'$ is with $n \times p$ dimension matrix of the observed regressors. We refer to the first stage procedure of estimation proposed by Li et al. (2020).

The least square estimates of both factor loadings $(\hat{\mathbf{B}})$ and the factors $(\hat{\mathbf{F}}_t)$ minimize:

$$V(\mathbf{F}, \mathbf{B}) = \sum_{i=1}^{p}\sum_{t=1}^{n}(z_{it} - \lambda_i'\mathbf{F}_t)^2 = tr\left[\left(\mathbb{Z} - \mathbf{F}\mathbf{B}'\right)\left(\mathbb{Z} - \mathbf{F}\mathbf{B}'\right)'\right], \tag{3}$$

where $tr(\cdot)$ is defined as a matrix trace. The principal components are calculated in terms of eigenvectors and eigenvalues from $\mathbb{Z}\mathbb{Z}'$, which are normalized by dividing by $np$. Following that, the estimated matrix of common factors $\widehat{\mathbf{F}}$ is constructed by using the $k_0$ eigenvectors multiplied by $\sqrt{n}$ multiplied with the estimated $k_0$ largest eigenvalues, and then the matrix of factor loadings can be constructed as $\widehat{\mathbf{B}} = \mathbb{Z}'\widehat{\mathbf{F}} / n$. It should be noted that the number of factors $k_0$ is usually unknown, and the reasonable value of $k_0$ can be determined by applying optimized techniques, which will be described in the subsequent section.

### 3.2. Methods for determining the number of common factors

This research aims to estimate the number of common factors $k_0$ in the factor model (2) by comparing five different methodologies that rely on the eigenvalues of the covariance matrix.

The first method denotes the ER estimator is presented by Lam and Yao (2012), and it involves determining the number of factors by using the ratios of two adjacent eigenvalues. This method estimates the number of common factors by maximizing the criterion:

$$\hat{k}_{ER} = \arg \max_{1 \le i \le r_{\max}} \left\{ \hat{\lambda}_i / \hat{\lambda}_{i+1} \right\}.$$

The second method denotes the GR estimator is suggested by Ahn and Horenstein (2013). This method focuses on the mean of squared residuals that is defined by $V_i = \sum_{j=i+1}^{p} \hat{\lambda}_j$, which is analyzed in terms of the growth rates of residual variances as follows:

$$\hat{k}_{GR} = \arg \max_{1 \le i \le r_{\max}} \log\left(V_{i-1} / V_i\right) / \log\left(V_i / V_{i+1}\right).$$

The third method denotes the ACT estimator is concentrated on is the determination method, which draws on the eigenvalues of the correlation matrix of the data. Fan et al. (2022) developed a technique to estimate the number of common factors using modified eigenvalues thresholding. The objective function is defined as:

$$\hat{k}_{ACT} = \max\left\{ j : \hat{\lambda}_j^C > s,\ j \in \left[ r_{\max} \right] \right\}, \quad s = 1 + \sqrt{p/(n-1)},$$

where $\hat{\lambda}_j^C$ is a bias corrected estimator of the $i^{\text{th}}$ largest eigenvalue of sample correlation matrix $\hat{\lambda}_i$, $p$ is the dimension and $n$ is the sample size.

The fourth method denotes ON estimator is focused on the maximum eigengap technique, which was introduced by Onatski (2010). For determining the number of common factors $(k_0)$, Onatski (2010) presented the eigenvalue difference criterion as follows:

$$\hat{k}_{ON} = \max\left\{ i \le r_{\max} : \hat{\lambda}_i - \hat{\lambda}_{i+1} \ge s \right\},$$

where $s$ being a given threshold and $\hat{\lambda}_i$ is the $i^{\text{th}}$ largest eigenvalue of the sample covariance matrix.

The last method denotes EV estimator relies on calculating the total variation of the select few biggest eigenvalues. The number of factors is determined by choosing the initial eigenvectors of the covariance matrix in a descending sequence. The number of determining factors is labeled as "$\hat{k}_{EV}$", whereby using 80% of the total variation is defined as $\hat{k}_{EV\,80\%}$. The concept of this approach was presented by Jolliffe (2003) and implemented by Chen et al. (2018) and Li et al. (2020).

## 4. Simulation Study

In this simulation study, we performed Monte Carlo simulations in different situations. We have compared the performance of the methods for determining the factor number from the univariate FA-FCM. We will consider the true number of factors $k_0 = 3$ and $k_0 = 6$. In each case, our simulation study is divided into 3 parts. In the first part, we set the sample size $n = 50, 100, 500$, and the dimension $p = 0.5n, n, 2n, 3n, 4n$. Next, we generated a data $y_{t+1}$ from the following model:

$$y_{t+1} = \mathbf{F}_t'\boldsymbol{\beta}_1(u_t) + \mathbf{Y}_t'\boldsymbol{\beta}_2(u_t) + \epsilon_{t+1}, \tag{4}$$

when $u_t$ and $\epsilon_t$ are drawn from $U(0,1)$ and $N(0,1)$, respectively.

For the true number of common factors $k_0 = 3$, we set the lags of response $d_0 = 3$. So, the coefficient functions are defined by the vectors, $\boldsymbol{\beta}_1(u) = \left[\beta_{11}(u), \beta_{12}(u), \beta_{13}(u)\right]'$ and $\boldsymbol{\beta}_2(u) = \left[\beta_{21}(u), \beta_{22}(u), \beta_{23}(u)\right]'$, with $\beta_{11}(u) = \frac{1}{3}\left(\tan(u)\right) - u$, $\beta_{12}(u) = \frac{1}{5}\ln(2u+3)$, $\beta_{13}(u) = \frac{1}{6}\cos^{-1}(u) + u$, $\beta_{21}(u) = 2\sqrt{u}$, $\beta_{22}(u) = e^{2u} - 2$ and $\beta_{23}(u) = u^2 + 2\sin(u)$.

Let $z_t$ be i.i.d. from $N(0,1)$ and we generate the vector of latent factor $\mathbf{F}_t = (F_{t1}, F_{t2}F_{t3})$ from $AR(1)$ model as

$$F_{tk} = 0.6 \cdot F_{t-1,k} + z_{tk},\ k = 1,2,3. \tag{5}$$

In addition, for the true number of common factors $k_0 = 6$, we also set the lags of response $d_0 = 3$. Then, the coefficient functions are defined by the vectors, $\boldsymbol{\beta}_1(u) = \left[\beta_{11}(u), \beta_{12}(u), \beta_{13}(u), \beta_{14}(u), \beta_{15}(u), \beta_{16}(u)\right]'$ and $\boldsymbol{\beta}_2(u) = \left[\beta_{21}(u), \beta_{22}(u), \beta_{23}(u)\right]'$, with $\beta_{11}(u) = \frac{1}{3}\left(\sin^{-1}(u) + \tan(u)\right) + 2u^2$, $\beta_{12}(u) = \frac{1}{5}\arccos^2(u)$, $\beta_{13}(u) = 3u + \frac{\sqrt{u}}{2}$, $\beta_{14}(u) = 2\log(u+2) - u$, $\beta_{15}(u) = 3u^2 + u - 1$, $\beta_{16}(u) = e^u + 1$, $\beta_{21}(u) = \frac{1}{4}uu$, $\beta_{22}(u) = \frac{1}{2}\cos(u)$ and $\beta_{23}(u) = \frac{1}{2}\tan(u)$. In this case, the latent factor is generated as

$$F_{tk} = 0.4 \cdot F_{t-1,k} + z_{tk},\ k = 1,2,\ldots,6. \tag{6}$$

In the second part, through the use of principal component analysis, we were able to compute the eigenvalues of the correlation matrix. Following that, we use the five methods that were discussed in the previous part in order to determine the number of common factors, $\hat{k}_{ER}, \hat{k}_{GR}, \hat{k}_{ACT}, \hat{k}_{ON}$ and $\hat{k}_{EV}$. For the last method, we considered selecting the few biggest eigenvalues of the total variation 70%, 75%, 80%, 85%, 90% and 95% that were labeled as $\hat{k}_{EV\,70\%}, \hat{k}_{EV\,75\%}, \hat{k}_{EV\,80\%}, \hat{k}_{EV\,85\%}, \hat{k}_{EV\,90\%}$ and $\hat{k}_{EV\,95\%}$, respectively.

In the third part, we computed the percentages of 4 different values, accurate estimation value $(\hat{k} = k_0)$, overestimation value $(\hat{k} > k_0)$, and underestimation value $(\hat{k} < k_0)$ of the number of common factors, as well as the average of the factor numbers $(\bar{\hat{k}})$ in each method. We repeat all the parts 10,000 times for every simulation study case for both simulation experiments.

The simulation results for two different situations, $k_0 = 3$ and $k_0 = 6,$ in which the true factor numbers are shown in Tables 1-3 and 4-6, respectively.

**Table 1** The percentages of the determined number of common factors $(k_0 = 3)$ with $n = 50$

| $p$ | | ER | GR | ACT | ON | EV | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | 75% | 80% | 85% | 90% | 95% |
| $0.5n$ | $\hat{k} = k_0$ | 62.3300 | 88.7200 | 99.2100 | 98.2700 | 9.6700 | 33.6800 | 65.6100 | 48.9000 | 3.6000 |
| | $\hat{k} > k_0$ | 0 | 0 | 0.5300 | 1.6900 | 0 | 0.4000 | 7.1900 | 47.8800 | 96.3800 |
| | $\hat{k} < k_0$ | 37.6700 | 11.2800 | 0.2600 | 0.0400 | 90.3300 | 65.9200 | 27.2000 | 3.2200 | 0.0200 |
| | $\overline{\hat{k}}$ | 2.3670 | 2.8465 | 3.0028 | 3.0202 | 2.1051 | 2.3506 | 2.8223 | 3.8502 | 6.6409 |
| $n$ | $\hat{k} = k_0$ | 89.8600 | 99.0100 | 99.8100 | 98.4200 | 9.9000 | 40.9700 | 76.7000 | 37.7700 | 0.4100 |
| | $\hat{k} > k_0$ | 0 | 0 | 0 | 1.5800 | 0 | 0.1500 | 7.9800 | 61.6100 | 99.5900 |
| | $\hat{k} < k_0$ | 10.1400 | 0.9900 | 0.1900 | 0 | 90.1000 | 58.8800 | 15.3200 | 0.6200 | 0 |
| | $\overline{\hat{k}}$ | 2.8283 | 2.9887 | 2.9981 | 3.0191 | 2.1022 | 2.4144 | 2.9612 | 4.4960 | 9.0858 |
| $2n$ | $\hat{k} = k_0$ | 98.5400 | 99.9300 | 99.7200 | 99.0000 | 8.7500 | 45.4800 | 82.9600 | 29.3000 | 0.0500 |
| | $\hat{k} > k_0$ | 0 | 0 | 0 | 1.0000 | 0 | 0.0800 | 7.2000 | 70.5600 | 99.9500 |
| | $\hat{k} < k_0$ | 1.4600 | 0.0700 | 0.2800 | 0 | 91.2500 | 54.4400 | 9.8400 | 0.1400 | 0 |
| | $\overline{\hat{k}}$ | 2.9753 | 2.9993 | 2.9972 | 3.0128 | 2.0890 | 2.4571 | 3.0154 | 5.1413 | 11.6379 |
| $3n$ | $\hat{k} = k_0$ | 99.5300 | 99.9900 | 99.1600 | 99.2300 | 8.4500 | 47.4000 | 85.0000 | 26.0800 | 0.0100 |
| | $\hat{k} > k_0$ | 0 | 0 | 0 | 0.7700 | 0 | 0.0500 | 7.2600 | 73.8800 | 99.9900 |
| | $\hat{k} < k_0$ | 0.4700 | 0.0100 | 0.8400 | 0 | 91.5500 | 52.5500 | 7.7400 | 0.0400 | 0 |
| | $\overline{\hat{k}}$ | 2.9923 | 2.9999 | 2.9916 | 3.0103 | 2.0854 | 2.4757 | 3.0499 | 5.5503 | 12.9836 |
| $4n$ | $\hat{k} = k_0$ | 99.9200 | 99.9800 | 98.3400 | 99.2100 | 8.5500 | 48.0300 | 86.0100 | 24.7300 | 0.0100 |
| | $\hat{k} > k_0$ | 0 | 0 | 0 | 0.7900 | 0 | 0.0300 | 7.0800 | 75.2400 | 99.9900 |
| | $\hat{k} < k_0$ | 0.0800 | 0.0200 | 1.6600 | 0 | 91.4500 | 51.9400 | 6.9100 | 0.0300 | 0 |
| | $\overline{\hat{k}}$ | 2.9987 | 2.9998 | 2.9834 | 3.0117 | 2.0861 | 2.4810 | 3.0576 | 5.8221 | 13.7273 |

Table 1 illustrates the simulation results for estimating of the number of factors when $k_0 = 3$ in the case of small sample size $(n = 50)$. It is evident that the $\hat{k}_{ACT}$ works very well when the dimension size $(p)$ is less than or equal to the sample size $(n)$, as it achieves the greatest percentages of the estimated factor number (99.21%) and the average estimated factor number that is closest to the real factor number. Furthermore, we can see that the $\hat{k}_{GR}$ works extremely well when the dimension exceeds the sample size by obtaining the highest percentages of the estimated factor number and the average of the estimated factor number that is most equal to the true factor number. For the EV method by using 70%, 75%, 80%, 85%, 90% and 95% of the total variation, the $\hat{k}_{EV\,85\%}$ outperforms others by much bigger percentages and a reasonable average of the estimated factor number. Moreover, the estimator $\hat{k}_{ON}$ performs really well with the percentages of the estimated factor number are greater than 98% and the average estimated factor number that is closer to the true factor number as the dimension grows.

**Table 2** The percentages of the determined number of common factors $(k_0 = 3)$ with $n = 100$

| $p$ | | ER | GR | ACT | ON | EV | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | 75% | 80% | 85% | 90% | 95% |
| $0.5n$ | $\hat{k} = k_0$ | 98.7400 | 99.9300 | 99.4400 | 99.1400 | 9.0000 | 44.8500 | 83.0900 | 30.0900 | 0.0500 |
| | $\hat{k} > k_0$ | 0 | 0 | 0.5600 | 0.8600 | 0 | 0.1200 | 7.4900 | 69.7600 | 99.9500 |
| | $\hat{k} < k_0$ | 1.2600 | 0.0700 | 0 | 0 | 91.0000 | 55.0300 | 9.4200 | 0.1500 | 0 |
| | $\overline{\hat{k}}$ | 2.9791 | 2.9993 | 3.0056 | 3.0103 | 2.0914 | 2.4513 | 3.0257 | 5.1314 | 11.6187 |
| $n$ | $\hat{k} = k_0$ | 99.9900 | 100 | 100 | 99.2700 | 7.0900 | 51.0300 | 90.1600 | 19.0100 | 0 |
| | $\hat{k} > k_0$ | 0 | 0 | 0 | 0.7300 | 0 | 0.0100 | 5.8700 | 80.9700 | 100 |
| | $\hat{k} < k_0$ | 0.0100 | 0 | 0 | 0 | 92.9100 | 48.9600 | 3.9700 | 0.0200 | 0 |
| | $\overline{\hat{k}}$ | 2.9998 | 3 | 3 | 3.0096 | 2.0711 | 2.5105 | 3.0688 | 6.3838 | 15.0532 |
| $2n$ | $\hat{k} = k_0$ | 100 | 100 | 100 | 99.4000 | 6.0600 | 55.7200 | 93.1200 | 12.6200 | 0 |
| | $\hat{k} > k_0$ | 0 | 0 | 0 | 0.6000 | 0 | 0 | 5.2200 | 87.3800 | 100 |
| | $\hat{k} < k_0$ | 0 | 0 | 0 | 0 | 93.9400 | 44.2800 | 1.6600 | 0 | 0 |
| | $\overline{\hat{k}}$ | 3 | 3 | 3 | 3.0083 | 2.0606 | 2.5572 | 3.0887 | 7.8422 | 15.9153 |
| $3n$ | $\hat{k} = k_0$ | 100 | 100 | 100 | 99.4400 | 5.2400 | 56.3000 | 94.6500 | 10.3200 | 0 |
| | $\hat{k} > k_0$ | 0 | 0 | 0 | 0.5600 | 0 | 0 | 4.2700 | 89.6800 | 100 |
| | $\hat{k} < k_0$ | 0 | 0 | 0 | 0 | 94.7600 | 43.7000 | 1.0800 | 0 | 0 |
| | $\overline{\hat{k}}$ | 3 | 3 | 3 | 3.0060 | 2.0524 | 2.5630 | 3.0790 | 8.5795 | 15.9843 |
| $4n$ | $\hat{k} = k_0$ | 100 | 100 | 100 | 99.4500 | 5.1000 | 57.4300 | 94.7500 | 9.2700 | 0 |
| | $\hat{k} > k_0$ | 0 | 0 | 0 | 0.5500 | 0 | 0 | 4.3000 | 90.7300 | 100 |
| | $\hat{k} < k_0$ | 0 | 0 | 0 | 0 | 94.9000 | 42.5700 | 0.9500 | 0 | 0 |
| | $\overline{\hat{k}}$ | 3 | 3 | 3 | 3.0065 | 2.0510 | 2.5743 | 3.0829 | 9.1448 | 15.9954 |

Table 2 shows the results for estimating of the factor numbers when $k_0 = 3$ and $n = 100$. It is clear that the estimation method $(\hat{k}_{GR})$ shows very good performance for all different dimensions. It has the greatest percentage of the determined factor number, and its average is actually quite similar to the actual factor number. In addition, the performance of the $\hat{k}_{ACT}$ is very effective as well when the dimensions and sample size are the same. Furthermore, the $\hat{k}_{ON}$ exhibits excellent performance when the estimated factor number reaches 99% and when the average estimated factor number is within an acceptable range. In the case of the EV method, the $\hat{k}_{EV\,85\%}$ is significantly higher than others, in the same method, both in terms of percentages and the average.

**Table 3** The percentages of the determined number of common factors $(k_0 = 3)$ with $n = 500$

| $p$ | | ER | GR | ACT | ON | EV | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | 75% | 80% | 85% | 90% | 95% |
| $0.5n$ | $\hat{k} = k_0$ | 100 | 100 | 99.6200 | 99.5400 | 1.3400 | 68.0200 | 98.9600 | 1.6700 | 0 |
| | $\hat{k} > k_0$ | 0 | 0 | 0.3800 | 0.4600 | 0 | 0 | 1.0300 | 98.3300 | 100 |
| | $\hat{k} < k_0$ | 0 | 0 | 0 | 0 | 98.6600 | 31.9800 | 0.0100 | 0 | 0 |
| | $\overline{\hat{k}}$ | 3 | 3 | 3.0038 | 3.0061 | 2.0134 | 2.6802 | 3.0281 | 13.7152 | 16 |
| $n$ | $\hat{k} = k_0$ | 100 | 100 | 100 | 99.5700 | 0.3400 | 74.6200 | 99.6200 | 0.2500 | 0 |
| | $\hat{k} > k_0$ | 0 | 0 | 0 | 0.4300 | 0 | 0 | 0.3800 | 99.7500 | 100 |
| | $\hat{k} < k_0$ | 0 | 0 | 0 | 0 | 99.6600 | 25.3800 | 0 | 0 | 0 |
| | $\overline{\hat{k}}$ | 3 | 3 | 3 | 3.0064 | 2.0034 | 2.7462 | 3.0096 | 15.4015 | 16 |
| $2n$ | $\hat{k} = k_0$ | 100 | 100 | 100 | 99.4600 | 0.1800 | 79.4700 | 99.8500 | 0.0300 | 0 |
| | $\hat{k} > k_0$ | 0 | 0 | 0 | 0.5400 | 0 | 0 | 0.1500 | 99.9700 | 100 |
| | $\hat{k} < k_0$ | 0 | 0 | 0 | 0 | 99.8200 | 20.5300 | 0 | 0 | 0 |
| | $\overline{\hat{k}}$ | 3 | 3 | 3 | 3.0060 | 2.0018 | 2.7947 | 3.0048 | 15.8725 | 16 |
| $3n$ | $\hat{k} = k_0$ | 100 | 100 | 100 | 99.6700 | 0.1300 | 79.9400 | 99.8800 | 0.0400 | 0 |
| | $\hat{k} > k_0$ | 0 | 0 | 0 | 0.3300 | 0 | 0 | 0.1200 | 99.9600 | 100 |
| | $\hat{k} < k_0$ | 0 | 0 | 0 | 0 | 99.8700 | 20.0600 | 0 | 0 | 0 |
| | $\overline{\hat{k}}$ | 3 | 3 | 3 | 3.0040 | 2.0013 | 2.7994 | 3.0027 | 15.9228 | 16 |
| $4n$ | $\hat{k} = k_0$ | 100 | 100 | 100 | 99.6800 | 0.0300 | 81.9000 | 99.9400 | 0.0300 | 0 |
| | $\hat{k} > k_0$ | 0 | 0 | 0 | 0.3200 | 0 | 0 | 0.0600 | 99.9700 | 100 |
| | $\hat{k} < k_0$ | 0 | 0 | 0 | 0 | 99.9700 | 18.1000 | 0 | 0 | 0 |
| | $\overline{\hat{k}}$ | 3 | 3 | 3 | 3.0034 | 2.0003 | 2.8190 | 3.0012 | 15.9647 | 16 |

Table 3, in the case of $k_0 = 3$ and $n = 500$, presents that both estimating approaches, $\hat{k}_{ER}$ and $\hat{k}_{GR}$ behave excellent across all dimensions. It has the 100 percentage of the determined factor number, and the averages are equal 3. Furthermore, the performance of the $\hat{k}_{ACT}$ is very efficient in nearly every scenario, except when the dimension is smaller than the sample size. Regarding the EV method, the $\hat{k}_{EV\,85\%}$ exhibits a notably superior performance compared to others using the same method, both in terms of percentages and the average. Additionally, the $\hat{k}_{ON}$ works effectively with percentages of the estimated factor number that are above $99.4\%$ and maintains a reasonable average estimated factor number.

**Table 4** The percentages of the determined number of common factors $(k_0 = 6)$ with $n = 50$

| $p$ | | ER | GR | ACT | ON | EV | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | 75% | 80% | 85% | 90% | 95% |
| $0.5n$ | $\hat{k} = k_0$ | 93.1600 | 98.0500 | 45.8000 | 99.2400 | 0 | 0 | 0 | 0 | 3.8500 |
| | $\hat{k} > k_0$ | 0 | 0 | 0 | 0.7400 | 0 | 0 | 0 | 0 | 0 |
| | $\hat{k} < k_0$ | 6.8400 | 1.9500 | 54.2000 | 0.0200 | 100 | 100 | 100 | 100 | 96.1500 |
| | $\overline{\hat{k}}$ | 5.8581 | 5.9776 | 5.3809 | 6.0095 | 2.8065 | 3.1153 | 3.5758 | 4.0878 | 4.8352 |
| $n$ | $\hat{k} = k_0$ | 99.8100 | 99.9700 | 48.4400 | 99.3300 | 0 | 0 | 0 | 0.01 | 14.6600 |
| | $\hat{k} > k_0$ | 0 | 0 | 0 | 0.6700 | 0 | 0 | 0 | 0 | 0 |
| | $\hat{k} < k_0$ | 0.1900 | 0.0300 | 51.5600 | 0 | 100 | 100 | 100 | 99.99 | 85.3400 |
| | $\overline{\hat{k}}$ | 5.9969 | 5.9997 | 5.3890 | 6.0075 | 2.9698 | 3.3359 | 3.8205 | 4.3624 | 5.0998 |
| $2n$ | $\hat{k} = k_0$ | 100 | 100 | 30.7600 | 99.6100 | 0 | 0 | 0 | 0.0700 | 24.3400 |
| | $\hat{k} > k_0$ | 0 | 0 | 0 | 0.3900 | 0 | 0 | 0 | 0 | 0 |
| | $\hat{k} < k_0$ | 0 | 0 | 69.2400 | 0 | 100 | 100 | 100 | 99.9300 | 75.6600 |
| | $\overline{\hat{k}}$ | 6 | 6 | 5.08410 | 6.0062 | 3.0451 | 3.4625 | 3.9271 | 4.5059 | 5.227 |
| $3n$ | $\hat{k} = k_0$ | 100 | 100 | 17.4500 | 99.8800 | 0 | 0 | 0 | 0.0500 | 28.9200 |
| | $\hat{k} > k_0$ | 0 | 0 | 0 | 0.1200 | 0 | 0 | 0 | 0 | 0 |
| | $\hat{k} < k_0$ | 0 | 0 | 82.5500 | 0 | 100 | 100 | 100 | 99.9500 | 71.0800 |
| | $\overline{\hat{k}}$ | 6 | 6 | 4.7605 | 6.0012 | 3.0728 | 3.5081 | 3.9517 | 4.5642 | 5.2789 |
| $4n$ | $\hat{k} = k_0$ | 100 | 100 | 9 | 99.7300 | 0 | 0 | 0 | 0.0900 | 30.1700 |
| | $\hat{k} > k_0$ | 0 | 0 | 0 | 0.2700 | 0 | 0 | 0 | 0 | 0 |
| | $\hat{k} < k_0$ | 0 | 0 | 91 | 0 | 100 | 100 | 100 | 99.9100 | 69.8300 |
| | $\overline{\hat{k}}$ | 6 | 6 | 4.4487 | 6.0037 | 3.0829 | 3.5194 | 3.972 | 4.5779 | 5.2929 |

Table 4 reports our second simulation results, which aims to estimate the number of common factors $k_0 = 6$ in the situation of $n = 50$. It is obvious that both estimators, $\hat{k}_{ER}$ and $\hat{k}_{GR}$, give excellent performances in cases that the number of dimensions is larger than the number of sample size. Furthermore, when the number of dimensions is equivalent to the sample size, the $\hat{k}_{GR}$ also represents outstanding performance. On the other hand, when the number of dimensions is smaller the sample size, the ON estimator outperforms other estimators by achieving the highest percentage (99.24%) of accurately estimating the factor number, as well as having estimated common factor number that is closest to the actual factor number. Moreover, it has a high percentage of over 99.3% in estimating the factor number and the average estimated factor number closely approximates the true factor number when the dimension rises. The end result of this table, the performance of all the EV method are very poor across all dimensions.

**Table 5** The percentages of the determined number of common factors $(k_0 = 6)$ with $n = 100$

| $p$ | | ER | GR | ACT | ON | EV | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | 75% | 80% | 85% | 90% | 95% |
| $0.5n$ | $\hat{k} = k_0$ | 100 | 100 | 100 | 99.6200 | 0 | 0 | 0 | 0.1100 | 46.2200 |
| | $\hat{k} > k_0$ | 0 | 0 | 0 | 0.3800 | 0 | 0 | 0 | 0 | 0 |
| | $\hat{k} < k_0$ | 0 | 0 | 0 | 0 | 100 | 100 | 100 | 99.8900 | 53.7800 |
| | $\overline{\hat{k}}$ | 6 | 6 | 6 | 6.0040 | 3.3528 | 3.8443 | 4.1661 | 4.8530 | 5.4619 |
| $n$ | $\hat{k} = k_0$ | 100 | 100 | 100 | 99.6500 | 0 | 0 | 0 | 0.4000 | 70.1700 |
| | $\hat{k} > k_0$ | 0 | 0 | 0 | 0.3500 | 0 | 0 | 0 | 0 | 0 |
| | $\hat{k} < k_0$ | 0 | 0 | 0 | 0 | 100 | 100 | 100 | 99.6000 | 29.8300 |
| | $\overline{\hat{k}}$ | 6 | 6 | 6 | 6.0042 | 3.5569 | 3.9559 | 4.3512 | 4.9602 | 5.7017 |
| $2n$ | $\hat{k} = k_0$ | 100 | 100 | 100 | 99.7200 | 0 | 0 | 0 | 0.9900 | 80.8400 |
| | $\hat{k} > k_0$ | 0 | 0 | 0 | 0.2800 | 0 | 0 | 0 | 0 | 0 |
| | $\hat{k} < k_0$ | 0 | 0 | 0 | 0 | 100 | 100 | 100 | 99.0100 | 19.1600 |
| | $\overline{\hat{k}}$ | 6 | 6 | 6 | 6.0029 | 3.6671 | 3.9885 | 4.4588 | 4.9904 | 5.8084 |
| $3n$ | $\hat{k} = k_0$ | 100 | 100 | 100 | 99.7800 | 0 | 0 | 0 | 1 | 83.7000 |
| | $\hat{k} > k_0$ | 0 | 0 | 0 | 0.2200 | 0 | 0 | 0 | 0 | 0 |
| | $\hat{k} < k_0$ | 0 | 0 | 0 | 0 | 100 | 100 | 100 | 99 | 16.3000 |
| | $\overline{\hat{k}}$ | 6 | 6 | 6 | 6.0023 | 3.7021 | 4.001 | 4.5144 | 4.9968 | 5.8370 |
| $4n$ | $\hat{k} = k_0$ | 100 | 100 | 100 | 99.8100 | 0 | 0 | 0 | 1.2000 | 86.6500 |
| | $\hat{k} > k_0$ | 0 | 0 | 0 | 0.1900 | 0 | 0 | 0 | 0 | 0 |
| | $\hat{k} < k_0$ | 0 | 0 | 0 | 0 | 100 | 100 | 100 | 98.8000 | 13.3500 |
| | $\overline{\hat{k}}$ | 6 | 6 | 6 | 6.0019 | 3.7304 | 4.0087 | 4.5384 | 5.0026 | 5.8665 |

Table 5 shows the simulation results for the case $k_0 = 6$ and $n = 100$. Regardless of the dimensions, all three estimating methods, $\hat{k}_{ER}$, $\hat{k}_{GR}$, and $\hat{k}_{ACT}$, work equally well with one hundred percentage for determining the factor number, and the averages are consistently the true number of common factor. Additionally, the estimator $\hat{k}_{ON}$ works efficiently with a minimal proportion $(99.4\%)$ of the estimated factor number, while still maintaining an appropriate average estimated factor number. In the case of the EV method, the $\hat{k}_{EV\,95\%}$ is significantly higher than others, in the same method, both in terms of percentages and the average. In terms of both percentage and average value, the $\hat{k}_{EV\,85\%}$ demonstrates outstanding outcomes compared to others in same method.

**Table 6** The percentages of the determined number of common factors $(k_0 = 6)$ with $n = 500$

| $p$ | | ER | GR | ACT | ON | EV | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | 75% | 80% | 85% | 90% | 95% |
| $0.5n$ | $\hat{k} = k_0$ | 100 | 100 | 100 | 99.8600 | 0 | 0 | 0 | 63.0400 | 100 |
| | $\hat{k} > k_0$ | 0 | 0 | 0 | 0.1400 | 0 | 0 | 0 | 0 | 0 |
| | $\hat{k} < k_0$ | 0 | 0 | 0 | 0 | 100 | 100 | 100 | 36.9600 | 0 |
| | $\overline{\hat{k}}$ | 6 | 6 | 6 | 6.0020 | 4.0372 | 4.9244 | 5 | 5.6304 | 6 |
| $n$ | $\hat{k} = k_0$ | 100 | 100 | 100 | 99.8400 | 0 | 0 | 0 | 79.3700 | 100 |
| | $\hat{k} > k_0$ | 0 | 0 | 0 | 0.1600 | 0 | 0 | 0 | 0 | 0 |
| | $\hat{k} < k_0$ | 0 | 0 | 0 | 0 | 100 | 100 | 100 | 20.6300 | 0 |
| | $\overline{\hat{k}}$ | 6 | 6 | 6 | 6.0016 | 4.0852 | 4.9753 | 5 | 5.7937 | 6 |
| $2n$ | $\hat{k} = k_0$ | 100 | 100 | 100 | 99.8900 | 0 | 0 | 0 | 86.7000 | 100 |
| | $\hat{k} > k_0$ | 0 | 0 | 0 | 0.1100 | 0 | 0 | 0 | 0 | 0 |
| | $\hat{k} < k_0$ | 0 | 0 | 0 | 0 | 100 | 100 | 100 | 13.3000 | 0 |
| | $\overline{\hat{k}}$ | 6 | 6 | 6 | 6.0012 | 4.1312 | 4.9914 | 5 | 5.8670 | 6 |
| $3n$ | $\hat{k} = k_0$ | 100 | 100 | 100 | 99.8800 | 0 | 0 | 0 | 88.9500 | 100 |
| | $\hat{k} > k_0$ | 0 | 0 | 0 | 0.1200 | 0 | 0 | 0 | 0 | 0 |
| | $\hat{k} < k_0$ | 0 | 0 | 0 | 0 | 100 | 100 | 100 | 11.0500 | 0 |
| | $\overline{\hat{k}}$ | 6 | 6 | 6 | 6.0014 | 4.1543 | 4.9929 | 5 | 5.8895 | 6 |
| $4n$ | $\hat{k} = k_0$ | 100 | 100 | 100 | 99.8400 | 0 | 0 | 0 | 89.8500 | 100 |
| | $\hat{k} > k_0$ | 0 | 0 | 0 | 0.1600 | 0 | 0 | 0 | 0 | 0 |
| | $\hat{k} < k_0$ | 0 | 0 | 0 | 0 | 100 | 100 | 100 | 10.1500 | 0 |
| | $\overline{\hat{k}}$ | 6 | 6 | 6 | 6.0016 | 4.1731 | 4.9951 | 5 | 5.8985 | 6 |

The simulation results of $k_0 = 6$ and $n = 500$ are reported in Table 6. With the exception of the ON method, we can see all of the estimators, $\hat{k}_{ER}, \hat{k}_{GR}, \hat{k}_{ACT}$, and $\hat{k}_{EV\,95\%}$ provide outstanding performance by obtaining one hundred percent when it comes to estimate the true number of common factors, and the averages are always the true number of factors. In addition, the ON estimator $(\hat{k}_{ON})$ works efficiently with a minimum percentage $(99.84\%)$ of the estimated factor number, while still preserving a reasonable average the estimated number of the factor.

## 5. Application in Real Data

This section presents empirical evidence to illustrate the estimate techniques used to study a quarterly consumer price index (CPI) in Australia. The CPI is computed by weighted averaging the retail prices of eight capital cities: Canberra, Darwin, Hobart, Perth, Adelaide, Brisbane, Melbourne, and Sydney. The dataset used in this study is obtained from the official websites of the Australian Bureau of Statistics (ABS) (https://www.abs.gov.au/statistics/economy/price-indexes-and-inflation/consumer-price-index-australia/latest-release#data-downloads). It comprises of the expenditure category and selected analytical series index values that are seasonally adjusted. This encompasses the duration from the second quarter (Q2) of 2018 to the second quarter (Q2) of 2024, which consists of 25 sample sizes and 100 explanatory variables representing the main categories of index numbers such as food and non-alcoholic beverages, alcohol and tobacco, clothing and footwear, and insurance and financial services. In the beginning, the explanatory variables complete normalization before the

computation of the eigenvalues of the correlation matrix. Subsequently, the aforementioned procedures, as described in the previous section, are used to estimate the common factor number, $\hat{k}_{ER}, \hat{k}_{GR}, \hat{k}_{ACT}, \hat{k}_{ON}, \hat{k}_{EV75\%}, \hat{k}_{EV80\%}, \hat{k}_{EV85\%}, \hat{k}_{EV90\%}$, and $\hat{k}_{EV95\%}$. The results are reported in Table 7 as shown below.

**Table 7** The determined number of common factors

| | ER | GR | ACT | ON | EV | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | 75% | 80% | 85% | 90% | 95% |
| $\hat{k}$ | 1 | 1 | 2 | 2 | N/A | N/A | 2 | 3 | 5 |

According to the results in Table 7, the $\hat{k}_{ER}$ and $\hat{k}_{GR}$ are equal to 1, the $\hat{k}_{ACT}$, $\hat{k}_{ON}$, and $\hat{k}_{EV85\%}$ are all 2, and $\hat{k}_{EV90\%}$ and $\hat{k}_{EV95\%}$ gives values of 3 and 5, respectively. Nevertheless, both estimators, $\hat{k}_{EV75\%}$ and $\hat{k}_{EV80\%}$ are unable to determine (N/A). The analytical results indicate an appearance of three estimators: $\hat{k}_{ACT}, \hat{k}_{ON}$, and $\hat{k}_{EV85\%}$ that demonstrate significant characteristics to the same outperform estimators of simulation study in Table 1 under the condition of the small sample size where $p = 4n$. From the scree plot in Figure 1, the values obtained are quite acceptable. However, the ER and GR estimators ($\hat{k}_{ER}$ and $\hat{k}_{GR}$) show slight differences. This occurs because the first eigenvalue of the real dataset is more than ten times larger than the second eigenvalue, as depicted in Figure 1. Based on the principles of the ER and GR methods, this significant disparity in eigenvalues suggests that the appropriate number of common factors is one. Furthermore, both estimators ($\hat{k}_{EV75\%}$ and $\hat{k}_{EV80\%}$) are unable to determine the solution due to the substantial total variation of the few largest eigenvalues, which amounts to 0.9600. It is observed that the division of the largest eigenvalue (0.7687) and all the largest variation eigenvalues (0.9600) is above 0.80.

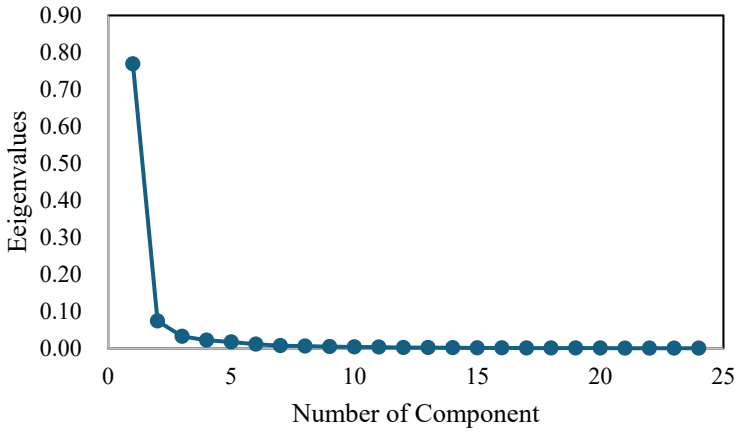



**Figure 1** The scree plot for the real dataset

## 6. Conclusions

This paper examines the estimation methods used for estimating the number of common factors in nonparametric time series models, specifically in cases when the dimensions are larger than the sample size. We analyze the performance of all methods for estimating by conducting Monte Carlo simulations, and we displayed the percentages as well as the average of the estimated factor numbers and percentages. Ultimately, we suggest that the user apply the GR estimator, which has shown exceptional performance in our simulation experiments for nonparametric time series data with a dimension that is more than or equal to the sample size. Moreover, it is recommended to use the ACT and ON estimators when working with small sample numbers that exceed the dimension.

## 7. Further Research

We could conduct further studies to provide a more comprehensive comparison of the number of common factors, such as by determining the factor number in the loading matrix using inferential statistics and rank hypothesis testing (Kong et al. 2019). Additionally, we are exploring the discussed approaches to estimate the number of common factors in various nonparametric high-dimensional time series models.

## Acknowledgements

The research was financially supported by Naresuan University under grant number R2564C003. The author expresses gratitude towards the editor and referees for their valuable comments and recommendations to the manuscript.

## References

Ahn SC, Horenstein AR. Eigenvalue ratio test for the number of factors. Econometrica. 2013 May; 81(3): 1203-1227.

Bai J, Ng S. Determining the number of factors in approximate factor models. Econometrica. 2002 Jan; 70(1): 191-221.

Bühlmann P, Van De Geer S. Statistics for high-dimensional data: methods, theory and applications. Berlin: Springer Science and Business Media; 2011.

Cai T, Han X, Pan G. Limiting laws for divergent spiked eigenvalues and largest nonspiked eigenvalue of sample covariance matrices. Ann Stat. 2020; 48 (3): 1255-1280.

Cattell RB. The scree test for the number of factors. Multivariate Behav Res. 1966; 1(2): 245-276.

Chamberlain G, Rothschild M. Arbitrage, factor structure, and mean-variance analysis on large asset markets. Econometrica. 1983; 51(5): 1281-1304

Chen J, Li D, Linton O, Lu Z. Semiparametric ultra-high dimensional model averaging of nonlinear dynamic time series. J Am Stat Assoc. 2018; 113(522): 919-932.

Chongcharoen S, Paksaranuwat P, Kingphai K. One-sided multivariate test for two population means with common unknown covariance matrices of high-dimensional data. Thailand Statistician. 2022; 20(2): 412-419.

Choosawat C, Reangsephet O, Srisuradetchai P, Lisawadi S. Performance comparison of penalized regression methods in Poisson regression under high-dimensional sparse data with multicollinearity. Thail Stat. 2020; 18(3), 306-318.

Fama EF, French KR. The cross-section of expected stock returns. J Finance. 1992; 47(2): 427-465.

Fan J, Guo J, Zheng S. Estimating number of factors by adjusted eigenvalues thresholding. J Am Stat Assoc. 2022; 117(538): 852-861.

Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. J R Stat Soc Series B Stat Methodol. 2008; 70(5): 849-911.

Hallin M, Liška R. Determining the number of factors in the general dynamic factor model. J Am Stat Assoc. 2007; 102(478): 603-617.

Hastie T, Tibshirani R, Wainwright M. Statistical learning with sparsity: the lasso and generalizations. New York: CRC press; 2015.

Jolliffe IT. Principal component analysis. Technometrics. 2003; 45(3): 276.

Kaiser HF. The application of electronic computers to factor analysis. Educ Psychol Meas. 1960; 20(1): 141-151.

Kapetanios G. A testing procedure for determining the number of factors in approximate factor models with large datasets. J Bus Econ Stat. 2010; 28(3): 397-409.

Kong XB, Liu Z, Zhou W. A rank test for the number of factors with high-frequency data. J Econom. 2019; 211(2): 439-460.

Lam C, Yao Q. Factor modeling for high-dimensional time series: inference for the number of factors. Ann Stat. 2012; 40(2): 694-726.

Li D, Tosasukul J, Zhang W. Nonlinear factor-augmented predictive regression models with functional coefficients. J Time Ser Anal. 2020; 41(3): 367-386.

Onatski A. Testing hypotheses about the number of factors in large factor models. Econometrica. 2009; 77(5): 1447-1479.

Onatski A. Determining the number of factors from empirical distribution of eigenvalues. Rev Econ Stat. 2010; 92(4): 1004-1016.