The non-robustness issue for estimating the number of factors in high dimensional data

Zhenhao Gong*

September 17, 2021

Abstract: The idea of factor analysis is that it can use a few latent factors to capture the variations of a large number of economic variables in a high dimensional data set. A critical question in factor analysis is to estimate the number of factors. Most methods for choosing the number of factors are based on the results from random matrix theory (RMT), which studies the distribution of sample eigenvalues and requires i.i.d and gaussian assumption on the error terms in the factor model. These restrictions may not appropriate when we want to apply those methods in practice. This paper aims to show that those methods are not robust by simulation when the error terms in the factor model are serially and cross-sectionally correlated or have non-gaussian distributions. Our simulation results provide useful recommendations to applied users for how to choose the estimation method in dealing with different types of data.

Keywords: factor analysis, strong and weak factors, high dimensional data, serially and cross-sectionally correlated errors, principal components analysis, random matrix theory

JEL codes: C1; C13; C18; C23; C58

*Department of Economic, University of Connecticut, Email: zhenhao.gong@uconn.edu

1 Introduction

With the more and more data that collected by government and private firms, economists have an opportunity to achieve a better estimation of economic effects and outcomes, using of these newly available large data sets on their researches. These large data sets often characterize as the dimension of variables that have the same order as, or possibly even larger than, the sample size. Handling such large and complex data sets was a challenge to economists and econometricians, however. The reason is that the classical asymptotic theories and standard econometric methods may not be applicable or even break down in such regimes. Thus, we need to develop new theories as well as new methods. One of the core methods for handling large data matrices and high dimensional data sets is through factor analysis. It uses a few latent factors to capture the variations of a large number of economic variables in a high dimensional data set, with wide applications in macroeconomics, finance, and other areas. For example, [25] [26] [27] use diffusion indices (similar as factors) constructed from a large number of macroeconomic series to forecast inflation. [9] measure the effects of monetary policies using a factor-augmented vector auto-regressive vector (FAVAR) approach. [20] consider factors as conditioning information to discuss the conditional mean and conditional volatility of excess stock market returns.

A critical question and also one of the big challenges in factor analysis is to estimate the number of factors. In the classical factor analysis setting, we assume that the cross-section units N is fixed with a relatively large number of time periods T. In such a setting, classical methods to estimate the number of factors includes the likelihood ratio test [19] [8] [4], scree test [10] [11], Kaiser's rule [17], and parallel analysis (PA) [15]. Those methods can not be applied to high dimensional data in which both N and $T \to \infty$, however. In high dimensional regime, factors can be classified into strong and weak factors according to their strengths and assumptions. Some of the most popular methods for estimating the number of strong factors are the information criteria based methods (IC) developed by Bai and Ng [**?**], the eigenvalue difference based method (ED) proposed by Onatski [22] and the eigenvalue ratio based method

(ER) developed by Ahn and Horenstein [1]. These methods for selecting the number of strong factors, however, may fail to detect weak factors in a high dimensional data set. For weak factor estimation, Nadakuditi and Edelman [21] proposed an information criteria based method (NE) to estimate the number of detectable weak factors in a high dimensional data set with white noise. Instead of estimating the number of detectable weak factors, Owen and Wang [24] developed a bi-cross-validation based method (BCV) to estimate the number of useful weak factors in a high dimensional data set with heteroscedastic noise.

Overall, most methods for choosing the number of strong and weak factors in a high dimensional data set are based on the results from random matrix theory (RMT), which studies the distribution of sample eigenvalues and requires i.i.d and gaussian assumption on the error terms in the factor model. These restrictions may not appropriate when we want to apply those methods in practice. Hence, this paper aims to show that those methods are not robust by simulation when the error terms in the factor model are serially and cross-sectionally correlated or have non-gaussian distributions. Our simulation results provide useful recommendations to applied users for how to choose the estimation method in dealing with different types of data.

2 Basic factor model and identification

Factor analysis is based on a model that separates the observed data into an unobserved systematic part (signal part) and an unobserved error part (noise part). The systematic part captures the main information of the data so that we want to separate it from noise part. Specifically, let Y_{it} be the observed data for the i-th cross-section unit at time t, for $i = 1, 2, \cdots, N$ and $t = 1, \cdots, T$. The factor model for Y_{it} is given by

$$Y_{it} = L_i^{'} F_t + e_{it}, \ i = 1, \ldots, N, t = 1, \ldots, T, \tag{2.1}$$

where F_t is a $(r_0 \times 1)$ vector of common factors, L_i is a $(r_0 \times 1)$ vector of loadings associated with F_t, and e_{it} is the idiosyncratic component (noise part) of Y_{it}. The number of true factors in the model is r_0. The product of $L_i^{'} F_t$ is called the common component (signal part) of Y_{it}.

The factors, their loadings, as well as the idiosyncratic errors are not observable. (2.1) can also be represented in a matirx form as

$$Y_t = LF_t + e_t, \tag{2.2}$$

with $Y_t = (Y_{1t}, Y_{2t}, \cdots, Y_{Nt})'$, $L = (L_1, L_2, \cdots, L_N)' \in \mathbb{R}^{N \times r_0}$, and $e_t = (e_{1t}, e_{2t}, \cdots, e_{Nt})'$. Note that in (2.2), L and F_t can not be identified from the product $L'F_t$ since we have $L'RR^{-1}F_t$ for any $r_0 \times r_0$ invertible matrix R, and R has r_0^2 free parameters. Thus, in order to identify L and F_t, we need at least r_0^2 restrictions. One common constraint to make L identifiable up to rotation is to assume $\text{cov}(F_t) = I_r$. This normalization on F_t implies that the latent factors are uncorrelated to each other, which gives $r_0(r_0+1)/2$ restrictions since a symmetric matrix contains $r_0(r_0+1)/2$ free parameters. To eliminate the rotation uncertainty, we can further assuming that LL' is diagonal with distinct entries, which contains $r_0(r_0-1)/2$ restrictions. Together, we have exactly r_0^2 restrictions on L and F_t. Note that there are many other ways to constraint L and F_t for identification. One can refer to [6] and [3] for more details.

In classical factor analysis, we assume fixed T and large N (panel studies) or fixed N and large T (multivariate time series models). In contrast, the high dimensional factor analysis characterizes as both large cross-section units N and large time dimensions T, and N is possibly much larger than T. Such a high dimensional framework greatly expands the application of the factor models into more realistic and modern data-rich economic environments. For example, in macroeconomics, Y_{it} represents the GDP growth rate for country i in period t, F_t is a vector of common shocks, L_i is the heterogeneous impact of the shocks, and e_{it} is the country-specific growth rate. In finance, Y_{it} is the return for asset i in period t, F_t is vector of systematic risks, L_i is the exposure to the factor risks, and e_{it} is the idiosyncratic return [5].

Note that, in this paper, we only consider the static factor model, where the relationship between observed data Y_{it} and its corresponding latent factor F_t is static. For econometrics applications, there are more methods to estimate the number of factors for dynamic factor models such as [2] [13] [14] , which allow Y_t to depend also on f_t with lags in time. Such dependency models are beyond the scope of this paper.

3 Literature review

Here we review the most commonly used methods for estimating the number of factors in a high dimensional data set. We begin with some recently developed methods from the econometrics community for choosing the number of strong factors. Then we consider a source of RMT based methods that are designed to choose the number of weak factors. Before introducing those methods, let's first briefly go over the strong and weak factor assumptions in the literature.

3.1 Strong and weak factor assumptions

Assuming factors F_t and noise e_t are uncorrelated and have zero mean, and normalization $\mathbb{E}(F_t F_t^{'}) = I_r$ for identification, then the population covariance matrix of the factor model (2.2) can be expressed as

$$\Sigma_Y = LL^{'} + \Sigma_e, \tag{3.1}$$

where Σ_Y and Σ_e are the $N \times N$ population covariance matrix of Y_t and e_t, respectively.

Assumption 3.1.1. (*Strong Factor assumption)*
For (3.1), we assumed that $L^{'}L/N \to \Sigma_L$ for some $r_0 \times r_0$ positive definite matrices Σ_L and all the eigenvalues of Σ_e are bounded as $N, T \to \infty$.

This is a standard assumption for factor models. Under this assumption, the top r_0 eigenvalues of Σ_Y are diverge at the rate $O(N)$ while the rest of its eigenvalues are bounded as $N, T \to \infty$. It ensures that PCA or MLE estimators for estimating factors and corresponding loadings in a factor model are consistent. It is also the critical assumption for those methods to consistently estimate the number of strong factors as $N, T \to \infty$.

Assumption 3.1.2. (*Weak Factor assumption)*
In contrast to strong factors, for the weak factors, we assumed that $L^{'}L \to \Sigma_L$ instead of $L^{'}L/N \to \Sigma_L$ and all the eigenvalues of Σ_e are bounded as $N, T \to \infty$.

Under this assumption, all the eigenvalues of Σ_Y are bounded as $N, T \to \infty$ and PCA or MLE estimators for estimating factors and corresponding loadings in a factor model are not

consistent. We can illustrate this by a simple example. For the basic factor model (2.1) we defined before, we can assume e_{it} are i.i.d with mean zero and variance σ^2 and let $r_0 = 1$ for simplicity. If L_i is known, then the OLS estimator for F_t is

$$\hat{F}_t = \frac{\sum_{i=1}^N L_i Y_{it}}{\sum_{i=1}^N L_i^2} = F_t + \frac{\sum_{i=1}^N L_i e_{it}}{\sum_{i=1}^N L_i^2},$$

so we have $E(\hat{F}_t) = F_t$ and $\text{Var}(\hat{F}_t) = \sigma^2 / \sum_{i=1}^N L_i^2$. Hence, for $\hat{F}_t$ to be consistent, we need $\sum_{i=1}^N L_i^2 \to \infty$ (the strong factor assumption) such that $\mathbb{P}(|\hat{F}_t - F_t| > \delta) \leq \text{Var}(\hat{F}_t)/\delta^2 \to 0$ as $N \to \infty$.

There are several reasons why we need to estimate the number of weak factors except for strong ones. First, in many real finance and macroeconomics data sets where both N and T are large, the empirical observations show that the eigenvalues of the sample covariance matrices of these data sets do not obviously separate into groups of large and small eigenvalues. We show the empirical evidence by collecting the real data in the Appendix A. Second, [24] show that the estimation error for recovering the common components (signal part) in (2.2) will decrease by including useful weak factors in the estimation. Third, as [12] showed, if we assume a factor structure for asset returns, an asset's risk premium is approximately equal to a linear combination of its factor loadings. The approximation error goes arbitrarily large as the number of assets increases, however, if we ignore relatively weakly influential factors from consideration.

3.2 Method for estimating strong factors

There are many methods to consistently estimate the number of factors under strong factor assumption as $N, T \to \infty$. Some of the most popular methods are the information criteria based methods (IC) developed by Bai and Ng [**?**]. Let $\hat{L}_r^{\text{pc}}$ and $\hat{F}_r^{\text{pc}}$ be the PCA estimators for loadings and factors (for the details of using PCA in factor analysis, please refer to the Appendix B). Define

$$V(r) = \frac{1}{NT} \left\| Y - \hat{L}_r^{\text{pc}} \hat{F}_r^{\text{pc}} \right\|_F^2, \tag{3.2}$$

and the following loss function:

$$\mathrm{IC}(r) = V(r) + rg(N,T) \quad \text{or} \quad \log(V(r)) + rg(N,T), \tag{3.3}$$

where the penalty function $g(N,T)$ satisfies two condition: (i) $g(N,T) \to 0$, and (ii) $C_{NT}^2 g(N,T)$ $\to \infty$ as $N,T \to \infty$, where $C_{NT} = \min\{\sqrt{N}, \sqrt{T}\}$. Define the estimator for the number of factors as $\hat{r}_{\mathrm{IC}} = \mathrm{argmin}_{0\le r\le r_{max}} \mathrm{IC}(k)$. Then the consistency: $\hat{r}_{\mathrm{IC}} \xrightarrow{p} r_0$, as $N,T \to \infty$, can be established under the strong factor assumption. We take this method as an example and explain why the strong factor assumption is critical in the Appendix C.

The other popular methods for estimating the number of strong factors in a high dimensional data set are the ED and ER methods we introduced before. Specifically, the ED estimator is defined as

$$\hat{r}_{\mathrm{ED}} = \max\left\{r \le r_{max} : \lambda_r - \lambda_{r+1} \ge \delta\right\},$$

where δ is some fixed number, λ_i is the i-th largest eigenvalue of $\hat{\Sigma}_Y$. This method estimates the number of factors by exploiting the structure of idiosyncratic terms using the results from RMT. It explicitly allows serially and cross-sectionally correlated error terms in the factor model in its assumptions. An advantage of this estimator comparing with the IC estimator [**?**] is that the consistency of the ED estimator can allow for much weaker strength of the factors: instead of growing in the order of $O(N)$, the smallest eigenvalue of $L'L$ are just required to diverge in probability as $N \to \infty$. The ER estimator is defined as

$$\hat{r}_{\mathrm{ER}} = \mathrm{argmin}_{0\le r\le r_{max}} \lambda_r/\lambda_{r+1},$$

with $\lambda_0 = \sum_{r=1}^{\min(N,T)} \lambda_r / \log\min(N,T)$. The intuition for this method to work is very simple: based on strong factor assumption, for any $r \neq r_0$ the ratio λ_r/λ_{r+1} converges to $O(1)$ as $N,T \to \infty$, while the the ratio $\lambda_{r_0}/\lambda_{r_0+1}$ diverges to infinity.

Remark. *To use $\hat{r}_{IC}$, $\hat{r}_{ED}$ and $\hat{r}_{ER}$, we need to determine the upper bound r_{max} for r. However, there is no theoretical result to guide choosing r_{max}.*

3.3 Methods for estimating weak factors

Instead of assuming $L^{'}L/N \to \Sigma_L$ for strong factors, it is assumed that $L^{'}L \to \Sigma_L$ as $N, T \to \infty$ for weak factors. The results from random matrix theory (RMT) [18] show that, even for white noise case $\Sigma_e = \sigma^2 I_N$ in (3.1), PCA or MLE estimators of the loadings and factors are inconsistent as $N, T \to \infty$. Specifically, there exists a phase transition phenomenon in the limit: if the k-th largest eigenvalue of population covariance matrix Σ_Y less than the threshold $(\sqrt{N/T}+1)\sigma^2$, it has little chance to detect of the k-th factor using PCA or MLE as $T, N \to \infty$. Define the number of detectable factors as $\#\{i \leq N : \xi_i > (\sqrt{N/T}+1)\sigma^2\}$, where ξ_i is the i-th largest eigenvalue of the population covariance matrix Σ_Y, then one goal is to estimate the number of detectable factors.

Nadakuditi and Edelman [21] developed an information criteria based method (NE) to estimate the number of detectable factors in a high dimensional data set with white noise using the results from RMT, which studies the distribution of the sample eigenvalues. Specifically, for fixed N and large T, Anderson and Rubin [4] characterized the distribution of the sample eigenvalues by large sample asymptotics. Their analysis suggests that the sample eigenvalues will be symmetrically centered around the population eigenvalues. This is not true when the dimensionality is large and the sample size is relatively small, however. New analytical results from RMT can precisely describe the spreading of the sample eigenvalues in the high dimensional regime. The idea of the NE method is that they use the distributional properties of the signal-free $(r = 0)$ sample eigenvalues to approximate the distributional properties of the $N-r$ sample eigenvalues in $\hat{\Sigma}_Y$, assuming the number of factors is r and $r \ll N$. The NE estimator is defined as

$$\hat{r}_{\mathrm{NE}} = \underset{0 \leq r < \min(N,T)}{\arg\min} \left\{ \frac{\beta}{4} \left[\frac{T}{N} \right]^2 t_r^2 + 2(r+1) \right\},$$

where

$$t_r = \left[(N-r) \frac{\sum_{i=r+1}^{N} \lambda_i^2}{\left(\sum_{i=r+1}^{N} \lambda_i \right)^2} - \left(1 + \frac{N}{T} \right) \right] N - \left(\frac{2}{\beta} - 1 \right) \frac{N}{T}.$$

Instead of estimating the number of detectable factors, one may prefer estimating the number of useful factors (including strong and useful weak factors). The number of useful factors can be used to recover an underlying signal matrix $X = LF$ in the factor model more precisely than using the true number of factors or detectable factors. Owen and Wang [24] proposed a method to estimate the number of useful factors based on bi-cross-validation, using randomly held-out submatrices of the data matrix. Their results are simulation based using guidance from random matrix theory (RMT). Specifically, their model is defined as:

$$Y = X + \Sigma^{\frac{1}{2}} E = LF + \Sigma^{\frac{1}{2}} E, \tag{3.4}$$

where $X \in \mathbb{R}^{N \times T}$ (signal matrix) is a product of loading matrix $L \in \mathbb{R}^{N \times r_0}$ and factor matrix $F \in \mathbb{R}^{r_0 \times T}$, and r_0 is the true number of factor. The noise matrix $E \in \mathbb{R}^{N \times T}$ with entries $e_{it} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$. The variance of each cross-section unit is given by $\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2, \cdots \sigma_N^2)$. The goal is to recover the signal matrix X despite the heteroscedastic noise using the criteria:

$$\text{Err}_X(\hat{X}) = \mathbb{E}\left(\|\hat{X} - X\|_F^2 \right). \tag{3.5}$$

Their algorithm for recovering the signal matrix X has two steps. First, they devised early stopping alternation (ESA) method to estimate X given the number of optimal factors r^*. Second, they proposed bi-cross-validation (BCV) method to estimate the number of optimal factors r^* based on the ESA method. For BCV method, the data matrix Y is partitioned into four blocks by randomly select N_0 rows and T_0 columns as the held-out block as below

$$Y = \begin{pmatrix} Y_{00} & Y_{01} \\ Y_{10} & Y_{11} \end{pmatrix},$$

where Y_{00} is the selected $N_0 \times T_0$ held-out block, and Y_{01}, Y_{10}, and Y_{11} are the other three held-in blocks. Correspondingly, X and Σ can be partitioned into four parts. The idea of BCV method is that, for each candidate r, we first use the three held-in blocks to estimate the held-out block

X_{00} (corresponding to Y_{00} in the factor model) and then select the optimal r^* by minimizing the BCV estimated prediction error, which is defined as

$$\mathbb{E}\left(\widehat{\mathrm{PE}}_r\left(Y_{00}\right)\right) = \mathbb{E}\left\{\frac{1}{N_0 T_0}\left\|Y_{00} - \hat{X}_{00}(r)\right\|_F^2\right\}. \tag{3.6}$$

Remark. *The randomness of (3.6) comes from the random partition of the original data matrix.*

4 Issue of non-robustness and simulation design

As we have introduced before, most methods (ED, NE, and BCV) for estimating the number of factors are based on the results from random matrix theory (RMT), which studies the distribution of sample eigenvalues and requires i.i.d and gaussian assumption on the error terms in the factor model. These restrictions may not appropriate when we want to apply them in practice. The purpose of this simulation design is to show that all of those methods we have discussed before are not robust when the error terms in the factor model are serially and cross-sectionally correlated or have non-gaussian distributions. We consider the five representative methods reviewed in Section 3.

4.1 Strong factors only

In this section, we only generate strong factors in our data generating process (DGP). We apply all of those methods (IC2, ED, ER, NE, and BCV) for estimating the number of factors in the factor model with serially, cross-sectionally correlated, or non-gaussian error terms. Note that IC2 is one of the six criterions proposed by [**?**] for choosing the number of strong factors. Serially and cross-sectionally correlated error terms in the factor model tend to cause overestimating the number of factors. Hence, we choose IC2 simply because it uses the largest penalty among the six criterions, so the probability of overestimation is the smallest.

When factors are "strong", we know that the IC2 method [**?**] [27] allows weak serial and cross-sectional dependence in the error terms for large N and T. This is because the depen-

dence of the factor structure will eventually dominate any weak dependence in the error terms asymptotically. The IC2 method does not allow high serial and cross-sectional dependence in error terms, however. Since in this case, the eigenvalues of the covariance matrix of the error terms may not be bounded as $N, T \to \infty$, which violates the strong factor assumption and causes overestimation. The ED and ER methods are designed to estimate the number of strong factors. The assumptions of ED and ER methods allow the error terms in the factor model to be serially and cross-sectionally correlated. The NE and BCV methods are designed to estimate the number of strong and weak factors in a high dimensional set data with white and heteroscedastic noise based on the results from RMT, which require i.i.d and gaussian assumption on the error terms in the factor model.

In this simulation, we are going to show that those methods are not robust when the error terms in the factor model are high serially and cross-sectionally correlated, or have non-gaussian distributions in finite samples. This simulation design follows the design of [7] and [23]. Specifically, we consider the following DGP:

$$\begin{aligned} &Y_{it} = \sum_{j=1}^{r} \lambda_{ij} F_{tj} + e_{it}, \quad \text{where} \\ &\lambda_{ij}, F_{tj} \overset{\text{iid}}{\sim} \mathcal{N}(0,1), \\ &e_{it} = \rho_1 e_{it-1} + (1-\rho_1^2)^{1/2} \xi_{it}, \\ &\xi_{it} = \rho_2 \xi_{i-1,t} + (1-\rho_2^2)^{1/2} \epsilon_{it}, \quad \epsilon_{it} \overset{\text{iid}}{\sim} \mathcal{N}(0,1). \end{aligned} \tag{4.1}$$

We let $r = 2$, and consider the three cases for e_{it} below:

Case I: high serial correlation only, $\rho_1 = 0.9$ and $\rho_2 = 0$;

Case II: mild cross-sectional correlation only, $\rho_1 = 0$ and $\rho_2 = 0.5$;

Case III: non-guassion distributions only, $\rho_1 = \rho_2 = 0$ with four types of distributions for e_{it}: normal, gamma, lognormal and chi-square with mean zero and variance 0.5.

4.2 Useful weak factor only and mixed strong and useful weak factors

In this section, we generate strong and useful weak factors in our DGP with **three scenarios**. In the first scenario, we only generate six useful weak factors. In the second scenario, we mixed five useful weak factors with one strong factor. In the third scenario, we mixed three useful weak factors with three strong factors. We then apply all of those five methods (IC2, ED, ER, NE, and BCV) for estimating the number of factors in the factor model with serially, cross-sectionally correlated, or non-guassian error terms. This simulation design follows the design of [24] and [23]. Consider the factor model as:

$$\begin{aligned} Y &= X + \Sigma^{\frac{1}{2}} E \\ &= \Sigma^{\frac{1}{2}}(\Sigma^{-\frac{1}{2}} X + E) = \Sigma^{\frac{1}{2}}(\sqrt{T}\hat{U}\hat{D}\hat{V}' + E), \end{aligned} \tag{4.2}$$

where $\sqrt{T}\hat{U}\hat{D}\hat{V}'$ is the singular value decomposition (SVD) for $\Sigma^{-\frac{1}{2}}X$ with $\hat{U} \in \mathbb{R}^{N\times \min(N,T)}$, $\hat{V} \in \mathbb{R}^{T\times \min(N,T)}$, $\hat{D} = \text{diag}(\hat{d}_1, \hat{d}_2, \cdots, \hat{d}_{\min(N,T)})$, $\hat{U}'\hat{U} = \hat{V}'\hat{V} = I_{\min(N,T)}$, and $\hat{d}_1 \geq \hat{d}_2 \geq , \cdots, \geq \hat{d}_{\min(N,T)}$. For the weighted signal matrix $\Sigma^{-\frac{1}{2}}X = \sqrt{T}\hat{U}\hat{D}\hat{V}'$, we can generate the factors with different strengths for the three scenarios by specified the entries in $\hat{D}$. For the error term $\Sigma^{\frac{1}{2}}E$ in (4.2), assuming homoscedastical noise $\Sigma = I_N$, we consider three cases below:

Case I: high serial correlation only,

$E = (e_{it})_{N\times T} : e_{it} = \rho_1 e_{it-1} + (1-\rho_1^2)^{1/2}\epsilon_{it}, \quad \epsilon_{it} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ with $\rho_1 = 0.9$;

Case II: mild cross-sectional correlation only,

$E = (e_{it})_{N\times T} : e_{it} = \rho_2 e_{i-1,t} + (1-\rho_2^2)^{1/2}\epsilon_{it},\ \epsilon_{it} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ with $\rho_2 = 0.5$;

Case III: non-guassion distributions only,

$E = (e_{itj})_{N\times T} : e_{it}$ is i.i.d with three types of non-guassion distributions: gamma, log-normal and chi-square with mean zero and variance 0.5.

Remark. *In the simulation results, we use* $E = (e_{it})_{N\times T} : e_{it} \overset{iid}{\sim} \mathcal{N}(0,1)$ *as our benchmark since the number of weak factors are hard to estimate precisely in finite small samples.*

5 Numerical results and discussion

5.1 Results for estimating strong factors only

Table 1 shows the finite sample performances of those five methods we discussed on choosing the number of strong factors in the factor model with white, high serially correlated, or mild cross-sectionally correlated error terms. We take 1000 replications for each pair of (N, T) with a specified type of noise in the simulation. From table 1, we can see that all of those five methods can almost precisely choose the number of factors (see the columns 3-5) when the factors are "strong" with white noise only in the factor model. However, when the error terms are high serially correlated, the IC2, NE, and BCV methods will almost inevitably overestimate the number of factors (see the columns 6-8). The ER and ED methods perform quite well when T is large. The results are consistent with the theory since the assumptions of the ER and ED methods allow serially and cross-sectionally error terms in the factor model as $N, T \to \infty$.

Also, when the error terms are mild cross-sectionally correlated, the IC2 method tends to overestimate the number of factors when N is small but it performs quite well when N is large (see the columns 9-11). It is consistent with the theory since the assumptions of the IC2 method allow weak serially and cross-sectionally correlated error terms in the factor model as $N, T \to \infty$. The ER and ED methods are robust to cross-sectionally correlated error terms in factor model for all pairs of N and T in the results of our simulation. The NE and BVC methods perform quite well when N is large. Besides, table 2 shows that, when the error terms in the factor model have non-gaussian distributions, all of those five methods have poor performances for choosing the number of strong factors in the factor model. Overall, if we only have strong factors to be estimated in the data and the data is serially and cross-sectionally correlated with gaussion distribution, our simulation results recommend to use the ER or ED method for estimating the number of factors in the data.

5.2 Results for estimating mixed strong and useful weak factors

Table 3 shows the performances of those five methods on estimating the number of useful weak factors only in the first scenario with white, high serially correlated, mild cross-sectionally correlated, or non-gaussion error terms in the factor model. From table 3, we can see that IC2 and ED methods actually perform very well for estimating the number of useful weak factors with white noise in the factor model, though they are not designed to estimate the number of weak factors. Also, when N is large, the IC2 method performs quite well when the error terms in the factor model are mild cross-sectionally correlated. However, the IC2 and ED methods have poor performances when the error terms in the factor model are high serially correlated or have non-gaussion error terms. The ER method fails to estimate the number of useful weak factors in all cases. The BCV and NE methods perform quite well when the error terms in the factor model are white, mild cross-sectionally correlated, or have chi-square distribution. They have poor performances when the error terms are high serially correlated or have lognormal and gamma distributions, however.

Table 4 shows the performances of those five methods on estimating the number of mixed strong and useful weak factors in the second scenario with white, high serially correlated, mild cross-sectionally correlated, or non-gaussion error terms in the factor model. Comparing table 4 with table 3, we can see that the performances of the IC2, NE, and BCV methods in the table 4 are the same as the results in table 3. The ED method performs similar to the results in table 3, except now it performs surprisingly well for all N and T pairs when error terms in the factor model are mild cross-sectionally correlated. The ER method still fails to choose the number of weak factors, but it is quite robust to separate the number of strong factors from weak ones in the data for all types of the noise terms in the factor model.

Table 5 shows the performances of those five methods on estimating the number of mixed strong and useful weak factors in the third scenario with white, high serially correlated, mild cross-sectionally correlated, or non-gaussian error terms in the factor model. Comparing table 5 with table 4, we can see that the IC2, ED, and NE methods have better performances for estimating the number of strong and useful weak factors in table 5 than in table 4 when the

error terms in the factor model have non-guassian distributions. The BCV method performs the same as the results in table 4. The ER method not only fails to choose the number of weak factors but also not perform well for estimating the number of strong factors for all types of error terms in the factor model. Overall, if we want to estimate the number of strong and useful weak factors in the data with white noise only, our simulation results recommend using the IC2, ED, NE, or BCV method. Also, if the data are mild cross sectionally-correlated with mixed strong and useful weak factors, our simulation results recommend using the ED, NE or BCV method. Besides, if the data has chi-square distribution, our simulation results recommend to use the NE or BCV method.

6 Conclusion

In this paper, we have shown that all of the methods we have discussed for choosing the number of factors in high dimensional data are not robust by simulation when the error terms in the factor model are serially and cross-sectionally correlated or have non-gaussian distributions. Our simulation results provide useful recommendations to applied users for how to choose the estimation method in dealing with different types of data. Specifically, if we only have strong factors to be estimated in the data and the data is serially and cross-sectionally correlated with gaussion distribution, our simulation results recommend to use the ER or ED method for estimating the number of factors. If we want to estimate the number of strong and useful weak factors in the data with white noise only, our simulation results recommend using the IC2, ED, NE, or BCV method. Also, if the data are mild cross sectionally-correlated with mixed strong and useful weak factors, our simulation results recommend to use the ED, NE or BCV method. Besides, if the data has chi-square distribution, our simulation results recommend using the NE or BCV method. Further researches will be needed to take care of the cross-sectional and serial correlation in error terms in the context of choosing of number of strong and weak factors in a high dimensional data set.

Table 1: Finite-sample performances of those five methods on choosing the number of strong factors with white, high serially correlated, or mild cross-sectionally correlated error terms in the factor model.

		White			Serial			Cross sectional		
N	T	$<$	$=$	$>$	$<$	$=$	$>$	$<$	$=$	$>$
IC2										
20	200	0	100	0	0	0	100	0	0.0	100.0
20	100	0	100	0	0	0	100	0	2.1	97.9
100	20	0	100	0	0	0	100	0	100.0	0.0
50	50	0	100	0	0	0	100	0	99.7	0.3
200	20	0	100	0	0	0	100	0	100.0	0.0
ER										
20	200	0.5	99.5	0	5.9	94.1	0.0	4.2	95.8	0
20	100	1.4	98.6	0	16.5	83.5	0.0	6.9	93.1	0
100	20	0.7	99.3	0	45.6	54.4	0.0	2.5	97.5	0
50	50	0.0	100.0	0	11.5	68.4	20.1	0.2	99.8	0
200	20	0.6	99.4	0	44.7	55.3	0.0	0.8	99.2	0
ED										
20	200	0	99.8	0.2	0.0	86.2	13.8	0.1	95.5	4.4
20	100	0	99.2	0.8	3.9	61.4	34.7	0.2	93.0	6.8
100	20	0	99.2	0.8	0.0	0.4	99.6	0.0	95.3	4.7
50	50	0	97.4	2.6	4.3	31.3	64.4	0.0	96.4	3.6
200	20	0	99.6	0.4	0.1	0.6	99.3	0.0	98.2	1.8
NE										
20	200	0	100	0	0	0.2	99.8	0	0.5	99.5
20	100	0	100	0	0	0.0	100.0	0	31.0	69.0
100	20	0	100	0	0	0.0	100.0	0	97.3	2.7
50	50	0	100	0	0	0.0	100.0	0	78.4	21.6
200	20	0	100	0	0	0.0	100.0	0	100.0	0.0
BCV										
20	200	0.0	92.1	7.9	0	8.8	91.2	0.0	16.1	83.9
20	100	0.0	92.4	7.6	0	0.1	99.9	0.0	23.1	76.9
100	20	0.0	95.1	4.9	0	0.0	100.0	0.2	85.9	13.9
50	50	0.0	99.8	0.2	0	0.0	100.0	0.0	83.3	16.7
200	20	0.1	95.1	4.8	0	0.0	100.0	0.0	91.1	8.9

Note: $>, =, <$*: overestimation, correct estimation, underestimation, respectively.*

Table 2: Finite-sample performances of those five methods on choosing the number of strong factors with non-guassion distributional error terms in the factor model.

		Lognormal			Gamma			Chi-square		
N	T	$<$	$=$	$>$	$<$	$=$	$>$	$<$	$=$	$>$
IC2										
20	200	41.3	36.8	21.9	0.1	10.3	89.6	0	0.0	100.0
20	100	33.6	35.3	31.1	0.7	12.1	87.2	0	0.2	99.8
100	20	34.3	35.4	30.3	0.4	14.0	85.6	0	0.3	99.7
50	50	49.4	35.3	15.3	0.2	8.4	91.4	0	0.0	100.0
200	20	44.0	36.3	19.7	0.1	9.4	90.5	0	0.0	100.0
ER										
20	200	97.4	2.6	0.0	58.6	41.4	0.0	20.8	79.2	0.0
20	100	93.5	6.5	0.0	59.4	40.6	0.0	22.4	77.6	0.0
100	20	96.3	3.7	0.0	58.3	41.7	0.0	23.7	76.3	0.0
50	50	94.1	5.1	0.8	7.1	12.2	80.7	2.5	48.4	49.1
200	20	96.7	3.3	0.0	63.0	37.0	0.0	19.5	80.5	0.0
ED										
20	200	53.8	22.4	23.8	1.2	4.8	94.0	0	0.4	99.6
20	100	58.6	22.6	18.8	11.3	14.9	73.8	0	1.8	98.2
100	20	59.8	22.7	17.5	11.4	14.3	74.3	0	1.3	98.7
50	50	51.2	24.4	24.4	3.3	3.7	93.0	0	0.3	99.7
200	20	52.6	22.6	24.8	1.7	5.1	93.2	0	0.0	100.0
NE										
20	200	26.6	47.9	25.5	0.0	12.1	87.9	0	0.2	99.8
20	100	34.8	42.2	23.0	1.1	22.6	76.3	0	0.5	99.5
100	20	12.6	32.9	54.5	0.1	4.7	95.2	0	0.0	100.0
50	50	5.5	27.9	66.6	0.0	0.4	99.6	0	0.0	100.0
200	20	11.9	35.3	52.8	0.0	1.7	98.3	0	0.0	100.0
BCV										
20	200	53.4	28.5	18.1	17.7	25.1	57.2	1.7	13.7	84.6
20	100	70.2	22.3	7.5	30.5	28.9	40.6	1.8	18.3	79.9
100	20	87.2	11.3	1.5	43.8	30.4	25.8	0.8	19.7	79.5
50	50	80.0	17.0	3.0	9.3	28.9	61.8	0.0	4.6	95.4
200	20	81.5	14.6	3.9	17.8	26.1	56.1	0.3	11.3	88.4

Note: $>, =, <$*: overestimation, correct estimation, underestimation, respectively.*

Table 3: Finite-sample performances of those five methods on choosing the number of useful weak factors only with different types of error terms in the factor model. There are six useful weak factors in total to be estimated in the model. REE is the estimation error for recovering the signal matrix defined in (3.5).

		White		Serial		Cross		Lognormal		Gamma		Chi-square	
N	T	REE	$\hat{k}$	REE	$\hat{k}$	REE	$\hat{k}$	REE	$\hat{k}$	REE	$\hat{k}$	REE	$\hat{k}$
IC2													
20	200	0.00	6.0	0.40	10	0.42	10.0	0.01	1.2	0.11	1.0	0.02	7.0
20	100	0.00	6.0	0.24	10	0.43	10.0	0.03	1.4	0.14	1.2	0.19	8.0
100	20	0.00	6.0	0.09	10	0.03	6.1	0.04	1.4	0.14	1.3	0.26	8.0
50	50	0.11	5.4	0.12	10	0.04	5.6	0.04	1.4	0.16	1.0	0.06	6.3
200	20	0.00	6.0	0.08	10	0.00	6.0	0.02	1.2	0.13	1.0	0.09	7.0
ER													
20	200	3.56	0.0	1.94	0.0	2.00	0.0	0.00	1.1	0.12	0.9	1.56	0.0
20	100	3.99	0.0	1.77	0.1	2.51	0.0	0.00	1.1	0.18	0.6	2.04	0.0
100	20	4.12	0.0	1.37	0.7	3.79	0.0	0.01	1.1	0.18	0.7	2.28	0.0
50	50	0.99	3.9	0.48	2.8	1.17	3.4	0.01	1.1	0.16	1.0	1.16	1.8
200	20	3.63	0.0	1.15	0.9	3.53	0.0	0.00	1.0	0.14	0.9	1.79	0.0
ED													
20	200	0.00	6	1.77	0.8	1.80	0.7	0.02	1.4	0.10	1.2	0.35	5.4
20	100	0.00	6	1.60	1.1	2.13	1.0	0.03	1.4	0.13	1.1	0.81	4.1
100	20	0.04	6	0.47	6.5	0.87	4.6	0.04	1.5	0.14	1.2	1.21	3.4
50	50	0.00	6	0.85	1.5	1.88	2.8	0.02	1.4	0.17	1.1	0.93	3.5
200	20	0.00	6	0.40	5.9	0.14	5.8	0.02	1.4	0.12	1.3	0.61	4.7
NE													
20	200	0.13	5.2	0.24	8.2	0.15	7.1	0.02	1.3	0.10	1.1	0.03	6.3
20	100	0.15	5.2	0.18	8.8	0.04	6.2	0.03	1.4	0.12	1.2	0.05	6.1
100	20	0.00	6.0	0.10	10.8	0.01	6.0	0.06	1.8	0.13	1.7	0.10	7.0
50	50	0.09	5.5	0.25	14.9	0.04	6.1	0.04	1.8	0.12	1.9	0.02	6.9
200	20	0.00	6.0	0.11	10.3	0.00	6.0	0.03	1.6	0.11	1.4	0.08	6.8
BCV													
20	200	0.19	5.7	0.22	7.8	0.23	6.8	0.05	1.5	0.16	2.4	0.18	6.2
20	100	0.24	5.4	0.21	9.1	0.21	6.7	0.05	1.4	0.13	2.1	0.23	5.7
100	20	0.22	5.3	0.10	10.8	0.20	6.7	0.05	1.1	0.22	1.7	0.28	5.0
50	50	0.11	5.5	0.20	12.6	0.23	6.8	0.07	1.3	0.36	2.1	0.12	5.8
200	20	0.18	5.7	0.10	12.3	0.24	6.8	0.05	1.3	0.37	2.3	0.19	6.3

Table 4: Finite-sample performances of those five methods on choosing the number of mixed strong and useful weak factors with different types of error terms in the factor model.There are six factors to be estimated in the model with one strong factor and five useful weak factors. REE is the estimation error for recovering the signal matrix defined in (3.5).

		White		Serial		Cross		Lognormal		Gamma		Chi-square	
N	T	REE	$\hat{k}$	REE	$\hat{k}$	REE	$\hat{k}$	REE	$\hat{k}$	REE	$\hat{k}$	REE	$\hat{k}$
IC2													
20	200	0.00	6.0	0.40	10	0.42	10.0	0.01	1.2	0.11	1.0	0.02	7.0
20	100	0.00	6.0	0.24	10	0.43	10.0	0.03	1.4	0.14	1.2	0.19	8.0
100	20	0.00	6.0	0.09	10	0.03	6.1	0.04	1.4	0.14	1.3	0.26	8.0
50	50	0.11	5.4	0.12	10	0.04	5.6	0.04	1.4	0.16	1.0	0.06	6.3
200	20	0.00	6.0	0.08	10	0.00	6.0	0.02	1.2	0.13	1.0	0.09	7.0
ER													
20	200	2.39	1	1.31	1	2.45	1	0.07	1.0	0.19	1.8	1.01	1
20	100	2.79	1	1.18	1	2.45	1	0.05	1.1	0.14	1.7	1.33	1
100	20	2.90	1	0.89	1	2.44	1	0.06	1.1	0.23	1.7	1.57	1
50	50	3.19	1	0.54	1	2.43	1	0.09	1.1	0.34	1.9	1.22	1
200	20	2.60	1	0.57	1	2.43	1	0.07	1.0	0.36	1.8	1.18	1
ED													
20	200	0.00	6	1.18	1.5	0.34	5.1	0.03	1.8	0.21	2.0	0.28	5.3
20	100	0.03	6	1.07	2.2	0.25	5.4	0.05	1.6	0.16	1.8	0.67	3.9
100	20	0.00	6	0.27	7.0	0.35	5.2	0.05	1.7	0.24	1.9	0.63	4.7
50	50	0.00	6	0.47	1.9	0.20	5.5	0.06	2.2	0.37	2.1	0.57	4.1
200	20	0.00	6	0.31	6.1	0.25	5.4	0.04	1.9	0.40	2.1	0.34	5.5
NE													
20	200	0.13	5.2	0.24	8.2	0.15	7.1	0.02	1.3	0.10	1.1	0.03	6.3
20	100	0.15	5.2	0.18	8.8	0.04	6.2	0.03	1.4	0.12	1.2	0.05	6.1
100	20	0.00	6.0	0.10	10.8	0.01	6.0	0.06	1.8	0.13	1.7	0.10	7.0
50	50	0.09	5.5	0.25	14.9	0.04	6.1	0.04	1.8	0.12	1.9	0.02	6.9
200	20	0.00	6.0	0.11	10.3	0.00	6.0	0.03	1.6	0.11	1.4	0.08	6.8
BCV													
20	200	0.19	5.7	0.22	7.8	0.23	6.8	0.05	1.5	0.16	2.4	0.18	6.2
20	100	0.24	5.4	0.21	9.1	0.21	6.7	0.05	1.4	0.13	2.1	0.23	5.7
100	20	0.22	5.3	0.10	10.8	0.20	6.7	0.05	1.1	0.22	1.7	0.28	5.0
50	50	0.11	5.5	0.20	12.6	0.23	6.8	0.07	1.3	0.36	2.1	0.12	5.8
200	20	0.18	5.7	0.10	12.3	0.24	6.8	0.05	1.3	0.37	2.3	0.19	6.3

Table 5: Finite-sample performances of those five methods on choosing the number of mixed strong and useful weak factors with different types of error terms in the factor model.There are six factors to be estimated in the model with three strong factor and three useful weak factors. REE is the estimation error for recovering the signal matrix defined in (3.5).

		White		Serial		Cross		Lognormal		Gamma		Chi-square	
N	T	REE	k_hat	REE	k_hat	REE	k_hat	REE	k_hat	REE	k_hat	REE	k_hat
IC2													
20	200	0.04	6.0	0.42	10	0.14	4.8	0.07	4.5	0.25	4.0	0.09	7.1
20	100	0.03	6.0	0.32	10	0.15	4.8	0.11	4.9	0.23	4.5	0.16	7.9
100	20	0.00	6.0	0.13	10	0.18	4.7	0.13	4.9	0.51	4.8	0.22	7.7
50	50	0.11	5.4	0.36	10	0.13	4.8	0.05	4.2	0.44	4.0	0.04	6.5
200	20	0.00	6.0	0.23	10	0.15	4.8	0.03	4.2	0.48	4.0	0.11	7.1
ER													
20	200	4.77	2	3.34	2	0.88	3	0.43	1.0	0.84	1.8	2.66	2
20	100	4.81	2	2.87	2	0.89	3	0.40	1.1	0.84	1.8	2.82	2
100	20	5.28	2	1.87	2	0.90	3	0.38	1.0	1.04	1.7	3.12	2
50	50	1.24	3	0.02	3	0.89	3	0.49	1.2	0.41	3.8	0.24	3
200	20	5.30	2	1.95	2	0.89	3	0.41	1.0	0.96	1.8	2.83	2
ED													
20	200	0.04	6.0	0.32	3.5	0.16	5.4	0.12	3.5	0.29	4.0	0.07	5.8
20	100	0.04	5.9	0.33	4.2	0.12	5.4	0.19	2.6	0.40	3.5	0.17	5.2
100	20	0.01	6.0	0.12	7.3	0.13	5.6	0.20	2.8	0.53	3.7	0.24	5.1
50	50	0.01	6.0	0.10	4.0	0.15	5.5	0.23	3.0	0.45	4.2	0.13	5.1
200	20	0.00	6.0	0.19	7.3	0.14	5.5	0.12	3.4	0.46	4.0	0.13	5.9
NE													
20	200	0.08	5.2	0.24	7.9	0.01	6.0	0.02	4.2	0.22	4.0	0.05	6.3
20	100	0.10	5.3	0.24	9.1	0.04	6.1	0.04	3.8	0.20	4.0	0.05	6.2
100	20	0.00	6.0	0.16	10.9	0.05	6.1	0.07	4.7	0.40	4.3	0.13	7.0
50	50	0.07	5.6	0.51	14.8	0.01	6.0	0.08	4.7	0.47	4.2	0.03	7.0
200	20	0.00	6.0	0.25	10.8	0.02	6.1	0.03	4.3	0.46	4.2	0.09	6.9
BCV													
20	200	0.19	5.7	0.22	7.8	0.23	6.8	0.05	1.5	0.16	2.4	0.18	6.2
20	100	0.24	5.4	0.21	9.1	0.21	6.7	0.05	1.4	0.13	2.1	0.23	5.7
100	20	0.22	5.3	0.10	10.8	0.20	6.7	0.05	1.1	0.22	1.7	0.28	5.0
50	50	0.11	5.5	0.20	12.6	0.23	6.8	0.07	1.3	0.36	2.1	0.12	5.8
200	20	0.18	5.7	0.10	12.3	0.24	6.8	0.05	1.3	0.37	2.3	0.19	6.3

References

[1] S. Ahn and A. Horenstein. Eigenvalue ratio test for the number of factors. *Econometrica*, 81(3).

[2] D. Amengual and M. W. Watson. Consistent estimation of the number of dynamic factors in a largenandtpanel. *Journal of Business Economic Statistics*, 25(1):91–96, 2007.

[3] T. W. Anderson. *An introduction to multivariate statistical analysis*. Wiley, 2003.

[4] T. W. Anderson and H. Rubin. *Statistical inference in factor analysis*. 1956.

[5] J. Bai and S. Ng. Large dimensional factor analysis. *Foundations and Trends® in Econometrics*, 3(2):89–163, 2008.

[6] J. Bai and S. Ng. Principal components estimation and identification of static factors. *Journal of Econometrics*, 176(1):18–29, 2013.

[7] B. H. Baltagi, C. Kao, and B. Peng. On testing for sphericity with non-normality in a fixed effects panel data model. *Statistics Probability Letters*, 98:123–130, 2015.

[8] M. S. Bartlett. Tests of significance in factor analysis. *British Journal of Statistical Psychology*, 3(2):77–85, 1950.

[9] B. S. Bernanke, J. Boivin, and P. Eliasz. Measuring the effects of monetary policy: A factor-augmented vector autoregressive (favar) approach. *The Quarterly Journal of Economics*, 120(1):387–422, Jan 2005.

[10] R. B. Cattell. The scree test for the number of factors. *Multivariate Behavioral Research*, 1(2):245–276, 1966.

[11] R. B. Cattell and S. Vogelmann. A comprehensive trial of the scree and kg criteria for determining the number of factors. *Multivariate Behavioral Research*, 12(3):289–325, 1977.

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

[13] M. Forni, M. Hallin, M. Lippi, and L. Reichlin. The generalized dynamic-factor model: Identification and estimation. *Review of Economics and Statistics*, 82(4):540–554, 2000.

[14] M. Hallin and R. Liska. Determining the number of factors in the general dynamic factor model. *Journal of the American Statistical Association*, 102(478):603–617, 2007.

[15] J. L. Horn. A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2):179–185, 1965.

[16] I. T. Jolliffe. *Principal component analysis*. Springer, 2011.

[17] H. F. Kaiser. The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20(1):141–151, 1960.

[18] S. Kritchman and B. Nadler. Non-parametric detection of the number of signals: Hypothesis testing and random matrix theory. *IEEE Transactions on Signal Processing*, 57(10):3930–3941, 2009.

[19] D. N. Lawley. Tests of significance for the latent roots of covariance and correlation matrices. *Biometrika*, 43(1/2):128, 1956.

[20] S. C. Ludvigson and S. Ng. The empirical risk return relation: A factor analysis approach. *Journal of Financial Economics*, 83(1):171–222, 2007.

[21] R. Nadakuditi and A. Edelman. Sample eigenvalue based detection of high-dimensional signals in white noise using relatively few samples. *IEEE Transactions on Signal Processing*, 56(7):2625–2638, 2008.

[22] A. Onatski. Determining the number of factors from empirical distribution of eigenvalues. *Review of Economics and Statistics*, 92(4):1004–1016, 2010.

[23] A. Onatski. Asymptotics of the principal components estimator of large factor models with weakly influential factors. *Journal of Econometrics*, 168(2):244–258, 2012.

[24] A. B. Owen and J. Wang. Bi-cross-validation for factor analysis. *Statistical Science*, 31(1):119–139, 2016.

[25] J. Stock and M. Watson. Diffusion indexes. *NBER Working Paper Series*, page 6702, Aug 1998.

[26] J. H. Stock and M. W. Watson. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460):1167–1179, 2002.

[27] J. H. Stock and M. W. Watson. Macroeconomic forecasting using diffusion indexes. *Journal of Business Economic Statistics*, 20(2):147–162, Aug 2002.

Appendix A

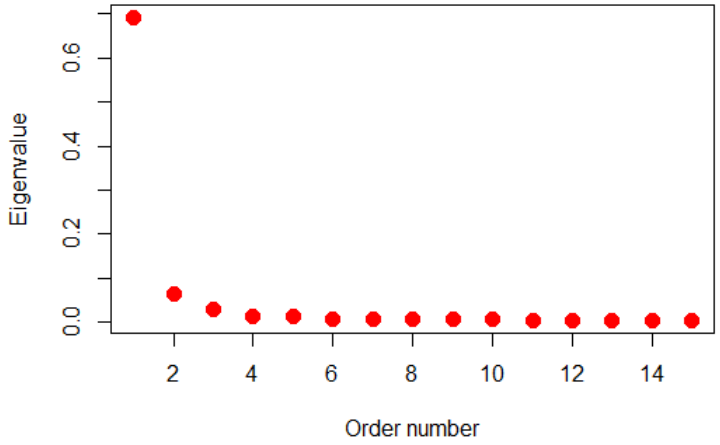
In this section, we show the empirical evidence that weak factors exist in the real finance data and strong factor assumption is not appropriate to assume when both N and T are large, as we mentioned in the section 3.1. That is, the eigenvalues of the sample covariance matrix of the data do not separate into groups of large and small eigenvalues as $N, T \to \infty$. Specifically, we collect the data on daily returns of 100 industrial portfolios from the web site of Kenneth French. The 100 industrial portfolios are formed as the intersection of ten portfolios based on the book to market ratio (BM) and the other ten portfolios based on the market equity (ME). The book to market ratio is the book value divided by market equity. The excess returns are calculated for the period from the Jan.20^{th}, 2015 to Dec.20^{th}, 2018 (T=1000) as follow:

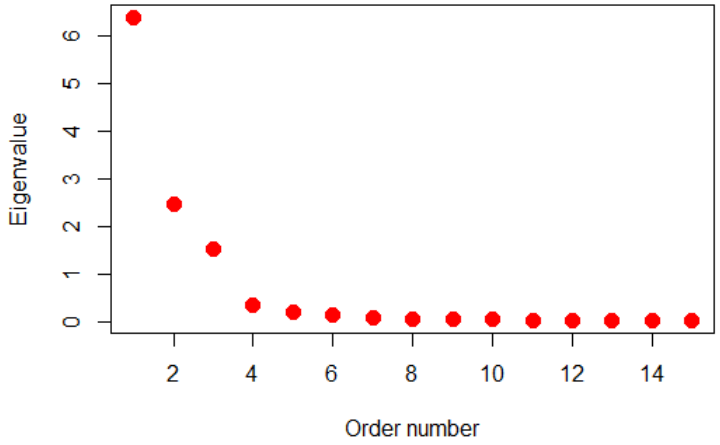
$$\tilde{R}_{it}^{real} = \frac{R_{it} - R_{i,t-1}}{R_{i,t-1}},$$

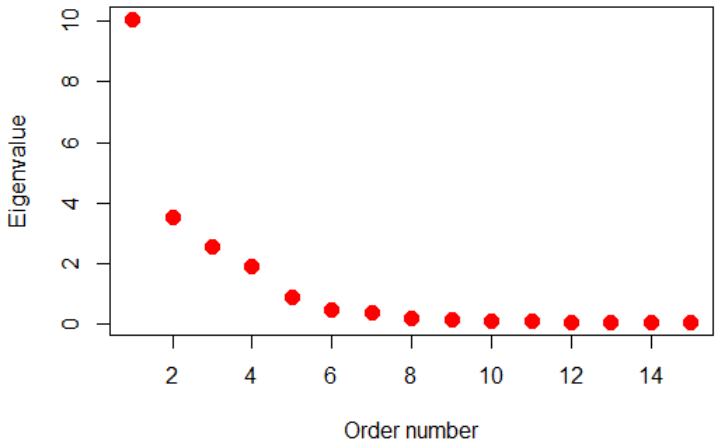
where R_{it} is the average value weighted returns of the portfolios formed on BM and ME.

It is commonly believed that such data contain at least three factors based on the standard Fama-French three-factor model. Hence, the strong factor assumption suggests the existence of a large gap between λ_3 and λ_4 as $N, T \to \infty$, where λ_i is the i-th largest eigenvalue of the sample covariance matrix of the data. However, if we fixed $N = 100$ and let $T = 100, 500, 1000$, Figure 1 clearly shows that except $i = 1$, there are no large gaps between eigenvalues i and $i + 1$ of the sample covariance matrix $\hat{\Sigma}_Y = Y'Y/N$ of the excess return data for $i = 1, 2 \cdots 15$. Also, instead of diverging at the rate $O(T)$, the largest eigenvalue λ_1 of $\hat{\Sigma}_Y$ is bounded between the detection threshold and estimation threshold that defined in RMT, rending it as a weak factor as $N, T \to \infty$. Therefore, the strong factor assumption does not appropriate to assume for this data set.

Figure 1: Top 15 sample covariance eigenvalues of 100 industrial portfolio data. We fixed N=100 and let T = 100, 500, and 1000 from top to below.







Appendix B

In this section, we go over the details of using PCA in factor analysis. As a common statistical method for dimension reduction of the data, principal component analysis (PCA) closely relate to factor analysis. The reason is that the PCA method is often used to estimate the loadings and factors in a factor model. Basically, PCA tries to maximize the sample variance by finding linear combinations of the observed variables. Specifically, let $\hat{\Sigma}_Y = YY'/T$ be the sample covariance matrix corresponding to the population covariance matrix $\Sigma_Y = E(Y_t Y_t')$, assuming $E(Y_i) = 0$ for $i = 1, 2, \cdots, N$. Let P_i be a $(N \times 1)$ independent orthogonal vector such that $P_i' P_i = 1$. Then the variance of $P_i' Y$ is

$$\begin{aligned} \text{Var}(P_i'Y) = E(P_i'Y)^2 &= E[(P_i'Y)(P_i'Y)'] \\ &= P_i' E(YY')P_i, \quad \text{for } i, = 1, 2 \cdots, N. \end{aligned} \tag{6.1}$$

This variance can be estimated by $P_i' \hat{\Sigma}_Y P_i$. To maximize $P_i' \hat{\Sigma}_Y P_i$ subject to $P_i' P_i = 1$, the standard approach is to use the technique of Lagrange multipliers. Maximize

$$P_i' \hat{\Sigma}_Y P_i - \lambda(P_i' P_i - 1), \tag{6.2}$$

where λ is a Lagrange multiplier. Differentiation with respect to P_i gives

$$(\hat{\Sigma}_Y - \lambda I_N) P_i = 0, \tag{6.3}$$

where I_N is the $(N \times N)$ identity matrix. Thus, λ is an eigenvalue of $\hat{\Sigma}_Y$ and P_i is the corresponding eigenvector. To decide which of the N eigenvectors P_i gives $P_i' Y$ with maximum variance, note that $P_i' \hat{\Sigma}_Y P_i = P_i' \lambda P_i = \lambda P_i' P_i = \lambda$, so λ must be as large as possible to maximize the variance of $P_i' Y$. Thus, P_1 is the eigenvector corresponding to the largest eigenvalue of $\hat{\Sigma}_Y$ and $P_1' Y$ is the first principle component (PC). In general, the k-th PC of Y is $P_k' Y$ and $P_k' \hat{\Sigma}_Y P_k = \lambda_k$, where λ_k is the k-th largest eigenvalue of $\hat{\Sigma}_Y$, and P_k is the corresponding eigenvector. Hence, in matrix form, let the eigenvalue decomposition of $\hat{\Sigma}_Y$ be $\hat{\Sigma}_Y = P \Lambda P'$,

where $\Lambda = \text{diag}(\lambda_1, \lambda_2, \cdots, \lambda_N)$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N \geq 0$ and P is an orthogonal $N \times N$ matrix. Then the columns of P are eigenvectors of $\widehat{\Sigma}_Y$, and the rows of $P'Y$ are called principle components (PCs).

Similar with PCA, one can use the singular value decomposition (SVD) of Y to derive eigenvectors and PCs. Let $Y = \sqrt{T}\hat{U}\hat{D}\hat{V}'$ where $\hat{U} \in \mathbb{R}^{N\times\min(N,T)}$, $\hat{V} \in \mathbb{R}^{T\times\min(N,T)}$, and $\hat{D} = \text{diag}(\hat{d}_1, \hat{d}_2, \cdots, \hat{d}_{\min(N,T)})$, with $\hat{U}'\hat{U} = \hat{V}'\hat{V} = I_{\min(N,T)}$, and $\hat{d}_1 \geq \hat{d}_2 \geq, \cdots, \geq \hat{d}_{\min(N,T)}$. Now, it is clear that the first r $(r \leq \min(N,T))$ eigenvectors of $\hat{\Sigma}_Y$ are columns of $\hat{U}$ and the first r principle components are columns of $\sqrt{T}VD$ and $\hat{d}_r^2 = \lambda_r$ for $r = 1, 2, 3, \cdots, \min(N,T)$. For more detail of PCA, one can refer to [16] and [3].

To use PCA in factor analysis, we essentially use the eigenvectors of the sample covariance matrix to estimate the linear space of factors. Then, we can estimate the corresponding loadings by the least square method. Specifically, let r be an arbitrary number such that $0 < r < \min\{N,T\}$, the least square method seeks $L_r = (L_1, L_2, \cdots, L_N)' \in \mathbb{R}^{N\times r}$ and $F_r = (F_1, F_2, \cdots, F_T) \in \mathbb{R}^{r\times T}$ such that

$$V(\hat{L}_r, \hat{F}_r) = \min_{F_r, L_r} \frac{1}{NT} \|Y - L_r F_r\|_F^2, \tag{6.4}$$

subject to the normalization

$$F_r F_r'/T = I_r, \quad \text{and} \quad L_r'L_r \text{ is diagonal for identification,} \tag{6.5}$$

where $\|A\|_F$ is denoted as Frobenius norm for matrix A, defined by $\|A\|_F = \text{tr}^{\frac{1}{2}}(AA')$. The matrix form for factor model (2.1) can be expressed as

$$Y = L_r F_r + e, \tag{6.6}$$

with $Y \in \mathbb{R}^{N\times T}$, $L_r \in \mathbb{R}^{N\times r}$, $F_r \in \mathbb{R}^{r\times T}$, and $e \in \mathbb{R}^{N\times T}$. Hence, for each given F_r, the least squares estimator of L_r is $\hat{L}_r = YF_r'/T$, using the constraint (6.5) on the factors. Substituting

this into equation (6.4), the objective function now becomes

$$\begin{aligned}\|Y - YF_r'F_r/T\|_F^2 &= \text{tr}(Y - YF_r'F_r/T)(Y - YF_r'F_r/T)' \\ &= \text{tr}(YY' - YF_r'F_rY'/T - YF_rF_r'Y'/T + YF_r'F_rF_r'F_rY'/T^2) \\ &= \text{tr}\{Y(I_N - F_r'F_r/T)Y'\},\end{aligned} \tag{6.7}$$

which is identical to maximizing $\text{tr}\{Y(F_r'F_r/T)Y'\} = \text{tr}\{F_r(Y'Y/T)F_r'\}$ due to cyclic property of the trace. In view of the equation (6.1) in matrix form, the estimated factor matrix, denoted by $\hat{F}_r^{\text{pc}'}$, is $\sqrt{T}$ times the eigenvectors corresponding to the r largest eigenvalues of $T \times T$ matrix $Y'Y$. Given $\hat{F}_r^{\text{pc}'}$, $\hat{L}_r^{\text{pc}} = Y\hat{F}_r^{\text{pc}\prime}/T$ is the corresponding matrix of loadings. Note that if we normalize $L_r'L_r/N = I_r$ instead of $F_rF_r'/T = I_r$ and assume F_rF_r' is diagonal in (6.5), then another solution is given by $(\bar{F}_r^{\text{pc}}, \bar{L}_r^{\text{pc}})$, where $\bar{L}_r^{\text{pc}}$ is constructed as $\sqrt{N}$ times the eigenvactors corresponding to the r largest eigenvalues of the $N \times N$ matrix YY'. Given $\bar{L}_r^{\text{pc}}$, $\bar{F}_r^{\text{pc}} = \bar{L}_r^{\text{pc}\prime}Y/N$ is the corresponding matrix of factors. The second set of calculation is computationally less expensive when $N < T$, while the first is less intensive when $T < N$.

Appendix C

In this section, we take the information criteria based method developed by Bai and Ng [**?**] as an example and explain why the strong factor assumption is critical for choosing the number of factors. The mechanism for this method to work is that the discrepancy $V(r)-V(r_0)$ converges to different values for $r < r_0$ and $r > r_0$. More specifically, Bai and Ng [**?**] have shown that $V(r)-V(r_0) \nrightarrow 0$ for $r < r_0$ and $V(r)-V(r_0) \to 0$ for $r > r_0$ at the rate C_{NT}^2 as $N, T \to \infty$. Note that in (3.2), the loss function $V(r)$ is decreasing as r increases while the penalty function $rg(N,T)$ is increasing in r, so IC(r) in (3.3) is minimized by balancing these two functions at the true factor number r_0 asymptotically. Therefore, under the two conditions for the penalty function $g(N,T)$, if $r < r_0$, we have $V(r) - V(r_0) \nrightarrow 0$ and $g(N,T) \to 0$ as $N, T \to \infty$, it is clear that IC(r) is not asymptotically minimized at a $r < r_0$; while if $r > r_0$, we have $V(r) - V(r_0) \to 0$ at the rate C_{NT}^2 and $C_{NT}^2 g(N,T) \to \infty$ as $N, T \to \infty$, then the penalty eventually becomes dominant and overfitting is prohibited.

The reason why strong factor assumption is critical for this method to work is that first, it ensures the PCA or MLE estimators for estimating factor loadings and common factors are consistent as we have shown before; second, from the equation (3.3), we can see that the loss function $V(r)$ in (3.3) is actually the sum of rest eigenvalues of the $T \times T$ matrix $Y'Y$ from λ_{r+1} to λ_T. Since we know that under the strong factor assumption, the top r_0 of the $T \times T$ matrix $Y'Y$ will goes to infinity as $N, T \to \infty$, so $V(r)$ will goes to infinity as $N, T \to \infty$ if $r < r_0$, while it is bounded if $r > r_0$. Hence, the strong factor assumption ensures that the discrepancy $V(r) - V(r_0)$ converges to different values for $r < r_0$ and $r > r_0$.