General Unilateral Loading Estimation

Yaping Li, Guangbao Guo

***Abstract*—Although factor model can extract effective common factors from a large number of data variables, it also meets several concerns in different datasets. For example, its estimation accuracy is not high, and the relationship between the variables and the common factors are difficult to explain. In this paper, a general unilateral loading method is proposed to solve the estimation problems of variance matrix and common factor in general factor model. It can not only explain the relationship between the original variables and the common factors, but also improve the precision of parameter estimation and shorten the estimation time. To evaluate the stability and sensitivity of the proposed method, simulation studies have been conducted. Furthermore, the method has also been applied to real data analysis.**

***Index Terms*—General factor model, parameter estimation, principal component method, unilateral loading.**

I. Introduction

FACTOR model has significant advantages in big data analysis and is widely used in the fields of social sciences and physical sciences. However, when facing financial investment and prediction datasets evenly distributed at the same level, there are still several concerns: the estimation accuracy may be limited, and the time cost of processing financial investment, and prediction datasets are also relatively large.

Due to the complexity of data, as the dimension of the variables increases, the relationship between the original variables and common factors becomes challenging to interpret, and finding meaningful common factors among many variables also becomes difficult. Furthermore, for different types of datasets, using only a single method will reduce the efficiency and accuracy of parameter estimation. Therefore, it is crucial to synthesize the characteristics of different methods and different data sets that are suitable for processing.

This paper proposes a unilateral loading method for general factor model. Firstly, the proposed method has a two-layer structure, and the double loadings are obtained through matrix decomposition, cf. Gao and Tsay. [6] (2021a). It can effectively reduce the difference in the estimation process and improve the estimation accuracy. The advantage of this method is in that case that both the error term and the expected value of the common factor are zero in prediction datasets with the same level of uniform distribution. The proposed method not only helps to gain a deeper understanding of the relationship between the original variables and the extracted common factors, but also offers solid backing for solving real-world problems.

Manuscript received 2 June, 2023; revised 27 October, 2023.

This work was supported by a grant from National Social Science Foundation Project under project ID 23BTJ059, a grant from Natural Science Foundation of Shandong under project ID ZR2020MA022, and a grant from National Statistical Research Program under project ID 2022LY016.

Yaping Li is a postgraduate student of Mathematics and Statistics, Shandong University of Technology, Zibo, China. (e-mail: lyplyplyps-dlg@163.com).

Guangbao Guo is a professor of Mathematics and Statistics, Shandong University of Technology, Zibo, China (Corresponding author to provide phone:15269366362; e-mail: ggb11111111@163.com).

A. General factor model

Assuming that the matrix $X_0 \in R^{n\times p}$ is observed, we standardize the data matrix X_0, and obtain

$$X = \sqrt{n-1}(X_0 - \bar{X}_0)[\mathrm{diag}((X_0 - \bar{X}_0)^\top (X_0 - \bar{X}_0))]^{\frac{-1}{2}}, \tag{1}$$

where

$$\bar{X}_0 = \frac{e_n {e_n}^\top X_0}{n}, e_n = (1, \cdots, 1)^\top,$$

the standardized matrix is

$$X = (X_{1\cdot}^\top, X_{2\cdot}^\top, \cdots, X_{n\cdot}^\top)^\top = (X_{\cdot 1}, X_{\cdot 2}, \cdots, X_{\cdot p}) \in R^{n\times p}.$$

Assuming that general smooth function $g(\cdot)$ on common factor matrix $(n < m)$

$$F_1 = (F_{1,\cdot 1}, F_{1,\cdot 2}, \cdots, F_{1,\cdot m}) \in R^{n\times m}$$

is $g(F_1)$, the general factor model is expressed as

$$X = g(F_1)A_1^\top + V_1, \tag{2}$$

where $A_1 = (a_{ij}) \in R^{p\times m}$ is the factor loading array, a_{ij} are factor loadings;

$$V_1 = (V_{1\cdot 1}, V_{1\cdot 2}, \cdots, V_{1\cdot p})$$

is a special factor of X and $g(\cdot)$ is a known smooth function. For example $g(F_1)$ can be equivalent to F_1, F_1^2, $\log(F_1)$ or $\exp\{F_1\}$, and meet for $j_m = 1, \ldots, m$ and $j = 1, \ldots, p$, an n-by-n unit matrix $I_{n\times n}$,

$$E(g(F_1)_{\cdot j_m}) = E(V_{1,\cdot j}) = 0, \mathrm{cov}(g(F_1)_{\cdot j_m}, V_{1,\cdot j}) = 0,$$

$$\mathrm{var}(g(F_1)_{\cdot j_m}) = I_{m\times m}, \mathrm{var}(V_{1,\cdot j}) = \mathrm{diag}(\sigma_{11}, \ldots, \sigma_{1p}),$$

cf. Bai [10] (2003), Bai and Li [11] (2012) and Gao et al. [8] (2020), for other conditions. The general factor model differs from the factor model in that it makes F_1 and V_1 meet the expectation equal to 0, while the covariance is equal to the condition of the unit array, and the common factor and the error term are uncorrelated, cf. Bai et al. [12] (2013). Then $A_1^T A_1 = I_{m\times m}$, cf. Gao et al. [7] (2021b). Here we constrain not only the load, but also the error term, which makes the general factor model more applicable and more stable than the factor model, and the estimation is also more accurate.

For the problem of dimensionality reduction in factor models, the classical principal component analysis has some drawbacks. For example, it isn't easy to interpret the results of PCA in factor models. and the accuracy of estimation is not high. Fan et al. [3] (2016) proposed a projection principal component (PPCA) method in factor models which eliminates the noise component and can estimate unknown latent factors more accurately. For large datasets updated in real-time, Guo et al. [16] (2023) proposed a new sparse online principal component (SOPC) method for factor models

that can identify sparse solutions by iterative online updating to obtain a consistent and easily interpretable solution. Bai and Ng [9] (2002) proposed a factor model-based principal component method to estimate the number of factors. As the data dimension increases and the cost increases, Fan et al. [1] (2013) proposed a sparse principal component method based on an approximate factor model to study the convergence rate, and Fan et al. [4] (2019) proposed a distributed principal component (FanPC) algorithm based on Heterogeneous factor models, effectively reduce the computing cost of large datasets. Gao and Tsay [6] (2021) proposed a special unilateral loading distributed principal component(GaoPC) method based on a distributed factor model to analyze time series data, and to estimate the load and common factors.

The concerns with the above methods are as follows. First, the maximum eigenvalue of the load is no longer a consistent estimator when the feature dimension of the dataset grows at the same rate or larger than the sample size. Second, in high-frequency datasets, it is very important to increase the constraints on the load. Otherwise, the error of the load estimation will become large. The estimation of loadings in Gao and Tsay [6] (2021) is obtained by selecting the eigenvectors corresponding to the first K eigenvalues of the sample covariance, ignoring the influence of the error terms V_{1t} and V_{2t} on the estimation process. When estimating the load in Fan et al. [4] (2019), it is unnecessary to satisfy the identification conditions.

B. Our work

The main work of this paper is as follows.

Firstly, based on the above concerns, we propose a GulPC method for general factor model. The proposed method focuses on processing i.i.d. non-time series data when considering the impact of error terms on the estimation process to address estimation problems in factor models. The application of this method can provide effective support for improving the estimation accuracy and solving practical problems

Secondly, we investigate the impact of explanatory variables, sample size, and dimensionality on the GulPC. The study found that the estimation accuracy of the GulPC method increased with the increase in sample size; as the dimensions increased, the estimation accuracy of the GulPC also increased. Additionally, we compared the effectiveness of the GulPC with other methods (PCA, PPC, SOPC, FanPC, and GaoPC). Through comparison with other methods, we further verified the effectiveness and advantages of the GulPC.

Ultimately, we found that the proposed method not only improves the efficiency of dimensionality reduction, but also improves the estimation accuracy and significantly reduces the amount of calculation.

II. General unilateral loading in general factor model

A. General unilateral load decomposition

In Equation (2), assuming that there are two layers decomposed, X allows the existence of an underlying factor structure $g(F_1)$, cf. Gao and Tsay. [6] (2021a). First, the second layer factor structure of $g(F_1)$ can be expressed as

$$g(F_1) = g(F){A_2}^\top + V_2. \tag{3}$$

then

$$X = g(F)A_g + E, A_g = {A_2}^\top {A_1}^\top, E = V_2{A_1}^\top + V_1. \tag{4}$$

Here $g(F) = (F_{\cdot 1}, F_{\cdot 2}, \cdots, F_{\cdot p_c}) \in R^{n\times p_c}$ is the general term for the common factor, $A_g \in R^{p_c\times p}$ is the loading of second layer in Equation (4), and

$$V_2 = (V_{2,\cdot 1}, V_{2,\cdot 2}, \cdots, V_{2,\cdot m}) \in R^{n\times m}$$

is the special factor of the second layer in Equation (4). The above steps constitute the decomposition process for general unilateral loading. And it is necessary to meet the following conditions.

$$E(g(F)_{\cdot i_m}) = E(V_{2,\cdot i}) = 0, \mathrm{cov}(g(F)_{\cdot i_m}, V_{2,\cdot i}) = 0,$$
$$\mathrm{var}(g(F)_{\cdot i_m}) = I_{p_c\times p_c}, \mathrm{var}(V_{2,\cdot i}) = \mathrm{diag}(\sigma_{21}, \ldots, \sigma_{2m}),$$

for $i_m = 1, \ldots, p_c$ and $i = 1, \ldots, m$.

B. Estimation of the load and the common composition

Firstly, for loadings A_1 and A_2, we need to meet $A_1A_1^\top = I_{p\times p}$ and $A_2A_2^\top = I_{m\times m}$. Assuming that the sample covariance matrix

$$\hat{\Sigma}_1 = \frac{1}{np}\sum\nolimits_{i=1}^{n}(X - \bar{X}_{i\cdot})^\top(X - \bar{X}_{i\cdot}),$$

there are eigenvalue-eigenvector pairs (λ_i, Q_i) for $i = 1, 2, \cdots, p$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. We extract the first m, m is determined according to the cumulative contribution rate, when the $p-m$ eigenvalues are very small. This matrix $\hat{\Sigma}_1$ can be approximately decomposed, and based on the sample covariance, we can obtain estimates of the load $\hat{A}_1$

$$\begin{aligned}\hat{\Sigma}_1 &= \lambda_{1\cdot 1}Q_{1\cdot 1}{Q_{1\cdot 1}}^\top + \cdots + \lambda_{1\cdot m}Q_{1\cdot m}{Q_{1\cdot m}}^\top \\ &\quad + \lambda_{1\cdot m+1}V_{1\cdot m+1}{Q_{1\cdot m+1}}^\top + \cdots + \lambda_{1\cdot p}Q_{1\cdot p}{Q_{1\cdot p}}^\top \\ &= \lambda_{1\cdot 1}Q_{1\cdot 1}{Q_{1\cdot 1}}^\top + \cdots + \lambda_{1\cdot m}Q_{1\cdot m}{Q_{1\cdot m}}^\top + D_1 \\ &\approx \hat{A}_1\hat{A}_1^\top + \hat{D}_1.\end{aligned} \tag{5}$$

Here $\bar{X}_{i\cdot} = \sum_{i=1}^{n} X_{i\cdot}/n$, the expression of the mean has two forms

$$\bar{X}_{i\cdot} = (\bar{X}_{\cdot 1}, \bar{X}_{\cdot 2}, \ldots, \bar{X}_{\cdot p}), \bar{X}_{\cdot j} = (\bar{X}_{1\cdot}, \bar{X}_{2\cdot}, \ldots, \bar{X}_{n\cdot})^\top.$$

Here $\hat{\Sigma}_1$ is approximated by $S_1^2 = \hat{A}_1\hat{A}_1^\top + \hat{D}_1$, see Fan et al. [2] (2015). Here m is known. Similar to the arguments of Harris [5] (1997).

Thus, the estimated loading matrix $\hat{A}_1$ consists of the first m eigenvalues with the first m eigenvectors.

$$\hat{\Lambda}_1 = \mathrm{diag}(\hat{\lambda}_{1\cdot 1}, \cdots, \hat{\lambda}_{1\cdot k_1}, \cdots, \hat{\lambda}_{1\cdot m}),$$

and $\hat{Q}_1 = Q^\top = (\hat{Q}_{1\cdot 1}, \cdots, \hat{Q}_{1\cdot m})$, can be obtained through the eigenvalue decomposition $\hat{\Sigma}_{X_{i\cdot}} = Q\Lambda Q^{-1}$, and so to

$$\begin{aligned}&\hat{A}_1 = \hat{\Lambda}_1^{1/2}\hat{Q}_1 = (\hat{\Lambda}_{1\cdot 1}^{1/2}\hat{Q}_{1\cdot 1}, \cdots, \hat{\Lambda}_{1\cdot m}^{1/2}\hat{Q}_{1\cdot m}), \\ &\hat{D}_1 = \mathrm{diag}(\hat{\sigma}_1^2, \hat{\sigma}_2^2, \ldots, \hat{\sigma}_m^2), \\ &\hat{\sigma}_i^2 = 1 - \sum_{j=1}^{m}\hat{a}_{ij}^2, i = 1, 2, \cdots, p.\end{aligned} \tag{6}$$

The first layer common factor $g(\hat{F}_1)$ is then obtained by least squares estimation and can be expressed as

$$g(\hat{F}_1) = X\hat{A}_1(\hat{A}_1\hat{A}_1^\top)^{-1}. \tag{7}$$

Next, we define the normalized first-layer common factor $g(\hat{F}_1)^* = g(\hat{F}_1)/\sqrt{m}$, and

$$g(\hat{F}_1)^* = (\hat{F}^*_{1,\cdot 1}, \hat{F}^*_{1,\cdot 2}, \cdots, \hat{F}^*_{1,\cdot m}).$$

The same method was then used to estimate the load A_2 based on

$$E(g(\hat{F}_1)^*_i) = 0, \mathrm{cov}(g(\hat{F}_1)^*_i) = \Sigma_{g(\hat{F}_1)^*},$$
$$\hat{\Sigma}_2 = \frac{1}{n}\sum_{i=1}^{n}(g(\hat{F}_1)^*_i - g(\bar{\hat{F}}_1)^*_i)^\top(g(\hat{F}_1)^*_i - g(\bar{\hat{F}}_1)^*_i). \tag{8}$$

Similarly, we can get that $\hat{\Sigma}_2$ is approximated by

$$S_1^2 = \hat{A}_1\hat{A}_1^\top + \hat{D}_1,$$

where $g(\bar{\hat{F}}_1)^*_{\cdot i} = \frac{1}{m}\sum_{i=1}^{m} g(\hat{F}_1)^*_{\cdot i}$, and write this here because the expression of the mean has two forms

$$g(\bar{\hat{F}}_1)^*_{\cdot i} = (\bar{\hat{F}}^*_{1,\cdot 1}, \ldots, \bar{\hat{F}}^*_{1,\cdot m})$$
$$\text{and } g(\bar{\hat{F}}_1)^*_{\cdot j} = (\bar{\hat{F}}^{*\top}_{1,1\cdot}, \ldots, \bar{\hat{F}}^{*\top}_{1,n\cdot})^\top.$$

Let $E(\hat{\Sigma}_{g(\hat{F}_1)^*}) = \Sigma_{g(F_1)}$, the estimator $\hat{A}_2$ consists of the first p_c maximum eigenvalues and the corresponding p_c eigenvectors, which can be obtained by eigenvalue decomposition $\hat{\Lambda}_2$, $\hat{Q}_2$, then

$$\begin{aligned}\hat{A}_2 &= \hat{\Lambda}_2^{1/2}\hat{Q}_2 = (\hat{\Lambda}_{2\cdot 1}^{1/2}\hat{Q}_{2\cdot 1}, \cdots, \hat{\Lambda}_{2\cdot m}^{1/2}\hat{Q}_{2\cdot m}),\\ \hat{D}_2 &= \mathrm{diag}(\hat{\sigma}_1^2, \hat{\sigma}_2^2, \ldots, \hat{\sigma}_{p_c}^2).\end{aligned} \tag{9}$$

For $\hat{\sigma}_i^2 = 1 - \sum_{j=1}^{p_c}\hat{a}_{ij}^2, i = 1, 2, \cdots, m$. Then

$$\hat{A}_g = \hat{A}_2^\top\hat{A}_1^\top.$$

Firstly, in the process of estimating the loadings by adding constraints on the error terms such that

$$\mathrm{var}(V_{l,\cdot j}) = \mathrm{diag}(\sigma_{l1}, \ldots, \sigma_{lp}), E(V_{l,\cdot j}) = 0$$

for $j = 1, \ldots, p; l = 1, 2$.

We use the principal component method to estimate $\hat{D}_1$ and $\hat{D}_2$, and the loading matrix $\hat{A}_1$ and $\hat{A}_2$, and get the $\hat{A}_g \in R^{p_c\times p}$ and the common factors $g(\hat{F}_1)$ and $g(\hat{F})$ are estimated by the least squares method.

III. Numerical Analysis

A. Preparation

Mean squared error (MSE) is considered in simulation to evaluate the deviation of the true value from the estimators

$$\begin{aligned}&\mathrm{MSE}_{\hat{\Sigma}} = \frac{1}{p^2}\|\hat{\Sigma} - S^2\|_F^2, \mathrm{MSE}_{\hat{A}_g} = \frac{1}{p_c p}\|A_g - \hat{A}_g\|_F^2.\\ &\mathrm{MSE}_{\hat{D}} = \frac{1}{p^2}\|D - \hat{D}\|_F^2.\end{aligned} \tag{10}$$

Let A_g and $\hat{A}_g$ be the real loading matrix and the loading matrix estimator. Both of them satisfy $A_g A_g^\top = I_{p_c\times p_c}$. We define the distance between loading spaces as

$$D(A_g, \hat{A}_g) = \sqrt{1 - \frac{1}{pp_c^2}\mathrm{tr}(\hat{A}_g^\top A_g A_g^\top \hat{A}_g)}. \tag{11}$$

It is easy to see that $D(A_g, \hat{A}_g)$ always takes values in the interval [0, 1]. The smaller the value of $D(A_g, \hat{A}_g)$, the more accurate the estimated load is.

B. Simulation

In this subsection, we verify the stability and sensitivity of the GulPC in general factor models.

1) Stability analysis: We now explore the sample size effect of the proposed method. Fixing (p, m) = (15, 3) and generating X, $A = (a_{ij}) \in R^{p\times m}$, $g(F) \in R^{n\times m}$, $E \in R^{n\times p}$ according to Equation (2), such that n = (800, 1000, 1200, 1400, 1600, 1800, 2000, 2400, 2800, 3000), we examine the trend of the performance of six methods with n, and the line chart of the comparison results is as follows.

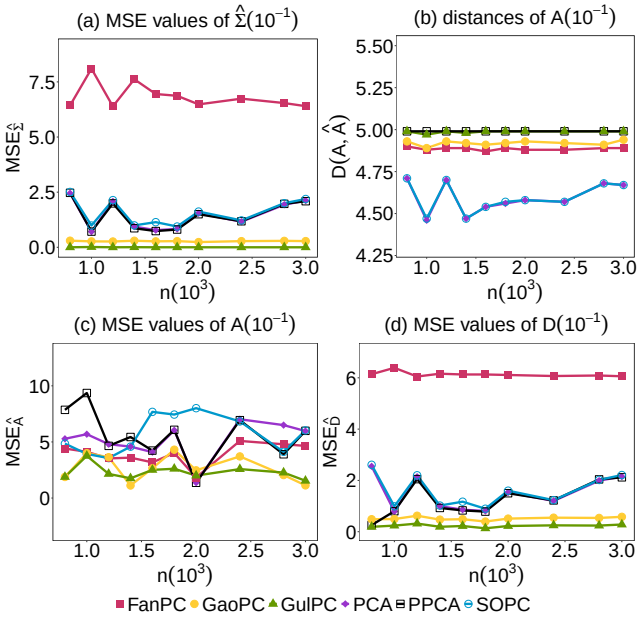

Fig. 1. The performance of the GulPC under different n in simulation

Figure 1 summarizes the comparison results of $\mathrm{MSE}_{\hat{\Sigma}}$, load space distance, $\mathrm{MSE}_{\hat{A}}$ and $\mathrm{MSE}_{\hat{D}}$ of the six methods at different n. From Fig. 1(a), we can see that the $\mathrm{MSE}_{\hat{\Sigma}}$ values of six methods tend to decrease as n increases, which indicates that the estimation accuracy of them is increasing as the sample size increases; from Fig. 1(b), the spatial distance between the true and estimated load values of the PCA and SOPC fluctuates greatly, while the fluctuation range of the other four methods is relatively small, indicating a certain degree of stability; from Fig. 1(c), the $\mathrm{MSE}_{\hat{A}}$ values of the six algorithms fluctuate to some extent, but they generally show a downward trend; from Fig. 1(d), for the $\mathrm{MSE}_{\hat{D}}$ values of the six methods, the estimation accuracy value of FanPC is always the highest, while the estimation accuracy values of PCA, PPCA, and SOPC fluctuate greatly. Only the estimation value of GulPC remains relatively small, with a small variation range and a stable trend.

We also explore the effect of dimension size p on the proposed method. Fixing (n, m) = (1000, 3) and generating X, $A = (a_{ij}) \in R^{p\times m}$, $g(F) \in R^{n\times m}$, $E \in R^{n\times p}$ according to Equation (2), such that p = (7, 9, 11, 13, 15, 17, 19, 21, 23, 25), We test the trend of the performance of six methods with p, and the line chart of the test results is as follows.

Figure 2 summarizes the comparative results of $\mathrm{MSE}_{\hat{\Sigma}}$, load space distance, $\mathrm{MSE}_{\hat{A}}$ and $\mathrm{MSE}_{\hat{D}}$ of six methods in different dimensions p. From Panel (a) of Fig. 2, the $\mathrm{MSE}_{\hat{\Sigma}}$

values of the FanPC, PPCA, PCA, and SOPCA increase with the increase of p, while GaoPC and GulPC show a downward trend overall, indicating that the estimation accuracy continues to improve with the increase of dimension p; from Fig. 2(b), the spatial distance between the true and estimated load values of the PCA and SOPC fluctuates greatly, while the fluctuation range of the other four methods is relatively small, indicating change relatively smoothly; from Fig. 2(c), the $\mathrm{MSE}_{\hat{\mathrm{A}}}$ values of the six methods. Although the $\mathrm{MSE}_{\hat{\mathrm{A}}}$ values of the six methods fluctuate greatly, except for the PPCA which has an increasing trend, the estimated values of the other five methods still tend to decrease. From Figure 2(d), for the $\mathrm{MSE}_{\hat{\mathrm{D}}}$ values of six methods, the estimates of the FanPC are in an upward trend, and the rest of the methods are in a downward trend as a whole, and the estimation of the GulPC is the smallest.

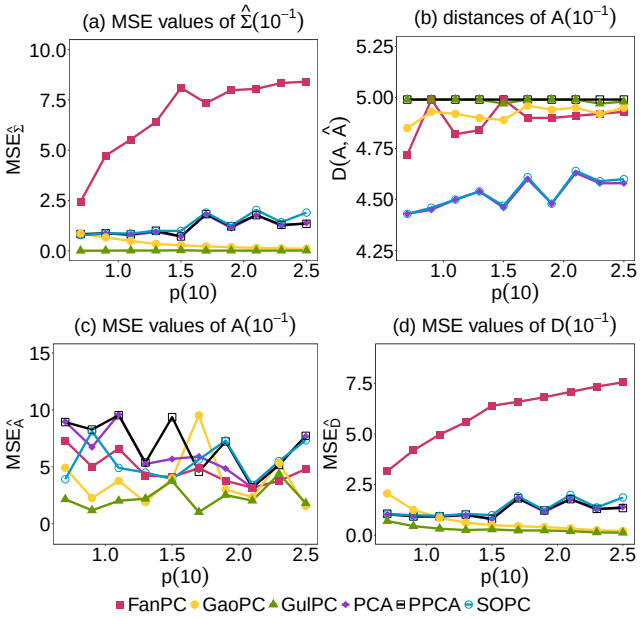

Fig. 2. The performance of the GulPC under different p in simulation

2) Sensibility analysis: We explore the effect of the proposed method. Fixing (n, p) = (1000, 15) and generating X, $A = (a_{ij}) \in R^{p\times m}$, $g(F) \in R^{n\times m}$, $E \in R^{n\times p}$ in (2), for m = (3, 4, 5, 6, 7, 8, 9, 10, 11, 12), we examine the trend of the performance of six methods with m, and the line chart of the examined results is as follows.

Figure 3 summarizes the comparison results of $\mathrm{MSE}_{\hat{\Sigma}}$, load space distance, $\mathrm{MSE}_{\hat{\mathrm{A}}}$ and $\mathrm{MSE}_{\hat{\mathrm{D}}}$ for the six algorithms under different m. From Fig. 3(a), the $\mathrm{MSE}_{\hat{\Sigma}}$ values of the six methods gradually decrease as m increases, indicating that the estimation accuracy is increasing; from Fig. 3(b), the spatial distance between the true and estimated load values of the PCA and SOPC fluctuates greatly, while the fluctuation range of the other four methods is relatively small, indicating change relatively smoothly; from Fig. 3(c), the values of $\mathrm{MSE}_{\hat{\mathrm{A}}}$ of the six methods fluctuate a lot, but the overall trend is still decreasing; from Fig. 3(d), the values of $\mathrm{MSE}_{\hat{\mathrm{D}}}$ of the six methods, of which the FanPC fluctuates more, and the other methods fluctuate very little and tend to be stable. In general, The estimation accuracy of the six methods increases with the increase of m, with GulPC having the best estimation accuracy.

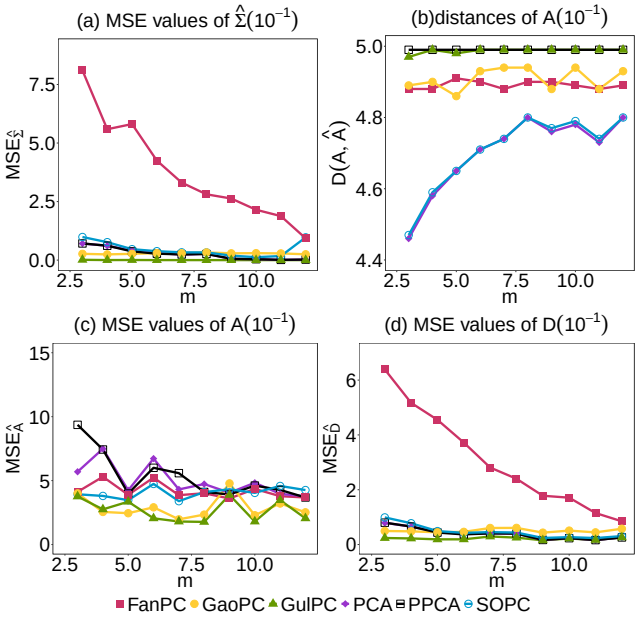

Fig. 3. Sensitivity line graph of GulPC in simulation

C. Real data analysis

In this section, we analyze real datasets to examine the performance of the GulPC by using $\mathrm{MSE}_{\hat{\Sigma}}$ and $\mathrm{MAE}_{\hat{\Sigma}}$. The datasets are as follows.

(1) Riboflavin (RIB) production with the B. subtilis datasets. DSM (Kaiseraugst, Switzerland) (see also Lee et al. [17] (2001) and Zamboni et al. [18] (2005)) has kindly provided these data.

(2) Istanbul Stock Exchange (ISE) datasets. The dataset consists of data for eight complete indices of the Istanbul Slok Exchange from 2009 to 2011.

(3) Stock Portfolio Performance (SPP) datasets. This dataset includes data on the performance of weighted score stock portfolios from the U.S. stock market historical database for the period from 1990 to 2010.

1) RIB datasets: First, we fit the RIB datasets (n, p, m) = (4088, 71, 3) to examine the trends of six methods. Figure4 (a)–(b) summarizes the results of the numerical comparison of $\mathrm{MSE}_{\hat{\Sigma}}$ and $\mathrm{MAE}_{\hat{\Sigma}}$, for the GulPC under RIB datasets. From Panels (a)–(b) of Figure4, the GulPC has the best performance results, GaoPC has the second best performance results, and PCA, PPCA, and FanPC have average estimation accuracy in the medium range, while SOPC has the worst performance results.

2) ISE datasets: Next, we fit the ISE datasets, where (n, p, m) = (536, 8, 5) to test the trend of these six methods. Figure 4(c)–(d), summarizes the comparison results of $\mathrm{MSE}_{\hat{\Sigma}}$ and $\mathrm{MAE}_{\hat{\Sigma}}$, for the GulPC. From Figure 4(c)–(d), GulPC still has the best performance results, followed by PCA, PPCA, SOPC, GaoPC has a medium range of average estimation accuracy, and FanPC has the worst performance results.

3) SPP datasets: Next, we fit the SPP datasets here (n, p, m) = (126, 10, 4) to test the trend of these six methods. Figure 4(e)–(f), summarizes the comparison results of $\mathrm{MSE}_{\hat{\Sigma}}$ and $\mathrm{MAE}_{\hat{\Sigma}}$, for the GulPC. From Figure 4(e)–(f), GulPC still has the best performance results, PCA, PPCA, and GaoPC have the second best performance results, FanPC has

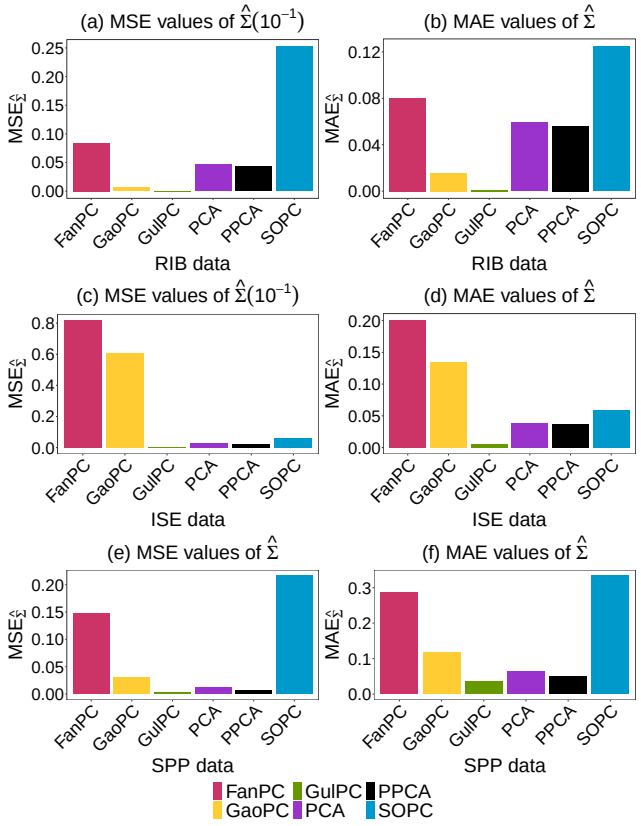

Fig. 4. Line graph comparing the stability of six methods under RIB, ISE, and SPP data

a medium range of average estimation accuracy, and SOPC still has the worst performance results.

IV. Summary and Outlook

With the rapid development of technology and networks, data are becoming more and more diverse. How to effectively deal with these large-scale datasets has attracted extensive attention from researchers. We propose a parameter estimation method for general factor model, to reduce the dimensionality of data and improve the estimation accuracy.

Simulation studies show that the GulPC is more sensitive and stable for estimation, and real data analysis also demonstrates its excellent performance. It is also found in the research that it has higher estimation accuracy and excellent performance. In the future, we will pay more attention to theoretical support and conduct in-depth research on the convergence theories of it.

References

[1] J. Fan, Y. Liao, and M. Mincheva, "Large covariance estimation by thresholding principal orthogonal complements," *Journal of the Royal Statistical Society: Series (Statistical Methodology)*, vol. 75, no. 4, pp. 603-680. 2013.

[2] J. Fan, Y. Liao, and X. Shi, "Risks of large portfolios," *Journal of Econometrics*, vol. 186, no. 2, pp. 367-387. 2015.

[3] J. Fan, Y. Liao, and W. Wang, "Projected principal component analysis in factor models," *Annals of Statistics*, vol. 44, no. 1, pp. 219-254. 2016.

[4] J. Fan, D. Wang, K. Wang and Z. Zhu, "Distributed estimation of principal eigenspaces," *Annals of Statistics*, vol. 47, no. 6, pp. 3009-3031. 2019.

[5] D. Harris, "Principal components analysis of cointegrated time series," *Econometric Theory*, vol. 13, no. 4, pp. 529-557. 1997.

[6] Z. Gao, and R. Tsay, "Divide-and-Conquer: A Distributed Hierarchical Factor Approach to Modeling Large-Scale Time Series Data," arXiv: 2103.14626.

[7] Z. Gao, C. Yuan, B. Jing, H. Wei, and J. Guo, "A two-way factor model for high-dimensional matrix data," arXiv: 2103.07920.

[8] Z. Gao, and R. Tsay, "Modeling high-dimensional time series: a factor model with dynamically dependent factors and diverging eigenvalues," *Journal of the American Statistical Association*, vol. 117, no. 539, pp. 1398-1414. 2022.

[9] J. Bai, and S. Ng, "Determining the number of factors in approximate factor models," *Econometrica*, vol. 70, no. 1, pp. 191-221. 2002.

[10] J. Bai, "Inferential theory for factor models of large dimensions," *Econometrica*, vol. 71, no. 1, pp. 135-171. 2003.

[11] J. Bai, and Y. Li, "Statistical analysis of factor models of high dimension," *Annals of Statistics*, vol. 40, no. 1, pp. 436-465. 2012.

[12] J. Bai, and S. Ng, "Principal components estimation and identification of the factors," *Journal of Econometrics*, vol. 176, no. 1, pp. 18-29. 2013.

[13] G. Guo, C. Wei, and G. Qian, "Sparse online principal component analysis for parameter estimation in factor model," *Computational Statistics*, vol. 38, no. 2, pp. 1095-1116. 2022.

[14] G. Guo, Y. Sun and X. Jiang, "A partitioned quasi-likelihood for distributed statistical inference,"*Computational Statistics*, vol. 35, no. 4, pp. 1577–1596, 2020.

[15] G. Guo, H. Song, and L. Zhu, "ISR: The Iterated Score Regression-Based Estimation Algorithm", 2022.

[16] G. Guo, C. Wei, and G. Qian, "SOPC: The Sparse Online Principal Component Estimation Algorithm," 2022.

[17] M. Lee, S. Zhang, S. Saha, S. Anna, C. Jiang, J. Perkins, J, "RNA expression analysis using an antisense Bacillus subtilis genome array," *Bacteriol*, vol. 183, no. 2, pp. 7371-80. 2002.

[18] N. Zamboni, E. Fischer, A. Muffler, M. Wyss, PM. Hohmann, U. Sauer, "Transient expression and flux changes during a shift from high to low riboflavin production in continuous cultures of Bacillus subtilis," *Biotechnol*, vol. 89, no. 2, pp. 219-32. 2005.