Comparison and Empirical Analysis between Principal Component Analysis and Factor Analysis

Lu Xu[1] Yingying Zhang[1]

[1] Department of Statistics and Actuarial Science, College of Mathematics and Statistics, Chongqing University, China

Abstract

We carry out an empirical analysis based on the study of the principal component analysis and the factor analysis, the data is in the 2012 China Statistical Yearbook[1] with 31 provinces and 9 major economic indicators. First by comparing three kinds of methods of factor analysis, we find out that the principal factor analysis method has the minimum sum of squared errors. Then we use the principal component analysis and the principal factor analysis to extract two principal components and two principal factors from the economic indicators. Third, we calculate principal component scores, factor scores, and their comprehensive scores of the provinces under two methods. Finally, we find that the rankings of the provinces under the two methods are different, but the overall trends are consistent, and the trends are also consistent with the actual situations.

Keywords: Principal component analysis; factor analysis; R software; China Statistical Yearbook.

1. Introduction

In scientific research or daily life, we often need to judge something, such as superiority and its law of development and so on. And the factors influencing the characteristics and law of development of things (indicators) are various. Therefore, in a study of things, to fully and accurately reflect the characteristics of it and its law of development, we should not only evaluate them from a single index or aspect. In contrast, we should take into account the related factors in many aspects. Multivariate large sample data undoubtedly provides the researchers or policy makers with a lot of valuable information. But in the analysis of multivariate problems, often there are correlations between the variables, so the information reflected by the observations exists overlapping phenomenon. Thus to avoid overlapping information and reduce the workload, we hope to be able to find out a few unrelated variables as much as possible to reflect the original data for the most part of information. Principal component analysis (PCA) and factor analysis (FA) are two multivariate statistical analysis methods[2] to solve such a problem.

1.1. Symbol explanations

X_1: GDP (￥100 million);

X_2: GDP per capita (￥100 million / 10,000 persons);

X_3: Fixed asset investments (￥100 million);

X_4:Total imports and exports (\$10,000);

X_5: RMB deposits (￥100 million);

X_6: Citizen consumption level (￥);

X_7: Total retail sales of social consumer goods (￥100 million);

X_8: Total output values of regional construction industry (￥10,000);

X_9: Fiscal income (￥100 million).

2. Statistical methods[3]

2.1. Principal component analysis

In practice, the population covariance matrix $\boldsymbol{\Sigma}$ is usually unknown, the sample principal components will be calculated from the sample correlation matrix $\hat{\boldsymbol{R}}$. Suppose $\lambda_1^* \geq \lambda_2^* \geq \cdots \geq \lambda_p^* \geq 0$ be the eigenvalues of the sample correlation matrix $\hat{\boldsymbol{R}}$, $\boldsymbol{a}_1^*, \boldsymbol{a}_2^*, \cdots, \boldsymbol{a}_p^*$ be the corresponding orthonormal eigenvectors. Let $\boldsymbol{Q}^* = \left(\boldsymbol{a}_1^*, \boldsymbol{a}_2^*, \cdots, \boldsymbol{a}_p^*\right)$, then

$$\boldsymbol{Q}^{*\mathrm{T}} \hat{\boldsymbol{R}} \boldsymbol{Q}^* = \boldsymbol{\Lambda}^* = \begin{pmatrix} \lambda_1^* & & & \\ & \lambda_2^* & & \\ & & \ddots & \\ & & & \lambda_p^* \end{pmatrix}.$$

In addition, $\hat{\boldsymbol{R}}$ owns a spectrum decomposition:

$$\begin{aligned} \hat{\boldsymbol{R}} &= \boldsymbol{Q}^* \boldsymbol{\Lambda}^* \boldsymbol{Q}^{*\mathrm{T}} \\ &= \left(\boldsymbol{a}_1^*, \boldsymbol{a}_2^*, \cdots, \boldsymbol{a}_p^*\right) \begin{pmatrix} \lambda_1^* & & & \\ & \lambda_2^* & & \\ & & \ddots & \\ & & & \lambda_p^* \end{pmatrix} \begin{pmatrix} \boldsymbol{a}_1^{*\mathrm{T}} \\ \boldsymbol{a}_2^{*\mathrm{T}} \\ \vdots \\ \boldsymbol{a}_p^{*\mathrm{T}} \end{pmatrix} \\ &= \sum_{i=1}^{p} \lambda_i^* \boldsymbol{a}_i^* \boldsymbol{a}_i^{*\mathrm{T}}. \end{aligned}$$

If we select m principal components, then the residual matrix $\boldsymbol{E} = \left(e_{ij}\right)_{p \times p}$ of the principal components is

$$\boldsymbol{E} = \hat{\boldsymbol{R}} - \sum_{i=1}^{m} \lambda_i^* \boldsymbol{a}_i^* \boldsymbol{a}_i^{*\mathrm{T}}. \qquad (1)$$

Moreover, the sum of the squared elements of the residual matrix (briefly known as the sum of squared errors) $Q_0(m)$ is

$$Q_0(m) = \sum_{i=1}^{p} \sum_{j=1}^{p} e_{ij}^2. \qquad (2)$$

2.2. Factor analysis

The expression of factor analysis matrix is:

$$\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{F} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{X} = (X_1, X_2, \cdots, X_p)^{\mathrm{T}}$ is the observable random vector, and $E(\boldsymbol{X}) = \boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_p)^{\mathrm{T}}$, $\mathrm{var}(\boldsymbol{X}) = \boldsymbol{\Sigma} = (\sigma_{ij})_{p \times p}$, $\boldsymbol{A} = (a_{ij})_{p \times m}$ is the factor loading matrix, $\boldsymbol{F} = (f_1, f_2, \cdots, f_m)^{\mathrm{T}}$ is the common factor vector, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_p)^{\mathrm{T}}$ is the special factor vector. We usually assume

$$E(\boldsymbol{F}) = \boldsymbol{0}, \ \mathrm{var}(\boldsymbol{F}) = \boldsymbol{I}_m,$$

$$E(\boldsymbol{\varepsilon}) = \boldsymbol{0}, \ \mathrm{var}(\boldsymbol{\varepsilon}) = \boldsymbol{D} = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \cdots, \sigma_p^2),$$

$$\mathrm{cov}(\boldsymbol{F}, \boldsymbol{\varepsilon}) = \boldsymbol{0}.$$

As seen from the assumptions above, the common factors are unrelated with each other and have unit covariance matrix, the special factors are unrelated with each other and are unrelated with the common factors.

Starting from sample correlation matrix $\hat{\boldsymbol{R}}$, we need to estimate the factor loading matrix $\hat{\boldsymbol{A}}^* = (\hat{a}_{ij}^*)_{p \times m}$ and the special variance matrix $\hat{\boldsymbol{D}}^* = \mathrm{diag}(\hat{\sigma}_1^{*2}, \hat{\sigma}_2^{*2}, \cdots, \hat{\sigma}_p^{*2})$ in order to set up a factor model. There are three usual methods used for parameter estimation: the principal component analysis

(pca) method, the principal factor analysis (pfa) method, and the maximum likelihood estimation (mle) method. The residual matrix $\boldsymbol{E} = \left(e_{ij}\right)_{p\times p}$ of factor analysis is

$$\boldsymbol{E} = \hat{\boldsymbol{R}} - \left(\hat{\boldsymbol{A}}^{*}\hat{\boldsymbol{A}}^{*\mathrm{T}} + \hat{\boldsymbol{D}}^{*}\right). \qquad (3)$$

The sum of squared errors $Q(m)$ is

$$Q(m) = \sum_{i=1}^{p}\sum_{j=1}^{p} e_{ij}^{2}. \qquad (4)$$

It is proved that

$$Q(m) \le \lambda_{m+1}^{2} + \cdots + \lambda_{p}^{2}.$$

When we choose the number of factors (m) properly, the $Q(m)$ will be very small.

3. Empirical analysis

R software is used for computing in this section.

3.1. Sum of squared errors and scree plot

The sum of squared errors using (1), (2), (3), and (4) are listed in Table 1.

Table 1. The sum of squared errors of the principal component analysis and the factor analysis methods.

Methods	Principal component analysis	Factor analysis		
		pca	pfa	mle
Sum of squared errors	0.3331946	0.15648377	0.09875134	0.12203535

Due to different calculating methodologies between the principal component analysis and the factor analysis, the results can not be compared directly. In factor analysis, the sum of squared errors of the principal factor analysis method is the least, so we can compare the results of the principal factor analysis method and the principal component analysis. The scree plots of the principal component analysis and the principal factor analysis method are shown in Figures 1 and 2, respectively. From the figures we see that they are close.

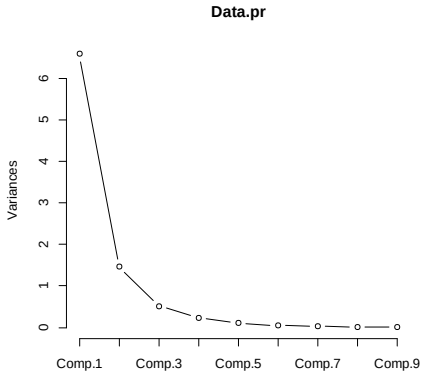

Figure 1. The scree plot of the principal component analysis.

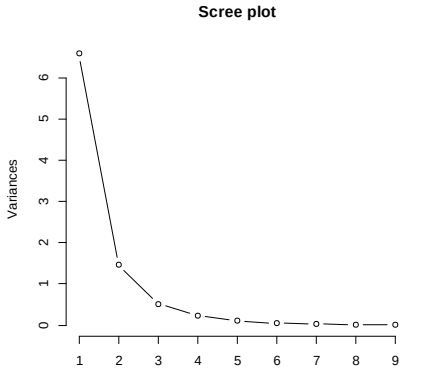

Figure 2. The scree plot of the principal factor analysis method.

3.2. Explanations of loading matrices and principal components/factors

The loading matrices of the principal component analysis and the principal factor analysis method are show in Table 2.

Table 2. The loading matrices of the principal component analysis and the principal factor analysis method.

variable	PCA $\boldsymbol{Q}^{*}$		FA $\hat{\boldsymbol{A}}^{*}$	
	Z_1^{*}	Z_2^{*}	F_1^{*}	F_2^{*}
X_1^{*}	-0.372	-0.224	0.983	0.219
X_2^{*}	-0.242	0.593	0.234	0.821

X_3^*	-0.309	-0.399	0.881	0.002
X_4^*	-0.338		0.703	0.460
X_5^*	-0.373		0.882	0.375
X_6^*	-0.250	0.615	0.179	1.041
X_7^*	-0.371	-0.212	0.968	0.232
X_8^*	-0.328		0.731	0.341
X_9^*	-0.383		0.846	0.511

The cumulative contribution rate of the first two principle components already reaches 89.5%, thus the rest can be omitted to reduce the dimensionality. The loading values of the first principal component is around -0.3, which reflects the regional economic capability, so this component is the economic strength principal component, the smaller this value, the stronger the regional economic capability. The second component has positive correlation with X_2 (GDP per capita) and X_6 (citizen consumption level), but negative correlation with X_1 (GDP), X_3 (fixed asset investments), and X_7 (total retail sales of social consumer goods), which indicates the advancement level, so this component is the advancement level principal component, the bigger this value, the more advanced this region.

The cumulative contribution rate of the first two common factors reaches 87.5%, and the factors have practical significances. Among the first common factor, variables with big absolute values are: X_1 (GDP), X_3 (fixed asset investments), X_4 (total imports and exports), X_5 (RMB deposits), X_7 (total retail sales of social consumer goods), X_8 (total output values of regional construction industry), and X_9 (fiscal income), these variables show the economic aggregate in a region, so the first common factor is the economic aggregate factor, the bigger this value, the bigger the economic aggregate in this region. Among the second common factor, variables with big absolute values are: X_2 (GDP per capita), X_6 (citizen consumption level), these variables reflect average wealth per capita, so the second common factor is the average rich-poor factor, the bigger this value, the richer the average in this region.

3.3. Scores and rankings

The scores of the first two principal components are calculated as follows:

$$Z_1^* = -0.372X_1^* - 0.242X_2^* - 0.309X_3^* - 0.338X_4^* - 0.373X_5^* - 0.250X_6^* - 0.371X_7^* - 0.328X_8^* - 0.383X_9^*,$$

$$Z_2^* = -0.224X_1^* + 0.593X_2^* - 0.399X_3^* + 0.615X_6^* - 0.212X_7^*.$$

Using the factor score formula $\boldsymbol{F}^* = \hat{\boldsymbol{A}}^{*\mathrm{T}}\hat{\boldsymbol{R}}^{-1}\boldsymbol{X}^*$ in regression method, we get the scores for the first two common factors.

$$F_1^* = 8.043X_1^* + 0.307X_2^* - 3.822X_3^* - 2.731X_4^* - 0.867X_5^* - 1.148X_6^* - 2.113X_7^* - 0.164X_8^* + 2.142X_9^*,$$

$$F_2^* = -1.576X_1^* - 1.136X_2^* + 1.731X_3^* + 1.043X_4^* - 0.126X_5^* + 2.673X_6^* - 0.097X_7^* - 0.052X_8^* - 1.063X_9^*.$$

The scatter plots of the principal component analysis and the principal factor analysis method are shown in Figures 3 and 4, respectively.

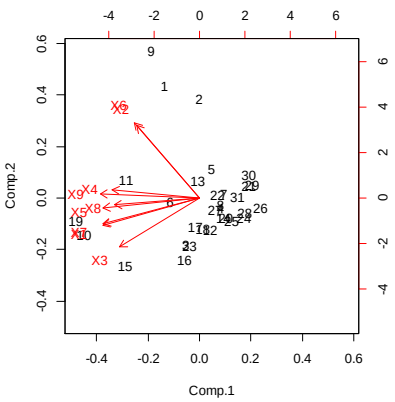

Figure 3. The scatter plot of the first two principal scores of the principal component analysis.

From Figure 3 we see the categorization of variables: X_2 (GDP per capita) and X_6 (citizen consumption level) form one category, and the other variables form another category. Regarding the first principal component, 19 (Guangdong) and 10 (Jiangsu) have smaller values, which means the two provinces are stronger in economic strength; while 26 (Tibet), 29 (Qinghai) have larger values, indicating their weaker economy strength. As to the second principal component, smaller ones are 15 (Shandong), 16 (Henan), 23 (Sichuan), and 3 (Hebei), so these provinces are less advanced; while 9 (Shanghai), 1 (Beijing), and 2 (Tianjin) have larger values, suggesting they are more advanced.

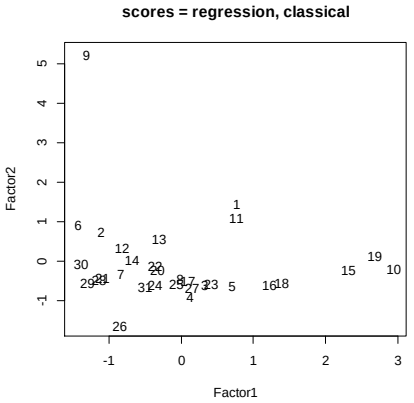

Figure 4. The scatter plot of the first two factor scores of the principal factor analysis method.

Regarding the first factor in Figure 4, 6 (Liaoning) and 30 (Ningxia) have smaller values, indicating lower aggregate of economy; 10 (Jiangsu) and 19 (Guangdong) have bigger values, indicating larger aggregate of economy. Regarding the second factor, 26 (Tibet) has a smaller value, which means lower wealth per capita; while 9 (Shanghai) has a bigger value, which means higher wealth per capita.

In principal component analysis, the comprehensive ranking[4,5] is based on the scores from the first two principal components, weighed by the contribution rates and then summed. Because the coefficients of the first principal component are negative, the calculation will use the opposite numbers. In factor analysis, the ranking utilizes the same rule. The rank comparisons between the principal component analysis and the principal factor analysis method are given in Table 3. In the table, P1 is short for the economic strength, P2 is short for the advancement level, P is short for the comprehensive rank of principal components, F1 is short for the economic aggregate, F2 is short for the average rich-poor, and F is short for the comprehensive rank of factors.

Table 3. Rank comparisons between the principal component analysis and the principal factor analysis method.

Province	Principal Component Analysis			Factor Analysis (pfa)		
	P1	P2	P	F1	F2	F
Jiangsu	2	27	2	1	12	1
Guangdong	1	23	1	2	8	2
Shandong	3	31	4	3	13	3
Beijing	6	2	6	6	2	4
Zhejiang	4	6	3	7	3	5
Shanghai	5	1	5	29	1	6
Hunan	14	25	14	4	21	7
Henan	8	30	10	5	24	8
Neimenggu	16	4	15	8	27	9
Sichuan	10	29	11	9	23	10
Hebei	9	28	9	10	26	11
Fujian	12	7	12	16	6	12
Hubei	11	24	13	13	19	13
Shǒ nxi	17	17	17	11	29	14
Heilongjiang	19	14	19	14	17	15
Shō nxi	20	15	20	12	30	16
Yunnan	24	22	24	15	22	17
Chongqing	18	11	18	18	11	18
Guangxi	23	20	23	17	14	19
Anhui	15	26	16	22	7	20
Guizhou	26	21	26	19	25	21
Jiangxi	21	19	22	21	9	22
Tianjin	13	3	8	26	5	23
Xinjiang	25	12	25	20	28	24
Liaoning	7	13	7	31	4	25
Jilin	22	10	21	23	15	26
Hainan	29	9	29	25	16	27
Gansu	27	18	27	27	18	28
Ningxia	28	5	28	30	10	29
Qinghai	30	8	30	28	20	30
Tibet	31	16	31	24	31	31

Although the principle factor analysis method is similar to the principle component analysis, in the arithmetic process, factors are rotated but principal compo-

nents are not. Meanwhile, the sum of squared errors of the principal factor analysis method is smaller, so the final ranking will be based on the principal factor analysis method. The provinces with top comprehensive ranks are Jiangsu, Guandong, Shandong, Beijing, Zhejiang, and Shanghai, and the bottoms are Tibet, Qinghai, Ningxia, Gansu, and Hainan.

4. Conclusions

Both principal component analysis and factor analysis reduce the dimensionality, take out common parts to analyze and process so as to get conclusions.

The differences between these two analysis lie in that factor analysis turns variables into linear combination of factors, while principal component analysis turns principal components into linear combination of variables; the number of factors needs to be chosen by the analyzer, while the number of principal components is equal to the number of variables. Because the factor analysis methods explain factors with the help of rotation, so they are better than the principal component analysis in terms of factor explainations.

The rankings of the principal component analysis and the principal factor analysis method are similar: Jiangsu, Guangdong, Shandong, Beijing, Zhejiang, and Shanghai rank top in the list while Tibet, Qinghai, Ningxia, Gansu, and Hainan rank bottom in the list. Although the exact rankings of the provinces of the two methods are different, the overall trends are the same, and the results conform with the reality. So in this paper there is no better or worse between the two methods. Anyway, we should have clear understandings on these two methods and use them in accordance with different situations so as to make the best of each method in practical analysis.

5. References

[1] The national bureau of statistics of the People's Republic of China. *China Statistical Yearbook* [J]. China Statistics Press, 2012.

[2] Hu Yang, Qiongsun Liu, Bo Zhong. *Mathematical Statistics* [M]. Beijing: Higher Education Press, 2004.

[3] Yi Xue, Liping Chen. *Statistical Modeling and R Software* [M]. Beijing: Tsinghua University Press, 2007.

[4] Haiming Lin, Wenlin Zhang. The similarities and differences between the principal component analysis and factor analysis and SPSS software [J]. *Statistical Research*, 2005 (3).

[5] Baoren Li, Zhenrong Wang. Profitability and capital structure of listed companies in China empirical analysis [J]. *Quantitative and Technical Economics*, 2003 (4).