RCES REVIEW OF COMPUTER ENGINEERING STUDIES

ISSN: 2369-0755 (Print), 2369-0763 (Online)
Vol. 4, No. 2, June 2017, pp. 82-85
DOI: 10.18280/rces.040208

A publication of IIETA

IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

http://www.iieta.org/Journals/RCES

Principal component analysis of income sources of urban households in China

Minghua Wu*, Xiaogang Xia

School of Science, Xi'an University of Science and Technology, Xi'an 710054, China

Email: mhWu_1978@126.com

ABSTRACT

Based on the principal component analysis method of multivariate statistical analysis, this paper constructs various models for the income sources of urban households in China by means of MATLAB and SPSS. The status quo of the income sources of urban households in China is objectively analyzed by adopting the factor analysis to categorize the 31 provinces [1], municipalities and autonomous regions in China by income sources. Moreover, the author analyzes the income sources of and correlations between urban residents in different regions of China in 2015 and draws some useful conclusions. Some rational suggestions are presented to further improve the income of residents.

Keywords: Income Sources of Residents, Principal Component Analysis, Factor Analysis.

1. INTRODUCTION

As a vast and populous country, China faces uneven levels of economic development and huge gaps in the income and expenditure of urban residents in different regions. Under the combined effect of various influencing factors (e.g. technological development, market environment, talent flow, etc.), the gaps are widening at an accelerated rate. Whereas income is the basis of consumption, it is of great importance to conduct an objective, accurate and effective analysis of the income sources of urban households in China. Such an analysis may shed light on the formulation of macro-control policies, improvement of the living standards of residents, and implementation of the Belt and Road Initiative.

2. MATHEMATICAL MODEL AND CALCULATION STEPS OF PRINCIPAL COMPONENT ANALYSIS

2.1 Principal component analysis model

The principal component analysis (PCA) of principal component model is one of the most popular ways to determine factor variables in factor analysis [2]. In the PCA, the p original relevant variables xi are linearly converted into a set of irrelevant variables via coordinate transformation. The transformation process can be expressed as:

$$\begin{cases} F_1 = u_{11}x_1 + u_{21}x_2 \cdots + u_{p1}x_p \\ F_2 = u_{12}x_1 + u_{22}x_2 \cdots + u_{p2}x_p \\ \cdots \\ F_p = u_{1p}x_1 + u_{2p}x_2 \cdots + u_{pp}x_p \end{cases} \tag{1}$$

where $u_{1k} + u_{2k} + \cdots + u_{pk} = 1, (k = 1,2,3,..,p)$; $F_1, F_2, .., F_P$ are the first, second, … and the p-th principal components, respectively. Specifically, F_1 takes up the largest proportion of the total variance and boasts the strongest ability to synthesize the original variables, while the remaining principal components account for increasingly smaller proportions and gradually weakening abilities.

2.2 Steps of the PCA

(1) Data standardization

Let there be $x_{ij}^* = (x_{ij} + \bar{x}_j)/S_j$, where i=1, 2, …, n (n is the number of samples); $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$; j=1, 2, …, p (p is the number of sample variables). For the sake of convenience, we have:

$$p\left[x_{ij}^*\right]_{n\times p} = \left[x_{ij}\right]_{n\times p}$$

(2) Calculate the covariance matrix R of the data $[x_{ij}]_{n\times p}$.

(3) Find the first m eigenvalues of R: $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$, as well as the corresponding eigenvalue vectors $u_1, u_2 \ldots, u_m$.

(4) Find the factor loading matrix of the m variables.

$$A=\begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{p1} & \cdots & a_{pm} \end{pmatrix}=\begin{pmatrix} u_{11}\sqrt{\lambda_1} & \cdots & u_{1m}\sqrt{\lambda_m} \\ \vdots & \ddots & \vdots \\ u_{p1}\sqrt{\lambda_1} & \cdots & u_{pm}\sqrt{\lambda_m} \end{pmatrix}$$

(5) Calculate factor scores

First, express each factor variable as a linear combination of the original variables, i.e.:

$$F_j=\beta_{j1}x_1+\beta_{j2}x_2+\cdots+\beta_{jp}x_p\left(j=1,2,\cdots,m\right) \quad (2)$$

Then, assign different weights to these variables, and run the comprehensive judgment formula below:

$$F=a_1F_1+a_2F_2+\cdots+a_mF_m \quad (3)$$

Thus, the comprehensive score is obtained [3].

3. ANALYSIS OF THE INCOME SOURCES OF URBAN HOUSEHOLDS IN CHINA

According to the net income data in the Per Capita Annual Income of Urban Households in Different Regions 2013, the China Statistical Yearbook, the author selects such four indicators of income sources as y_1 : wage income, y_2 : operating income, y_3 : transfer income, and y_4 : property income. The indicators are processed in SPSS. The outputted results are listed below:

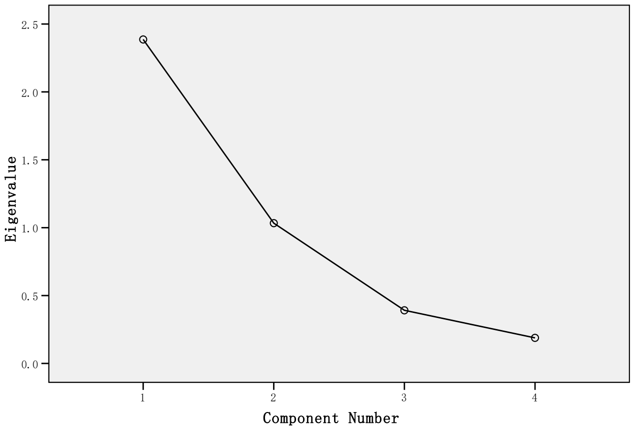
Table 2-1. KMO and Bartlett's test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.545
Bartlett's Test of Sphericity	Approx. Chi-Square	47.400
	df	6
	Sig.	.000

Table 2-2. Factor analysis results (variance explanation)

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Varia nce	Cumulative %	Total	% of Variance	Cumulative %
1	2.386	59.660	59.660	2.386	59.660	59.660	1.739	43.482	43.482
2	1.033	25.835	85.495	1.033	25.835	85.495	1.681	42.013	85.495
3	.392	9.790	95.285						
4	.189	4.715	100.00						

The contents in Table 2-1 demonstrate the applicability of factor analysis to this research [4]. According to the principal components listed in Table 2-2, the first principal component has a characteristic root of 2.386 and a variance contribution rate of 59.66%. The cumulative contribution rate of the first two principal components stands at 85.495%. Two factors are selected considering the factor extraction condition: the eigenvalues should be greater than 1.



As can be seen from the figure, the characteristic roots of the two principal components are greater than 1.

Table 2-3. Factor loading matrix

	Component	
	1	2
y_1	.841	-.391
y_2	.680	.632
y_3	.839	.376
y_4	.716	-.583

The factor expression of each variable is listed below:

Table 2-4. Factor loading after rotation

	Component	
	1	2
y_1	.877	.299
y_2	.054	.927
y_3	.346	.852
y_4	.920	.074

$$y_1=0.841F_1-0.391F_2$$
$$y_2=0.680F_1+0.632F_2$$
$$y_3=0.839F_1+0.376F_2$$
$$y_4=0.716F_1-0.583F_2$$

The factor expression of each variable after rotation is listed below:

$$y_1 = 0.877F_1' + 0.299F_2'$$

$$y_2 = 0.054F_1' + 0.927F_2'$$

$$y_3 = 0.346F_1' + 0.852F_2'$$

$$y_4 = 0.920F_1' + 0.074F_2'$$

Table 2-5. Factor transformation matrix

Component	1	2
1	.722	.692
2	-.692	.722

Table 2-6. Factor score coefficient matrix

	Component	
	1	2
y_1	.516	-.029
y_2	-.217	.639
y_3	.002	.506
y_4	.607	-.200

It can be seen that the first principal factor is mainly determined by the first and fourth variables (i.e. wage income and property income), while the second one is mainly determined by the middle two variables [5] (i.e. operating income and transfer income).

The factor loading matrix after rotation is obtained by multiplying the original factor loading matrix with the factor transformation matrix.

The factor score expressions are obtained as below:

$$F_1' = 0.516y_1 - 0.217y_2 + 0.002y_3 + 0.607y_4$$

$$F_2' = -0.029y_1 + 0.639y_2 + 0.506y_3 - 0.200y_4$$

Table 2-7. Covariance matrix of factor scores

Component	1	2
1	1.000	.000
2	.000	1.000

The above table indicates that the two common factors extracted are not correlated [6].

The regions are ranked as follows by the two different factors:

Table 2-8. Ranking by the score of the first principal component

Index	Score	Index	Score	Index	Score	Index	Score	Index	Score
Beijing	1.1767	Fujian	0.7146	Yunan	0.5234	Guangxi	0.4925	Hainan	0.4748
Shanghai	1.1416	Shandong	0.6483	Hubei	0.5229	Jilin	0.4902	Xinjiang	0.4674
Zhejiang	0.86	Chongqing	0.6224	Hunan	0.5152	Qinghai	0.4873	Xizang	0.4655
Guangdong	0.8128	Liaoning	0.5559	Anhui	0.5092	Jiangxi	0.4848	Ningxia	0.4624
Tianjin	0.7839	Shanxi	0.5428	Shaanxi	0.5039	Sichuan	0.4845	Heilongjiang	0.4442
Jiangsu	0.7196	Hebei	0.5255	Neimeng	0.4961	Gansu	0.4829	Guizhou	0.4334
				Henan	0.4938				

Table 2-9. Ranking by the score of the second principal component

Index	Score	Index	Score	Index	Score	Index	Score	Index	Score
Zhejiang	0.6848	Huna	0.0194	Henan	-0.156	Shandong	-0.2349	Shanxi	-0.382
Guangdong	0.2062	Ningxia	-0.013	Sichuan	-0.157	Hebei	-0.2353	Shaanxi	-0.414
Neimeng	0.079	Hainan	-0.053	Anhui	-0.168	Hubei	-0.3164	Liaoning	-0.434
Guizhou	0.0538	Xinjiang	-0.116	Jilin	-0.182	Chongqing	-0.3484	Tianjin	-0.771
Guangxi	0.0253	Fujian	-0.120	Jiangsu	-0.184	Gansu	-0.3576	Shanghai	-0.806
Heilongjiang	0.0225	Xizang	-0.142	Jiangxi	-0.199	Qinghai	-0.3628	Beijing	-1.309
				Yunnan	-0.212				

As shown in Table 2-8, the different regions can be roughly divided into three groups based on the household income level of urban residents.

1) The first group mainly includes Beijing, Shanghai, Zhejiang, Guangdong, Tianjin, Jiangsu, Fujian, etc. With massive urban markets, advanced scientific and technological power, convenient transportation and rich market information, Beijing and Shanghai provide exceptionally good conditions for market economy and abundant opportunities for the employment and development of the residents [8]. However, the two municipalities lag far behind other places in operating income under the influence of the regional environment and market conditions (Table 2-9). The uniqueness of Beijing and Shanghai residents is reflected by the closeness between the two municipalities and the remote western regions in the score of the second principal component [9]. In other traditionally developed regions in China, such as the coastal provinces of Zhejiang, Guangdong, and Fujian, the per capita household operating income of residents are higher than that of the other regions. These regions are also the frontier of the reform and opening up. In some western provinces, namely Inner Mongolia, Guizhou and Guangxi, the residents also enjoy significant increases in operating income, as evidenced by the high scores of the second principal component.

2) The second group consists of agricultural heavyweights like Shandong, Chongqing, Liaoning, Shanxi, Hebei, Yunnan, Hubei, Hunan, Anhui, Shaanxi, Inner Mongolia, Henan, Guangxi and Jilin. Thanks to the diversification of the income structure, the income level of the residents has been greatly improved in recent years. Nevertheless, the scores of the second principal component are varied due to the difference in regional environments.

3) The third group covers Qinghai, Jiangxi, Sichuan, Gansu, Hainan, Xinjiang, Tibet, Ningxia, Heilongjiang, Guizhou and other places [10]. Most of these regions are located in the border areas, featuring backward economy, large poor population, simple economic structure, limited source of income and low wage level. Owing to these features, the scores of the first principal component are relatively low. In contrast, the scores of the second principal component rank high by virtue of the implementation of the China Western Development strategy.

The analysis and study of the different income sources of urban households in various regions are very meaningful for the country to issue macro-control policies and set the policy orientation. This research helps to formulate more reasonable policies and rationalize policy formulation and implementation, thereby supporting the implementation of the Belt and Road Initiative. It also guides the economic development and improvement of residents' living standards in China.

REFERENCES

[1] National Statistics Bureau. (2015). *National Bureau of Statistics of China*, China Statistical Publishing House.

[2] Fang K.T. (1988). Practical regression analysis, *Science Press*, No. 1, p. 170.

[3] He X.Q., Liu W.Q. (2002). Applied regression analysis, *China Renmin University Press*, No. 1, pp. 116-118

[4] Tang S.Z. (1984). Multivariate statistics analysis method, *China Forestry Publishing House*, No. 1, pp. 90-101

[5] Wang X.M. (2002). Comparative analysis on urban residents' living expenditure by regions, pp. 64-69.

[6] Picard D. (1995). Testing and estimating change-point in time series, *Advances in Applied Probability*, Vol. 17, pp. 841-867. DOI: 10.2307/1427090

[7] Inclán C., Tiao G.C. (1994). Use of cumulative sums of squares for retrospective detection of changes of variances, *Journal of the American Statistical Association*, No. 89, pp. 913-923.

[8] Andrews D.W.K., Ploberger W. (1994). Optimal tests when a nuisance parameter is present only under the alternative, *Econometrica*, No. 62, pp. 1383-1414. DOI: 10.2307/2951753

[9] Jach A., Kokoszka P. (2004). Subsampling unit root tests for heavy-tailed observations, *Methodology and Computing in Applied Probability*, No. 6, pp. 73-94. DOI: 10.1023/B:MCAP.0000012416.28866.c5

[10] Hamilton J.D. (1994). *Time Series Analysis*, Princeton University Press, New Jersey.

[11] Bai J., Perron P. (2003). Computation and estimation of multiple structural change models, *Applied Econometrics*, No. 18, pp. 1-22.