Application of principal component analysis in evaluation of epidemic situation policy implementation

Jiaqi Chen¹, Fen Gong², Shaojun Xiang¹ and Ting Yu¹*

¹School of Science, Wuhan University of Science and Technology, Wuhan, Hubei, 430000, China
²School of Information Science and Engineering, Wuhan University of Science and Technology, Wuhan, Hubei, 430000, China
*Corresponding author’s e-mail: yuting@wust.edu.cn

Abstract. In order to study the impact of various policies issued by the government on epidemic prevention and control, we first considered them from three aspects: the degree of leniency of policies, the degree of stringency of policy implementation and the degree of self-consciousness of individuals, broken down into 10 indicators, designed a questionnaire to collect relevant data. Using Principal Component Analysis (PCA), 4 principal components were extracted from 10 indexes, and the scores and total scores of each principal component were obtained. Finally, the corresponding suggestions are given.

1. Principal Component Analysis

1.1 The basic idea of principal component analysis
Principal Component Analysis (PCA) is to try to replace the original index by a new set of irrelevant composite index. The internal structure of several variables is revealed through a few principal components. That is, to derive a few principal components from the original variables, so that they retain as much information as possible about the original variables and are not related to each other[1].

1.2 Mathematical model of principal component analysis
Assuming that there are n samples and p observation indexes(p<n), the original data matrix \( X = (X_1, X_2, \cdots, X_p) \) is obtained, The correlation coefficient matrix is R. The usual method in mathematics is to make a linear combination of the original indexes as a new comprehensive index[2]. Note that these new comprehensive indicators are \( Z_1, Z_2, \cdots, Z_k \). The most classic method is to use variance to express. These new indexes are not correlated with each other, and the variance decreases. Therefore, if the eigenvalue of the correlation coefficient matrix is \( \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \), and the vector \( l_1, l_2, \cdots, l_p \) is the corresponding unit eigenvector, then the \( i-th \) principal component is:

\[
Z_i = l_i^T X \quad (i = 1, 2, \cdots, p)
\]

Generally, the first k principal components are selected according to the cumulative contribution. In most cases, the first several principal components represent most of the information of the original indicators[3].
2. Application example - 2020 China's COVID-19 in Hubei Province

2.1 Data collection

In order to study the impact of various policies issued by different city governments on epidemic prevention and control[1], we selected 10 representative indicators, which were "Time to Return to Work" "Mode of Travel" "Measures for Returning Persons" "Requirements for Body Temperature Measurement for Returning Persons" "Isolation Period for Returning Persons" "Number of telephone searches for returning persons" "Requirements for daily necessities" "Frequency of residents going out to buy necessities" "Visit of relatives and friends" "Urban control".

A total of 885 questionnaires were distributed to the residents of Hubei Province, where the epidemic situation is high in China. 834 questionnaires were valid, and the effective rate was 94.20%. Among the valid questionnaires, 58% were male and 42% were female[2,4].

2.2 Calculation by principal component analysis

2.2.1 Data standardization and relevance determination. The ten indexes are represented by $X_1, X_2, \ldots, X_{10}$ respectively, the scores of each indicator in nine regions are standardized and the correlation matrix is calculated (as shown in Table 1).

|     | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $X_9$ | $X_{10}$ |
|-----|-------|-------|-------|-------|-------|-------|-------|-------|-------|---------|
| $X_1$ | 1.000 | .341  | -.026 | .594  | -.229 | .129  | -.062 | -.146 | .243  | -.308   |
| $X_2$ | .341  | 1.000 | -.770 | -.297 | -.392 | -.294 | -.452 | .370  | .061  | .084    |
| $X_3$ | -.026 | -.770 | 1.000 | .459  | .612  | .627  | .094  | -.286 | .200  | -.073   |
| $X_4$ | .594  | -.297 | .459  | 1.000 | .097  | .014  | .183  | -.655 | .130  | -.378   |
| $X_5$ | -.229 | -.392 | .612  | .097  | 1.000 | .641  | -.601 | -.279 | -.264 | -.252   |
| $X_6$ | .129  | -.294 | .627  | .014  | .641  | 1.000 | -.446 | .267  | .117  | .112    |
| $X_7$ | -.062 | -.452 | .094  | .183  | -.061 | -.446 | 1.000 | -.209 | .323  | .150    |
| $X_8$ | -.146 | .370  | -.286 | -.655 | -.279 | .267  | -.209 | 1.000 | -.005 | .436    |
| $X_9$ | .243  | .061  | .200  | .130  | -.264 | .117  | .323  | -.005 | 1.000 | .678    |
| $X_{10}$ | -.308 | .084  | -.073 | -.378 | -.252 | .112  | .150  | .436  | .678  | 1.000 |

From the correlation matrix of the output results, it can be seen that there is a significant correlation between the ten indicators. Therefore, the principal component analysis method can be used for further research.

2.2.2 Determination of the number of principal components. Common factor variance is the cumulative contribution rate of several common factor variances. The higher the cumulative contribution rate, the higher the representativeness or interpretation rate of the extracted common factors for the original variables, the better the overall effect[6, 7].

|     | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $X_9$ | $X_{10}$ |
|-----|-------|-------|-------|-------|-------|-------|-------|-------|-------|---------|
| Initial | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000   |
| Extract | 0.931 | 0.939 | 0.940 | 0.897 | 0.917 | 0.917 | 0.915 | 0.727 | 0.886 | 0.881   |

According to the common factor variance of the output results, the data in the last column of Table 2 are between 70% - 95%, indicating that the extracted principal components have a high degree of explanation for each variable.
Gravel map is a broken line map which arranges feature roots from large to small. The first four eigenvalues change obviously, and the fifth eigenvalue becomes more gentle, so this paper extracts four principal components.

From the total variance of the output result graph, the sum of variance of the first four principal components accounted for 89.488% of the total variance. Therefore, the first four principal components can basically retain the information of the original indicators, thus reducing the number of indicators from the original 10 to 4, thus reducing the dimension.

2.2.3 Naming principal components. The absolute value of the correlation coefficient represents the degree of correlation. From this data, the information of the original variables represented by each principal component can be obtained. On this basis, the principal component is named[5].

In this example, the first principal component mainly covers the measures taken for the returned people from Wuhan, the requirements for the temperature measurement of the returned people from Wuhan, and the isolation period for the returned people from Wuhan, which can be named "the degree of intervention for the returned people from Wuhan". The second principal component mainly covers the purchasing requirements for daily necessities and the frequency of going out to purchase necessities, which can be named "the degree of intervention for the returned people from Wuhan". The third principal component mainly covers the visits of relatives and friends during the Chinese New Year and the control strength of the city, which can be named "policy implementation strength". The fourth principal component mainly covers the return to work time, travel mode and the number of phone calls, which can be named "policy strictness".

2.2.4 Principal component expression. The principal component expression can be obtained from the component score coefficient matrix. The original variables after standardization are recorded as $z_{x_1}, z_{x_2}, \cdots, z_{x_{10}}$. The four principal components obtained in this example are denoted as $f_1, f_2, f_3, f_4$. The principal component expression is as follows:
The standardized data of each city is substituted into the principal component expression to get the score of each principal component and the total score of each city. The score results are shown in Table 3.

| City     | Component 1 | Component 2 | Component 3 | Component 4 | Total score |
|----------|-------------|-------------|-------------|-------------|-------------|
| Ezhou    | 0.771       | 1.441       | -0.185      | 2.059       | 4.086       |
| Huanggang| -1.060      | 0.300       | 0.360       | -0.345      | -0.745      |
| Huangshi | -0.095      | -1.086      | -1.902      | 0.425       | -2.657      |
| Jingmen  | 0.050       | 1.662       | -1.010      | -1.695      | -0.993      |
| Jingzhou | 0.682       | -1.058      | -0.354      | -0.154      | -0.884      |
| Suizhou  | -1.548      | 0.059       | 0.940       | 0.387       | -0.162      |
| Xiangyang| -0.652      | -0.370      | 0.453       | 0.133       | -0.437      |
| Xiaogan  | 0.134       | -0.775      | 0.334       | -0.222      | -0.528      |
| Yichang  | 1.718       | -0.173      | 1.364       | -0.588      | 2.321       |

It can be seen from the table that Ezhou and Yichang do the best in policy intervention. Yichang, Ezhou and Suizhou do a better job in the intervention of returning people in Wuhan; Ezhou and Jingmen do a better job in the intervention of daily purchasing; Suizhou and Yichang do a better job in the implementation; Ezhou, Huangshi and Suizhou have more strict policies.

3. Summary

The idea and method of principal component analysis are applied to the evaluation of epidemic prevention and control policies in various regions. Using the actual data, the first principal component is "the degree of intervention to the returned people in Wuhan", which reflects the importance of various regions to the returned people in Wuhan. The second principal component is "the degree of intervention in daily purchasing", which reflects the control of residents' daily travel in different regions. The third principal component is "policy implementation strength", which reflects the degree of responsibility implementation of governments at all levels. The fourth principal component is "policy strictness", which reflects the leniency and strictness of policy-making in different regions.

Using the method of principal component analysis to evaluate the implementation of policies in various regions, comprehensively weighing the scores of the four principal components in each region, governments at all levels can have a more impartial and objective grasp of the intervention of policies in various regions, then carry on the adjustment to the policy pertinently.
Acknowledgements
Fund item: Innovation training program for college students(provincial level) Impact of policy intervention on epidemic prevention and control; Item number:S202010488046

References
[1] Geng, H., Xu, A. D., Wang, X. Y. (2020) Analysis of the role of related interventions in COVID-19 outbreaks based on SEIR models. Journal of Jinan University(natural science and medicine), 41(02): 175-180.
[2] Li, Z. C., Guo, Q. (2010) Investigation and analysis on the core values of top-notch innovative talents. Party Building and Ideological Education in Schools, 22: 24-26.
[3] Liu, J. (2014) The principal components analysis of mathematical modeling. Science and Technology Vision, 000(015): 223-224.
[4] Wu, Z. Y. (2011) A study on the relationship between peer groups and the values of contemporary college students-An empirical study based on six universities in Guizhou. Party Building and Ideological Education in Schools, 07: 43-45.
[5] Wang, L. W., Zeng, Q. L. (2006) Application of principal component analysis in economic benefit evaluation of enterprises. Journal of United University of Beijing (Natural Science Edition), 20(3): 48-50.
[6] Wang, E. P., Yao, L., Sun, X. F., Ma, H. P. (2021) Cause analysis of coal mine accidents in Shandong province based on principal component analysis. Shaanxi Coal, 40(01): 57-60.
[7] Guo, L. L., Fu, Z. q., Yi, Q. J. (2021) Application of principal component analysis in the analysis and evaluation of students’ achievement. Higher Education Journal, 2021(03): 88-91.