Curriculum and Teaching Methodology (2021)
ISSN 2616-2261 Volume 4 Issue 5
DOI: 10.23977/curtm.2021.040501
Clausius Scientific Press, Canada

# *Research on the Higher Education System Based on Principal Component Analysis*

**Wenyuan Cai[1,*], Yingying Zhang[2], Jiayao Yin[3]**

[1]*School of Economics and Trade, Huashang College, Guangdong University of Finance and Economics, Guangzhou, Guangdong, 511300*

[2]*School of Finance, Huashang College, Guangdong University of Finance and Economics. Guangzhou, Guangdong, 511300*

[3]*School of Data Science, Huashang College of Guangdong University of Finance and Economics. Guangzhou, Guangdong, 511300*

*Corresponding author

***Keywords:*** principal component analysis, higher education system, SPSS

***Abstract:*** The higher education system is an important part of a country's education of citizens, and has important value to the country's economic development. Firstly, this paper conducted detailed research and analysis on the healthy education system and collected five indicators from eight representative countries (students from specific countries, scientific journal articles in higher education, and higher education staff, higher education enrollment rate, higher education sector's percentage of research and development expenditures) for ten years or more. In order to evaluate quantitatively, this paper establishes a principal component analysis model to evaluate the health status of the national higher education system, and uses SPSS to calculate the final comprehensive evaluation score.

## 1. Introduction

A healthy higher education system needs to measure factors such as cost, access, funding degree value, quality of education, level of research, and talent exchange. In the current epidemic situation, countries are reflecing on the health of their higher education systems. Our team has built a comprehensive principal component analysis model based on historical data to measure and assess the health of any country's higher education system at the national level, giving a good direction for the future development of education in each country at the time of the epidemic.

## 2. Principal Component Analysis Model

### 2.1 Model analysis

First of all, we conducted a detailed research and analysis of the healthy education system, and collected five indicators (students from specific countries (outbound migrant students) from eight representative countries (China, the United States, Japan, Australia, Singapore, Thailand,

Brazil and Switzerland), higher education science and technology journal articles, higher education faculty (ISCED 5 to 8), higher education enrollment rate, higher education sector's percentage of research and development (GERD) expenditure) for ten years or more. The key to the problem is to build a model that can assess the health of any country's higher education system. In order to be able to quantitatively evaluate, we use SPSS to perform principal component analysis, and finally get a comprehensive evaluation score.

China and Japan are located in the eastern part of Asia, the United States is located in the middle of North America, Australia is located between the South Pacific and the Indian Ocean, Singapore and Thailand are located in Southeast Asia, and Switzerland is located in central and southern Europe. The above eight countries are not only different in geographic location and national system. There are also distinctive features of the education system, which is why we chose these eight countries, and also makes our model suitable for evaluating the health of the education system of any country in the world.

## 2.2 Models Establishments

*Table 1: Data Source Collation.*

| *Database Names* | *Database Websites* | *Obtain Data Type* |
|---|---|---|
| WORLD BANK | https://data.worldbank.org.cn | Education |
| UN | http://data.un.org | Education |
| KNOEMA | https://cn.knoema.com | Education |

*Table 2: Data Table of Health Indicators for Higher Education in Eight Countries.*

| | *OMS* | *SJP* | *AST* | *ERH* | PES |
|---|---|---|---|---|---|
| China | 432413.9286 | 396222.1 | 1546005.4 | 27.30663238 | 9.177491053 |
| Singapore | 21353.71429 | 11106.33 | 16356 | 63.724068 | 25.71785722 |
| United States of America | 54303.5 | 426201.4 | 1425196.3 | 83.9565235 | 13.30058111 |
| Japan | 51689.42857 | 105251.1 | 527821.5714 | 55.74125 | 13.57390842 |
| Austria | 8986.357143 | 49999.89 | 39027.735 | 75.930625 | 26.795715 |
| Brazil | 23069.92857 | 51538.67 | 359489.1429 | 27.547899 | 14.92109294 |
| Switzerland | 10012.07143 | 21216.89 | 35580.1 | 49.84175453 | 24.48019 |
| Thailand | 25679.07143 | 8896 | 74604.75 | 45.21968034 | 29.71918375 |

Based on the covariance matrix

Data standardization:

$$\begin{cases} \mu = \dfrac{N_1 + N_2 + \cdots + Nn}{n} \\ \sigma = \sqrt{\dfrac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2} \\ X_i = Z_x = \dfrac{(x - \mu)}{\sigma} \end{cases} \tag{1}$$

In practical problems, the covariance of X is usually unknown, and the sample has:

$$X_1 = (x_{i1}, x_{i2}, \cdots, x_{ip})(i = 1,2, \cdots, n) \tag{2}$$

$$\sum_x = \sum \tag{3}$$

**Step 1:** From the covariance of X, find its characteristic roots, that is, solve the equation $|\Sigma-\lambda I|=0$, and get the characteristic roots[1]:

$$\lambda_1 > \lambda_2 > \cdots > \lambda_p \geq 0 \tag{4}$$

**Step 2:** Find the corresponding eigenvectors $U_1, U_2, \cdots U_P$,

$$U_j = (U_{1j}, U_{2j}, \cdots, U_{pj}) \tag{5}$$

**Step 3:** Calculate the cumulative contribution rate and give the appropriate number of principal components:

$$F_j = U_j^{'} X,\ j = 1, 2, \cdots, k(k \leq p) \tag{6}$$

**Step 4:** Calculate the scores of the selected k principal components. The centralization value of the original data:

$$X_i^* = X_i - \overline{X} = (x_{i1} - \overline{x_1}, x_{i1} - \overline{x_2}, \cdots, x_{ip} - \overline{x_p}) \tag{7}$$

Substitute the expressions of the first k principal components, calculate the scores of the k principal components of each unit, and rank them according to the score value.

**Step 5:** Principal component expression.

Two principal components and $Y_1$ and $Y_2$ expressions are obtained from the principal component loading matrix:

$$Y_1 = U_{11}X_1 + U_{21}X_2 + U_{31}X_3 + U_{41}X_4 + U_{51}X_5 \tag{8}$$

$$Y_2 = U_{12}X_1 + U_{22}X_2 + U_{32}X_3 + U_{42}X_4 + U_{52}X_5 \tag{9}$$

It should be noted that before calculating the variables, the original variables need to be standardized. The $X_1 \sim X_5$ in the above $Y_1$ and $Y_2$ expressions are the standard scores of the original variables.

**Step 6:** Calculate the comprehensive evaluation score.

$$Y = \omega_1 * Y_1 + \omega_2 * Y_2 \tag{10}$$

## 2.3 Models Solution

Step 1: KMO and BARTLETT sphere inspection [2]

**KMO and Bartlett's Test**

| | | |
|---|---|---|
| Kaiser-Meyer-Olkin Measure of Sampling Adequacy. | | .537 |
| Bartlett's Test of Sphericity | Approx. Chi-Square | 36.706 |
| | df | 10 |
| | Sig. | .000 |

*Figure 1: KMO and Bartlett test.*

Since KMO is 0.537>0.5, it shows that the data is suitable for factor analysis, and the significance P value of Bartlett's sphere test is 0.000, which also shows that the data is suitable

for factor analysis.

Step 2: The total variance decomposition table

As can be seen from the figure below, two principal components with eigenvalues greater than 1 are extracted, and the variance contribution rates of the two principal components are 67.365% and 23.924%, respectively. The cumulative variance contribution rate is 91.288%. The two eigenvalues are respectively 3.368 and 1.196.

**Total Variance Explained**

| Component | Initial Eigenvalues | | | Extraction Sums of Squared Loadings | | |
|---|---|---|---|---|---|---|
| | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % |
| 1 | 3.368 | 67.365 | 67.365 | 3.368 | 67.365 | 67.365 |
| 2 | 1.196 | 23.924 | 91.288 | 1.196 | 23.924 | 91.288 |
| 3 | .350 | 6.996 | 98.284 | | | |
| 4 | .083 | 1.667 | 99.951 | | | |
| 5 | .002 | .049 | 100.000 | | | |

Extraction Method: Principal Component Analysis.

*Figure 2: Total Variance Explained.*

Step 3: Factor loading matrix

The transformation matrix of principal component analysis, that is, the mathematical relationship between principal component loading matrix U, factor loading matrix A and eigenvalue λ is as follows:

$$U_i = A_i/\sqrt{\lambda_i} \tag{11}$$

Therefore, the principal component load matrix U can be obtained from the two by calculating variables:

*Table 3: Principal Component Loading Matrix U.*

| $U_1$ | $U_2$ |
|---|---|
| 0.462 | -0.302 |
| 0.5 | 0.334 |
| 0.528 | 0.202 |
| -0.145 | 0.87 |
| -0.486 | 0.017 |

Step 4: The principal component expression

Two principal components and $Y_1$ and $Y_2$ expressions are obtained from the principal component loading matrix:

$$Y_1 = X_1 * 0.462 + X_2 * 0.500 + X_3 * 0.528 - X_4 * 0.145 - X_5 * 0.486 \tag{12}$$

$$Y_2 = -X_1 * 0.302 + X_2 * 0.334 + X_3 * 0.202 + X_4 * 0.87 + X_5 * 0.017 \tag{13}$$

It should be noted that before calculating the variables, the original variables need to be standardized. The $X_1 \sim X_5$ in the above $Y_1$ and $Y_2$ expressions are the standard scores of the original variables.

Step 5: Calculate the comprehensive evaluation score.

$$Y=\omega_1 * Y_1 + \omega_2 * Y_2 \tag{14}$$

$$Y = \frac{(Y_1 * 67.365 + Y_2 * 23.924)}{91.288} \tag{15}$$

## 3. Model Evaluations

The model uses principal component analysis, which is objective and reasonable. In the comprehensive evaluation function, the weight of each principal component is its contribution rate, which reflects the proportion of the principal component containing the original data information to the total information, so that the weight is objective and reasonable.

But the interpretation of the principal component is a bit vague, not as clear and precise as the meaning of the original variable.

## 4. Conclusion

The higher education system is an important part of a country's education of citizens, and has important value to the country's economic development. Under the epidemic situation, all countries are reflecting on the health status of higher education in their countries. This paper proposes a series of models such as principal component analysis method to evaluate the health status of higher education.

We conducted detailed research and analysis on the healthy education system and collected five indicators from eight representative countries (students from specific countries (outbound migrant students), scientific journal articles in higher education, and higher education staff (ISCED 5 to 8), higher education enrollment rate, higher education sector's percentage of research and development (GERD) expenditures) for ten years or more, a principal component analysis model was established. In order to evaluate quantitatively, this paper establishes a principal component analysis model to evaluate the health status of the national higher education system, and uses SPSS to calculate the final comprehensive evaluation score. We obtain the comprehensive evaluation score formula: $Y = \frac{(Y_1 * 67.365 + Y_2 * 23.924)}{91.288}$.

## References

*[1] Shoukui Si, Zhaoliang Sun. Mathematical Modeling Algorithms and applications [M]. BEIJING: National Defense Industry Press, 2015.5:399-402.*

*[2] Qi Yong, Ou Lingyan. Research on the Quality Evaluation Index System of my country's Higher Education Development [J]. Education Modernization, 2020 (6): 107-113.*