DOI:10.18686/ahe.v7i20.9592

# Application of Principal Component Method in Student Achievement Analysis Based on R Language

**Manli Zhang, Rui Chen, Zhenhuan Kang, Lu Liu**
Taishan University,Tai'an271000, Shandong,China

**Abstract:** The principal component method is a frequently used method in dimensionality reduction, which usually targets data characterized by a certain correlation between variables, so that some major components can be identified from the intricate relationship of things, the redundant parts can be removed, and the new variables identified are mutually unrelated to each other as a comprehensive indicator, so that a large amount of statistical data can be effectively used for quantitative analysis to reveal the This paper introduces the idea of principal components and its application in the analysis of student performance. Principal components play an important role in dimensionality reduction and are a powerful tool for comprehensive multivariate evaluation.
**Keywords:** Principal component approach; Dimensionality reduction; Performance analysis

**Fund Project:** Taishan University Teaching Reform and Research Project(Project number:4101210116).

## 1. Principal component analysis methods

Principal components (abbreviated PCA) was introduced by Pearson (1901) and later developed by Hotelling (1933) [1]. An intuitive metaphor for principal components is to give data weight loss by retaining a valid minority of variables in place of the original redundant multiple variables.

The basic mathematical principle is as follows: if we write down the original variable of the thing under study as $Y_1,...Y_p$ Component analysis attempts to define a set of variables that are unrelated to each other $Z_1,...Z_p$ are called the principal components of the original variables, each of which is $Y_1,...Y_p$ Linear combinations of:

$$Z_1 = a_{11}Y_1 + a_{12}Y_2 ..... + a_{1p}Y_p$$
$$Z_2 = a_{21}Y_1 + a_{22}Y_2 ..... + a_{2p}Y_p$$
$$\vdots$$
$$Z_p = a_{p1}Y_1 + a_{p2}Y_2 ..... + a_{pp}Y_p$$

The principal components can be determined by solving for the representation coefficients of each principal component, which are derived in order of variance contribution: i.e., the first principal component $Z_1 = a_1^{'}y$ ,In meeting the restrictions $a_1^{'}a_1 = 1$ Time,Maximising the variance $\mathrm{var}(a_1^{'}y)$ ,Second principal component $Z_2 = a_2^{'}y$ ,In meeting the restrictions $a_2^{'}a_2 = 1$ Time,and $\mathrm{cov}(a_1^{'}y, a_2^{'}y) = 0$ Time maximized variance $\mathrm{var}(a_2^{'}y)$ , all the way to the jth principal component $Z_j = a_j^{'}y$ ,In meeting the restrictions $a_j^{'}a_j = 1$ Time,and $\mathrm{cov}(a_k^{'}y, a_j^{'}y) = 0$ Time maximized variance $\mathrm{var}(a_j^{'}y)$ ,In this way, j principal components are selected, and the process of calculation is obtained by using a decomposition theorem for orthogonal matrices of linear algebra:

note the covariance matrix M of the original variables, and note that its eigenvalues are $\lambda_1,...\lambda_p \geq 0$ The corresponding orthogonalized eigenvectors are $e_1,...e_p$ ,then the variable $Y_1,...Y_p$ The jth principal component of is given by the following equation: $Z_j = e_{j1}Y_1 + e_{j2}Y_2 + ... + e_{jp}Y_p, j = 1,..., p,$ of which $\text{var}(Z_j) = \lambda_j; \text{cov}(Z_j, Z_k) = 0$ .Principal component variables $Z_1,...Z_p$ and the sum of the variances of the original variables $Y_1,...Y_p$ The sum of the variances of the components is equal, ensuring the integrity of the information. In determining the number of principal components, we usually choose to explain about 80% of the original variance to determine the number of principal components, or we can use a gravel plot to determine the number of principal components selected.

## 2. Example application of the principal component method

When carrying out the example validation, the theory of sample principal component analysis is actually applied specifically. At the end of each semester, school teachers always do a flip-flop assessment of students' grades. Compared to overall grades and averages, we want to use a linear combination of several better raw variables to maximise differentiation between students; assessing principal components for each subject grade can give a new analytical assessment method to help teachers or parents get a deeper grasp of how well students are learning. (a) Firstly, the results of a class examination are selected; the performance data chosen here include mathematics, physics, chemistry, language, history and English;respectively, using the variables $y_1, y_2, y_3, y_4, y_5, y_6$ said. The data included a total of 52 students' results in six subjects. The correlation coefficient matrix is shown in Table 1: Table 1 shows that the correlation between the variables is quite high, which indicates that there is redundancy between the variables. The next step is to reduce the dimensionality using principal component analysis, which in R uses the function princomp(). From the results of the analysis, the cumulative contribution of two principal components reached 82.9%, so the selection of two principal components is sufficient, we can also choose the number of principal components by gravel plot.

Of course, in software it is common to have several variables to derive several principal components, and then the corresponding principal component variables are selected according to the principal component scores, and the loadings of the principal components are shown in Table 2.

Table 1: correlation coefficient table

| variable / coefficient | Y1 | Y2 | Y3 | Y4 | Y5 | Y6 |
|---|---|---|---|---|---|---|
| Y1 | 1 | 0.647 | 0.696 | -0.561 | -0.456 | -0.439 |
| Y2 | 0.647 | 1 | 0.573 | -0.503 | -0.351 | -0.458 |
| Y3 | 0.696 | 0.573 | 1 | -0.38 | -0.274 | -0.244 |
| Y4 | -0.561 | -0.503 | -0.38 | 1 | 0.813 | 0.835 |
| Y5 | -0.456 | -0.351 | -0.274 | 0.813 | 1 | 0.819 |
| Y6 | -0.439 | -0.458 | -0.244 | 0.835 | 0.819 | 1 |

Table 2. Principal component payloads

| principal component / variable | Comp.1 | Comp.2 | Comp.3 | Comp.4 | Comp.5 | Comp.6 |
|---|---|---|---|---|---|---|
| Y1 | 0.412 | 0.376 | 0.216 | 0.788 | 0.145 | 0 |
| Y2 | 0.381 | 0.357 | -0.806 | -0.118 | -0.212 | -0.141 |
| Y3 | 0.332 | 0.563 | 0.467 | -0.588 | 0 | 0 |
| Y4 | -0.461 | 0.279 | -0.599 | 0.59 | 0 | 0 |
| Y5 | -0.421 | 0.415 | -0.25 | 0.738 | 0.205 | 0 |
| Y6 | -0.43 | 0.407 | 0.146 | 0.134 | -0.222 | -0.749 |

From Table 2, we can obtain the expressions for the first two principal components:

$$Z_1 = 0.412Y_1 + 0.381Y_2 + 0.332Y_3 - 0.461Y_4 - 0.421Y_5 - 0.43Y_6$$

$$Z_2 = 0.376Y_1 + 0.357Y_2 + 0.563Y_3 + 0.279Y_4 + 0.415Y_5 + 0.407Y_6$$

In the raw data $y_1, y_2, y_3, y_4, y_5, y_6$ The first principal component is defined as bias because the first three variables have positive coefficients and are coefficients in front of the three subjects of mathematics, physics and chemistry, while the last three variables are language, history and English and have negative coefficients, which can be seen as the difference between subjects with a science bias and those with an arts bias. The second principal component, which is positive, can be seen as a weighted average and is therefore defined as balanced performance. We can further explore the principal component scores of a sample of typical students and then analyse their bias and balance of performance across subjects. For example, we pulled out the principal component scores of students with serial numbers 6, 7, 45, 30, 49, 26, 33 and 8 as shown in Table 3:

Table 3: Principal component scores

| 得分序号 | Comp.1 | Comp.2 | Comp.3 | Comp.4 | Comp.5 | Comp.6 |
|---|---|---|---|---|---|---|
| 6 | 3.518 | -0.82 | -1.072 | -0.156 | 0.763 | -0.166 |
| 7 | 3.516 | 0.104 | 0.101 | -0.574 | 0.011 | 0.08 |
| 45 | 3.975 | 1.054 | 0.147 | 0.349 | -0.252 | -0.049 |
| 30 | -4.49 | 0.693 | -0.62 | 0.832 | 0.054 | 0.029 |
| 49 | -4.622 | 0.997 | -0.236 | -0.724 | -0.289 | -0.465 |
| 26 | 0.841 | 2.117 | 0.544 | -0.156 | 0.07 | 0.192 |
| 33 | -0.345 | 2.187 | 0.414 | 0.225 | -0.018 | 0.854 |
| 8 | 0.982 | -2.326 | -1.292 | -0.017 | -0.093 | -0.052 |

From Table 3 we can see that the first principal component scores of students No. 6, No. 7 and No. 45 are all positive and quite large, revealing that they are better in science than in arts. Table 4 presents the true scores of the students, and we can see that students No. 6, No. 7 and No. 45 are indeed better in science than in arts. From Table 3 we can see that the first principal component scores of students No. 30 and No. 49 are negative and large, revealing that they are better in Arts. From the true scores presented in Table 4 we can also see that the two students, No. 30 and No. 49, are indeed much better in Arts teaching Science. From Table 3 we can see that the second principal component scores of students No. 26 and No. 33 are very high and positive. From the meaning of the second principal component is the overall score we can see that the overall score of students No. 26 and No. 33 should be good, and from the true scores in Table 4 we can see that these two students No. 26 and No. 33 do have good scores in all subjects. From Table 3 we can see that the second principal component for student number 8 is negative, and from the second principal component meaning this student should have a low true score.

Table 4: Students' real grades

| variable / Serial Number | Y1 (mathematics) | Y2 (physics) | Y3 (chemistry) | Y4 (Chinese) | Y5 (history) | Y6 (English) |
|---|---|---|---|---|---|---|
| 6 | 83 | 100 | 79 | 41 | 67 | 50 |
| 7 | 86 | 94 | 97 | 51 | 63 | 55 |
| 45 | 99 | 100 | 99 | 53 | 63 | 60 |
| 30 | 64 | 61 | 49 | 100 | 99 | 95 |
| 49 | 52 | 62 | 65 | 100 | 96 | 100 |
| 26 | 87 | 84 | 100 | 74 | 81 | 76 |
| 33 | 86 | 78 | 92 | 87 | 87 | 77 |
| 8 | 67 | 84 | 53 | 58 | 66 | 56 |

# References:


[1] Wang Binhui. Multivariate statistical analysis and R language modeling [M]. Guangzhou: Jinan University Press.2010.

[2] Mi Changlin, Ma Aigong, Zhang Xiaodong, Sun Jingguang, Yang Xuelian. Application of principal component analysis in remote sensing image data[J]. Shandong Land Resources,2013,07:69-71+76.


**About the author:**


Manli Zhang (1985.09-), Mujer, Nacionalidad Han, Tai'an City, Shandong Province, Estudiantes de posgrado, Research direction: Applied statistics.