

Application of factor analysis method in establishing exam result analysis model

Hongmei Nie

Math, Physics and Information Engineering College
Zhejiang Normal University
Jinhua, China
nhm@zjnu.cn

Abstract—In order to overcome shortcomings that the other analysis methods are not easy to reduce dimensions, difficult to establish a model, or not easy to analyze underlying factors, the paper presents to use factor analysis method to establish an exam result analysis model. Furthermore, the model can meet the actual needs of teachers carrying out the comprehensive analysis for the underlying factors on an exam. First, using factor analysis method, the paper gives the demand analysis of establishing an exam result analysis model so as to determine the main goal is to obtain the factor loading. Secondly, for the main goal, puts forward the related algorithm. Then, by collecting and processing data, verifies the realization of the algorithm, and establishes an exam result analysis model. Finally, for the exam result analysis model, the paper expounds the practical significance.

Keywords—factor analysis; exam result analysis model; underlying factors

I. INTRODUCTION

Currently most of student' result management systems in our education institutions provide scores recording, querying, analyzing, and printing functions. Exam result analysis modules in all kinds of student' result management systems provide analysis methods are quite different. By summing up, we know they are mainly four kinds of analysis methods: principal component analysis, cluster analysis, association analysis, classification analysis.

The above four methods exist themselves shortcomings. Using principal component analysis method, for factors that affect the result, we can reduce dimensions. However, principal component analysis method is not easy to establish a model. Furthermore, if we hope that principal factors provide an explanation conforming to actual background and significance, this is usually difficult [1][2].

For dealing with noise data and reducing dimensions, clustering analysis method is usually weak. Therefore, using this method, we possibly get some abnormal results [3][4].

Association analysis method can only get the relationship between factors and comprehensive result, can't obtain direct quantitative analysis result. So this isn't conducive to the analysis of a result [5][6].

In classification analysis method, the factors that affect a result can be observed. However, for the underlying factors that are not easy to be observed, it is hard for classification analysis method to get an objective interpretation [7][8][9].

In order to overcome the shortcomings of the above methods, we hope to establish an exam result analysis module which has the following four advantages. 1. The module is easy to reduce dimensions; 2. The module is easy to establish an exam analysis model, and the model is easy to give an explanation conforming to actual background and significance; 3. For the underlying factors that influence the result, the model can provide comprehensive explanation; 4. For the exam result, the module can timely make analysis and evaluation.

In order to achieve this goal, first, using factor analysis method, the paper gives the demand analysis of establishing an exam result analysis model. Secondly, on the basis of the demand analysis, puts forward the main algorithm processes. Then, by collecting and processing data, verifies the realization of the algorithm, and establishes an exam analysis model. Finally, for the exam result analysis model, the paper expounds the practical significance.

II. THE DEMAND ANALYSIS OF AN EXAM RESULT ANALYSIS MODEL

An exam course has g item types, such as: filling in the blanks, choice, calculation, application, practice, proof, higher level, etc. Basically all kinds of types can be attributed to the students' interest in learning, learning attitude, learning methods and so on m aspects. Every aspect is called a factor. The relationship between g kinds of scores and m factors can be described as follows:

$$x_i = a_{i1}h_1 + a_{i2}h_2 + \dots + a_{im}h_m + e_i, i = 1, 2, \dots, g \quad (1)$$

Among them, x_i ($i=1, 2, \dots, g$) represent g types of scores. h_1, h_2, \dots, h_m represent m factors. Each score can be expressed as a linear combination of h_1, h_2, \dots, h_m . a_{ij} is called a loading of x_i on the factor h_j , reflecting the importance of factor h_j about x_i . e_i is called an error.

Formula (1) can also be represented in matrix and vector as follows:

$$x = Ah + e \tag{2}$$

In the formula (2), $x=(x_1, x_2, \dots, x_g)'$, $h=(h_1, h_2, \dots, h_m)'$, $e=(e_1, e_2, \dots, e_g)'$, $A=(a_{ij})_{g,m}$ is a loading matrix. Usually we suppose that

$$E(h) = 0, E(e) = 0, V(h) = I, \text{cov}(h, e) = 0.$$

That is, the formula (2) is an orthogonal factor model. There is no correlation between h_1, h_2, \dots, h_m . There is no correlation between h and e .

On both sides of formula (2), we calculate the covariance, get formula (3):

$$V(x) = AV(h)A' + V(e) \approx AA' \tag{3}$$

According to the above analysis, our main goal is to get the estimate of the factor loading matrix $A=(a_{ij})_{g,m}$, thus, we can establish an exam result analysis model. Aiming at this goal, we present the following algorithm.

III. ALGORITHM PROCESS

For establishing an exam result analysis model, we give the following specific algorithm process:

Step1. x_1, x_2, \dots, x_g is a set of n samples. We calculate the average value of x in formula (2), namely,

$$\bar{x} = \frac{1}{g} \sum_{i=1}^g x_i.$$

Calculate the estimate value of covariance $V(x)$, namely,

$$S = \frac{1}{g-1} \sum_{i=1}^g (x_i - \bar{x})(x_j - \bar{x}), i, j = 1, 2, \dots, g.$$

Step2. Calculate all eigenvalues $\lambda_i (1, 2, \dots, g)$ in $|S-\lambda I|=0$.

Step3. By putting $\lambda_i(1, 2, \dots, g)$ into $(\lambda_i I - S)t_i = 0$, Calculate the eigenvectors $t_i(1, 2, \dots, g)$ corresponding to the eigenvalues $\lambda_i(1, 2, \dots, g)$. Then we transform into the eigenvectors $t_i(1, 2, \dots, g)$ into the unit orthogonal eigenvectors $T_i(1, 2, \dots, g)$.

Step4. The eigenvalues are sorted: $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_g$. Calculate the accumulative contribution rate C of the main factors, making

$$C = \frac{\sum_{i=1}^m x_i}{\sum_{i=1}^g x_i} \geq 0.85.$$

Step5. Due to

$$V(x) \approx AA' \approx S \\ \approx \lambda_1 T_1 T_1' + \lambda_2 T_2 T_2' + \dots + \lambda_m T_m T_m' = A^* A^{*'},$$

Calculate

$$A^* \approx (\sqrt{\lambda_1} T_1, \sqrt{\lambda_2} T_2, \dots, \sqrt{\lambda_m} T_m) = (\hat{a}_{ij}).$$

So, A^* is the estimate of the factor loading matrix $A=(a_{ij})_{g,m}$.

IV. THE ESTABLISHMENT OF AN EXAM RESULT ANALYSIS MODEL

A. Collect data and set variables

The author uses some students' higher mathematics scores in a class as the source data. The item types of the higher mathematics exam involve: filling in the blanks, choice, calculation, application, practice, proof, and higher level. Let x_1, x_2, \dots, x_7 respectively be the seven kinds of scores, and let $\bar{x}_i (i=1, 2, \dots, 7)$ be the average value of x_1, x_2, \dots, x_7 . Table 1 shows the sample data.

TABLE 1 THE HIGHER MATHEMATICS SCORES

| ID | x ₁ | x ₂ | x ₃ | x ₄ | x ₅ | x ₆ | x ₇ |
|-------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| 1 | 8 | 20 | 10 | 10 | 6 | 2 | 2 |
| 2 | 6 | 6 | 12 | 6 | 4 | 4 | 2 |
| 3 | 10 | 16 | 16 | 18 | 8 | 8 | 7 |
| 4 | 8 | 18 | 18 | 16 | 8 | 5 | 2 |
| 5 | 2 | 6 | 6 | 2 | 2 | 0 | 0 |
| 6 | 8 | 18 | 20 | 20 | 8 | 10 | 8 |
| 7 | 6 | 20 | 12 | 10 | 2 | 1 | 1 |
| 8 | 4 | 14 | 10 | 8 | 4 | 6 | 0 |
| 9 | 6 | 10 | 10 | 12 | 6 | 8 | 4 |
| 10 | 6 | 16 | 12 | 8 | 4 | 2 | 0 |
| 11 | 4 | 6 | 12 | 4 | 4 | 0 | 0 |
| 12 | 6 | 6 | 2 | 6 | 4 | 0 | 1 |
| 13 | 2 | 8 | 10 | 4 | 4 | 1 | 1 |
| 14 | 8 | 18 | 8 | 2 | 4 | 3 | 2 |
| 15 | 6 | 12 | 6 | 10 | 6 | 2 | 0 |
| \bar{x}_i | 6.00 | 12.93 | 10.93 | 9.07 | 4.93 | 3.47 | 2.00 |

B. Process the data

By processing the data in table 1, we get table 2 and table 3.

TABLE 2 THE CORRELATION MATRIX

| | x_1 | x_2 | x_3 | x_4 | x_5 | x_6 | x_7 |
|-------|--------|--------|--------|--------|--------|--------|--------|
| x_1 | 1.0000 | 0.6647 | 0.4604 | 0.6813 | 0.6996 | 0.5430 | 0.6533 |
| x_2 | 0.6647 | 1.0000 | 0.4999 | 0.5459 | 0.3866 | 0.3659 | 0.3421 |
| x_3 | 0.4604 | 0.4999 | 1.0000 | 0.7336 | 0.6119 | 0.6593 | 0.6370 |
| x_4 | 0.6813 | 0.5459 | 0.7336 | 1.0000 | 0.8389 | 0.7789 | 0.7703 |
| x_5 | 0.6996 | 0.3866 | 0.6119 | 0.8389 | 1.0000 | 0.7267 | 0.7191 |
| x_6 | 0.5430 | 0.3659 | 0.6593 | 0.7789 | 0.7267 | 1.0000 | 0.8244 |
| x_7 | 0.6533 | 0.3421 | 0.6370 | 0.7703 | 0.7191 | 0.8244 | 1.0000 |

TABLE3 THE EIGENVALUE AND EIGENVECTOR OF THE CORRELATION MATRIX

| Eigenvector | T_1 | T_2 | T_3 | T_4 | T_5 | T_6 | T_7 |
|------------------------|---------|---------|---------|---------|---------|---------|--------|
| | -0.4396 | 0.2604 | 0.4101 | -0.0520 | 0.5310 | -0.3884 | 0.3677 |
| | 0.3532 | -0.0572 | -0.3089 | -0.1904 | -0.2277 | -0.7786 | 0.2869 |
| | -0.0875 | 0.2323 | 0.4729 | 0.1786 | -0.7399 | 0.0620 | 0.3628 |
| | -0.4411 | -0.6815 | -0.2670 | 0.2882 | -0.0495 | 0.0567 | 0.4255 |
| | 0.4910 | 0.2462 | -0.1355 | 0.6379 | 0.2958 | 0.1629 | 0.3988 |
| | -0.2246 | 0.4545 | -0.5379 | -0.4236 | -0.0558 | 0.3425 | 0.3922 |
| | 0.4340 | -0.3787 | 0.3657 | -0.5097 | 0.1602 | 0.3034 | 0.3963 |
| Eigenvalue λ_i | 0.0923 | 0.1409 | 0.2258 | 0.3208 | 0.5397 | 0.8768 | 4.8037 |

Table 2 shows the correlation matrix of the sample data, and Table 3 shows the eigenvalue and eigenvector of the correlation matrix.

In table 3, The eigenvalues are sorted: $\lambda_7 > \lambda_6 > \lambda_5 > \lambda_4 > \lambda_3 > \lambda_2 > \lambda_1$. Then, we get the accumulative

contribution rate $C = (\lambda_7 + \lambda_6 + \lambda_5) / (\lambda_7 + \lambda_6 + \lambda_5 + \lambda_4 + \lambda_3 + \lambda_2 + \lambda_1) > 0.85$. So, we set the factor number $m = 3$. Get the main factors, as shown in table 4.

TABLE4 THE MAIN FACTORS

| | h_1 | h_2 | h_3 |
|-------|--------|---------|---------|
| x_1 | 0.8059 | -0.3637 | 0.3901 |
| x_2 | 0.6288 | -0.7291 | -0.1673 |
| x_3 | 0.7952 | 0.0581 | -0.5436 |
| x_4 | 0.9326 | 0.0531 | -0.0364 |
| x_5 | 0.8741 | 0.1525 | 0.2173 |
| x_6 | 0.8596 | 0.3207 | -0.0410 |
| x_7 | 0.8686 | 0.2841 | 0.1177 |

Therefore, table 4 is the exam result analysis model.

C. Analyze the model

Table 4 shows that $x_1, x_2, x_3, x_4, x_5, x_6, x_7$ have larger positive loadings on the factor h_1 , thus the factor h_1 can be called the learning attitude factor; On the factor h_2 , x_6, x_7, x_5 have moderate positive loadings, x_3, x_4 have smaller positive loadings, x_2 has larger negative loadings, x_1 has moderate negative loadings, thus h_2 can be referred to as the learning interest factor; On the factor h_3 , x_1, x_5, x_7 have moderate positive loadings, x_2, x_3 have moderate negative loadings, x_4, x_6 have smaller negative loadings, thus factor h_3 can be called a learning method factor.

Through the analysis of the above data, we can draw the following conclusion. If you want to get good grades in a mathematics subject, your learning attitude is the most critical factors. If you hope to make remarkable achievement in a mathematics subject, to maintain interest in the study of this

subject is imperative. In the process of learning, of course, you should also pay attention to cultivating your own learning methods.

By analysing every exam, teachers should excavate the real underlying learning factors so as to change the way of teaching in accordance with students' practical condition.

V. CONCLUSION

In this paper, factor analysis method is adopted to establish an exam result analysis model, which overcomes the shortcomings of the other methods being not easy to reduce dimensions, difficult to build a model, or not easy to analyse underlying factors. Moreover, this method can meet every teacher's actual demand for each exam result analysis.

In the paper, the model has the objective analysis results, is easy to transplant, and fully contact with practice. However, the described method needs to be improved, that is we need to

further calculate scores of each sample x_i on m factors. In this way, we can provide more specific analysis for each student's learning situation. In other words, for each student's learning situation, we can analysis in detail. In the next step of research, we need to add this part so as to make the whole result analysis module be improved.

REFERENCES

- [1] Meifang Ding. "PCA method application in the analysis of student performance," Journal of Technology College Education, vol.22, 2003, pp.71-74.
- [2] Nengfu Yu,Bangkao Chen. Application of principal component analysis in the analysis of student achievement.Journal of Anhui Construction Industry Institute, vol. 15,2007, pp.73-75.
- [3] Hongbo Gu,Wanping Zhao. Application of MMD clustering algorithm in performance analysis. Journal of Hebei University of Engineering, vol. 27, 2010, pp.96-99.
- [4] Xianghong Song, Mingyu Bai. Student achievement analysis based on clustering analysis. Computer Knowledge and Technologe, vol. 8, 2012, pp.519-521.
- [5] Qing Tan. Research of association rules APriori algorithm in the paper analysis. Journal of Xinyang Normal University, vol.22,2009, pp.300-303.
- [6] Xuewen Cui. The application of association rules Apriori algorithm in student performance analysis. Journal of Hebei North University, vol. 27, 2011, pp.44-47.
- [7] Yongmei Wang. The application of ID3 algorithm in the computer performance analysis. Journal of Hefei College , vol. 20,2010,pp.34-38.
- [8] Haixia Huang,Xianglin Xu. The research and analysis of student's grade based on C4.5 algorithm. Computer Knowledge and Technologe vol. 7,2011,pp.3405-3409.
- [9] Shufeng Xiong, Liming Nie. Student achievement analysis based on C5.0 algorithm. Scientific and Technological Information 2010,pp.24-25.