Available online at http://scik.org

J. Math. Comput. Sci. 9 (2019), No. 6, 702-706

https://doi.org/10.28919/jmcs/4212

ISSN: 1927-5307

COMPARISON OF MULTIVARIATE STATISTICAL ANALYSIS METHODS

WU WENQI, WANG JIAN*

Mathematical and Statistical Institute, Shandong University of Technology, Zibo, China

Abstract: In order to solve the problems more efficient, it is necessary for us to compare the statistical analysis methods. Firstly, this paper compares two common dimensionality reduction methods- principal component analysis and factor analysis. Then four discriminant analysis methods are compared in applicability. It is a try of helping to clearly choose the suitable statistical analysis methods.

Keywords: principal component analysis and factor analysis; discriminant analysis.

2010 AMS Subject Classification: 62H25.

1. INTRODUCTION

Some methods for dealing with two-dimensional normal population appeared in the 19th century. In the 20th century, the methods systematically processing the overall statistical analysis multi-dimensional probability distribution problems have come up. People often use the derivation of the Wishart distribution as the sign of the independent discipline of multidisciplinary analysis. In the mid-1950s, with the development and popularization of electronic computers, it has been widely used in many fields such as geology, biology, and economic analysis, which also promoted the development of theory ([1]).

*Corresponding author

E-mail address: wjzhenhua@126.com

Received July 8, 2019

2. COMPARISON BETWEEN PRINCIPAL COMPONENT ANALYSIS AND FACTOR ANALYSIS

2.1 Principal component analysis

Principal component analysis is to linearly transform the original variables to obtain new variables. The amount of the original variable information contained in the new variables expressed by the variance. New variable which has the largest variance is the first principal component. Except the overlapping information in the first principal component, the second principal component has the largest variance. The principal components are sequentially obtained so that the principal components can contain most of the information of the original variables.

2.2 Factor analysis

The basic idea of factor analysis is to integrate the original variables into common factors and special factors. The number of factors is smaller than the number of original variables, and the common factors are not related to the special factors. The original variables are represented by common factors and special factors. The common factors have effects on all variables, and the special factors only work on specific variables.

2.3 Comparison between principal component analysis and factor analysis

The basic idea of principal component analysis is to linearly transform the original data to form new variables, so that new variables can reflect the most information of the original variables. These new variables are principal components, and the principal components are not related to each other. Factor analysis is the synthesis of original variables into fewer factors, to find the relationship between variables and factors.

The coefficients of each principal component in principal component analysis are uniquely determined and orthogonal, and cannot rotate; the coefficients of the factors in the factor analysis are not unique and can rotate, and the coefficients indicate the degree of correlation between variables and factors. Factor analysis can make the factor have a naming explanatory factor through factor rotation, and principal component analysis cannot ([2]).

In principal component analysis, the principal component can be directly obtained from the sample variable, that is, there is reversibility; the load matrix in the factor analysis is irreversible,

and the unobservable common factor can only be estimated from the sample variable ([3]).

However, both are produced under the idea of dimensionality reduction, which is obtained by transforming the original variables; the number of principal components and factors is less than the number of original variables, and the number of them can be determined by the contribution degree.

3. COMPARISON OF DISCRIMINANT ANALYSIS METHODS

3.1 Markov distance discrimination

The Mahalanobis distance discrimination is the most common distance discrimination method in distance discrimination. The basic idea of Mahalanobis distance is simply summarized as follows: The Mahalanobis distance between the sample and the total is the smallest, the sample is judged to the overall.

3.2 Bayes discriminant

Bayes statistical thought always assumes that there is a certain understanding of the research object before the discrimination. Generally, the prior distribution is used to describe this understanding, and then randomly select a sample, and the extracted samples are used to correct the existing knowledge. The prior distribution probability is obtained by obtaining the posterior probability distribution, then establishing the loss function, and then discriminating the attribution of the sample according to the principle of minimum loss ([4]).

3.3 Fisher discriminant

The basic idea of Fisher's discriminant is projection. The k sets of m dimension data are projected into a certain direction so that the group and the group are separated as much as possible after the projection. The method of measuring whether groups are separated from each other is based on the idea of variance analysis.

3.3 Stepwise discrimination

The basic idea of stepwise discriminant is to introduce variables one by one. It is necessary to examine whether the discriminant effect of the variable on the whole is significant to decide

whether to introduce the variable into the discriminant, and then introduce the variable with the strongest discriminative ability. After the introduction of variables, the discriminating ability of the original variables may be weakened by the introduction of new variables. At this time, it is necessary to eliminate them, and the process is repeated until no variables can be introduced and no variables can be eliminated.

3.4 Comparison of discriminant analysis methods

The Mahalanobis distance discrimination is based on the assumption that the distribution of the sample is normally distributed and then directly uses the distance between the sample and the overall distance to discriminate the classification. The overall distribution is not required and the calculation method is simple. However, the problem is that the Mahalanobis distance discrimination does not consider the probability of occurrence of each overall and does not consider the misjudgment of losses. For example, in the identification of "disease" and "no disease", the probability of occurrence of the two situations is different, and the "disease" is judged as "not disease" and "not disease" is judged as "disease". The losses in the two cases are also different.

The Bayes discriminant method is to calculate the loss judged as each overall under the premise that the probability of occurrence of each overall is known, and then judged as the overall which has the smallest loss. When the probability of occurrence is equal and the variances of all kinds are equal, the Markov distance discrimination and the Bayes discrimination have the same efficiency.

In essence, Fisher discriminant is to do the transformation first and then to make the distance discriminant. It is simple to apply when the overall mean vector collinearity is high, and it does not require the overall distribution and sample data, but does not consider the overall appearance probability and is not able to estimate the loss of the wrong judgment.

Markov distance discrimination, Bayes discriminant and Fisher discriminant are all using the variables to establish discriminant functions, but some of the variables in these variables have a great effect on the discriminant effect, and some have little effect, more than a number of

variables increase the amount of computation and may cause a reduction in computational accuracy. Gradual discriminant can be selected to discriminate the significant variables to establish a discriminant function for classification.

On the whole, each method has advantages and disadvantages. When using it, the application should analyze whether the overall distribution is known, whether it is required to misjudge the loss, and so on, and choose the most appropriate method.

Conflict of Interests

The authors declare that there is no conflict of interests.

REFERENCES

[1] https://baike.so.com/doc/6311666-6525255.html

[2] http://blog.csdn.net/ysuncn/archive/2007/12/08/1924502.aspx

[3] Manjie Qiao, Huihui Lu, Panpan Wu. Analysis of principal component analysis and factor analysis. Wisdom Health, 4 (36) (2018), 41-42.

[4] Huixuan Gao. Multivariate statistical analysis. Peking University Press 2005.