Algorithm Implementation Process and Mathematical Model Construction for Factor Analysis

Xiuli Xu

Zaozhuang Vocational College of Science & Technology,Tengzhou, 277599, China

xuxiuli1020@163.com

Keywords: factor analysis; basic principles; algorithm implementation process; mathematical model construction

Abstract. Factor analysis is based on a statistical analysis method. It integrates the variables of intricate and complex relationships into a small number of factors, reproduces the relationship between original variables and factors, and realizes the interpretation of complex economic and social issues. Based on the basic principle of factor analysis, this paper studies the process of factor analysis algorithm, including constructing initial matrix, data normalization, calculating correlation matrix, calculating Eigen values, determining the number of factors, calculating factor loading matrix, establishing factor analysis model, and factor variable name explains; according to the algorithm process, the mathematical model of factor analysis is constructed. By pursuing the basic structure of variables, simplifying the system under study, reducing the number of variables, and using a few variables to interpret the complex issues under study.

Introduction

In practical scientific research, it is often desirable to collect as much data as possible, and to grasp and understand the issue in a more comprehensive and complete manner. However, increasing the number of variables will increase the complexity of the analysis problem, because there may be a correlation between variables, resulting in information overlap between multiple variables. In the face of high dimensionality, large scale, and complex structure data, data reduction is particularly important. Data reduction, also known as dimensionality reduction, is to reduce the dimensions of data. On the one hand, it can solve "dimensional disasters", alleviate the "information richness, lack of knowledge" status and reduce complexity; on the other hand, it can better understand And understand the data. Factor analysis is based on the idea of data dimensionality reduction, and a multivariate statistical analysis method that reduces variables with intricate and complex relationships into a few comprehensive factors. From the dependence of research variables on the internal correlation, the correlation is higher, that is, the links are more closely classified in the same category, and the correlation between different types of variables is lower. Each type of variable represents a basic structure, which is a common factor. It is combining variables with intricate and complex relationships into several smaller factors to reproduce the relationship between original variables and factors, simplifying the analysis.

Basic Principles for Factor Analysis

Factor analysis uses less independent factor variables instead of most of the information of the original variables. Factor analysis method and principal component analysis method are both based on statistical analysis methods, but there is a big difference between them. Principal component analysis method is based on coordinate transformation to extract principal components, that is, to transform a group of variables with correlation into one. Group-independent variables express the principal component as the linear combination of original observation variables; factor analysis method constructs the factor model and decomposes the original observation variables into linear combinations of factor variables. Therefore, factor analysis is the development of principal

Copyright © 2018, the Authors. Published by Atlantis Press.
This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

component analysis. The narrow factor analysis method and the principal component analysis method have similarities in processing methods. All variables must be normalized and the correlation matrix after normalization of the original variables should be found. The main difference is the method used to establish the linear equations. The factor analysis expresses the variable as a linear combination of factor variables in the form of a regression equation, and the number of factors is smaller than the original variable, which simplifies the model structure. The core goal of factor analysis is to concentrate the original variable extraction factor. The basic principle is described as follows:

(1) Prerequisites. One of the tasks of factor analysis is to concentrate the original variables, that is, to extract the overlapping parts of the original variables and synthesize them into factors, thereby reducing the number of variables. Therefore, factor analysis requires a strong correlation between variables. If the original variables are independent of each other, they cannot be condensed and no factor analysis can be performed. Therefore, it is necessary to check whether the original data is suitable for factor analysis. Kaiser-Meyer-Olkin measures test statistics are commonly used.

(2) Factor extraction. Extracting and synthesizing factors on the basis of sample data, the key is to solve factor load matrix through sample data, usually using principal component analysis method, using correlation matrix for analysis, and extracting based on Eigen values. Spindle analysis, maximum likelihood, least squares, factor extraction, and image analysis are also used.

(3) Making the factors naming interpretable. After the original variables are synthesized into a few factors, if the actual meaning of the factors is not clear, it is not conducive to further analysis. Therefore, the factors to be extracted are analyzed in depth to make the naming interpretable and clear to people.

(4) Calculate the factor score of each sample. Usually, the regression method is used, and the resulting factor score has an average value of 0. The variance is equal to the square of the multivariate correlation between the estimated factor value and the real factor value. The output can be saved as a variable or a matrix of score coefficients to lay the foundation for further analysis.

Algorithm Implementation Process for Factor Analysis

The algorithmic process of factor analysis is shown in Fig. 1:

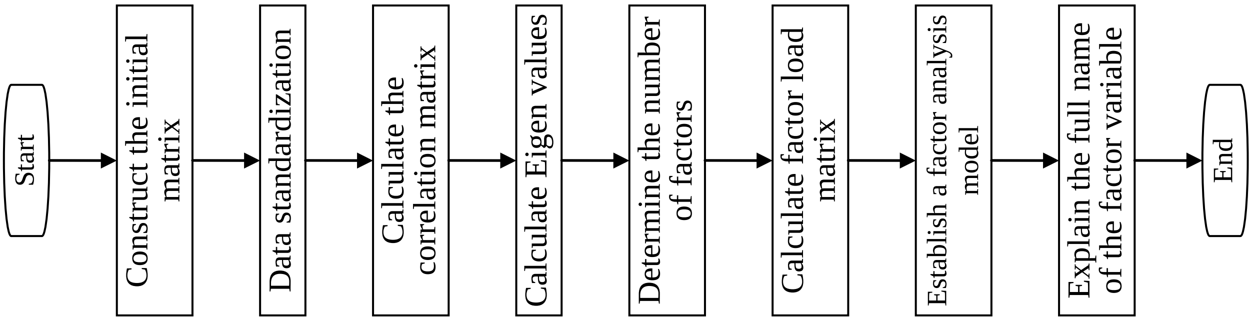
Start
Construct the initial matrix
Data standardization
Calculate the correlation matrix
Calculate Eigen values
Determine the number of factors
Calculate factor load matrix
Establish a factor analysis model
Explain the full name of the factor variable
End

Fig. 1. Algorithm implementation process for factor analysis

Mathematical model construction for factor analysis

The mathematical model constructed according to the factor analysis algorithm shown in Fig. 1 is as follows:

Step 1: Construct the initial matrix

Let *m* samples, *n* indicators, the *j*-th index of *i*-th sample is y_{ij} $(i=1,2,\cdots,m \quad j=1,2,\cdots,n)$, and all the sample values constitute the initial matrix, expressed as follows:

$$Y = \left[y_{ij} \right]_{m \times n} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ y_{m1} & y_{m2} & \cdots & y_{mn} \end{bmatrix} \tag{1}$$

Step 2: Data standardization

There are many ways to standardize the data, including *Z*-score, full-range from -1 to 1, full-distance from 0 to 1, 1 of maximum and average is 1, standard deviation is 1, etc. This article selects "full-range from 0 to 1".

The average of the *j*-th column (factor) data is:

$$\overline{y_j} = \sum_{i=1}^{m} y_{ij} \Big/ m \tag{2}$$

The new sequence after data averaging is:

$$\overline{y_{ij}} = y_{ij} \Big/ \overline{y_j} \tag{3}$$

The data is averaged and normalized. The formula is as follows:

$$x_{ij} = \frac{\overline{y_{ij}} - \min(\overline{y_{i1}}, \overline{y_{i2}}, \cdots, \overline{y_{in}})}{\max(\overline{y_{i1}}, \overline{y_{i2}}, \cdots, \overline{y_{in}}) - \min(\overline{y_{i1}}, \overline{y_{i2}}, \cdots, \overline{y_{in}})} \tag{4}$$

The standardized data matrix is:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \tag{5}$$

Step 3: Calculate the correlation matrix

The correlation matrix is represented by *R* and consists of the correlation coefficient between each two indicators. The calculation formula of the correlation coefficient is:

$$r_{ij} = \frac{1}{m-1} \sum_{k=1}^{m} X_{ik} X_{jk} \quad (i, j = 1, 2, \cdots, n) \tag{6}$$

The correlation matrix consisting of correlation coefficients is expressed as:

$$R = \frac{1}{m-1} X'X = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{bmatrix} \tag{7}$$

R is a real symmetric matrix, element values on the main diagonal $r_{ii} = 1$.

Step 4: Calculate Eigen values

The Eigen values of the correlation matrix can be found by the characteristic polynomial of *R*. Let *E* be the identity matrix and the characteristic polynomial be expressed as follows:

$$\left| \lambda E - R \right| = \begin{vmatrix} \lambda - r_{11} & -r_{12} & \cdots & -r_{1n} \\ -r_{21} & \lambda - r_{22} & \cdots & -r_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ -r_{n1} & -r_{n2} & \cdots & \lambda - r_{nn} \end{vmatrix} = \lambda^n + r_1 \lambda^{n-1} + r_2 \lambda^{n-2} + \cdots + r_{n-1} \lambda + r_n \tag{8}$$

$f(\lambda) = \left| \lambda E - R \right| = \lambda^n + r_1 \lambda^{n-1} + \cdots + r_n = 0$, it is an *n*-degree algebraic equation called the characteristic equation of *R*. The root of the characteristic equation is called the characteristic root (Eigen value) of *R*.

Step 5: Determine the number of factors

The number of factors is determined by the cumulative contribution rate. The factor contribution rate indicates the ratio of the degree of variation of each factor to the degree of variation of all factors. The calculation formula is:

$$c_i = \lambda_i \Big/ \sum_{i=1}^{n} \lambda_i \tag{9}$$

The cumulative factor contribution rate of the first *k* factors is expressed as:

$$d_k = \sum_{i=1}^{k} c_i \tag{10}$$

When the cumulative contribution rate exceeds 80% or the Eigen value λ is not less than 1, the number of factors is determined.

Step 6: Calculate factor load matrix

The eigenvector *U* with *p*-dimension can be decomposed into:

$$U = [U_1 U_2 \cdots U_k U_{k+1} U_{k+2}] = \begin{bmatrix} \underset{n\times k}{U(1)} & \underset{n\times(n-k)}{U(2)} \end{bmatrix} \tag{11}$$

The basic equations for factoring *U* into factor analysis are:

$$\underset{n\times m}{x} = \underset{n\times n}{U}\ \underset{n\times m}{f} = [U_1 U_2]\begin{bmatrix} f(1) \\ f(2) \end{bmatrix} = \underset{n\times k}{U(1)}\ \underset{k\times m}{f(1)} + \underset{n\times(n-k)}{U(2)}\ \underset{(n-k)\times m}{f(2)} = U(1) f(1) + e \tag{12}$$

Among them: $f(1)$ is the main factor $f(2)$ is a special factor and *e* is the residual part. Factor analysis expresses the original variable x_i as a linear combination f_j of new factors and requires that the variance of the specific factor be as small as possible.

The main factor is selected. After the residual *e* is omitted, the factor expression is:

$$\begin{cases} x_1 = u_{11} f_1 + u_{12} f_2 + \cdots + u_{1k} f_k \\ x_2 = u_{21} f_1 + u_{22} f_2 + \cdots + u_{2k} f_k \\ \vdots \qquad\quad \vdots \qquad\qquad \vdots \\ x_n = u_{n1} f_1 + u_{n2} f_2 + \cdots + u_{nk} f_k \end{cases} \tag{13}$$

In order to make the sum of the squares of the coefficients u_{ij} of *k* factors in each equation close to 1, the normalization process is needed, and the factor load $u_{ij}\sqrt{\lambda_j}$ is taken as the coefficient a_{ij}, that is, the load of *i*-th variable on the *j*-th main factor is:

$$a_{ij} = u_{ij}\sqrt{\lambda_j} \tag{14}$$

Therefore, the factor load matrix *A* is:

$$A = (a_{ij}) = \begin{bmatrix} u_{11}\sqrt{\lambda_1} & u_{12}\sqrt{\lambda_2} & \cdots & u_{1k}\sqrt{\lambda_k} \\ u_{21}\sqrt{\lambda_1} & u_{22}\sqrt{\lambda_2} & \cdots & u_{2k}\sqrt{\lambda_k} \\ \vdots & \vdots & \cdots & \vdots \\ u_{n1}\sqrt{\lambda_1} & u_{n2}\sqrt{\lambda_2} & \cdots & u_{nk}\sqrt{\lambda_k} \end{bmatrix} \tag{15}$$

Step 7: Establish a factor analysis model

The adjusted factor model is expressed as:

$$\begin{cases} x_1 = a_{11} f_1 + a_{12} f_2 + \cdots + a_{1k} f_k + a_1 \varepsilon_1 \\ x_2 = u_{21} f_1 + u_{22} f_2 + \cdots + u_{2k} f_k + a_2 \varepsilon_2 \\ \vdots \qquad\quad \vdots \qquad\qquad\qquad \vdots \\ x_n = u_{n1} f_1 + u_{n2} f_2 + \cdots + u_{nk} f_k + a_n \varepsilon_n \end{cases} \tag{16}$$

The last item in the above formula is only related to x_i, ε_i is the special factor. The matrix of the factor model is expressed as:

$$x = Af + a\varepsilon \tag{17}$$

Step 8: Explain the full name of the factor variable

Observing the elements in the factor load matrix, a_{ij} the absolute value of may have a large value on many columns of a row, or may have a large value on many rows of a column, indicating that an original variable may be at the same time, it has a large correlation with several factor variables. That is, the information of an original variable needs to be explained by several factor variables. Although a variable can explain much variable information, it can only explain a small part of the information of a variable, not a typical representation of a variable, and obscure the actual meaning of a factor variable. In order to make the factors better reflect the relationship between variables, under the premise of keeping the factorial axes orthogonal, the factorial axis is rotated, and the covariance of the covariance is minimized. To do this, select the orthogonal matrix *T* so that:

$$AT^{-1} = B, \qquad f^T T^T = g^T \tag{18}$$

Therefore:

$$Bg = AT^{-1}Tf = Af \tag{19}$$

In the above formula, *B* is a new factor load matrix and *g* is a new common factor. The choice of orthogonal matrix makes the elements of *B* as close to 1 or 0 as possible, so that some factors can reflect several variables, and other factors reflect other variables.

Conclusion

Factor analysis grouping indicators according to the correlation between indicators makes the correlation between the indicators within the group higher, and the correlation between different indicators is lower. Each set of indicators is replaced by a factor called the public factor; use less The linear function constructed by the number of common factors to describe each index of the original observation, through the analysis of these common factors to achieve the interpretation of complex economic and social issues.

References

[1] Akram Rostami, Hamid Abdollahi, Marcel Maeder, "Enhanced target factor analysis," Analytica Chimica Acta, vol. 911, no. 10, pp. 35-41, 2016.

[2] Marsha Simon, Youn-Jeng Choi, "Using factor analysis to validate the Clance Impostor Phenomenon Scale in sample of science, technology, engineering and mathematics doctoral students," Personality and Individual Differences, vol. 121, no. 15, pp. 173-175, 2018.

[3] Brecht Desplanques, Kris Demuynck, Jean-Pierre Martens, "Adaptive speaker diarization of broadcast news based on factor analysis," Computer Speech & Language, vol. 46, no. 9, pp. 72-93, 2017.

[4] Jianhua Zhao, Lei Shi, "Automated learning of factor analysis with complete and incomplete data," Computational Statistics & Data Analysis, vol. 72, no. 1, pp. 205-218, 2014.

[5] M. Y. Lu, Y. Pan, L. An, etal., "Comprehensive evaluation of quality traits of processed Apple Based on factor analysis," Jiangsu Journal of Agricultural Sciences, vol. 34, no. 1, pp. 130-137, 2018.

[6] S. W. Wang, P. Yan, "Research on Evaluation System of Public Science and Technology Service Ability in Hubei Province Based on Factor Analysis," Science and Technology Management Research, vol. 35, no. 2, pp. 58-63, 2015.