Development of a program to reduce the dimension of multidimensional data arrays by the method of principal components

E.A. Golovkova¹ and N.M. Ermakova²

¹ FSEI HPE «Angarsk State Technical University», Angarsk, Russia
² FSEI HPE «Irkutsk State Transport University», Irkutsk, Russia

E-mail: temnikova_ea@bk.ru

Abstract. In the article, the author discusses the importance of factor analysis methods, including the method of the main components, when studying multivariable information in the process of statistical analysis, with the aim of further modeling and predicting random processes. Key problems are described, when solving which they often turn to the method of the main components, as well as its algorithm, taking into account the fact that those interested in this article are familiar with the essence of statistical entimes. There is given a graphical representation of the author's software, which reads the file specified by the user with a sample of data, made in the form of a table, which automatically prepares them for further analysis and search of the main components. The program functionality allows: to automatically calculate the mathematical expectation, variance of a multidimensional random variable; a covariance matrix for describing the shape of a random variable from which it is possible to obtain its dimensions by automatically finding eigenvectors and eigenpars; reduce the size of the sample; recover data for further hypothesis building and modeling. The algorithm of the principal component method is presented schematically.

1. Introduction
In practice, it is often necessary to analyze the dependencies between different values, while the qualitative value of these dependencies tends to be expressed in quantitative form. To do this, they resort to the measurement of connections - functional dependence, which is essentially just abstraction. You can distinguish two main functions of measuring a relationship: determining the degree of dependence of a function on arguments based on a large sample of data and the degree of influence of other factors on it. This kind of relationship is usually represented as a correlation matrix, and in the case of multivariate random values in the form of a covariance matrix. It happens that in the correlation matrix there are non-diagonal values close in absolute magnitude to unity, which leads to a high degree of conventionality of the matrix, or degeneracy [1]. In such situations, factor analysis methods will be most effective.

2. The main tasks are solved by the principal component method
The main goals of factor analysis are: reducing the number of variables (reduction of data) and determining the structure of relationships between variables.

Knowing the dependencies and their power, you can express several features through one, thus simplifying the model. At the same time, it is simply impossible to avoid the loss of information, but it can be minimized due to the method of the main components [2]. This method approximates a
multidimensional set of observations to an ellipsoid, the semi-axes of which will be the main components. When projected onto such axes (dimensioning), the largest amount of information is retained.

3. Main component method algorithm used as the basis for software implementation

There is a source data matrix $X$ (multidimensional selection as a table). The number of monitoring observations (product of the number of observations by the number of their parameters) determines the number of rows of the table ($n$), number of variables – dimension of the feature space ($p$) [3]. In our work, matrix $X$ has a size $130 \times 9$.

To describe a random variable, we use mathematical expectation $\overline{x}_{ij}$ (sets the position of the random variable) and variance (standard deviations $\sigma_j$, which sets its spread). The dispersion is very sensitive to scaling. If the characteristic units vary widely in order, you must standardize them.

$$\overline{x}_{ij} = \frac{1}{n} \sum_{i=1}^{n} x_{ij},$$ 

(1)

where $x_{ij}$ – takeoff values (at the intersection of rows and columns);

$$\sigma_j = \left(\frac{\sum_{i=1}^{n} (x_{ij} - \overline{x}_{ij})^2}{(n-1)}\right)^{1/2},$$

(2)

The calculation of the standardized $z_{ij}$ values of the $Z$ data matrix is carried out according to the formula given below:

$$z_{ij} = \left(\frac{x_{ij} - \overline{x}_{ij}}{\sigma_j}\right).$$

(3)

To describe the form of a random vector, a covariance matrix is needed in which the ($j, k$) element is a correlation of features ($x_j, x_k$). See the formula (4).

$$\text{cov}(x_j, x_k) = \frac{\sum_{i=1}^{n} (x_{ij} - \overline{x}_{ij})(x_{ik} - \overline{x}_{ik})}{(n-1)},$$

(4)

$$r_{jk} = \frac{\text{cov}(z_j, z_k)}{\sigma_j \sigma_k} = \frac{\text{cov}(x_j, x_k)}{\sigma_j \sigma_k}.$$ 

(5)

The covariance matrix describes the spread of a random variable, as well as the dispersion. Calculate the correlation matrix by the formula

$$R = \left(\frac{Z - Z^T}{n-1}\right),$$

(6)

where $Z^T$ – transposed matrix of standardized values.

The result of automatic matrix $R$ calculation in the program is shown in figure 1.

Figure 2 show that there are strong correlations; this confirms the possibility of representing our sample in a space smaller than nine-dimensional size.

Further transformations are aimed at finding an optimal vector subspace reflecting sample data with minimal loss.

Next, we need to find vectors in which the variance of the projection of our sample on them is maximized. To do this, do the following: calculation of characteristic matrix and eigenvalues ($\lambda$).

A number $\lambda$ is called the eigenvalue of the matrix $R$, in the case of the existence of a non-zero vector $x$, under which the condition $Rx = \lambda x$ is me [4].
Let's compose the characteristic polynomial of the matrix $R$, taking the $\gamma$ from its elements on the main diagonal, then find the determinant of the correlation matrix (the program implementation of this process is shown in figure 2). So we get the expression:

$$
P(\lambda) = -\lambda^9 + 9\lambda^8 - 25.58305\lambda^7 + 34.24075\lambda^6 - 24.43720\lambda^5 + 9.59643\lambda^4 - 2.02409\lambda^3 + 0.21277\lambda^2 - 0.00937\lambda + 0.00013 = 0.
$$

The result of automatic isolation of the roots of the characteristic polynomial is shown in figure 3 (this is the eigenvalues of the $R$ matrix).
Figure 3. The matrix eigenvalues.

Now we find the eigenvectors $x_1, x_2, ..., x_n$, solving a system of linear equations of the form: $Rx_i = \lambda_i x_i$ for each $i$ eigenvalue. Which can be converted to this view ($E$ is the unit matrix):

$$\text{Det}(R - \lambda_i E) = 0$$

(7)

The direction of maximum dispersion in the projection always coincides with the eigenvector having a maximum eigenvalue equal to the value of this dispersion. The components of the eigenvectors are loads on the source variables, which allows analyse the contribution of each characteristic to the value of the corresponding main component. Result of mathematical calculations in the program is shown in figure 4.

Figure 4. Eigenvectors of the matrix, calculated using the software application.

Note that the sum of the squares of the eigenvector components must be one. We have it, which indicates that the contribution to the magnitude of the main component can be estimated by calculating the square of its coefficient, and by the sign in front of it. Therefore, the direction of maximum dispersion in the projection always coincides with the eigenvector having a maximum eigenvalue equal to the value of this dispersion.

Further for decrease in dimension of sample it is necessary to find the equation and to build the line of regression as the greatest vector has the direction similar to the line of regression and, having projected on him our sample, we will lose information comparable with the sum of residual members of regression.

The flow diagram of the program operation algorithm is shown in figure 5.
4. Conclusion
In our article, we described and described the algorithm of the method of the main components, presented its software implementation and evaluated its operability using the example of real data. By analysing the results, we can confidently say that using the method of the main components, it is possible to significantly reduce the volume of the multivariable data matrix without losing their significant importance (essence).

And the program we developed made it possible to speed up and automate the process of labour-intensive calculations, as well as reduce the likelihood of an error in the results.

![Block Diagram](image)

**Figure 5.** The block diagram of the program operation algorithm.
References

[1] Mokeev V V 2014 Method of main components and method of eigenstates in problems of analysis and forecasting (Chelyabinsk. SUSU) p 138

[2] Dobrynina N F 2008 Modern science-intensive technologies 4 pp 95-6

[3] Zabuga G A, Temnikova E A, Efimova N V and Grebenshchikova V I 2017 System analysis. Modeling 1 (53) pp 104-10

[4] Rush E A and Davydenko A Yu 2004 Complex geocological assessment of the impact of an impact source of mercury pollution on environmental objects (Irkutsk: Irstu) p 111