Bulletin of the *Transilvania* University of Braşov
Series V: Economic Sciences • Vol. 8 (57) No. 2 - 2015

The generalized approach to multidimensional scaling

Ludmila MALYARETZ[1], Oleksandr DOROKHOV[1],
Vladimir PONOMARENKO[1]

Abstract: *Often in practice of solving economic problems there is a need to analyse objects with properties that are measured in different non-metric scales. Most popular mathematical approach of processing that properties is multidimensional scaling method. There are metric and non-metric scaling. The most reasonable of them is the Torgerson metric method, but it is supposed to have metric scale input variables. The paper proposes a modified method of factor analysis, which allows non-metric data as input. The paper presents arguments that the proposed modification is the improvement of the method of multidimensional scaling. The proposed method is less time-consuming than the Torgerson method, but is mathematically justified as well. So the mathematical explanation of the generalized multidimensional scaling method has been described. This approach broadens the range of possible ways to describe objects in economics by means of factors measured on different scales.*

Key-words: *methodology of multidimensional scaling, generalization, universality, metric factors, non-metric factors.*

1. Introduction

In conditions of indeterminacy, objects in economics are described by means of non-metric scales. The further research of these objects requires the use of mathematical methods. It leads to certain problems caused by the limited choice of methods that allow ordinal numbers or nominations as the original data (Malyaretz 2010, Ponomarenko 2007). The multidimensional scaling is the method that makes it possible to perform the analysis of non-metric factors. The multidimensional scaling can be metric and non-metric (Egorshin 2007).

Torgerson's scaling is the most valid of them. The elements of this method are also used in other methods of multidimensional scaling (Ponomarenko 2009). The existing algorithms of the multidimensional scaling method have a lot of drawbacks: the heuristic validity, laboriousness, the limit of two stimuli, the visual presentation on a plane only etc (Malyaretz 2007, Manly 2004).

[1] Department of Informations Systems Kharkiv National University of Economics (KhNUE), Kharkiv, Ukraine; aleks.dorokhov@meta.ua

The purpose of multidimensional scaling correlates with factor and cluster analysis. Like with the principal components method, the goal of multidimensional scaling lies in the construction of reduced metric space with generalized coordinates.

The principal components method is mathematically valid, and any deviations from it meant for non-metric data generalization can only hinder the results of the analysis (Davison 1988, Davison 2009, Stern 2004, Härdle 2007). Thus, the principal components method can be considered a standard for comparison with any other multidimensional scaling method, if data allows using both methods.

2. First variant of the principal components method

It is known that standardization is used in the principal components method, and any other variants can only worsen the exhaustion principle (Egorshin 1998, Malyaretz 2006).

Let consider the matrix of initial metric data with size $n \times m$ (n – number of objects or observations, m – number of attributes of the objects). For it may be calculated standardized data matrix Z with the same dimensions $n \times m$ and Z'- the transposed matrix with matrix size $m \times n$.

The $F_i, i = \overline{1,m}$ components are linear combinations of standardized factors $Z_j, j = \overline{1,m}$ and coefficients u_{ij}, which are the eigenvectors of the correlation matrix between factors $R = \dfrac{Z'Z}{n}$:

$$F_i = u_{i1}Z_1 + u_{i2}Z_2 + ... + u_{im}Z_m,$$

where n - is the number of observations (objects);

m - the number of factors;

R - a matrix the size of $m \times m$;

U - matrix with dimensions $m \times m$ of eigenvectors of the correlation matrix, which are defined by the matrix equation $RU = UD(\lambda)$, where $D(\lambda)$ is a diagonal matrix with eigenvalues λ_i of the correlation matrix R (Malyaretz, 2010).

Unlike the initial standardized factors Z_j, the dispersions of the components F_i are different and are the eigenvalues λ_i. It is known that the correlation matrix is positively defined, meaning that its eigenvalues are nonnegative and in total are equal to the numbers of factors m.

The first components are more important than any separate factor, their dispersions (eigenvalues λ_i) are greater than 1. On the contrary, the rest of the components almost don't vary, their dispersions (eigenvalues) are close to zero. We

can discard these components and get a limited space of the principal components, the number of which is less than the number of the initial factors $k \le m$.

If we leave the first k columns, corresponding to the main components, in the matrix U, than we can calculate the values of these components with the matrix product of $F=ZU_T$ (the dimensions of the matrixes are: $n \times k$, $n \times m$, $m \times k$ and U_T is corresponding truncated matrix). It is known that the eigenvectors of the correlation matrix are mutually orthogonal. Hence, they are normalized so that the norms of the vectors were equal 1. In the factor analysis, the eigenvectors are normalized so that their norms were equal the eigenvalues λ_i. The matrix of vectors normalized in such a fashion is called the matrix of factor stresses $A_i = U_i \lambda_i$ and is used to classify factors in such groups.

3. Second variant of the principal components method

There is another variation of the principal components method, which due to the number of mathematical operations is considerably more labor-consuming than the main one stated above, but this modification makes it possible to synthesize the principal components method for other kinds of data, e.g. for non-metric data.

According to the other variant of the method, after the initial standardization of the factors it is necessary to build a correlation matrix between objects $RO = \dfrac{ZZ'}{m}$, to find its eigenvectors V_i and to normalize them so that the new norms were equal to the eigenvalues of the correlation matrix RO, e.g. we must calculate its matrix of factor stresses. The dimensions of the matrix RO are $n \times n$, which is considerably greater than $m \times m$ of the correlation matrix R between the factors.

To prove that the non-nil eigenvalues for the matrixes $Z'Z$ and ZZ' are same, we use the singular decomposition of a random rectangular matrix (Forsythe 1977).

The singular figures σ_i of the rectangular matrix Z with dimensions $n \times m$ are defined by the following system of matrix equations:

$$\begin{cases} ZU = V\Delta(\sigma) \\ Z'V = U\Delta'(\sigma) \end{cases}$$

where U - is a square matrix with dimensions $m \times m$;

V - is a square matrix with dimensions $n \times n$;

$\Delta(\sigma)$ - a rectangular matrix with dimensions $n \times m$, in which only the

elements of the main diagonal, where the singular figures σ_i are, are other than zero.

We multiply the first matrix equation of the system of the singular figures calculation by Z' and consider the second matrix equation of the system:

$$Z'ZU = Z'V\Delta(\sigma) = U\Delta'(\sigma) = UD\left(\sigma^2\right)$$

From the matrix equation $(Z'Z)U = UD(\lambda)$ we see that $\lambda_i = (\sigma_i)^2$ are eigenvalues, and U_i are eigenvectors of the symmetrically positive defined matrix $Z'Z$. Therefore, all the eigenvalues λ_i are actual nonnegative, and the eigenvectors U_i are mutually orthogonal.

We multiply the second matrix equation of the system for the singular figures calculation on the left by Z and consider the first matrix equation of the system:

$$Z'ZV = Z'U\Delta(\sigma) = V\Delta(\sigma)\Delta'(\sigma) = VD^*\left(\sigma^2\right)$$

where $\Delta(\sigma)\Delta'(\sigma) = D^*\left(\sigma^2\right)$ - is a diagonal matrix with dimensions $n \times n$, on the main diagonal of which there is m squares of singular figures, and the other elements equal zero.

From the matrix equation $(ZZ')V = VD^*\left(\sigma^2\right)$ we see that $\lambda_i = (\sigma_i)^2$ are non-nil eigenvalues, and V_i are the eigenvectors of the symmetrically positive defined matrix ZZ' . Therefore, all the eigenvalues λ_i are actual nonnegative, and the eigenvectors V_i are mutually orthogonal.

Thus, it is proven that non-nil positive values of matrixes $Z'Z, ZZ'$ are similar and equal the squares of the singular figures. For the correlation matrix between the factors $R = \frac{Z'Z}{n}$ the eigenvalues will be n times less and in total will equal m; for the correlation matrix between the objects $RO = \frac{ZZ'}{m}$ the eigenvalues will be m times less and in total will equal n, the eigenvectors will stay the same.

Now we can formulate the most important conclusion that the matrix of factor stresses for the matrix RO is in fact the matrix of the values of the main factors $F = ZU$. Indeed, form the first matrix equation of the system we get $F = ZU = V\Delta(\sigma)$, and $F_i = V_i\sqrt{\lambda_i}$ - the columns of the matrix of factor stresses for $ZZ' \sim RO$. Therefore, we have shown equivalence of the calculation results for both variants of the principal components method.

However, the second universal variant is much more labor-consuming than the first one, but it may be used for non-metric data. For ordinal data some prefer the cumbersome iterative Kruskal-Wish method with the non-metric series of Lingoes and Guttman (Terekhina, 1986). But there is no theoretical foundation for this procedure, furthermore this method requires a great amount of calculations but also is the heuristic.

4. Proving of the generalized approach of multidimensional scaling

The Torgerson method in the metric multi-dimensional scaling is congruent almost completely with the other variants of the principal components method, but, form the calculating point of view, those methods are not that similar (Davison 1988).

In the Torgerson method, after the matrix of metric data X with dimensions $n \times m$ we build a matrix of squares of Euclidean distances D^2 with dimensions $n \times n$. Initially, the values of all the factors must be normalized.

If we take the dispersions of the factors as the norms, the elements of the matrix D^2 will be found through the formula:

$$d_{ij}^2 = \sum_{k=1}^{m} \left(z_{ik} - z_{jk} \right)^2$$

where z_{ik}, z_{jk} are the values of the standardized factor Z_k for the objects i, j (additional centering of the factors doesn't change the distance between the objects).

For the matrix D^2 we find the averages of the rows $d_{i\bullet}^2$ and the columns $d_{\bullet j}^2$ and the mutual average $d_{\bullet\bullet}^2$; then by means of double centering we turn to the matrix D^*, the elements of which are defined through the formula:

$$d_{ij}^* = -\frac{1}{2}\left(d_{ij}^2 - d_{i\bullet}^2 - d_{\bullet j}^2 + d_{\bullet\bullet}^2 \right)$$

Now let us prove that the matrix D^* is the matrix ZZ'.

We express the averages $d_{i\bullet}^2$, $d_{\bullet j}^2$, $d_{\bullet\bullet}^2$ through the initial data (standardized). We transform the data by the rows of the matrix $d_{i\bullet}^2$:

$$d_{i\bullet}^2 = \frac{1}{n}\sum_{j=1}^{n} d_{ij}^2 = \frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{m}\left(z_{ik} - z_{jk}\right)^2 =$$

$$= \frac{1}{n}\sum_{k=1}^{m}\sum_{j=1}^{n} z_{ik}^2 + \frac{1}{n}\sum_{k=1}^{m}\sum_{j=1}^{n} z_{jk}^2 - \frac{2}{n}\sum_{k=1}^{m}\sum_{j=1}^{n} z_{ik}\, z_{jk} =$$

$$= \sum_{k=1}^{m} z_{ik}^2 + \sum_{k=1}^{m} z_{\bullet k}^2 - 0,$$

where with $z_{\bullet k}^2$ the average squares of standardized values of each factor are designated:

$z_{\bullet k}^2 = \frac{1}{n}\sum_{j=1}^{n} z_{jk}^2$; the last item in the formula $d_{i\bullet}^2$ equals zero $\sum_{k=1}^{m}\sum_{j=1}^{n} z_{ik} z_{jk} = 0$

since for standardized factors $\sum_{j=1}^{n} z_{jk} = 0$.

Similarly the averages of the columns $d_{\bullet j}^2$ are transformed.

$$d_{\bullet j}^2 = \frac{1}{n}\sum_{i=1}^{n} d_{ij}^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{m}\left(z_{ik} - z_{jk}\right)^2 =$$

$$= \frac{1}{n}\sum_{k=1}^{m}\sum_{i=1}^{n} z_{ik}^2 + \frac{1}{n}\sum_{k=1}^{m}\sum_{i=1}^{n} z_{jk}^2 - \frac{2}{n}\sum_{k=1}^{m}\sum_{i=1}^{n} z_{ik} z_{jk} =$$

$$= \sum_{k=1}^{m} z_{\bullet k}^2 + \sum_{k=1}^{m} z_{jk}^2 - 0.$$

The mutual average $d_{\bullet\bullet}^2$ can be found through one of the following equivalent formulas:

$$d_{\bullet\bullet}^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{i=1}^{n} d_{ij}^2 = \frac{1}{n}\sum_{i=1}^{n} d_{i\bullet}^2 = \frac{1}{n}\sum_{j=1}^{n} d_{\bullet j}^2$$

The transformation of any of these formulas leads to the following expression:

$$d_{\bullet\bullet}^2 = \sum_{k=1}^{m} z_{\bullet k}^2 + \sum_{k=1}^{m} z_{\bullet k}^2 = 2\sum_{k=1}^{m} z_{\bullet k}^2$$

Finally we transform the elements of the matrix D^* (after double centering of the matrix D^2):

$$d_{ij}^{*} = -\frac{1}{2}\left(d_{ij}^{2} - d_{i\bullet}^{2} - d_{\bullet j}^{2} + d_{\bullet\bullet}^{2}\right) =$$

$$= -\frac{1}{2}\left[\begin{array}{l}\left(\sum_{k=1}^{m} z_{ik}^{2} + \sum_{k=1}^{m} z_{jk}^{2} - 2\sum_{k=1}^{m} z_{ik} z_{jk}\right) - \\ -\left(\sum_{k=1}^{m} z_{ik}^{2} + \sum_{k=1}^{m} z_{\bullet k}^{2}\right) - \left(\sum_{k=1}^{m} z_{\bullet k}^{2} + \sum_{k=1}^{m} z_{jk}^{2}\right) + 2\sum_{k=1}^{m} z_{\bullet k}^{2}\end{array}\right] = \sum_{k=1}^{m} z_{ik} z_{jk}.$$

Therefore, it turned out that the elements of the matrix D^{*} are equal to scalar products of the rows of the matrix Z (the matrix of standardized factors). To put it simple, the matrix D^{*} can be presented as a matrix product of mutually transposed matrixes: $D^{*} = ZZ'$. The further computations repeat the steps of the universal method (the second variant of the principal components method).

For the matrix D^{*} we find eigenvalues that correspond to the eigenvectors V_i, which are normalized so that their norms would be equal to the eigenvalues. Several first normalized vectors (usually the first two) are taken for the limited vector space. This issue has already been cleared. It is necessary to leave enough generalized coordinates that the sum of their eigenvalues wouldn't be less than 70% of mn; in addition, all these eigenvalues must be no less than m.

In scientific sources it is noted that with other factors normalization or other measure of distances, the matrix D^{*} loses positive definiteness, and negative eigenvalues appear (Mukhopadhyay, 2008). It is considered acceptable, if these negative values are low. We should use the expert opinion on the admissibility of minor abnormalities. By the way, the Kruskal-Wish method also infringes upon the condition of the positive definiteness of the similarity matrixes.

4. Conclusions

As the initial matrix in the multidimensional scaling method for metric factors it is recommended to use the correlation matrix, for ordinal factors – Spearman's rank correlation matrix, for nominal data – the similarity matrix based on Hamming's measure.

As the result of the calculations of the generalized method of multidimensional scaling we get a new space with new coordinates, where the analysis of the objects can be continued with the methods of cluster analysis. Thereby, we have proven the foundation of the generalized method of multidimensional scaling for the analysis of objects in economics, originally described in the space of different factors, particularly nominal ones.

5. References

Davison, M. 1988. *Multidimensional scaling*. N.Y.: Jonh Wiley & Sons.

Egorshin, O. 1998. *Methods of multivariate statistical analysis*. Kiev: IZMN. (in ukrainian).

Egorshin, O. 2007. "The universal method of multidimensional scaling." *Economics of Development* 4: 43–47 (in ukrainian).

Forsythe, G. 1977. *Computer methods for mathematical computations*. N.Y.: Prentice-Hall.

Härdle, W. 2007. *Applied multivariate statistical analysis.* Berlin: Springer.

Malyaretz, L. 2006. "Application of methods of nominal quality attributes statistical analysis of the objects in the economy." *Economic, Management, Entrepreneurship* 16: 17–30 (in ukrainian).

Malyaretz, L. 2006. *Measuring attributes of the objects in the economy*. Kharkiv: KhNUE (in ukrainian).

Malyaretz, L. 2007. "Development of a universal, objective measures closeness of the connection attributes of the objects in the economy." *Business Inform* 9: 114–118 (in russian).

Malyaretz, L. 2010. "Current problems in the development of econometric models." *Economics of Development* 6: 154–158 (in ukrainian).

Malyaretz, L. 2010. "Theoretical basis of universal method of multidimensional scaling." *Business Inform* 4: 51–53 (in ukrainian).

Manly, B. 2004. *Multivariate statistical methods*. N.Y.: Chapman & Hall/CRC press.

Mukhopadhyay, P. 2008. *Multivariate statistical analysis.* Singapore: World Scientific Publishing Co.

Ponomarenko, V. 2007. "Summary of methodology of mathematical modeling identification of socio-economic systems." *Economics of Development* 3: 43–47 (in ukrainian).

Ponomarenko, V. 2009. *The analysis of data in studies of socio-economic systems*. Kharkiv: INGEC Publishing House (in ukrainian).

Sengupta, A. 2009. *Advances in multivariate statistical methods*. Singapore: World Scientific Publishing Co.

Stern, R. 2004. *Good statistical practice for natural resources research*. Cambridge: CABI Publishing.

Terekhina, A. 1986. *The data analysis methods of multidimensional scaling*. M.: Nauka (in russian).