NEURAL NETWORK METHOD USED FOR INTERPRETATION OF ANALYTICAL CONCENTRATION DATA

C. Oprea[1], I.A. Oprea[1,2]

[1]*Joint Institute for Nuclear Research (JINR), Dubna 141980, Russian Federation*
[2]*Bucharest State University, 76900 Bucharest, Romania*

Abstract. The present paper study how the neural network approach can be practically and efficiently applied to factor analysis. Until now several methods of factor analysis such as the centroid method, α factor analysis, principal component analysis, maximum likelihood factor analysis and image factor analysis are known. In order to find a compromise among many criteria in evaluation of the results given by the above approximations a factor analysis extended by neural network model has to be developed.

The neural network technique is a computer procedure which simulates the informational processes from biological systems. It was established that the factor analysis problem can be solved using an energy function of a neural network.

Factor analysis based neural-network approach was developed in view of characterizing chemical composition of environmental samples evaluated by multielemental analytical procedures. The approach developed in our research enabled the separation between various sources of elements "cached" inside the large concentration data sets.

Keywords: neural network, factor analysis,

INTRODUCTION.

Factor analysis is a part of statistical multivariate analysis and it is based on the multidimensional normal distribution of the studied properties and characteristics of the analyzed object. Factor analysis extracts the inner hidden properties of the objects from the covariant and correlation matrix obtained from measurements.

For the studied object (phenomena or system) a series of *n* measurement of *m* properties are realized. These measurements together with the properties form a matrix $\{Y\}$ with $m \times n$ dimension. The values of Y are multidimensional and are conditioned by some objective causes as named "factors". These factors are however hidden and it is not possible to measure them in a direct way but factor analysis is one of the methods that can evidence them.

The main purposes of the factor analysis are:

a) to determine the number of factors and their relative intensity;
b) to evidence the factor structure of the studied object;
c) to determine the influence of the extracted factors to the properties of the studied object;
d) to recreate in the factors coordinate space the image of the object using the calculated values of the factors.

Any type of factor problem is starting with the Y matrix of the initial data. From the initial Y matrix using simple mathematical procedure one obtains the standard (or normalized) Z matrix of the initial data.

$$Z = A \cdot P \tag{1}$$

where:

Z=standard $m \times n$ matrix, A=factor pattern $m \times r$ matrix, P= factor loadings $r \times n$ matrix, r= factors number

After that from standard matrix of initial data we obtain the correlation matrix R $(m \times m)$. The next stage is the so called “problem of communalities” and the obtained values for communalities will replace the diagonal of the R matrix (On the diagonal of the R matrix we have only elements equal to *1*). The new obtained matrix is R_h. Using different methods from the reduced correlation matrix R_h we obtain the A matrix from where the factors are extracted. The relation between A and R_h is given down (A^T is the transposed A matrix) and represents the fundamental theorem of factor analysis:

$$R_h = A \cdot A^T \qquad (2)$$

The relation (2) has many solutions and from them we must choose one solution using different procedures using a factor rotation. As a result of rotation procedure we obtain a new A matrix usually noted with V.

$$V = A \cdot G \qquad (3)$$

with:

$$R_h = A \cdot A^T = A \cdot I \cdot A^T = A \cdot \left(G \cdot G^T\right) \cdot A^T = (A \cdot G) \cdot (G \cdot A)^T = V \cdot V^T \qquad (4)$$

$G =$ orthogonal rotation matrix with the properties:

$G \cdot G^{-1} = G \cdot G^T = I$ ($I =$ the unit matrix).

After rotation in the space of factors the communalities are not changing. There are many solutions for the G rotation matrix but from the set of solutions we choose only those which simplifying the factor structure. The rotation problem in factor analysis consists in finding an orthogonal transformation of the initial factor loadings so that by rotation we obtain a simple structure that can be easily interpreted. The most popular orthogonal transformations are the quartimax and varimax procedure with Kaiser normalization.

The last stage is the problem of the estimation of the factors values for each measurement (P factor loading matrix determination).

There are a few methods of factor analysis:

- the centroid method (used before computer era);
- alpha factor analysis;
- principal component analysis;
- maximum likelihood factor analysis;
- image factor analysis.

One of the most used and effective factor analysis methods is the principal component analysis based on the eigenvalue and eigenvector technique. The maximum likelihood factor analysis is based on the minimization of a function representing the “distance” between the observed covariance matrix and predicted values of this matrix. The relations (1-4) very schematic describe the principal component analysis.

The listed methods give different results and some authors suggest that in order to obtaining reliable results it is better to use several of the above methods.

The SunFA method.

Based on the above suggestion we will use the principal component analysis and a neural network approach (as the SunFA approach).

It is well-known that the neural network is a computer procedure which simulates the informational processes from biological systems. The biological system takes the information from outside through the nervous system which consists from neurons. The neural networks were applied with success in various fields of science.

The factor analysis problem can be solved using an energy function of a neural network. This energy function is obtaining starting from fundamental theorem of factor analysis (2) and according to [7] has the form (*{r}* and *{a}* are the matrix elements of correlation matrix *R* and factors matrix *A* respectively):

$$E = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \left(r_{ij} - \sum_{k=1}^{r} a_{ik} a_{jk} \right)^2 \tag{5}$$

For a better understanding of relation (5) we give the energy function in the case of one factor *(r=1)* [7]:

$$E = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \left(r_{ij} - a_{i1}a_{j1}\right)^2 = \left(r_{12} - a_{11}a_{21}\right)^2 + \left(r_{13} - a_{11}a_{31}\right)^2 + ... + \left(r_{1m} - a_{11}a_{m1}\right)^2 + \\ + \left(r_{23} - a_{21}a_{31}\right)^2 + ... + \left(r_{2p} - a_{21}a_{p1}\right)^2 + ... + \left(r_{p-1p} - a_{p-11}a_{p1}\right)^2 \tag{6}$$

In the SunFA method we start with analyze of the correlation matrix obtained from the initial data in order to determine the number of factors. In the beginning the factor loadings are considered random numbers in the interval *(0,1)* and a temporary correlation matrix is created. From this temporary correlation matrix by usual procedure is obtained the factor matrix. With the factor loadings the energy function is calculated according to relation (5). After this energy is minimized using a procedure described in [6]. A new set of factor loadings and a new value for energy are obtained. This procedure is repeated until a so called "stable state" is obtained. When we reach the stable state it is possible that the number of factors to be different in comparison with the number of factor obtained by usual factor analysis and rotation procedures. It is considered that the stable state is reached when the energy is closed to zero and for this value of energy a new number of factors and a new matrix of factor loadings can be obtained.

SOME APPLICATIONS.

We now will proceed to the analysis of some applications and first will be the classical problem in factor analysis the boxes problem. In the boxes problem it is supposed that you don't now the main properties of a rectangular box and with the help of factor analysis we try to find them. The properties of a box are the three sides *{x, y, z}* – length, width and height.

Making some measurements of type *{x^2, y^2, z^2, xy, xz, yz, xyz, e^x, e^y, e^z,…….}* to a set large number of rectangular boxes *(20, 30, 40,…..)* and applying one of the method of FA (in our case the principal component analysis) we obtain *3* factors. These 3 factors can be interpreted as the main properties of a rectangular box and they are the *3* sides of a box – the length, width and height. Further also we have applied the SunFA method in order to see if the number of factor will be modified. As it is expected the number of factor remains the same.

Another application more close to the reality and suggested by the boxes problem is that of system of Gaussian source of pollution. This system of pollution sources are situated at some distance between them and each source emits a spectrum of elements but each spectrum is different to each source and they have no common elements. Using the distribution of the concentration of emitted elements we simulate a set of sample collection. Depending on the area of collection the factor analysis give us the number of the sources of pollution and this is the expected results. However the results of factor analysis depend on the form of area collection, how far is the collection area from the system of pollution sources. For example if the collection area departs from sources system the number of factor is decreasing starting from the number of sources until to one when the collection area is to far from the sources. Here SunFA method help us to improve our results. Applying this new method for the cases when the number of factors is decreasing by factors analysis procedure we note that number of factors remains the same for a large number of areas collection and is equal to the number of sources which is the expected results. Naturally if the collection area is to far from the sources the SunFA method indicates the decreasing of the number of factors but this is happen in fewer cases that in the case of applying only factor analysis.

CONCLUSIONS.

In this paper we have presented only preliminary results. The analyzed cases were simulated on the computer and for this was created computer programs for factors analysis by principal component analysis and factor rotation. These procedures was used after in SunFA method in the minimization of so called "energy function".

The factor analysis and neural network approach led to the SunFA procedure that is a powerful method for describing and characterizing the chemical composition of environmental samples. The generality of the methods allows to be applied not only in the field of nuclear analytical method but also in other field of sciences.

References

[1] A. I. Bahtin, N. M. Nizamutdinov, N. M. Hasanova, E. M. Nurieva, Factornyi Analiz v Geologyi, Tutorial for Laboratories Lessons, University of Kazan, Kazan, 2007 (in Russian)

[2] K. Uberla, Faktornyi Analiz, Matematicheskie Metody za Rubezhom Publishing House, Moskva, 1980 (in Russian)

[3] H. F. Kaiser, Psyhometrika, vol.23, (3), September, 1958

[4] L. Knusel, http://www.stat.uni-muenchen.de/~knuesel/rotation/knuesel.pdf

[5] L. L. Thurstone, Multiple Factor Analysis, Chicago, University of Chicago Press, 1947

[6] J. O. Kim, C. W. Muller, Introduction to factor analysis: what it is and how to do it. Newbury Park, London: Sage Publications, 1978