Comparison of Principal Analysis and the Tucker3 Model

A Case Study

Helena Morais[a,b], Cristina Ramos[a,b], Esther Forgács[c], Annamaria Jakab[c], Tibor Cserháti[a]*, José Oliviera[b], Tibor Illés[d] and Zoltán Illés[e]

[a] National Agronomical Station, Quinta do Marques, 2784-505 Oeiras, Portugal
[b] New University of Lisbon, 2825 Monte da Caparica, Portugal
[c] Institute of Chemistry, Chemical Research Center, Hungarian Academy of Sciences, P.O. Box 17, 1525 Budapest, Hungary
[d] Department of Operations Research, Eötvös Loránt University of Sciences, Budapest, Hungary
[e] Central European University, Budapest, Hungary

Abstract

A three-ways array data matrix consisting of the activity data of laccase enzyme has been evaluated by both principal component analyses (PCA) and Tucker3 model. Activity data have been determined in 28 culture media, at 6 sampling times and by four strains of *Lentinus edodes*. PCA has been carried out three times one of the factors being always the variables and the other two factors being the observations. The dimensionality of the matrices of loadings calculated by PCAs and those of component matrices of Tucker3 model has been reduced to two by the nonlinear mapping technique. It has been found that the dimensionality of component matrices for Tucker3 model can be predicted from the results of PCAs. Linear regression analyses indicated that the distribution of the original data points on the two-dimensional nonlinear maps considerably depends on the fact that the data have been calculated by PCA or by Tucker3 model.

1 Introduction

Principal component analysis (PCA) a versatile and easy-to-use multivariate mathematical-statistical method has been developed to contribute to the extraction of maximal information from large data matrices containing numerous columns and rows [1]. As PCA is only suitable for the elucidation of the similarities and dissimilarities among the columns and rows of two-dimensional data matrices it cannot be employed for the evaluation of matrices of higher dimensions. Theoretically, a three dimensional matrix composed of factors I, II, and III can be evaluated by carrying out PCA three times: 1. Factor I being the variable and factors II and III the observations; 2. Factor II being the variable and factors I and III the observations; 3. Factor III being the variable and factors I and II the observations. It can be assumed that the matrices of PC loadings of the three PCAs may be similar to the corresponding matrices calculated by three-dimensional calculation method. However, this procedure is time consuming and to the best of our knowledge the results of such a pseudo three dimensional PCA have never been compared with those of real three-dimensional techniques. The development of various multiway mathematical statistical methods such as parallel factor analysis (PARAFAC) [2], its tutorial [3], and application in chromatography [4], a direct fitting algorithm [5], and canonical decomposition (CANDECOMP) [6] have been reported. The Tucker3 model of three-way PCA [7, 8] and its improved version [9] are also suitable for this purpose. The Tucker3 model computes three orthogonal matrices with lower dimensions than the original data matrix such the variance explained by the reduced matrices being as high as possible. The usefulness of these calculation methods have been proven in many fields of up-to-date data evaluation such as in the analysis of retention data in normal phase high-performance liquid chromatography [10], environmental analysis [11], person perception analysis in psychology [12], medium-rank second order calibration [13], and enzyme production [14, 15].

As the resulting matrices of loadings and variables of both PCA and Tucker3 model are generally multidimensional they cannot be evaluated by visual methods except in the cases when the first two sets of PC loadings and components

* To receive all correspondence: Tibor Cserháti, Institute of Chemistry, Chemical Research Center, Hungarian Academy of Sciences, P.O. Box 17, 1525 Budapest, Hungary. Phone: 36-1-325-7900; Fax: 36-1-325-7554; E-mail: forgacs@cric.chemres.hu

Key words: laccase activity, PCA, Tucker3 model, nonlinear mapping

  1611-020X/03/0406-0449 $ 17.50+.50/0

Table 1. Ratio of variance explained of the three dimensional data matrix by two-dimensional principal component analyses (PCA). L.e. = strains of *Lentinus edodes*

PCA1: variables = 28 compositions of fermentation media; observations = 6 sampling times × 4 strains of L.e.			
No of principal component	Eigenvalues	Variance explained%	Total variance explained%
1	22.08	78.88	78.88
2	2.06	7.36	86.24
3	1.40	5.01	91.25
4	1.05	3.74	94.98
5	0.60	2.15	97.13
PCA2: variables = 6 sampling times; observations = 28 compositions of fermentation media × 4 strains of L.e.			
No of principal component	Eigenvalues	Variance explained%	Total variance explained%
1	3.82	63.63	63.63
2	1.04	17.33	80.96
3	0.60	10.01	90.98
PCA3: variables = 4 strains of; observations = 28 compositions of fermentation media × 6 sampling times			
No of principal component	Eigenvalues	Variance explained%	Total variance explained%
1	1.41	35.26	35.26
2	1.04	26.05	61.32
3	0.94	23.57	84.89

explain the overwhelming majority of variance. Nonlinear mapping technique (NLMAP) and varimax rotation have been developed for the reduction of the dimensionality of such multidimensional matrices [16]. Two dimensional NLMAP projects the points scattered in the multidimensional space on a plane such a manner that the distances among the points in the multidimensional space and on the plane being as similar as possible. Varimax rotation modifies the original orthogonality of the axes of PCA loadings obtaining maximal PCA loadings around new orthogonal axes.

The objectives of the study were the application PCA and Tucker3 model for the evaluation of the dependence of the production of laccase enzyme on the type of white rot fungi, fermentation time, and the composition of the culture media, and the comparison of the results of PCA and Tucker3 model. To the best of our knowledge the comparison of traditional PCA and Tucker3 model has never been previously carried out using data matrices of biotechnological importance. We are well aware that two factors of the three-ways array are relatively small. However, the time-consuming and intricate character of experimental work in fermentation studies justifies the complex three-ways evaluation of even small data matrices of considerable biotechnological impact.

As the aim of the work was the comparison of the efficacy of two- and three-ways data evaluation methods and not the comparison of various 3D methods, PARAFAC has not been employed for the evaluation of the original 3D data matrix.

2 Experimental

The laccase production capacity of 4 strains of the white rot fungi *Lentinus edodes* (further L.e. 1 – 3; L.e. am; L.e. 16 – 3, and L.e. 6 – 6) has been determined in 28 different culture media at 6 sampling times (10, 20, 30, 40, 50 and 60 days after inoculation). The conditions of fermentation, the method used for the measurement of laccase activity and the activity data have been previously reported [17]. PCA has been three times employed for the activity data. The original data matrix for PCA1 contained the 28 various culture media as variables and 6 sampling times × 4 strains as observations (altogether 24 observations). Matrix for PCA2 consisted of 6 sampling times as variables and 28 culture media × 4 strains as observations (altogether 112 observations). The four strains were the variables and 28 culture media × 6 sampling times were the observations (altogether 168) for PCA3. The variance explained was set to 99% in each instance. The dimensionality of the matrices of principal component loadings representing the culture media, sampling time and white rot fungi has been reduced to two by the nonlinear mapping technique. The iteration has been carried out to the point where the difference between the last two iterations taking into consideration all of scores was lower than 10^{-8}. The dimensions of the same matrices have been also reduced by the varimax rotation around two axes in order to compare the performance of the methods.

The Tucker3 model has been employed for the direct evaluation of the three-dimensional data matrix. The

Table 2. Main parameters of Tucker3 model

Variance explained by Tucker3 models using various component matrices			
Dimensions of component matrices			Variance explained %
5	3	3	96.51
4	3	2	95.93
4	2	3	85.86
4	2	2	85.34
3	2	2	85.33
Parameters of Tucker3 model using component matrix 4, 3, 2			
No. of iterations: 3			
No	Factors of component matrices		Variance explained of the total matrices variance explained %
	I	II	III	
1	1	1	1	66.34
2	3	2	1	10.63
3	2	1	2	10.23
4	2	3	1	9.64
5	2	2	1	1.40
6	3	3	1	1.28

original data matrix used for the Tucker3 procedure consisted of the laccase activities determined at 28 different culture medias (factor I), at 6 sampling times (factor II) by 4 strains of *Lentinus edodes* (factor III) giving a 3-way array with dimensions 28, 6, 4. To be sure that the maximal information will be extracted from the original data, the dimensions of component matrices have been composed from the principal components of PCAs which have an eigenvalue higher than one + the next principal component. I.e. the eigenvalue of the fourth PC of PCA1 is 1.05 (see data in Table 1) the dimension of factor 1 will be 5. Similarly, the dimensions of factors II and III have been defined as 3. In order to assess the effect of the reduction of dimensions of component matrices Tucker3 model has been performed on other matrices compiled in Table 2. The dimensionality of the resulting component matrices calculated from the three-ways array 4, 3, 2 has been reduced by the nonlinear mapping technique as described above. Varimax rotation has not been carried out on the results of Tucker3 model.

The similarities and dissimilarities among the results of PCAs and Tucker3 model using three-ways array 4, 3, 2 have been elucidated by linear regression analysis. Linear correlation were calculated between the first and second coordinates of the two-dimensional nonlinear maps computed from the loadings of PCAs, the first and second coordinates of the varimax rotation, and the first and second coordinates of the nonlinear maps calculated from Tucker3 model (4, 3, 2). Calculation have been separately carried out for factors I, II, and III.

Software for Tucker3 model was taken from N-WAY TOOLBOX, http:\\newton.mli.kvl.dk\foodtech.html prepared by Dr. C. A. Andersson and Dr. R. Bro. To the best of our knowledge this software has been withdrawn and replaced by http://www.models.kvl.dk/source. Software for PCA, varimax rotation and nonlinear mapping technique has been prepared by Dr. Barna Bordás, Plant Protection Institute, Hungarian Academy of Sciences (Budapest, Hungary).

3 Results and Discussion

The ratios of variance explained by the three PCAs are compiled in Table 1. The data clearly show that the overwhelming majority of total variance can be explained by a fairly low number (3 or 4) of theoretical (backgound) variables with a negligible amount of unexplained variance. Unfortunately, PCA does not define these variables as concrete physicochemical entities only indicates their mathematical possibility. It has to be marked that PCA considerably reduced the number of variables in the case of the 28 culture media while the reduction of variables in considerably lower in the case of sampling times and strains of white rot fungi indicating the more marked similarity of culture media.

The variances explained by Tucker3 model performed on different three-way arrays and the impacts of the individual component matrices of the three-way array 4, 3, 2 are compiled in Table 2. The data in the first part of Table 2 demonstrate that the dimensions of component matrices can be successfully approximated by using the variances explained by the PCAs. However, the data also indicate that the dimensions can be further decreased from 5, 3, 3 to 4, 3, 2 with only 0.57% loss of variance explained which cannot be concluded from the results of PCAs. The high ratio of variance explained by the reduced component matrix 4, 3, 2 indicates that this matrix is also suitable for the evaluation of the information present in the original data matrix with

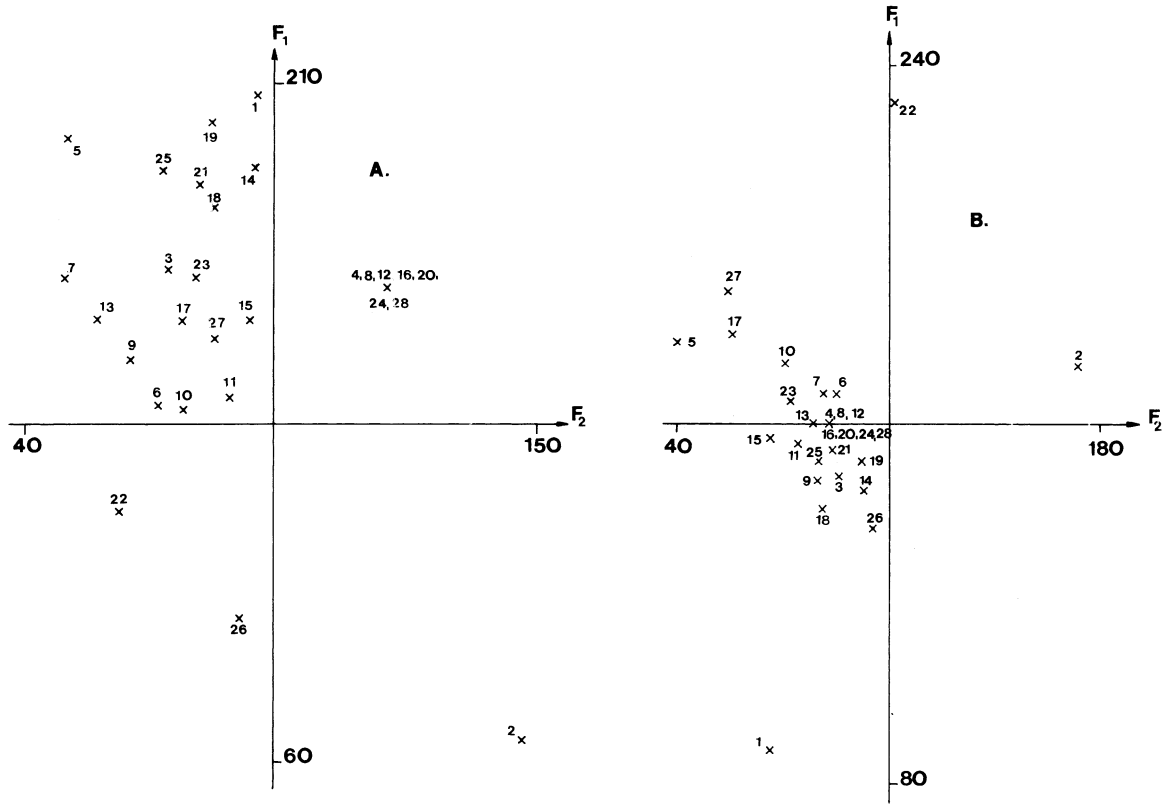

Figure 1. A: two dimensional nonlinear map of principal component loadings calculated from PCA1. No. of iterations: 660; maximal error: $2.74.10^{-2}$. B: two dimensional nonlinear map of the corresponding component matrix calculated from Tucker3 model 4, 3, 2. No. of iterations: 851; maximal error: $1.22.10^{-2}$. For symbols see Experimental.

4.07% of variance unexplained. It can be also observed from the data that further reduction of the dimensions results in considerable decrease in the amount of variance explained. The results compiled in the second part of Table 2 demonstrate that the first three factors of Tucker3 model 4, 3, 2 explain the majority of variance present in the original data matrix of 28, 6, 4 illustrating the basical similarities among the effects of culture media, sampling time and the type of strains on the production of laccase activity. However, other factors also considerably contribute to the variance explained, therefore, their inclusion in the calculation of nonilinear maps is justified.

The two dimensional nonlinear maps of 28 culture media calculated from the corresponding matrices of PCA and Tucker3 model are shown in Fig. 1A and 1B, respectively.

Only cations added at the highest concentration to the fermentation broth (points 4, 8, 12, 16, 20, 24 and 28) form a clear-cut cluster, the other media representing various cations at different concentrations are widely scattered on the maps. This finding can be tentatively explained by the supposition that both the chemical character and concentration of cations exert a similar influence on the production of laccase and distribution observed is the result of the interplay of both parameters. The data also demonstrate that the laccase production of these white rot fungi can be easily modified by the addition of various salts in different concentrations the differences between the effect of salts being higher when the points are far from each other on the maps. It can be further observed that the distances between the points on the maps calculated by PCA and Tucker3 model are not identical suggesting that the mode of calculation may exert a considerable effect on the scattering of data points on the two-dimensional maps. We assume that the results of these methods are not complementary and because of the rapidity the use of Tucker3 procedure is more advantageous than that of the three corresponding PCAs.

The two dimensional nonlinear maps of 6 sampling times calculated from the corresponding matrices of PCAs and Tucker3 model are depicted in Fig. 2A and 2B, respectively.

As it can be expected the distribution of fermentation times on the maps indicates that the laccase production of *Lentinus edodes* strains considerably depend on the fermentation time. It can be also concluded that the scattering of the sampling times show high variations according to the method of calculation. This result supports the conclusions drawn from the distribution of points on Fig. 1A and 1B that the results of these methods are different and they are not complementary. The two-dimensional non-linear maps of strains of *Lentinus edodes* also display considerable deviations as illustrated in Fig. 3A and 3B. The differences among the enzyme yield of *Lentinus edodes* strains emphasizes the importance of the selection of the adequate strain for the

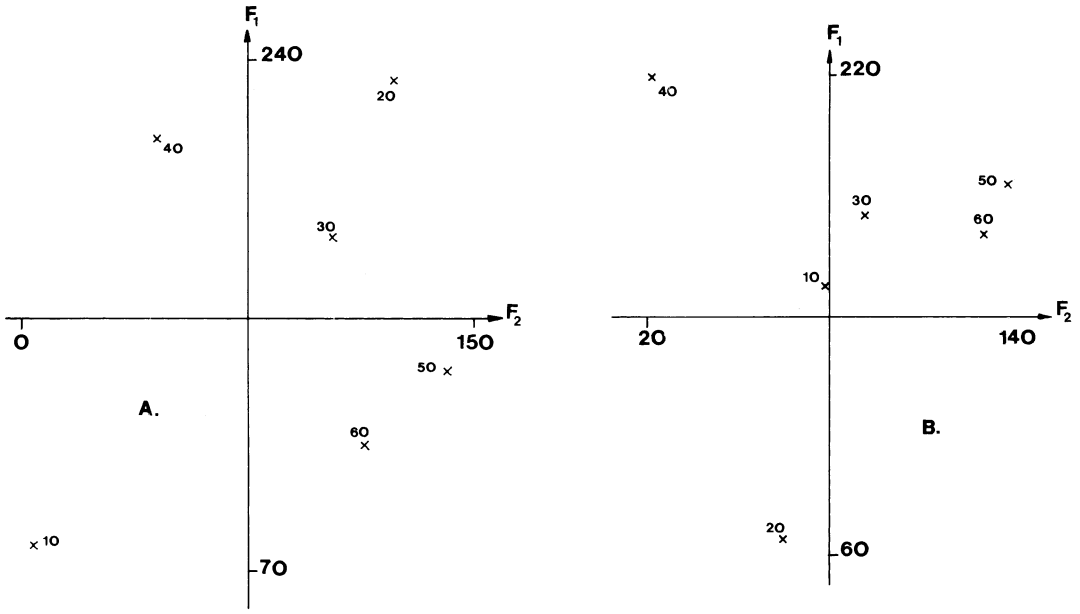

Figure 2. A: two dimensional nonlinear map of principal component loadings calculated from PCA2. No. of iterations: 92; maximal error: $2.93.10^{-4}$. B: two dimensional nonlinear map of the corresponding component matrix calculated from Tucker3 model 4, 3, 2. No. of iterations: 272; maximal error: $5.74.10^{-3}$. For symbols see Experimental.

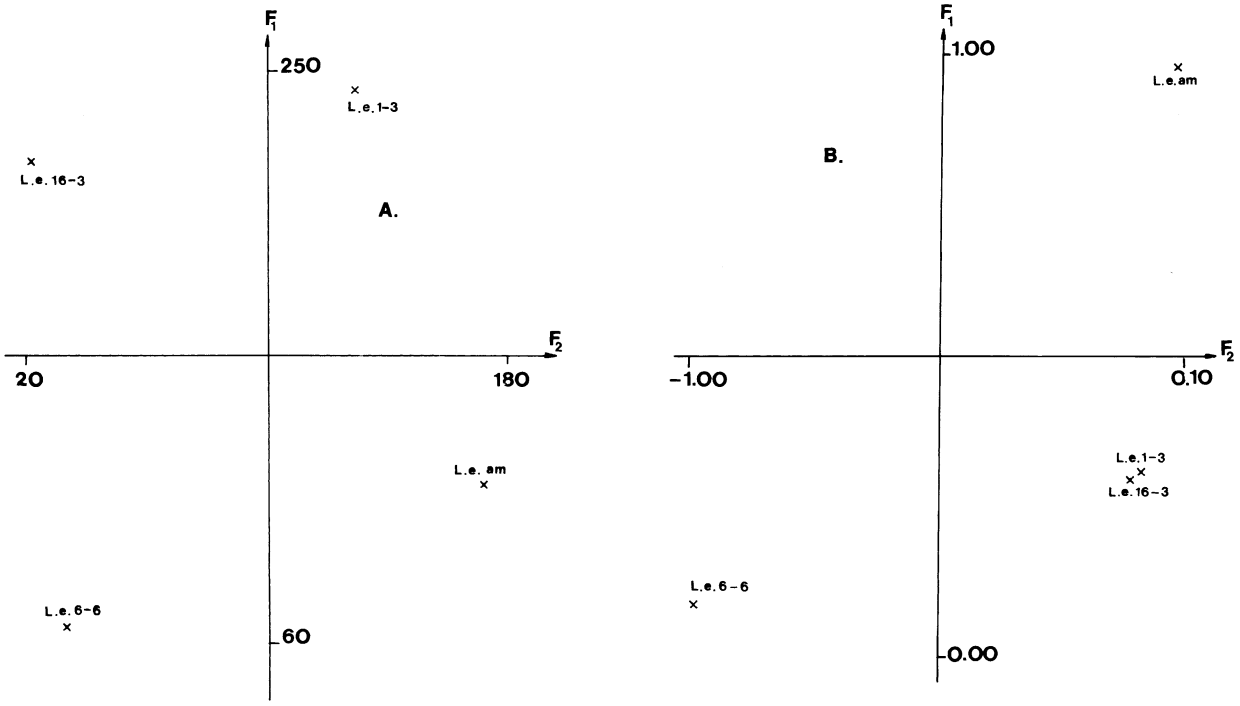

Figure 3. A: two dimensional nonlinear map of principal component loadings calculated from PCA3. No. of iterations: 20; maximal error: $1.58.10^{-2}$. B: plot of the corresponding component matrix calculated from Tucker3 model 4, 3, 2. For symbols see Experimental.

large-scale biotechnological production of laccase. Similarly to Figs. 1 and 2 the differences in the scattering of points on the two-dimensional maps indicates again the different character of the mode of calculations.

The regression coefficients of linear relationships between the coordinates of nonlinear mapping technique and varimax rotation are compiled in Table 3 (1 = first coordinate of the nonlinear map calculated from the loadings of PCA; 2 = second coordinate of the nonlinear map calcu-

Table 3. Regression coefficients of linear relationships between the coordinates of nonlinear mapping technique and varimax rotation. Significant relationship are underlined.

28 various culture media:					
	1	2	3	4	5
2	−0.2274				
3	0.2580	−0.9275			
4	−0.5415	−0.4546	0.2962		
5	−0.4164	−0.1891	0.0963	0.2281	
6	−0.5630	0.4678	−0.5049	0.3891	0.0418
6 sampling times:					
	1	2	3	4	5
2	0.3137				
3	0.5763	0.8096			
4	−0.7891	−0.5057 −0.8772			
5	−0.2375	−0.2612	0.1333	−0.1981	
6	−0.5871	0.5800	0.1953	0.2622	−0.0186
4 strains of *Lentinus edodes*:					
	1	2	3	4	5
2	0.0400				
3	0.9913	−0.0699			
4	−0.3377	−0.8915	−0.2631		
5	−0.4292	0.7297	−0.4703	−0.6829	
6	0.6738	0.5787	0.6433	−0.8785	0.3725

lated from the loadings of PCA; 3 = first coordinate of the varimax rotation calculated from the loadings of PCA; 4 = first coordinate of the varimax rotation calculated from the loadings of PCA; 5 = first coordinate of the nonlinear map calculated from the loadings of Tucker 3 model; 6 = second coordinate of the nonlinear map calculated from the loadings of Tucker3 model). Regression coefficients indicating significant relationship at a significance level of 95% are underlined. As the degree of freedom was different for each data set the numerical value of the regression coefficients indicating significant relationships are also different. The results in Table 3 entirely support the previous qualitative conclusions, the linear correlations between the corresponding coordinates of the various data reducing methods are not always significant. This finding indicates that the selection of method used for the reduction of the dimensionality of matrices calculated by both PCA and Tucker3 model modifies the similarities and dissimilarities among the data points leading to misinterpretation of the results. This discrepancy can be tentatively explained by the structure and solution method of Tucker3 model. This discrepancy can be explained by the fact that the uniqueness of the solution of two-dimensional PCA comes from the underlying structure of the least square optimization problem which belongs to the class of convex optimization problems. Oppositely, the optimization version of Tucker3 model is not a convex but a global optimization problem. Global optimization problems are hardly solvable because not all local minima are global. Therefore, there might be several local solutions. The local minima found by an algorithm may depend on the starting data points used to run the method, consequently, different local minima can be calculated from different starting points. When the nonlinear mapping technique is applied to these results, the obtained maps may show differences. At such cases the evaluation of the results has to be carried out very cautiously. We have to emphasize that the conclusions discussed above are not based on exact mathematical treatments, they are empirical ones and it is possible that they are valid only for this set of data. Therefore, any extrapolation of the conclusions to other data matrices has to be performed very carefully.

It can be concluded from the results that the dimensionality of the component matrices of Tucker3 model can be predicted by using PCA. The method used for the reduction of the dimensionality of matrices calculated by both PCA and Tucker3 model modifies the similarities and dissimilarities among the elements of the original matrix and may considerably influence the reliability of visual evaluation.

Acknowledgement

This work was supported by the Portugues-Hungarian cooperation project 'Development of new methods and their application for the assessment of the effect of environmental conditions on the stability of color pigments in foods'. The research of T. Illés has been partially supported by the grant OTKA T 029775. T. Illés thanks the Bolyai Farkas Research Fellowship of the foundation

"Arany János Közalapitvány a Tudományért". The authors express their gratitude to Dr. R. Bro and Dr. C. A. Andersson to make available the N-WAY TOOLBOX.

References

[1] Mardia, K. V., Kent, J. T., and Bibby, J. M., *Multivariate Analysis,* Academic Press, London 1979, pp. 213 – 254.
[2] Harshman, R. A., Foundations of the PARAFAC procedure: model and conditions for an 'explanatory' multi-mode factor analysis, *UCLA Working Papers in Phonetics, 16*, 1 – 84 (1970).
[3] Bro, R., PARAFAC: Tutorial and applications, *Chemom. Intell. Lab. Syst. 38*, 149 – 171 (1997).
[4] Bro, R., Andersson, C. A., and Kiers, H. A. L., PARAFAC2: Part II. Modeling chromatographic data with retention time shifts, *J. Chemom. 13*, 295 – 309 (1999).
[5] Kiers, H. A. L., Ten Berge, J. M. F., and Bro, R., PARAFAC2: Part I. A direct fitting algorithm for the PARAFAC model, *J. Chemom. 13*, 275 – 294 (1999).
[6] Carrol, J. D., and Chang, J., Analysis of individual differences in multidimensional scaling via an N-way generalization of and Eckhardt-Young decomposition, *Psychometrika 35*, 283 – 319 (1970).
[7] Tucker, R. L., Some mathematical notes on the three mode factor analysis, *Psychometrika 31*, 279 – 311 (1966).
[8] Tucker, R. L., Relations between multidimensional scaling and three-mode factor analysis, *Psychometrika 37*, 3 – 27 (1972).
[9] Andersson, C. A., and Bro, R., Improving the speed of multiway algorithms: Part I. Tucker3, *Chemom. Intell. Lab. Syst. 42*, 93 – 103 (1998).
[10] Ligny, C. de, Spajner, M., Houwelingen, J. C. van, and Weesie, H. M., Three-mode factor analysis of data on retention in normal phase high-performance liquid chromatography, *J. Chromatogr. 301*, 311 – 324 (1984).
[11] Gemperline, P. J., Miller, K. H., West, T., Weinstein, E., Hamilton, J. C., and Bray, J. T., Principal component analysis, trace elements and blue crab shell disease, *Anal. Chem. 64*, 523 – 532 (1992).
[12] Kloot, W. A. van der, and Kronenberg, P. M., Extremal analysis with three-mode principal component analysis, *Psychometrika 50*, 479 – 494 (1985).
[13] Smilde, A. K., Wang, Y., and Kowalski, B. R., Theory of medium-rank second order calibration with restricted Tucker models, *J. Chemom. 8*, 21 – 36 (1994).
[14] Morais, H. Ramos, C., Forgács, E., Cserháti, T., Oliviera, T., and Illés, T., Three-dimensional principal component analysis employed for the study of the β-glucosidase production of *Lentinus edodes* strains, *Chemom. Intell. Lab. Syst. 57*, 57 – 64 (2001).
[15] Morais, H., Ramos, C., Forgács, E., Jakab, A., Cserháti, T., Oliviera, J., Illés, T., and Illés, Z., Threedimensional principal component analysis used for the study of enzyme kinetics. An empirical approximation for the determination of the dimensions of component matrices, *Quant. Struct.-Act. Relat. 20*, 241 – 247 (2001).
[16] Sammon, J. W. Jr., A nonlinear mapping for data structure analysis, *IEEE Trans. Comput. C18*, 401 – 407 (1969).
[17] Morais, H., Ramos, A. C., Cserháti, T., Forgács, E., Darwish, Y., and Illés, Z., Effect of the composition of culture media on the laccase production of *Lentinus edodes* strains, *Acta Biotechnol. 21*, 307 – 320 (2001).

Received on June 27, 2002; Accepted on January 15, 2003