ELSEVIER

Analytica Chimica Acta 462 (2002) 133–148

ANALYTICA CHIMICA ACTA

www.elsevier.com/locate/aca

Three-way principal component analysis applied to food analysis: an example

V. Pravdova [a], C. Boucon [b], S. de Jong [b], B. Walczak [a,1], D.L. Massart [a,*]

[a] *ChemoAC, FABI-VUB, Laarbeeklaan 103, 1090 Brussels, Belgium*
[b] *Unilever Research Vlaardingen, P.O. Box 114, 3130 AC Vlaardingen, The Netherlands*

Received 3 October 2001; received in revised form 16 April 2002; accepted 16 April 2002

Abstract

The purpose of the study is to show how the interpretation of a complex multivariate data array can significantly be improved by the application of *N*-way principal component analysis (PCA). Two food related three-way data sets were studied; a sensory and a chromatographic data array. The Parafac and the Tucker3 models were applied and results were compared. Both *N*-way models presented here allow visualization of the data structure and give detailed information about the data set, notably allowing to understand relationships between objects and variables.

Keywords: *N*-way PCA; Tucker3 model; PARAFAC model; Maillard reaction

1. Introduction

The main purpose of exploratory data analysis is to learn about interrelationships between objects and variables. As data sets are becoming larger and more complex, visualization is needed to achieve this aim. A very well known method for visualisation of the data patterns is principal component analysis (PCA). The aim of this study is to show that applying the *N*-way approach (an extension of PCA to higher orders) can significantly improve visualisation and interpretation of complex multivariate data compared to other data analysis methods previously applied to the same datasets [1,2]. Tucker3 and PARAFAC models are two of the most popular *N*-way methods. There are many applications of them in the literature, but mainly for spectroscopic data, e.g. [3–5]. Some non-spectroscopic applications can be found [6–9], which show that *N*-way tools can be very useful and often can give more information about the data than standard multivariate methods such as two-way PCA. Here, we describe an application of *N*-way analysis and visualization to data from food chemistry.

The Maillard reaction is one of the most studied reactions in food chemistry. It is a reaction between a reducing sugar and an amino compound, usually an amino acid, resulting in a wide variety of reaction products. These contribute to the smell and color of the food when it is prepared. A full factorial experimental design with 11 amino acids (AA) and 6 sugars (S) was carried out. For each combination of an amino acid and a sugar, gas chromatographic (GC) and sensory (smell) data were obtained. This way, the data sets are organized in natural three-way order, namely AA × S × GC or AA × S × smells.

* Corresponding author. Tel.: +32-24774737; fax: +32-24774735.
E-mail address: fabi@vub.vub.ac.be (D.L. Massart).
[1] On leave from Silesian University, Katowice, Poland.

0003-2670/02/$ – see front matter
PII: S0003-2670(02)00318-5

2. Theory

2.1. Notation

Two-way data arrays (i.e. matrices) are denoted with bold capital letters, **X**. To distinguish notation of matrices from three or more way data arrays, the latter are indicated by an underlined bold $\underline{\mathbf{X}}$. The elements of a three-way data array $\underline{\mathbf{X}}$ are denoted by x_{ijk}, where indices can change in the following ranges: $i = 1, \ldots, I$, $j = 1, \ldots, J$, $k = 1, \ldots, K$. The number of factors in each mode is denoted as M, N and L, respectively.

2.2. N-way models

2.2.1. Tucker3 model

The Tucker3 method decomposes the 3-way data arrays $\underline{\boldsymbol{X}}$ (see Fig. 1) into three orthonormal loading matrices, denoted as $\boldsymbol{A}$ $(I \times L)$, $\boldsymbol{B}$ $(J \times M)$, $\boldsymbol{C}$ $(K \times N)$ and the core matrix $\underline{\boldsymbol{Z}}$ $(L \times M \times N)$, which describes the interactions among $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$.

The largest squared elements of the core matrix $\underline{\boldsymbol{Z}}$ indicate the most important factors that describe $\underline{\boldsymbol{X}}$. The number of factors in each mode is not necessarily the same, i.e. $L \neq M \neq N$. Because of rotational freedom of the model, the core matrix is needed for their interpretation [10].

The determination complexity of N-way models, i.e. the number of factors in each mode, is a very important part of the data analysis. Several tools can be used for this purpose, for instance, core consistency analysis [11] (only for PARAFAC model), half split analysis [12], cross validation [13] or percentage of explained data variance. In our study, core consistency analysis and percentage of explained data variance have been used to identify the right number of factors in each mode. The results are usually presented in the form of loading plots.

2.2.2. PARAFAC model

PARAFAC is a decomposition method, which can be considered as a constrained version of Tucker3 model, i.e. the number of factors in all modes is equal

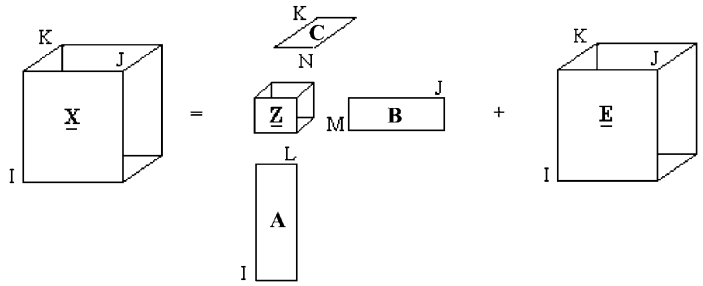

Fig. 1. Tucker3 model—graphical explanation.

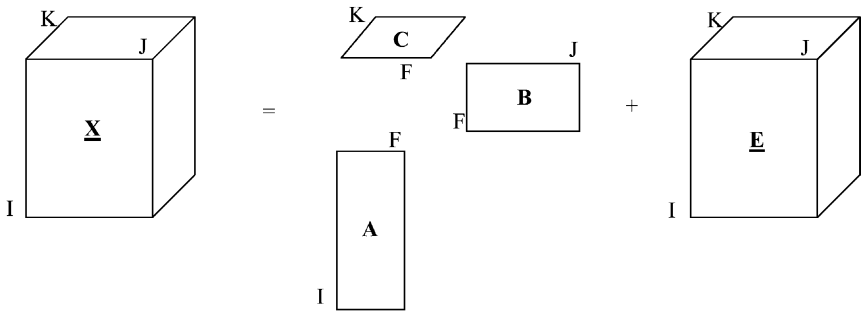

Fig. 2. PARAFAC model—graphical explanation.

and the diagonal elements of the core matrix are equal to 1. The decomposition of e.g. three-way data is made into triads or tri-linear components (multi-linear if $N > 3$), but instead of one score vector and one loading vector as in bilinear PCA, each component consists of one score vector and two loading vectors. In multi-way analysis the distinction between score and loading vectors is often not made and one also uses the term loadings for all modes. A PARAFAC model of a three-way array is then given by three loading matrices (A, B and C) with elements a_{if}, b_{jf} and c_{kf} (Fig. 2).

Generally, the PARAFAC model is easier to interpret and many authors prefer to apply PARAFAC when it is possible (i.e. the complexity in all dimensions is equal). In this work the core consistency tool was used to obtain the optimal complexity of the PARAFAC model.

3. Data

3.1. Data description

The Maillard reaction was carried out at fixed pH 3, with all possible combinations of one amino acid (AA: alanine, asparagine, arginine, cysteine, glutamine, glutamate, glycine, lysine, methionine, proline and threonine) and one sugar (S: fructose, glucose, lactose maltose, rhamnose and xylose).

The samples were then analyzed with high speed gas chromatography using a narrow bore column and FID detector. Besides, samples were evaluated sensorically by a small panel of 3–5 trained panellists. Samples were scored individually for nine sensory descriptors (SM: sulfur, meaty, caramel, burnt, nutty, popcorn, jammy, potato, aldehyde) on a scale from 0 to 4. The average was taken into consideration for further analysis. This yields two three-way data matrices with dimensionality 11AA × 6S × 159GC (chromatographic data matrix) and 11AA × 6S × 9SM (sensory data matrix). No pre-processing method was used before the data analysis.

3.2. Software

All programs used were written in Matlab 5.0 (the MathWorks) computing environment. The N-way toolboxes used are freely accessible on the Internet [14].

4. Results and discussion

In the subsequent discussion, the factors of the different modes will be noted as follows. Amino acid mode (A): $A1, A2, \ldots,$ AL, sugar mode (B): $B1, B2, \ldots,$ BM, peak or sensory mode and (C): $C1, C2, \ldots,$ CN.

4.1. Chromatographic data set

4.1.1. Tucker3 model

Generally, the optimal complexity of the Tucker3 model is the one that requires the smallest number of factors, but still describes relatively high fraction of data variance. All possible models, with different number of factors in each mode ($L, M, N = 1, \ldots, 5$), have been evaluated. In order to visualize how many percent of explained variance is gained by adding factors, the explained variance was plotted versus increasing value of the product ($L \times M \times N$) (Fig. 3). To distinguish different models with the same product $L \times M \times N$ (e.g. for models with complexities (1,2,2), (2,1,2), (2,2,1)), the models are ordered so that the model with the smallest L is shown first, followed by the model with the smallest M and finally that with the smallest N.

The optimal complexity of the Tucker3 model was considered to be (1,2,2), i.e. 1 factor in the first mode (AA), 2 factors in the second mode (sugars) and 2 in the third mode (peaks). This model explains 78.3% of data variance. Two groups of AA can be observed on the plot of $A1$ loadings. Alanine, methionine and threonine have high $A1$-values while the remaining AA have low $A1$-value (Fig. 4).

In the loading plot $B1$ versus $B2$ the distribution of sugars is presented. Almost no variability can be observed along $B1$. Along the $B2$ axis, fructose, glucose, lactose and maltose form one cluster, whereas, rhamnose and xylose are separated from them, rhamnose to a higher degree than xylose (Fig. 5).

Fig. 6 shows the loadings plot of $C1$ versus $C2$. Like for sugars, most variability is observed along the second axis, $C2$. Most peaks are grouped in a large, elongated cluster. However, a few peaks (e.g. numbers 55, 142, 143, 56) are revealed and can be interpreted by looking at the interactions with the other modes, as will be explained below.

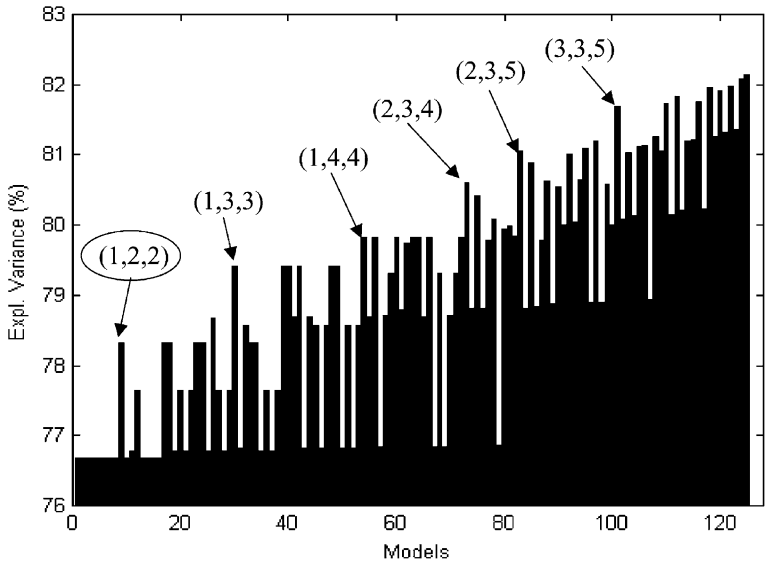

Fig. 3. Chromatographic data set: variance explained (%) for different Tucker3 models. Several complexities are identified and indicated by arrows.

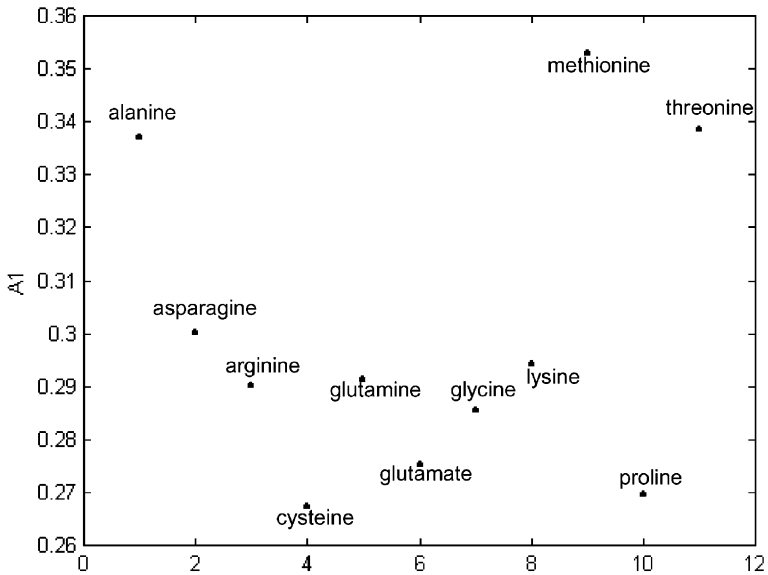

Fig. 4. Tucker3 model for chromatographic data matrix: plot of *A*1 loadings for 'amino acids' mode.

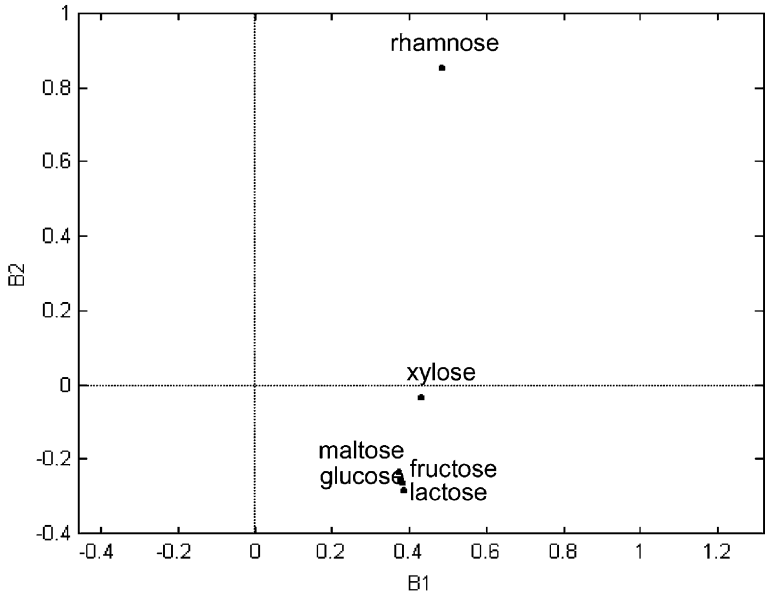

Fig. 5. Tucker3 model for chromatographic data matrix: loading plot (*B*1 vs. *B*2) for 'sugars' mode.

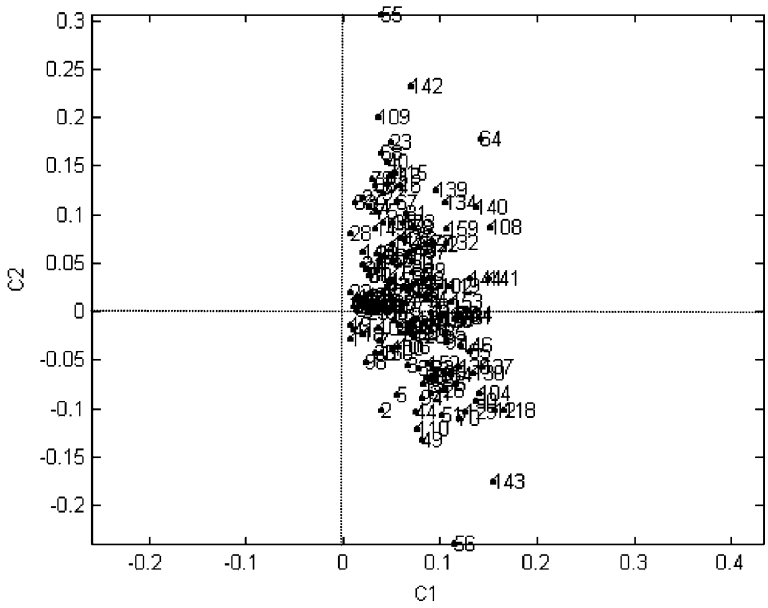

Fig. 6. Tucker3 model for chromatographic data matrix, loading plots for 'peak intensities' mode. *C*1 vs. *C*2.

Table 1
Core matrix with dimensionality (1,2,2)

	*C*1		*C*2	
	*B*1	*B*2	*B*1	*B*2
*A*1	**151**	0	0	**22**

Bold letters indicate important core elements.

The core matrix gives information about interactions among the three modes. The core matrix (Table 1) has the same dimensionality as the complexity of the model, i.e. (1,2,2).

The dominant core elements (bold) are: (1,1,1) and (1,2,2), where the numbers indicate factors participating in the interactions.

The first important core element (1,1,1) indicates interaction among factors *A*1, *B*1 and *C*1. The axis *C*1 describes peak intensities of reaction products. Thus, peaks with highest loadings along *C*1, i.e. 140, 64, 108, 141, 118, 143, 12 are usually the most intense ones in all samples. This fact is most evident when methionine (high *A*1 value) is present in the Maillard reaction.

The core element (1,2,2) expresses the interactions among factors *A*1, *B*2 and *C*2. Axis *A*1 differentiates to the highest degree methionine, threonine and alanine from the rest of the amino acids. Fructose, glucose, lactose and maltose have low negative *B*2 values, while rhamnose has very high positive *B*2 value. Peaks 142 and 55 are the most extreme peaks with very high positive *C*2 loadings, whereas peaks 56 and 143 have the lowest *C*2 loadings. Therefore, it can be concluded that a high intensity of peaks 142 and 55 is observed for the reaction products of methionine, threonine or alanine with rhamnose. Two alternative conclusions could be drawn for peaks 56 and 143. Peaks 56 and 143 have high intensities for the reaction of fructose, glucose, maltose and lactose with all AA or alternatively that peak 56 has a low intensity when rhamnose is present in reaction. Those facts are the most evident for methionine, which posses the highest *A*2 loading.

4.1.2. PARAFAC model

Fig. 7a,b shows the core consistency plots for two and three components extracted. It can clearly be seen that the optimal complexity is obtained for a two components model explaining 78.6% of data variance.

The loadings plot *A*1 versus *A*2 (Fig. 8) reveal four groups of amino acids. The first one (group I) contains cysteine, proline and glutamate with low *A*1 loadings. Asparagine, arginine, glutamine, glycine and lysine form the second group (group II). Group III contains alanine and threonine with high *A*1 loadings and low *A*2 loadings. Along *A*2, methionine (group IV) is the amino acid with the highest loading, while arginine, glycine, glutamate, lysine and asparagine have the lowest *A*2 loadings. The loading plot *B*1 versus *B*2 (Fig. 8b) shows separation of xylose and rhamnose from the rest of sugars mainly along *B*2 axis in similar way as in Fig. 5. The peaks are presented on the loading plot *C*1 versus *C*2 (Fig. 8c). The extreme points are distributed in similar way as for the Tucker3 model and similar conclusions can be made. Axis *C*1 reflects peak intensity as in the Tucker3 model. Peaks 140, 64 and 108 have generally high intensity in all samples, the highest intensities are observed when methionine and rhamnose participated in the Maillard reaction. This corresponds to the position of methionine (high *A*1 loadings) and rhamnose (high *B*1 loadings) in loading plots *A*1 versus *A*2 (Fig. 8a) and *B*1 versus *B*2 (Fig. 8b). Along *C*2 axis, peaks 142 and 55 have the highest positive values, while peaks 56, 51 and 143 posses the lowest loadings.

Both models (i.e. the Tucker3 and Parafac models) reveal similar pattern in all modes. For AA mode, the methionine, alanine and threonine are separated from cysteine, proline and glutamate. For the sugar mode, rhamnose and xylose are differentiated from the rest of sugars, rhamnose to higher degree than xylose. Peaks are situated in one cluster in both models, some peaks position are interchanged, e.g. 142 and 55, 109 and 23 or 64 and 140, when comparing loading plot *C*1 versus *C*2 from Tucker3 and Parafac models. However, the conclusions obtained from interpretation of these models are not in conflict.

4.2. Sensory data matrix

4.2.1. Tucker3 model

The number of factors in each mode was chosen in the same way as for the chromatographic data set, the plot of explained variance versus product $L \times M \times N$ is presented in Fig. 9. The 4 factors in the first mode (amino acids), 1 factor in the second mode (sugars) and 4 factors in the third mode (smells) were chosen as the best complexity for the Tucker3 model. It explains

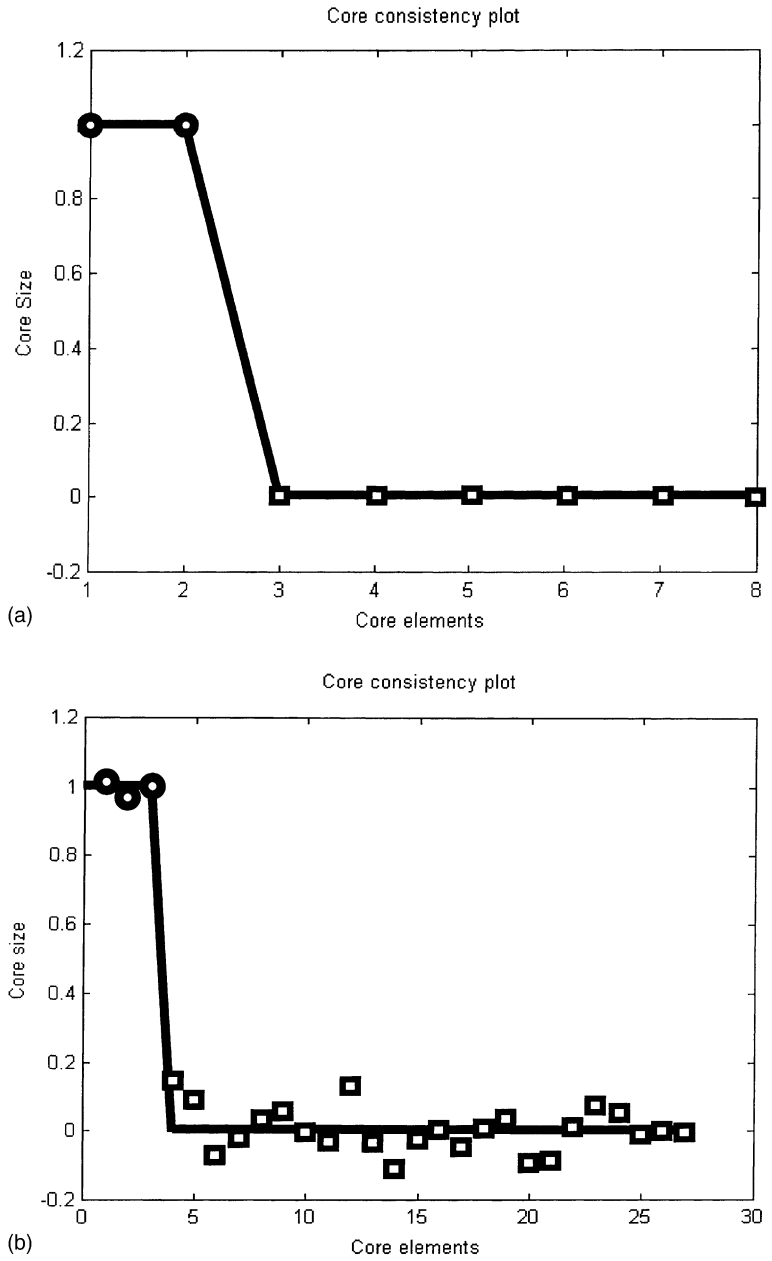

Fig. 7. Core consistency plot (□/○: zero/non-zero core elements) for the chromatographic data matrix with (a) 2 factors extracted; (b) 3 factors extracted. If the right number of factors is extracted, the superdiagonal core elements (○) should be close to 1 and off-superdiagonal elements (□) should be close to 0.

77.7% of data variance. The rather high number of factors in the third mode containing sensory data (smells) is caused by relatively low correlation between the studied variables, which were indeed selected to describe different directions of the sensory space.

The interpretation is based on loading plots of all three modes combined with the information in core matrix. The loading plots for the amino acid mode (*A*1 versus *A*2, and *A*3 versus *A*4) are presented in the Fig. 10a,b.

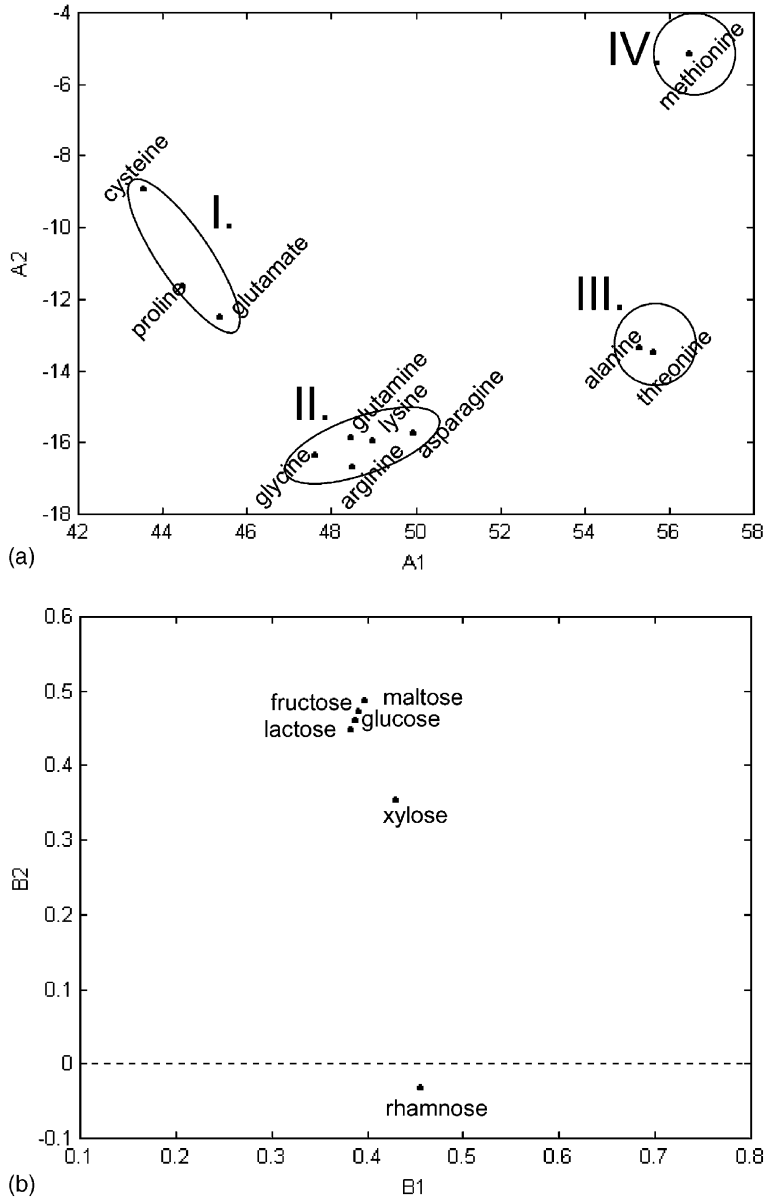

Fig. 8. PARAFAC model for chromatographic data matrix. Loading plots for (a) 'amino acids' mode; (b) 'sugars' mode and (c) 'GC peaks' mode.

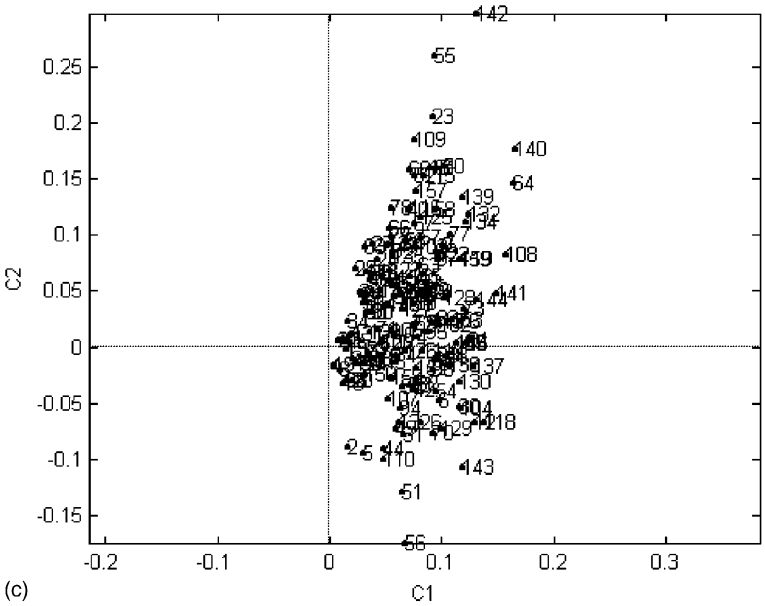

Fig. 8. (*Continued*).

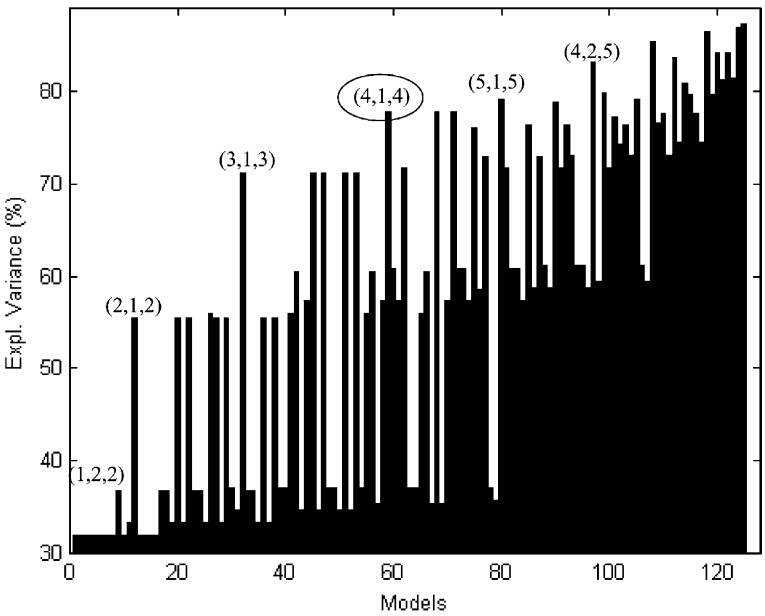

Fig. 9. Sensory data set: variance explained (%) for different Tucker3 models. Several complexities are identified and indicated by arrows.

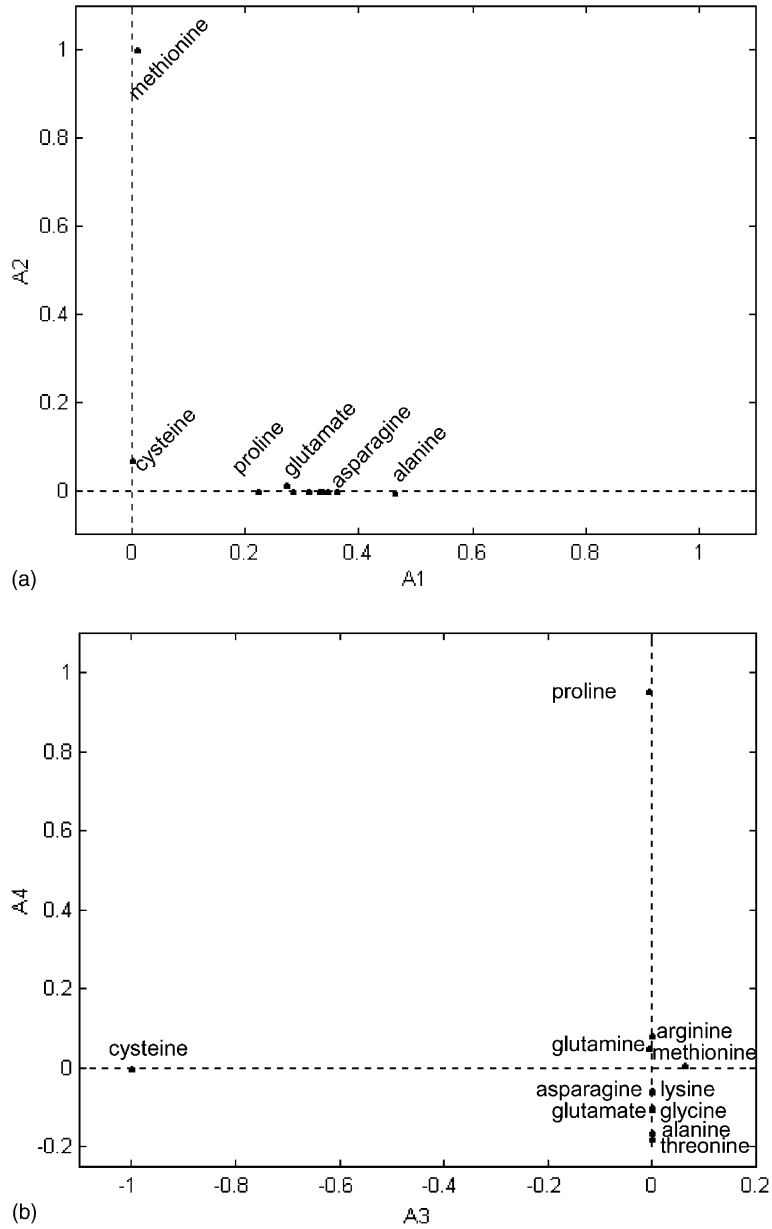

Fig. 10. Tucker3 model for sensory data matrix: (a) loading plot *A*1–*A*2; (b) loading plot *A*3–*A*4.

Sugars are depicted in plot of *B*1 loadings in Fig. 11 and smells can be seen in loading plots *C*1 versus *C*2 and *C*3 versus *C*4 (Fig. 12a,b, respectively).

The core matrix is presented in Table 2. Bold numbers distinguish the important core elements that express interaction among the factors. The marked core elements can be denoted as follows: (1,1,1), (2,1,2), (3,1,3) and (4,1,4) where the numbers indicate factors. The values of those core elements are approximately the same so that the interactions between factors have also the same importance.

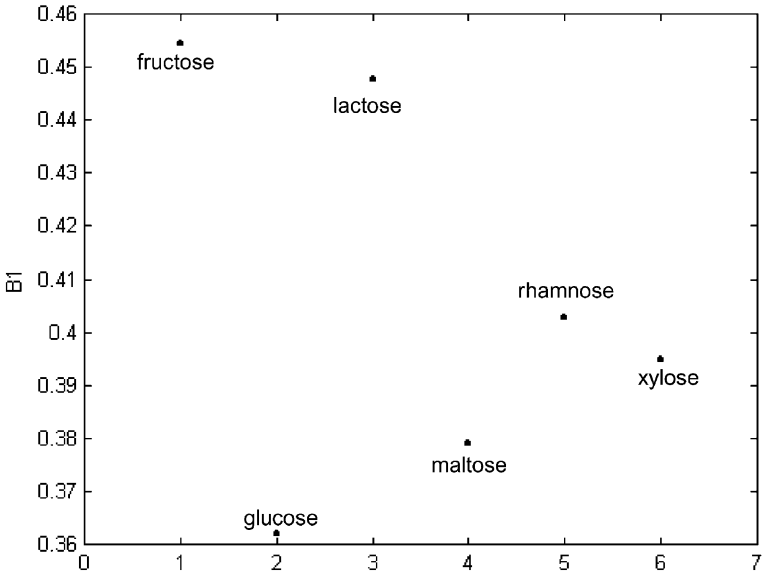

Fig. 11. Tucker3 model for sensory data matrix. Plot of *B*1 loadings.

Table 2
Core matrix with dimensionality (4,1,4)

	*B*1			
	*C*1	*C*2	*C*3	C4
*A*1	**10**	0	0	0
*A*2	0	**8.9**	0	0
*A*3	0	0	**−7.2**	0
*A*4	0	0	0	**4.8**

Bold letters indicate important core elements.

Core element (1,1,1) describes the interaction among factors *A*1, *B*1 and *C*1. Axis *A*1 differentiates to the highest degree cysteine and methionine from alanine. All sugars have more or less similar value along *B*1, only small differentiation along *B*1 axis is observed. Lactose and fructose can be distinguished from glucose and maltose, rhamnose and xylose are situated between them according *B*1 value. Along the *C*1 axis sweet smells (jammy, caramel) are discriminated from savoury (popcorn, potato, sulfur) and nutty smells. The value of the core element is positive; it leads to the conclusion that alanine in reaction with lactose or fructose give jammy and caramel smells.

Interaction among *A*2, *B*1 and *C*2 is expressed by core element (2,1,2). Methionine is very different on the *A*2 axis compared to the rest of the amino acids. Axis *B*1 is already described. The *C*2 axis differentiates to a high degree the potato smell from the remaining ones. Taking into account the positive value of the core element for this interaction, it can be concluded that all sugars (positive loadings) contribute to the potato smell in combination with methionine, and this fact is most marked for fructose and lactose that posses the highest loadings on *B*1.

Core element (3,1,3) indicates *A*3, *B*1 and *C*3 interaction. Cysteine (with negative *A*3 value) is different from the rest of the AA along axis *A*3. Sulfur and meaty smells have a higher positive value than the rest of the smells along the *C*3 axis. The core element has a negative value, so that cysteine in combination with fructose and lactose cause the sulfur and meaty smells.

The interaction among *A*4, *B*1 and *C*4 is described by core element (4,1,4). The value of this core element is positive. Proline has a unique value on the *A*4 axis. The *C*4 axis reflects the differences between nutty and jammy smells. Reaction of proline with fructose or/and lactose leads to products with a nutty smell.

4.2.2. PARAFAC model

To determine the optimal complexity of the PARAFAC model for the sensory data matrix the core consistency plots were drawn for models when 2 (Fig. 13a) and 3 factors (Fig. 13b) were extracted.

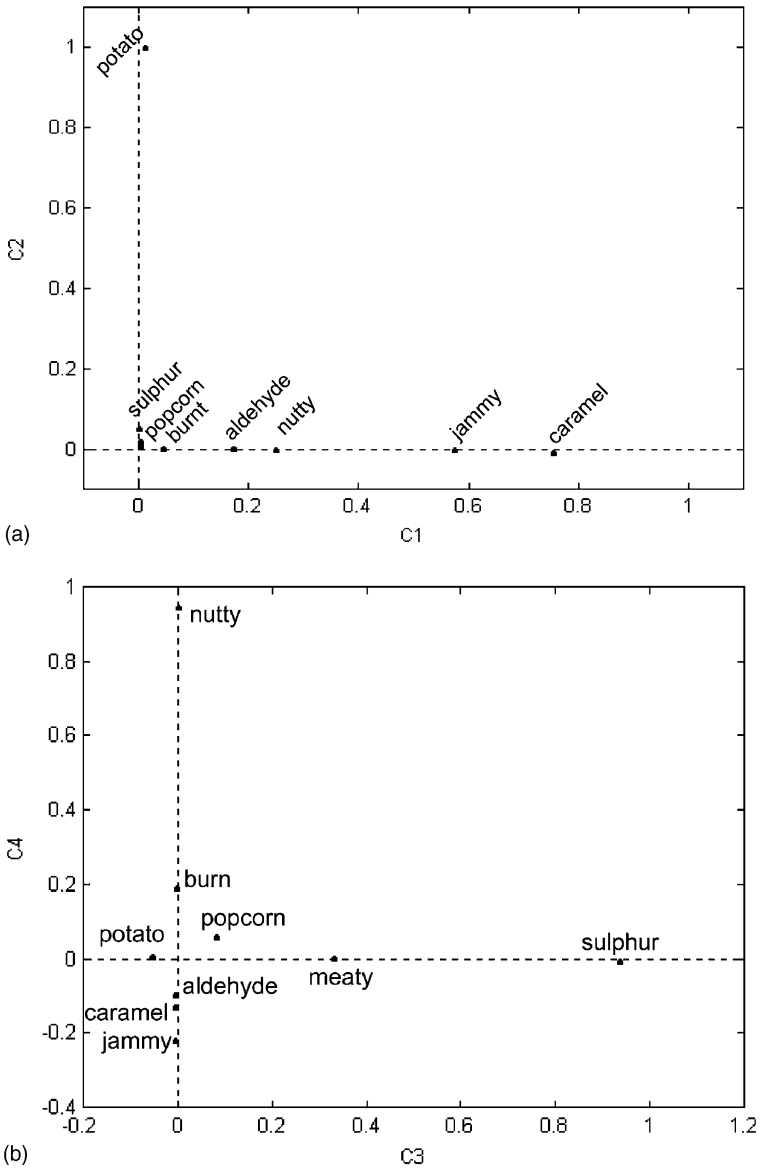

Fig. 12. Tucker3 model for sensory data matrix: (a) Loading plot *C*1–*C*2; (b) loading plot *C*3–*C*4.

The plots clearly show that the best PARAFAC model can be constructed by extracting 2 factors. This model explains 60% of data variance. The loadings plots (*A*1 versus *A*2, *B*1 versus *B*2 and *C*1 versus *C*2) are presented in Fig. 14a,b and c, respectively.

In the loading plot *A*1 versus *A*2 (see Fig. 14a) a similar pattern is observed compared to the Tucker3 model. Alanine has the highest value along *A*1 axis and methionine is highly differentiated from the rest of the AA along the *A*2 axis. The sugars can be viewed in the loading plot *B*1 versus *B*2. The values of *B*1 and *B*2 loadings are relatively small compared to the loadings of *A*1, *A*2. This observation suggests the conclusion that AA have a dominant effect on the smell of the reaction products. Along *B*1 axis fructose and lactose have the highest value. Xylose, rhamnose and lactose

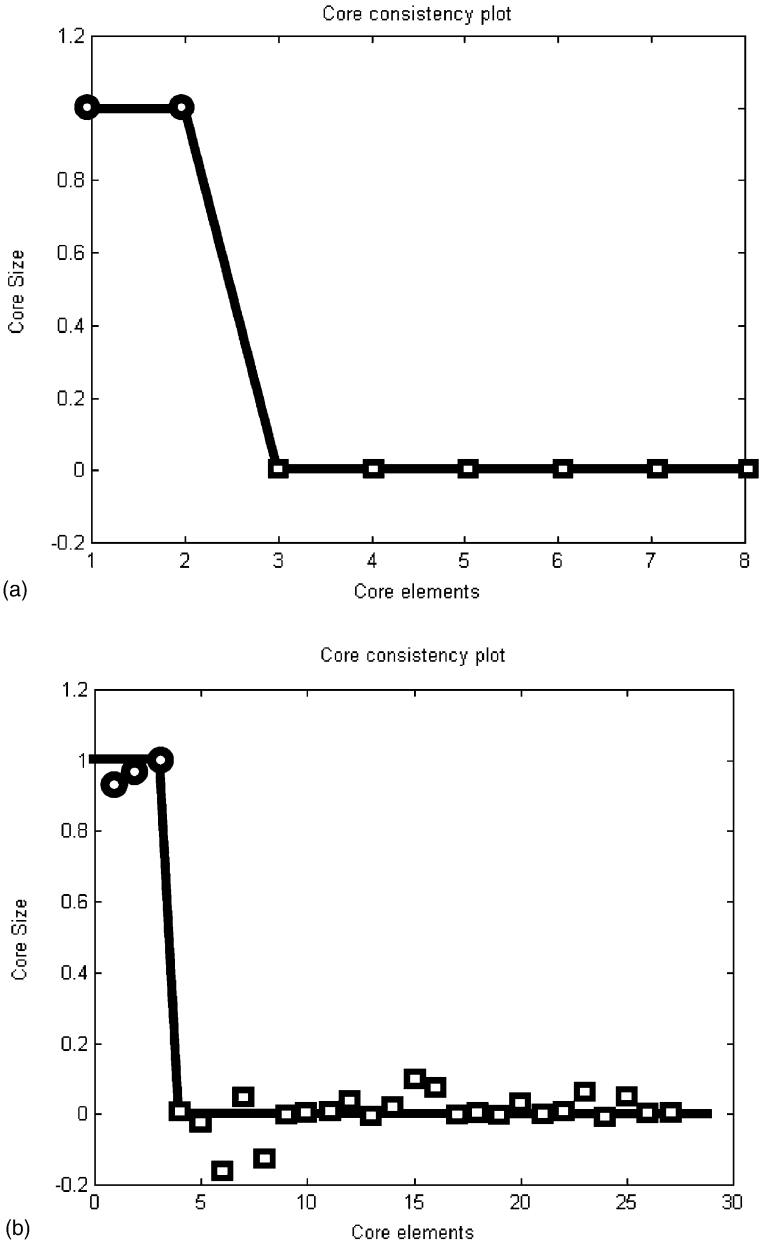

Fig. 13. Core consistency plot (■/●: zero/non-zero core elements) for the sensory data matrix with (a) 2 factors extracted; (b) 3 factors extracted. If the right number of factors is extracted, the superdiagonal core elements (●) should be close to 1 and off-superdiagonal elements (■) should be close to 0.

have the highest values along *B*2 axis. The loading plot *C*1 versus *C*2 shows that the 'sweet' smells have a high *C*1 value and that they are distributed along this axis in the same way as in the Tucker3 model. The *C*2 axis differentiates the potato smell from the rest of the smells to a high degree. Similar conclusions as in the case of the Tucker3 model can be extracted from the loading plots. The reaction products of all sugars and

alanine have a caramel or jammy smell, which it is not so for cysteine and methionine. This is most evident for fructose and lactose. The reaction of methionine with all sugars, but mostly with xylose and rhamnose, leads to a potato smell.

Both data matrices should contain similar information but gas chromatography is expected to give a more detailed description of reaction products than sensory analysis (as many of the detected substances do not contribute to the odor of the samples). There-

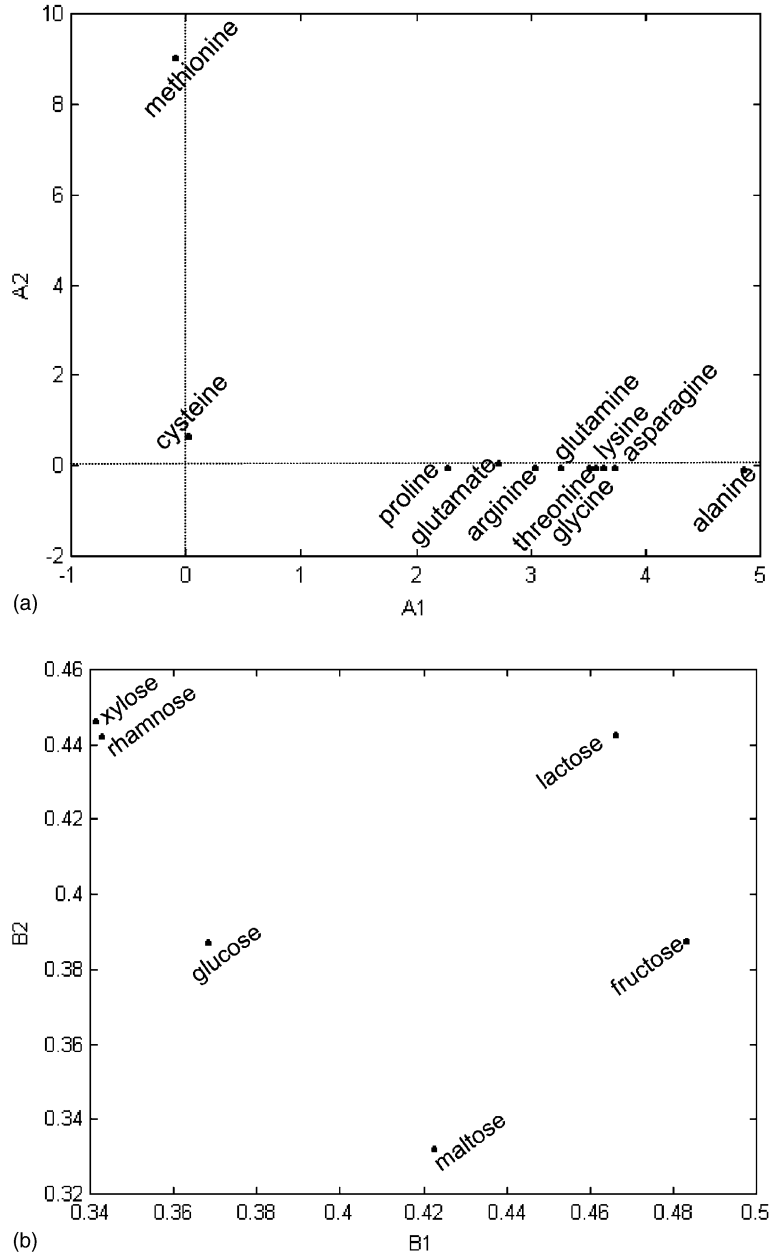

Fig. 14. PARAFAC model for the sensory data matrix. Loading plots for (a) ‘amino acids’ mode; (b) ‘sugars’ mode and (c) ‘smell’ mode.

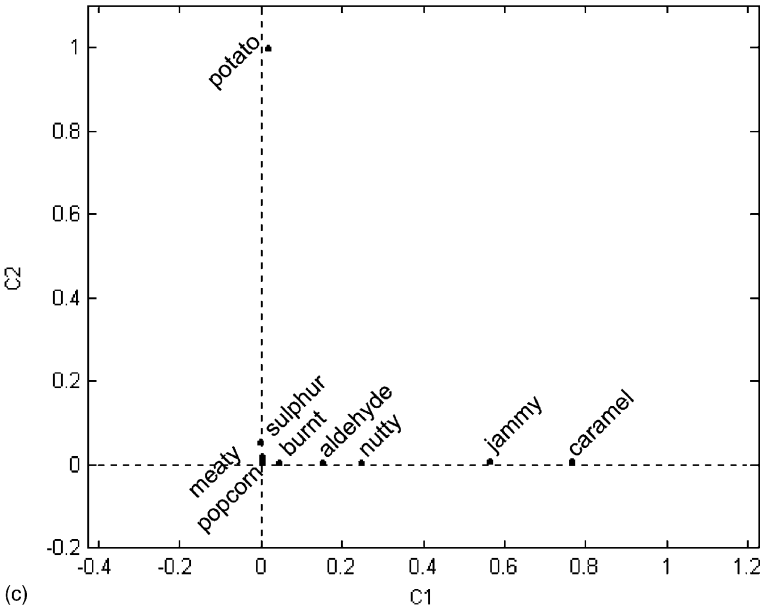

Fig. 14. (*Continued*).

fore, the model of the sensory data matrix should be less complex than the model for the chromatographic data matrix. The obtained results however give a different picture. The complexity of the Tucker3 model of the sensory data is higher than that for chromatographic data and the loading plots contain different patterns. For the sensory data set, for the mode associated with nine smells, 4 factors are interpretable. Therefore, it is not straightforward to relate peaks with smells based on the presented models. Moreover, no pre-processing has been applied to the data set so that a few large peaks, which are not necessarily related to odorous compounds, can dominate in GC data set. It is possible that small peaks (or even substances which are not detected by GC) have more impact on smell than large peaks such as 142. Therefore, the information carried by the chromatograms is difficult to relate to the sensory information. Pre-processing in *N*-way data analysis is much less evident than in 2-way PCA and was therefore not included in this research. The effect of pre-processing on the Tucker3 model will be investigated in further research. The PARAFAC and the Tucker3 model gave the same information for the chromatographic data set. For the sensory data matrix the Tucker3 model allows more freedom in choosing the number of factors in each mode. Four factors extracted in the first and third mode were interpretable and obviously related. However, the PARAFAC confirmed the conclusions obtained by interpretation of the first core elements of Tucker3 model.

5. Conclusion

The presented results show that *N*-way methods are useful for the study of these types of data. *N*-way models offer detailed information about the data set and allow visualization of the data structure.

In this case, the Tucker3 model seems to be a more suitable tool for exploratory data analysis than the PARAFAC model. Indeed, it allows to choose a different number of factors in each mode which is more appropriate for the type of data, where one mode (here the amino acids) has a more dominant effect than the other mode (i.e. sugars).

For the Tucker3 model, the interpretation is based on significant core elements that express interactions among modes. Sometimes more than one conclusion can be obtained by interpreting certain core element. In that case it is necessary to check the conclusions by considering the original data.

The interpretation of PARAFAC model is performed in ways similar to the better known two-way PCA that many scientists are already familiar with, which is an advantage of this method.

References

[1] F. Questier, Q. Guo, B. Walczak, D.L. Massart, C. Boucon, S. de Jong, Sequential projection pursuit using genetic algorithms for data mining of analytical data, Anal. Chem. 72 (2000) 2846–2855.
[2] F. Questier, Q. Guo, B. Walczak, D.L. Massart, C. Boucon, S. de Jong, The neural gas network for classifying analytical data, Chemomet. Intell. Lab. Syst. 61 (2002) 105–121.
[3] P. Barbieri, C.A. Andersson, D.L. Massart, S. Predonzani, G. Adami, E. Reisenhofer, Modeling bio-geochemical interactions in the surface waters of the Gulf of Trieste by three-way principal component analysis (PCA), Anal. Chim. Acta 398 (1999) 227–235.
[4] S.P. Reinikainen, P. Laine, P. Minkkinen, P. Paatro, Factor analytical study on water quality in lake Simaa, Finland, Fresenius J. Anal. Chem. 369 (2001) 727–732.
[5] P. Paatero, S. Juntto, Determination of underlying components of a cyclical time series by means of two-way and three-way factor analytic techniques, J. Chemomet. 14 (2000) 241–259.
[6] D. Baunsgaard, C.A. Andersson, A. Arndal, L. Munck, Multi-way chemometrics for mathematical of fluorescent colorants and color precursors from spectrofluorimetry of beet sugar and beet sugar thick juice as validated by HPLC analysis, Food Chem. 70 (2000) 113–121.
[7] C.A. Andersson, L. Munck, R. Henrion, G. Henrion, Analysis of *N*-dimensional data arrays from fluorescence spectroscopy of an intermediary sugar product, Fresenius J. Anal. Chem. 359 (1997) 138–142.
[8] R. Leardi, C. Armanino, S. Lanteri, L. Albertonanza, Three-mode principal component analysis of monitoring data from Venice lagoon, J. Chemomet. 14 (2000) 197–211.
[9] R. Henrion, G. Henrion, G.C. Onuoha, Multi-way principal components analysis of a complex data array resulting from physico-chemical characterization of natural waters, Chemom. Intell. Lab. Syst. 16 (1992) 87–94.
[10] R. Henrion, *N*-way principal component analysis. Theory, algorithms and applications, Chemomet. Intell. Lab. Syst. 25 (1994) 1–23.
[11] R. Bro, C.A. Andersson, H.A.L. Kiers, PARAFAC2. Part II. Modeling chromatographic data with retention time shifts, J. Chemomet. 13 (1999) 295–309.
[12] R.A. Harshman, W.S. de Sarbo, An application of PARAFAC to a small sample problem, demonstrating preprocessing, orthogonality constraints, and split-half diagnostic techniques, Research Methods for Multimode Data Analysis, Praeger, New York, 1984, p. 602.
[13] D.J. Louwerse, A.K. Smilde, H.A.L. Kiers, Cross-validation of multiway component models, J. Chemomet. 13 (1999) 491–510.
[14] http://www.models.kvl.dk/source/.