ELSEVIER

Chemometrics and Intelligent Laboratory Systems 65 (2003) 83–95

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemometrics

Structure preserving feature selection in PARAFAC using a genetic algorithm and Procrustes analysis

W. Wu[a,*], Q. Guo[b], D.L. Massart[b], C. Boucon[c], S. de Jong[c]

[a] *Safety Assessment, GlaxoSmithKline Pharmaceuticals, The Frythe, Welwyn, Hertfordshire, AL6 9AR, UK*
[b] *ChemoAC, Pharmaceutical Institute, Vrije Universiteit Brussel, Laarbeeklaan 103, B-1090 Brussels, Belgium*
[c] *Unilever R&D Vlaardingen, PO Box 114, 3130 AC Vlaardingen, The Netherlands*

Received 18 June 2002; received in revised form 28 August 2002; accepted 30 August 2002

Abstract

In this paper, a method is proposed to select subsets of variables in parallel factor analysis (PARAFAC), such that information in the complete multi-way data set is preserved as much as possible. The information retained is measured by means of the percentage of consensus in Procrustes analysis. The best *N*-way subset is obtained by applying a genetic algorithm (GA) to optimize the consensus between the subset and the complete *N*-way data set in order to prevent exhaustive searching. The method was applied to two industrial data sets: a three-way sensory data set and a four-way gas chromatography (GC) data set. The results showed that the proposed method successfully identified structure-bearing variables in both data sets and that it led to better subsets of variables than feature selection based on loadings.

Keywords: Feature selection; Multi-way analysis; Parallel factor analysis (PARAFAC); Genetic algorithm; Procrustes analysis

1. Introduction

Multi-way analysis is gaining more and more interests in practical applications [1–5]. Hyphenated analytical techniques often lead to multi-way data arrays [4], although they are sometimes not recognized as such. In general, multi-way (also known as multi-mode or multi-order) data are described by several sets of variables measured in a crossed way [2,6]. For instance, a typical three-way sensory data set represents measures of a set of attributes (variables) on samples by different judges. As the size of such data sets tend to become larger and larger, demands to simplify the model for easy interpretation constantly increase.

Parallel factor analysis (PARAFAC) is one of the multi-way decomposition methods [1]. Other methods are the Tucker3 method and unfolding two-way principal component analysis (PCA) [2]. The latter converts the multi-way structure of the data into a two-way structure and then works with a bilinear decomposition. Tucker3 is a truly multilinear method that takes into account the multi-way structure, so that it is more parsimonious than the unfolding method. As a simplified version of Tucker3, by forcing the core array to a (super)diagonal form, the PARAFAC model is even more condensed than that of Tucker3. In chemomet-

[*] Corresponding author. Tel.: +44-1438-782940; fax: +44-1438-782582.
E-mail address: wen_2_wu@gsk.com (W. Wu).

0169-7439/02/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0169-7439(02)00105-3

rics, PARAFAC is more frequently applied to explore multi-way data because of its properties of uniqueness and simplicity [2,4,5]. It is a natural extension of traditional two-way PCA. In PARAFAC, multi-way data are decomposed into sets of scores and loadings with the same number of columns (latent variables). The number of latent variables (factors or components) is much lower than the number of original variables, so that the data can be visualised in a reduced dimensional space. However, similar to PCA, PARAFAC does not eliminate the original variables, as it uses all the original variables to generate the new latent variables. For interpretation purposes or future investigation, it would often be very useful to reduce the number of variables. Nowadays, due to the rapid development of analytical techniques, the number of variables usually is very large, which implies a lot of redundant information. In a previous study, we proposed new feature selection methods in PCA [7] and sequential projection pursuit (SPP) [8]. In the present paper, the feature selection methods are further extended from two-way to three- or N-way analysis method using PARAFAC. The method selects the best N-way subset that keeps, as much as possible, the structure information of the complete multi-way data set when all variation of the complete data is interesting.

2. Theory

2.1. Notation and nomenclature

$\underline{\boldsymbol{X}}$	$I \times J \times K$, a three-way data array with I rows, J columns and K layers
$\underline{\boldsymbol{X}_{\mathbf{s}}}$	$I \times J \times K_s$, a three-way subset data array with I rows, J columns and K_s selected layers
$\mathbf{A}$	$I \times F$, the loading matrix of $\underline{\boldsymbol{X}}$ in the first mode
$\mathbf{A}_s$	$I \times F$, the loading matrix of $\underline{\boldsymbol{X}_{\mathbf{s}}}$ in the first mode
$\mathbf{B}$	$J \times F$, the loading matrix of $\underline{\boldsymbol{X}}$ in the second mode
$\mathbf{B}_s$	$J \times F$, the loading matrix of $\underline{\boldsymbol{X}_{\mathbf{s}}}$ in the second mode
$\mathbf{C}$	$K \times F$, the loading matrix of $\underline{\boldsymbol{X}}$ in the third mode
$\mathbf{C}_s$	$K_s \times F$, the loading matrix of $\underline{\boldsymbol{X}_{\mathbf{s}}}$ in the third mode
$\mathbf{Y}_{\mathbf{A}}$	$I \times F$, the prescaled matrix of $\mathbf{A}$ (total variance = 50)
$\mathbf{Y}_{\mathbf{A}_s}$	$I \times F$, the prescaled matrix of $\mathbf{A}_s$ (total variance = 50)
$\mathbf{Y}_{\mathbf{B}}$	$J \times F$, the prescaled matrix of $\mathbf{B}$ (total variance = 50)
$\mathbf{Y}_{\mathbf{B}_s}$	$J \times F$, the prescaled matrix of $\mathbf{B}_s$ (total variance = 50)

In PARAFAC, to deal with several modes or ways, it is necessary to specify the modes where subsets of variables will be selected. We will assume that variable selection applies to the last mode(s), the one(s) with a large number of variables that are not of primary interest. Such modes are called selection modes. In contrast, the first mode(s) are of primary interest. These are called the structure modes.

2.2. PARAFAC

PARAFAC is a natural extension of PCA [4], which can be described as a bilinear decomposition of data. The decomposition can be modeled as

$$x_{ij} = \sum_{f=1}^{F} a_{if} b_{jf} + e_{ij} \tag{0}$$

where x_{ij} is the element at the ith row and jth column of the data matrix $\underline{\boldsymbol{X}}$. A three-way PARAFAC model is a trilinear decomposition of three-way data (x_{ijk}) [4], which can be expressed as

$$x_{ijk} = \sum_{f=1}^{F} a_{if} b_{jf} c_{kf} + e_{ijk} \tag{1}$$

Similarly, an N-way data matrix can be modeled by multilinear decomposition. With PARAFAC, a set of matrices is obtained corresponding to each way (mode or order). The first matrix (**A**) is often referred to as score matrix, and the other matrices (**B**, **C**, ...) are called loading matrices. The score matrix corresponding to samples, e.g. the first mode, is not normalized, while each loading matrix is scaled to a sum of squares equal to one. In order to simplify the notations, hereafter, both score and loading matrices will be called loading matrices, as often done by authors publishing in this field [2–4].

The number of factors (F) in the PARAFAC model can be determined by various methods [2,9], such as the percentage of variation, which is a simple and effective way. A new efficient method developed by Bro and Kiers [10], called the core consistency diagnostic (CORCONDIA), for determining the number of components in PARAFAC models is used to estimate F.

2.3. Feature selection method based on loadings

One straightforward way to locate important original variables is based on loadings. For any factor, high loadings in absolute value indicate that corresponding variables contribute more to the factor than other variables. Usually the first few, e.g. F, factors are regarded as significant. Then, for each selection mode, one chooses the subset of variables that exhibit the highest loadings among the F factors.

2.4. Feature selection using Procrustes analysis and genetic algorithm

Generalized Procrustes analysis (GPA) [12] is a statistical method to match data sets that are measured from the same samples or objects with different variables. In the special case, when only two data sets are involved, it is called Procrustes analysis. GPA has been widely used in food science to analyze sensory data [21], and recently, it has been applied in mining genomic and proteomic data [22]. In analytical chemistry, Procrustes analysis has been applied for discriminant analysis [23,24], calibration transfer [25], variable selection [11,26,27] and spectral analysis [28,29]. In Brereton's group, it was applied to analyze mass spectral data [30,31] and high-performance liquid chromatography data [32]. In our group, Procrustes analysis was combined with genetic algorithm (GA) for feature selection in PCA [7] and SPP [8]. Here, the method is further modified for PARAFAC.

The previously described method approximates the loadings of the selection modes by setting low loading values equal to zero. It is not clear how this affects the loadings of the other mode(s) if the reduced data set is subjected to a new PARAFAC analysis. Ideally, one would like to retain the information of the structure modes as it is contained in the corresponding PARAFAC loading matrices. In the following methods, we aim at selecting the N-way subset that best retains the structure information of the complete N-way data.

The similarity between the structure information of the complete set and the subset is quantified by a Procrustes criterion [11]. In PCA, the similarity between the loading matrices of a candidate subset and the complete data is measured by a percentage of consensus value after optimal Procrustes matching [7,8]. Here, for each structure mode, a similar procedure is applied to assess the similarity between the two PARAFAC loading matrices of an N-way candidate subset and the complete multi-way data. The total structure information retained by the candidate N-way subset can be obtained by considering the consensus values of all structure modes.

The method proposed here seeks for subset selections that least affect the loading patterns of the structure modes viewed separately. For exploratory purposes, it usually suffices to produce and interpret the loading patterns of the separate modes, and the present method is perfectly geared to retain such information with a reduced set of variables in the selection mode. One should notice, however, that a good Procrustes match of the loading patterns is a weaker requirement than asking to preserve the complete PARAFAC model structure. If the data are generated by a process closely obeying a PARAFAC model, then one expects only small departures in the structure mode when variables in the selection mode are omitted. A good Procrustes match of the loading patterns is still a prerequisite for the stronger requirement of preserving the complete PARAFAC structure of the structure modes. The prescaling, translation, rotation and reflection in Procrustes match are completely different from the preprocessing of data before PARAFAC. These operations are only on loading matrix after PARAFAC to calculate the searching criterion and have no effect on the N-way structure in PARAFAC.

In Fig. 1, a three-way data array $\underline{\boldsymbol{X}}$ $(I \times J \times K)$ is shown as an example to describe the procedure for the estimation of the Procrustes criterion, where $\underline{\boldsymbol{X}}_{\mathbf{s}}$ $(I \times J \times K_{\mathrm{s}})$ is a three-way subset of $\underline{\boldsymbol{X}}$ with K_{s} selected variables $(F \leq K_{\mathrm{s}} < K)$ in the third mode. Suppose it is desired to preserve the structure information in both the first and second modes, with the first F factors after PARAFAC modeling of $\underline{\boldsymbol{X}}$ and $\underline{\boldsymbol{X}}_{\mathbf{s}}$, respectively, $\mathbf{A}$ and $\mathbf{A}_{\mathrm{s}}$ are the loading matrices for the first mode, and $\mathbf{B}$ and $\mathbf{B}_{\mathrm{s}}$ are the corresponding loading matrices for the second mode. For mode 1, prescaling is introduced to pretreat $\mathbf{A}$ and $\mathbf{A}_{\mathrm{s}}$, ensuring both configurations ($\mathbf{Y_A}$ and $\mathbf{Y}_{\mathbf{A}_{\mathrm{s}}}$) have equal variances of 50, giving rise to total variance of 100. To measure the agreement between the two configurations, $\mathbf{Y_A}$ and $\mathbf{Y}_{\mathbf{A}_{\mathrm{s}}}$ are subjected to Procrustes analysis [12]. After optimally matching the two matrices by means of translation, rotation and reflection, a percentage consensus value [7,8] is

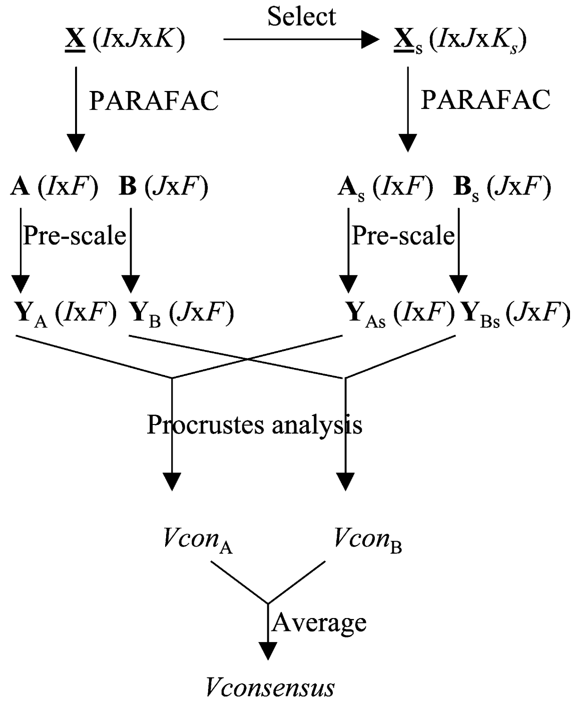

Fig. 1. The flow chart for a three-way feature selection method.

calculated as the measurement of the similarity between $\mathbf{A}$ and $\mathbf{A}_s$ by

$$V_{\mathrm{con_A}} = 2(\lambda_1 + \lambda_2 + \ldots + \lambda_F) \tag{2}$$

where λ_1, λ_2, ..., λ_F are the singular values of the matrix $\mathbf{Y}_\mathbf{A}^\mathrm{T}\mathbf{Y}_{\mathbf{A}_s}$. The same procedure is followed for the other predefined structure modes, e.g. mode 2, to obtain the consensus value ($V_{\mathrm{con_B}}$) expressing the percentage of preserved information in the second mode. The total retained structure information is measured by the average of the two consensus values:

$$V_{\mathrm{consensus}} = (V_{\mathrm{con_A}} + V_{\mathrm{con_B}})/2 \tag{3}$$

$V_{\mathrm{consensus}}$ is a percentage value like $V_{\mathrm{con_A}}$ and $V_{\mathrm{con_B}}$. In the extreme case that the three-way subset preserves all the structure information of the entire three-way data, the consensus is equal to 100. The procedure can be extended to more than three-way situations.

The best N-way subset is the one with highest $V_{\mathrm{consensus}}$ among all possible subsets. An exhaustive searching procedure is not computationally feasible. In order to search for the best subset efficiently, a genetic algorithm (GA) [13,14] is applied. This is designated as GA–PARAFAC.

Genetic algorithms provide a powerful means to search for a global optimum in a high-dimensional space [15,16]. We applied the one developed by Leardi et al. [13,14], originally applied to select the best subset of variables to build a multiple linear regression model in calibration. In PARAFAC, feature selection can be made according to one mode or multiple modes. To select variables from a mode, a string vector is produced, containing the same number of elements as the total number of variables in the predefined feature selection mode. To select variables from multiple modes such as modes 2 and 3, the vector consists of two subvectors corresponding to the two modes. The length of the string vector is equal to the total number of variables in modes 2 and 3. In the vector, the first J elements in the first subvector correspond to the J variables of mode 2 and the last K elements in the second subvector of those in mode 3. Each element is coded as 1 if the corresponding variable was chosen and 0 if otherwise. The initial population of solutions consists of a certain number of strings (e.g. 30), which were randomly created and evaluated for their fitness. The fitness measures the structural information preserved by each string (subset) in the population, as evaluated by Eqs. (2) and (3). Pairs of parent strings are selected according to a probability value that is proportional to the quality of their fitness. Then they undergo ‘reproduction’ by applying a crossover operator (e.g. probability of 0.5) and, less frequently, a mutation operator (e.g. probability of 0.02). This results in two children strings for each pair of parents. They become new candidates and join the population forming a new generation. This procedure is repeated a (prespecified) number of times (e.g. 500). In the genetic algorithm (GA) [13], the maximal number of variables retained (the 1s) must be predefined by the user. Usually, it is decided according to experience. Alternatively, one may try to decide according to a penalty function; since the more variables selected, the higher consensus value could be obtained.

The proposed GA–PARAFAC algorithm is summarized as follows:

(1) Apply PARAFAC on $\underline{\boldsymbol{X}}$ to obtain loading matrices and define loading matrices (e.g. **A**, **B**, **C**, ...) with F significant factors as structure modes.
(2) Prescale the loading matrices **A**, **B**, **C**, ... yielding $\mathbf{Y_A}$, $\mathbf{Y_B}$, $\mathbf{Y_C}$, ... as the target matrices in Procrustes analysis.
(3) Create a candidate subset $\underline{\boldsymbol{X}}_{\mathbf{s}}$ with randomly chosen variables for each of the selection modes.
(4) Apply PARAFAC on $\underline{\boldsymbol{X}}_{\mathbf{s}}$ to obtain the corresponding loading matrices ($\mathbf{A}_s$, $\mathbf{B}_s$, $\mathbf{C}_s$, ...) with F factors.
(5) Prescale the loading matrices $\mathbf{A}_s$, $\mathbf{B}_s$, $\mathbf{C}_s$, ... yielding $\mathbf{Y}_{\mathbf{A}_s}$, $\mathbf{Y}_{\mathbf{B}_s}$, $\mathbf{Y}_{\mathbf{C}_s}$, ...
(6) Calculate $V_{\text{consensus}}$ according to Eqs. (2) and (3) as the fitness of the candidate subset.
(7) Repeat steps 3–6 to construct a set of candidate subsets as the initial population and incorporate GA to search for the best subset.

3. Experimental

3.1. Data

3.1.1. Three-way bread sensory data

The three-way data set (10 × 8 × 11) was given by Bro [19] as a case study for comparing unfolding PCA and PARAFAC. Ten bread samples were assessed by eight judges, each scoring 11 attributes: bread odor, yeast odor, off-flavor, color, moisture, tough, salt taste, sweet taste, yeast taste, other taste and total. The 10 samples are pairwise replicates. Bro [20] concluded that PARAFAC performs better than unfolding-PCA for both simplicity and interpretation of the model.

3.1.2. Four-way GC data for food samples

A set of gas chromatographic (GC) data of reaction product mixtures was obtained for investigation of the Maillard reaction. This data set has been studied as two-way data for several purposes [7,8,18]. The data set comprises chromatograms with 199 detected peaks. Each sample is the reaction product of one of six sugars: fructose, glucose, lactose, maltose, rhamnose

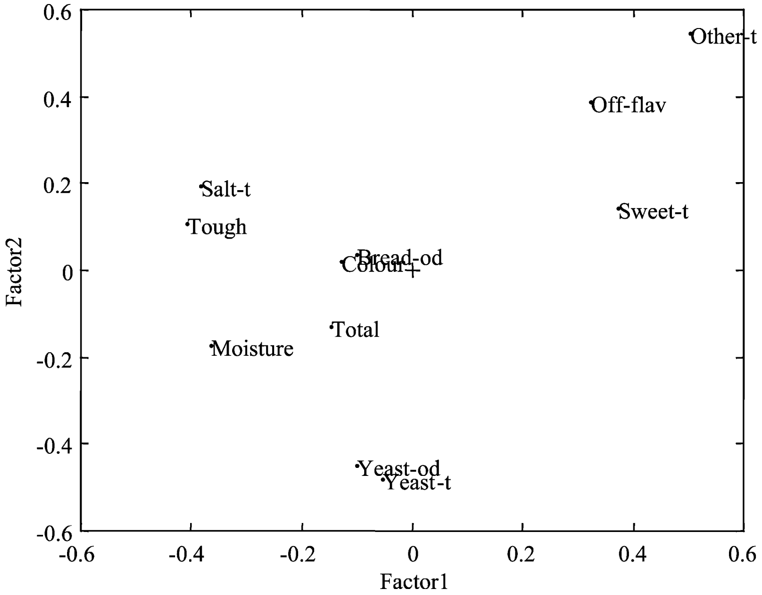

Fig. 2. Loading plot for the mode of attributes using complete data. The 11 attributes are: (1) bread odor (Bread-od), (2) yeast odor (Yeast-od), (3) off-flavor (Off-flav), (4) color, (5) moisture, (6) tough, (7) salt taste (Salt-t), (8) sweet taste (Sweet-t), (9) yeast taste (Yeast-t), (10) other taste (Other-t) and (11) total.

and xylose, and one of nine amino acids: alanine, asparagine, glutamine, glycine, threonine, arginine, cysteine, lysine and glutamate. The reaction was carried out at two different pH levels (pH = 6.5 and 7.9). Here, a subset of the data with 81 samples is rearranged into a four-way array (6 sugars × 9 amino

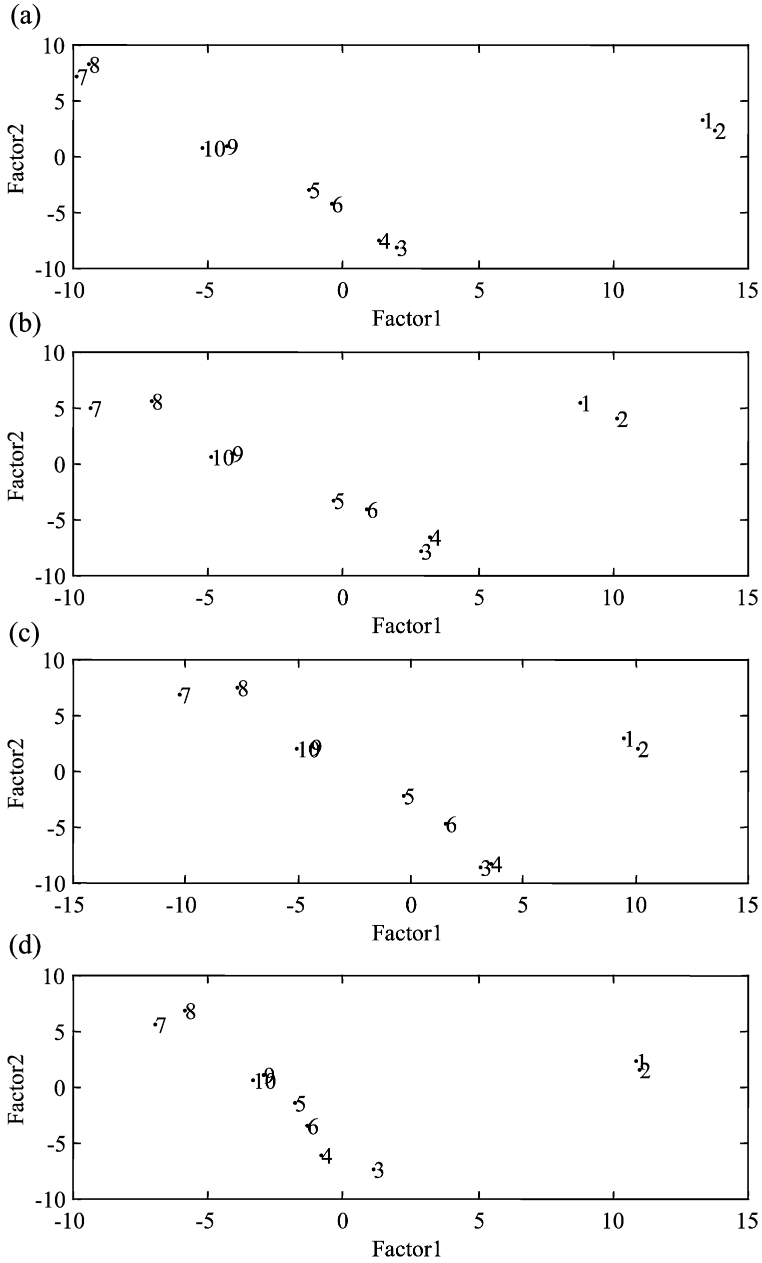

Fig. 3. Loading plot for the mode of samples for three-way bread sensory data using (a) all variables, (b) five attributes selected by GA-PARAFAC with two structure modes (sample and judge), (c) five attributes selected by loadings, (d) five attributes by GA-PARAFAC with one structure mode (sample), (e) five attributes and four judges selected by GA-PARAFAC and (f) five attributes and four judges selected by loadings.

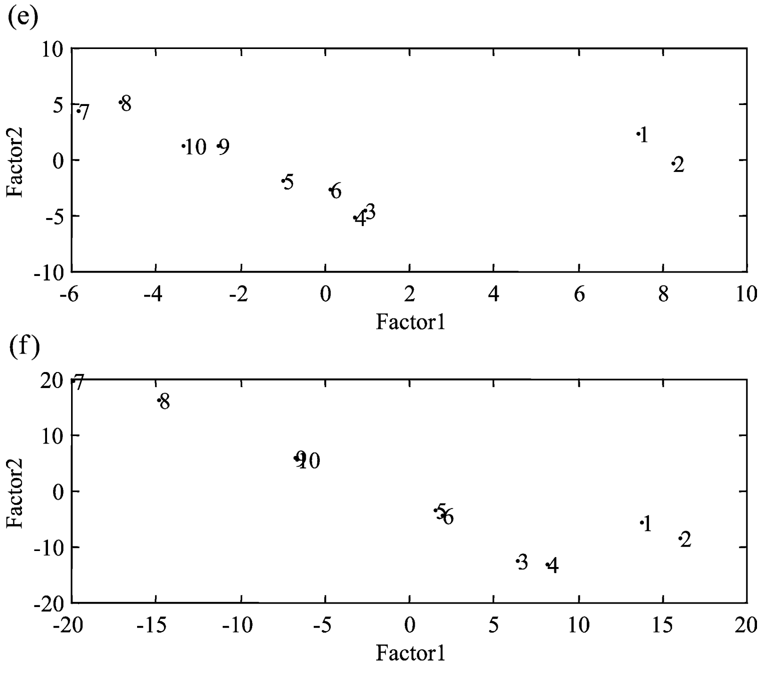

Fig. 3 (*continued*).

acids × 2 pH levels × 199 peaks). There are 27 samples missing, i.e. 27 × 199 missing elements in the four-way array.

3.2. Software

The data are analyzed using a program in MATLAB (version 4.0). The m-files for PARAFAC were from Bro and Andersson [17]. They can handle missing data and a variety of constraints. Other programs were developed by ourselves.

4. Results and discussion

4.1. Three-way bread sensory data

The three-way data are centered across the first mode (the samples). Two PARAFAC factors are sufficient for the centered data according to the CORCONDIA criterion [10]. The loading plot of mode 2 (Fig. 2) shows that the attributes "salt taste", "tough", "color" and "bread odor" are located more or less in the same direction, indicating high correlation between them. This is also the case for the attributes "total" and "moisture", for "yeast odor" and "yeast taste" and for "off-flavor" and "other taste". The attribute "sweet taste" is relatively different from others. Therefore, the 11 attributes could be grouped into five clusters according to their correlation. Five attributes should be able to well represent all 11 attributes since five is higher than the number of significant factors.

The loading plot of the sample mode (Fig. 3a) demonstrates that the 10 samples are grouped in five pairs. This indicates that the PARAFAC model correctly identifies the five pairs of replicate samples. All pairs of samples are located on a line except samples 1 and 2. Fig. 4a shows that the judges 1–7 give quite similar assessments and that judge 8 is located relatively far from the other panelists.

From the results in these loading plots, one might expect to find a three-way subset by selecting five attributes instead of all attributes, but still to represent the same structure information in both modes of samples and judges. Or one might want the three-way subset to retain only the information in the mode of samples. In practice, it is also

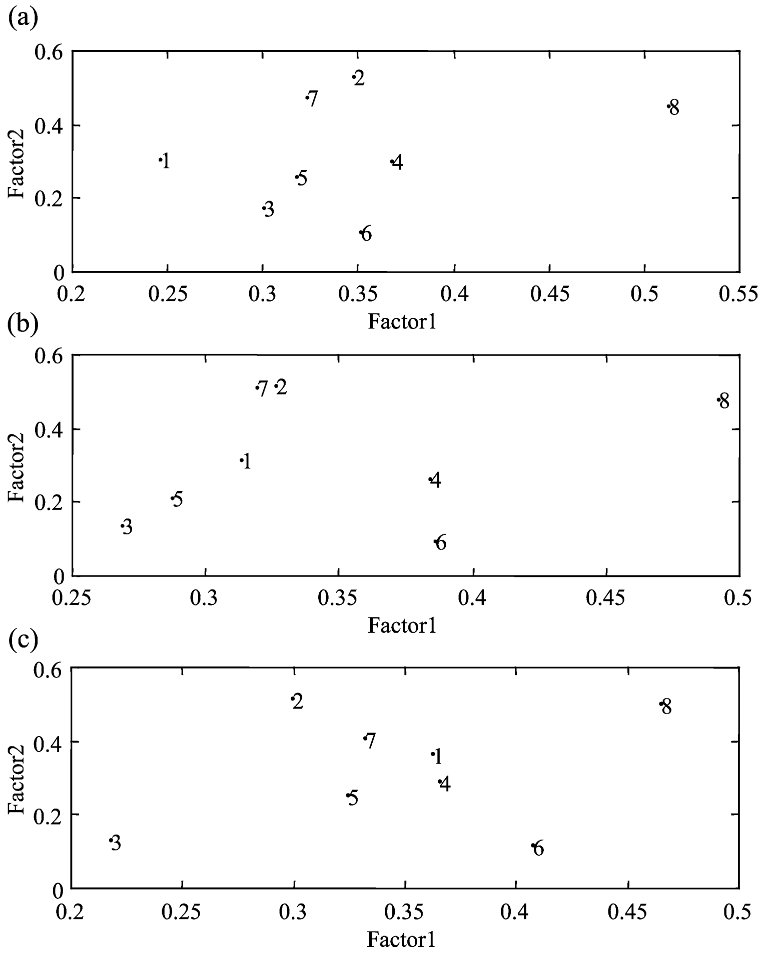

Fig. 4. Loading plot for the mode of judges for three-way bread sensory data using (a) all variables, (b) five attributes selected by GA-PARAFAC with two structure modes (sample and judge) and (c) five attributes selected by loadings.

desirable to reduce both attributes and number of judges, but to keep the same information in the sample mode.

Feature selection is systematically studied by both GA–PARAFAC and the loading method. The numerical results in Table 1 show that all the

Table 1
The results obtained with the two feature selection methods for the three-way bread sensory data with different combinations of feature selection modes and structure modes

Method	Selection mode(s)	Structure mode(s)	$V_{\text{consensus}}$ (total)	V_{con_A} (sample mode)	V_{con_B} (judge mode)	Selected attributes	Selected judges
GA-PARAFAC	C	A and B	97.95	98.84	97.06	3, 6, 7, 9, 10	not applicable
Loading	C	A and B	93.99	97.71	90.27	2, 6, 7, 9, 10	not applicable
GA-PARAFAC	C	A	99.31	99.31	not applicable	2, 6, 8, 9, 10	not applicable
Loading	C	A	97.71	97.71	not applicable	2, 6, 7, 9, 10	not applicable
GA-PARAFAC	B and C	A	99.29	99.29	not applicable	4, 5, 7, 9, 10	2, 4, 6, 7
Loading	B and C	A	90.45	90.45	not applicable	2, 6, 7, 9, 10	2, 4, 7, 8

selected subsets in general have high total consensus values. They all retain more than 90% of the structural information in the complete three-way data. Among the five selected attributes in all subsets, two attributes ((9) yeast taste and (10) other taste) are always selected. For all the three situations, GA–PARAFAC always retains more structural information than the loading method, as expected. But one should always examine loading plots in order to avoid bias.

For the situation with one feature selection mode and two structure modes, the five attributes selected by GA–PARAFAC retain 98.8% information in the sample mode and 97.1% information in the judge mode, resulting in 98.0% total structural information of the complete three-way data. The sample structure in Fig. 3b is indeed quite similar to that in Fig. 3a, and the clustering pattern of judges in Fig. 4b is also like that obtained from the complete data in Fig. 4a. Figs. 3c and 4c are the loading plots of the sample and judge modes of the subset selected by the loading method. Fig. 3c matches Fig. 3a well with a consensus value of 97.7. There is no apparent difference between Fig. 3c and b and their consensus values are very close (97.7 and 98.8). The judges pattern in Fig. 4c differs from that in Fig. 4a, with the consensus value dropping from 97.1 to 90.3 compared to Fig. 4b.

For the second situation with one feature selection mode and one structure mode, GA–PARAFAC leads to a subset of attributes ((2) yeast odor, (6) tough, (8) sweet taste, (9) yeast taste and (10) other taste) with the highest consensus value (99.3). Fig. 3d shows that the 10 samples are distributed similarly as in Fig. 3a, except that the distance between the paired samples 3 and 4 becomes larger. The loading method gives the same subset ((2) yeast odor, (6) tough, (7) salt taste, (9) yeast taste and (10) other taste) as in the first situation (the second row in Table 1), since the selected attributes only depend on the selection mode and are independent on the structure mode.

For the third situation with two feature selection modes and one structure mode, the subset obtained from GA–PARAFAC explained 99.3% information of the complete data by using five selected attributes and four selected judges. The retained information is about 9% higher than that obtained by the loading method, which gives the lowest consensus value (90.5). The loading plots in Fig. 3e–f shows that they both are generally consistent with Fig. 3a, but the replicated samples 1 and 2 lie more closely to the line of the other samples in Fig. 3f than in the other plots (Fig. 3a–e).

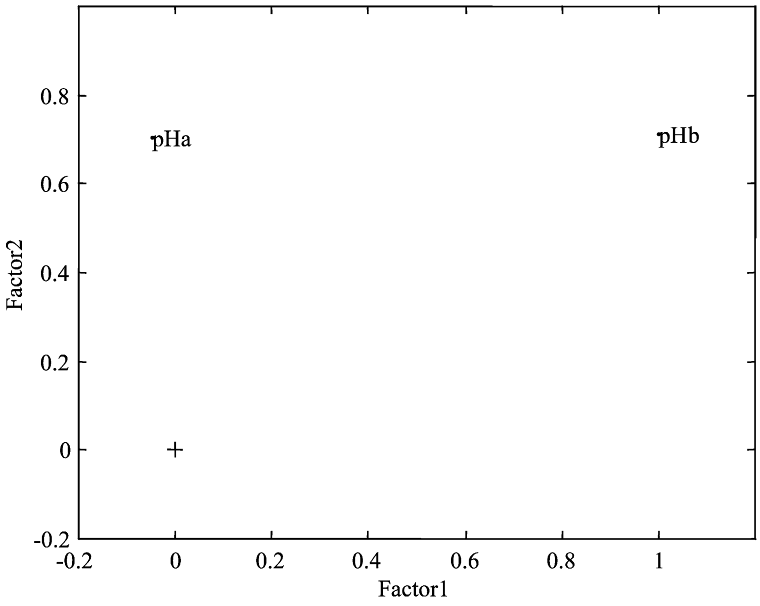

Fig. 5. Loading plot for the pH mode for four-way GC data using all variables.

4.2. Four-way GC data for food samples

Since the samples are arranged into more than one mode and all samples have been preprocessed by baseline subtraction, the four-way data are not centered. For this data set, the aim is to select about 10% of the GC peaks that represent the same information about sugars and amino acids in the complete data. Two factors are detected as significant by the CORCONDIA criterion [10] in the four-way model of PARAFAC. Fig. 5 shows that the first component mainly explains the difference between the two pH levels.

In Fig. 6a, the six sugars are observed in four clusters. Rhamnose is clearly separated from other sugars on the upper right; maltose and lactose are clustered on the left; fructose and glucose locate in the middle and xylose in the lowest part along factor 2. Quite similar clusters are found in Fig. 6b from a subset of 20 peaks selected by GA–PARAFAC. The similarity amounts to 99.7% (Table 2), and the four clusters in Fig. 6b are separated even more clearly than in Fig. 6a. Such cluster separation appears less obvious in Fig. 6c, which retains 96.6% information in Fig. 6a. This indicates that the loading method catches less structural information than GA–PARAFAC.

Fig. 7a shows the behavior of the nine amino acids in five groups in the model of the complete data. Along factor 1, a group of lysine, glutamate and arginine clearly separates from the others on the right. Along factor 2, alanine and asparagine are clustered on the

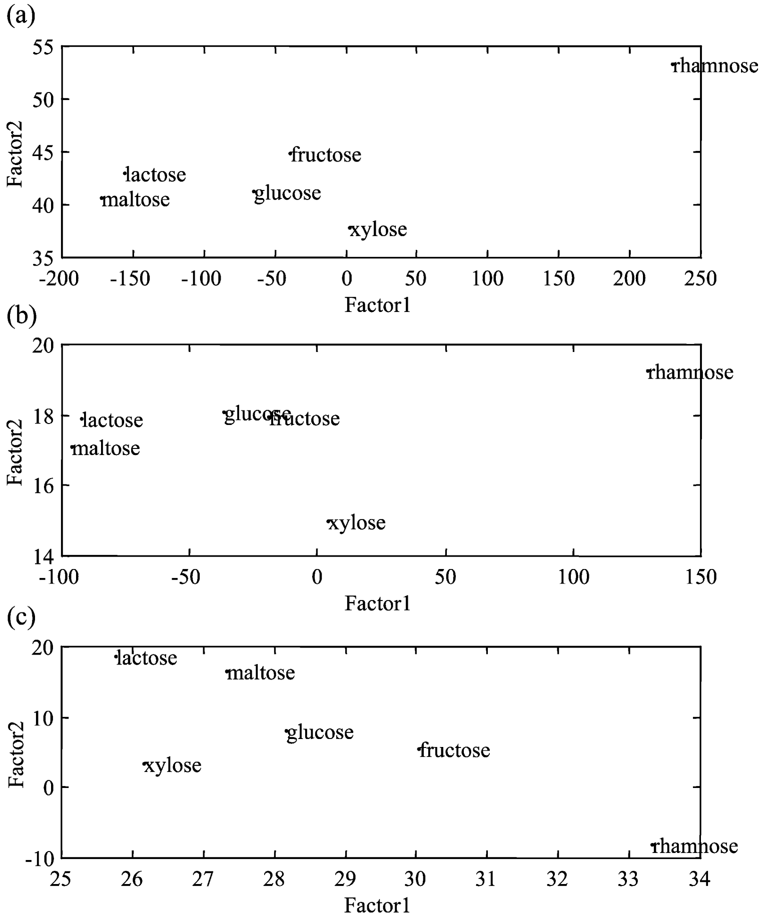

Fig. 6. Loading plot for the mode of sugars for four-way GC data using (a) all variables, (b) 20 peaks selected by GA-PARAFAC and (c) 20 peaks selected by loadings.

Table 2
The consensus values and selected variables for the two feature selection methods of the four-way GC data

Method	$V_{\mathrm{consensus}}$ (total)	V_{con_A} (sugar mode)	V_{con_B} (amino acid mode)	20 selected peaks
GA-PARAFAC	99.83	99.72	99.94	12, **32**, 35, **47**, 57, 58, 63, 71, 79, 88, 90, **103**, **112**, **117**, **118** , 128, 143, 149, **151**, 153
Loading	80.96	96.59	65.32	18, 25, 29, 31, **32**, 33, 43, **47**, 60, **103**, **112**, **117**, **118**, 137, 141, 142, 145, 148, **151**, 152

The bold numbers indicate peaks common for both methods.

top, and glutamine and threonine next. Glycine is located between cysteine (which stays on the bottom of factor 2) and the cluster of glutamine and threonine. A similar and even more clear clustering pattern is observed in Fig. 7b, implying that GA–PARAFAC performs well. The subset of peaks preserves 99.9% information by choosing only 20 instead of all 199. The result of the loading method (Fig. 7c) shows a different distribution of the amino acids in Fig. 7c from that in Fig. 7a, and their similarity is only 65.3%. The loading method fails to catch the structural information in the amino acid mode. The comparison results in Table 2

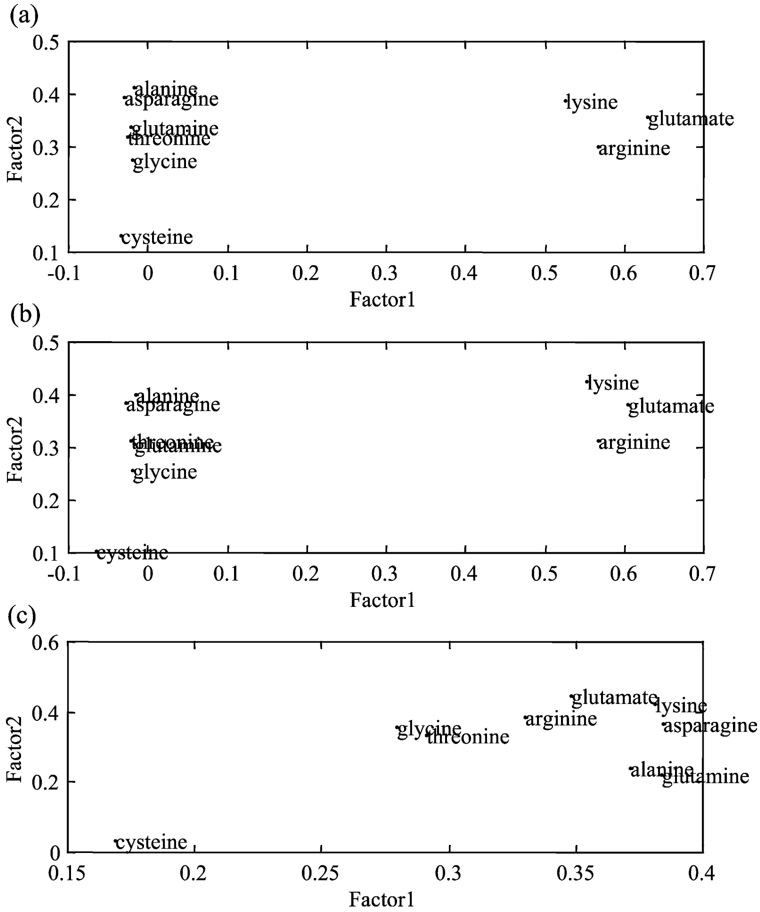

Fig. 7. Loading plot for the mode of amino acids for four-way GC data using (a) all variables, (b) 20 peaks selected by GA-PARAFAC and (c) 20 peaks selected by loadings.

show that GA–PARAFAC retains 99.8% total structural information, which is about 19% higher than the loading method. The two subsets of variables selected by the two methods (Table 2) vary a lot, and there are only seven peaks in common.

5. Conclusion

A new method is proposed to select features with high-dimensional *N*-way data. The performance of the method was studied on a three-way sensory data set and a four-way food chemistry (FC) data set. The results show that the proposed method leads to a better *N*-way subset than the loading method. For the sensory data, a three-way subset with five attributes and four judges preserves more than 99% general features presented in the three-way data with all 11 attributes and eight judges. For the GC data, 20 selected peaks account for up to 99.8% consensus of the complete data with 199 original peaks. Similar to the feature selection methods in two-way PCA [7], the proposed method uses the full model from the complete data as target. When the full model is not reasonable because of the large amount of irrelevant information, one needs to spend extra efforts to explore data to find a reasonable model and uses it to replace the full model in the method. For two-way data, the sequential projection pursuit [8] can be used to find a reasonable model. But for *N*-way data, it is still under investigation. Like most feature selection methods in two-way analysis, the number of selected variables should be predefined. The proposed method cannot be applied if one wants to reduce the number of variables lower than the number of factors. This feature selection methodology can also find applications in other multi-way methods such as *N*-way PLS.

Acknowledgements

The authors thank the EU for the financial assistance on the *N*-way methods project.

References

[1] B.G.M. Vandeginste, D.L. Massart, L.C.M. Buydens, S. de Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics: Part B, Elsevier, Amsterdam, 1998.

[2] R. Bro, PARAFAC. Tutorial and applications, Chemometrics and Intelligent Laboratory Systems 38 (1997) 149–171.

[3] R. Bro, Multiway calibration. Multilinear PLS, Journal of Chemometrics 10 (1996) 47–61.

[4] R. Bro, J.J. Workman, P.R. Mobley, B.R. Kowalski, Review of chemometrics applied to spectroscopy: 1985–95, Part 3—multi-way analysis, Applied Spectroscopy Reviews 32 (1997) 237–261.

[5] A.W. Czarnik, Combinatorial chemistry, Analytical Chemistry 70 (1998) 378A–386A.

[6] P. Geladi, Analysis of multi-way (multi-mode) data, Chemometrics and Intelligent Laboratory Systems 7 (1989) 11–30.

[7] Q. Guo, W. Wu, D.L. Massart, C. Boucon, S. de Jong, Feature selection in principal component analysis of analytical data, Chemometrics and Intelligent Laboratory Systems 61 (2002) 123–132.

[8] Q. Guo, W. Wu, F. Questier, D.L. Massart, C. Boucon, S. de Jong, Sequential projection pursuit using genetic algorithms for data mining of analytical data, Analytical Chemistry 72 (2000) 2846–2855.

[9] D.J. Louwerse, H.A.L. Kiers, A.K. Smilde, Cross-validation of multi-way component models, Journal of Chemometrics 13 (1999) 491–510.

[10] R. Bro, H.A.L. Kiers, A new efficient method for determining the number of components in PARAFAC models, Journal of Chemometrics (in press).

[11] W.J. Krzanowski, Selection of variables to preserve multivariate data structure, using principal components, Applied Statistics 36 (1987) 22–33.

[12] J.C. Gower, Generalised Procrustes analysis, Psychometrika 40 (1975) 33–51.

[13] R. Leardi, R. Boggia, M. Terrile, Genetic algorithms as a strategy for feature selection, Journal of Chemometrics 6 (1992) 267–281.

[14] R. Leardi, Application of genetic algorithms to feature selection under full validation conditions and to outlier detection, Journal of Chemometrics 8 (1994) 65–79.

[15] D.E. Goldberg, Genetic Algorithms in Search, Optimisation and Machine Learning, Addison-Wesley Publishing, Reading, MA, 1989.

[16] C.B. Lucasius, G. Kateman, Understanding and using genetic algorithms: Part 1. Concepts, properties and context, Chemometrics and Intelligent Laboratory Systems 19 (1993) 1–33.

[17] Available at: http://www.models.kvl.dk/research/source.

[18] Q. Guo, W. Wu, D.L. Massart, C. Boucon, S. de Jong, Feature selection in sequential projection pursuit, Analytica Chimica Acta 446 (2001) 85–96.

[19] Available at: http://www.models.kvl.dk/research/data.

[20] R. Bro, Multi-way analysis in the food industry. Models, algorithms, and applications, PhD thesis, University of Amsterdam (NL) and Royal Veterinary and Agricultural University (DK), 1998, pp. 196–203.

[21] W. Wu, Q. Guo, S. de Jong, D.L. Massart, Randomisation test for the number of dimensions of the group average space in generalised Procrustes analysis, Food Quality and Preference 13 (2002) 191–200.

[22] W. Wu, Validation of consensus between proteomic/genomic

expression and clinical chemical data by a new randomisation *F*-test in generalised Procrustes analysis, Oral Presentation at the Eighth Chemometrics in Analytical Chemistry Conference, 9/22–26/2002, Seattle, USA.

[23] D. Gonzalez-Arjona, G. Lopez-Perez, A.G. Gonzalez, Holmes, a program for performing Procrustes transformations, Chemometrics and Intelligent Laboratory Systems 57 (2001) 133–137.

[24] D. Gonzalez-Arjona, G. Lopez-Perez, A.G. Gonzalez, Performing Procrustes discriminant analysis with HOLMES, Talanta 49 (1999) 189–197.

[25] C.E. Anderson, J.H. Kalivas, Fundamentals of calibration transfer through Procrustes analysis, Applied Spectroscopy 53 (1999) 1268–1276.

[26] J.R. King, D.A. Jackson, Variable selection in large environmental data sets using principal components analysis, Environmetrics 10 (1999) 67–77.

[27] G. Scarponi, I. Moret, G. Capodaglio, M. Romanazzi, Cross-validation, influential observations and selection of variables in chemometric studies of wines by principal component analysis, Journal of Chemometrics 4 (1990) 217–240.

[28] L. Scarminio, M. Kubista, Analysis of correlated spectral data, Analytical Chemistry 65 (1993) 409–416.

[29] M. Kubista, A new method for the analysis of correlated data using Procrustes rotation which is suitable for spectral analysis, Chemometrics and Intelligent Laboratory Systems 7 (1990) 273–279.

[30] C. Bessant, R.G. Brereton, S. Dunkerley, Integrated processing of triply coupled diode array liquid chromatography electrospray mass spectrometric signals by chemometric methods, Analyst 124 (1999) 1733–1744.

[31] C. Demir, P. Hindmarch, R.G. Brereton, Procrustes analysis for the determination of number of significant masses in gas chromatography mass spectrometry, Analyst 121 (1996) 1443–1449.

[32] R.G. Brereton, D.V. McCalley, Procrustes analysis for the comparison of test methods in reversed-phase high performance liquid chromatography of basic compounds, Analyst 124 (1999) 227–238.