Received: 5 June 2007, Revised: 4 October 2007, Accepted: 26 October 2007, Published online in Wiley InterScience: 2007

(www.interscience.wiley.com) DOI: 10.1002/cem.1106

New exploratory clustering tool

Evrim Acar[a]*, Rasmus Bro[b] and Bonnie Schmidt[c]

This paper describes a clustering method on three-way arrays making use of an exploratory visualization approach. The aim of this study is to cluster samples in the object mode of a three-way array, which is done using the scores (sample loadings) of a three-way factor model, for example, a Tucker3 or a PARAFAC model. Further, tools are developed to explore and identify reasons for particular clusters by visually mining the data using the clustering results as guidance. We introduce a three-way clustering tool and demonstrate our results on a metabolite profiling dataset. We explore how high performance liquid chromatography (HPLC) measurements of commercial extracts of St. John's wort (natural remedies for the treatment of mild to moderate depression) differ and which chemical compounds account for those differences. Using common distance measures, for example, Euclidean or Mahalanobis, on the scores of a three-way model, we verify that we can capture the underlying clustering structure in the data. Beside this, by making use of the visualization approach, we are able to identify the variables playing a significant role in the extracted cluster structure. The suggested approach generalizes straightforwardly to higher-order data and also to two-way data. Copyright © 2007 John Wiley & Sons, Ltd.

Keywords: data mining; clustering; multiway models; higher-order data; visualization

1. INTRODUCTION

Clustering is a well-studied problem, where the goal is to divide data into groups of similar objects using an unsupervised learning method. Many types of clustering techniques exist, for example, partitional, hierarchical, density-based or grid-based clustering, as well as various clustering algorithms for these clustering strategies. Most clustering algorithms, however, are based on the assumption that the dataset is a two-way array. Recently, a few clustering schemes have been proposed for cluster analysis in three-way datasets. For instance, TriCluster [1] combines subspace clustering with graph-based approaches to capture coherent clusters in three-way arrays. Similarly, multi-way distributional clustering (MDC) [2] relies on subspace clustering and introduces an extension of two-way clustering to multiway arrays. The underlying principle in these techniques is to cluster all modes of the data array simultaneously.

In this paper, we focus on obtaining the clusters in one of the modes based on all the chosen variables. This problem has been addressed in the literature in a couple of studies. One of the proposed methods [3] relies on decomposing a data matrix as a membership matrix and a centroid matrix using an Alternating Least Squares-based algorithm and generalizes this scheme to multiway arrays. Another study [4] proposes to form a set of dendrograms by applying cluster analysis on each individual slice of a three-way array independently and discusses how to come up with the best consensus of these dendrograms in order to identify the clusters in a particular mode of a three-way data.

Compared to these approaches, our clustering method relies on three-way factor models, for example, a PARAFAC [5] or a Tucker3 [6] model, and applies clustering on the reduced subspace rather than the raw data. Thus, this approach is particularly appropriate in situations where these models provide a meaningful description of the data. When this is the case, it is indirectly implied that these models provide a parsimonious compression of the data and, hence, if clusters among samples are sought, using the scores of the PARAFAC or Tucker3 model is suitable.

Clustering decisions on the results of three-way factor models are often based on the visual interpretation of scatter plots, where objects in the mode of interest are projected onto the space spanned by components in the component matrix corresponding to that mode. The main drawback of such an approach is that scatter plots can consider only two or three components at a time. However, for models with more than two or three factors, clustering decisions relying on all factors could be more accurate than those based on scatter plots of pair-wise factors. Spectral clustering [7] can handle this problem by applying clustering methods, for example, k-means, on all factors extracted from the mode of interest. In a recent study [8], three-way factor models have been followed by fuzzy c-means clustering on the reduced subspace in order to find the groups of users talking to each other in a chat room community. Users are clustered into groups based on all the factors extracted from the user mode of a three-way array with modes: users × keywords × time samples. This approach has been shown to capture user groups successfully. Nevertheless, apart from user groups, the clustering method has

* Department of Computer Science, Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY 12180, USA.
E-mail: acare@cs.rpi.edu

[a] E. Acar
Department of Computer Science, Rensselaer Polytechnic Institute, Troy, New York, USA

[b] R. Bro
Department of Food Science, Faculty of Life Sciences, Copenhagen University, Copenhagen, Denmark

[c] B. Schmidt
Department of Medicinal Chemistry, Faculty of Pharmaceutical Sciences, Copenhagen University, Copenhagen, Denmark

not revealed any insight about the clusters, for example, which users are clustered together based on what type of activity.

1.1. Our contributions

The main goal of this work is to capture the hidden group structure among the objects in one mode of a three-way array and explore what variables in other modes account for the cluster structure. This aspect of identifying the underlying causes is often the primary goal of clustering but has traditionally received less attention. This paper has several contributions:

- We apply a three-way factor model, for example, a PARAFAC or Tucker3 model, to model a three-way array and apply hierarchical clustering techniques based on standard similarity measures on the score matrix. Rather than scatter plots, we exploit dendrograms to represent the cluster structure captured in the scores.
- We introduce a visualization tool for exploratory analysis of clusters. We enable graphical display of differences and similarities in the variable modes among clusters by the use of this visualization tool.
- We apply the proposed clustering scheme on a metabolite profiling dataset. We verify that clusters can be meaningfully extracted and the variables playing a significant role in the extracted cluster structure are successfully identified.

We briefly explain three-way arrays and the most common multiway models in the Section 'Background'. In the Section 'Theory', we discuss the construction of a distance matrix and hierarchical clustering using a score matrix extracted by a three-way model. We also describe how the results of the clustering can be visualized in order to understand the background for a specific clustering in this section. In the Section 'Materials and Methods', we introduce a metabolite profiling dataset arranged as a three-way array. The cluster analysis on this dataset is demonstrated and clustering results are discussed along with the exploratory power of the visualization tool in the Section 'Visual Cluster Analysis'.

2. BACKGROUND

2.1. Notations and terminology

Multiway arrays, also referred to as tensors, are higher-order generalizations of vectors and matrices. Higher-order arrays are represented as $\underline{X} \in \mathbb{R}^{I_1 \times I_2 \cdots \times I_N}$, where the order of $\underline{X}$ is $N(N > 2)$ while a vector and a matrix are arrays of order 1 and 2, respectively. In the higher-order terminology, each dimension of a multiway array is called a mode (way) and the number of variables in each mode is used to indicate the dimensionality of a mode. For instance, $\underline{X} \in \mathbb{R}^{I_1 \times I_2 \cdots \times I_N}$ is a multiway array with N modes (called an N-way array or an Nth-order tensor) with $I_1, I_2, \ldots I_N$ dimensions in the first, second,. . .Nth mode, respectively.

A third or higher-order array can also be rearranged as a matrix. This procedure is called matricization (or unfolding, flattening). The mode-n matricization of a higher-order dataset, for example, $\underline{X}$, denoted by $X_{(n)}$, unfolds the data in the nth mode. There are multiple definitions of matricization in the literature and we refer interested users to see the References [13,14].

We denote higher-order arrays using underlined capital letters, for example, $\underline{X}$, following a similar notation as in the Reference [13]. Matrices and vectors are represented by uppercase, for example, X, and lowercase letters, for example, x, respectively. Scalars are denoted by lowercase or uppercase italic letters, for example, x and X. Vector, matrix and tensor entries are represented by lowercase letters with subscripts, for example, x_i, x_{ij} or x_{ijk}. We mention explicitly if other notations are used throughout the paper.

2.2. Multiway models

Data in numerous disciplines including chemometrics, psychometrics, signal processing, neuroscience, data mining and computer vision is arranged as multiway datasets and analyzed using multiway models. This section describes the two most common multiway models, that is, PARAFAC and Tucker3, which we also use in our analyses.

PARAFAC [5] is an extension of bilinear factor models to multiway data. A PARAFAC model can be represented as the decomposition of a tensor as a linear combination of rank-one tensors. Let $\underline{X} \in \mathbb{R}^{I \times J \times K}$ be a three-way array. Then an R-component PARAFAC model on $\underline{X}$ can be expressed as in Eq. 1.

$$\underline{X} = \sum_{r=1}^{R} a_r \circ b_r \circ c_r + \underline{E} \tag{1}$$

where a_r, b_r and c_r are the rth columns of the component matrices $A \in \mathbb{R}^{I \times R}$, $B \in \mathbb{R}^{J \times R}$ and $C \in \mathbb{R}^{K \times R}$ in the first, second and third mode, respectively. The array $\underline{E} \in \mathbb{R}^{I \times J \times K}$ contains the residuals corresponding to each data entry. The symbol $\bigcirc$ denotes the vector outer product. Vector outer product is defined as follows. Let x, y and z be column vectors of size $I \times 1$, $J \times 1$ and $K \times 1$ and $\underline{M}$ be a tensor of size $I \times J \times K$, then $\underline{M} = x \circ y \circ z$ if and only if $m_{ijk} = x_i y_j z_k$. It is also possible to express a PARAFAC model in matrix notation as in Eq. 2.

$$X_{(1)} = A(C \odot B)^T + E_{(1)} \tag{2}$$

where $X_{(1)}$ and $E_{(1)}$ are matricized arrays in the first mode and the symbol $\odot$ denotes the Khatri-Rao product [9]. The PARAFAC model is somewhat unusual in the sense that every dataset cannot be meaningfully modeled by PARAFAC [9]. The data has to approximately possess low-rank trilinear variation in order for PARAFAC to be able to model the data. This is different from Principal Component Analysis (PCA) [19,20] that can always fit a desired percentage of the variation by including sufficiently many components.

A more flexible model compared to PARAFAC is a Tucker3 [6] model. Similar to PARAFAC, Tucker3 is also a generalization of bilinear factor models to higher-order datasets. However, unlike a PARAFAC model, a Tucker3 model can extract different number of components from each mode and the factors from different modes can interact with each other. The interactions between factors are quantified by the entries in a core array. A (P, Q, R)-component Tucker3 model on $\underline{X} \in \mathbb{R}^{I \times J \times K}$ can be represented in matrix notation as in Eq. 3.

$$X_{(1)} = AG_{(1)}(C \otimes B)^T + E_{(1)} \tag{3}$$

where $A \in \mathbb{R}^{I \times P}$, $B \in \mathbb{R}^{J \times Q}$ and $C \in \mathbb{R}^{K \times R}$ are the component matrices in the first, second and third mode, respectively and $G_{(1)}$ represents the core array, $\underline{G} \in \mathbb{R}^{P \times Q \times R}$, matricized in the first mode. The symbol $\otimes$ denotes the Kronecker product [9]. Tucker3, as PCA, can always represent a dataset to an arbitrary extent by including sufficiently many components, so Tucker3 does not suffer from the numerical and mathematical problems that sometimes make applications of the PARAFAC model difficult.

3. THEORY

Three-way models decompose an array into component matrices corresponding to each mode, a possible core array and an error term containing the residuals. Our goal is to cluster objects in the sample mode of a three-way array. Therefore, we make use of the score matrix to find out how objects cluster; in other words, instead of trying to cluster according to distances in the original metric, we aim at finding clusters in terms of distances of the underlying latent variables. For example, rather than trying to cluster according to similarity of elution profiles, we here aim at using the similarity of scores (which are representative of relative concentrations), which is often a more interesting metric in relation to the application.

3.1. Distance matrix and hierarchical clustering

Let $\underline{X} \in \mathbb{R}^{I \times J \times K}$ be a three-way array and suppose that the goal is to explore the grouping in the first mode of the dataset without loss of generality. Using a three-way factor model, $\underline{X}$ can be modeled as $X_{(1)} = AZ^T + E_{(1)}$, where $Z^T = (C \odot B)^T$ in the case of a PARAFAC model and $Z^T = G_{(1)}(C \otimes B)^T$ for a Tucker3 model [9]. Each row of the score matrix A corresponds to an object/sample and each column of A contains either oblique or orthogonal factors depending on the model. We compute pair-wise distances between rows using standard distance measures, for example, Euclidean or Mahalanobis, and construct a symmetric distance/dissimilarity matrix, whose (*i*, *j*)th entry contains the distance between the *i*th and *j*th rows of A. Note that scaling of the columns of PARAFAC models may be useful in certain applications as the columns can have quite different scales especially when many components are computed. In Tucker3, this is usually not an issue as the vectors are conventionally normalized as an intrinsic part of the algorithm.

We employ agglomerative hierarchical clustering [10] on the constructed distance matrix. This clustering scheme starts by assigning each object to a separate cluster and merges clusters based on the selected linkage condition iteratively, for example, single (nearest neighbor), average and complete (farthest neighbor). The type of linkage condition determines the way clusters are merged. While complete linkage clustering produces very tight clusters of similar objects, single linkage clustering may form long chains of loose clusters. At intermediate stages of the clustering method, there exists a series of partitions of the object set. Eventually, all objects are assigned to a single cluster at the last step of the clustering scheme and a complete hierarchical tree, that is, a dendrogram is formed. Although hierarchical clustering algorithms are, in general, computationally expensive, our motivation behind using hierarchical clustering is to clearly display the degrees of similarity and distant relationships among clusters using dendrograms.

3.2. Visualization

Various clustering algorithms organize objects into groups and assign them certain cluster memberships. However, it is still difficult to discover cluster structures and assimilate why some objects are clustered together and others are grouped into a different cluster. For instance, a chemist working on the cluster analysis of chemical profiling data of drug samples would want to see how some compounds differ across samples in some clusters in order to have a better understanding of the underlying differences or similarities between samples. To facilitate that kind of study, we illustrate a visualization approach, which enables exploratory analysis of clusters.

Mostly, the result of clustering is presented, for example, as a dendrogram, which then forms the final output of the clustering. Such a result provides few means for exploring the deeper structure of the data and we show how simple visualization tools can help making clustering a truly exploratory tool.

- *Sample/class labels and colors*: In most situations, prior information is given on the samples, for example, the factories where the sugar samples are from in the dataset given in Figure 1. It is helpful to use the prior information in order to verify what is reflected in the clustering. This can be done through labeling but even more cognitively appealing by using simple coloring of the clustering results. If a certain cluster contains only samples from one class, for example, from the same factory, we produce the clustering results with a unique color. Here 'class' indicates the group of samples based on some prior information while 'cluster' is the grouping captured by the methodology described in the previous section. We also allow clusters to be colored by their main class even when they have a certain number of samples from different classes to get an idea about approximate clusters in the data. Unless samples from different classes exceed the specified number, all samples in that cluster are colored by the color of the class of samples that form the majority of the cluster. If samples belong to several different types of classes, or in other words if we have different types of prior information, any of these classes can be used in coloring samples and clusters. In Figure 1, an example is provided to demonstrate how using different thresholds can help identify when clusters coincide with prior classes.

  In addition to coloring, we also enable different ways of labeling on the dendrogram. One way is to use labels such as sample ids, for example, Figure 1A or B, while another way is to label the samples based on their class or any given prior information, for example, Figure 8, where extracts that are in the same preparation are labeled with the same preparation id. This feature enables the comparison of clustering results with any given prior information.
- *Plots of average and individual profiles in one variable mode*: When validating clustering results, for example, determining the right number of clusters or understanding why a certain clustering occurs or does not occur, it is important to be able to relate the clusters to the raw data. We show several plots that are helpful in this respect when implemented in a simple point-and-click fashion. For example, when two clusters are chosen, a plot can be made of the average profile of each of the two clusters (Figure 2A) as well as the individual profiles (Figure 2B) in one variable mode, for example, a plot of average elution profile.
- *Changing between variable modes*: Furthermore, by marking individual points, for example, a certain emission wavelength, on the plots given in Figure 2, a plot of the other variable mode can be made, for example, the corresponding excitation spectra, as shown in Figure 3.

With such tools, it is possible to perform meaningful data mining and learn new facets of the data. We illustrate the main

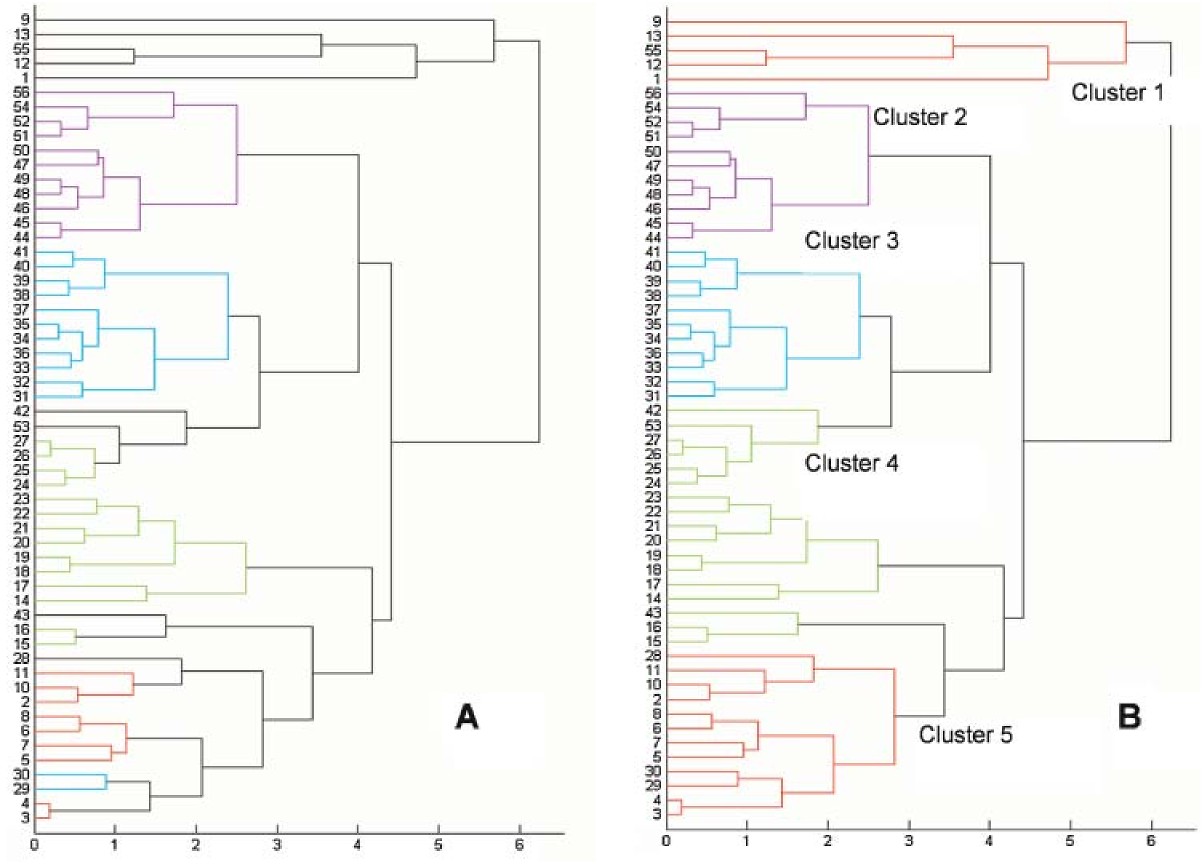


Figure 1. A dendrogram constructed by hierarchical clustering on the scores of a Tucker3 model of a three way fluorescence dataset of 56 sugar samples. Different colors are used to represent sugar samples from different factories. To the left, the immediate output is given, while to the right, the same clustering is shown after allowing coloring a cluster even with up to three samples belonging to a different factory. This figure is available in colour online at www.interscience.wiley.com/journal/cem

features of the visualization tool in Figure 4 and in the Section 'Visual Cluster Analysis', we further explain and exemplify the use of its features.

It is important to realize that we suggest using plots of the raw data here rather than using the reconstructed data. However, we should emphasize that the actual clustering is based on the model of the data, not the actual data. Hence, a direct link between the clustering results and the raw data is only applicable for a well-fitting model. An over or under-fitted model could lead to clustering results that would be difficult to recognize in the data. This should be avoided by first validating the actual multi-way model before proceeding to the clustering. As an indirect

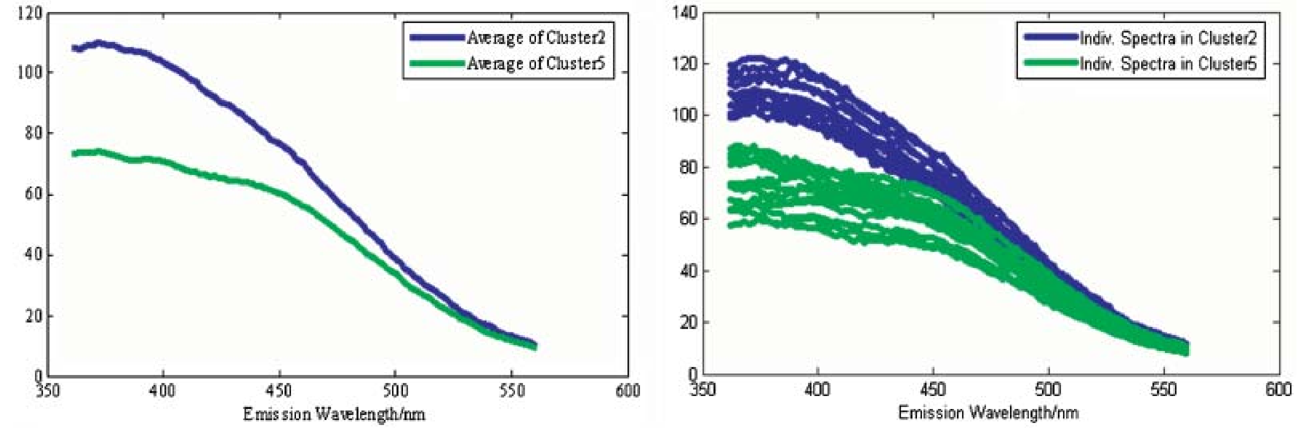


Figure 2. (A) The average emission spectra are shown for the fluorescence data given in Figure 1 by selecting Cluster 2 and 5, respectively. (B) From the average profiles shown, it is easy to see the main difference and this difference can be verified by subsequently showing all individual spectra rather than the averages. This figure is available in colour online at www.interscience.wiley.com/journal/cem

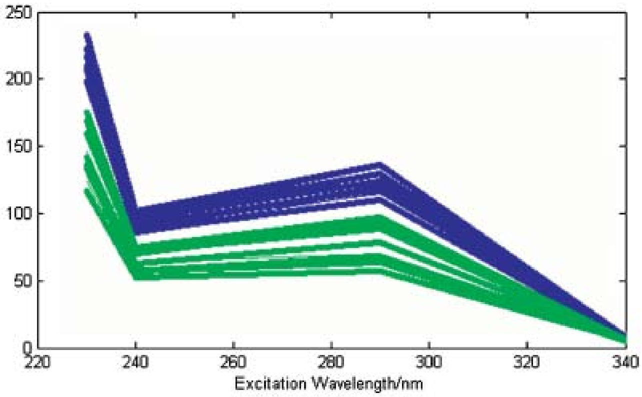


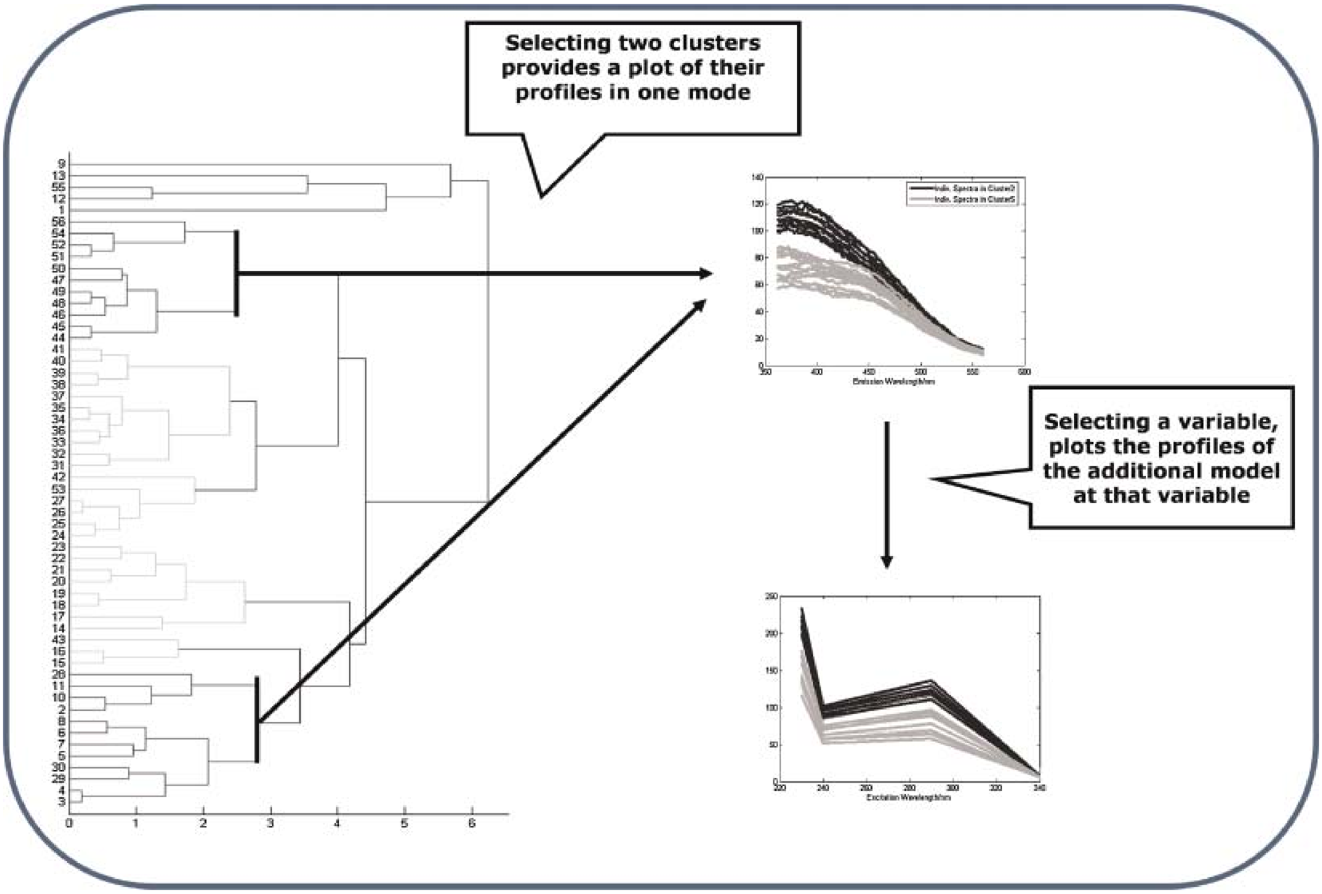
Figure 3. Individual excitation spectra corresponding to a certain emission wavelength for the samples in Cluster 2 and Cluster 5. This figure is available in colour online at www.interscience.wiley.com/journal/cem

check of the validity, it is also possible to build in functionality so that the user can choose to plot the reconstructed data if preferred, thus, making it possible, for example, to diagnose why a certain clustering result cannot be verified in the actual data. However, as long as the model is valid, such a feature would not be needed.

4. MATERIALS AND METHODS

We apply the proposed clustering scheme on a metabolite profiling dataset containing HPLC measurements of commercial extracts of St. John's wort. HPLC-PDA (HPLC with photodiode-array detection) profiles were obtained from 24 different samples of St. John's wort from several continents (Africa, Asia, Europe and North America). All samples originated from commercial suppliers. For one of the brands, two different batches were obtained. The number of samples from each continent varies between 2 and 12. The HPLC-PDA profiles in replicates of three or four for each sample were obtained using a system consisting of an A eluent (acetonitrile:water 5:95 + 0.1% HCOOH) and a B eluent (acetonitrile:water 95:5 + 0.1% HCOOH) with a linear gradient. The chromatography was monitored between 190 nm and 620 nm. Two regions of the chromatographic data were chosen for analysis and reduced to steps of 3 nm in the UV-mode (260–550 nm) and steps of 1.32 seconds in the retention time mode. Each region was separately warped using correlation optimized warping (COW), correcting for shifts in retention time mode according to criterion working on whole spectrum from 260 nm to 520 nm to correct for unwanted peak shifting [11].

The clustering and visualization tools running under MATLAB have been implemented by making use of the algorithms of



Selecting two clusters provides a plot of their profiles in one mode

Selecting a variable, plots the profiles of the additional model at that variable

Figure 4. The screenshot illustrating the main features of the visualization tool. When two clusters are selected from the dendrogram, the plot of their profiles in one variable mode appears on the right hand-side. Both average profiles of clusters and individual profiles of samples in each cluster can be plotted. If a certain variable is selected on this plot, then the tool plots the profiles of the clusters in the other mode at that particular variable. This figure is available in colour online at www.interscience.wiley.com/journal/cem

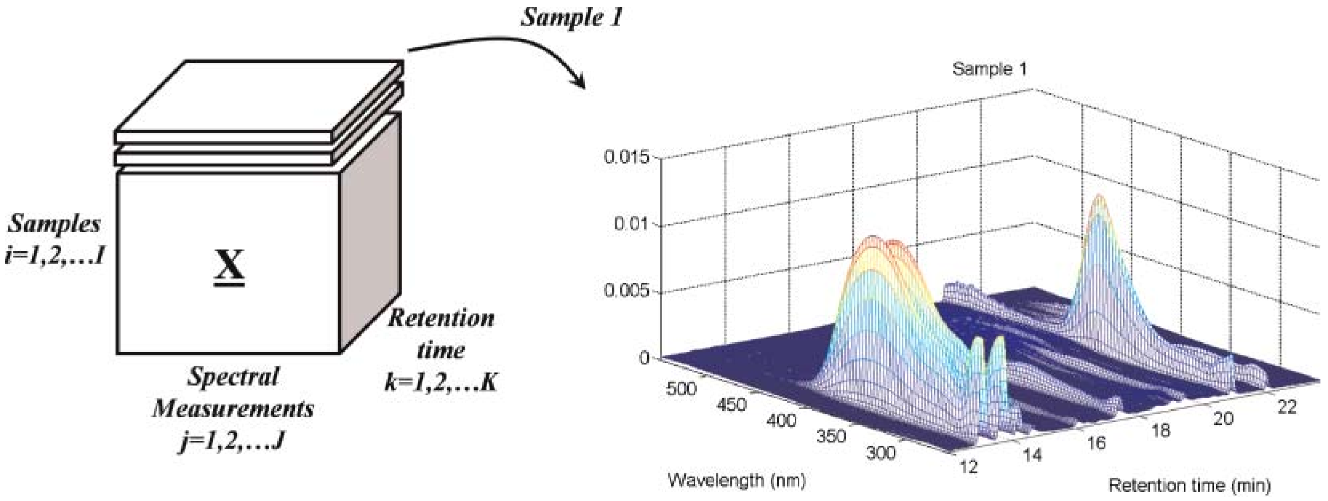


Figure 5. The HPLC measurements of commercial extracts arranged as a three-way array with modes: samples, spectral measurements and retention time. Each horizontal slice of a three-way array corresponds to a sample and we illustrate as an example what Sample 1 looks like. This figure is available in colour online at www.interscience.wiley.com/journal/cem

three-way factor models, that is, PARAFAC and Tucker3, in PLS_Toolbox (Eigenvector Research Inc.). The software for cluster analysis and visualization can be downloaded from www.models.life.ku.dk/source/ (Sept 2007).

5. VISUAL CLUSTER ANALYSIS

In this section, we perform cluster analysis on a metabolite profiling dataset arranged as a three-way array. First, we fit a Tucker3 model to the HPLC measurements of commercial extracts of St. John's wort arranged as a three-way array as in Figure 5. We then use the scores to cluster the samples.

5.1. Parameter selection

In order to determine a suitable complexity of the Tucker model, models of increasing complexity were fitted to the data (using the same number of components in each mode). The resulting fit values are shown in Figure 6 indicating that no more than a (5, 5, 5) model is needed to describe the majority of the data.

The (5, 5, 5) Tucker model has been further investigated by looking at the core elements of the most important factor combinations of a model rotated to maximum simplicity. These core elements are indicative of the important variation in the data. The 10 most important combinations are shown in Table I. Table I shows that components up to factor 4, 3 and 5, respectively, are participating in the important factor combinations. Hence a (4, 3, 5) model is fitted and found to fit 98.2% of the data as opposed to 98.6% for the (5, 5, 5) model. The simpler model is capable of explaining most of the variation in the (5, 5, 5) model. Furthermore, visual interpretation of the model parameters clearly indicates that the model reflects the systematic variation in the data. Therefore, (4, 3, 5) model has been chosen for further analysis. There are several other techniques used in the literature for determining the component number in PARAFAC models, that is, core consistency diagnostic [15] and Tucker3 models, for example, Difference in Fit (DIFFIT) [16,17] and a more recent approach based on convex hulls [18].

5.2. Modeling and interpretations

After selecting the number of components, we model the data illustrated in Figure 5 using a Tucker3 model and extract four score vectors. We initially demonstrate the cluster structures captured by scatter plots using the first two scores. Scatter plots are used to plot two or three loadings against each other but they may not be sufficient by themselves to unravel the underlying cluster structure. For instance, although clusters observed in the scatter plot in Figure 7 are promising and the tendency of replicate extracts being close to each other is obvious, we should consider many scatter plots together to find the clusters of replicates. Even then, we may not observe all existing clusters clearly.

On the other hand, graphical representation using dendrograms is accurate and simple if the underlying model (Tucker3 or PARAFAC) is appropriate. Moreover, dendrograms demonstrate

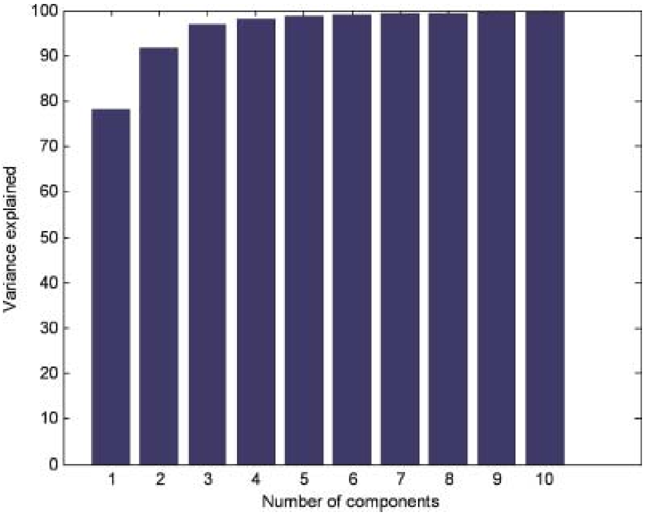


Figure 6. Model complexity versus explained variance. The explained variance with a Tucker3 model with (*i*, *i*, *i*) components is given on the y-axis corresponding to $x = i$. This figure is available in colour online at www.interscience.wiley.com/journal/cem

Table I. The analysis of 5 × 5 × 5 core array

Component number	Value	Squared value	Variance (%)	Cumulative variance (%)
[1, 1, 1]	−2.34	5.48	79.26	79.26
[2, 1, 2]	−0.81	0.65	9.40	88.66
[1, 2, 2]	0.43	0.19	2.73	91.40
[1, 2, 3]	−0.42	0.18	2.58	93.97
[1, 3, 3]	−0.37	0.14	1.98	95.96
[1, 3, 2]	−0.32	0.10	1.50	97.45
[3, 1, 5]	−0.13	0.02	0.25	97.70
[2, 3, 2]	−0.12	0.01	0.22	97.92
[1, 3, 4]	0.12	0.01	0.21	98.13
[4, 1, 5]	−0.12	0.01	0.20	98.33

both clusters and the degree of similarity between samples/clusters. In order to construct the dendrogram for the HPLC measurements arranged as a three-way array, we first construct a symmetric distance matrix based on the Euclidean distance to compute the pair-wise distances between samples using all four scores. Complete linkage clustering is then employed for merging samples into clusters and forming the dendrogram in Figure 8.

Once we have the dendrogram constructed, we can explore common and different structures among clusters using the visualization tools. The tools enable us to mark two lines, each representing a cluster, on the hierarchical tree and plot their profiles in the elution mode or spectral mode. For instance, we select the clusters corresponding to preparations 2 and 3 and observe the differences and similarities of the samples in these clusters (Figure 9). In Figure 9A, it is observed that there are significant differences in the average elution profiles of samples in preparations 2 and 3. Especially, the peaks around retention time 13.1, 13.7, 14.3, 20.9 and 22 minutes (corresponding to rutin, hyperoside, isoquercetin, quercetin and biapigenin, respectively) can easily be used to separate the samples in the two preparations. This is verified by adding individual profiles of all samples in both preparations in Figure 9B. We observe that

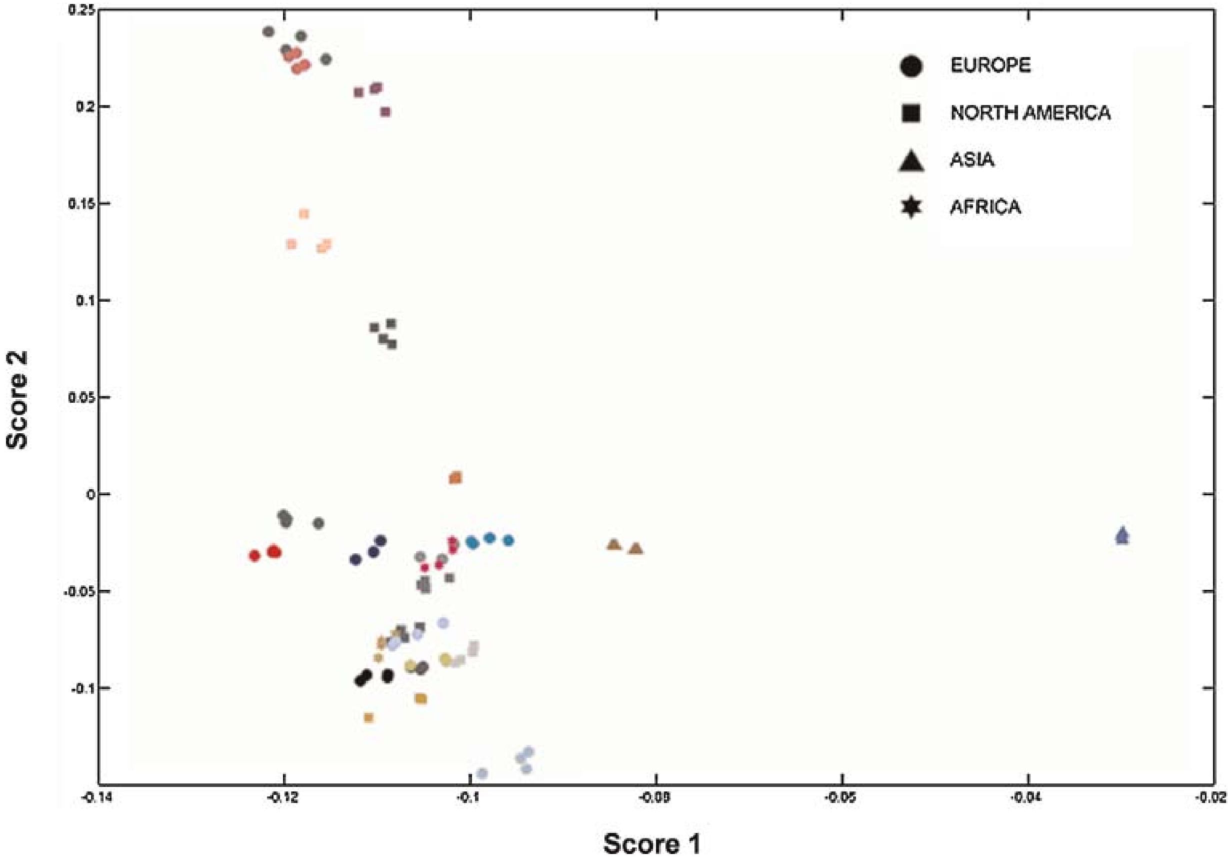


Figure 7. Scatter plot using score 1 versus score 2. Dataset contains 24 different extracts and three to four replicates of each extract are included. The points are colored by replicates and represented by circles, squares, triangles or stars depending on their origin. This figure is available in colour online at www.interscience.wiley.com/journal/cem

individual samples are quite similar within each preparation and well represented by their average. The compounds observed to be responsible for the differences between preparations 2 and 3 are in accordance with earlier observations from PCA analysis of NMR spectra of the same preparations [12]. In that study the separation of preparation 2 from preparation 3 amongst other preparations was caused by a higher content of mainly hyperoside (retention time 13.7) and to a lesser extend rutin (retention time 13.1). The separation of preparation 3 from preparation 2 amongst other preparations was caused mainly by an increased content of quercetin (retention time 20.9). The visualization tool allows us to discover other compounds also contributing to the differences between the observed preparations.

Preparations 2 and 3 behave quite differently around retention time 20.9 minutes, and we inquire their spectral differences by selecting a point close to this elution time and examine the spectra of each sample in these two preparations at that particular time point. Figure 10 suggests that there is remarkable variation in the spectra of samples in different clusters, whereas within-cluster variation is negligible at elution time 20.9 minutes.

An interesting set of samples is the one containing the samples from preparations 5, 7 and 8. Preparations 7 and 8 are different batches of the same brand, and preparation 5 was sold under a different brand name but with the same label insert as the other two. We observe that preparations 5 and 8 cluster together in the dendrogram in Figure 8. In order to perform further analysis, we select two clusters from the dendrogram: one formed by merging preparations 5 and 8 and the other containing only samples from preparation 7. Figure 11A demonstrates that average profiles of samples in all preparations across elution mode are quite similar.

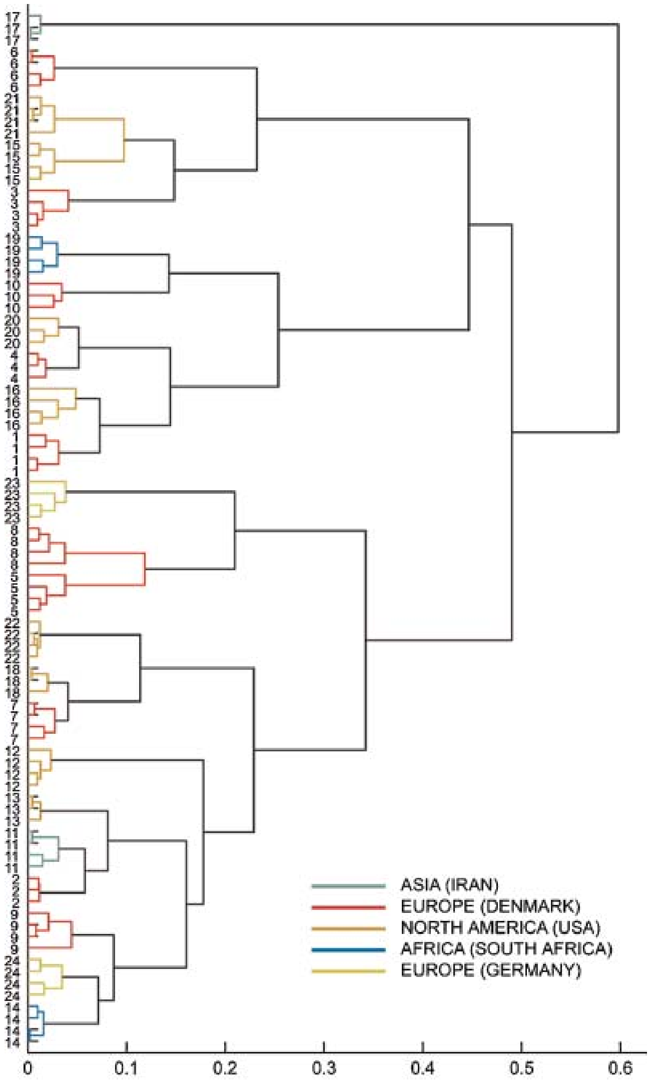


Figure 8.

Figure 8. Dendrogram representation of hierarchical clustering on the scores of a Tucker3 model on three-way chromatographic data of 89 commercial extracts of St. John's wort. Preparation numbers are shown on the y-axis and the cost of clustering samples is shown on the x-axis. Extracts represented by the same color are preparations originating from the same country. In the dendrogram, replicate extracts are merged together at the first level successfully. This figure is available in colour online at www.interscience.wiley.com/journal/cem

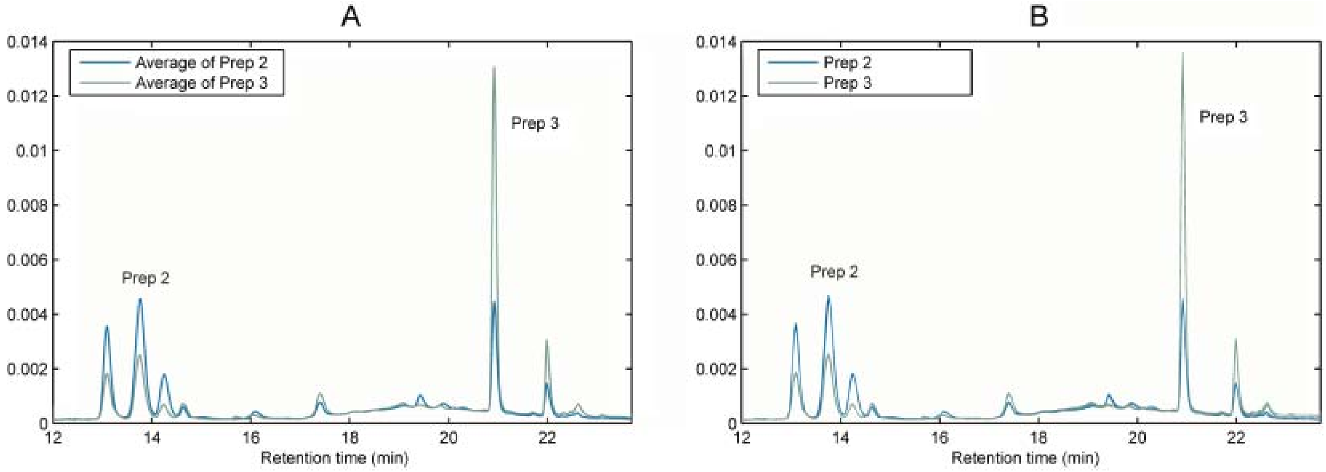


Figure 9. (A) Average sample profiles across elution mode for the clusters representing preparations 2 and 3. (B) Average sample profiles across elution mode for the clusters representing preparations 2 and 3 together with the individual profiles of samples in each cluster. This figure is available in colour online at www.interscience.wiley.com/journal/cem

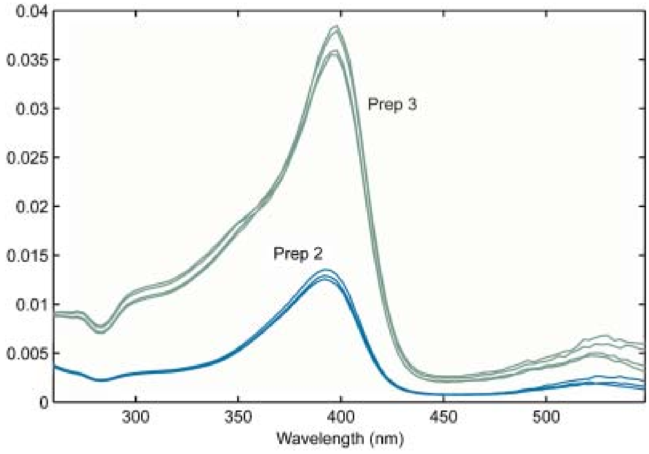


Figure 10. Sample profiles across spectral mode at elution time 20.9 minutes (corresponding to quercetin) for preparations 2 and 3. This figure is available in colour online at www.interscience.wiley.com/journal/cem

Nevertheless, we observe that there are slight differences in the average profiles of preparations 5 and 8 when compared with that of preparation 7. For instance, preparations 5 and 8 show a lower baseline around retention from 18 minutes to 21 minutes and have a higher content of the peak eluting at 20.9 minutes corresponding to quercetin (Figure 11). In an earlier study based on NMR spectra of these three preparations [12], it was shown that the preparations clustered very close to each other and they were hard to separate. Using the described hierarchical clustering techniques on the HPLC-PDA profiles of the preparations, we are now able to separate these samples and the new visualization tool shows exactly where in the profiles the preparations differ.

Extracts representing the same country are colored by the same color in the dendrogram in Figure 8. We can see that a number of preparations cluster together with other preparations from the same country, for example, preparations 15 and 21 originating from USA and preparations 5 and 8 originating from Denmark. Further analysis of these two sets of samples indicates that nearly all peaks across elution mode differ between

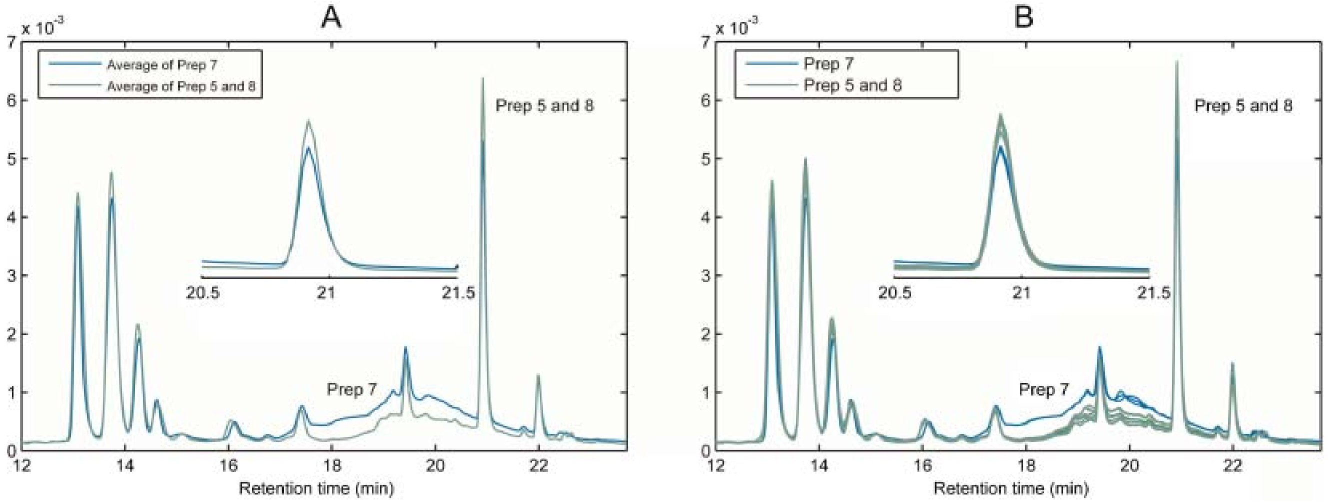


Figure 11. (A) Average profiles across elution mode for the cluster representing preparation 7 and the combined cluster representing preparations 5 and 8. (B) Average profiles across elution mode for samples in preparation 7 and preparations 5 and 8 including individual profiles of each sample. This figure is available in colour online at www.interscience.wiley.com/journal/cem

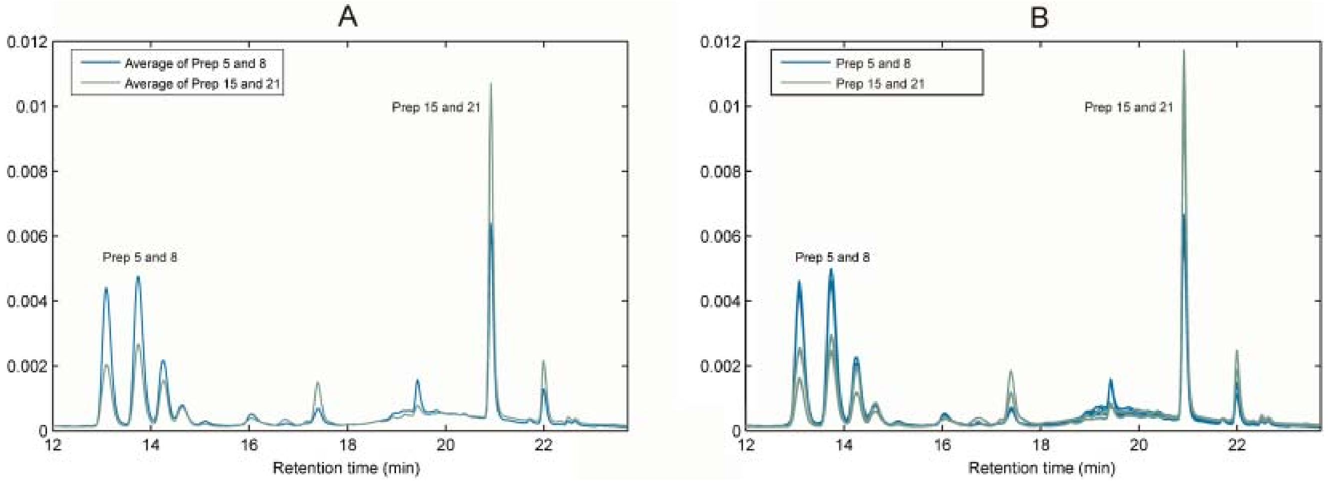


Figure 12. (A) Average profiles across elution mode for the combined clusters representing preparations 5 and 8 (Denmark) and preparations 15 and 21 (USA). (B) Average profiles across elution mode for the combined clusters of samples in preparations 5 and 8 and preparations 15 and 21 including individual profiles of each sample. This figure is available in colour online at www.interscience.wiley.com/journal/cem

preparations from the two countries represented by the chosen clusters as illustrated in Figure 12. The standardization of St. John's wort according to the European Pharmacopoeia monograph only relies on the total content of hypericins, whereas as described in the United States Pharmacopeia the standardization is based on the content of hypericins as well as hyperforin. Therefore, it is not surprising that other compounds, such as rutin (rt = 13.1), hyperoside (rt = 13.7), isoquercetin (rt = 14.3), quercetrin (rt = 17.4), quercetin-3-*O*-(2-acetyl)-β-D-galactoside (rt = 19.4), quercetin (rt = 20.9) and biapigenin (rt = 22) vary between countries, since none of the preparations are standardized to these compounds. The visualization tools quickly allow us to discover exactly which compounds correspond to the observed differences.

6. CONCLUSION

A new combined tool that allows clustering on three-way arrays and exploring clustering results has been developed and exemplified. The premise for meaningful clustering is that a suitable three-way model can be found, but this is generally possible using Tucker3 modeling. In some cases, a PARAFAC model can be used to adequately describe the data and then the interpretation may be further simplified by the simple interpretation of such models. With the clustering of scores and especially the suggested visualization tool, the interpretation of clustering results is greatly eased particularly in relation to tracing results back to the raw data and understanding what the different clustering results represent. This is of utmost importance in, for example, data mining, metabonomics and similar exploratory studies.

REFERENCES

1. Zhao L, Zaki M. TriCluster: an effective algorithm for mining coherent clusters in 3D microarray data. *Proc. ACM SIGMOD Int. Conf. Management of Data* 2005; 694–705.
2. Bekkerman R, El-Yaniv R, McCallum A. Multi-way distributional clustering via pairwise interactions. *Proc. Int. Conf. Machine Learning (ICML)* 2005; 41–48.
3. Depril D, Van Mechelen I. One-mode additive clustering of multiway data. *Proc. Int. Symp. Applied Stochastic Models and Data Analysis* 2005; 724–729.
4. Vichi M. One-mode classification of a three-way data matrix. *J. Classif.* 1999; **16**: 27–44.
5. Harshman RA. Foundations of the PARAFAC procedure: models and conditions for an Explanatory Multimodal Factor Analysis. *UCLA Work. Pap. Phonetics* 1970; **16**: 1–84.
6. Tucker LR. Some mathematical notes on three-mode factor analysis. *Psychometrika* 1966; **31**: 279–311.
7. Ng AY, Jordan MI, Weiss Y. On spectral clustering: analysis and an algorithm. *Adv. Neural Inf. Process. Syst.* 2001; **14**: 849–856.
8. Acar E, Camtepe SA, Krishnamoorthy MS, Yener B. Modeling and multiway analysis of chatroom tensors. *Lect Notes Comput Sci* 2005; **3495**: 256–268.
9. Smilde A, Bro R, Geladi P. Multi-way Analysis. Applications in the Chemical Sciences. Wiley: England, 2004.
10. Johnson SC. Hierarchical clustering schemes. *Psychometrika* 1967; **32**: 241–254.
11. Tomasi G, van den Berg F, Andersson C. Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *J. Chemometrics* 2004; **18**: 231–241.
12. Rasmussen B, Cloarec O, Tang H, Stærk D, Jaroszewski JW. Multivariate analysis of integrated and full-resolution 1H-NMR spectral data from complex pharmaceutical preparations: St. John's Wort. *Planta Med.* 2006; **72**: 556–563.
13. Kiers HAL. Towards a standardized notation and terminology in multiway analysis. *J. Chemometrics* 2000; **14**(3): 105–122.
14. De Lathauwer L, De Moor B, Vandewalle J. A multilinear singular value decomposition. *SIAM J. Matrix Anal. Appl.* 2000; **21**(4): 1253–1278.
15. Bro R, Kiers HAL. A new efficient method for determining the number of components in PARAFAC models. *J. Chemometrics* 2003; **17**(5): 274–286.
16. Timmerman ME, Kiers HAL. Three mode principal component analysis: choosing the number of components and sensitivity to local optima. *Br. J. Math. Stat. Psychol.* 2000; **53**(1): 1–16.
17. Kiers HAL, Der Kinderen A. A fast method for choosing the number of components in Tucker3 analysis. *Br. J. Math. Stat. Psychol.* 2003; **56**(1): 119–125.
18. Ceulemans E, Kiers HAL. Selecting among three-mode principal component models of different types and complexities: a numerical convex-hull based method. *Br. J. Math. Stat. Psychol.* 2006; **59**(1): 133–150.
19. Jackson JE. Principal components and factor analysis: part I—principal components. *J. Qual. Technol.* 1980; **12**: 201–213.
20. Jackson JE. Principal components and factor analysis: part II—additional topics related to principal components. *J. Qual. Technol.* 1981; **13**: 46–58.