Talanta 83 (2011) 1088–1097

Contents lists available at ScienceDirect

Talanta

journal homepage: www.elsevier.com/locate/talanta

Methods for the exploratory analysis of two-dimensional chromatographic signals

M. Daszykowski*, B. Walczak

Department of Chemometrics, Institute of Chemistry, The University of Silesia, 9 Szkolna Street, 40-006 Katowice, Poland

ARTICLE INFO

Article history:
Available online 15 September 2010

Keywords:
Rv coefficient
Comparing data tables
Chromatographic fingerprints
Alignment
HPLC-DAD
PARAFAC
PARAFAC2

ABSTRACT

In this article several approaches for the exploratory analysis of two-dimensional chromatographic signals (fingerprints) are presented. Their usefulness is illustrated on experimental chromatographic data obtained from high performance liquid chromatography using the photodiode-array detector (HPLC-DAD). Among the methods discussed are principal component analysis (PCA), hierarchical clustering methods and several N-way techniques such as PARAFAC, PARAFAC2 and Tucker3. In addition to the N-way methods, other approaches that allow for comparing samples represented by two-dimensional fingerprints are also presented (the Rv coefficient, the STATIS approach and 'fuzzy' variants of the similarity matrix). Exploratory analysis of the HPLC-DAD data with peak shifts is also discussed.

1. Introduction

In recent years, much attention has been paid to the characterisation of samples using their chemical fingerprints [1]. Such a strategy assumes that instrumental signals capture unique information about the chemical composition of samples and can later be used for the purpose of their comparative analysis. This approach is also referred to as non-targeted analysis in contrast to targeted analysis where samples are described by a few carefully selected and quantified chemical components. However, when complex samples are analysed, e.g. herbal extracts, environmental and biological samples, the targeted approach has a rather limited application because it requires chemical standards and some preliminary knowledge about the systems being studied.

Non-targeted analysis has the advantage of presenting a comprehensive view of the composition of the sample represented by a certain type of instrumental signals, and requires no chemical standards in order to compare samples. Therefore, the chemical fingerprints enable exploration of the system being analysed at the chemical level. Once the chemical fingerprints are appropriately registered, an exploratory analysis follows in order to display the differences among the samples and to identify certain signal components contributing most to these differences.

There are several important aspects that have to be addressed appropriately in order to obtain valuable information from chromatographic fingerprints. The first is related to maximising the information content of the chromatographic fingerprints and their quality by appropriate sample preparation and identifying the optimal chromatographic conditions.

The second aspect is related to data exploration and successful extraction of the chemically relevant information. Different chemometric techniques are available to compare and explore a collection of chromatographic signals. In this arsenal there are also methods well suited for correcting certain deficiencies in instrumental signals including noise and background levels as well as for synchronizing time axes in chromatograms (the alignment techniques [2]). Regardless of the type of chromatographic data at hand, their exploration is considered to be the first step in the discovery of knowledge.

Projection methods [3] such as e.g. principal components analysis (PCA) and clustering approaches (including hierarchical clustering techniques) are widely used to explore chromatographic data containing one-dimensional HPLC chromatograms. In fact, due to the large dimensionality of such data, projection methods are highly valued because they help to summarise data structures using only a few latent variables. As illustrated in [4] it is also possible to adopt the same strategy for an exploratory analysis of chromatographic data constituted by data tables obtained when a multi-channel detector is used (e.g. the photo-diode array detector, DAD). They are constituted by measurements of a certain property

* Corresponding author.
E-mail address: mdaszyk@us.edu.pl (M. Daszykowski).

0039-9140/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2010.08.032

registered for two different features, for instance at a given retention time and wavelength, and are referred to as two-dimensional signals. Their collection is viewed as a three-way data array, **X**, with the dimensions (*the number of spectral channels* × *the number of sampling points on retention time axis* × *the number of samples*) representing three data modes.

The major goal of this article is to present different chemometric approaches for the exploratory analysis of chromatographic data represented by a set of two-dimensional chromatographic signals. The article will be limited to the exploratory analysis of HPLC-DAD signals only, but the approach discussed in this article can also be used for exploring other types of two-dimensional signals as well. Our interest in the HPLC-DAD method arises mostly from its popularity as a fingerprinting technique and its wide availability.

In the theory section we will review some basic chemometric methods for exploratory data analysis. Later, their applications to two-dimensional HPLC-DAD signals will be presented. We will also focus on dealing with the issue of peak shifts in two-dimensional HPLC-DAD signals and will indicate some possible ways to process them in the context of their exploratory analysis.

2. Theory

Projection and clustering techniques, individually or in a combination, are usually used for the exploratory analysis of chromatographic signals. In the literature many examples of exploratory analysis on a set of one-dimensional signals can be found. Applications of these methods to two-dimensional chromatographic signals are reported less frequently. Depending on the type of chromatographic signal and the presence or absence of peak shifts, different exploratory approaches are available (see Table 1).

Regardless of the type of chromatographic signals (one- or two-dimensional signals) when peak shifts are observed they require special pre-treatment (synchronisation of their time axes [2]) or special approaches to data exploration that are unaffected by peak shifts. Such exploratory approaches involve the construction of a similarity matrix that is insensitive to peak shifts (as described in the following section) used as input to projection and/or hierarchical clustering techniques. It is assumed that peak shifts in two-dimensional chromatographic signals are only observed along time dimension.

2.1. Scoring similarities among samples represented by two-dimensional signals

Similarities between two samples i and j, characterised by two-dimensional chromatographic signals, **X** and **Y** respectively, can be presented as a similarity matrix, with elements s_{ij}:

$$s_{ij} = \mathrm{vec}(\mathbf{X})^{\mathrm{T}} \cdot \mathrm{vec}(\mathbf{Y}) \quad (1)$$

where 'vec(·)' is the column-wise unfolding of **X** and **Y**.

The **S** matrix is a square, positive and semi-definite matrix with diagonal elements larger than or equal to zero. The entries of the similarity matrix can be additionally normalised to remove scaling effects by dividing each element s_{ij} by the square root of the product of $s_{ii} \cdot s_{jj}$.

When peak shifts in signals along the time dimension are present, the similarities have to be scored differently. To compare samples characterised by two-dimensional chromatographic signals, **X** and **Y**, the so-called Rv coefficient can be used as follows [5]:

$$\mathrm{Rv}_{ij} = \frac{\mathrm{trace}(\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{Y}^{\mathrm{T}})}{\sqrt{\mathrm{trace}(\mathbf{X}\mathbf{X}^{\mathrm{T}}) \cdot \mathrm{trace}(\mathbf{Y}\mathbf{Y}^{\mathrm{T}})}} \quad (2)$$

where $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ and $\mathbf{Y}\mathbf{Y}^{\mathrm{T}}$ are the Gram matrices for two samples described by two-dimensional chromatographic signals **X** and **Y** with the dimensions *the number of spectral channels* × *the number of sampling points on retention time axis*, 'trace(·)' denotes the sum of the squared diagonal elements of a matrix. (Remark: in general, the Gram matrix is a square, symmetric and positive matrix obtained either as $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ or $\mathbf{X}^{\mathrm{T}}\mathbf{X}$).

Values of the Rv coefficient are between zero and one indicating no similarity and the highest similarity between two samples, respectively.

In addition to the already presented measures of similarity between two-dimensional chromatographic signals, another concept exists that can be used when peak shifts are present. It relies on the construction of a so-called 'fuzzy' similarity matrix obtained for 'blurred' or 'semi-blurred' data representations: $\mathbf{U}^{*}\mathbf{U}^{*\mathrm{T}}$ and $\mathbf{U}\mathbf{U}^{*\mathrm{T}}$. Rows of matrix **U** contain unfolded two-dimensional signals, and the asterisk denotes that they were 'blurred' along their time dimension. The signal's 'blurring' is achieved by averaging the values of its elements within a window of a specified size along the time dimension (for more details see Ref. [6]).

To enable a straightforward visualisation of the inter-sample similarities, the constructed similarity matrices according to description provided in this section may be further explored using PCA and/or hierarchical clustering methods.

2.2. Projection methods

There are many different projection methods which can be used for exploring the structure of multidimensional [3] and multi-mode [7] chemical data. Principal component analysis (PCA) [8] plays a fundamental role among these methods. Its aim is to compress a collection of one-dimensional chromatographic signal data into a few new variables, constructed as linear combinations of the original variables maximising the description of the data variance. PCA is a decomposition model and presents data matrix **X** as a product of scores and loadings matrices.

PARAFAC [9], PARAFAC2 [10,11] and Tucker3 [12] methods, known in the literature as the N-way methods, are similar in spirit to PCA. Their aim is to facilitate the exploration of a set of two-dimensional signals (a three-way data array, **X**, with three data modes A, B, and C denoting for the HPLC dimension, the DAD data spectral dimension and samples, respectively). Like PCA, the N-way approaches decompose three-way data into three sets of latent variables **A**, **B** and **C** for each data mode, called loadings. Each of the above-mentioned models is constructed to maximise the descrip-

Table 1
Overview of different approaches used for exploration of two-dimensional chromatographic signals.

Type of chromatographic signals	Require prior signal alignment	No-alignment approach
One-dimensional	Similarity matrix	Similarity matrix insensitive to peak shifts: Gram matrix, Rv coefficient, 'fuzzy' similarity matrix
	PCA	PCA on Rv or 'fuzzy' similarity matrix
	Clustering methods	Clustering on similarity matrix insensitive to peak shifts
Two-dimensional	Tucker3	STATIS
	PARAFAC	PARAFAC2

tion of the data variance. Depending on the method applied, the number of factors in each data mode may differ. In PARAFAC and PARAFAC2 models the same number of factors is extracted for each data mode in contrast to the Tucker3 model where a different number of factors in each data mode is allowed. To achieve a good compromise between a model's complexity and the amount of variance it explains, the number of model factors has to be optimised [7] as well as applying a suitable model. In general, the PARAFAC and Tucker3 models are used for exploring HPLC-DAD data when retention shifts are negligible or when the time axes of signals are synchronised prior to the construction of the N-way model. The PARAFAC model is relatively simple, but it may describe data variance insufficiently due to the equal number of loadings in each data mode. The Tucker3 model is more flexible in this respect. When data peak shifts are observed in the HPLC-DAD, each data slab can be represented as a Gram matrix. It is constructed as a cross-product of the data matrix in such a way that the dimension corresponding to retention time 'disappears'. This is done in PARAFAC2 approach.

If the N-way model describes a substantial part of the data variability, the so-called C-loadings summarise similarities among individual samples well. Additionally, they can be clustered with the hierarchical clustering methods to summarise information contained in a few C-loadings.

STATIS [13], like PARAFAC, PARAFAC2 and Tucker3, can also be used to handle three-dimensional data arrays. The relation of those methods to STATIS has been described in [13]. STATIS does not construct a typical three-way model, but uncovers similarities among samples represented by two-dimensional chromatographic signals on a compromise plot. The compromise in STATIS is obtained as the weighted sum of cross-product matrices constructed for individual samples of the dimensions *the number of spectral channels* × *the number of spectral channels*. A detailed description of this method is provided in [13].

Other approaches to exploratory analysis are necessary when peak shifts are present in two-dimensional signals. For instance, it is possible to replace the original representation of a two-dimensional signal, **X**, with a Gram-like matrix $\mathbf{XX}^{\mathrm{T}}$, where the problematic time dimension 'disappears', as is exploited in PARAFAC2. STATIS has properties similar to PARAFAC2 in handling two-dimensional signals with peak shifts. A quick comparison of two-dimensional signals is also possible with the similarity matrices insensitive to peak shifts (e.g. containing the Rv coefficients, 'fuzzy' similarity matrix) and visualising these similarities on a PCA score plot and/or dendrogram.

2.3. Hierarchical clustering methods

The purpose of hierarchical clustering methods is to group similar samples [14] and to visualise these similarities. The similarity matrix or PCA score space or C-loadings space from N-way techniques can be used as input data for clustering. Clustering of samples is achieved through a stepwise procedure where at each step the two most similar clusters (according to a certain similarity measure, e.g. Euclidean distance, correlation coefficient, etc.) are joined. In this way, a hierarchy of sample similarities is established and displayed as a dendrogram. A sequence of joined samples is listed along the horizontal axis of the dendrogram, whereas the vertical axis provides information about similarities among the samples. Samples grouped in the lowest dendrogram branches are the most similar. Hierarchical clustering methods differ with respect to the linkage principle applied. Among the most popular are single linkage (the linkage distance is equal to the shortest distance observed between points of two clusters), complete linkage (the linkage distance is equal to the furthest distance observed between points of two clusters) and the Ward's method (the linkage distance is equal to the minimal increase of information loss in terms of sum of squares criterion). A comprehensive overview of different clustering approaches is provided in [15].

3. Data set

For illustrative purposes, a collection of 89 HPLC-DAD signals (metabolite profiles of St. John's wort) will be used. Samples of St. John's wort were obtained from commercial suppliers in Africa, Asia, Europe and North America. All samples were analysed using an HPLC system, equipped with a photo-diode array detector (DAD) with two eluents as a mobile phase with a linear gradient: an A eluent (acetonitrile:water 5:95 + 0.1% COOH) and a B eluent (acetonitrile:water 95:5 + 0.1% HCOOH). At each retention time (549 sampling points from 12 to 23.7 min in steps of 1.32 s), a spectrum was registered every 3 nm from 260 to 548 nm. The final chromatographic data contains 89 two-dimensional HPLC-DAD fingerprints with the dimensions (97 × 549) obtained for 24 samples for which a different number of replicates is available. The replicate samples are denoted with capital letters from A to D. A detailed description of the data set is given in [4], and the data set is available from [16].

4. Results and discussion

In this section, we will discuss some issues related to the exploratory analysis of two-dimensional chromatographic fingerprints, namely the HPLC-DAD fingerprints.

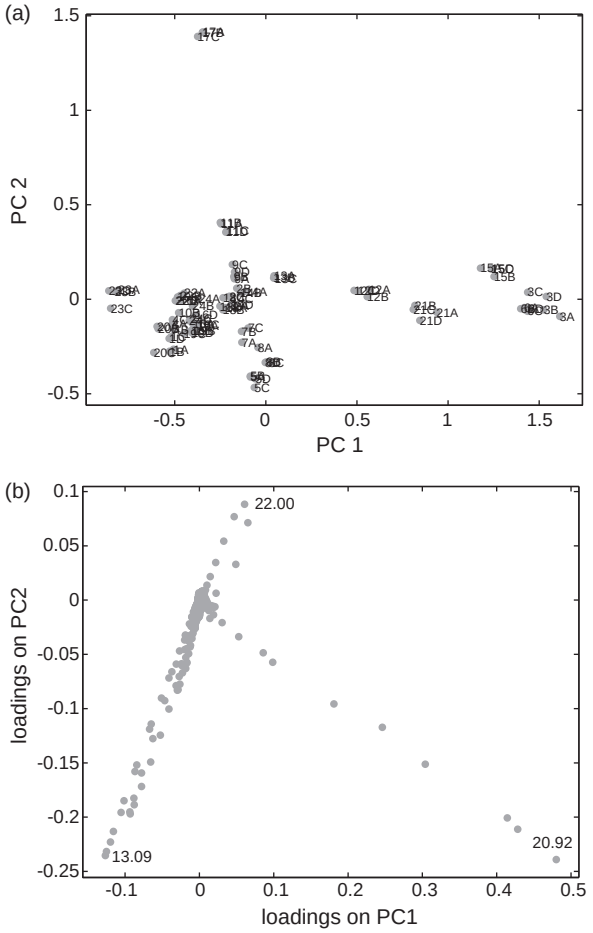

Fig. 1. (a) Score plot of one-dimensional chromatograms—the total spectra chromatograms of 89 HPLC-DAD fingerprints of St. John's wort extracts (two first principal components) with indicated sample number (1–24) and replicate (from A to D), and (b) corresponding loading plot where variables with the largest absolute contribution to a given PC (certain retention times) are indicated.

4.1. Exploring a set of signals with negligible peak shifts

For the sake of presentation, a relatively simple approach of handling two-dimensional HPLC-DAD signals with no peak shifts will be presented. Any two-dimensional chromatographic signals can be transformed into one-dimensional signals, for instance by summing or averaging signal intensities. Although this approach may lead to a great reduction in data, often due to the large size of certain type of signals, this step may be required prior to their further analysis [17]. A simpler data representation is more convenient to process and often allows for general conclusions to be drawn about similarities among samples, for instance by applying PCA. In Fig. 1a a projection of 24 samples and their replicated (A–D) in the space of two PCs is shown. The two PCs explain more than 86.24% of the total data variance and their projection is accompanied by a corresponding loading plot (Fig. 1b).

When the structure of a data set is summarised effectively by a few principal components, PCs (the PCA model explains a large part of the variability of the data), frequently low-dimensional projections of PCs uncover interesting data structures. A score plot allows similarities among data samples to be studied, whereas a loading plot provides information about the contributions of individual variables to a given principal component. A simultaneous analysis of both the score and loading plot allows the contribution of original variables (chromatographic peaks eluting at a certain retention time) to be attributed to the observed pattern of samples on the score plot. The largest differences among samples were revealed along PC1 (see Fig. 1a and b). For samples nos. 3, 6, and 15 a relatively high concentration of quercitine, eluting at around 20.9 min, was observed, whereas rutine, eluting at around 13.0 min, was present at a relatively low concentration. The unique character of sample no. 17 with respect to the remaining samples is explained by the loading values on the PC2. This sample has a relatively high content of biapigenine, eluting at around 22 min, whereas it contains relatively low amounts of quercitine and rutine.

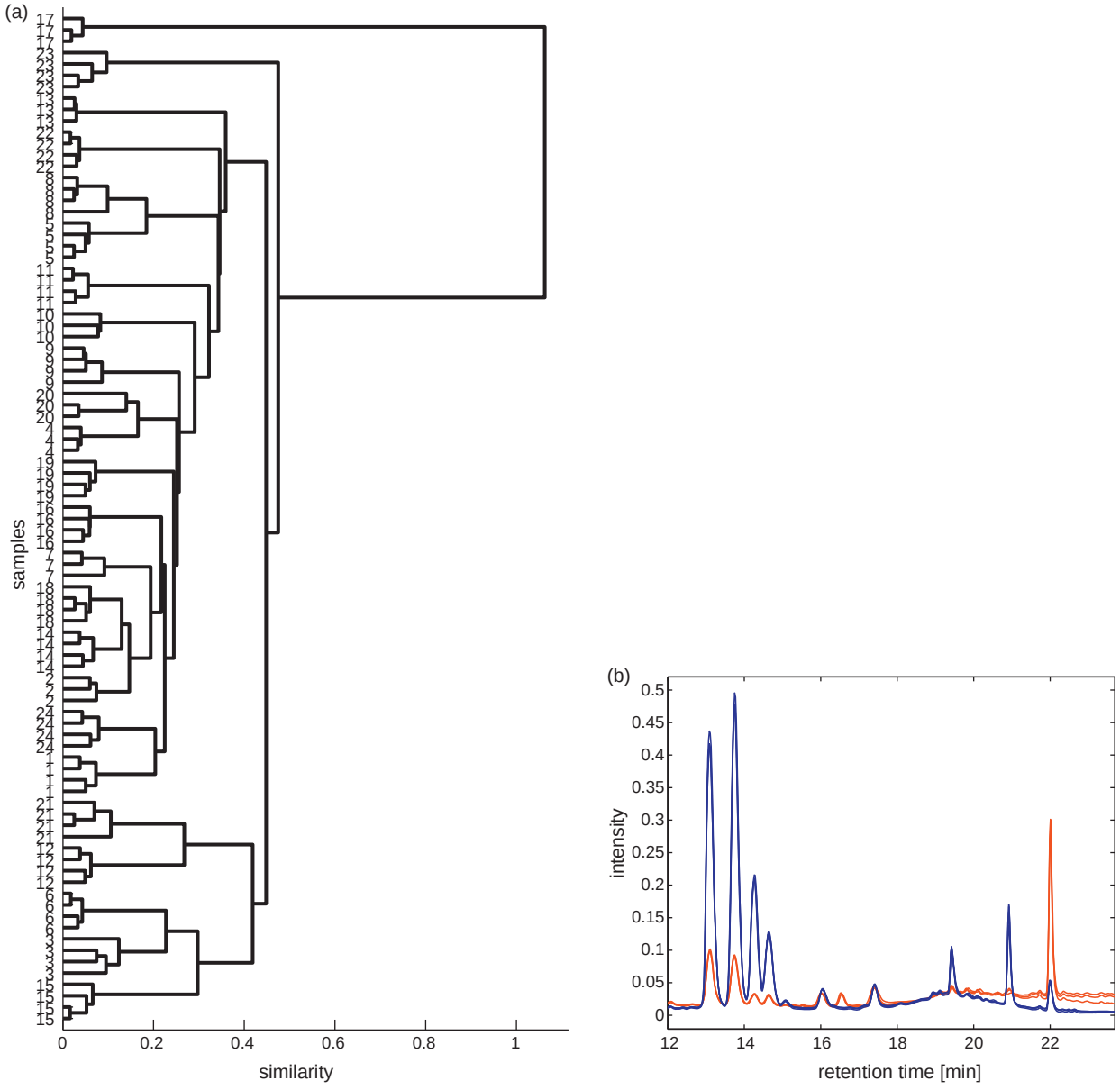

Fig. 2. (a) Single linkage dendrogram of 89 St. John's wort extracts represented by 6 PCs (explaining above 96.95% of total data variance) with the Euclidean distance as a similarity measure, and b) chromatographic profiles of replicates of two samples: 17 (a red line) and 23 (a blue line). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article.)

When the PCA compression is somewhat less efficient, several PCs are often required to explain a substantial part of the data variance, and thus a large number of score plots have to be evaluated simultaneously. To quickly gain an insight into the data structure summarised by a few of the PCs, hierarchical clustering methods can be used with the PCs as input variables. Hierarchical clustering methods present a hierarchy of sample similarities in the form of a dendrogram (also called a tree) the interpretation of which is relatively intuitive. In Fig. 2a a single linkage dendrogram with the Euclidean distance as a similarity measure is presented. It was constructed by clustering samples in the space of the first six PCs explaining more than 96.95% of the total data variance. The differences between certain groups of samples are explained by the analysis of the corresponding signals of certain samples. For instance, in Fig. 2b replicates of samples no. 17 and no. 23 are presented. It is apparent that the largest quantitative differences are characteristic for chromatographic peaks eluting at ca. 13.1, 13.7, 14.3, 14.7, 19.4, 20.9, and 22.0 min. Some of these peaks were identified and correspond to particular chemical substances present in St. John's wort samples as described in [4].

Another approach to exploring a collection of two-dimensional HPLC-DAD signals is to study the similarity matrices obtained from the unfolded individual two-dimensional signals or scoring similarities among individual data tables with the Rv coefficient (see Eq. (2)). Regardless of the similarity measure applied, the final results are presented as a positive and square matrix, known as a similarity matrix, **S**, with the dimensions (*samples* × *samples*). The similarity matrix can also be presented as a heat map or a colour map. Each pixel of a colour map represents a similarity between the *i*-th and the *j*-th sample, and the intensity of colour is proportional to the similarity level.

In Fig. 3a, an example colour map presenting the Rv coefficients for the HPLC-DAD data is shown. An element of the colour map, representing a similarity between two samples, has a certain colour assigned from a colour scale and proportional to the Rv value. Low similarity between samples is represented by dark blue colour, whereas high similarity by intensive red, as indicated by a colour bar in Fig. 3a.

To quickly evaluate the information content of this similarity matrix PCA analysis is a straightforward option (as illustrated later

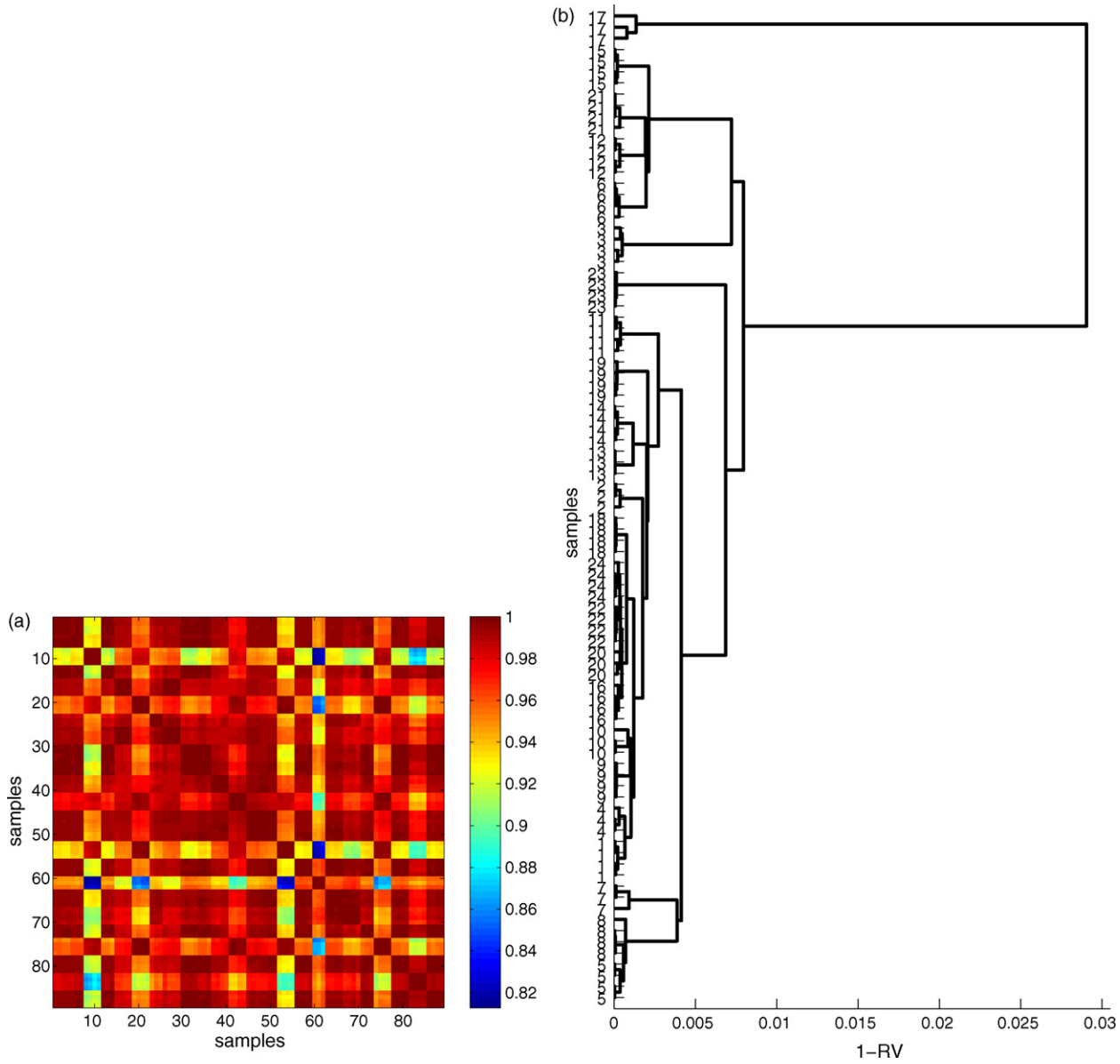

Fig. 3. (a) Colour map of the Rv coefficients for 89 St. John's wort extracts represented by the two-dimensional HPLC-DAD signals and (b) corresponding single linkage dendrogram with (1 − Rv)-similarity measure.

on) as is hierarchical clustering. Since any hierarchical clustering technique works on the similarity matrix, the dendrogram is obtained directly from the similarity matrix or its several significant PCs. Fig. 3b depicts the situation where a single linkage dendrogram with $1 - Rv_{ij}$ was used as a similarity measure. In the dendrogram, all of the replicate samples are grouped together on low branches.

When chromatographic data are arranged as a three-way data array, such a data representation requires the use of N-way methods for their further exploration. The N-way methods, like PCA, act as data dimensionality reduction techniques. They help in constructing a set of latent variables (loadings) describing the relationships among three different data modes. One set of loadings, namely C-loadings, is of particular interest in the context of exploratory analysis because it expresses differences among samples. Several different N-way techniques can potentially be used to summarise the structure of a three-way data set, including e.g. PARAFAC, PARAFAC2, Tucker3 [7] and STATIS [13] approaches. As explained in [4], a set of several loadings capturing a large part of the data variance obtained from a certain N-way data model can also be used as input data for hierarchical clustering approaches. As an example, let us present a single-linkage dendrogram (the Euclidean distance used as a similarity measure) obtained by grouping the C-loadings of PARAFAC (see Fig. 4a). The two-factor model explained more than 95.32% of the total data variability.

As indicated in the dendrogram in Fig. 4b, there is a relatively good agreement between replicates of the samples. Sample no. 17 is unique (located far away from the remaining samples) which also is confirmed when other dendrograms are studied. This sample has relatively high contents of substances eluting between 12 and 15 min as well as between 18 and 22.5 min (see Fig. 2b).

4.2. Peak shifts issue

When chromatographic fingerprints are explored the final conclusions can be significantly affected by the presence of retention shifts in the data. These are manifested as a result of unstable experimental conditions over a longer period of the chromatographic analysis (e.g. the problem of degradation during the stationary phase—column ageing; using a new solvent purchased from other vendor for preparing mobile phase, small pH variations, etc. [18]).

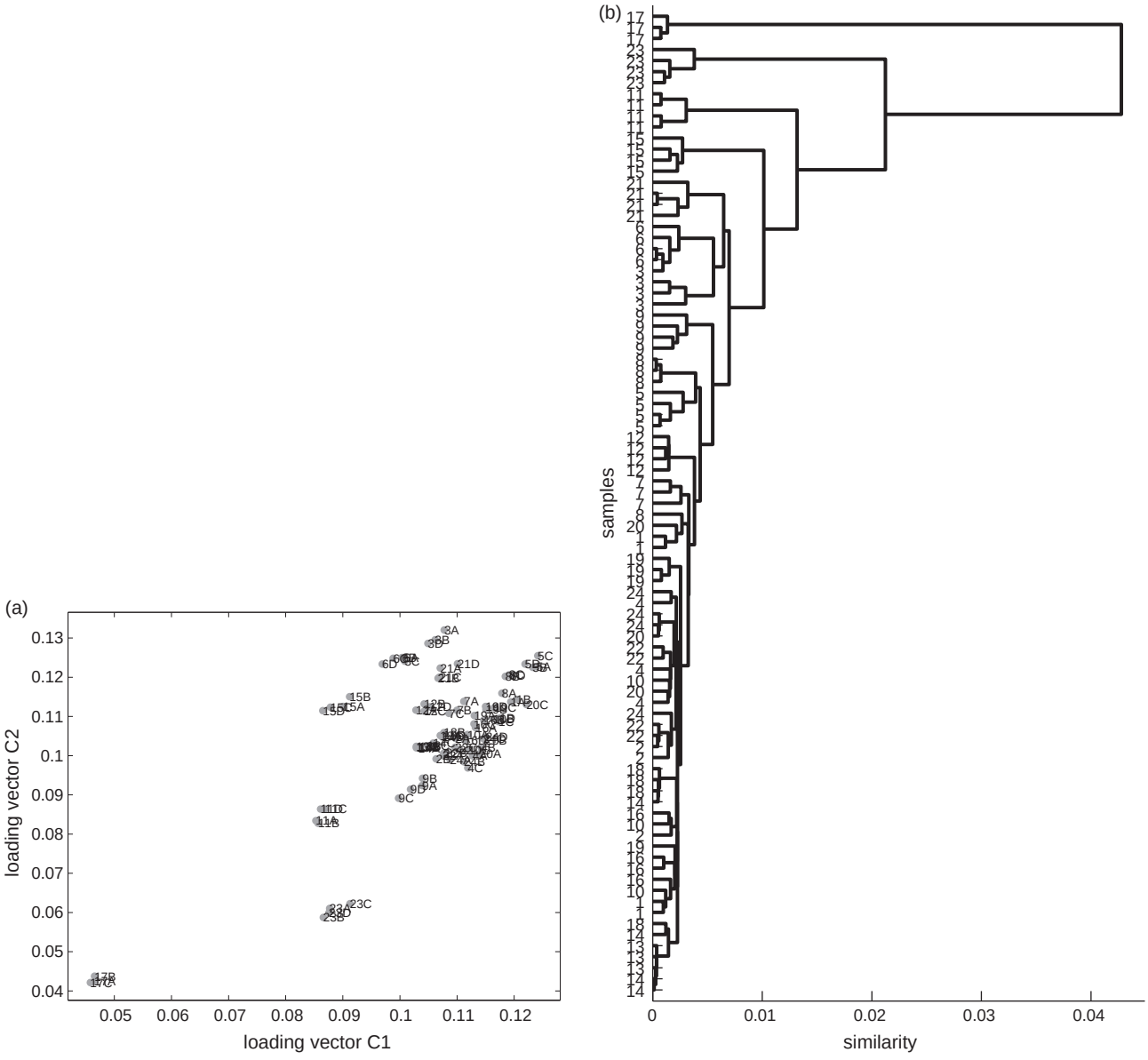

Fig. 4. (a) Projection of two first C-mode loadings obtained from PAFARAC2 method of 89 St. John's wort extracts represented by two-dimensional HPLC-DAD signals (the two-component PARAFAC2 model explained above 95.32% of the total data variance) and (b) the corresponding single linkage dendrogram of the two C-mode loadings.

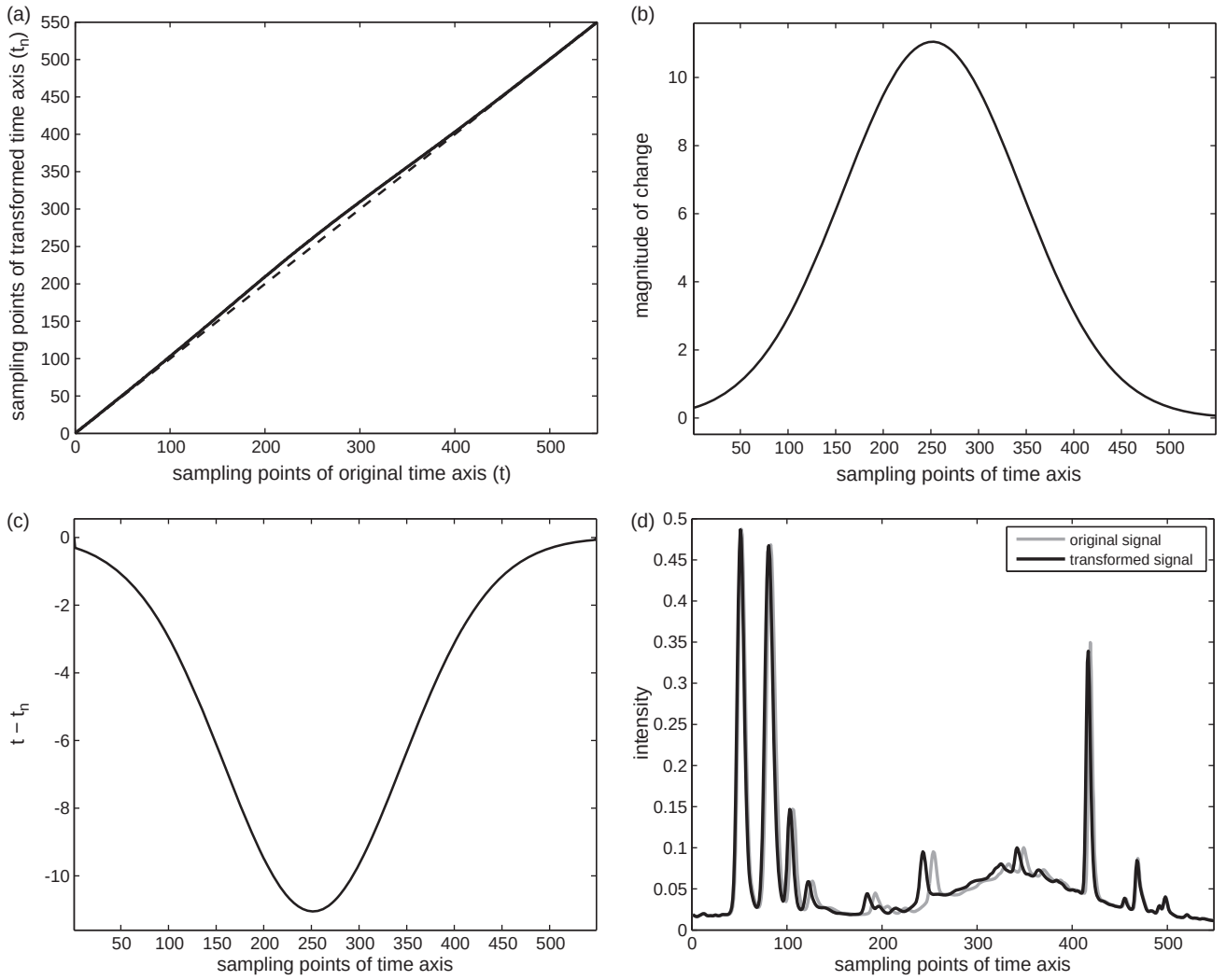

Fig. 5. (a) Plot of sampling points of two time axes in original '*t*' and transformed signal 't_n' (dashed line is the ideal relation when no peak shifts are observed between two signals), (b) shape of the warping function simulated to transform signal of the 1st sample, (c) differences between sampling points on the retention time axis in the original and transformed signals and (d) example of original and transformed signal for sample no. 1 using the warping function illustrated in panel (b) of this figure.

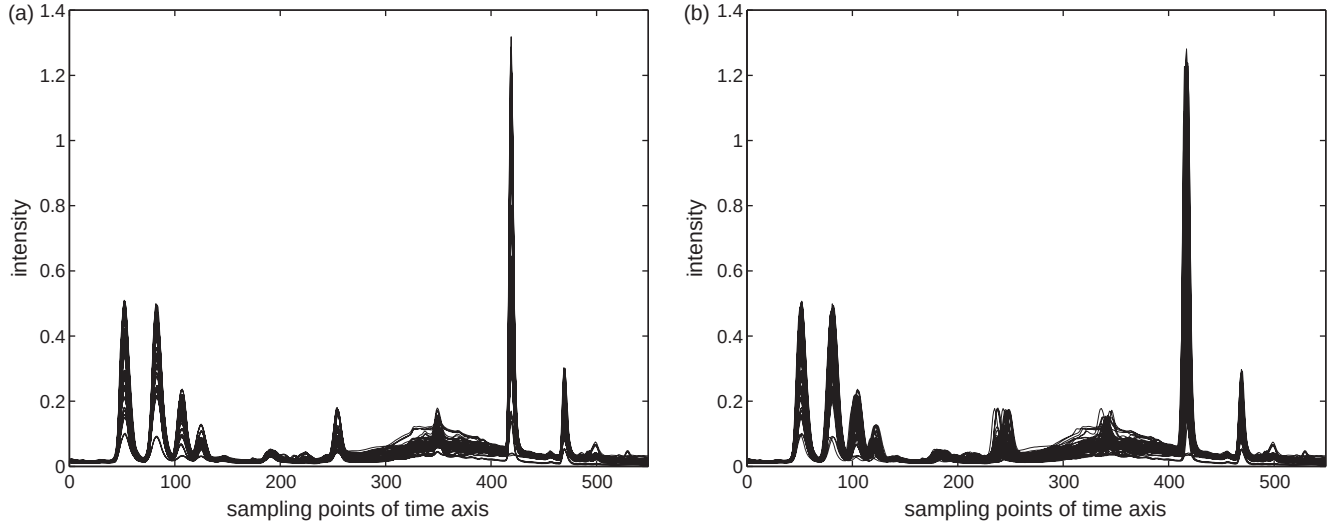

Fig. 6. One-dimensional total spectra chromatograms of 89 St. John's wort extracts for: (a) original signals and (b) signals with simulated retention shifts.

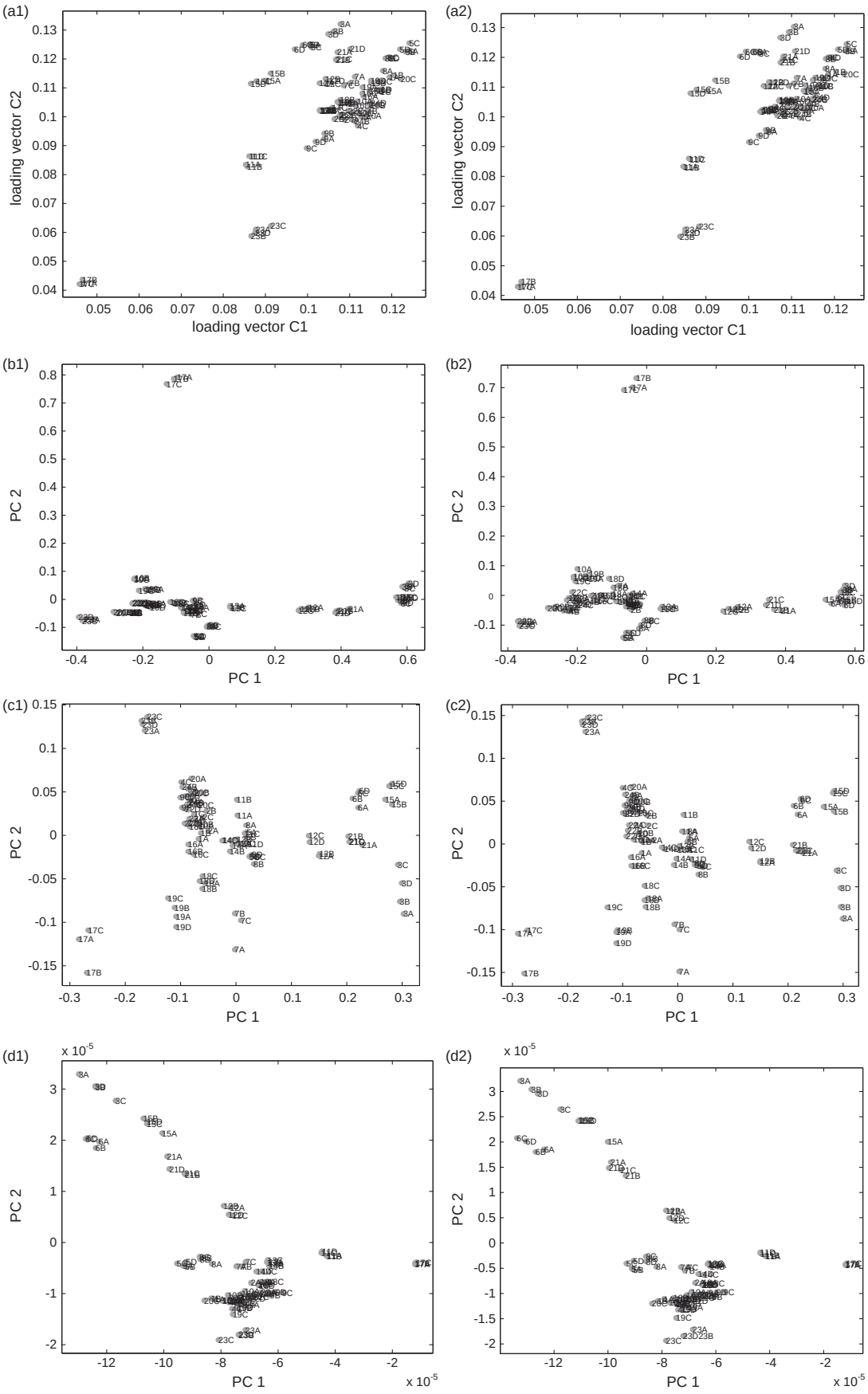

Fig. 7. Projection of the two first factors for 89 HPLD-DAD of signals St. John's wort extracts with negligible shifts and substantial peak shifts (introduced deliberately) obtained from such approaches as: (a) the Rv coefficient, (b) PARAFAC2, (c) 'blurred' correlation matrix, and (d) STATIS.

Two types of approaches are usually used to diminish the effects of peak shifts in chromatographic data. For one- and two-dimensional signals, the classic approach relies on the alignment of the time axes of signals using one of them as a target. For instance, correlation optimised warping (COW) [19] is widely used to account for peak shifts in both types of chromatographic signals. Once the time axes in signals are synchronised, the above-described approaches can be utilised to explore a collection of chromatographic signals. Another possibility is the use of the so-called no-alignment approaches. These implement a similarity matrix insensitive to peak shifts: 'fuzzy' similarity matrices ('blurred' and 'semi-blurred' variants) or use some kind of Gram matrix (constructed in such a way that the time dimension is removed).

In order to illustrate the performance of these two approaches let us consider an artificial HPLC-DAD data set with simulated peak shifts. They were induced by transforming each HPLC-DAD signal according to a simulated warping function (describing a position of a sampling point on the new time axis). The warping function, w, corresponded to a unimodal Gaussian signal with the number of sampling points equal to the number of sampling points on the retention time axis in the original instrumental signal. Initially, a Gaussian peak was placed at sampling point $t = 250$, the peak height $h = 1$ and the peak width $w = 80$. The parameters of the Gaussian peak, h, t, and w, were the subject of a random variation in the course of simulation. A number, drawn from uniform distribution multiplied by a positive constant, was added to each of these parameters. Then, a new signal was obtained by transforming each chromatogram registered at a given spectral channel by the linear interpolation of signal intensities to sampling points on the new time axis, $t_n = w(t)$.

In an ideal case, when no peak shifts are observed between two chromatographic signals, the sampling points of two signals are located on a line with a unit slope. Otherwise, deviation from this perfect relationship is observed. In Fig. 5a, a scatter plot of the time sampling points for original and transformed time axis of signal no. 1 is presented. The time axis was transformed using a simulated warping function (see Fig. 5b) leading to changes in the sampling points on the retention time axis with respect to the original time axis. The negative values of the differences (see Fig. 5c) indicate an earlier elution of the peaks in the transformed signal (a left shift) compared to the elution time of the same peaks in the original signal as demonstrated for sample no. 1. The total spectra chromatograms (i.e. intensities summed over each retention time point) are shown in Fig. 5d.

Fig. 6a and b present the total spectra chromatograms of all HPLC-DAD samples for the original and the transformed signals, respectively. In the set of transformed signals, peak shifts are greatly pronounced, especially around sampling point $t = 250$ on the retention time axis, whereas in the original data they are negligible.

In this section, we will compare the performance of different exploratory approaches (PARAFAC2, STATIS, Rv and 'blurred' similarity matrix constructed using window size equal to one) with respect to their robustness in handling two-dimensional chromatographic signals with peak shifts. Peak shifts in signals were introduced as described in the previous section. As the target, the projection of original data samples (without peak shifts) obtained from a given approach will be used and then compared with the corresponding projection for signals with simulated peak shifts. Two projections for each method examined are presented in Fig. 7. They are derived from the original and the transformed signals (with simulated peak shifts). A general conclusion for this data is that all of the methods discussed in this section are relatively insensitive to peak shifts. This is confirmed by very similar loading projections for the original and transformed signals. Moreover, very little scatter was observed among the replicate samples. In order to quantify the level of agreement between pairs of projections, the Procrustes similarity measure, D (bearing in mind possible scaling and rotation between two projections) was calculated in the space of two factors as follows:

$$D = \frac{\sum_{i=1}^{m}\sum_{j=1}^{2}(f_{ij}^{*} - f_{ij})^2}{\sum_{i=1}^{m}\sum_{j=1}^{2}(f_{ij} - \bar{f}_j)^2} \quad (3)$$

where, f_{ij}^{*} are scores of the first two factors after the Procrustes transformation [20] and $\bar{f}_j$ is the mean value for the j-th factor.

The smallest differences between projections for the original and transformed signals were observed for the Rv approach ($D = 0.0015$), then for STATIS ($D = 0.0031$), PARAFAC2 ($D = 0.0054$) and finally for the 'blurred' similarity matrix approach ($D = 0.1055$). Despite a larger D value for the 'blurred' similarity matrix approach, the two projections look virtually the same with a slightly larger spread of samples compared to the scatter of points in the projection of the original signal. It is interesting to note that on the projections from PARAFAC2 and the 'fuzzy' similarity matrix approach three replicates A, B, and C of sample no. 17 are different from the remaining ones; however, on the projection obtained from the Rv approach their outlying character is less pronounced. It is also apparent from Fig. 7a–d that depending on the approach applied to construct a projection, the patterns of samples seem to be quite different at the first sight. When clusters of replicate samples are carefully inspected, in general, similar conclusions about sample similarities can be drawn. For instance, clusters of replicate samples nos. 3, 6, 12, 15 and 21 are close regardless of the method used to obtain a projection (cf. Figs. 7).

Another important problem that requires further attention is associated with the lack of straightforward interpretability of projections in terms of the original data variables when the data are represented as some kind of the Gram matrix. Unfortunately, peak shifts in signals do not allow for a simple data projection in order to obtain the contributions of the original variables to latent factors as can be done in the kernel variant of PCA [21].

5. Conclusions

Data exploration is an important step in the discovery of knowledge. When two-dimensional chromatographic HPLC-DAD signals are explored, they may require a different exploratory treatment than one-dimensional signals. In certain cases, exploration of the HPLC-DAD signals requires their prior pre-processing (background removal and noise suppression—when necessary). A serious obstacle in the exploration of such signals is caused by retention shifts. If these are negligible in the data being studied, several options are available. For instance, the two-dimensional signal can be reduced to one-dimensional signal and treated using different projection methods including PCA and hierarchical methods (e.g. in the space of a few significant PCs) later on. The HPLC-DAD signals can also be explored with the N-way projection methods followed by a further data hierarchical clustering on a set of C-mode loadings. Interpretation of clusters revealed on low-dimensional projections in terms of individual variables is achieved by examining PCA loading plots and chromatographic profiles characteristic for a given cluster of samples.

The presence of peak shifts in the HPLC-DAD signals have to either be corrected with the signal alignment methods or the no-alignment approaches (the Rv coefficients, 'fuzzy' similarity matrix, and PARAFAC2) should be considered.

References

[1] Y. Heyden, LC–GC Europe 21 (2008) 438–443.
[2] R. Jellema, in: S. Brown, B. Walczak, R. Tauler (Eds.), Comprehensive Chemometrics, vol. 2, Elsevier, Amsterdam, 2009, pp. 85–108.

[3] M. Daszykowski, B. Walczak, D.L. Massart, Chemometrics and Intelligent Laboratory Systems 65 (2003) 97–112.
[4] E. Acar, R. Bro, B. Schmidt, Journal of Chemometrics 22 (2008) 91–100.
[5] M. Daszykowski, R. Danielsson, B. Walczak, Journal of Chromatography A 1192 (2008) 157–165.
[6] R. Danielsson, D. Bäckström, S. Ullsten, Chemometrics and Intelligent Laboratory Systems 84 (2006) 33–39.
[7] P. Kroonenberg, Applied Multiway Data Analysis, Wiley, Hoboken, NJ, 2008.
[8] S. Wold, K. Esbensen, P. Geladi, Chemometrics and Intelligent Laboratory Systems 2 (1987) 37–52.
[9] R. Bro, Chemometrics and Intelligent Laboratory Systems 38 (1997) 149–171.
[10] H. Kiers, J. Ten Berge, R. Bro, Journal of Chemometrics 13 (1999) 275–294.
[11] R. Bro, C. Andersson, H. Kiers, Journal of Chemometrics 13 (1999) 295–309.
[12] P. Geladi, Chemometrics and Intelligent Laboratory Systems 7 (1989) 11–30.
[13] I. Stanimirova, B. Walczak, D. Massart, V. Simeonov, C. Saby, E. Di Crescenzo, Chemometrics and Intelligent Laboratory Systems 73 (2004) 219–233.
[14] N. Bratchell, Chemometrics and Intelligent Laboratory Systems 6 (1987) 105–125.
[15] D. Massart, L. Kaufman, The Interpretation of Analytical Chemical Data by the use of Cluster Analysis, Robert E. Krieger Publishing Company, Malabar, FL, 1989.
[16] http://www.models.life.ku.dk/Bonnie.
[17] M. Daszykowski, W. Wu, A. Nicholls, R. Ball, T. Czekaj, B. Walczak, Journal of Chemometrics 21 (2007) 292–302.
[18] G. Malmquist, R. Danielsson, Journal of Chromatography A 687 (1994) 71–88.
[19] N. Nielsen, J. Carstensen, J. Smedsgaard, Journal of Chromatography A 805 (1998) 17–35.
[20] J.M. Andrade, M.P. Gomez-Carracedo, W. Krzanowski, M. Kubista, Chemometrics and Intelligent Laboratory Systems 72 (2004) 123–132.
[21] W. Wu, D.L. Massart, S. de Jong, Chemometrics and Intelligent Laboratory Systems 36 (1997) 165–172.