Chemometrics and Intelligent Laboratory Systems 114 (2012) 36–43

Contents lists available at SciVerse ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Extracting homologous series from mass spectrometry data by projection on predefined vectors

Johan E. Carlson [a,b,*], James R. Gasson [a], Tanja Barth [a], Ingvar Eide [c]

[a] Department of Chemistry, University of Bergen, Allégaten 41, NO-5007 Bergen, Norway
[b] Department of Computer Science, Electrical and Space Engineering, Luleå University of Technology, SE-971 87 Luleå, Sweden
[c] Statoil Research Centre, NO-7005 Trondheim, Norway

ARTICLE INFO

Article history:
Received 2 November 2011
Received in revised form 18 January 2012
Accepted 13 February 2012
Available online 23 February 2012

Keywords:
Chemometrics
Compound classes
Mass spectrometry
Fingerprint
Principal components
Bio oil

ABSTRACT

Multivariate statistical methods, such as Principal Component Analysis (PCA), have been used extensively over the past decades as tools for extracting significant information from complex data sets. As such they are very powerful and in combination with an understanding of underlying chemical principles, they have enabled researchers to develop useful models. A drawback with the methods is that they do not have the ability to incorporate any physical / chemical model of the system being studied during the statistical analysis. In this paper we present a method that can be used as a complement to traditional chemometric tools in finding patterns in mass spectrometry data. The method uses a pre-defined set of equally spaced sequences that are assumed to be present in the data. Allowing for some uncertainty in the peak locations due to the uncertainties for the measurement instrumentation, the measured spectra are then projected onto this set. It is shown that the resulting scores can be used to identify homologous series in measured mass spectra that differ significantly between different measured samples. As opposed to PCA, the loading vectors, in this case the pre-defined homologous series, are readily interpretable.

1. Introduction

Multivariate analysis of mass spectrometry (MS) data provides valuable insight into systematic variations in experimental data sets. Techniques such as, for example, Principal Component Analysis (PCA) are widely used for this purpose [1]. PCA is optimal in the sense that it compresses (linear) experimental variation into as few components as possible, thus efficiently reducing the dimensionality of the problem.

A drawback with these techniques is, however, that it is difficult to make chemical interpretations of the results, since the actual dimensionality reduction and decomposition of the experimental data do not necessarily reflect specific underlying chemical principles. MS data of complex mixtures, e.g. petroleum or bio-oil products, contain regularly spaced signals which reflect classes of chemical compounds that vary in a regular manner. These are not reflected in an interpretable format in the loading vectors when using PCA for data analysis. In this paper we propose an alternative strategy to make use of these traits exploiting fundamental properties of the molecular composition of the samples being studied.

Our alternative approach gives a set of relatively few components that are sufficient to discriminate samples which have different chemical compositions. The components that are generated immediately lend themselves to interpretations in terms of the underlying chemical composition of the samples.

An example comprising MS analysis data on a screening set of bio-oils is used to demonstrate the algorithm and how the results can be interpreted. The results will also be compared to traditional PCA and relations to other techniques will be discussed.

2. Background

2.1. The evolution of mass spectrometry

Recent developments in MS techniques have lead to a revolutionary revival of one of the oldest analytical techniques associated with the identification of chemical components in complex mixtures such as petroleum. Long since has MS become an elementary part of most analytical organic laboratories as an analyser especially in hyphenated instrumentation such as Gas Chromatography (GC)-MS or Liquid Chromatography (LC)-MS. Hard ionisation techniques such as Electron and Chemical Ionisation (EI and CI), which are typically used within this kind of instrumentation to analyse small molecules, produce characteristic fragmentation patterns of the analytes, enabling their identification [2]. This is supported by modern library search interfaces/engines, (e.g. NIST MS Search Program), using a

* Corresponding author at: Department of Computer Science, Electrical and Space Engineering, Luleå University of Technology, SE-971 87 Luleå, Sweden. Tel.: +46 920 492517.
E-mail address: Johan.Carlson@ltu.se (J.E. Carlson).

0169-7439/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2012.02.007

wide variety of different algorithm approaches coupled with extensive databases, the two dominant of which today are NIST 11 Mass Spectral Library and Wiley's Registry of Mass Spectral Data [3,4].

Developments in soft ionisation techniques, such as in Matrix Assisted Laser Desorption/Ionisation (MALDI)–MS or Electrospray Ionisation (ESI)–MS and high resolution analysers have opened new pathways of analysis especially in terms of macro-molecules and complex mixtures, which until now largely relied on bulk property analysis methods such as density, acidity, Infra-Red (IR), or Ultra Violet (UV)/Visible (VIS) spectroscopy [2]. Combining the avoidance of fragmentation of molecular ions and the increased accuracy in mass determination allow the assignment of elemental compositions and subsequent identification of the single compounds. This is supported by the analysis of isotopic peak ratios and complementary identification of lower homologue of the same chemical compound class [5]. This approach has enabled a new way of thinking within *-omics* sciences, enabling a fingerprint type detection of large amounts of smaller molecules (<1500 Da) in complex mixtures using a single or only a small number of analyses.

2.2. Fingerprinting

Characteristic patterns may already be seen in fingerprint mass spectra, even without statistical evaluation. Characteristics can be observed for both dimensions, i.e. both for the spacings between signals along the mass/charge (m/z) axis for key fragmentations or structural analogues, as well as abundance ratios in respect to isotopic patterns. This has been exemplified e.g. using GC-MS data to extract chlorinated components from mixtures [6,7]. In addition, classification methods for double bond positional isomers have also been established [8]. Using high resolution equipment, numerical identifiers can play a strong role in supporting the identification of molecules of the same functional class. When translating all atomic masses in a mass spectrum of crude oil, for example, from the commonly used Dalton mass-scale to the Kendrick mass-scale, the latter of which sets $CH_2 = 14.0000$ Da, a periodicity of reoccurring numerical values in the sub-integer area, so called mass defects, describing molecules of the same identical base-structure with varying aliphatic chain lengths attached, become visible [9,10]. These homologous series are of considerable interest, as they cannot only reflect the effectiveness of, for example, a catalyst on selected species dependent of the length of attached substituents, but the abundance spread of the compounds may also be used as an indicator towards physical properties, such as the boiling point range. Choice of the ionisation method can in addition enable a more precise focus on the species of interest. ESI, for example, which commonly ionises only polar components, gives a selective picture of this heteroatomic polar fraction, which was used e.g. to analyse thermal oxidative stability of aviation fuels [11].

Fig. 1 shows a crude bio-oil ESI mass spectrum taken from our application example (see Section 4). The reoccurring spacings in the range of 14 Da, which are equivalent to one additional CH_2 group, are clearly visible in the magnified part of the spectrum. Resolution restrictions of the instrument do not permit more precise specifications of molecular weights or spacings. In addition, 2 Da spacings are also observed, as also noted in prior work, which give an indication to the loss of H_2 and thus the replacement of a saturated hydrocarbon bond with a double bond [12–14].

2.3. Statistical evaluation of mass spectrometry data

As our example suggests, lower resolution units, without the benefits of elemental association to accurate mass numbers, are nonetheless suitable to explore these kind of periodical signatures in combination with PCA to yield good results [12,15]. The additional information gained by statistical analysis is quite considerable and implementation is generally easier than for chromatographic data.

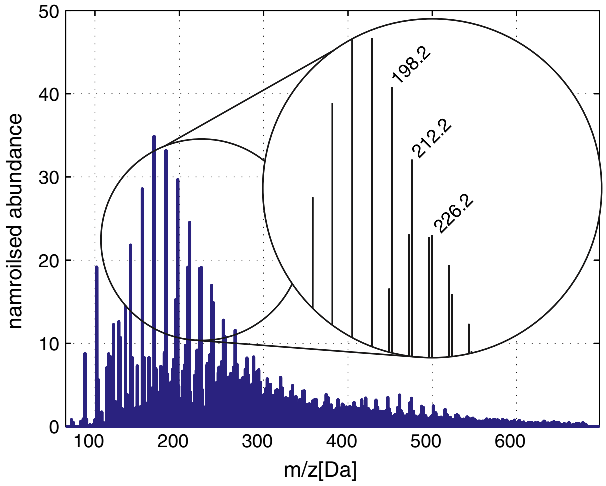

Fig. 1. Positive ESI—mass spectrum of the first replicate of LtL-Oil F04t from the application example. The crude sample was dissolved in dichloromethane and analysed by full scan mass spectrometry on an Agilent 1100 Series LC/MSD system using an acetonitrile-aqueous ammonium acetate (50 mM) 9:1 solution as a mobile phase. The analysis was performed without prior separation over a chromatographic column. Periodical reoccurring spacings of both 2 and 14 Da are clearly visible within this spectrum.

Direct injection MS avoids some of the typical complications, such as chromatographic effects and peak matching issues, which demand complex curve resolution approaches [16]. The constraints imposed by e.g. GC-MS analysis can reach even further than just data-processing complications. Critical points are also the instrument run-time per analysis and possible restrictions as to analysable components, e.g. boiling point limitations of the method.

Direct injection MS fingerprinting and profiling analysis in combination with clustering methods present a novel opportunity to classify and also access non-trivial correlations in mixtures, exemplified amongst others by the analysis of different whiskeys, beers and honeys [17,18]. High resolution measurements have also been attempted and the knowledge of the elemental composition of the single analysed components gives further information [19]. These findings are frequently supported by complementary analytics, which do not necessarily require to be directly tied into the chemometric evaluation [20].

Statistically supported analysis of MS data can allow new insights, especially when working with complex data-sets, given correct pretreatment and statistical evaluation of the data. A certain degree of awareness when trying to classify recurring periodical peaks, in homologous series as well as adducts, is furthermore essential [21,22]. Wold and Christie point out that whilst use of pattern recognition for MS data-sets clearly results in a larger amount of information, the introduction of a class specification is necessary to yield significantly balanced information [22].

Extraction of information from MS data has mainly been accomplished on the basis of library search systems, wherein only limited approaches based on homologous series have been undertaken. One example is the library search system SISCOM, which extracts homologous series not only on the basis of a formal method, such as peaks with a relative intensity over a constant threshold and a specific consecutive interval of length between them, but combines these typically implemented restrictions with a search algorithm focussed on characteristic ions observed not in the immediate neighbourhood of ion peak, but from the neighbouring homologous ions [23].

3. Theory

3.1. PCA in comparison to 14 Da model-based analysis

Data compaction of MS data is commonly achieved using the methodology of principal components. It is common understanding that a good PCA result concludes the smallest number of principal components that describe the largest part of the variance and thus give the highest degree of data compaction and the lowest dimensionality. This purely mathematical approach succeeds with regard to data compaction, and highlights numerical correlations which require further interpretation. This can be obtained both from the scores and loading vectors/plots of the single PCs.

PCs purely aim to display maximal systematic variance on a numerical basis without any consideration of background information. If there is collinearity in the data, singular PCs can combine several effects, which the user has to try and separate based on his knowledge of the inherent properties of the analysed system. This challenges the direct use of the loadings vectors for modelling and prediction purposes [24]. In our own research this problem has frequently been observed, when aiming to predict abundances of sets of series in MS data upon alteration of process parameters [13].

In this paper, we propose an alternative method for analysis of our MS data. This approach is not purely based on mathematical or statistical principles, but also exploits the chemical principles involved, and directly implements these principles to treat the data in such a manner that the information which is being sought becomes more easily accessible. We therefore introduce new orthogonal components based on regularly spaced peaks, in this case 14 Da, thus directly applying restrictions based on the chemical properties to yield more targeted results. In comparison to PCA, this allows a more direct correlation with the chemistry of the experimental values and will therefore provide a higher degree of understanding of the system in itself. The aim is not to describe as much of the total variance as possible, but rather to describe variation between samples based on this pre-defined pattern. Because the loadings are pre-defined, the method does not suffer from problems stemming from collinearity in the data.

It should be stressed that the primary aim is not to be able to model the observed data based on the analysis, but rather to reveal patterns that can be used for fingerprinting of different samples. The proposed method does not yield a transformation of data in such a way that the original spectra can be approximated using the scores and loadings, which is the case for PCA.

3.2. The Dalton sequence analysis

This section will describe the principle for projecting MS data onto a set of mutually orthogonal 14 Da spaced sequences. We will start with the ideal case, assuming that the peaks are located at exactly 14 Da intervals. In practice, however, peak locations may drift slightly, due to the fact that the m/z values are also a measured quantity and thus subject to uncertainties in the data acquisition. After describing the ideal case, we will propose a method for taking some of this uncertainty into account.

The general idea behind the analysis of MS data in terms of different 14 Da sequences can be seen as a correlation between the original spectrum and a set of candidate 14 Da spaced sequences. Let $\mathbf{x}$ be an $N\times 1$ vector containing a measured spectrum, and let $\mathbf{U}$ be a matrix which columns consist of M orthogonal 14 Da sequences of unit length. In the ideal case, such a sequence is a vector of the same length as the spectrum, constructed as

$$\mathbf{w}_1 = [1 \quad 0 \quad 0 \quad \cdots \quad 1 \quad 0 \quad 0 \cdots]^T, \tag{1}$$

i.e. a 1 followed by 13 zeros, followed by a 1, 13 zeros, and so on. The second candidate sequence is formed in the same way, but shifted one step, so that the first element is zero. By construction, this will lead to a set of orthogonal vectors $\{\mathbf{w}_i\}$, for $i=1,2,\ldots,M$. These vectors are then normalised to unit length by

$$\mathbf{u}_i = \frac{\mathbf{w}_i}{\sqrt{\mathbf{w}_i^T\mathbf{w}_i}}. \tag{2}$$

We now have a set of orthonormal vectors that can be stored as columns of the matrix $\mathbf{U}$ as

$$\mathbf{U} = \begin{bmatrix} \vdots & \vdots & \cdots & \vdots \\ \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_M \\ \vdots & \vdots & \cdots & \vdots \end{bmatrix}, \tag{3}$$

where

$$\|\mathbf{u}_i\|_2 = 1, \tag{4}$$

for $i=1, 2, \ldots, M$, and

$$\mathbf{u}_i^T\mathbf{u}_j = 0, \tag{5}$$

for $i\neq j$.

The number of possible sequences $\mathbf{u}_i$ is determined by the resolution of the measurement equipment. If, for example, the instrument can only measure at integer m/z values and we are constructing 14 Da spaced sequences, there are only 14 unique candidate sequences.

We can now obtain scores (i.e. the weights from this new basis) by forming the product

$$\mathbf{t} = \mathbf{U}^T\mathbf{x}. \tag{6}$$

In the case of K measured spectra, these can be stored as columns of a matrix $\mathbf{X}$ as

$$\mathbf{X} = \begin{bmatrix} \vdots & \vdots & \cdots & \vdots \\ \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_K \\ \vdots & \vdots & \cdots & \vdots \end{bmatrix}, \tag{7}$$

Prior to the analysis, the columns of $\mathbf{X}$ are standardised to unit variance. This to done to avoid any potential variations of scale in the measurements to influence the interpretation. The scores are then obtained as columns of the matrix $\mathbf{T}$ by modifying Eq. (6) to

$$\mathbf{T} = \begin{bmatrix} \vdots & \vdots & \cdots & \vdots \\ \mathbf{t}_1 & \mathbf{t}_2 & \cdots & \mathbf{t}_K \\ \vdots & \vdots & \cdots & \vdots \end{bmatrix} = \mathbf{U}^T\mathbf{X}. \tag{8}$$

After applying Eq. (8) scores and the basis vectors (columns of the matrices $\mathbf{T}$ and $\mathbf{X}$) are sorted in order of decreasing variance of the columns in $\mathbf{T}$. By doing so, the first score vector will be associated with the 14 Da sequence that differs the most between the measured spectra, similar to how scores and loadings are sorted when using PCA.

It is worth noting here, that for PCA mean centering of the data set prior to analysis is important. The reason being that the loading vectors are determined by the eigenvectors of the covariance matrix of the data set. Any large offset would therefore result in a PC pointing to the mean value of the data set, which is generally not of interest. For the method proposed here, the mean centering is not necessary, since it would only shift the computed scores. The relative variance of the scores is not affected, since the loading vectors are not data dependent.

In order to merge several measured spectra into a matrix $\mathbf{X}$ as in Eq. (7), all spectra must be pre-processed so that they share a common m/z vector. Several approaches have been proposed for doing this [25,26]. In this work, the measured m/z values are first rounded to nearest ±0.05 Da for each spectrum. The spectra are then re-sampled corresponding to a uniformly sampled m/z vector with a 0.05 Da sampling step. It is worth noting that some pre-processing like this is necessary for all analysis methods (e.g. PCA) that require multiple spectra, which are not uniformly sampled, to be stored in a common matrix.

In practice, however, the peak locations may be shifted by a small fraction of a Da either to the left or to the right of the expected location. If we were to use the projection in Eq. (8), this would lead to misleading results, since a slight shift of the peak will affect the score corresponding to a different basis vector than expected. To account for this, we need to allow for a certain spread of the peaks around the ideal location, such that misaligned peaks will still contribute to the same score. In this work we propose to replace each discrete peak in $\mathbf{w}_i$ with a peak shaped like a Gaussian probability density function (i.e. a bell-shaped curve). The width, which is a design parameter of the algorithm, is specified by the standard deviation (given in Da) of the Gaussian probability distribution. In order to maintain the mutual orthogonality of the basis vectors, however, the Gaussian curve has to be truncated at some distance from the mean value (i.e. ideal peak location). In this work, the Gaussian peaks are cut at a $\pm 4\sigma$ distance from the mean value, since at this point the Gaussian peak has decayed to almost zero.

As a consequence of widening the peaks, the shift between consecutive basis vectors must be made larger, otherwise the orthogonality will no longer hold. Fig. 2 shows a section of the first three basis vectors, $\mathbf{u}_1$, $\mathbf{u}_2$, and $\mathbf{u}_3$. For a peak width of $\pm 4\sigma$, the resulting shift between the vectors becomes 8σ. Again, the resulting basis vectors are normalised to unit length according to Eq. (2).

The main rationale behind using the Gaussian shaped peaks is that by this approach, any offset or drift in the measured spectra will still result in a contribution to the same score. The score will, however, be weighted by a factor proportional to the Gaussian probability density function, so that the further away from the ideal peak location we get, the less importance is attributed to that spectral component. Other shapes of the peak could of course be used, which would affect the results. For example, using a rectangular peak with a certain width, would be the same as assuming a uniform probability density function for the uncertainty of the true peak locations. As a consequence, peaks located away from the ideal location would contribute as much to the scores as peaks at or very close to it. This would yield the same results as rounding of decimals along the *m/z* axis, assuming the resolution of the instrument is much lower than it actually is. From an instrumentation and measurement perspective, this is not realistic, and therefore this has not been investigated further in this paper.

As mentioned earlier, the peak width, defined by the standard deviation σ is a design parameter that has to be set by the user. If the parameter is chosen too small, more basis vectors will be required and the scores will be spread across these, even if the spectral components originate from the same compound. If the peaks are made too wide, this will result in a loss of resolution in the analysis, as peaks resulting from different chemical compounds will be attributed to the same candidate Da spaced sequence.

In the application example in the next section, the peak width was defined by $\sigma = 0.05$ Da. Given an assumed uncertainty of the measurement of ± 0.1 Da, this would correspond to $\pm 2\sigma$. Noting that for a Gaussian probability density function, the probability of being within $\pm 2\sigma$ from the mean is around 95%, which we considered to be a reasonable trade-off between resolution and uncertainty. The peaks spacing is still 8σ, in order for the Gaussian distribution to decay to almost zero.

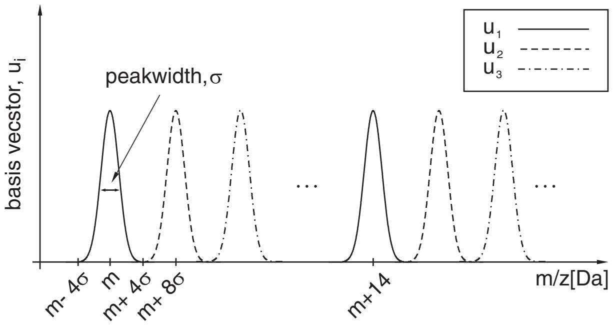

Fig. 2. Definitions of the basis vectors used in the analysis. The figure shows a section of the three first vectors only.

3.3. Relation to other techniques

We started the discussion by describing how PCA can be used for fingerprinting purposes. Although PCA is not primarily designed for this task, it has been shown to work well [12–14]. In addition to revealing underlying patterns (i.e. latent variables), PCA can also be used to develop models of the measured data, based on the principal components. This is usually done by means of Principal Component Regression (PCR) [1]. For exploring and modelling non-linear variance, Non-Linear PCA (NLPCA) based on neural networks, can be used [27,28].

The proposed method is, in contrast to the other techniques, designed only to reveal variability caused by specific patterns in the data, believed to be important when the task is to discriminate between samples. An important constraint is also that the patterns should have chemical meaning, which is not the case for any of the other methods mentioned above. Hence, it is not possible to approximate the original data set based on the scores and loadings of this method. As a consequence, it cannot be used for regression modelling and prediction. To clarify this point further, let us again look at what scores and loadings mean in the context of PCA and the 14 Da sequence analysis, respectively. In PCA, the original matrix $\mathbf{X}$ can be expressed as

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E}, \tag{9}$$

where the columns of $\mathbf{P}$ are the loading vectors, given by the eigenvectors of $\mathbf{X}^T\mathbf{X}$. $\mathbf{T}$ are the corresponding scores, or *weights*, which can be seen as the projection of the original data onto the loading vectors, thus describing *how much* of each loading vector is found in each of the measured spectra. The matrix $\mathbf{E}$ contains the residuals, resulting from discarding of less significant PCs. In the proposed method, the scores have the same meaning, i.e. they describe how well the spectra correlate with each of the candidate 14 Da sequences. The sequences themselves (the columns of the matrix $\mathbf{U}$) are here called loadings, as an analogy to PCA. Expressing the original data set in terms of the scores and the candidate sequences in the same way as in Eq. (9) is, however, not possible.

Another technique, that similarly to the proposed method, is looking for components that can be immediately interpreted in terms of the underlying chemical composition, is the Alternating Regression (AR) method [29]. AR aims at expressing observed spectra as a linear combination of a set of underlying spectra. In this way it is similar to the proposed method. AR does, however, not use any pre-defined set of components, but estimates these iteratively, starting with some random values. As such the technique is sensitive to the starting point of the iteration, and there is no guarantee that the solution is unique. It is also sensitive to collinearity in the observed data, which the method proposed in this paper is not. Since the components are not pre-defined in AR, it can, however, detect compounds that the proposed method will not detect. Again, the objectives of AR and the proposed method differ. While AR tries to model the observations, we aim at fingerprinting based on readily interpretable patterns.

4. Application example

4.1. Introduction

An emerging field of analysis of complex mixtures is comprised by 1st and 2nd generation bio-oils and fuels. These biomass derived liquids can be similar to crude oils in their degree of complexity, due to the variable chemical mixtures obtained and the number of possible combinations of input biomass, process conditions and possible upgrading treatments [30]. Optimisation of process conditions at lab-scale is challenging as many of the common analytical approaches for conventional fossil fuels and their transfer analogues to bio-oil applications cannot be performed on small product volumes. The large number of compounds within the product do however require an analytical methodology that can evaluate the chemical composition as precisely as possible, for example in terms of evaluating the catalytic efficiency of deoxygenation of different relevant chemical species within the oils for fuel blending purposes. Fingerprinting with soft-ionisation techniques has the potential to deliver valuable information as to the reactivity of different compound-classes to such a treatment. A low-resolution analogue to the high-resolution petroleomics approach, developed by Eide and Zahlsen has shown great potential as a low-cost approach to fingerprinting of bio-oils [12]. ESI-MS fingerprinting, backed up by complementary analytics both on chemical and physical property measurement basis, can provide a powerful combination to assess and optimise these bio-oil pioneering processes.

4.2. Sample set

A half-factorial experimental design sample set of bio-oils was produced from lignin-rich waste material in a high temperature and pressure hydrodeoxygenation solvolysis process. The process approach, termed Lignin-to-Liquid (short: LtL), uses *in-situ* hydrogen donation to produce a highly depolymerised and deoxygenated bio-oil from lignin, which may be suitable for fuel-blending [31]. The chosen set of experiments was used to investigate selected critical process parameters in an optimisation approach [13]. The in-situ hydrogenation in this set of experiments was accomplished by thermal degradation of formic acid, which decomposes via two major pathways, the more significant one of which under these conditions yields CO_2 and H_2, the less prominent one CO and H_2O. The sample set comprised $2^{(4-1)}=8$ experiments in addition to two centre points. Variables were the mole ratio of the co-solvents *iso*-propanol/ethanol, the mole ratio of the hydrogen donor formic acid / solvents, the mole ratio of added water/solvents and the variation of temperature between 370 and 390 °C. The amount of lignin-rich biomass was kept constant throughout all experiments. The experimental values used in the experiments are summarised in Table 1. 75 mL high pressure and high temperature non-stirred stainless steel (SS 316) batch reactors of the 4740 series, from Parr Instrument Co. were used to conduct the experiments in a Carbolite LHT oven. The produced oils comprise a wide range of chemical species such as phenols, methoxy-phenols, esters and ketones with aliphatic substituents of varying length. The compositions of the oils span from products rich in saturated hydrocarbons and aliphatic ketones that give spontaneous separation into an oil and an aqueous phase to products rich in phenolic components that give a single phase product dissolved in the reaction solvent medium. The distribution of products between the compound classes has a direct influence on critical physical properties of the oil, such as miscibility or storage stability. It is therefore of relevant interest to investigate their abundances based on a convenient and fast analytical approach.

Table 1
List of conducted experiments with input material amounts and reaction conditions. All reactions were run for approx. 16 h.

Experiment	Formic acid, mmol	Ethanol, mmol	*iso*-Propanol, mmol	Water, mmol	Lignin, g	*T*, °C
F01	65.9	359.2	35.9	4.0	3.75	370
F02	59.2	177.7	177.7	3.6	3.75	390
F03	268.2	243.8	24.4	2.7	3.75	390
F04	249.3	124.6	124.6	2.5	3.75	370
F05	64.3	350.7	35.1	38.6	3.75	390
F06	58.0	173.9	173.9	34.8	3.75	370
F07	263.8	239.8	24.0	26.4	3.75	370
F08	245.5	122.8	122.8	24.6	3.75	390
F09	112.4	224.8	112.4	16.9	3.75	380
F10	112.4	224.8	112.4	16.9	3.75	380

MS fingerprinting gives a higher degree of chemical compound resolution and is a promising approach for this kind of data set. Other analytical approaches using bulk property analytics, such as IR, have also shown to deliver successful results, e.g. in assessing biodegradation levels of petroleum oils [32]. IR analysis, however, was less suitable for the closely related samples from our presented dataset, as dominant (broad) bands from chemical functionalities such as the vibrational —OH band from water or alcohol inclusion concealed other less dominant underlying bands of interest, thus hindering discrimination between the different samples. Initial clustering evaluation of positive ESI-MS analysis showed a strong contribution of several series of equally (14 Da) spaced peaks contributing to the first two loadings of the PCA [13]. The separation of these series, the signals of which ideally originate from molecules of the same individual class and relating these to the applied process parameters, can be a significant step forward to further the understanding and allow fine-tuning of the product spectrum from these complex systems.

In this paper, we are using the existent positive ESI-MS data set as presented by Kleinert et al. for testing of the proposed data analysis and fingerprinting procedure based on 14 Da spacings to evaluate the quality of information extraction [13]. Further information on methods and procedures, such as work-up and data acquisition are to be found in the same reference. Identical identifiers for the single samples are used to enable cross-referencing.

4.3. Results

4.3.1. PCA

For comparison purposes, we performed both PCA and 14 Da basis vectorial analysis of the set of positive ESI-MS data. A visualisation of the scores from the first two principal components, describing a total of 67.5% of the variance (PC1 = 43.5%, PC2 = 24.0%), is given in Fig. 3.

The raw data obtained were processed identically both for the PCA as for the 14 Da methods and some variations to the observations made by Kleinert et al. are thus explainable. Fig. 3 shows that both

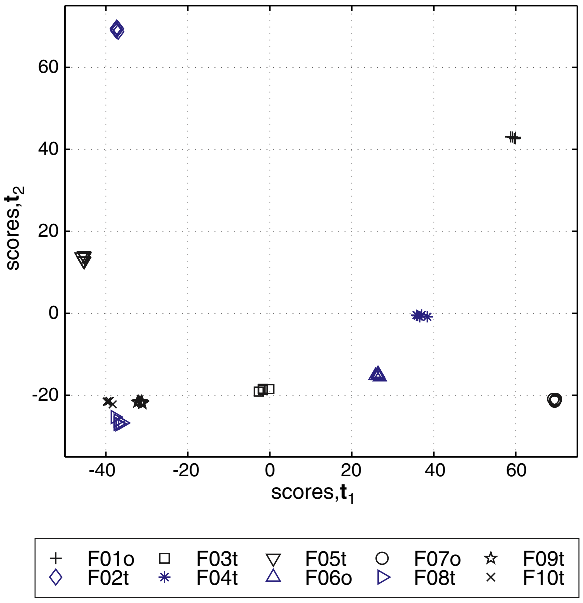

Fig. 3. Score plot of all 10 LtL-oil samples analysed with positive ESI-MS as described by the first two component vectors of the PCA. Five replicate analyses of each sample are included.

the repeatability of analysis, as well as separability between samples, is good, as is indicated by the close clustering of the replicates as well as the centre-point experiments of the experimental design (F09t and F10t). Samples F01o, F06o, F07o and F04t, the “o” describing a one phase product oil, whereas "t" describes the analysis of the organic top phase of a two phase separated (one oil and one aqueous phase) product oil, are plotted in the right half of the plot indicating a difference within the described variance of $\mathbf{t}_1$. This is expected to be due to either a varying composition based on water-soluble components which are less prominent in the organic top phase or suppression of the ionisation of other components due to the existence of more readily ionisable components in the one phase product.

The loading line plots of these first two principal components are given in Fig. 4. Series of 14 Da spaced peaks are easily identifiable within these loadings. These series are connected to different base structures of compounds such as phenols, ketones and esters that have been identified in prior work [14]. The contributions of these multiple series to the loadings are dominant. However, even if identification of single series can be accomplished, the single effects of these on the sample positioning in the scores plot are not easily accessible. This illustrates the limitations of the PCA for isolation of these series. The PCA, being based on illustrating the largest degree of variance in as few components as possible, groups several chemically different compound classes in the same loading vector. Also, a 14 Da spaced sequence stemming from the same compound could end up being distributed over several loading vectors. The loading vectors of the PCA are by design orthogonal, but since no concern is given to underlying chemical structures, interpretation in terms of such patterns is difficult.

4.3.2. 14 Da based analysis

Projecting the measured spectra onto the 14 Da spaced basis vectors, as described in Section 3.2, showed that the first three components account for 71% of the total variation (of the scores). These scores are shown in Fig. 5. The 3D-plot in the top left summarises the clustering which is separately illustrated in the three 2D-subplots in the same figure. Evaluation of the scores of the first three 14 Da basis vectors show that both the clustering quality of replicates as well as the centre-points are largely retained. F08t, F09t and F10t plot closely together in all sub-plots, thus illustrating that there is no larger variance between the first three basis vectors within these three samples. By comparison, F03t and F05t plot closely together on the basis of $\mathbf{t}_1$ and $\mathbf{t}_3$, however are separated on the basis of $\mathbf{t}_2$. This enables a more precise allocation to existent similarities but also variance between these samples, in this case to be found in the 14 Da series, described by $\mathbf{t}_2$. It must again be stressed that the clustering is not expected to be identical to the scores for the PCA. The isolation of the 14 Da series restricts the amount of possible explainable variance to these sets of static signals. This implies that the PCA does have the ability to explain a larger degree of variance within one single component and we do not necessarily expect to be able to compact the multidimensionality of such a complex dataset as is used here to that of the PCA.

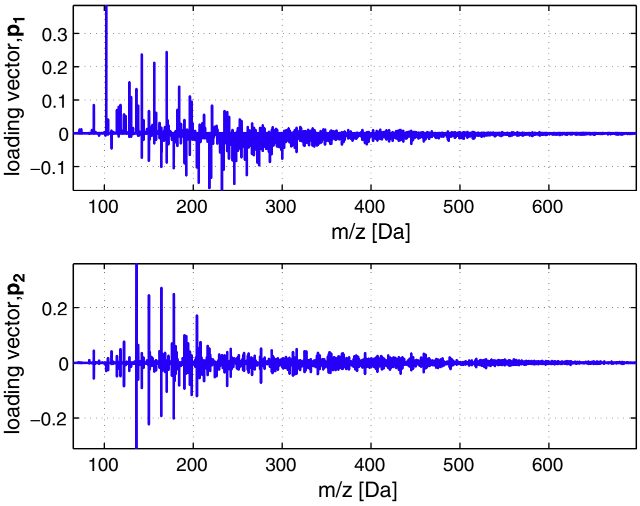

Fig. 4. Line plots of the first two loading vectors from the PCA showing a correlating dominance of 14 Da spaced signal sets.

The 14 Da series basis vectors are given in Fig. 6. These series have already been identified as major contributors to the PCA for this data set [13]. However, they are now directly singularly accessible and their influence on the different samples is visualised.

The first three 14 Da basis vectors, as illustrated in Fig. 6, show even numbered m/z values with a varying number of CH_2 units, illustrated by 14 Da spacings. A difference of −2 Da between the series illustrated by vectors $\mathbf{u}_1$ and $\mathbf{u}_3$ suggests the substitution of two hydrogen atoms for a double bond, retaining the same molecular base-structure. Identification of lower homologues via complementary analysis on GC-MS show a large variation between the samples in respect to non-methoxylated phenols (bulk formula $C_6H_6O + n\ CH_2$, i.e. $94 + n\ 14$) and substituted ketones (bulk formula $Cn\ H_{2n}\ O$, i.e. $58 + n\ 14$), which can also be matched in respect to the positioning of the different samples in the score plots from the 14 Da based analysis in Fig. 5, suggesting these base-structures for the vectors $\mathbf{u}_1$ and $\mathbf{u}_3$. However, the different possibly occurring adduct species of the analysed ions (alkali- and ammonium-ions as well as possibly occurring aromatic clusters), require a more in-depth investigation, increase the uncertainty of allocating the individual series unequivocally.

It is also worth noting that the third 14 Da sequence corresponds to the one readily identified by visual inspection of the raw data in Fig. 1.

5. Discussion

One property of PCA is that the decomposition into components is completely based on minimising the data at hand, and no information about underlying chemical patterns can be taken into consideration. This sometimes makes PCA unsuitable for some types of chemical data sets, where patterns based on some known property are to be expected. The described method exploits the known chemistry in terms of signal spacings and should be seen as a complement, not an alternative, to existing tools, aimed directly at finding specific patterns in the data. In the example presented in this paper, we are looking for sequences spaced by 14 Da, since these spacings are connected to multiple CH_2 groups being appended to the base-compounds. Other experimental variations in the data will not be captured by the algorithm, so in terms of compacting variation into as few components as possible, it is not an alternative to traditional tools. However, in some applications, such as the example presented here, the variation in these 14 Da sequences is an important discriminating property when it comes to fingerprinting of different samples and evaluating the analytical data.

Due to uncertainties in measured data concerning the m/z axis, the exact peak locations may vary slightly. In order to take this into account, the basis vectors were constructed as a sequence of narrow Gaussian shaped peaks instead of discrete peaks. The width of these peaks is a design parameter of the algorithm. Whether or not it is possible to derive some optimal criterion for how to determine this is left for future research. In this paper, the peak width was set so that 95% of the area below the peaks should cover the m/z interval corresponding to an assumed uncertainty of the instrument being used for the specific data set. It is based on these uncertainties that transformations such as Fourier transformation were dismissed within the data pretreatment.

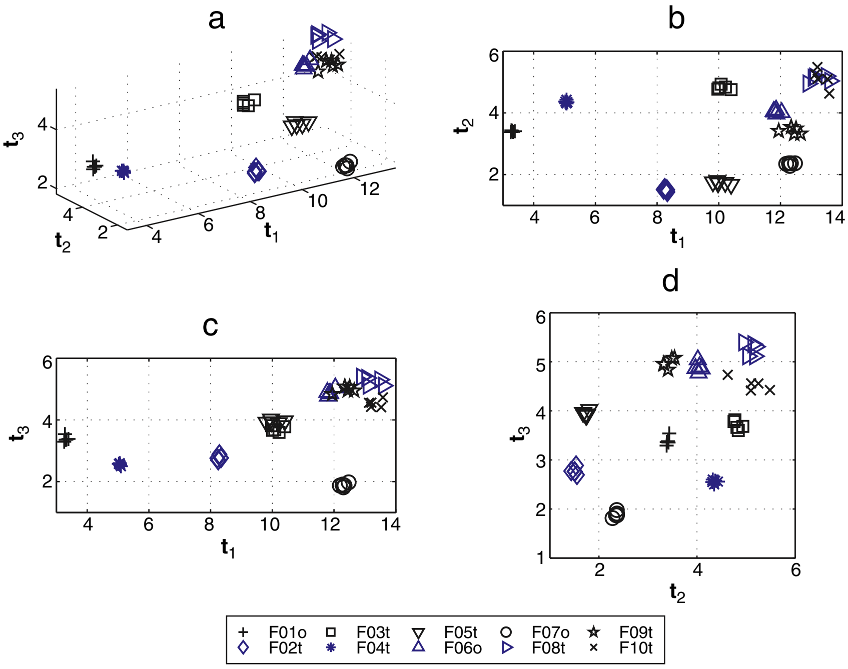

Fig. 5. Score plots for the three most significant 14 Da basis vectors.

A limitation of the algorithm in its current form is that the loading vectors are not localised along the *m*/*z* axis. As a consequence, if several compounds from the same chemical class (spaced by *n* 14 Da) are present in the same sample, these would contribute to the same score. Depending on the data set at hand, this could be a problem, but any subsequent analysis, where the chemist returns to the original data for interpretations would reveal this. For our data-set, by performing analysis of selected oils on high resolution equipment, showing that no further signals than the ones detected on the low resolution equipment were found, we were able to rule out this eventuality. A significant error would be expected when analysing a more complex sample set, e.g. crude petroleum oils, with the same set-up.

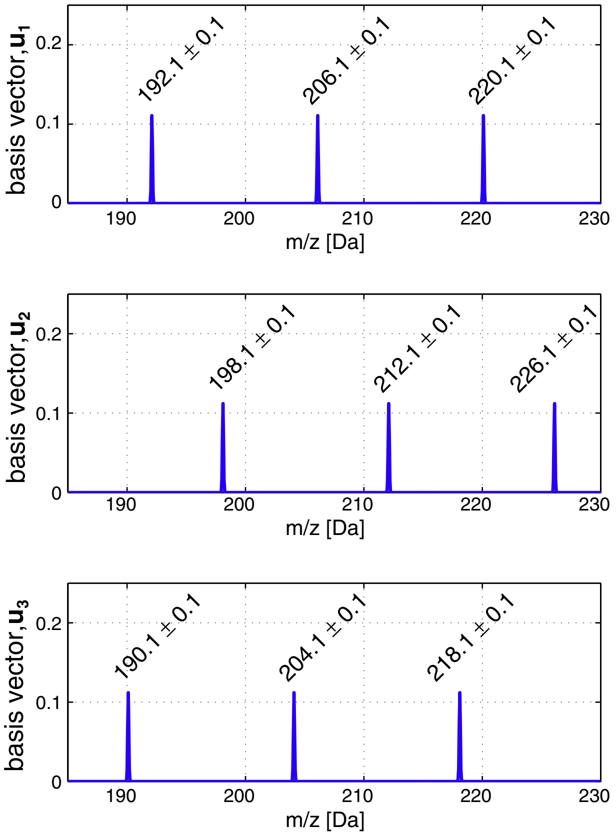

Fig. 6. Top to bottom: parts of the three most significant 14 Da basis vectors, with the corresponding *m*/*z* values for the peak locations.

6. Conclusions

In this paper, we have described a new method for extracting chemically relevant information from mass spectrometry data that can serve as a valuable complement to traditional multivariate tools such as PCA. The proposed method projects measured mass spectra onto a set of basis vectors constructed to represent 14 Da spaced homologous series. The corresponding scores were shown to cluster in a similar way to that of PCA, in the sense that samples with similar chemical properties appear close while others end up more distant. The main difference to other tools is that with the proposed method, the loading vectors are constructed based on existing chemical patterns. As such the interpretation of the results is significantly facilitated in comparison to PCA.

The application example also shows that valuable information can be extracted using relatively few components, although the proposed method does not possess any inherent variation compaction properties. When using PCA, this property is optimised by design. In other words, the proposed method does not aim at maximising the amount of experimental variation described, but rather to reveal patterns that are important for interpreting the results and are otherwise difficult to extract.

Acknowledgements

The authors would like to thank Statoil's VISTA scholarship programme, administered by the Norwegian Academy of Science & Letters, the Michelsen Centre, and Christian Michelsen Research AS for funding of this work. In addition, Gunhild Neverdal is greatly acknowledged for carrying out the ESI-MS measurements at Statoil Research Centre, Trondheim, Norway. The authors would also like to express their sincerest gratitude towards the reviewers of the original manuscript for their constructive comments.

References

[1] I.T. Jolliffe, Principal Component Analysis, 2nd edition Springer Verlag, New York, 2002.

[2] S. Borman, H. Russell, G. Siuzdak, A mass spec timeline, Today's Chemist at Work (September 2003) 47–49.

[3] NIST 11 Mass Spectral Library, National Institute of Standards and Technology, Gaithersburg, MD., USA, 2011.

[4] Wiley Registry of Mass Spectral Data, ninth edition John Wiley & Sons, Inc., New York, USA, 2007.

[5] C. Hughey, R. Rodgers, A. Marshall, Resolution of 11,000 compositionally distinct components in a single electrospray ionization Fourier transform ion cyclotron resonance mass spectrum of crude oil, Analytical Chemistry 17 (2003) 4145–4149.

[6] S. Johnsen, K. Kolset, The mass-selective detector as a chlorine-selective detector, Journal of Chromatography A 438 (1988) 233–242.

[7] P. Jurasek, M. Slimak, M. Kosik, Determination of isotope cluster patterns in mass spectra of GC-MS analyses by a chemometric detector, Mikrochimica Acta 110 (1993) 133–142.

[8] Y. Gu, A fuzzy classification for identification of double bond position in dodecadienic compounds based on mass spectral data, Organic Mass Spectrometry 23 (6) (1988) 487–491.

[9] E. Kendrick, A mass scale based on $CH_2 = 14.0000$ for high resolution mass spectrometry of organic compounds, Analytical Chemistry 35 (13) (1963) 2146–2154.

[10] C. Hughey, C. Hendrickson, R. Rodgers, A. Marshall, Kendrick mass defect spectroscopy: a compact visual analysis for ultrahigh-resolution broadband mass spectra, Analytical Chemistry 73 (2001) 4676–4681.

[11] M. Commodo, I. Fabris, C. Groth, O. Güler, Analysis of aviation fuel thermal oxidative stability by electrospray ionization mass spectrometry (ESI-MS), Energy & Fuels 25 (2011) 2142–2150.

[12] I. Eide, K. Zahlsen, A novel method for chemical fingerprinting of oil and petroleum products based on electrospray mass spectrometry and chemometrics, Energy & Fuels 19 (2005) 964–967.

[13] M. Kleinert, J. Gasson, I. Eide, A.-M. Hilmen, T. Barth, Developing solvolytic conversion of lignin to liquid (LtL) fuel components: optimisation of quality and process economical factors, Cellulose Chemistry and Technology 45 (1–2) (2011) 3–12.

[14] G. Gellerstedt, J. Li, I. Eide, M. Kleinert, T. Barth, Chemical structures present in biofuel obtained from lignin, Energy & Fuels 22 (2008) 4240–4244.

[15] R. Catharino, R. Haddad, L. Cabrini, I. Cunha, A. Sawaya, M. Eberlin, Characterization of vegetable oils by electrospray ionization mass spectrometry fingerprinting: classification, quality, adulteration, and aging, Analytical Chemistry 77 (2005) 7429–7433.

[16] I. Eide, G. Neverdal, B. Thorvaldsen, B. Grung, O. Kvalheim, Toxicological evaluation of complex mixtures by pattern recognition: correlating chemical fingerprints to mutagenicity, Environmental Health Perspectives 110 (Suppl. 6) (2002) 985–988.

[17] T. Cajka, K. Riddellova, M. Tomaniova, J. Hajslova, Ambient mass spectrometry employing a DART ion scource for metabolomic fingerprinting / profiling: a powerful tool for beer origin recognition, Metabolomics 4 (2011) 500–508.

[18] T. Cajka, J. Hajslova, F. Pudil, K. Riddellova, Traceability of honey origin based on volatiles pattern processing by artificial neuronal networks, Journal of Chromatography A 1216 (2009) 1458–1462.

[19] I. Yeo, J. Lee, S. Kim, Application of clustering methods for interpretation of petroleum spectra from negative-mode ESI FT-ICR MS, Bulletin of the Korean Chemical Society 31 (11) (2010) 3151–3155.

[20] J. Möller, R. Catharino, M. Eberlin, Electrospray ionization mass spectrometry fingerprinting of whisky: immediate proof of origin and authenticity, The Analyst 130 (2005) 890–897.

[21] K. Vamurza, Chemometrics in mass spectrometry, International Journal of Mass Spectrometry and Ion Processes 118/119 (1992) 811–823.

[22] S. Wold, O. Christie, Extraction of mass spectral information by a combination of autocorrelation and principal components models, Analytica Chimica Acta 165 (1984) 51–59.

[23] H. Damen, D. Henneberg, B. Weimann, SISCOM—a new library search system for mass spectra, Analytica Chimica Acta 103 (1978) 289–302.

[24] T. Næs, B.-H. Mevik, Understanding the collinearity problem in regression and discriminant analysis, Journal of Chemometrics 15 (2001) 413–426.

[25] J. Wong, C. Durante, H. Cartwright, Specalign—processing and alignment of mass spectra datasets, Bioinformatics 21 (2005) 2088–2090.

[26] W. Yu, B. Wu, N. Lin, K. Stone, K. Williams, H. Zhao, Detecting and aligning peaks in mass spectrometry data with applications to MALDI, Computational Biology and Chemistry 30 (1) (2006) 27–38.

[27] W.W. Hsieh, Nonlinear principal component analysis of noisy data, Neural Networks 20 (2007) 434–443.

[28] B.-W. Lu, L. Pandolfo, Quasi-objective nonlinear principal component analysis, Neural Networks 24 (2011) 159–170.

[29] E.J. Karjalainen, U.P. Karjalainen, Component reconstruction in the primary space of spectra and concentrations. Alternating regression and related direct methods, Analytica Chimica Acta 250 (1991) 169–179.

[30] D. Mohan, Pyrolysis of wood/biomass for bio-oil: a critical review, Energy & Fuels 20 (2006) 848–889.

[31] M. Kleinert, T. Barth, Towards a lignincellulosic biorefinery: direct one-step conversion of lignin to hydrogen-enriched biofuel, Energy & Fuels 22 (2008) 1371–1379.

[32] O. Abbas, C. Refuba, N. Dupuy, A. Permanyer, J. Kister, Assessing petroleum oils biodegradation by chemometric analysis of spectroscopic data, Talanta 75 (2008) 857–871.