Available online at www.sciencedirect.com

ScienceDirect

Journal of Magnetic Resonance 190 (2008) 26–32

JMR
Journal of
Magnetic Resonance

www.elsevier.com/locate/jmr

Quantitative analysis of NMR spectra with chemometrics

H. Winning *, F.H. Larsen, R. Bro, S.B. Engelsen

Quality and Technology, Department of Food Science, Faculty of Life Sciences, University of Copenhagen, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark

Received 11 April 2007; revised 8 October 2007
Available online 14 October 2007

Abstract

The number of applications of chemometrics to series of NMR spectra is rapidly increasing due to an emerging interest for quantitative NMR spectroscopy e.g. in the pharmaceutical and food industries.

This paper gives an analysis of advantages and limitations of applying the two most common chemometric procedures, Principal Component Analysis (PCA) and Multivariate Curve Resolution (MCR), to a designed set of 231 simple alcohol mixture (propanol, butanol and pentanol) ^{1}H 400 MHz spectra. The study clearly demonstrates that the major advantage of chemometrics is the visualisation of larger data structures which adds a new exploratory dimension to NMR research. While robustness and powerful data visualisation and exploration are the main qualities of the PCA method, the study demonstrates that the bilinear MCR method is an even more powerful method for resolving pure component NMR spectra from mixtures when certain conditions are met.

Keywords: Principal component analysis; Multivariate curve resolution; Experimental design; Signal overlap

1. Introduction

NMR is a unique and versatile spectroscopic method capable of measuring samples in the solid, liquid and gas phases. No other spectroscopic method contains equally detailed structural and dynamic information about chemical systems under investigation. However, a serious challenge in NMR spectroscopy lies between the technical capacity to generate data (such as in NMR metabonomics) and the human capacity to interpret and integrate these data [1]. In complex systems such as biofluids, a wide range of components (metabolites, acids, proteins, carbohydrates, etc.) are present with a majority of overlapping resonances distributed over several thousand data points [2]. This amount of data is difficult, if not impossible, to interpret.

The study of more complex systems, such as biofluids is characterised by many hidden relationships. To find these hidden relationships in complex data, experimental design, unsupervised data exploration and data mining techniques are required. Chemometrics is a multivariate data analysis field using statistics to compute models for extracting chemical information from large two-dimensional multivariate data sets. Development of chemometric data models requires a minimum of assumptions and the relationships may be visualised by intuitive illustrations by the graphic computer interface.

We have chosen a ternary model design with three simple linear water soluble alcohols containing different amounts of hydrocarbons with highly overlapping resonances. Using this design we can explore subtle differences in the methylene peak—a simplified simulation of one of the major metabolomic applications of NMR, namely lipoprotein profiling of blood. Besides lipid and lipoprotein resonances, the 0.7–1.5 ppm chemical shift region in blood plasma is characterised by many overlapping signals from small organic species [3]. Spectral assignments in this region have been limited by the extensive chemical shift overlap and by the broadness of the signals. Similar spectral problems may be encountered in organic or pharmaceutical samples when identifying impurities that mimic the compound of interest.

* Corresponding author.
E-mail address: haw@life.ku.dk (H. Winning).

1090-7807/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmr.2007.10.005

The first application of chemometrics to NMR spectra appeared in 1983 by Johnels et al. [4]. In the early nineties Gartland et al. [5] introduced PCA to classify proton NMR spectra of urine. The same group also introduced the research branch *metabonomics* defined as: "understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data" [6] or just "metabolic processes studied by NMR spectroscopy of biofluids" [1]. Impetus for the coupling of NMR spectroscopy with multivariate data analysis was clearly the terribly complex metabolic system in body fluids that gives rise to equally complex NMR spectra. Chemometrics were also applied early on for exploration of solid-state MAS NMR spectra [7]. Now chemometrics is rapidly gaining momentum in the analysis of NMR spectra [8–10] and this work aims to provide an understanding of some of the useful property of basic chemometric tools by their application to a designed set of alcohol mixture NMR spectra.

2. Materials and methods

Principal component analysis and multivariate curve resolution

Principal component analysis (PCA) [11] is the fundamental method in chemometrics. In PCA the data collected on a set of samples is resolved into principal components. The first principal component is defined by the spectral profile (loading) in the data which describes most of the variation, the second principal component is the profile describing the second most of the variation orthogonal to the first, and so on. Later components describe less variation and are more uncertain than the first components, because the systematic variation is primarily described in the first components. Deciding the right number of components is a most important issue and will be described for both PCA and MCR although for PCA, the choice is often less critical especially in exploratory studies because the first and most important component will not change as a function of the number of components chosen. The principal components are composed of so-called scores and loadings. Loadings contain information about the variables (chemical shifts) in the data set and the scores hold information on samples (concentrations) in the data set. For a given principal component, the loading vector is a spectral profile and the score for each sample is the amount of that particular loading in the sample in a least squares sense. Thus, the sum of loadings weighted by a certain sample's score values will provide an approximation of the spectrum of that sample. The similarity to Beers law is apparent as each measured spectrum is hence described by varying amounts of the same few underlying spectral loadings. The individual loadings, though, will mostly not resemble real chemical spectra due to orthogonality constraints of the scores and loadings, but the peaks in a loading are indicative of large spectral variation in the data. Thus, the loadings indicate which parts of the spectrum represent the main variation amongst the samples. The scores, on the other hand, provide information about the extent to which the spectral information represented by the loadings are high or low for particular samples. Hence, the scores can be considered as concentrations of multivariate, so-called *latent*, variables.

As the number of spectral components in a data set is typically much lower than the number of chemical shifts, the whole data set can for the most part be represented by a few (typically much less than 10–20) components that still represent the full chemical variation in the data. The scores are often plotted against each other in a scatter plot giving a 'map' of all the samples in the score plot. Samples that are grouped in a score plot are spectrally similar with respect to the selected principal components. One of the strengths of PCA is to provide a quick unsupervised view of the samples and thereby to identify samples that exhibit deviating features (outliers) or discover trends and groups in the samples. Prior to PCA modelling, data are centred by subtracting from every chemical shift value the average value at that particular shift calculated across all samples.

An alternative method to PCA is decomposition of the data matrix by multivariate curve resolution (MCR) using alternating least-squares (ALS) [12] which has also been named molecular factor analysis (MFA) along with a number of other names [13–15]. Principal component analysis is mostly used for exploration and classification and cannot normally provide direct estimates of real chemical spectra and concentrations because the loadings and scores are constrained to be orthogonal. MCR-ALS on the contrary can offer resolution of the spectra into the 'true' underlying components, i.e. the pure spectra. Huo et al. have proved that multivariate curve resolution was able to provide unique pure spectra and pure decay profiles from DOSY NMR data [16].

An appealing property of MCR is that the solution often looks much more 'chemical' than a PCA solution, because imposed non-negativity constraints make the spectral and sample profiles be positive. This often leads to oversimplifying interpretations where the solution is assumed to be real estimates of chemical spectra and their relative concentrations. However, the MCR solutions are generally not unique, hence the solution can be assumed to be just *one* arbitrary solution out of an infinity of equally well-fitting possible nonnegative solutions. The problem is due to the so-called rotational ambiguity [17] and even though imposing non-negativity helps removing some ambiguity it is not enough to guarantee uniqueness in general. This can only be achieved if the data have certain characteristics such as selective variables where only some analytes are present or samples where some analytes are absent [17].

Hence, for any specific MCR solution, uniqueness must be assessed before the solution can be assumed to be providing estimates of real chemical analytes. Uniqueness can be assessed in different ways, but in this investigation the model was simply restarted several times using different sets of random initial parameters and was verified to provide the same solution. There are other tools for multivariate analysis with a similar aim as MCR, such as direct exponential curve resolution algorithm (DECRA) [18] but these are not applicable to the type of matrix-data (2D) discussed here.

2.1. NMR data

The experimental design is a ternary design of mixtures of the linear alcohols: propanol, butanol and pentanol [13,19]. Each alcohol component (50 mM) has 21 concentration levels in increments of five from 0% to 100%. The samples were prepared from 495 µl of the mixture and 55 µl of D_2O (with 5.8 mM of TSP-d_4 (per-deuterated 3-trimethylsilyl propionate sodium salt) (Fig. 1).

^1H NMR spectra were recorded for each of the 231 mixtures. The spectra were acquired on a Bruker Avance Ultra Shield 400 spectrometer (Bruker Biospin Gmbh, Rheinstetten, Germany) operating at 400.13 MHz using a broad band inverse detection probe head equipped with 5 mm (o.d.) NMR sample tubes. Data were accumulated at 298 K employing a pulse sequence using presaturation of the water resonance during the recycle period followed by a composite 90° pulse [20] with an acquisition time of 4 s, a recycle delay of 20 s, eight scans and a sweep width of 8278.15 Hz, resulting in 64k complex data points. All samples were individually tuned, matched and shimmed. Prior to Fourier transformation, each FID was apodised by Lorentzian line broadening of 0.30 Hz and the corresponding spectra were automatically phased and baseline corrected and referenced to TSP-d_4. In order to secure quantitative measurements the receiver gain was set constant for all the samples.

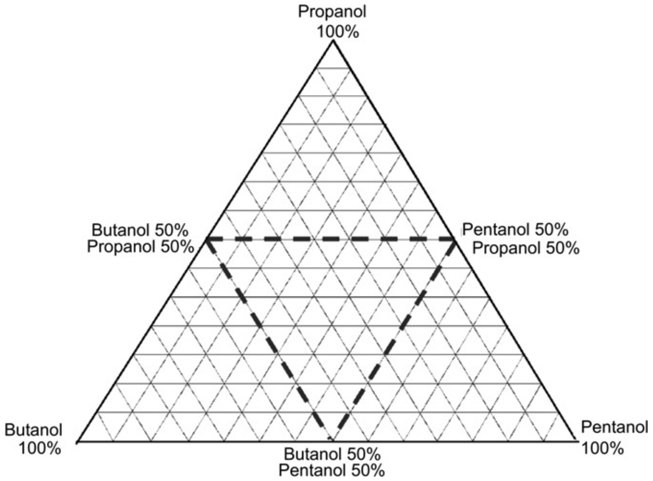

Fig. 1. Tri-axial experimental design of propanol, butanol and pentanol. Each alcohol component has 21 different levels in increments of 5 from 0% to 100%, 231 samples in total. The corners of the triangle represent 100% of the pure alcohol. The triangle with the dashed line shows the reduced experimental design of 66 samples.

Prior to the chemometric analysis the raw proton NMR spectra data matrix to be investigated had the dimensions (231 × 65,536) but was reduced to 14,000 data points (3.85 –0.65 ppm) in order to remove the water signal and make the investigation more efficient. No further pre-processing or alignment of the data such as co-shifting and warping [21] proved necessary.

The NMR spectra of the 231 alcohol mixtures results in only four specific signals (Fig. 2). The spectrum of pure propanol yields a triplet at 0.90 ppm from the CH_3, a quintet at 1.55 ppm from CH_2 and a triplet at 3.57 from the CH_2 next to the OH group. Similar assignments apply to butanol and pentanol, but they also contain aliphatic CH_2s with chemical shift in the range 1.30–1.35 ppm. The spectra of pure propanol, butanol and pentanol are displayed in the bottom of Fig. 2.

3. Results

Using PCA the raw 231 ^1H NMR spectra was decomposed into principal components to describe the systematic variation in the spectra. The data is mean centred prior to PCA, which means that the mean spectrum is subtracted from the individual sample spectra. This simple pre-transformation provides spectra that show the deviation from the average spectrum. PCA results in an almost perfect recovery of the ternary experimental design by the two first PCs, as seen by the score plot in Fig. 3.

Fig. 3 shows the scores and the loadings (of component one and two) of the PCA model where the scores are coloured by the propanol content. The first two components together describe 98% of the variation in the spectra. The 2% of the variation that remains to be explained appears non-systematic, hence due to noise. The loadings of the first two components (PC#1 and PC#2) are displayed in the corresponding loading plot. The first loading describes the overall data structure of the NMR spectra which before mean centering of the spectra are similar to the average spectrum. Upon mean centering, the first loading will change to describe the main variation of all centred data. The fact that the scores are negative is due to the centering as well as the imposed orthogonality that also indirectly causes the loadings to be negative. This clearly illustrates that PCA does not provide estimates of real chemical analytes. However it is also clear that the scores are (linearly) related to the true concentrations (compare Fig. 3, left and Fig. 1), and it is also clear that the loadings reflect the underlying spectral variation.

Apparently only two principal components are necessary to describe all the variation in the spectra, but this is due to the principle of closure specific to these data, i.e. that the concentration of any chemical component in a sample is defined by the remaining two because they add up to 100%. The ternary experimental design is reflected in the score plot which reveals the direct proportional signal intensities with analyte concentrations. Had the data

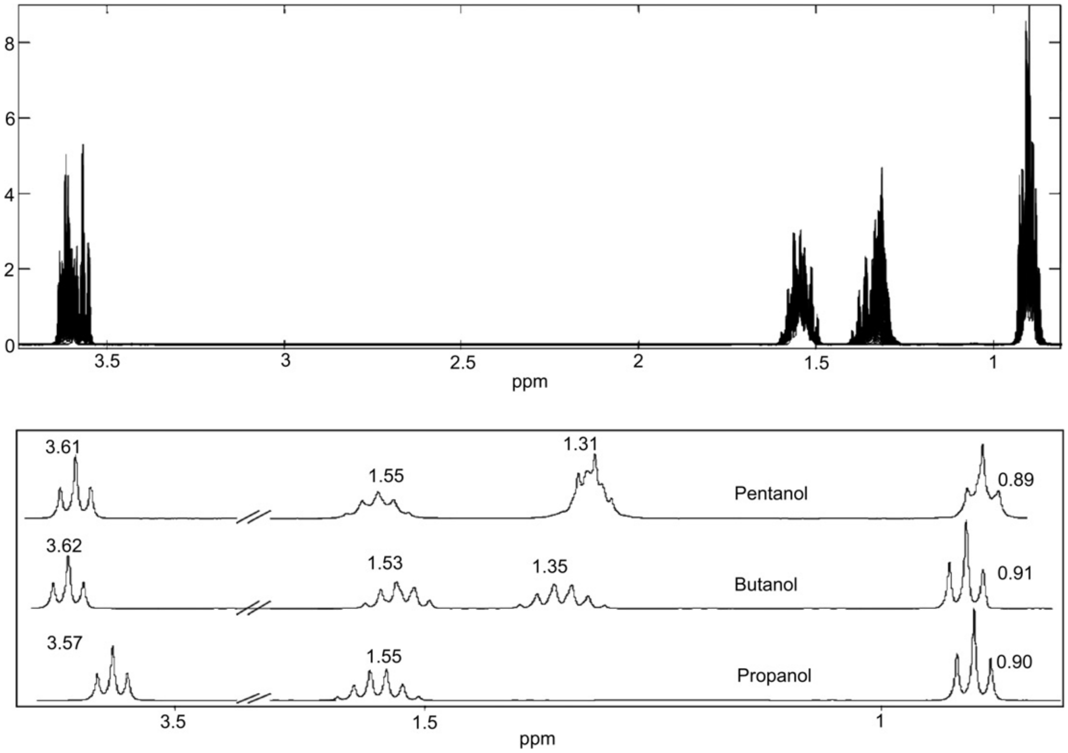

Fig. 2. (Top) NMR spectra of the 231 alcohol mixtures from 3.85 to 0.65 ppm. The NMR spectra of mixtures show highly overlapping signals. (Bottom) The ^{1}H NMR spectra of the pure alcohol samples of propanol, butanol and pentanol.

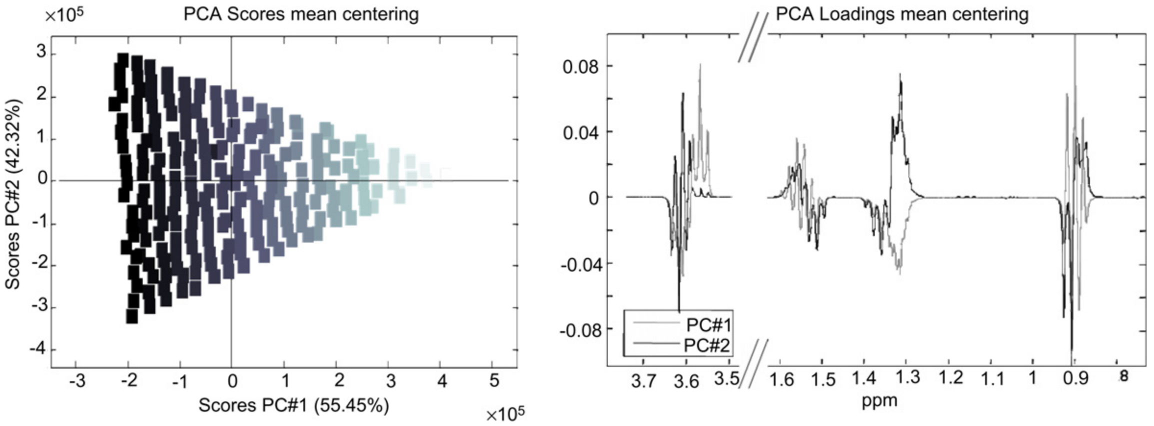

Fig. 3. Scores and loadings plot of the first two principal components from a PCA model calculated on mean-centred NMR spectra. For increased interpretability the score plot is coloured according to the propanol content. The first two principal components explain 97.8% of the variation.

not been centred, three components would be needed to describe the data [22].

In order to pursue the purpose of resolving overlapping resonances from CH_2 and CH_3 groups, the focus of the analysis is restricted to the methyl groups with chemical shifts around 0.9 ppm. As is obvious from Fig. 2, the resonances of the alcohols differ slightly in chemical shift as well as in line width. Our strategy is to perform the chemometric analysis on the restricted data set (0.85–0.95 ppm) representing a region with significant spectral overlap and compare it with results obtained on the full spectrum (3.85–0.65 ppm). The result in Fig. 4 is convincing; the PCA model recovers the ternary experimental design based on spectra of the methyl groups alone. Still, as evidenced by the loading plot, PCA cannot provide real estimates of the pure analyte spectra and concentrations.

MCR is an alternative multivariate data analytical method that can potentially decompose a data set into pure spectra and concentration profiles. The number of components to be extracted can be assessed by looking at the explained variance as a function of number of components, similar to how the number of components is often determined in PCA. Furthermore, visual interpretation of the results is often used as a practical guide in assessing the

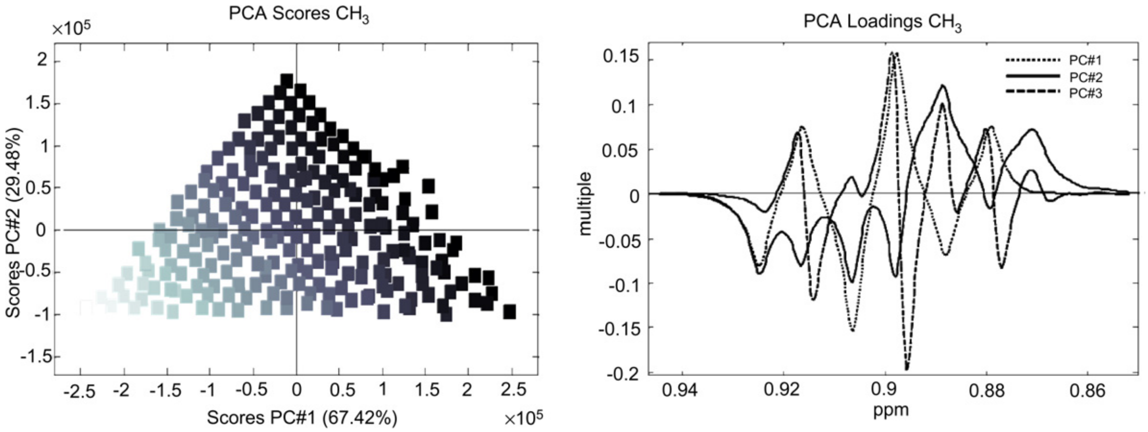

Fig. 4. Scores and loadings plot of the first two principal components from a PCA model calculated on mean-centred NMR spectra (0.85–0.95 ppm). For increased interpretability the score plot is coloured according to the propanol content. The first two principal components explain 96.9% of the variation.

validity of a given model. From Fig. 5 it is obvious that a model with three components is optimal, considering that the variance explained is over 99% and almost remains constant when using more than three components. This is also consistent with the fact that the samples are mixtures of three analytes. That 99% variance explained is adequate can be further assessed and validated by comparing with the intrinsic noise in the data (not shown).

The MCR model with three components is calculated without mean centering the data as is usual in MCR. Looking at a scatter plot of the scores from the MCR model, the triangle now shows perfect concentrations (Fig. 6). The slight non-ideality observed in the triangle, can be attributed to noise in the spectra and small uncertainties in the alcohol concentration. Non-negativity of estimated concentrations and spectra is imposed as part of the model.

The loadings from MCR shown in Fig. 6 resemble spectra of each of the pure alcohol compounds. By repeating the estimation of the MCR model many times from different random starting points, it is verified that the same fit and solution is obtained (results not shown). Hence, the solution can be assumed to be unique. Like PCA, each sample has a score for each of the loadings. The score is simply the amount of the corresponding loading, and as the loadings can be considered as estimates of real spectra, then the scores are then (relative) estimates of the concentrations. These scores are compared with the 'true' value (i.e. the concentration of the three alcohols) by plotting them against each other, yielding three correlation coefficients higher than 0.99.

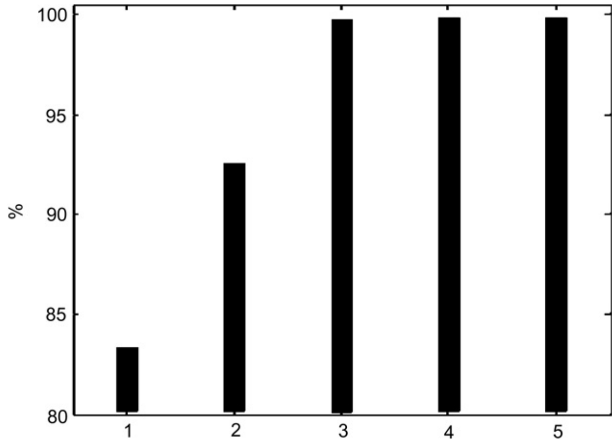

Fig. 5. The bar plot show the percentage explained variance as a function of number of components. Using three components, 99.7% of the variation is explained.

The results above are encouraging and imply that complex mixture NMR spectra can be separated mathematically into the underlying constituents. However, the main reason that the results are as good as they are is the presence of pure samples in the sample set. The presence of pure samples adds selectivity in the data. Selectivity means samples or variables for which only one analyte is present. This is one of the key requisites for obtaining uniqueness in MCR.

To demonstrate how well MCR can model more complex data, a model is calculated on a reduced experimental design of only 66 samples where no spectra of pure alcohols exist. The simplest samples in this reduced design consist of mixtures of at least two of the alcohols. The result is that the triangle of the experimental design is fully recovered and the three concentration profiles still yield correlations over 0.99 to the true concentrations.

However, the loading are not as perfectly resolved as in the full design, which is due to the overlap of the signals from the alcohols. This is particularly apparent in the spurious propanol peak in Fig. 7 at 0.8 ppm. Apparently the MCR model determined on this dataset is unique. Repeating the model estimations more than 1000 times from different starting points all lead in 74% of the cases to the same local solution which is spectrally correct (Fig. 7, left). However in 19% of the cases the global model, which explains almost the same variance but is spectrally incorrect, is the result (Fig. 7, right). This result represent the Achilles' heal of MCR when applied to unknown systems. The best model may not be the correct one in the physical sense.

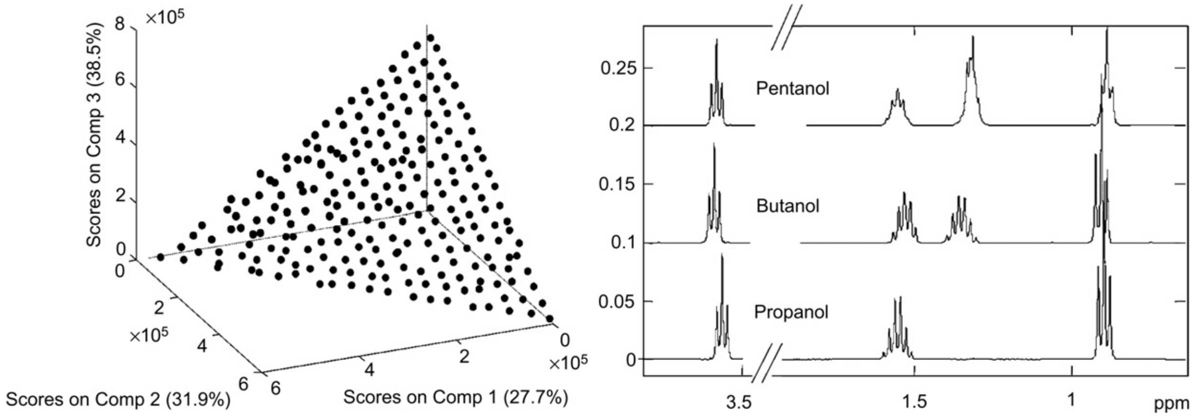

Fig. 6. Scores of the first three components (left) and loading plot (right) of the three components from the MCR model, obtained on the NMR spectra. The two components explain 98.1% of the variation. The score plot shows a perfectly equilateral triangle.

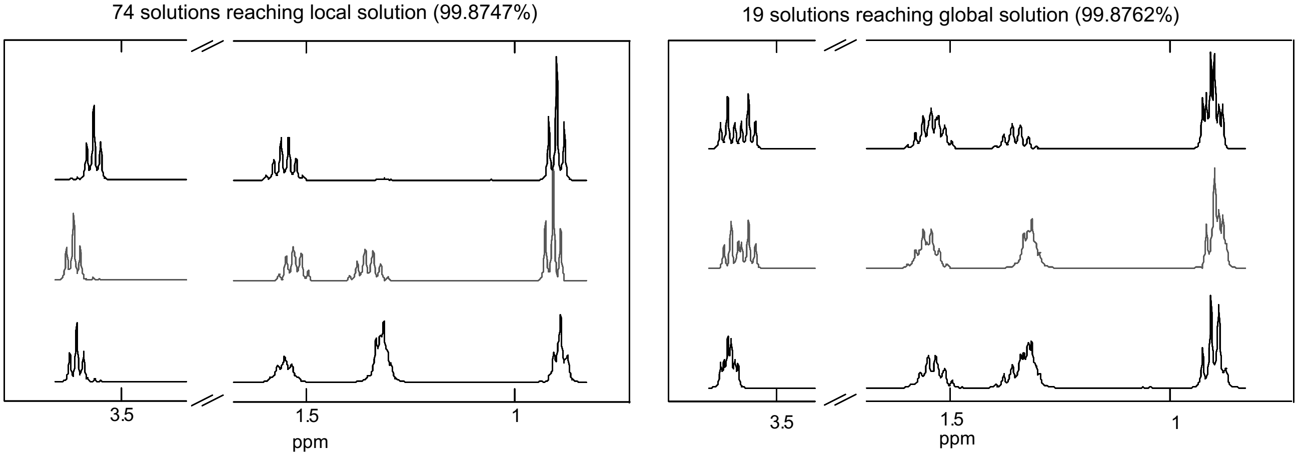

Fig. 7. Plot to the left show the loading of the three components from the local MCR model. This local solution is spectrally correct and is the result in 74% of the cases. The loading plot to the right is obtained from the global model which is the result in 19% of the cases. These loadings are mixtures of the loadings of the three pure alcohols.

4. Conclusions

The main objective of this work was to show how principal component analysis and multivariate curve resolution can be useful in the investigation of highly overlapping data from NMR studies. While it has been shown that PCA can be used to provide a comprehensive overview of complex data with many variables, it was also shown that there are some limits on the usefulness of PCA. The MCR method was demonstrated to possess the powerful ability to separate mixtures into pure spectra and concentrations even for much reduced designs. By applying the basic chemometric methods to a well defined ternary experimental design of ^{1}H NMR spectra the potential and characteristics of chemometric multivariate data analysis were demonstrated. It should be obvious that perhaps the greatest advantage of chemometrics is the simplicity by which even large data structures are analysed and visualised and thereby adding an exploratory dimension to modern NMR science. We have shown which results can be expected when applying quantitative chemometric methods to multivariate high resolution NMR data and our future research will focus on how MCR can perform on more complex metabonomic data.

Acknowledgments

The Ministry of Food, Agriculture and Fisheries is greatly acknowledged for financial support to the ASYNCRON project. Gilda Kischinovsky is acknowledged for proof reading.

References

[1] J.C. Lindon, J.K. Nicholson, E. Holmes, J.R. Everett, Metabonomics: metabolic processes studied by NMR spectroscopy of biofluids, Concepts Magn. Reson. 12 (2000) 289–320.

[2] J.L. Griffin, The potential of metabonomics in drug safety and toxicology, Drug Discov. Today Technol. 1 (2004) 285–293.

[3] J.D. Bell, P.J. Sadler, A.F. Macleod, P.R. Turner, A. Laville, ^{1}H-NMR studies of human–blood plasma assignment of resonances for lipoproteins, FEBS Lett. 219 (1987) 239–243.

[4] D. Johnels, U. Edlund, H. Grahn, S. Hellberg, M. Sjostrom, S. Wold, S. Clementi, W.J. Dunn, Clustering of aryl C-13 nuclear magnetic-resonance substituent chemical-shifts—a multivariate data-analysis using principal components, J. Chem. Soc., Perkin Trans. 2 (1983) 863–871.

[5] K.P.R. Gartland, C.R. Beddell, J.C. Lindon, J.K. Nicholson, Application of pattern-recognition methods to the analysis and classification of toxicological data derived from proton nuclear-magnetic-resonance spectroscopy of urine, Mol. Pharmacol. 39 (1991) 629–642.

[6] J.K. Nicholson, J.C. Lindon, E. Holmes, 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data, Xenobiotica 29 (1999) 1181–1189.

[7] J.M. Koons, P.D. Ellis, Applicability of factor-analysis in solid-state NMR, Anal. Chem. 67 (1995) 4309–4315.

[8] H.C. Keun, T.M.D. Ebbels, H. Antti, M.E. Bollard, O. Beckonert, E. Holmes, J.C. Lindon, J.K. Nicholson, Improved analysis of multivariate data by variable stability scaling: application to NMR-based metabolic profiling, Anal. Chim. Acta 490 (2003) 265–276.

[9] S. Halouska, R. Powers, Negative impact of noise on the principal component analysis of NMR data, J. Magn. Reson. 178 (2006) 88–95.

[10] R. Stoyanova, T.R. Brown, NMR spectral quantitation by principal component analysis, NMR Biomed. 14 (2001) 271–277.

[11] H. Hotelling, Analysis of a complex of statistical variables into principal components, J. Educ. Psychol. 24 (1933) 417–441.

[12] R. Tauler, Multivariate curve resolution applied to second order data, Chemometrics Intel. Lab. Syst. 30 (1995) 133–146.

[13] M.K. Alam, T.M. Alam, Multivariate analysis and quantitation of O-17-nuclear magnetic resonance in primary alcohol mixtures, Spectrochim. Acta A-Mol. Biomol. Spectrosc. 56 (2000) 729–738.

[14] C.D. Eads, C.M. Furnish, I. Noda, K.D. Juhlin, D.A. Cooper, S.W. Morrall, Molecular factor analysis applied to collections of NMR spectra, Anal. Chem. 76 (2004) 1982–1990.

[15] T.M. Alam, M.K. Alam, Chemometric analysis of NMR spectroscopy data: a review, Ann. Rep. NMR Spectrosc. 54 (2005) 41–80.

[16] R. Huo, C. Geurts, J. Brands, R. Wehrens, L.M.C. Buydens, Real-life applications of the MULVADO software package for processing DOSY NMR data, Magn. Reson. Chem. 44 (2006) 110–117.

[17] R. Manne, On the resolution problem in hyphenated chromatography, Chemometrics Intel. Lab. Syst. 27 (1995) 89–94.

[18] W. Windig, B. Antalek, L.J. Sorriero, S. Bijlsma, D.J. Louwerse, A.K. Smilde, Applications and new developments of the direct exponential curve resolution algorithm (DECRA). Examples of spectra and magnetic resonance images, J. Chemometrics 13 (1999) 95–110.

[19] T. Næs, T. Isaksson, The chemometric space: Interpreting principal components in NIR spectroscopy, NIR News 3 (1992) 7.

[20] A. Bax, A spatially selective composite 90-degrees radiofrequency pulse, J. Magn. Reson. 65 (1985) 142–145.

[21] F.H. Larsen, F. Berg, S.B. Engelsen, An exploratory chemometric study of ^{1}H NMR spectra of table wines, J. Chemometrics 20 (2006) 198–208.

[22] R. Bro, A.K. Smilde, Centering and scaling in component analysis, J. Chemometrics 17 (2003) 16–33.