Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Magnetic Resonance 169 (2004) 257–269
JMR Journal of Magnetic Resonance
www.elsevier.com/locate/jmr

Improved DOSY NMR data processing by data enhancement and combination of multivariate curve resolution with non-linear least square fitting

R. Huo, R. Wehrens, and L.M.C. Buydens*

Laboratory of Analytical Chemistry, University of Nijmegen, Toenooiveld 1, 6525 ED Nijmegen, The Netherlands

Received 23 February 2004; revised 27 April 2004
Available online 5 June 2004

Abstract

The quality of DOSY NMR data can be improved by careful pre-processing techniques. Baseline drift, peak shift, and phase shift commonly exist in real-world DOSY NMR data. These phenomena seriously hinder the data analysis and should be removed as much as possible. In this paper, a series of preprocessing operations are proposed so that the subsequent multivariate curve resolution can yield optimal results. First, the baseline is corrected according to a method by Golotvin and Williams. Next, frequency and phase shift are removed by a new combination of reference deconvolution (FIDDLE), and a method presented by Witjes et al. that can correct several spectra simultaneously. The corrected data are analysed by the combination of multivariate curve resolution with non-linear least square regression (MCR–NLR). The MCR–NLR method turns out to be more robust and leads to better resolution of the pure components than classic MCR.

Keywords: DOSY NMR; Reference deconvolution; Phase and frequency shift; Baseline correction; Multivariate curve resolution; Non-linear least square regression

1. Introduction

DOSY NMR data are obtained by a 2-D NMR experiment with the pulse field gradient (PFG) employed [1]. The original data consist of a series of NMR spectra, in which the intensities attenuate with the increase of the gradient strengths. The intensities of a specific component follow an exponential decay, depending on its diffusion coefficient. The DOSY NMR experiment results in a 2-D spectrum, displaying chemical shifts on one axis and the calculated diffusion coefficients on the other. It can be used as a qualitative method to identify the molecular components in a mixture and simultaneously obtain the physical properties of the system such as size, structure, and so on [2–5].

A DOSY NMR data set is a summation of several diffusion components and forms a bilinear data matrix. It is the bilinear characteristic that makes it possible to identify the pure components in a DOSY NMR data set by multivariate curve resolution (MCR) [6]. Previous research has discussed the difficulty in analysing DOSY data by regular single channel methods and it has been revealed that MCR can be a relatively general way to deal with DOSY NMR data [7]. To obtain reasonably good results from MCR, the baseline, frequencies, and phases of each spectrum need to be more or less consistent. However, this is often not the case. In a series of NMR spectra, the baseline offset, the position (frequency) and the phase of the corresponding resonance peaks are almost never identical due to experimental variations. This can significantly affect the performance of MCR to find the pure components. Consequently, DOSY NMR data need to be corrected to minimise the baseline drift, frequency shift, and phase shift in order to gain the best results from MCR. This paper proposes a strategy to do so.

* Corresponding author. Fax: +31-24-3652653.
E-mail address: lbuydens@sci.kun.nl (L.M.C. Buydens).

1090-7807/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmr.2004.04.019

The baseline offset is usually recognised by polynomial regression of a line through the baseline regions, which are free of resonance peaks. The baseline is then corrected by subtracting the constructed polynomial regression line from a NMR spectrum. This is a routine method for baseline correction. However, polynomial regression is not able to deal with the large baseline distortion in different regions of the spectrum. Golotvin and Williams [8] proposed a novel method to deal with this problem. The first step is to recognise whether the points in a spectrum are in the baseline and the second step is to model a baseline by using the smoothed spectrum. This technique is simple, and it can remove even severe baseline distortions effectively. Therefore, it is used as a baseline correction routine to remove the baseline offset of the NMR spectra in a DOSY data set in this paper.

Besides the baseline shift, frequency shift and phase shift also exist in the spectra of a DOSY NMR data set. The well-known method, named FIDDLE (free induction decay deconvolution for lineshape enhancement), also called reference deconvolution, is usually used to enhance NMR signals [9,10]. The principle behind it is to select a single peak as reference and define the desired lineshape of that peak. The difference between the original peak and the desired peak transfers the whole spectrum to an improved form. However, FIDDLE is designed to improve a 1-D NMR spectrum. In a DOSY NMR data set, there are at least 16 spectra (usually 32, and sometimes 64 spectra). It is very cumbersome to define the optimal lineshape and position for the reference peak in each spectrum. Witjes et al. [11,12] propose an automatic method to align all the peaks to the same position and with the same line shape of different spectra, which is an improved version of the Brown and Stoyanova method [13]. The procedure is automatic and quick since the user does not have to estimate the ideal lineshape one by one for each spectrum. The drawback of this method is that it can only perform the peak alignment for a single peak each time and cause peak distortion of overlapping peaks. Moreover, the peak-by-peak correction can give rise to discontinuities in the baseline. In this paper, a new combination of the two methods described above is proposed to minimise the peak shape and position problems, while no additional artefacts are introduced. The correction procedure mainly follows the FIDDLE method and the desired lineshape is obtained by Witjes method. It is shown that the combination of the two methods, together with the baseline correction techniques, improves the quality of a DOSY NMR data significantly.

Once the quality of DOSY NMR data is improved, it can be analysed by multivariate methods, like MCR. Previous research has indicated that MCR with a good initial estimation decay profile is a general method that can provide reasonably well-resolved spectra and decay profiles [7]. In this paper, the initial guess of MCR is obtained by orthogonal projection algorithm (OPA) [14–16] because it is more easily interpreted and implemented. However, even for a data set with relatively good quality, the classic MCR will still have difficulties in the data separation if there are overlapping regions and the diffusion coefficients are similar. One solution is to explicitly force the decay profiles to follow an exponential curve. This is often called hard modelling, since a parametric model is kept fixed, and only values for the parameters (in this case diffusion coefficients) are estimated for the data. The combination of hard modelling steps with soft modelling to improve the performance of MCR has been reported by many papers. For example, Bezemer and Rutan [17,18] incorporated MCR (soft modelling) with non-linear fitting of a kinetic model (hard modelling) into it. Bijlsma et al. [19] proposed to combine MCR with Levenberg–Marquardt algorithm [20] to estimate reaction rate constants from UV–vis spectroscopic data. The Levenberg–Marquardt algorithm is used in each iteration to update the decay profiles so as to reduce the ambiguity problem in MCR. It is shown that the MCR–NLR algorithm with non-negativity constraints is more robust and flexible than the classic MCR algorithm to analyse DOSY NMR data.

2. Theory

2.1. Data preprocessing

The goal of data preprocessing is to remove effects that will deteriorate the subsequent multivariate analysis. We propose a two-step strategy: first correct for baseline drifts and secondly remove frequency and phase shift.

2.1.1. Baseline correction

The baseline correction method applied here was proposed by Golotvin and Williams [8] and includes two steps. The first step is to recognise the baseline. This is done by placing the ith point in the centre of a rectangle window with a width of N spectral points. For each window the standard deviation is calculated and the smallest value is taken as the noise standard deviation (σ_{noise}). If the difference of the minimal and maximal values in the ith window is less than the noise standard deviation multiplied a predefined value n, then the ith point is considered to be in the baseline as described in the following equation:

$$(y_i^{\max} - y_i^{\min}) < n\sigma_{\text{noise}}. \quad (1)$$

The parameters, the window width N and the noise multiplier n, can be chosen to adapt to different data

sets. The second step is to calculate the smoothed spectrum by a moving average method. The baseline is defined by replacing the points in the spectrum found in Eq. (1) by the value of the smoothed spectrum and connecting the baseline fragments by straight lines. Finally, the calculated baseline is subtracted from the original spectrum. This baseline correction method is automatic and successfully removes most baseline distortion in DOSY data.

2.1.2. Phase and frequency shifts correction

As already mentioned, this paper proposes a combination of FIDDLE (also called reference deconvolution) [9,10] and a method presented by Witjes et al. [11,12].

The procedures of FIDDLE are performed in the time domain. The experimental time domain is transferred to frequency domain by the Fourier transform. The reference peak is extracted by replacing other peaks in the spectrum with noise and transformed back to the time domain by the inverse Fourier transform. An ideal Lorentzian peak shape of the reference peak is defined and also transformed to time domain. The corrected signal is obtained by dividing the original signal in time domain by the reference time domain signal and then multiplying the desired estimation of the reference peak. Finally the corrected NMR spectrum is obtained by a Fourier transformation of the resulted time domain signals. This is a classic method for lineshape enhancement of a 1-D NMR spectrum. The drawback of FIDDLE, especially for the application of the DOSY NMR data, is that the ideal lineshape of the reference peak needs to be estimated for every spectrum. It is not possible to have the same estimate of the reference peak in every spectrum because the intensities in different spectra are not the same, i.e., they attenuate exponentially.

The phase and frequency correction method proposed by Witjes et al. is designed to deal with the phase and frequency shift problems for a series of spectra. It is a method based on principal component analysis (PCA). All instances of a particular peak in the series of spectra are analysed by PCA to obtain a PC1 spectrum which can be regarded as the average spectrum. Each spectrum can be approximately represented by a linear combination of the real and imaginary values of PC1 and one or more of their derivatives. The information of phase and frequency shifts is contained in the regression coefficients, which can be obtained by using classic least squares. Finally the phase correction is done in the frequency domain while the frequency shift is corrected in the time domain. The whole procedure is repeated until the shifts reach insignificantly small values. As a result, every peak is aligned in the same frequency position and phase value as the PC1 spectrum. This method is simple and automatic. It does not need the estimation of lineshape function one by one spectra. The shortcoming is that it can only perform the alignment to a single peak, one by one. This can give rise to discontinuities on the baseline. Moreover, if it is applied to multi-peak correction, peak distortion can be caused (see below).

Considering the Witjes method is able to correct the peak shifts and phase shifts of single peaks in a series of spectra, while FIDDLE can correct the peaks of a spectrum simultaneously, we propose to combine those two methods to preprocess the original data set. The basic procedures are based on FIDDLE. The main difference lies in the step to obtain the desired lineshape. Instead of estimating it spectrum by spectrum, the desired lineshape of the reference peaks in the spectra can be obtained by the Witjes method at once. The reference peaks should have the same peak shape and the same frequency. The procedures of the combination method are summarised in Fig. 1.

When this method is used, it is assumed that the peaks in the same spectrum have almost the same global phase and frequency shift. This is usually the case in the DOSY NMR data. The reference peak should be a single peak and contain signals from the first spectrum to the last one. The combined method then can correct the global shifts for the whole data set and therefore decrease the experimental artefacts significantly without introducing new artefacts such as discontinuities and new distortion.

2.2. Data analysis by MCR–NLR

2.2.1. Finding the initial guess

Several methods have been proposed to find good initial guesses for MCR. In [7] IPCA is applied for this purpose. In this paper, the orthogonal projection approach (OPA), adapted to the application of DOSY NMR data, is employed instead because OPA is more easily implemented and interpreted than IPCA while the resulted pure variables are nearly the same or even better. Orthogonal projection approach (OPA) is a stepwise procedure based on orthogonalisation method [14,15]. It was initially used to check peak purity in a chromatogram. Later OPA was employed to find purest spectra in HPLC-DAD data and then trigger MCR-ALS to explore pure components in a mixture [16]. In each step of OPA, dissimilarities between all spectra and reference spectrum are calculated and the spectrum that has the highest dissimilarity is selected as the pure spectrum. The stepwise procedures continue until the dissimilarity spectra show random structure. In the case of the DOSY NMR data, a pure spectrum is only possibly present in the last few spectra because the intensities of some components vanish with the increase of the applied gradient. Therefore, it is logical to search pure variables on the frequency (chemical shift) dimension, i.e., search purest decay profiles. Before actually running OPA, the

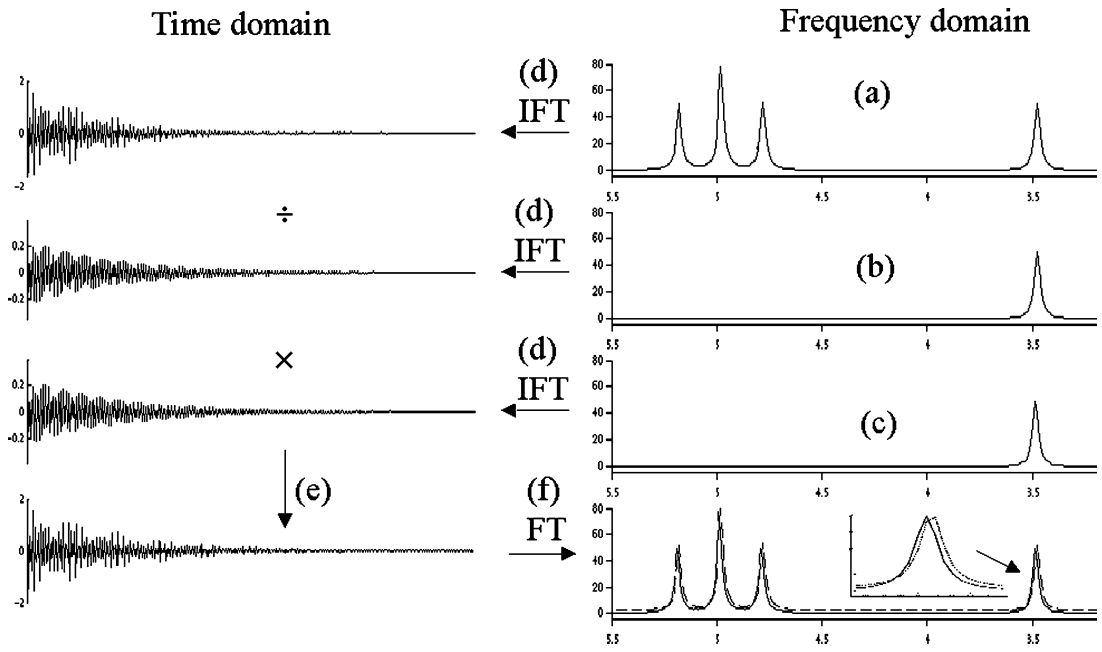

Fig. 1. The combination of data enhancement procedures. (a) Baseline correction of the original data set; (b) extract the reference peak by substituting other peaks with backgrounds noises; (c) using Witjes method to align the reference peaks; (d) inverse Fourier transform (IFT) of the corresponding spectra into time domain; (e) calculation in time domain to obtain the corrected signal; (f) Fourier transform (FT) into the corrected spectra (dashed line in zoomed-in image is the original spectrum).

data set needs to be pre-treated by selective normalisation [14], i.e., variables with mean values greater than a predefined threshold are normalised to unit length while other variables remain unchanged. Selectively normalising variables can eliminate the effects of the intensities. As a result, the amount of dissimilarity only depends on the angle between an individual vector and reference vectors. In the experimental data, it is difficult to obtain random dissimilarity spectrum even with a high number of pure variables. Hence, it is recommended to search more pure variables than the real number of components in a mixture and then plot the pure variables with the unit length to 1, i.e., normalised pure variables. If some pure variables show the same decay curves, then it can be considered that they account for the same component and hence only one of those pure variables is used to represent this component in the initial guess matrix.

In some cases there is no pure variable available for some of or all of the components in the original data set. Consequently, only using OPA is not able to find all the pure variables. Windig et al. [21–23] developed a method combining original with second-derivative data to solve this problem. Second-derivative data are obtained by Savitzky–Golay method. The pure variables are then searched in the conventional data and second-derivative data sequentially by OPA. The resulted pure variables are normalised to unit length of 1 and plotted in the same figure. The number of pure components is determined visually by the number of the different decay profiles in this plot.

2.2.2. MCR-ALS combined with non-linear least square regression

Generally, MCR-ALS with non-negativity constraints is capable of resolving pure spectra and decay profiles, provided a good initial guess is present. However, for a mixture that contains many components and hence may have many overlapping peaks, there is more ambiguity. What is more, even small experimental artefacts can also affect the performance of MCR to such an extent that it is difficult to obtain unique solutions to the separation of pure components. Fortunately, MCR is very flexible so that extra constraints can be applied. Because the signals of DOSY NMR spectra attenuate exponentially with the increase of gradient levels, non-linear least square regression on decay profiles for each iteration of ALS can be used to reduce the rotation ambiguity of MCR. This can be done by the Levenberg–Marquardt algorithm with a pre-defined exponential function [19,20].

The attenuation of signal of each component in DOSY NMR measurement is described by Eq. (2) [24]. In the exponential part of Eq. (2), $D(n)$ is diffusion coefficient of the nth component (m^2/s). δ is the duration of gradient pulses (s) and Δ is the diffusion time (s), both of which are experiment constants set by the user. K is multiplication of γ, the gyromagnetic ratio of the ^1H nucleus (rad s^{-1} T^{-1}), g, the gradient strength (T), and δ. $I_0(n)$ is the intensity before gradient strength g is employed. For a system with N components, the measured spectra are the sum of the intensities of the components and it can be expressed by the following equation:

$$I(n,g^2) = I_0(n)\exp[-D(n)(\Delta - \delta/3)K^2], \quad (2)$$

$$K = \gamma g \delta,$$

$$I(g^2) = \sum_{n=1}^{N} I(n,g^2). \quad (3)$$

According to Eqs. (2) and (3), a DOSY NMR data set of mixture is actually a bilinear data matrix as described in Eq. (4):

$$I = C \cdot S^{\mathrm{T}}, \quad (4)$$

where C contains N column vectors, every of which accounts for a pure decay profile of a component, and S matrix contains the pure spectra, i.e.,

for nth column in C,

$$C(n) = \exp[-D(n)(\Delta - \delta/3)K^2] \quad (5)$$

and the nth row in S, $\quad S(n) = I_0(n). \quad (6)$

Therefore, MCR-ALS can be applied to resolved DOSY NMR data. Firstly, pure variables are selected by OPA and placed in matrix C. Then the corresponding spectra are calculated by least square regression as Eq. (7). A new set of decay profiles is obtained from the calculated spectra according to Eq. (8). The non-negativity constraints are applied to Eqs. (7) and (8). This alternating procedure proceeds until the residuals reach to a predefined convergence criterion:

$$S = I^{\mathrm{T}} \cdot C \cdot (C^{\mathrm{T}} \cdot C)^{-1}, \quad (7)$$

$$C = I \cdot S \cdot (S^{\mathrm{T}} \cdot S)^{-1}. \quad (8)$$

Before starting a new iteration, Leverberg–Marquardt algorithm is applied to update the decay profiles C with an exponential function based on Eq. (5). By a transformation of natural logarithm, Eq. (5) is actually a linear regression problem with the increase K^2 or g^2. Polynomial fitting of the transformed equation can be used to obtain the initial estimation of parameters of Leverberg–Marquardt algorithm. The advantage of combining hard and soft modelling is that it can reduce the ambiguities of MCR and hence make the model more robust. In addition, the non-negativity constraints are also employed to obtain chemically and physically meaningful pure spectra and decay profiles. Finally, the relative root of sum of squared differences (RRSSQ) is used to assess the similarity between the reconstructed data and the original data:

$$\%\mathrm{RRSSQ} = 100 \times \sqrt{\frac{\sum (I_{\mathrm{reconstructed}} - I_{\mathrm{original}})^2}{\sum (I_{\mathrm{original}})^2}}. \quad (9)$$

3. Experimental

3.1. Simulated data

One simulated data are constructed to examine the combination of FIDDLE and the Witjes method. This data set has been used in [7]. It contains three components with the diffusion coefficients of 5.0×10^{-7}, 1.0×10^{-6}, and 1.0×10^{-7} cm^2/s. Thirty-two gradient levels from 64×104–1.9321×108 are employed. The two experimental constants Δ and δ are 100 and 5 ms, respectively. Additionally, in each spectrum of the data set, the peaks contain a frequency shift of -0.1 to 0.1 data point and a phase shift of -0.5 to 0.5°. This is a global shift for each spectrum, which means the peaks in the same spectrum have the same frequency and phase shifts within the range. To make the data more realistic, completely random small shifts (1% of the global shift) are added to the data as well. In addition, the data also contain normally distributed noise with a standard deviation of 0.035% of the highest peak intensity.

A second simulated data set contains four diffusion components whose diffusion coefficients are 5.0×10^{-7}, 1.0×10^{-7}, 2.0×10^{-7}, and 0.8×10^{-7}. It has the same gradients levels and experimental parameters as the one described above and also contains noise with a standard deviation of 0.035% of the highest peak intensity. It is supposed that this data set has been preprocessed by the methods mentioned above. Thus, there is only a small amount of frequency and phase shift, i.e., ± 0.02 data point and ± 0.01°, respectively. This is a more complex data set in which there are more overlapping peaks and the diffusion coefficients of the components are closer to each other. It is used to examine the difference between the classic MCR and MCR–NLR.

3.2. Experimental data

The mixture is made by Océ-Technologies BV, Venlo. It contains Tinuvin 328 (0.5323 g), M_vM_v (0.3279 g), ethylene glycol (0.1356 g), pyrazine (0.1370 g) dissolved in water, and $CDCl_3$. Two data sets of this mixture are recorded by Organon and Philips, respectively. The same sample measured at two different locations with different conditions leads to slightly different results.

The first data set, named EXP1, measured by N.V. Organon, Oss, contains small peak and phase shifts, as well as baseline distortion and baseline drifts. Therefore, EXP1 can be used to examine the preprocessing method. The data were measured by a Bruker 400 Hz NMR spectrometer. A bipolar gradient simulated echo pulse sequence was used. The applied diffusion time (Δ) is 100 ms and the duration of gradient pulses (δ) is 1.2 ms. The maximum gradient is 53.5 Gauss/cm and it varies with 32 levels. Therefore, the data contain 32 spectra and in each spectrum

there are 8192 points on the chemical shift dimension (size: 32×8192).

The second experimental data set (EXP2), measured by Philips CFT, Eindhoven, almost has no shift problem but only a small baseline drifts are present. These data are also recorded by a Bruker 400 Hz NMR instrument with the use of a bipolar gradient simulated echo pulse sequence. The maximum gradient applied is 54.4 Gauss/cm and there are 32 gradient levels used in the DOSY experiment. The size of the data EXP2 is also 32×8192. The applied diffusion time (Δ) is the same as those of data EXP1 but the duration of gradient pulses (δ) is smaller (0.6 ms), so the intensities decay slowly and hence the regression coefficients of the exponential curves (relative diffusion coefficients) are closer to each other. This may cause difficulties to resolve the data if only applying classic MCR. Therefore, data EXP2 are employed to evaluate the performance of classic MCR and MCR–NLR.

A DOSY spectrum of this mixture, calculated using the commercial XWINNMR Software (Bruker, Germany) [25], is shown in Fig. 2A, where the components and their molecular weights are also displayed. This DOSY spectrum is obtained based on the algorithm of a single channel method, i.e., mono-exponential fitting. It reveals the components in the mixture reasonably and can be used as a reference of the resolved spectra resulted from the multivariate methods. The single channel method is able to gain good separation in this case because there is almost no overlapping peak contained in

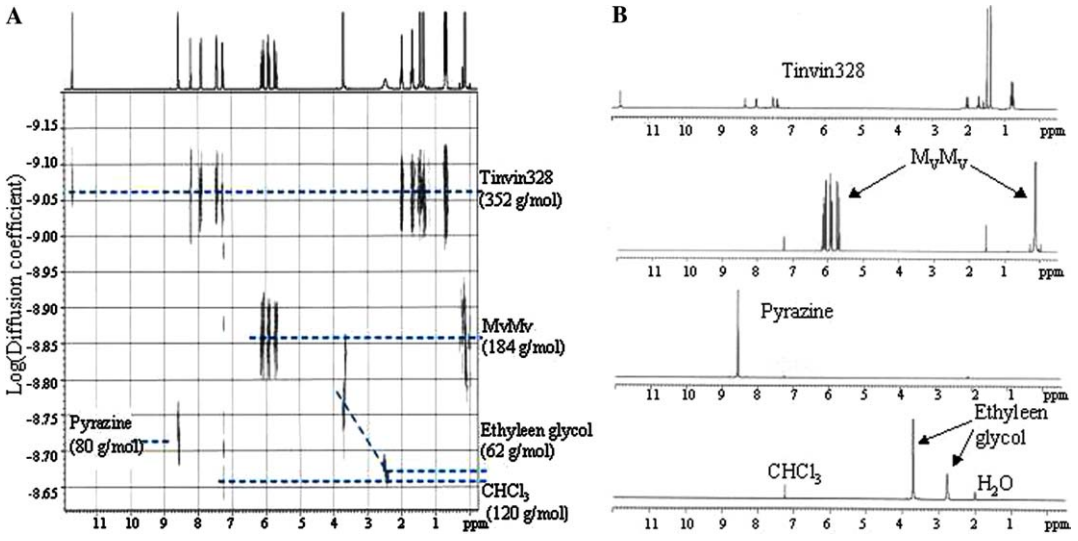

Fig. 2. Components of the chemical mixture. (A) The DOSY spectrum of experimental data and (B) the corresponding pure spectra.

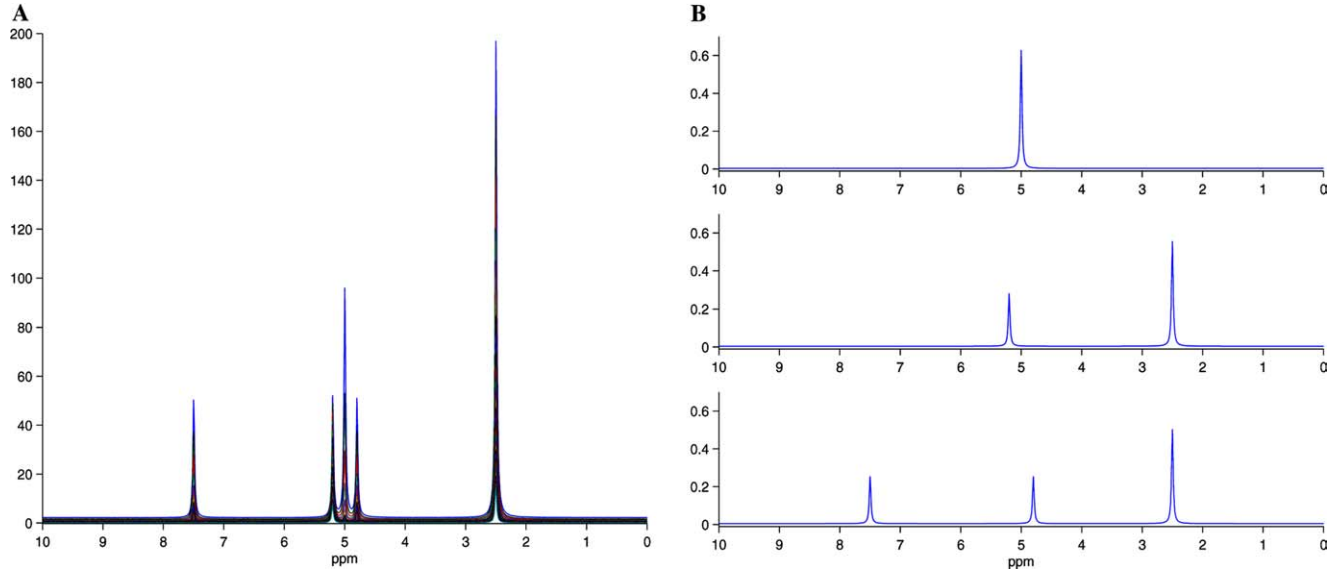

Fig. 3. (A) The simulated DOSY NMR data and (B) the reference pure spectra of the corresponding components.

the mixture spectrum. However, one can see that the calculated diffusion coefficient of each peak varies in a considerable range even there is only one component contributing to one peak. This is also the main disadvantage of all the single channel methods [7]. If there are more overlapping peaks in a sample, it may be difficult to use the single channel methods and therefore applying multivariate methods is necessary. This is also the purpose of this paper to explore a more general algorithm with the multivariate methods. The experimental data sets used for the evaluation of MCR and MCR–NLR contain six components. However, there are only four components that can be resolved because the diffusion coefficients of ethylene glycol and pyrazine are very close to each other. Besides, the signal produced by the OH group of water at 1.8 ppm in chloroform and the OH group of ethylene glycol at 2.45 ppm are very dependent on the condition of the total solution. Because of the exchangeable nature of these protons, the signals can change in position in time. This can also be another reason why the two experimental data sets result in different pure spectra for the last two components (see below). The peak near 2.4 ppm is a combination of those two OH groups. Hence, it appears to interfere on the diffusion coefficient axis with chloroform in Fig. 2A. Fig. 2B gives the NMR pure spectra of the chemical compounds measured in the solvent of chloroform.

3.3. Software

The data analysis is accomplished using MATLAB_6.0 from Math works [26]. The MCR algorithm used in this paper is modified based on the MCR function in PLS_Toolbox 2.2. All calculations are done on a Sun UNIX workstation. The software package used for the calculation in this paper will soon become available on our website: http://www.cac.sci.kun.nl/.

4. Results and discussions

4.1. Assessment of preprocessing method

4.1.1. Simulated data1

The first simulated data set and the "true" pure spectra are plotted in Fig. 3. There are two separated peaks and three peaks partially overlapping with one another. The problem of the Witjes method lies in the

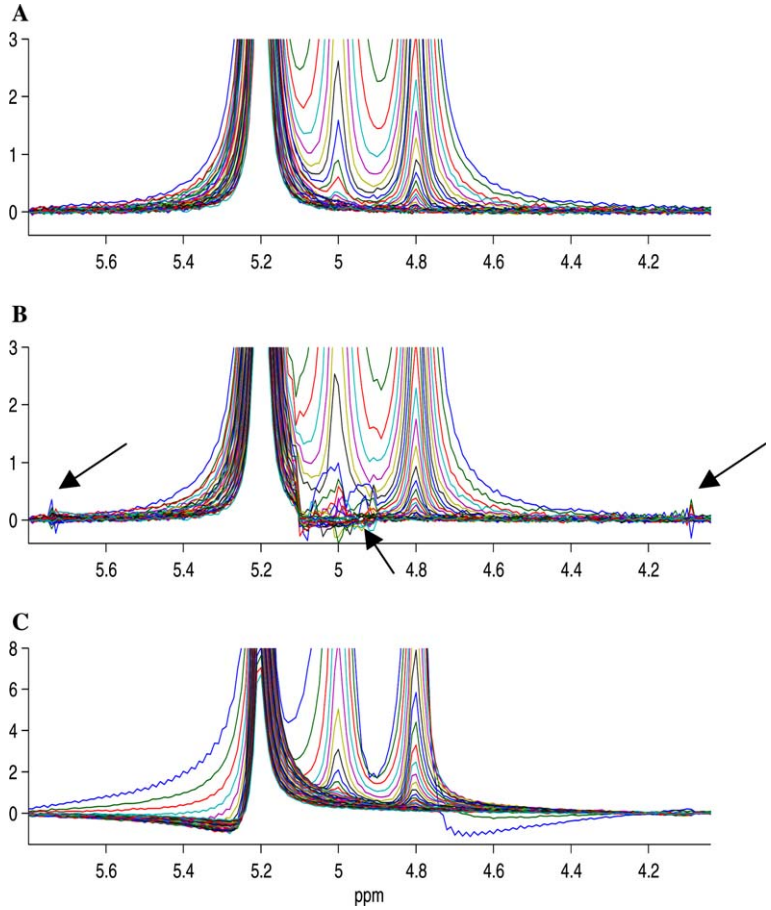

Fig. 4. Problem of the PCA-based correction for frequency and phase shifts. (A) The zoomed-in image of the partially overlapping region in the original data; (B) data correction peak by peak; and (C) data correction with the three peaks together.

overlapping region. The zoomed-in image of the overlapping region is given by Fig. 4A. If the frequency and phase correction are done peak-by-peak separately, then discontinuities will appear on the baseline, as indicated by arrows in Fig. 4B. If the three peaks are corrected altogether, the peak shape can be distorted (see Fig. 4C). On the other hand, when the data are corrected by the combination of FIDDLE and Witjes method, the phase and peak shifts are corrected and no new artefacts are introduced (see Fig. 5). To examine how the combined preprocessing method improves a DOSY NMR data, MCR is applied to the simulated data1 before and after correction. The resolved pure spectra and the corresponding decay profiles are shown in Fig. 6. In Fig. 6A, one can see that the pure spectra of the first and second component are not resolved correctly. There are some peaks that are contributions from other components. This is because the position and the phase of the corresponding peak in different spectra are not consistent. Also, the corresponding decay profiles are not smooth. On the other hand, the pure spectra and the pure decay profiles are better resolved after data correction, as indicated in Fig. 6B.

4.1.2. EXP1

The data EXP1 are analysed by OPA and classic MCR. The pure spectra and decay profiles of the data set before and after correction resolved by MCR are plotted in Figs. 7 and 8, respectively. In Fig. 7A, one can see that the peaks around 6 ppm from the second component are also present in other pure spectra, whereas these errors are reduced in the corrected spectra, as can be seen in Fig. 7B. This is also the case for the peak around 0 ppm and around 1 ppm. Moreover, the last spectrum in Fig. 7B contains lower intensities of the peaks from other components. The decay profiles from the data with correction are also smoother (see Fig. 8). By comparing the DOSY spectrum in Fig. 2A and the pure spectra in Fig. 2B, the peak around 3.5 ppm in the last spectrum in Fig. 7B dose not belong to any of the components in the

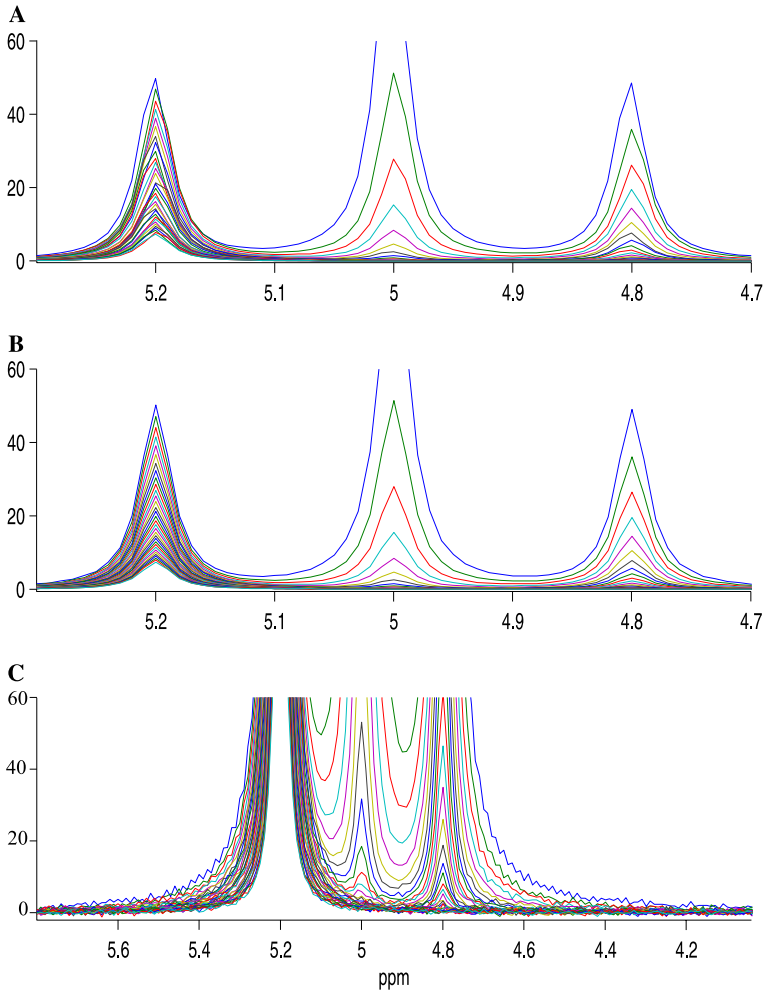

Fig. 5. Illustration of the data correction by combining FIDDLE and the Witjes method. (A) The partially overlapping region in the original data; (B) the partially overlapping region after correction; and (C) zoomed-in image after correction.

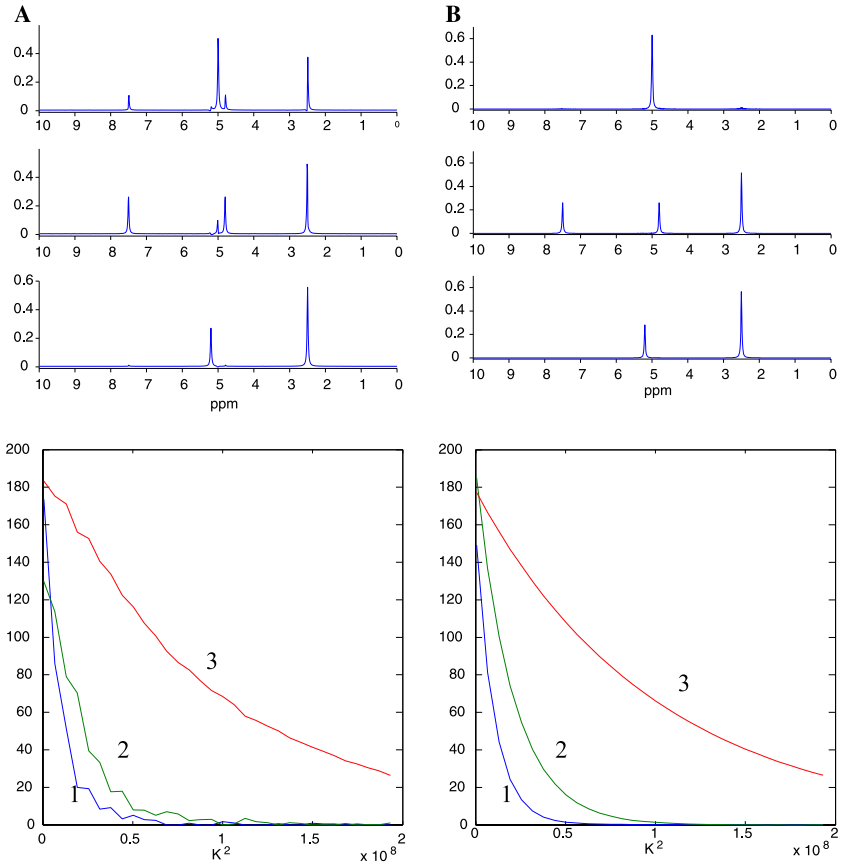

Fig. 6. The resolved pure spectra and pure decay profiles by MCR for simulated data1. (A) Obtained by the data before correction and (B) obtained by the data after correction.

mixture because it has a much slower decay behaviour (see Fig. 8). This could be caused by the formation of an unknown impurity, which may lead to the mixed peak of water and ethylene glycol that is too small to be separated. Moreover, the intensities of the peak near 7.25 ppm accounting for the chloroform are very low because of evaporation.

4.2. Comparison of MCR vs MCR–NLR

4.2.1. Simulated data2

The second simulated data set contains more overlapping region and four components (see Fig. 9). This more complex data set is used here to examine the performance of classic MCR and the combination of soft and hard modelling method, i.e., MCR–NLR. The pure variables are found by OPA and second-derivative method, as described already in Section 2. The resolved pure spectra obtained from the two methods are given by Fig. 10. It shows that the combination method MCR–NLR can gain much better resolution of the pure spectra than the classic MCR in which only non-negativity constraint is applied. The calculated diffusion coefficients (D) and the RRSSQ values are shown by Table 1. The D values acquired by both methods are very similar, although those values from MCR–NLR are a little closer to the corresponding reference values. Also, the RRSSQ is a little bit better with MCR–NLR. From the results above, it can be seen that the classic MCR has difficulties in dealing with data that contain overlapping peaks in the presence of even a small amount of artefacts. Also, the similarity of the diffusion coefficients is another reason why the classic MCR is not able to resolve the pure components properly. The disadvantage of most curve resolution method is that the solutions are not unique; i.e., there are infinite pure spectra and the decay profiles that can fulfil Eq. (4) with the same residuals between the constructed data and the original data [27]. This disadvantage can be overcome by applying non-negativity constraints to the solutions, which is described as the classic MCR in this paper. However, as there are more overlapping peaks in the data, non-negativity constraints can only remove part of the non-uniqueness problem. By applying NLR in each iteration of the classic MCR, the pure decay profiles are forced to follow exponential decay more strictly and hence results in unique solutions.

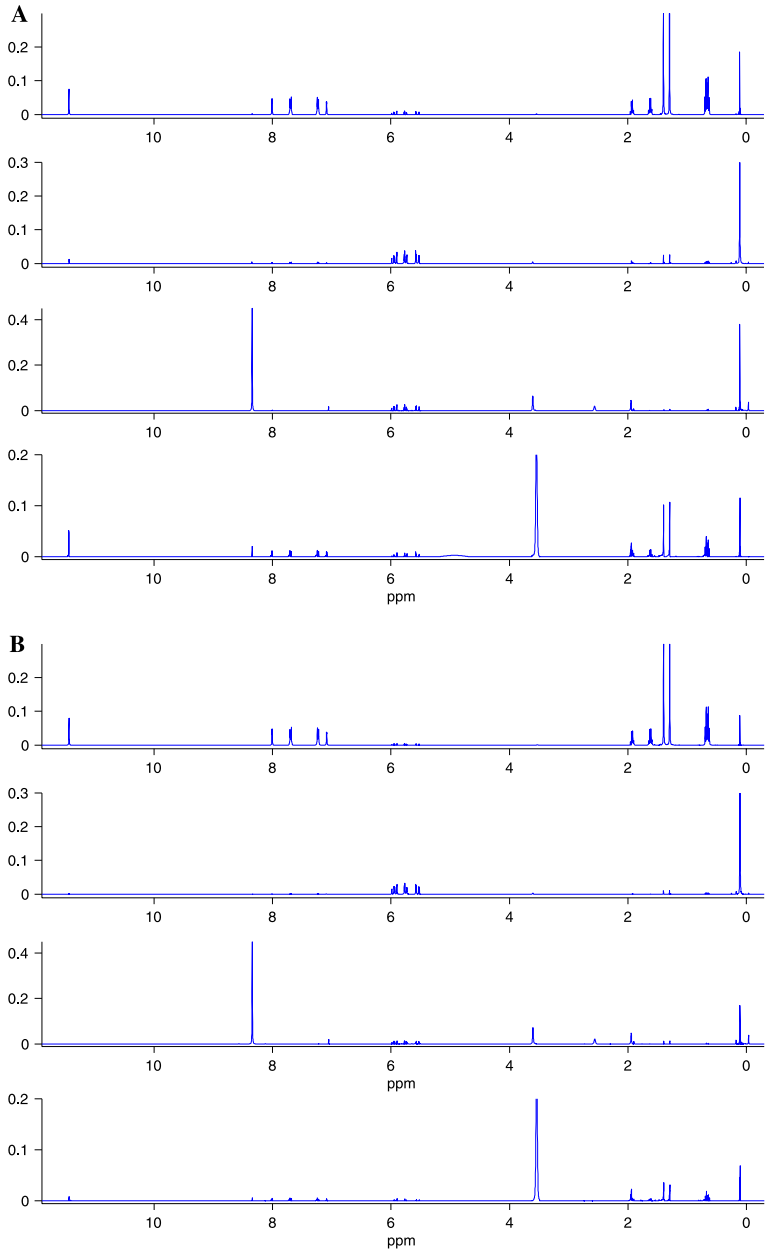

Fig. 7. Pure spectra of data EXP1: (A) before correction and (B) after correction.

4.2.2. Exp2

Following the same procedures as described before, i.e., after the data correction, first using OPA to find the pure variables, and then running the multivariate methods, four pure spectra and decay profiles are found. Since the data Exp2 do not contain distinct peak and phase shifts, the preprocessing procedure mainly correct the baseline shift problem and the peak and phase position remain more or less the same after correction. The resolved pure spectra are displayed in Fig. 11. Compared to the DOSY spectrum in Fig. 2A and the pure spectra in Fig. 2B, one can see that the pure spectra from MCR–NLR are better resolved than those obtained from the classic MCR. The classic MCR can only reasonably resolve the first component, Tinuvin 328, but fails to separate the others. Table 2 presents the relative diffusion coefficients of the components calculated by the two methods and the RRSSQ values. Again, one can see that there is only a tiny change in diffusion coefficients, which indicates that the diffusion coefficients are relatively more stable than the pure spectra. An interesting thing is that the RRSSQ value from MCR–NLR are higher than that from the classic MCR. This is because the classic MCR minimise the residuals as much as possible while imposing a kinetic model on the data is trying to correct an imperfect exponential decay profile. As a result, more residuals can be introduced.

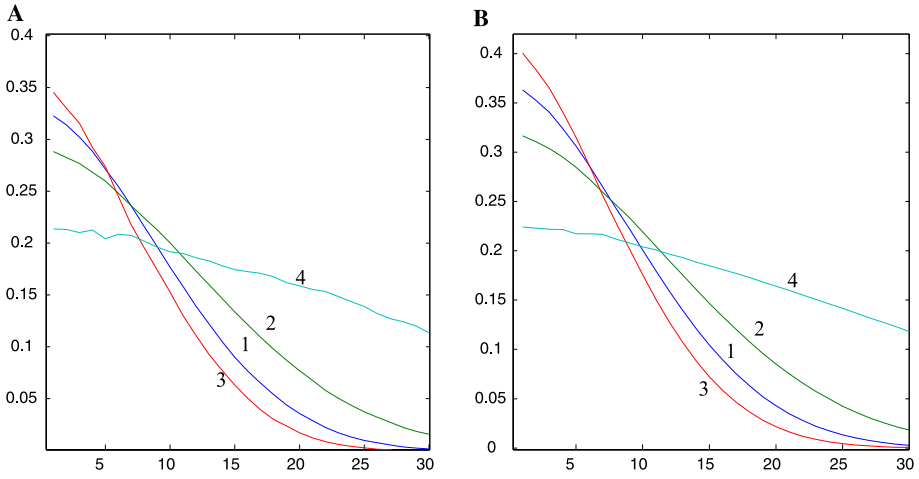

Fig. 8. The corresponding decay profiles of data EXP1: (A) before correction and (B) after correction.

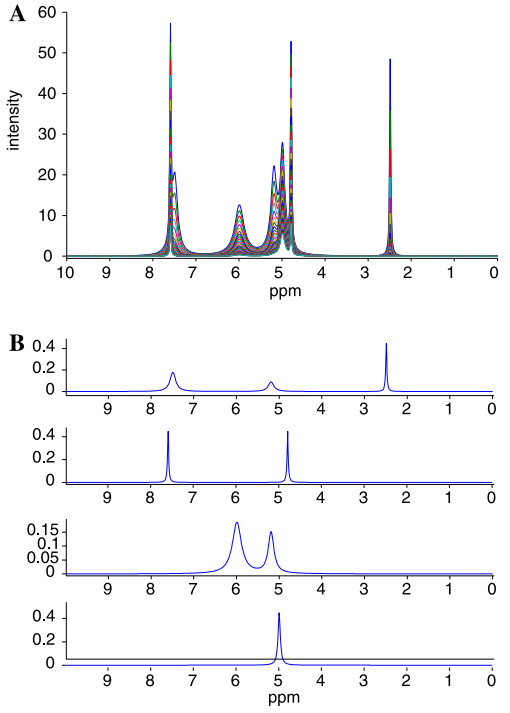

Fig. 9. The more complex simulated data2. (A) The raw data and (B) the reference pure spectra.

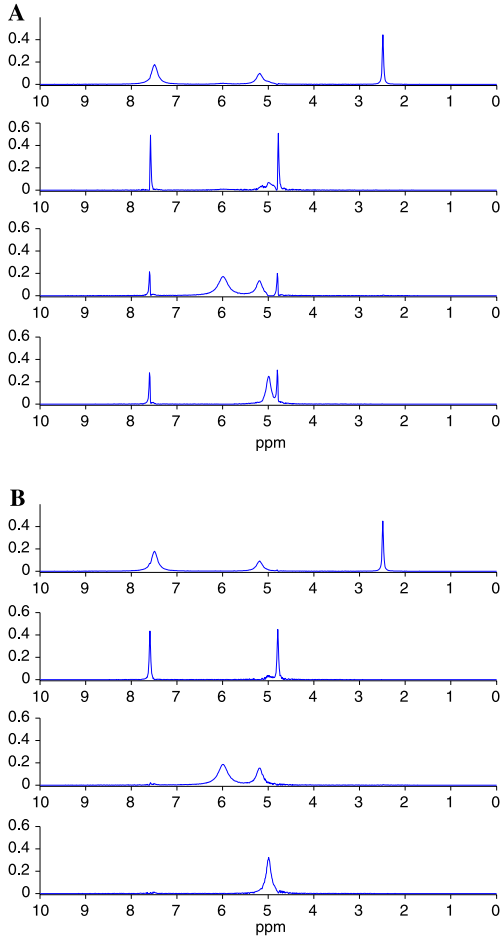

Fig. 10. Resolved pure spectra of the simulated data2 by: (A) MCR and (B) MCR–NLR.

5. Conclusion

The quality of DOSY NMR data can be improved by a set of carefully selected preprocessing methods. Baseline distortion and baseline drift can be eliminated by the automatic baseline correction method. The frequency and phase shift problem can be reduced by the combination of FIDDLE and the Witjes method. When using this combined method, it is assumed that the data have the same global shift of the peaks in the same spectrum and the small random variation can be ignored. For the data that have large dynamic shift in the whole spectrum, the data can be divided with two or

Table 1
Diffusion coefficients ($\times 10^{-7}$) of the simulated data obtained from MCR and MCR–NLR

	Reference value	MCR	MCR–NLR
Comp.1	5.000	5.040	5.021
Comp.2	1.000	1.006	0.994
Comp.3	2.000	1.984	1.986
Comp.4	0.800	0.769	0.778
RRSSQ	0	1.03%	0.88%

Table 2
Relative diffusion coefficients of the experimental data obtained from MCR and MCR–NLR

	MCR	MCR–NLR
Tinuvin	0.0173	0.0174
M_VM_V	0.0230	0.0228
EG and pyrazine	0.0274	0.0274
Water and $CHCl_3$	0.0454	0.0481
RRSSQ	0.31%	0.94%

several parts on the chemical shift dimension and the combined method is applied to part of the data each time. Preprocessing of the original data can improve the data quality and hence help to identify pure components more easily from MCR. As the data set is getting complex, i.e., many overlapping peaks, similar diffusion

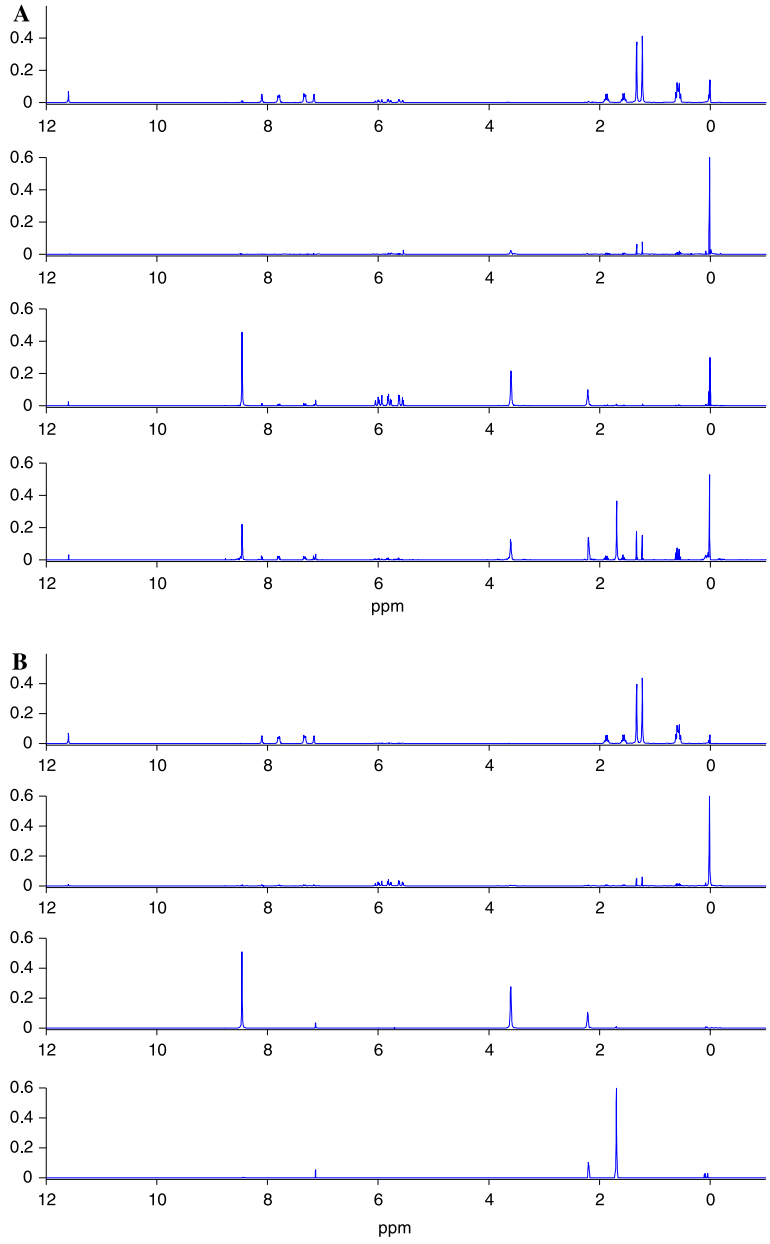

Fig. 11. Resolved pure spectra of the data EXP2 by: (A) MCR and (B) MCR–NLR.

coefficients, and so on, the solution of MCR is not unique any more. In this difficult situation, MCR–NLR, the combination of soft and hard modelling method, can be applied to eliminate the ambiguities and result in more reasonable resolution.

Acknowledgments

The authors are grateful to S.T.W. (790.35.393) for the financial support of this research. The authors thank Chris Geurt (Océ-Technology BV, Venlo) for providing the chemical mixture and its DOSY spectrum and the pure spectra, as well as Cathelijne Kloks (N.V. Organon, Oss) and Roland van de Molengraaf (Philips CFT, Eindhoven) for recording the DOSY NMR data. Also, Age Smilde, Hans Boelens, and Sabina Bijlsma (University of Amsterdam) are thanked for the suggestions of using NLR method.

References

[1] K.F. Morris, C.S. Johnson Jr., Diffusion-ordered two-dimensional nuclear magnetic resonance spectroscopy, J. Am. Chem. Soc. 114 (1992) 3139–3141.
[2] K.F. Morris, P. Stilbs, C.S. Johnson Jr., Analysis of mixtures based on molecular size and hydrophobicity by means of diffusion-ordered 2D NMR, Anal. Chem. 66 (1994) 211–215.
[3] D.A. Jayawickrama, C.K. Larive, E.F. McCord, D. Christopher Roe, Polymer additives mixture analysis using pulsed-field gradient NMR spectroscopy, Magn. Reson. Chem. 36 (1998) 755–760.
[4] A. Jerschow, N. Müller, Diffusion-separated nuclear magnetic resonance spectroscopy of polymer mixtures, Macromolecules 31 (1998) 6573–6578.
[5] K.F. Morris, B.J. Cutak, A.M. Dixon, C.K. Larive, Analysis of diffusion coefficient distributions in humic and fulvic acids by means of diffusion ordered NMR spectroscopy, Anal. Chem. 71 (1999) 5315–5321.
[6] R. Tauler, Multivariate curve resolution applied to second order data, Chemom. Intell. Lab. Syst. 30 (1995) 133–146.
[7] R. Huo, R. Wehrens, J. van Duynhoven, L.M.C. Buydens, Assessment of techniques for DOSY NMR data processing, Anal. Chim. Acta 490 (2003) 231–251.
[8] S. Golotvin, A. Williams, Improved baseline recognition and modelling of FT NMR spectra, J. Magn. Reson. 146 (2000) 122–125.
[9] G.A. Morris, Compensation of instrumental imperfections by deconvolution using an internal reference signal, J. Magn. Reson. 80 (1988) 547–552.
[10] A. Gibbs, G.A. Morris, Reference deconvolution. Elimination of distortions arising from reference line truncation, J. Magn. Reson. 91 (1991) 77–83.
[11] H. Witjes, W.J. Melssen, H.J.A. in 't Zandt, M. van der Graaf, A. Heerschap, L.M.C. Buydens, Automatic correction for phase shifts, frequency shifts, and lineshape distortions across a series of single resonance lines in large spectral data sets, J. Magn. Reson. 144 (2000) 35–44.
[12] H. Witjes, M. van den Brink, W.J. Melssen, L.M.C. Buydens, Automatic correction of peak shifts in Raman spectra before PLS regression, Chemom. Intell. Lab. Syst. 52 (2000) 105–116.
[13] T.R. Brown, R. Stoyanova, NMR spectral quantitation by principal-component analysis II. Determination of frequency and phase shifts, J. Magn. Reson. B 112 (1996) 32–43.
[14] F.C. Sánchez, B. van de Bogaert, S.C. Rutan, D.L. Massart, Multivariate peak purity approaches, Chemom. Intell. Lab. Syst. 34 (1996) 139–171.
[15] F.C. Sánchez, B. van de Bogaert, S.C. Rutan, D.L. Massart, Resolution of multicomponent overlapped peaks by the orthogonal projection approach, evolving factor analysis and window factor analysis, Chemom. Intell. Lab. Syst. 36 (1997) 153–164.
[16] A. Garrido Frenich, D. Picón Zamora, J.L. Martínez Vidal, M. Martínez Galera, Resolution (and quantitation) of mixtures with overlapped spectra by orthogonal projection approach and alternating least squares, Anal. Chim. Acta 449 (2001) 143–155.
[17] E. Bezemer, S.C. Rutan, Multivariate curve resolution with nonlinear fitting of kinetic profiles, Chemom. Intell. Lab. Sys. 59 (2001) 19–31.
[18] E. Bezemer, S.C. Rutan, Resolution of overlapped NMR spectra by two-way multivariate curve resolution alternating least squares with imbedded kinetic fitting, Anal. Chim. Acta 459 (2002) 277–289.
[19] S. Bijlsma, H.F.M. Boelens, H.C.J. Hoefsloot, A.K. Smilde, Constrained least square methods for estimating reaction rate constants from spectroscopic data, J. Chemom. 16 (2002) 28–40.
[20] W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling, Numerical Recipes, Cambridge University Press, New York, 1988.
[21] W. Windig, D.A. Stephenson, Self-modeling mixture analysis of second derivative near infrared spectral data using the SIMPLISMA approach, Anal. Chem. 62 (2002) 28–40.
[22] W. Windig, The use of second-derivative spectra for pure-variable based self-modeling mixture analysis techniques, Chemom. Intell. Lab. Sys. 23 (1994) 71–86.
[23] W. Windig, B. Antalek, J.L. Lippert, Y. Batonneau, C. Brémard, Combined use of conventional and second-derivative data in the SIMPLISMA self-modeling mixture analysis approach, Anal. Chem. 74 (2002) 1371–1379.
[24] K.F. Morris, C.S. Johnson Jr., Resolution of discrete and continuous molecular size distributions by means of diffusion-ordered 2D NMR spectroscopy, J. Am. Chem. Soc. 115 (1993) 4291–4299.
[25] Bruker, Xwin-NMR software manual, part1: general features and data processing, H9366, Brucker Analytik GmbH (1997).
[26] Mathworks, MATLAB version 6.0, Natick, MA (1999).
[27] A.K. Smilde, H.C.J. Hoefsloot, H.A.L. Kiers, S. Bijlsma, H.F.M. Boelens, Sufficient conditions for unique solutions within a certain class of curve resolution models, J. Chemom. 15 (2001) 405–411.