Available online at www.sciencedirect.com
SCIENCE DIRECT®
Analytica Chimica Acta 490 (2003) 231–251
ANALYTICA CHIMICA ACTA
www.elsevier.com/locate/aca

Assessment of techniques for DOSY NMR data processing

R. Huo [a], R. Wehrens [a], J. van Duynhoven [b], L.M.C. Buydens [a,*]

[a] *Laboratory of Analytical Chemistry Toernooiveld 1, 6525 ED Nijmegen, The Netherlands*
[b] *Central Analytical Sciences Unilever Research and Development Vlaardingen, The Netherlands*

Accepted 4 June 2003

Abstract

Diffusion-ordered spectroscopy **(**DOSY) NMR is based on a pulse-field gradient spin-echo NMR experiment, in which components experience diffusion. Consequently, the signal of each component decays with different diffusion rates as the gradient strength increases, constructing a bilinear NMR data set of a mixture. By calculating the diffusion coefficient for each component, it is possible to obtain a two-dimensional NMR spectrum: one dimension is for the conventional chemical shift and the other for the diffusion coefficient. The most interesting point is that this two-dimensional NMR allows non-invasive "chromatography" to obtain the pure spectrum for each component, providing a possible alternative for LC-NMR that is more expensive and time-consuming. Potential applications of DOSY NMR include identification of the components and impurities in complex mixtures, such as body fluids, or reaction mixtures, and technical or commercial products, e.g. comprising polymers or surfactants.

Data processing is the most important step to interpret DOSY NMR. Single channel methods and multivariate methods have been proposed for the data processing but all of them have difficulties when applied to real-world cases. The big challenge appears when dealing with more complex samples, e.g. components with small differences in diffusion coefficients, or severely overlapping in the chemical shift dimension. Two single channel methods, including SPLMOD and continuous diffusion coefficient (CONTIN), and two multivariate methods, called direct exponential curve resolution algorithm (DECRA) and multivariate curve resolution (MCR), are critically evaluated by simulated and real DOSY data sets. The assessments in this paper indicate the possible improvement of the DOSY data processing by applying iterative principal component analysis (IPCA) followed by MCR-alternating least square (MCR-ALS).

Keywords: DOSY NMR; Diffusion NMR; SPLMOD; CONTIN; Multivariate curve resolution; Alternating least squares; Factor analysis; DECRA

1. Introduction

Diffusion-ordered spectroscopy (DOSY) NMR is a two-dimensional NMR experiment, in which the signal decays exponentially according to the self-diffusion behaviour of individual molecules [1,2]. This leads to two dimensions; one dimension accounts for conventional chemical shift and the other for diffusion behaviour. Because the diffusion behaviour is related to properties of an individual molecule, such as size, shape, mass and charge as well as its surrounding environment, such as solution, temperature and aggregation state, each component in a mixture can be pseudo-separated, based on its own diffusion coefficient on the diffusion dimension. The strength of DOSY is that it can be used as a non-invasive

* Corresponding author. Tel.: +31-243-653-192; fax: +31-243-652-653.
E-mail address: lbuydens@sci.kun.nl (L.M.C. Buydens).

0003-2670/$ – see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0003-2670(03)00752-9

method to obtain both physical and chemical information. The easy and cheap implementation is another advantage of DOSY. It could be an important alternative to LC-NMR. DOSY NMR was proposed for about 10 years. Both experiments and data processing methods have been developed to obtain as much information as possible. High quality DOSY data can be obtained by the state-of-the-art NMR instrument. However, the main challenge still stands at the data processing techniques, which limits the wide use of DOSY in practice. There are two classes of techniques: single channel methods and multivariate methods.

Single channel methods employ only part of the spectra by considering a limited chemical shift range. In principle, they can find reasonable diffusion coefficients of a large number of components in a mixture, provided that the resonance peaks in the NMR spectra are well resolved. Nevertheless, overlapping regions are very commonly present in the spectra of real-life mixture and hence difficulty of interpretation by single channel methods is inevitable [3]. Two single channel methods, called SPLMOD and continuous diffusion coefficient (CONTIN) [4–7], have been implemented in the commercial NMR software, as mentioned in the literature [2]. SPLMOD [4] is applicable to discrete diffusion components. Discrete samples are defined as those containing monodisperse species and the signal can be expressed by sum of pure exponentials, i.e. each component should have one value of its corresponding diffusion coefficient. This technique is suitable for a channel containing up to two components and decay rates differing by a factor of at least two from experience. The other single channel method is CONTIN, especially used for polydisperse components, where the diffusion coefficient is not a single value but a range with a normal distribution [5–7]. The major problem of CONTIN is caused by the essential smoothing features that broaden all the peaks, even those of monodisperse components. Therefore, single channel methods find it difficult to deal with complex industrial mixtures containing both discrete and continuous distribution components and overlapping peaks at the chemical shift dimension. Both single channel methods can be applied to different peaks of the spectra (channels). For each channel, separate diffusion coefficients are found. Different peaks from the same components often lead to different diffusion coefficients (the fluctuation problem), which further complicates the DOSY data processing.

Unlike single channel methods, multivariate methods analyse the total information available in the data set simultaneously. This class of methods is probably preferable because it can eliminate the fluctuation problem and therefore give more easily interpretable pure spectra for all components. Moreover, it can handle overlapping regions which are difficult to deal with by single channel methods. However, multivariate methods can only resolve a limited number of components (up to 4–5) [3]. Two chemometrics methods, direct exponential curve resolution algorithm (DECRA) [8] and multivariate curve resolution (MCR) [10], have been proposed in the literature. The advantage of DECRA is that one experimental data set is split into two in such a way that the generalised rank annihilation method (GRAM) [8] can be applied to analyse the decaying exponential contributions of a mixture (see later). This avoids problems arising from comparing two experiments, such as peak shift and gradient level shift [1]. Moreover, DECRA uses an exact solution and therefore it only takes a few seconds to resolve pure spectrum and decay profile of each component even with the existence of overlapping peaks. The assumption of DECRA is that the data follow a highly pure (discrete) exponential decay [3]. On the other hand, MCR is based on alternating least square regression (ALS) [10]. It depends on the unique intensity variance of each component with the gradient levels and hence can tolerate deviation from the ideal exponential decay. The quality of initial guess as the input of MCR is crucial to obtain pure spectra and decay profiles.

In this article, it is intended to illustrate the difficulty of the DOSY data processing by critically evaluating the methods mentioned above, particularly focusing on the methods of multivariate analysis. The strengths and the weaknesses of each method will be highlighted, using data from simulations and a real sample. The conditions under which they can present reasonable results are investigated. These assessments will be the basis of development of several diagnostic methods and general strategies of DOSY data analysis in future studies. The most general of all methods considered here, multivariate curve resolution (MCR), is enhanced by a better initialisation method called iterative principal component analysis (IPCA) [17]

which aims to find pure or most pure variables as the initial guess to start MCR-ALS.

2. Theory

2.1. Principle of DOSY

A diffusion-ordered NMR spectrum is obtained by the use of pulsed magnetic field gradient spin-echo NMR (PFGSE-NMR) according to different diffusion of the components in the mixture. Diffusion behaviour is a measure of the translational motion of a molecule. According to the Debye–Einstein equation, it is related to the size, shape of individual molecules and specific molecule systems, such as aggregates [1]. DOSY measurements are acquired by means of either gradients in the main magnetic field, or gradients in radio frequency fields. The signal contribution of each component from the DOSY experiment is described by Eq. (1) [2]:

$$I(i, g^2) = I_0(i) exp\left[-D(i)\,(\Delta - \delta/3)\,K^2\right] \quad (1)$$

$$K = \gamma g \delta$$

where $I(i)$ is the signal amplitude of component i, $I_0(i)$ the amplitude with no gradients applied, γ the gyromagnetic ratio of the ^{1}H nucleus (rad S^{-1} T^{-1}), g the gradient strength (T), δ the duration of gradient pulses (s), Δ the diffusion time (s), and finally, $D(i)$ the diffusion coefficient of ith component (m^2/s). According to Eq. (1), if Δ and δ are the experimental constants, then the signal of a DOSY experiment attenuates depending on the gradient strengths (g^2) and diffusion coefficients (D) of individual components. A stacked plot of a simulated mixture of three compounds is shown in Fig. 1. At each gradient level, the measured spectrum is the sum of three combinations as in Eq. (1).

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

Eq. (1) can also be represented in a bilinear way,

$$I = CS^{\mathrm{T}} \quad (3)$$

where $I(r \times c)$ is a matrix with r rows and c columns, $C(r \times n)$ represents the pure decay profiles of the n components and S^{T} $(n \times c)$ contains the corresponding pure NMR spectra.

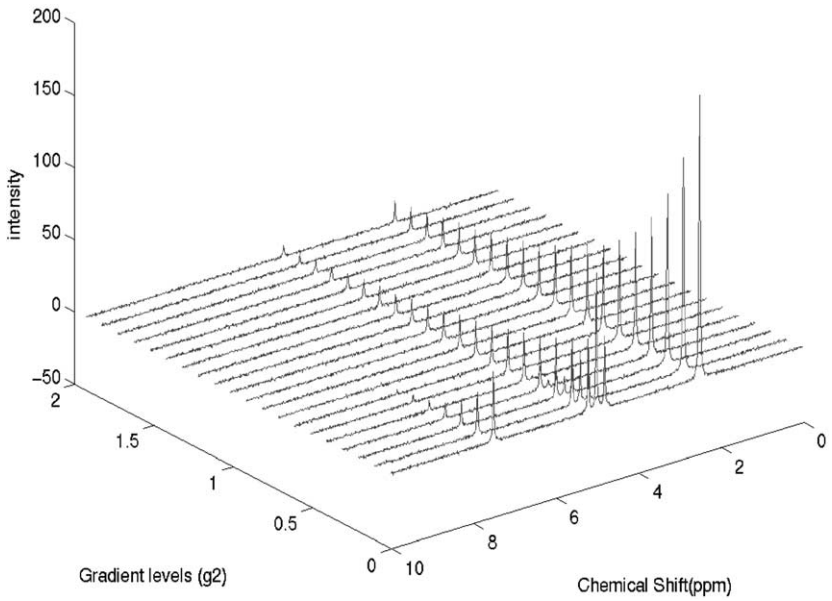

Fig. 1. Decaying NMR data with increase of gradient.

The goal is to find pure decay profile C and spectra S^{T}, given the measured data set I and appropriate constraints.

2.2. Single channel methods

2.2.1. Method for discrete diffusion coefficient: SPLMOD

SPLMOD is a single channel method which is restricted to analyse a system with discrete diffusion coefficients. SPLMOD intends to analyse sums of pure exponentials by performing least square fit of Eq. (4),

$$I(v, s) = \sum_{i=1}^{n} I_0(i) \exp(-\lambda_i s) + E \tag{4}$$

where n is the number of components, $\lambda = D(i)(\Delta - \delta/3)$, $s = K^2$ and E accounts for noise. I is the intensity of a specific frequency channel v and its variation of exponential decaying depends on the increase of s [2–4]. Resolution of discrete components using SPLMOD with certain rejection criteria has been presented by Morris and Johnson [2]. However, with the benefit of remedial constraints, SPLMOD still suffers from the overlap problem, i.e. it is difficult to separate more than two components in one single channel. It is also very sensitive to noise and hence SPLMOD often overestimates the number of the components.

2.2.2. Method for continuous diffusion coefficients: CONTIN

Some samples are composed of components with continuous distributions of diffusion coefficients, e.g. polymers and aggregates. For a specific frequency channel of the data of polydisperse system, the signal can be described by Eq. (5):

$$I(v, s) = \int_{\min\lambda}^{\max\lambda} g(\lambda) \exp(-\lambda s)\, \mathrm{d}\lambda + E \tag{5}$$

$g(\lambda)$ represents the 'spectrum' of diffusion coefficients and can be obtained by an inverse laplace transform (ILT) [5–7]. A method called CONTIN, a constrained regularisation program, attempts to solve this ILT problem and obtain the Laplace spectrum of the diffusion coefficients. The constraints are based on the non-negativity of the signal and decay constant, statistical prior knowledge and parsimony, which is used to solve the ill-posed problem. Usually, the smoothest solution with the minimum number of peaks is selected. There is no need to provide the number of components and only the threshold value is chosen as input for the program. An advantage of CONTIN is that it allows broad and narrow distributions and therefore it can be used to analyse an unknown system without any knowledge of whether the diffusion coefficient follow discrete or continuous distribution. If the distribution is narrow, then it can be reanalysed by SPLMOD if desired. The major problem is caused by the essential smoothing features which broaden all the peaks, even those of monodisperse components. This problem can lead to two monodisperse components to be described by one continuous diffusion coefficient. Thus, CONTIN often presents an incorrect number of components, due to oversmoothing. Other limitations of CONTIN include the requirement of high S/N in the data [1,2].

2.3. Multivariate methods

2.3.1. Direct exponential curve resolution algorithm (DECRA)

DECRA [8] is a multivariate method to calculate the pure spectrum and the corresponding decay profile of each component based on the GRAM [8,9]. For more details of the algorithm of DECRA, see reference [8]. GRAM is suitably applied to the two data sets with a correlation like the following:

$$A = CS^{\mathrm{T}}, \qquad B = C\alpha S^{\mathrm{T}} \tag{6}$$

α is a diagonal matrix, whose elements contain the constants accounting for the correlation of the two data sets. Previously, these two data sets were obtained from two PFGSE-NMR experiments with a slight change of experimental conditions [11]. By using DECRA, however, only one experiment is needed and the data set obtained is split into two sub-matrices, A and B. Matrix A is obtained by leaving out the last row (spectrum) of the data set; in matrix B the first spectrum is left out. Eq. (6) can be rewritten as a generalized eigenvalue problem:

$$AZa = BZ (Z \text{ is the pseudoinverse of } S^{\mathrm{T}}) \tag{7}$$

which can be solved by projecting both A and B in a common space, in this case the significant latent variables of A [8]. This procedure leads to estimation of C and S in Eq. (6).

DECRA is a fast algorithm to obtain information of pure component in a mixture. It can deal with spectra with overlapping regions within a few seconds. DECRA requires equally spaced gradient (g^2) to create exponential decay data. This is easy to implement experimentally. However, in the experimental data, the increase of g^2 can be non-linear due to the systematic error. Therefore, even if the experiment parameters are set to fulfil the requirement of getting equally spaced g^2, it is recommended that non-linear g^2 levels be checked and corrected before using the DECRA. The limitation of DECRA is that it can only be applied to discrete diffusion components with the range of about two orders of magnitude in the diffusion dimension due to the requirement of equal g^2 steps.

2.3.2. Multivariate curve resolution (MCR)

MCR applied to DOSY data was first described in reference [10]. DOSY data has bilinear structure and therefore the whole data set can be separated into two matrices, one representing the pure decay profiles and the other the pure spectra of the components. In MCR, DOSY data is first analysed by principal component analysis (PCA) to obtain the abstract factors or principal components. The loadings obtained from PCA, used as the abstract spectra are rotated by VARIMAX rotation, which maximises the simplicity of the abstract factor [12–14]. The rotated factors are used as the initial guess to start the alternating least square (ALS) [12]. By ALS, the rotated abstract spectra are used to calculate the abstract decay profiles as shown in Eq. (8) by least square regression. The new abstract decay profiles are then applied to get a new set of spectra as Eq. (9). This iteration with the application of non-negativity constraints continues until the convergence is achieved.

$$C = IS(S^{\mathrm{T}}S)^{-1} \tag{8}$$

$$S = I^{\mathrm{T}}C(C^{\mathrm{T}}C)^{-1} \tag{9}$$

MCR only depends on the change of intensity with the increase of the square of gradient strength g^2 as described in Eqs. (1) and (3), so there is neither a specific requirement of equidistant g^2 values nor any assumption of an exponential decay profile. The key requirement of MCR is a good initial guess of ALS. The main problem of MCR is in the VARIMAX rotation of the abstract factors. Obtaining the largest simplicity is the aim of VARIMAX rotation. However, a rotated factor with maximal simplicity does not necessarily mean that it is a good initial input of ALS.

Alternative methods to find starting values for MCR are simple-to-use interactive self-modelling mixture analysis (SIMPLISMA) [15,16] and IPCA [17]. Both aim to find pure or purest variables in a data matrix. For a DOSY data set, a pure variable means that the frequency on the chemical shift dimension of the NMR spectra has a pure decay profile from only one component contribution. In this article, IPCA is chosen to find the pure variables because it has the advantage that the constant for the noise correction is calculated from the real data error, i.e. from the eigenvalues, rather than choosing a random value within a certain range like in SIMPLISMA. IPCA is actually developed from the combination of key set factor analysis (KSFA) [18] and SIMPLISMA. Principal component analysis (PCA) is first applied to the original matrix. The first pure variable is the one with the highest purity, where purity is defined as

$$P_{1,i} = \frac{l_i}{(v_{1,i} + a)} \tag{10}$$

with l_i the length of the loading vector at frequency i, $v_{1,i}$ the loading of variable i on the first PC, and α a small offset, which is set to a value equivalent to real error [19] in the data. $P_{1,i}$ can be plotted in a so-called purity spectrum. A second purity spectrum, $P_{2,i}$, can be obtained by multiplying $P_{1,i}$ with a weight vector, which is obtained from the correlation between the loadings. The second pure variable is the maximum in this second purity spectrum. The process continues until all pure variables are found [17].

3. Data

To compare the different methods, they are applied to a simulated data set with varying noise levels, and an experimental one.

3.1. Simulated data

The simulated data set contains three components whose diffusion coefficients are 1.0×10^{-6}, 5.0×10^{-7}, $1.0 \times 10^{-7}\,\mathrm{cm^2\,s^{-1}}$. It is illustrated in Fig. 1. The first spectrum contains five Lorentzian

peaks. Twenty equally spaced gradient levels g^2 varying from 1.93×10^4–6.40×10^4 are used to construct twenty spectra. The two experimental constants Δ and δ are 100 and 5 ms, respectively. For each spectrum there are 1000 frequency points. Consequently, the size of the simulated data set is 20×1000. To illustrate the influence of the noise on the methods, two levels of normally distributed noise are added to the DOSY data, one with 0.05% and the other 0.25% of the highest peak intensity.

3.2. Experimental data set

The real data set has been measured by Unilever. A bipolar gradient simulated echo pulse sequence was used [20]. Spectra were recorded at a Bruker DMX600 spectrometer, using 64 gradient strengths, a spectral width of 12 ppm and 80 scans. A diffusion time of 400 ms was deployed, and the gradient durations were 1.5 ms. The sample contains three components, linear alkylbenzene sulfonate (LAS), sucrose and water. The concentrations of LAS and sucrose were 10 and 1 mM, respectively. Originally, the size of the whole data set is 64×32768. Because the signal of water is very high compared to the other signals, it is left out from each spectrum. To save computing time, only every fourth point is retained, and finally the size of the remaining data set is 64×8104. The data reduction does not affect the results of separation but reduces computing time considerably. Finally, baseline correction is applied so as to eliminate the baseline offset of each spectrum.

3.3. Software

The results from single channel methods are calculated by in-house modifications of CONTIN and SPLMOD. The original program of CONTIN and SPLMOD can be downloaded from the Internet [21]. The data analysis by multivariate methods uses MATAB 6.0 from Math works [22]. All calculations are done on a Sun UNIX workstation.

4. Results and discussion

4.1. Simulated data

4.1.1. Single channel methods

To illustrate the problem of single channel methods, the simulated data set is used here. The intensity of the NMR spectra decays with the gradient levels is shown in Fig. 1. The mean spectrum and the ideal DOSY plot are shown in Fig. 2. As can be seen in the mean spectrum, there are five peaks, accounting for

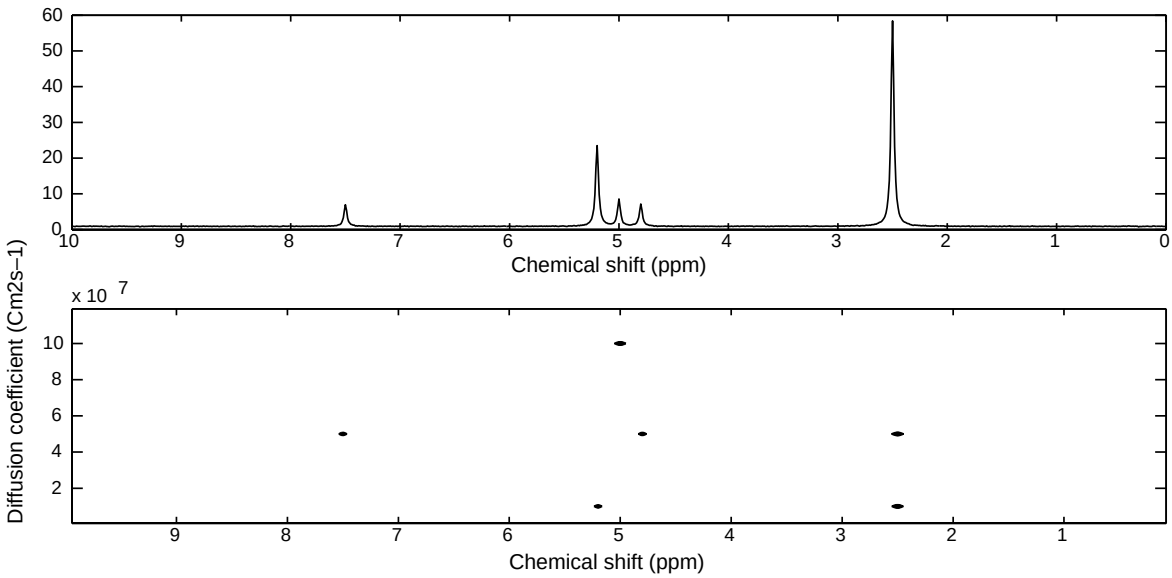

Fig. 2. Mean spectrum (upper) and DOSY plot (lower) of the simulated data.

Table 1
Diffusion coefficient, D ($cm^2\,s^{-1}$) by CONTIN

Noise level (%)	Expected D value	7.4–7.6 ppm	4.7–5.3 ppm	2.4–2.6 ppm
0.05	1.0×10^{-6}		NA[a]	
	5.0×10^{-7}	$3.66–6.80 \times 10^{-7}$	$0.496–1.24 \times 10^{-6}$	$4.06–6.08 \times 10^{-7}$
	1.0×10^{-7}		$0.723–1.62 \times 10^{-7}$	$0.886–1.20 \times 10^{-7}$
0.25	1.0×10^{-6}		NA[a]	
	5.0×10^{-7}	$2.70–8.25 \times 10^{-7}$	$0.496–1.24 \times 10^{-6}$	$3.66–7.45 \times 10^{-7}$
	1.0×10^{-7}		$0.653–1.62 \times 10^{-7}$	$0.723–1.47 \times 10^{-7}$

[a] NA: not available.

the three components of the mixture. One is a pure peak at 7.5 ppm and one is totally overlapped by two components at 2.5 ppm. At around 5 ppm, there are three peaks and they partially overlap. The integrals of the three regions at 2.4–2.6, 4.7–5.3 and 7.4–7.6 ppm from the simulated data set with different noise levels are calculated and are used as the input of SPLMOD and CONTIN. The results are shown in Tables 1 and 2. From Table 1, the results obtained by CONTIN all indicate continuous distributions with expanded ranges of diffusion coefficients, i.e. the diffusion coefficients are represented by normal distributions rather than a single value. They also reveal that CONTIN underestimates the number of components from the results at 4.7–5.3 ppm due to oversmoothing. With the increase of the noise, the range of diffusion coefficients is enlarged, which makes interpretation difficult.

The solution from SPLMOD with the standard error less than 100% of the diffusion coefficient is selected as the best solution and the results are shown in Table 2. In reference [2], the standard error in diffusion coefficient is less than 30%. The larger standard error used here allows more than two components to be resolved within one region, i.e. 4.7–5.3 ppm. One can see the fluctuation problem of single channel methods, even if the data contains little noise (0.05%): different diffusion coefficients in each calculation run for the same component. For example, component 2 occurs in the three chosen regions, the diffusion coefficient varies from 5.40×10^{-7}, 3.30×10^{-7}, to $5.10 \times 10^{-7}\,cm^2\,s^{-1}$, respectively. The different values of the diffusion coefficient for the same component can complicate the two-dimensional DOSY plot and therefore make it difficult to discern the pure spectra. In the region of 4.7–5.3 ppm, it can also be seen that the diffusion coefficient corresponding to component 2 (3.30×10^{-7}) is quite different from its reference value 5×10^{-7}. This also supports the view that "accuracy decreases when analysing more than two components by SPLMOD" [2]. Another artefact appears in the analysis of the 7.4–7.6 ppm region. According to the best solution of SPLMOD, two components were found but indeed there is only one component. As the noise increases, the calculated diffusion coefficients by SPLMOD are even farther from the real values.

Table 2
Diffusion coefficient, D ($cm^2\,s^{-1}$) by SPLMOD

Noise level	Expected D value	7.4–7.6 ppm	4.7–5.3 ppm	2.4–2.6 ppm
0.05%	1.0×10^{-6}		0.90×10^{-6}	
	5.0×10^{-7}	5.40×10^{-7}	3.30×10^{-7}	5.10×10^{-7}
	1.0×10^{-7}	2.50×10^{-7}[a]	0.94×10^{-7}	1.01×10^{-7}
0.25%	1.0×10^{-6}		0.83×10^{-6}	
	5.0×10^{-7}	4.70×10^{-7}	NA[b]	5.21×10^{-7}
	1.0×10^{-7}		1.30×10^{-7}	1.03×10^{-7}

[a] Indicates spurious value.
[b] NA: not available.

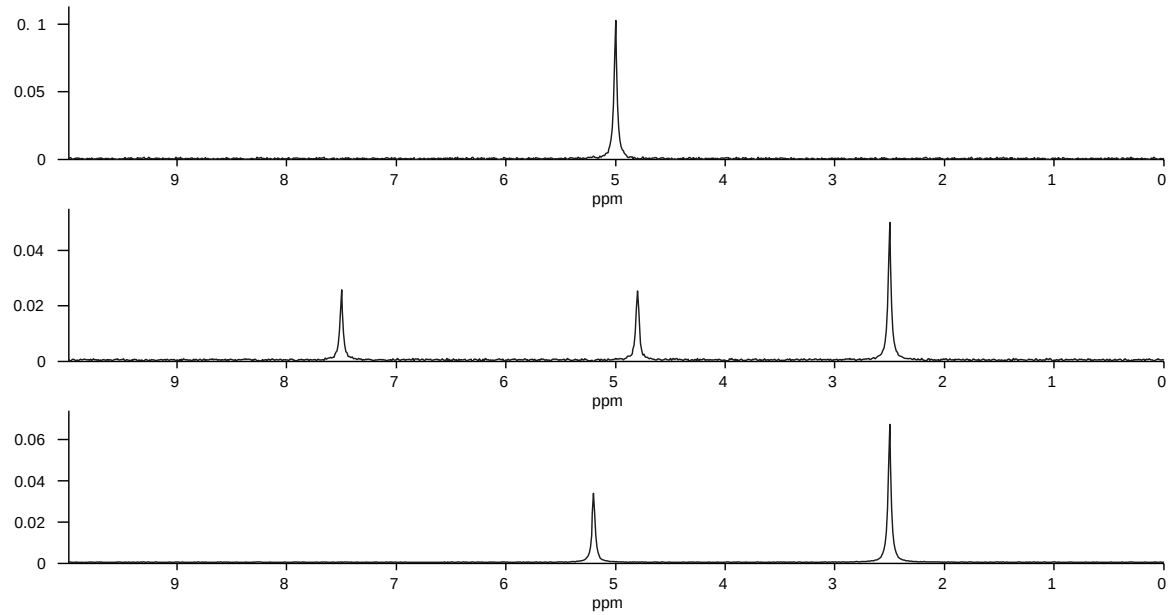

Fig. 3. Calculated spectra of simulated data by DECRA (noise level = 0.05%).

The results of this example show that single channel methods can provide correct diffusion coefficients in some cases when the basic assumptions are met, but cannot present reasonable results in general. The fluctuation problem and the sensitivity to the noise can be the main reason that results in incorrect diffusion profile. The drawbacks of single channel methods limit the widespread practical use of DOSY NMR and hence we explore multivariate methods to process DOSY data.

4.1.2. Multivariate methods

4.1.2.1. Direct exponential curve resolution algorithm (DECRA). The performance of DECRA data containing two noise levels is illustrated by Figs. 3 and 4 and Table 3. The same simulated data as used in the single channel method are used here. Fig. 3 shows the resolved pure spectra by DECRA from the relatively pure exponentially decay data, because the values of K^2 are spaced equally and only 0.05% noise is added to the data set, while Fig. 4 shows the resolved spectra of the DOSY data with larger noise (0.25%). The diffusion coefficient of each component is calculated from the least square fit of the corresponding resolved decay profile and is shown in Table 3. The correlation coefficients indicate how comparable the reference spectra and the resolved spectra are. It can be seen in Fig. 3 that the pure spectrum of each component is completely separated. However, with the increase of noise the separated spectra become more complicated to interpret. This is shown in Fig. 4. It can be

Table 3
Diffusion coefficients ($cm^2\ s^{-1}$) and correlation coefficients of simulated data by DECRA

Components	Reference value	Noise level = 0.05%		Noise level = 0.25%	
		Diffusion coefficient	Correlation coefficient	Diffusion coefficient	Correlation coefficient
1	1.00×10^{-6}	1.01×10^{-6}	0.9968	1.03×10^{-6}	0.9222
2	5.00×10^{-7}	5.02×10^{-7}	0.9978	4.72×10^{-7}	0.9406
3	1.00×10^{-7}	1.00×10^{-7}	0.9999	0.98×10^{-7}	0.9972

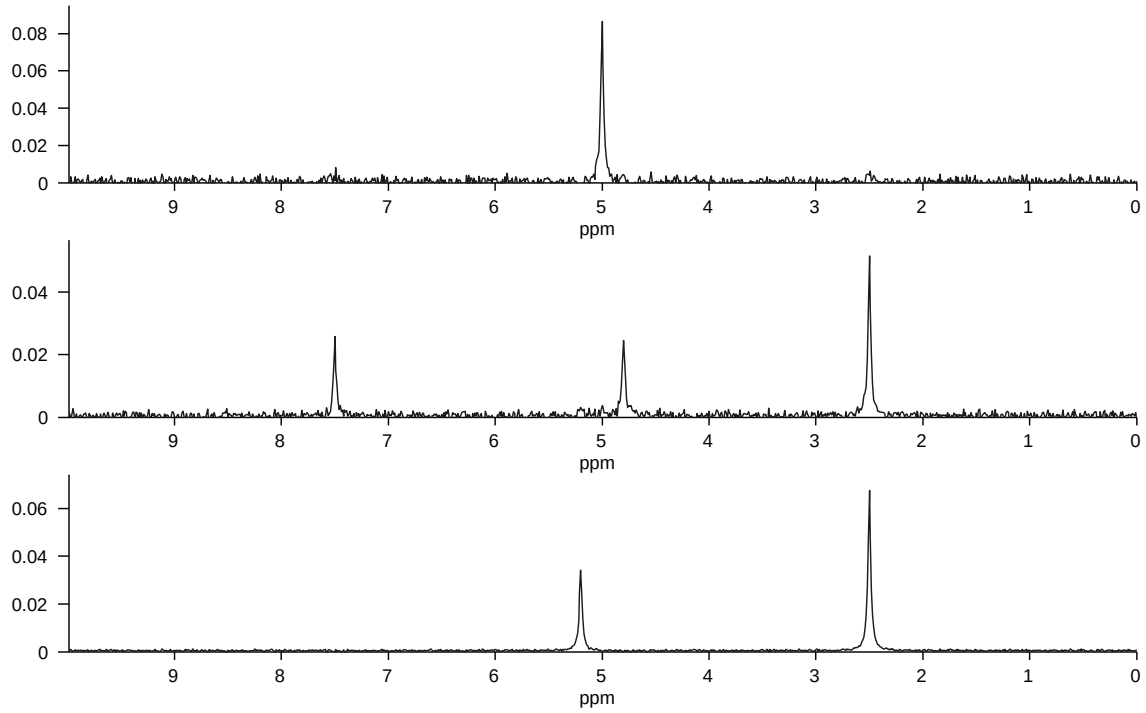

Fig. 4. Calculated spectra of simulated data by DECRA (noise level = 0.25%).

seen that small peaks from other components as well as the noise contributed to each of the resolved spectra. The diffusion coefficients are not as good as those obtained by the data with 0.05% noise. Moreover, the spectra correlation coefficients also decline with the increase of noise. This reveals that DECRA can be influenced by even a relatively small contribution of non-exponential noise.

4.1.2.2. Multivariate curve resolution.

PCA-VARIMAX-MCR. The same simulated data set with the same noise levels, i.e. 0.05 and 0.25%, is analysed by the algorithm of PCA-VARIMAX-MCR as described in reference [10]. To get meaningful spectra and decay profiles, non-negativity constraints are used to eliminate the negative values in both spectra and decay profiles [23]. The final spectra and decay profiles are shown in Figs. 5 and 6 for the data set with the noise level of 0.05%, and Figs. 7 and 8 for the one with the noise level of 0.25%. Similarly, the diffusion coefficients and spectra correlation coefficients in Table 4 illustrate the quality of the resolved spectra and decay profiles. Compared to the ideal DOSY spectrum, it can be seen that the resulted spectrum of component one has not been resolved completely.

Table 4
Diffusion coefficients ($cm^{-2}\,s^{-1}$) and correlation coefficients of simulated data by PCA-VARIMAX-MCR

Components	Reference value	Noise level = 0.05%		Noise level = 0.25%	
		Diffusion coefficient	Correlation coefficient	Diffusion coefficient	Correlation coefficient
1	1.00×10^{-6}	0.543×10^{-6}	0.6658	0.693×10^{-6}	0.8363
2	5.00×10^{-7}	2.96×10^{-7}	0.9906	3.71×10^{-7}	0.9319
3	1.00×10^{-7}	0.969×10^{-7}	0.9746	1.02×10^{-7}	0.9965

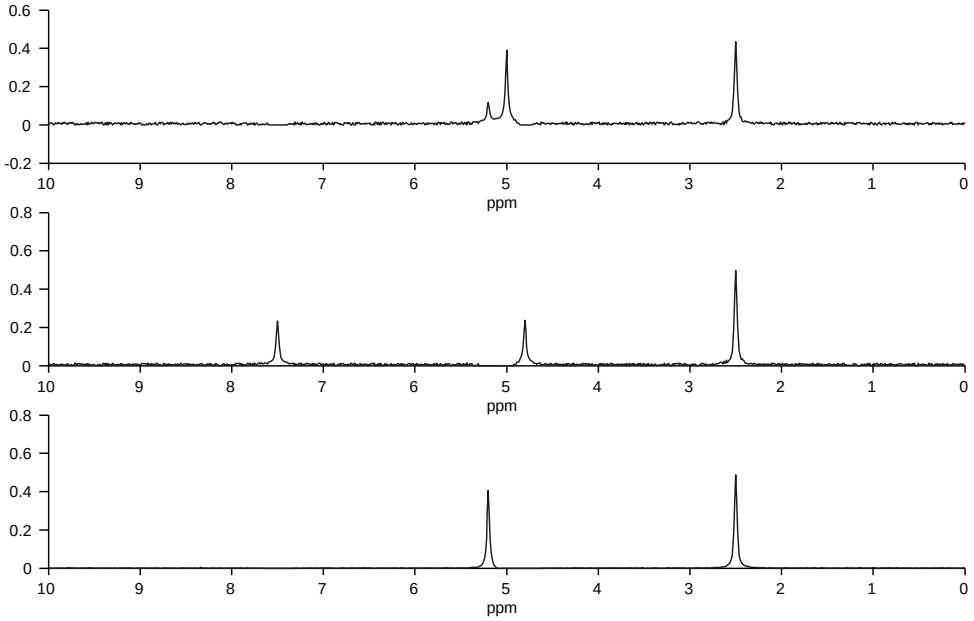

Fig. 5. Calculated spectra of simulated data by PCA-VARIMAX-MCR (noise level = 0.05%).

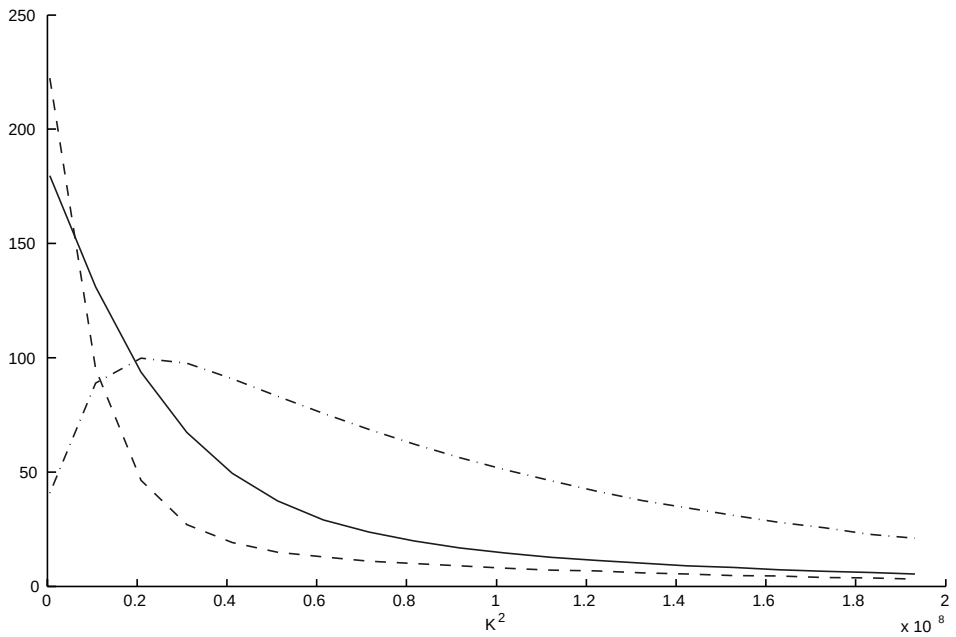

Fig. 6. Corresponding decay profile.

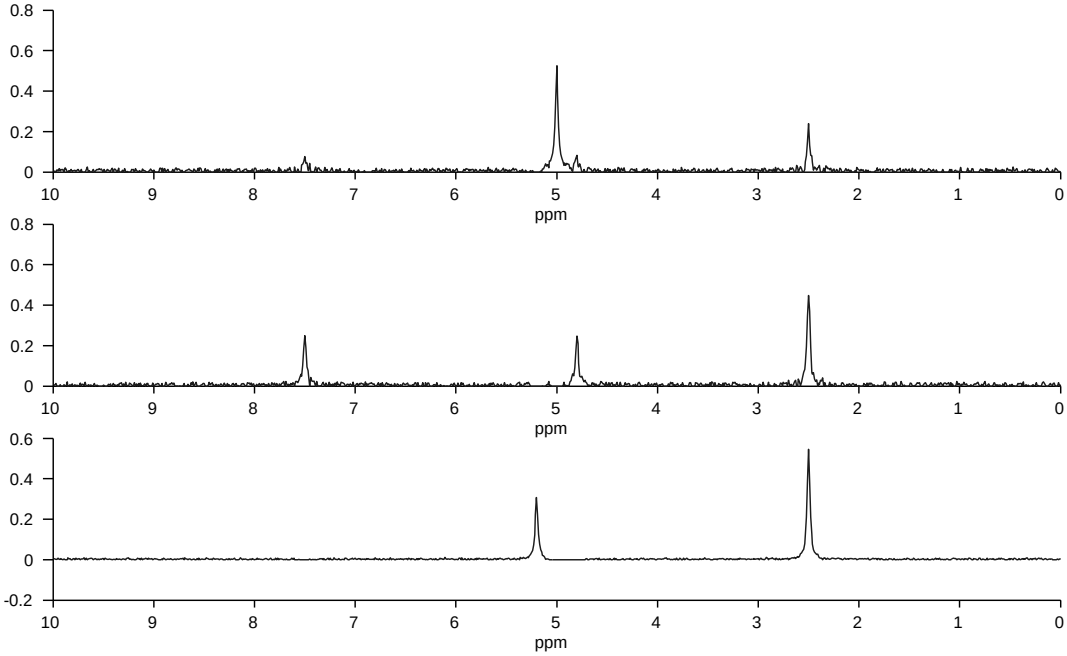

Fig. 7. Calculated spectra of simulated data by PCA-VARIMAX-MCR (noise level = 0.25%).

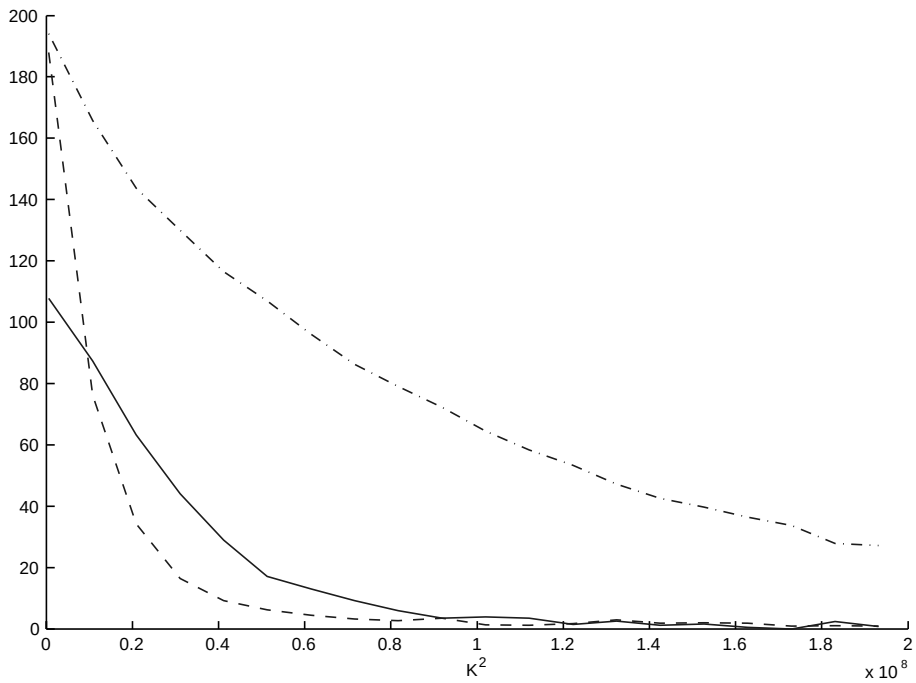

Fig. 8. Corresponding decay profile.

There should be only one peak in the first component but its resolved spectrum contains more than that at each different noise level. The decay profile of the third component shows deviation from the decay curve at the first four points (see Fig. 6). Hence, the diffusion coefficient of the third component at noise level of 0.05% is calculated by least square regression from the fifth to twentieth point. The results in Table 4 indicate that the method of PCA-VARIMAX-MCR is not influenced much by noise in the data set, but still has difficulty separating the pure spectra and decay profiles from a mixture. The original idea of VARIMAX is to rotate the abstract elution profile of the HPLC data until the greatest simplicity is reached so that each of the rotated elution profile has only one maximum. In the case of DOSY data, the abstract spectra are rotated instead of the abstract decay profiles because there is no maximum in decay profiles. However, it is rare that there is only one peak in each NMR spectrum, so the rotated spectra with most simplicity do not guarantee a good initial guess as the input to MCR. This can be the main reason that why this method does not present good resolution for a DOSY data set of a mixture in general cases.

IPCA-MCR. In this method, IPCA is used to find the pure variables. Only the purity spectra of the data with 0.05% noise are shown as an example in this article (see Fig. 9). In the purity spectra, the maximal peak occurs at 5.01, 4.81 and 5.21 ppm, respectively, and hence they are selected as the pure variables. The decay profiles for the pure variables are shown in Fig. 10. They show a nice exponential decay. The matrix as the initial guess of MCR is constructed by the column vectors of the original data sets at the chemical shift of the pure variables. After running MCR, the results of the spectra and decay profiles with the two levels of noise are shown in Figs. 11–14, respectively. Table 5 shows that calculated diffusion coefficients are similar to the reference values. The spectra correlation coefficients also indicate a good match between the calculated spectra and the corresponding real spectra. Compared with DECRA, IPCA-MCR does not show high accuracy in

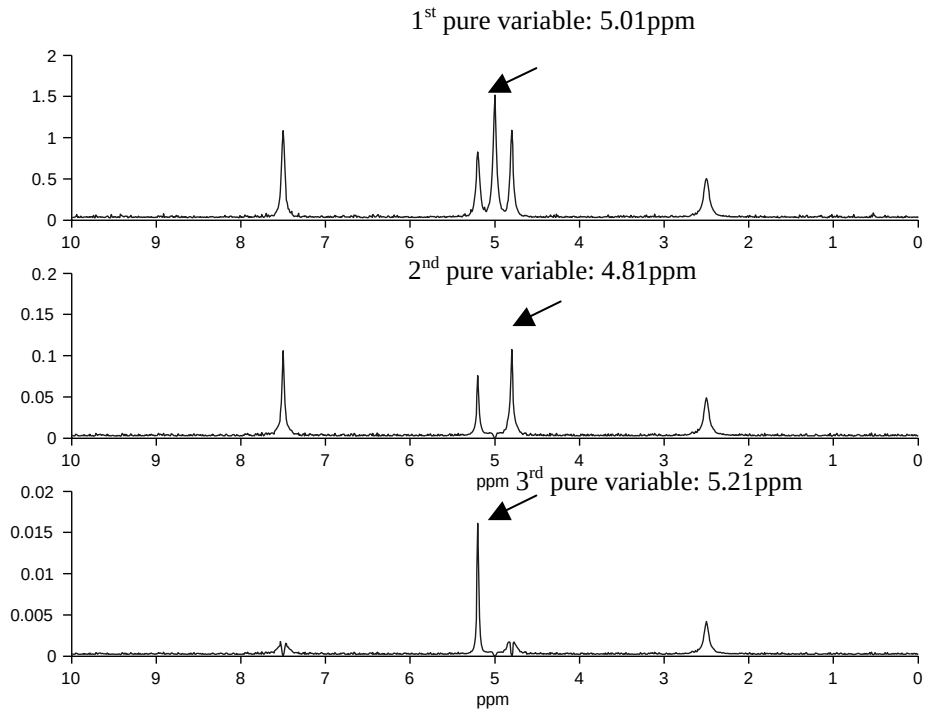

Fig. 9. Purity spectra of simulated data by IPCA (noise level = 0.05%).

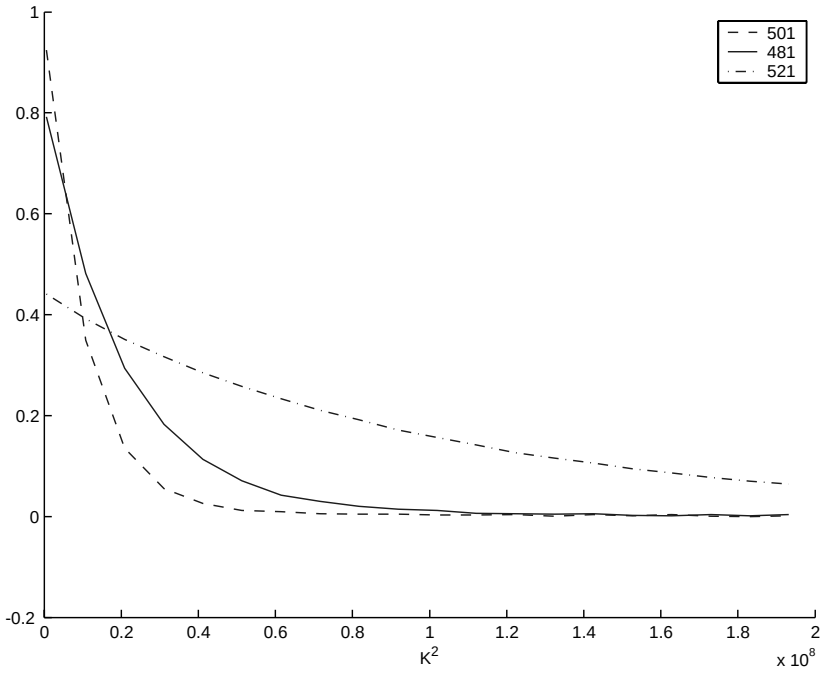

Fig. 10. Plot of pure variables at the index 501, 481, 521 at chemical shift domain.

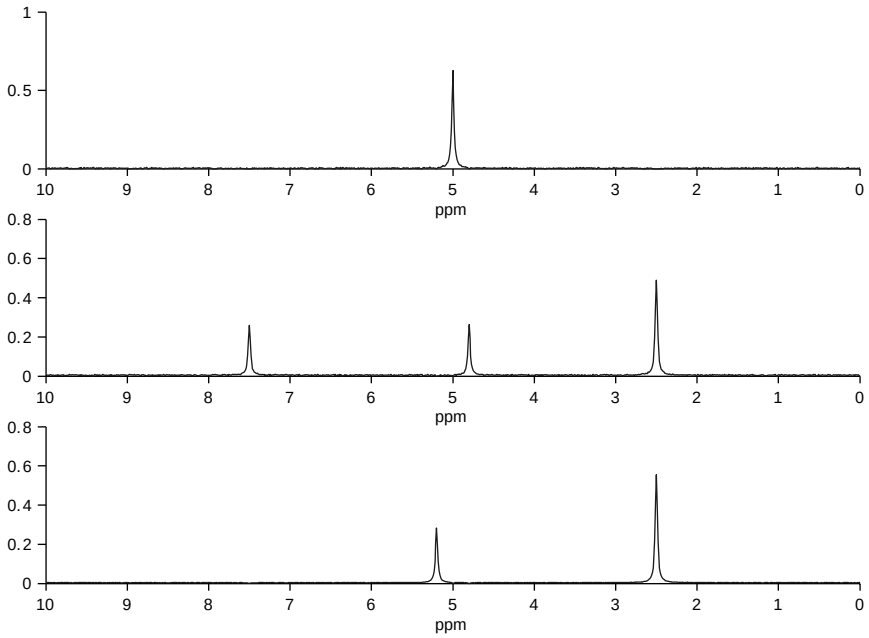

Fig. 11. Calculated spectra of simulated data by IPCA-MCR (noise level = 0.05%).

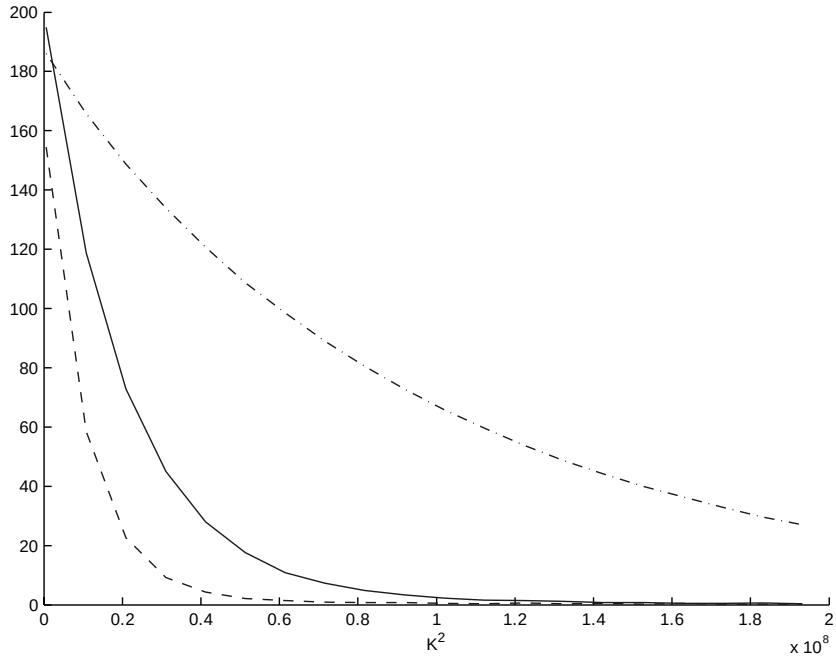

Fig. 12. Corresponding decay profile.

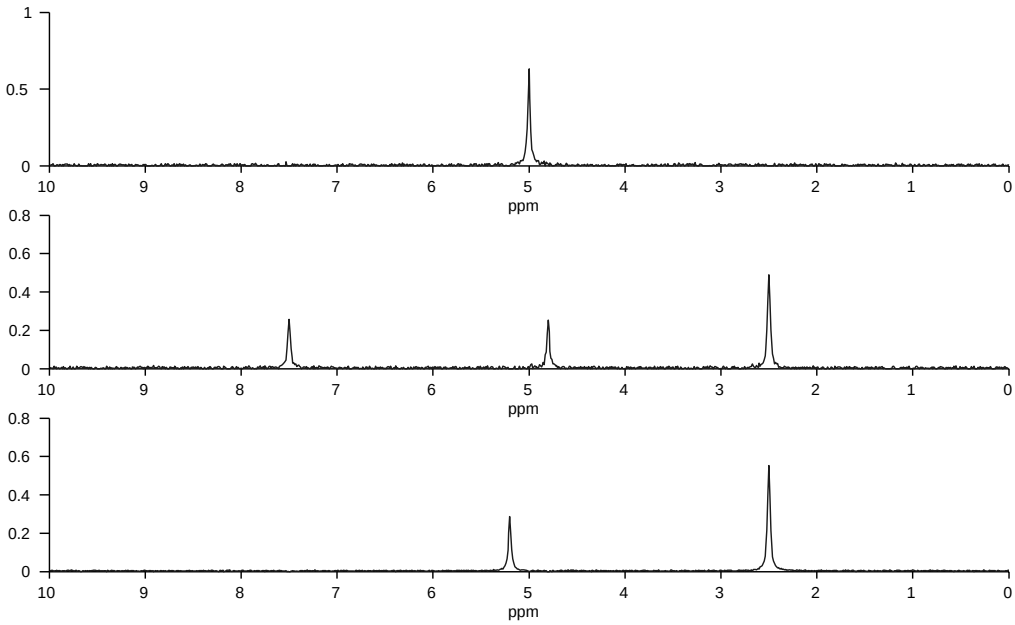

Fig. 13. Calculated spectra of simulated data by IPCA-MCR (noise level = 0.25%).

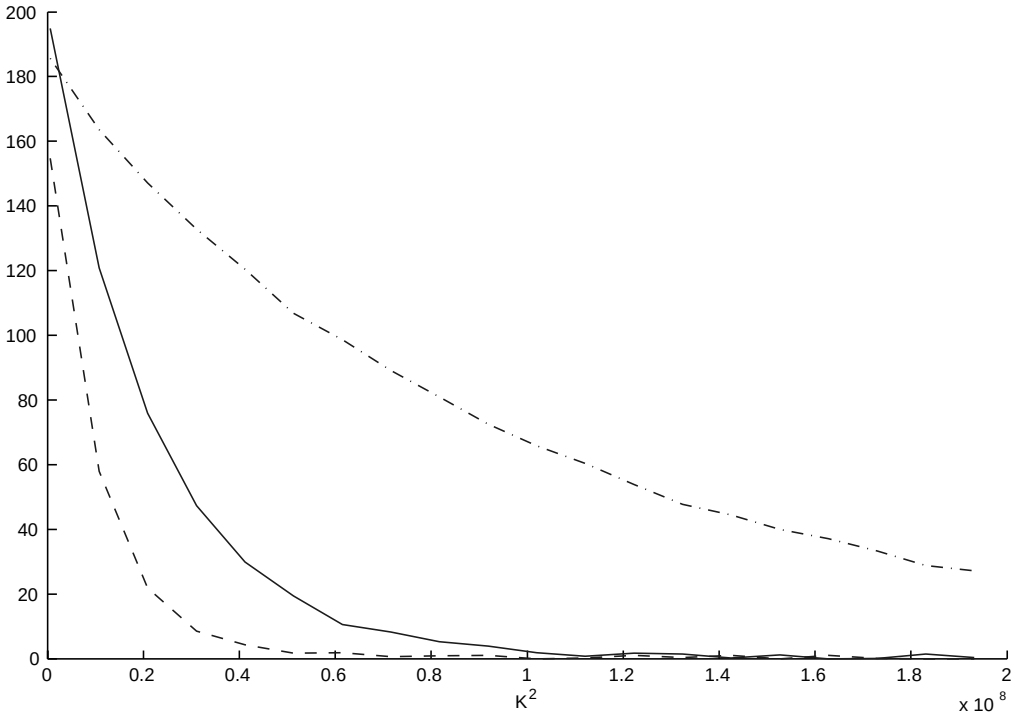

Fig. 14. Corresponding decay profile.

determining diffusion coefficient but it does present good resolved spectra with the higher noise level. This indicates that ICA-MCR is not sensitive to noise. Also, the figures of the resolved spectra visually show good separation of the pure components for the data set with both noise levels. From the results earlier, the performance of MCR is very dependent on the initial guess. Also, the choice of initial guess influences the computational speed significantly. If the initial guess of profile is similar to the true decay profile, then rather good results will be presented with rather short time; otherwise MCR may need much longer processing time to reach the convergence and lead to an incorrect interpretation. Therefore, in order to resolve DOSY data successfully by MCR, more user interactivity is needed.

4.2. Experimental data

So far, the simulated data have been used to illustrate the performance of the techniques of SPLMOD, CONTIN, DECRA and MCR. Now the real data set containing the LAS and sucrose mixture is used to evaluate the techniques further.

Table 5
Diffusion coefficients ($cm^2\,s^{-1}$) and correlation coefficients by IPCA-MCR

Components	Reference value	Noise level = 0.05%		Noise level = 0.25%	
		Diffusion coefficient	Correlation coefficient	Diffusion coefficient	Correlation coefficient
1	1.00×10^{-6}	0.934×10^{-6}	0.9972	0.970×10^{-6}	0.9858
2	5.00×10^{-7}	4.80×10^{-7}	0.9977	4.80×10^{-7}	0.9885
3	1.00×10^{-7}	1.01×10^{-7}	0.9998	1.00×10^{-7}	0.9980

Table 6
Diffusion coefficient, D ($\times 10^{-12}$ cm^2 s^{-1}) of LAS-sucrose mixture by SPLMOD and CONTIN

ppm	SPLMOD		CONTIN	
	LAS	Sucrose	LAS	Sucrose
7.69–7.59	3.94		2.86–5.14	
7.25–6.88	3.98		2.86–5.14	
5.36–5.34		32.9		29.3–39.3
4.30–4.26		32.1		29.3–39.3
4.14–4.09		32.0		29.3–39.3
3.98–3.79		31.9		29.3–39.3
3.77–3.70		29.7		21.9–39.3
3.64–3.59		31.4		25.4–39.3
3.55–3.50		32.0		25.4–39.3
2.77–2.30	3.80		1.85–7.94	
2.12–2.11		62.8[a]		34.0–70.0[a]
1.89–0.42	3.77		1.85–7.49	

[a] Indicates artefacts.

4.2.1. Single channel methods

The results from SPLOMD and CONTIN are shown in Table 6. The first row in the table represents the selected twelve chemical shift regions of the spectra. The second row and the third row are the diffusion coefficients found by SPLMOD and CONTIN, respectively. As can be seen, four regions have similar D values which indicate the peaks in these regions belong to the same compound. This component can be identified as LAS by comparing with the reference D value 2.6×10^{-12}. The reference D value is slightly different from the one obtained by SPLMOD because the D value could vary when the compound is in a different concentration or environment. The rest of the peak regions represent sucrose. From the results by SPLMOD, most of the D values belonging to the same components are very close except that in the chemical shift 2.12–2.11 ppm region, the D value is 62.8×10^{-12}. This is quite different from the other D values of sucrose and can easily be mistaken as representing another component. From the results of CONTIN, similar D values represent the same component. These results are comparable to those by SPLMOD. In 2.12–2.11 ppm region, the D value is also quite different from the others. From this example, it can be seen that fluctuation is a very common problem of the single channel method.

4.2.2. Multivariate methods

The experimental data set is analysed by the multivariate methods of DECRA, PCA-VARIMAX-MCR and IPCA-MCR. Fig. 15 shows the reference pure

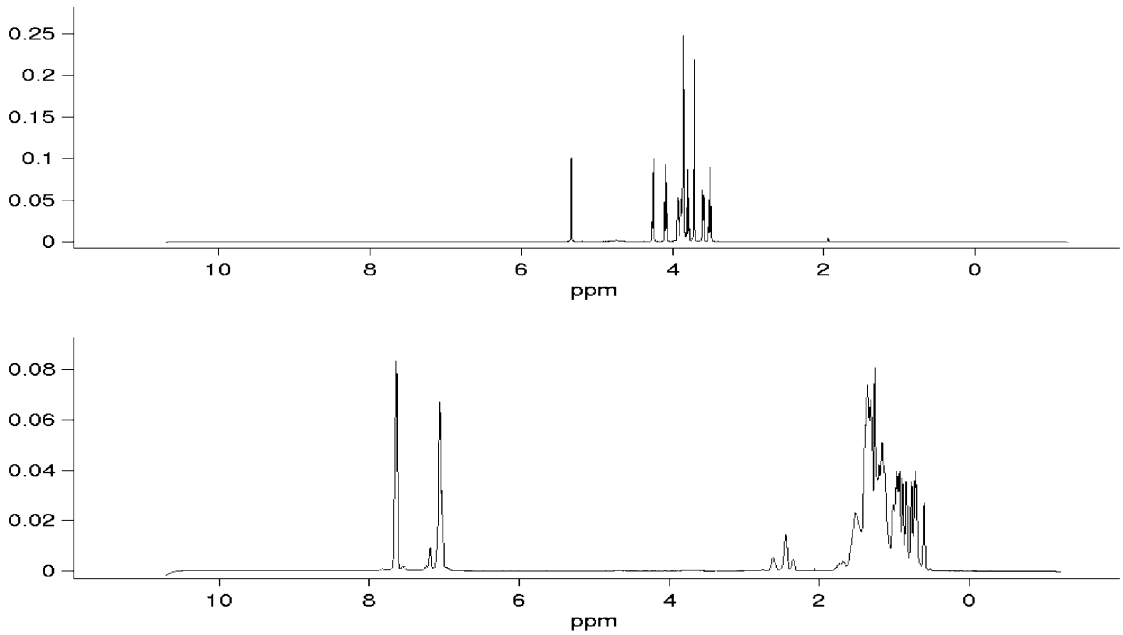

Fig. 15. Reference pure spectra of sucrose and LAS.

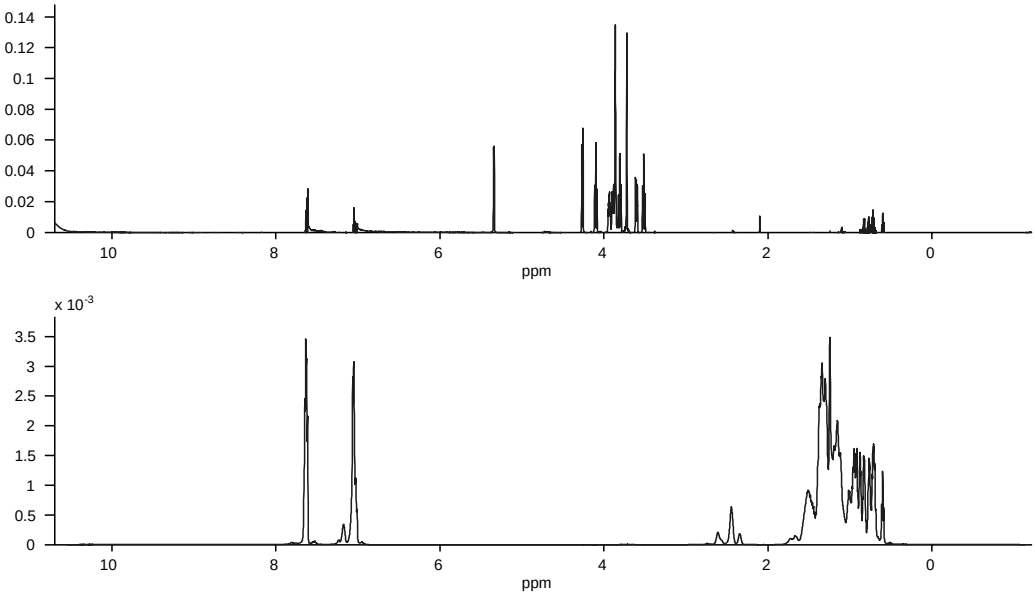

Fig. 16. Resolved spectra of sucrose-LAS mixture by DECRA.

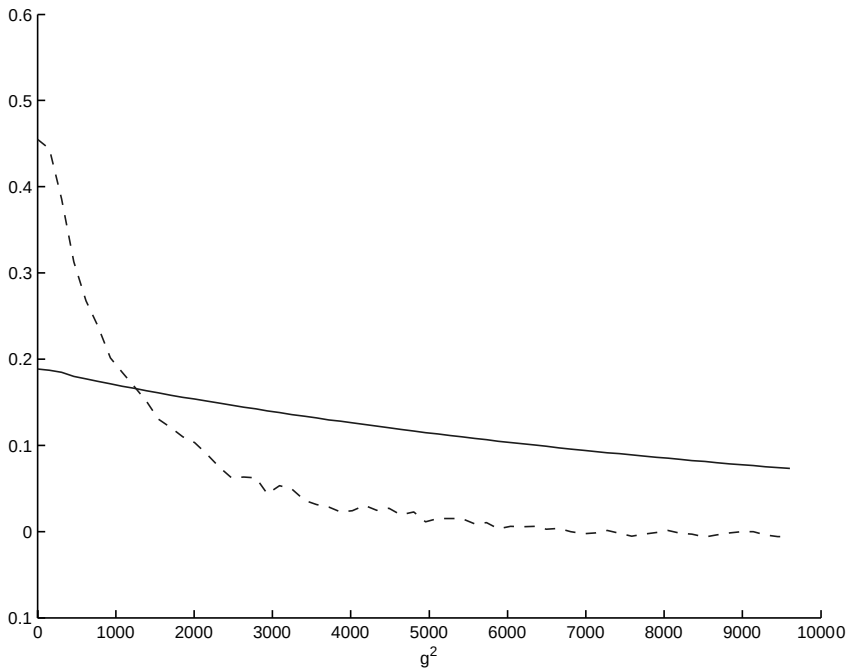

Fig. 17. Resolved pure decay profiles of sucrose (– – –) and LAS (—).

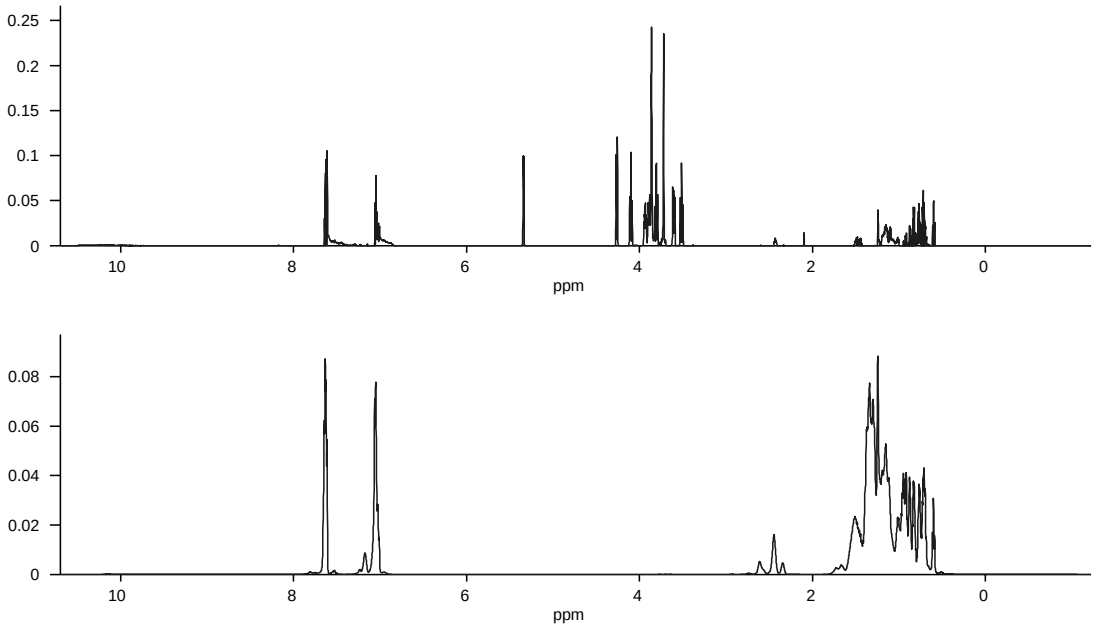

Fig. 18. Resolved spectra of sucrose-LAS mixture by PCA-VARIMAX-MCR.

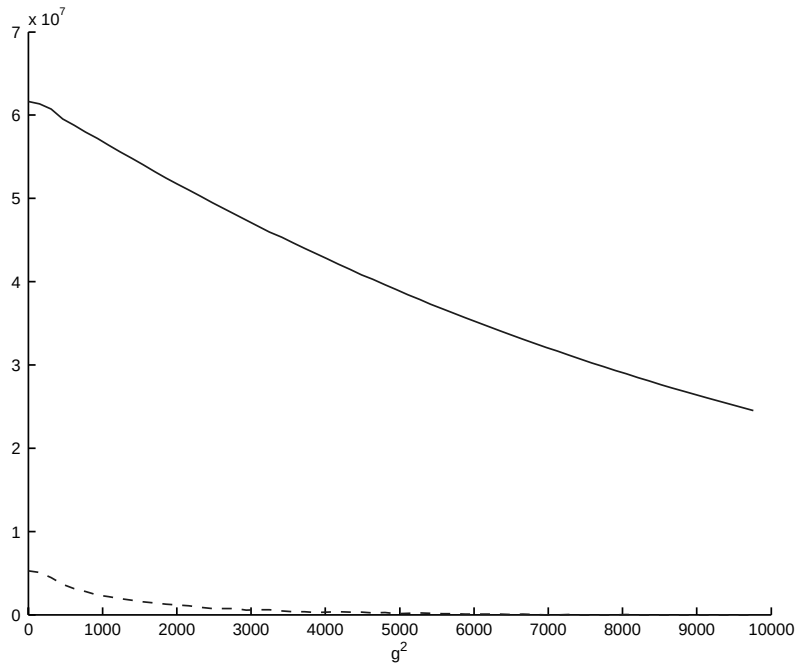

Fig. 19. Resolved pure decay profiles by PCA-VARIMAX-MCR, sucrose (– – –) and LAS (—).

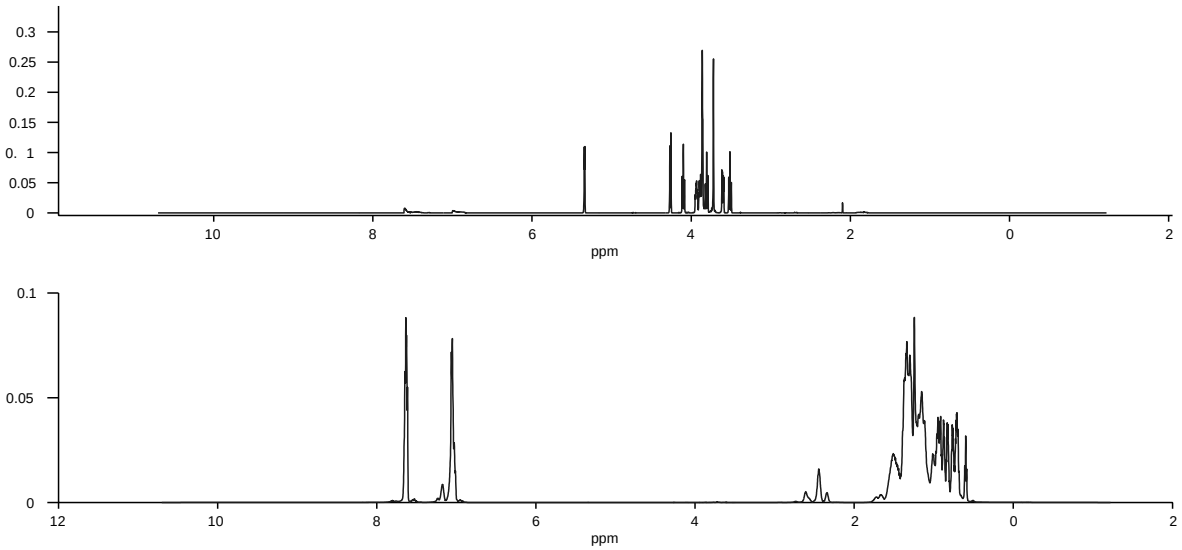

Fig. 20. Resolved spectra of sucrose-LAS mixture by IPCA-MCR.

spectra of sucrose and LAS. Figs. 16–21 show the resulted pure spectra and decay profiles of the three multivariate methods, respectively. Compared to the reference spectra in Fig. 15, the resolved spectrum of LAS by DECRA (lower one in Fig. 16) is quite reasonable. However, in the resolved sucrose spectrum it can be seen clearly that there are a few small peaks contributed from the other component, LAS, indicating

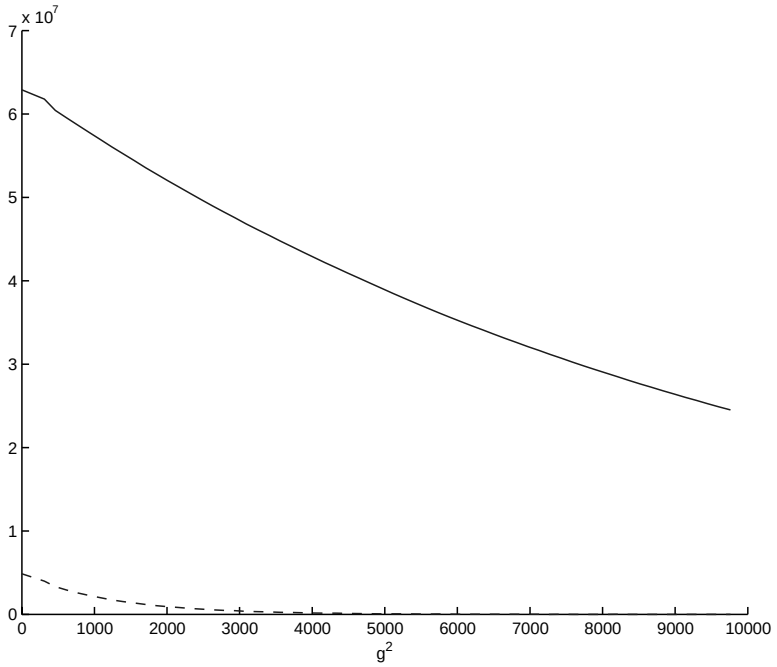

Fig. 21. Resolved pure decay profiles by IPCA-MCR, sucrose (– – –) and LAS (—).

Table 7
Diffusion coefficients ($cm^2\ s^{-1}$) and spectra correlation coefficient of sucrose-LAS mixture by multivariate methods

Methods	Component			
	Sucrose		LAS	
	Diffusion coefficient	Correlation coefficient	Diffusion coefficient	Correlation coefficient
DECRA	2.64×10^{-11}	0.8751	3.84×10^{-12}	0.9826
PCA-VARIMAX-MCR	2.62×10^{-11}	0.8615	3.73×10^{-12}	0.9826
IPCA_MCR	3.33×10^{-11}	0.9604	3.78×10^{-12}	0.9824

that DECRA did not separate the mixture completely. Since the data set is not recorded with equally spaced g^2, it is unfair to compare the results obtaining from DECRA with those from IPCA-MCR. However, this issue can arise in practice. The example used here only intends to illustrate that linearity and distance of g^2 step must be checked before using DECRA. In this case, g^2 steps have the spaces with the mean value of 154.92 and the standard deviation of 5.29. The results of PCA-VARIMAX-MCR also show incomplete separation of sucrose (see Fig. 18). On the other hand, Fig. 20 is the result from IPCA-MCR. After running IPCA, two pure variables at chemical shift 3.91 and 7.62 ppm are selected to construct an initial guess of decay profile. The resolved spectra shown in Fig. 20 are comparable with their reference spectra, which indicates MCR can resolve a mixture regardless of the spacing or the linearity of g^2, provided that a good initial guess of decay profile is available. The diffusion coefficients and the spectra correlation coefficients are summarised in Table 7. From the values of correlation coefficients in Table 7, it can also be seen that a relatively good resolution of the mixture is obtained by IPCA-MCR.

5. Conclusion

Single channel methods only analyse one part of the data set at each time and hence sometimes give ambiguous results for the DOSY data processing. However, they can be used to either obtain initial knowledge of a mixture before using multivariate methods or explore the distribution of diffusion coefficients after the pure decay profile of each component is resolved by multivariate methods. DECRA is a novel processing method to eliminate the synchronisation problem of two experiments by splitting a whole data set into two correlated sub-matrices. It provides a fast and easy way to analyse the mixture containing only discrete diffusion components. The limitation of DECRA is that it has difficulty analysing continuous diffusion components and it can be easily affected by noise. By using MCR, the DOSY intensities are not required to be of a pure exponential decay. There is no assumption of the exponential character and it is not sensitive to noise. A good initial guess of decay profile is the main requirement to present good separations for both pure spectra and decay profiles. The pure variable method IPCA can be one of the good options to find good initial guess for running MCR. This article only used rather simple examples to illustrate the different performance among the methods. Even so, one can see the different performance of each method. MCR, with a good pure variable method, is a relatively general way to deal with DOSY data. Further study will analyse more complex DOSY data from real-life samples, applying other pure variables methods with MCR and combining the results with single channel methods if necessary. Moreover, diagnostic methods will be studied, for one thing, to check the model, for the other, to identify and correct the experimental artefacts, such as line-shape distortion, non-linear increase g^2, etc. This can be used to develop pre-processing methods for DOSY data analysis.

Acknowledgements

The authors would like to thank STW (No. 790.35.393) for the financial support of this research. The authors also thank Jos Joordens and Henny Janssen for their help with the data conversion.

References

[1] C.S. Johnson Jr., Prog. NMR Spectrosc. 12 (1999) 203–256.
[2] K.F. Morris, C.S. Johnson Jr., J. Am. Chem. Soc. 115 (1993) 4291–4299.
[3] B. Antalek, Concepts Magn. Reson. 14 (2002) 225–258.
[4] R.H. Vogel, SPLMOD Users Manual, Data Analysis Group, EMBL-DA09, EMBL, Heidelberg, Germany, 1988.
[5] S.W. Provencher, Comput. Phys. Commun. 27 (1982) 213–227.
[6] S.W. Provencher, Comput. Phys. Commun. 27 (1982) 229–242.
[7] S.W. Provencher, CONTIN Users Manual (Version 2), Data Analysis Group, EMBL-DA07, EMBL, Heidelberg, Germany, 1984.
[8] W. Windig, B. Antalek, Chemom. Intell. Lab. Syst. 37 (1997) 241–254.
[9] E. Sanchez, B.R. Kowalski, Anal. Chem. 58 (1986) 496–499.
[10] L.C.M. Van Gorkom, T.M. Hancewicz, J. Magn. Reson. 130 (1998) 125–130.
[11] D. Schulze, P. Stilbs, J. Magn. Reson. 105 (1993) 54–58.
[12] M. Esteban, C. Ariño, J.M. Díaz-Cruz, M.S. Díaz-Cruz, R. Tauler, Trends Anal. Chem. 19 (2000) 49–61.
[13] M. Forina, C. Armanino, S. Lanteri, R. Leardi, J. Chemom. 3 (1988) 115–125.
[14] H.F. Kaiser, Psychometrika 23 (1958) 187–200.
[15] W. Windig, J. Guilment, Anal. Chem. 63 (1991) 1425–1432.
[16] W. Windig, C.E. Heckler, Chemom. Intell. Lab. Syst. 14 (1992) 195–207.
[17] B.U. Dongsheng, C.W. Brown, Appl. Spectrosc. 54 (2000) 1214–1221.
[18] E.R. Malinowski, Anal. Chim. Acta 134 (1982) 129–137.
[19] E.R. Malinowski, Factor Analysis in Chemistry, Wiley, New York, 1986.
[20] D. Wu, J. Chen, C.S. Johnson, J. Magn. Reson. 115A (1995) 260–264.
[21] http://s-provencher.com/, Steven Provencher, 2001.
[22] M.A. Natick, Mathworks, MATLAB Version 6.0, 1999.
[23] R. Bro, S. De Jong, J. Chemom. 11 (1997) 393–401.