Chemometrics and Intelligent Laboratory Systems 137 (2014) 47–56

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Blind separation of analytes in nuclear magnetic resonance spectroscopy: Improved model for nonnegative matrix factorization

Ivica Kopriva [a,*], Ivanka Jerić [b]

[a] Division of Laser and Atomic Research and Development, Ruđer Bošković Institute, Bijenička cesta 54, HR-10000 Zagreb, Croatia
[b] Division of Organic Chemistry and Biochemistry, Ruđer Bošković Institute, Bijenička cesta 54, HR-10000 Zagreb, Croatia

ARTICLE INFO

Article history:
Received 3 April 2014
Received in revised form 9 June 2014
Accepted 10 June 2014
Available online 18 June 2014

Keywords:
Nuclear magnetic resonance spectroscopy
(Non-)linear mixture model
Blind source separation
Nonnegative matrix factorization
Compound identification

ABSTRACT

We introduce an improved model for sparseness-constrained nonnegative matrix factorization (sNMF) of amplitude nuclear magnetic resonance (NMR) spectra of mixtures into a greater number of component spectra. In the proposed method, the selected sNMF algorithm is applied to the square of the amplitude of the NMR spectrum of the mixture instead of to the amplitude spectrum itself. Afterwards, the square roots of separated squares of the component spectra and the concentration matrix yield estimates of the true component amplitude spectrum and of the concentration matrix. The proposed model remains linear on average when the number of overlapping components is increasing, while the model based on the amplitude spectra of the mixtures deviates from the linear one when the number of overlapping components is increased. This is demonstrated through the conducted sensitivity analysis. Thus, the proposed model improves the capability of the sparse NMF algorithms to separate correlated (overlapping) component spectra from the smaller number of mixture NMR spectra. This is demonstrated in two experimental scenarios: extraction of three correlated component spectra from two ^{1}H NMR mixture spectra and extraction of four correlated component spectra from three COSY NMR mixture spectra. The proposed method can increase efficiency in a spectral library search by reducing the occurrence of false positives and false negatives. That, in turn, can yield better accuracy in biomarker identification studies, which makes the proposed method important for natural product research and the field of metabolic studies.

1. Introduction

Metabolites, low-molecular-weight compounds, are functional endpoints of metabolism and are a reflection of genetic and environmental perturbations of the system. Measurement of metabolites in biological fluids, typically urine and serum, is actually a measurement of a living system's responses to disease, drugs or toxins. Metabolic profiling is therefore an indispensable tool in drug development [1,2], toxicology studies [3], disease diagnosis [4,5], food, nutrition and environmental sciences [6,7]. Nuclear magnetic resonance (NMR) spectroscopy is emerging as a key technique in metabolomics to identify and quantify the individual compounds of which the biological fluids are composed [8–10]. The problem is notoriously difficult as a result of the presence of a large number of analytes in the studied samples. It is estimated that 2766 metabolites are to be derived from humans, and many of them are species independent [11]. Quantitative metabolomic profiling of patients with inflammatory bowel disease characterized 44 serum, 37 plasma, and 71 urine metabolites using ^{1}H NMR spectroscopy [12]. Because many analytes are structurally similar, their NMR spectra are highly correlated, with many overlapping peaks. It is thus the complexity of the samples that limits the identification of analytes, which is seen as one of the most challenging tasks in chemical biology [13]. Compound identification is often achieved by matching experimental spectra to spectra stored in the library [14,15], for example, the BioMagResBank metabolomic database [16] or, in the case of mass spectrometry, the NIST 11 Mass Spectral Library [17]. However, complexity (i.e., purity) severely hampers identification of individual compounds contained in the spectra of biological samples [15,18]. Thus, instead of analytes, their mixture is often compared with the reference components in the library. Algorithmic approaches to solve this problem may be grouped into three main categories. The scoring methods assess the matches between the experimental and theoretical spectra. To this end, similarity scores are developed to reduce the false alarm rate [19,20]. It is clear that this approach fails when the number of analytes in a mixture spectrum increases. Machine learning approaches try to learn a classifier using reference components from the library and apply it to experimental spectra [21,22]. Accuracy of this approach highly depends on representativeness and size of the training dataset (library). Thus, when the diversity of datasets is high or the number of spectra from a specific group is small, the accuracy of

* Corresponding author. Tel.: +385 1 4571 286; fax: +385 1 4680 104.
E-mail address: ikopriva@irb.hr (I. Kopriva).

http://dx.doi.org/10.1016/j.chemolab.2014.06.004
0169-7439/

analyte identification will deteriorate. Moreover, the accuracy will be affected further by the overlapping of analyte spectra. The third category of methods is known as source separation, or "deconvolution" methods.[1] The source separation methods, also known as multivariate curve resolution (MCR) methods, extract the concentration and spectra of individual components from multicomponent mixture spectra [24]. In particular, blind source separation (BSS) methods [25] refer to the class of multivariate data analysis methods capable of blind (unsupervised) extraction of analytes from mixture spectra, i.e., the concentrations of analytes are not required to be known when using the BSS algorithms. It is, however, clear that, under the stated conditions, the related inverse problem is severely ill-posed. To narrow down the infinite number of solutions to essentially unique one, constraints have to be imposed on the analyte spectra. Typically, constraints include uncorrelatedness, statistical independence, sparseness and nonnegativity. This, respectively, leads to principal component analysis (PCA) [26], independent component analysis (ICA) [27,28], sparse component analysis (SCA) [29,30] and nonnegative matrix factorization (NMF) [31]. These methods have already been applied successfully for analyte extraction from spectroscopic mixtures [32–39]. PCA, ICA and many NMF algorithms require that the *unknown* number of analytes be less than or equal to the number of mixture spectra available [32,33,36–39]. This is also true for many "deconvolution" methods [40]. This makes them inapplicable for the analysis of multicomponent mixture spectra, such as those acquired from biological samples. Sparseness-based approaches to BSS are currently a highly active research area in signal processing. Unlike PCA and ICA methods, SCA methods enable the solution of an underdetermined BSS problem, i.e., extraction of more analytes than there are mixtures available in 1D and 2D NMR spectroscopy [34,35]. Sparseness implies that at each frequency (in the case of NMR spectroscopy), only a small number of analytes are active. However, the majority of SCA algorithms require that each analyte is active at certain spectral region alone [34,35,41,42]. This assumption is increasingly hard to satisfy when the complexity of the mixture grows and when, due to reasons elaborated previously, multiple analytes become overlapped. Intuitively, it is clear that, when there are tens or hundreds of analytes in the mixture, it will be virtually impossible to isolate spectral regions where each analyte is active alone. Very recent developments in the blind separation of positive and partially overlapped sources require that each analyte be dominant, instead of active alone, at a certain spectral region [43]. Nevertheless, for complex multicomponent spectra, the same conclusion applies as above. The NMF algorithms, which in addition to nonnegativity also use a sparseness constraint, are capable of solving nonnegative underdetermined BSS problem without explicitly demanding existence of spectral regions where each analyte is active alone [44–48]. Thereby, the NMF algorithms that do not require *a priori* knowledge of the sparseness related regularization parameter are of practical value [44]. However, in the majority of cases, the NMF algorithms have been applied to extract a number of components that is smaller than the number of available mixture NMR spectra [37,38]. Herein, we demonstrate how sparseness constrained NMF ought to be applied to mixture NMR spectra to improve the quality of separation of correlated NMR component spectra. It is conjectured that the proposed method will be practically relevant for the extraction and identification of analytes in biomarker related studies. It could also increase efficiency in spectral library search procedures through reduced occurrence of false positives and negatives. Increased robustness of the linearity of the proposed method against the number of overlapping components is compared with the amplitude mixture spectra-based model and demonstrated through sensitivity analysis. The proposed method is further compared with state-of-the-art SCA algorithms. To this end, three highly correlated ^{1}H NMR component spectra were extracted from two mixtures [34], and four highly correlated COSY NMR component spectra were extracted from three mixtures [35].

2. Theory and method

2.1. Linear mixture model of multicomponent NMR spectra

The linear mixture model (LMM) is commonly used in chemometrics [24,32–39] in general and in NMR spectroscopy in particular [32,34–38]. It is the model upon which linear instantaneous BSS methods are based [25,28–31]. Taking into account the fact that NMR signals are intrinsically time domain harmonic signals with their amplitude decaying exponentially with some time constant, [49], the linear mixture model in the absence of additive noise reads as:

$$\mathbf{X} = \mathbf{AS} \tag{1}$$

where $\mathbf{X} \in \mathbb{C}^{N \times T} =: \{\mathbf{x}_n \in \mathbb{C}^{1 \times T}\}_{n=1}^{N}$ represents the mixture matrix such that each row of $\mathbf{X}$ contains one multicomponent temporal NMR mixture signal comprised of amplitude values at T time instants and symbol " =:" means "by definition". $\mathbf{A} \in \mathbb{R}_{0+}^{N \times M} =: \{\mathbf{a}_m \in \mathbb{R}_{0+}^{N \times 1}\}_{m=1}^{M}$ represents the mixture (a.k.a. concentration) matrix, whereas each column vector represents the concentration profile of one of the M analytes across the N mixtures. $\mathbf{S} \in \mathbb{C}^{M \times T} =: \{\mathbf{s}_m \in \mathbb{C}^{1 \times T}\}_{m=1}^{M}$ is a matrix where the rows represent NMR temporal signals of the analytes present in the mixture signals $\mathbf{X}$.[2] Thereby, it is assumed that $M > N$. This leads to the underdetermined BSS problem, in which case it is assumed that information regarding the concentration of analytes, stored in the mixing matrix, $\mathbf{A}$, is not known to the BSS algorithm. Thus, it is expected for the BSS method to estimate the matrix of analytes, $\mathbf{S}$, and the matrix of concentrations, $\mathbf{A}$, through use of a disposal matrix with recorded mixture signals, $\mathbf{X}$, only. However, amplitude spectra of the NMR signals, which are the actual interest, are amplitudes of the Fourier transform of the corresponding time domain NMR signals. Due to the linearity of the Fourier transform, it yields a linear mixture model in the frequency domain with the same structure as Eq. (1), whereas the T time domain instants are now interpreted as T frequencies. However, NMF algorithms are inapplicable of handling Eq. (1). That is because, in Fourier domain mixtures, $\{\mathbf{X}_n\}_{n=1}^{N}$ are complex numbers such that real and imaginary parts can be positive and negative.

[1] It is properly pointed out in [18] that the term "deconvolution" is essentially wrong, since it actually denotes inversion of a convolution, a particular kind of integral transform that describes input-output relations of linear systems with memory [23]. As opposed to that, extraction of analytes from mixtures of overlapped spectra is related to solving system of linear equations that describes a memoryless (instantaneous) system with multiple inputs (analytes) and multiple outputs (mixtures spectra).

[2] From the viewpoint of model (1), it is assumed that, in case of multidimensional NMR spectroscopy, either time- or frequency domain multidimensional signals are mapped onto their one-dimensional equivalents. It is also understood that, in transformation of time domain NMR signals into the frequency domain, the multidimensional Fourier transform is applied to the multidimensional time domain NMR signals mixture-wise: $\{\mathbf{X}_n =: F(\mathbf{x}_n)\}_{n=1}^{N}$, where F stands for Fourier transform of the appropriate dimension.

Nevertheless, amplitude spectra of the mixtures, $|\mathbf{X}| \in \mathbb{R}_{0+}^{N \times T} =: \{|\mathbf{X}_n|\}_{n=1}^{N}$, are nonnegative. Thus, an attempt is made to apply NMF to $|\mathbf{X}|$ assuming the linear mixture model [37]:

$$|\mathbf{X}| = \mathbf{B}|\mathbf{S}|. \tag{2}$$

We have purposefully denoted the mixing matrix in Eq. (2) by **B**, as opposed to **A** in Eq. (1). Because **A** stands for the matrix of concentrations, must **B** stand for something else? Actually, the NMR spectra of analytes, $|\mathbf{S}| =: \{|\mathbf{S}_m| =: |F(\mathbf{s}_m)|\}_{m=1}^{M}$, are related to the NMR spectra of mixtures, $\{|\mathbf{X}_n|\}_{n=1}^{N}$, through a nonlinear relation that, at the specific frequency ω_t reads as:

$$|\mathbf{X}_n(\omega_t)| = \sqrt{\sum_{m=1}^{M} a_{nm}^2 |\mathbf{S}_m(\omega_t)|^2 + 2\sum_{i\in I_k}\sum_{\substack{j\in I_k \\ j\neq i}} a_{ni}a_{nj}\left(\mathrm{Re}(\mathbf{S}_i(\omega_t))\mathrm{Re}\left(\mathbf{S}_j(\omega_t)\right) + \mathrm{Im}(\mathbf{S}_i(\omega_t))\mathrm{Im}\left(\mathbf{S}_j(\omega_t)\right)\right)} = \sqrt{\sum_{m=1}^{M} a_{nm}^2 |\mathbf{S}_m(\omega_t)|^2 + CT(k)} \tag{3}$$

$$0 \le k \le M,\ 1 \le t \le T, 1 \le n \le N$$

where $\mathrm{Re}(\mathbf{S}_i)$, resp. $\mathrm{Im}(\mathbf{S}_i)$, stand for the real, resp. imaginary, part of $\mathbf{S}_i$, I_k denotes an index set corresponding with the k pure components that are active at frequency ω_t and

$$CT(k) = 2\sum_{i\in I_k}\sum_{\substack{j\in I_k \\ j\neq i}} a_{ni}a_{nj}\left(\mathrm{Re}(\mathbf{S}_i(\omega_t))\mathrm{Re}\left(\mathbf{S}_j(\omega_t)\right) + \mathrm{Im}(\mathbf{S}_i(\omega_t))\mathrm{Im}\left(\mathbf{S}_j(\omega_t)\right)\right)$$

stands for the cross-terms that are explicitly dependent on k. Thus, the linear mixing model (2) does not hold. It is correct only at frequencies $\{\omega_{t(l)}\}_{l=1}^{L}$ where no analytes are active or where analyte m is active alone, that is, when $k < 2$, in which case the cross-terms $CT(k)$ equal zero:

$$\left|\mathbf{X}_n\left(\omega_{t(l)}\right)\right| = a_{nm}\left|\mathbf{S}_m\left(\omega_{t(l)}\right)\right| \quad l = 1,\ldots,L \text{ and } 1 \le t(l) \le T. \tag{4}$$

At all other frequencies, model (2) is approximate. Nevertheless, we can square the amplitude mixture spectra in Eq. (3), which yields:

$$|\mathbf{X}_n(\omega_t)|^2 = \sum_{m=1}^{M} a_{nm}^2 |\mathbf{S}_m(\omega_t)|^2 + CT(k) \qquad 0 \le k \le M,\ 1 \le t \le T,\ 1 \le n \le N. \tag{5}$$

Due to the square root operation in Eq. (3), it is intuitively clear that the linearity of model (5), defined in terms of the squares of the mixture coefficients $\{a_{nm}^2\}_{n,m=1}^{N,M}$ and squares of the amplitudes of pure components $\{|\mathbf{S}_m(\omega_t)|^2\}_{m=1}^{M}$, will be more robust with respect to (w.r.t.) the number of overlapping components k than the linearity of model (3). This statement is supported through sensitivity analysis in Section 3.1. Hence, the selected NMF algorithm should be applied to $|\mathbf{X}|^{squared} := |\mathbf{X}|. \times |\mathbf{X}|$, where $. \times$ denotes entry-wise multiplication, in order to estimate $\mathbf{A}^{squared} = \mathbf{A}. \times \mathbf{A} =: \{a_{nm}^2\}_{n,m=1}^{N,M}$ and $|\mathbf{S}|^{squared} := |\mathbf{S}|. \times |\mathbf{S}|$:

$$\left[\hat{\mathbf{A}}^{squared}, \left|\hat{\mathbf{S}}\right|^{squared}\right] = NMF\left(|\mathbf{X}|^{squared}\right) \tag{6}$$

Afterwards, estimates of **S** and **A** are obtained by:

$$\left|\hat{\mathbf{S}}\right| = \sqrt{\left|\hat{\mathbf{S}}\right|^{squared}}, \quad \hat{\mathbf{A}} = \sqrt{\hat{\mathbf{A}}^{squared}} \tag{7}$$

where the square-root operation is also performed entry-wise.

2.2. Sparseness constrained factorization

The underdetermined BSS problem (Eq. (6)) is ill-posed because matrix factorization suffers from indeterminacies: $|\mathbf{X}|^{squared} \approx \mathbf{A}^{squared}|\mathbf{S}|^{squared} = \mathbf{A}^{squared}\mathbf{D}\mathbf{D}^{-1}|\mathbf{S}|^{squared}$ for some $M \times M$ square invertible matrix **D**. Hence, it has an infinite number of possible solutions. Meaningful solutions of the instantaneous BSS problem are characterized by the permutation and scaling indeterminacies, in which case $\mathbf{D} = \mathbf{P}\mathbf{\Lambda}$, where **P** represents permutation and **Λ** represents the diagonal scaling matrix. Constraints are necessary to impose on $\mathbf{A}^{squared}$ and $|\mathbf{S}|^{squared}$ to obtain a solution of Eq. (6) unique up to permutation and scaling indeterminacies. For underdetermined BSS (uBSS) problems, of interest herein, the necessary constraint is sparseness of squares of analyte spectra stored in rows of $|\mathbf{S}|^{squared}$. Due to the character of the problem, a nonnegativity constraint is imposed on $\mathbf{A}^{squared}$ and $|\mathbf{S}|^{squared}$ as well, i.e., $\mathbf{A}^{squared} \ge \mathbf{0}$ and $|\mathbf{S}|^{squared} \ge \mathbf{0}$. While several methods are available for solving the sparseness constrained NMF problem (6) [44–48], in the experiments reported below we have used the nonnegative matrix under-approximation (NMU) algorithm [44], with MATLAB code available at [50]. The NMU method performs factorization of Eq. (6) in a recursive manner, extracting one component at a time. After identifying an optimal rank-one solution ($\mathbf{a}_1^{squared}$, $|\mathbf{s}_1|^{squared}$), the rank-one factorization is performed on the residue matrix $|\mathbf{X}|^{squared} \leftarrow |\mathbf{X}|^{squared} - \mathbf{a}_1^{squared}|\mathbf{s}_1|^{squared}$. To preserve the non-negativity of $|\mathbf{X}|^{squared}$, an underapproximation constraint is imposed on $\mathbf{A}^{squared}$ and $|\mathbf{S}|^{squared}$: $\mathbf{A}^{squared}|\mathbf{S}|^{squared} \le |\mathbf{X}|^{squared}$. It has been proven in theorem 1 in [44] that the number of non-zero entries of $\mathbf{A}^{squared}$ and $|\mathbf{S}|^{squared}$ is less than the number of non-zero entries of $|\mathbf{X}|^{squared}$. That is important in light of the very recent result proven in [51] (see Theorem 4 and Corollary 2) that uniqueness of some asymmetric NMF $\mathbf{X} = \mathbf{W}\mathbf{H}$ implies that each column of **W** (row of **H**) contains at least M-1 zeros, where M is nonnegative rank of **X**. A main reason for preferring the NMU algorithm over other sparseness constrained NMF algorithms [45,46,48] is that there are no regularization constants that require a tuning procedure. When performing NMU-based factorization of matrix $|\mathbf{X}|^{squared}$, the unknown number of analytes, M, needs to be

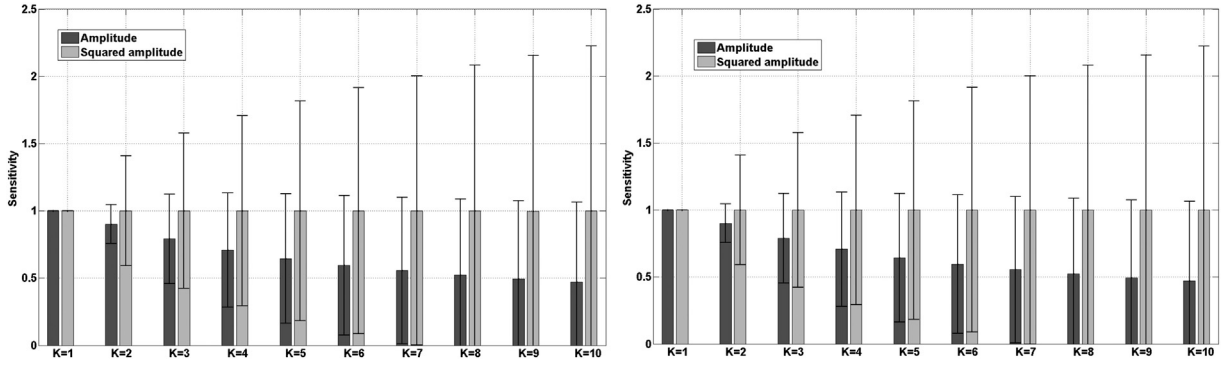

Fig. 1. Sensitivities (±standard deviation) (8) and (9) of, respectively, amplitude model (3) and squared amplitude model (5) vs. the number of analytes, k, present at some frequency ω_t. Simulation setup is described in Section 3.1. Phase of component 1: left-$\varphi_1 = \pi/4$, right-$\varphi_1 = 5\pi/7$.

given to the algorithm as an input. It is emphasized in [52] and recently in [38] that no criterion for determining the number of analytes is completely satisfactory when used alone. We, thus, do not treat this problem herein but assume that this information is available.

3. Experiment and materials

The proposed model/method was validated on a computational example related to the comparative sensitivity analysis of models (3) and (5) and two experiments: blind extraction of three analyte ^{1}H NMR spectra from two mixtures and blind extraction of four analyte COSY NMR spectra from three mixtures.[3] The first experiment has already been described in [34], and the second experiment in [35]. Both were designed to validate the SCA approach to blind extraction of analytes and their concentrations. The SCA approach explicitly demands observation points (not necessarily in the Fourier domain) where each analyte is active alone at least once. For this purpose, a wavelet basis had to be constructed in order to isolate such points [34,35]. Thereby, a data clustering procedure, the performance of which depends on tuning parameters, had to be used to estimate the matrix of concentrations, **A**. Afterwards, either a linear program or least square program regularized by ℓ_1-norm (implemented by the interior-point method) [53] had to be solved in the frequency domain to estimate the amplitude spectra of the analytes (the optimal value of the regularization constant has to be selected by the user). Please see [34,35] for detailed description of the SCA method. We demonstrate herein that the proposed methodology, which applies the NMU algorithm in Eq. (6) on squares of the mixture amplitude NMR spectra (the NMU-S), yields basically the same accuracy without explicitly demanding existence of "single analyte points" and being virtually free of the tuning parameters. In accordance with model (2)/(3), we also apply the NMU algorithm on the amplitude mixture NMR spectra (the NMU-A) in order to demonstrate deterioration in accuracy of the estimated analyte amplitude spectra. For the purpose of completeness, the experiments reported in [34,35] are briefly described here.

3.1. Numerical experiment: sensitivity analysis of mixture models (3) and (5)

The purpose of this numerical experiment is to comparatively validate sensitivity of the linearity of the mixture models (3) and (5) w.r.t. the number of analytes $0 \le k \le M$ simultaneously active at some frequency ω_t, $t = 1,..,T$. Therefore, we calculate a variation of $|\mathbf{X}_n(\omega_t)|$ in Eq. (3) w.r.t. $|\mathbf{S}_m(\omega_t)|$, as well as a variation of $|\mathbf{X}_n(\omega_t)|^2$ in Eq. (5) w.r.t. $|\mathbf{S}_m(\omega_t)|^2$, as follows:

$$\frac{\partial|\mathbf{X}_n(\omega_t)|}{\partial|\mathbf{S}_m(\omega_t)|} = \frac{a_{nm}^2|\mathbf{S}_m(\omega_t)| + a_{nm}\sum\limits_{\substack{j\in I_k\\ j\neq m}} a_{nj}\Big(\cos(\varphi_m(\omega_t))\mathrm{Re}\Big(\mathbf{S}_j(\omega_t)\Big) + \sin(\varphi_m(\omega_t))\mathrm{Im}\Big(\mathbf{S}_j(\omega_t)\Big)\Big)}{\sqrt{\sum_{m=1}^{M} a_{nm}^2|\mathbf{S}_m(\omega_t)|^2 + 2\sum\limits_{i\in I_k}\sum\limits_{\substack{j\in I_k\\ j\neq i}} a_{ni}a_{nj}\Big(\mathrm{Re}(\mathbf{S}_i(\omega_t))\mathrm{Re}\Big(\mathbf{S}_j(\omega_t)\Big) + \mathrm{Im}(\mathbf{S}_i(\omega_t))\mathrm{Im}\Big(\mathbf{S}_j(\omega_t)\Big)\Big)}} \tag{8}$$

$$\frac{\partial|\mathbf{X}_n(\omega_t)|^2}{\partial|\mathbf{S}_m(\omega_t)|^2} = a_{nm}^2 + a_{nm}\sum_{\substack{j\in I_k\\ j\neq m}} a_{nj}\left(\frac{\cos(\varphi_m(\omega_t))}{|\mathbf{S}_m(\omega_t)|}\mathrm{Re}\Big(\mathbf{S}_j(\omega_t)\Big) + \frac{\sin(\varphi_m(\omega_t))}{|\mathbf{S}_m(\omega_t)|}\mathrm{Im}\Big(\mathbf{S}_j(\omega_t)\Big)\right) \tag{9}$$

where $\varphi_m(\omega_t)$ stands for the phase of the pure component m at frequency ω_t. For $k = 1$, the linearity condition for Eq. (8), i.e., model (3), is established as:

$$\frac{\partial|\mathbf{X}_n(\omega_t)|}{\partial|\mathbf{S}_m(\omega_t)|} = a_{nm} \tag{10}$$

[3] To emphasize contribution of the proposed method in extraction of more component spectra than mixtures available, we point out the method recently introduced in [38]. There, a sparseness constrained NMF algorithm [48] has been used in three experiments to extract 3, 5 and 2 pure component spectra from, respectively, 30, 30 and 32 pulse field gradient ^{1}H NMR mixture spectra.

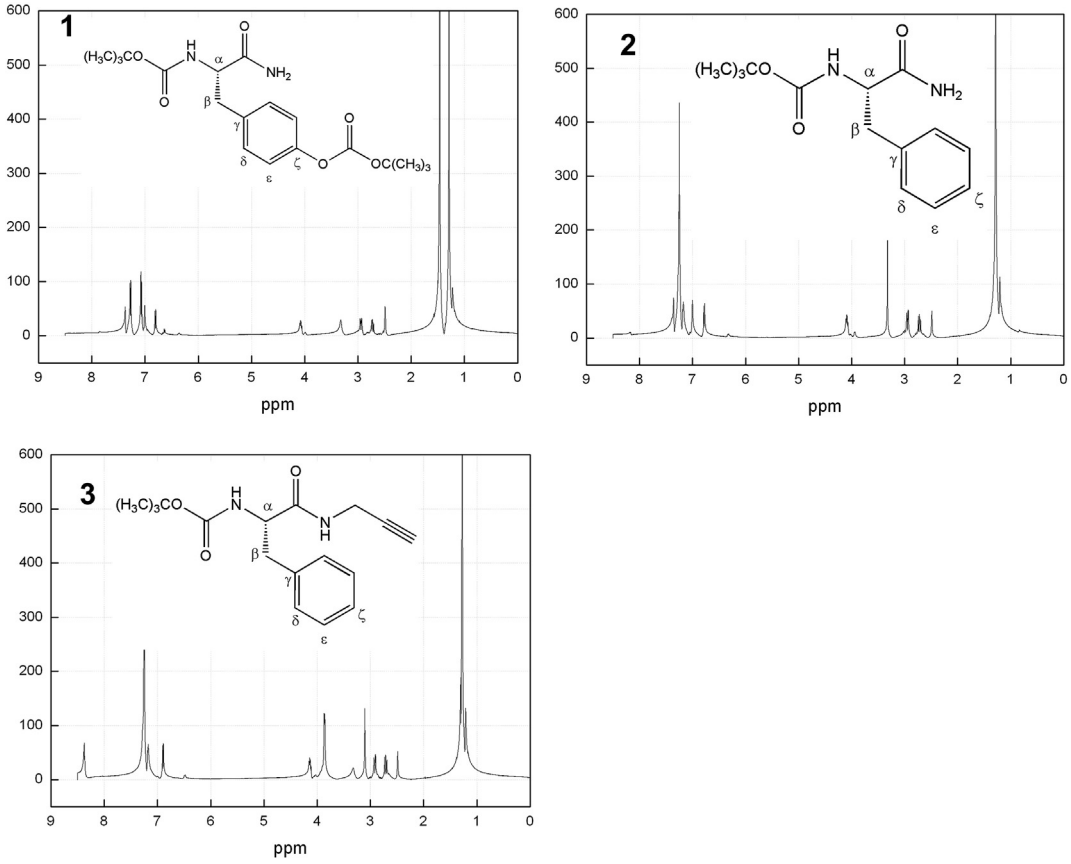

Fig. 2. ^{1}H NMR magnitude spectra and structures of pure components **1**–**3**.

and for Eq. (9), i.e., model (5), as:

$$\frac{\partial|\mathbf{X}_n(\omega_t)|^2}{\partial|\mathbf{S}_m(\omega_t)|^2} = a_{nm}^2 \tag{11}$$

In simulation of Eqs. (8) and (9), we have assumed that component $m = 1$ is dominantly active at frequency ω_t with amplitude $|\mathbf{S}_m(\omega_t)| = 1$ and arbitrary phase $\varphi_m(\omega_t) \in [0, 2\pi]$. Amplitudes and phases of other components, for $k \geq 2$, were drawn randomly with uniform distribution from (0,1] and [0,2π] intervals. 10^6 draws were executed for each value of k. Entries of the mixing vector were kept fixed at $\{a_{ni} = 1\}_{i=1}^{k}$. That is because strength of the presence of source $i = 2,..,k$ has been regulated by random amplitude $|\mathbf{S}_i(\omega_t)|$.

3.2. ^{1}H NMR measurements

Compounds Boc_2-Tyr-NH_2 (**1**), Boc-Phe-NH_2 (**2**) and Boc-Phe-NH-CH_2-C ≡ CH (**3**) were used for the preparation of two mixtures: $\mathbf{X}_1$ (**1**:**2**:**3** = 20 mg: 20 mg: 7 mg) and $\mathbf{X}_2$ (**1**:**2**:**3** = 10 mg: 25 mg: 15 mg). Mixtures were dissolved in 600 μL of DMSO-d_6. NMR experiments were carried out on a Bruker AV600 spectrometer equipped with a 5 mm BBO probe with z-gradient. The liquid-state ^{1}H spectra (600.13 MHz) were measured in DMSO-d_6 at 298 K.

3.3. COSY NMR measurements

Compounds 6-*O*-(*N*,*O*-bis-*tert*-butyloxycarbonyl-L-tyrosyl-L-prolyl)-D-glucopyranose (**4**), 6-*O*-(*N*,*O*-bis-*tert*-butyloxycarbonyl-L-tyrosyl-L-prolyl-L-phenylalanyl)-D-glucopyranose (**5**), 6-*O*-(*N*-*tert*-butyloxycarbonyl-L-prolyl-L-phenylalanyl-L-valyl)-D-glucopyranose (**6**) and 6-*O*-(*N*,*O*-bis-*tert*-butyloxycarbonyl-L-tyrosyl-L-prolyl-L-phenylalanyl-L-valyl)-D-glucopyranose (**7**), [54], were used for the preparation of three mixtures with different ratios of **4**-**7**: $\mathbf{X}_3$ (**4**:**5**:**6**:**7** = 1.1:1.7:2.7:1), $\mathbf{X}_4$ (**4**:**5**:**6**:**7** = 2.5:1.7:1.3:1) and $\mathbf{X}_5$ (**4**:**5**:**6**:**7** = 1:4:2.7:2.2). Compounds **4**-**7** and mixtures $\mathbf{X}_3$

Table 1
Normalized correlation coefficients between true and estimated pure components **1**–**3** ^{1}H NMR. Estimation error ε is defined in Eq. (12). The best values are in bold.

	$\mathbf{c}_{11}$	$\mathbf{c}_{22}$	$\mathbf{c}_{33}$	ε
SCA	**0.9254**	0.**9257**	**0.8473**	**0.1117**
NMU-S	0.9150	0.9160	0.8421	0.1140
NMU-A	0.7496	0.6595	0.6988	0.1994

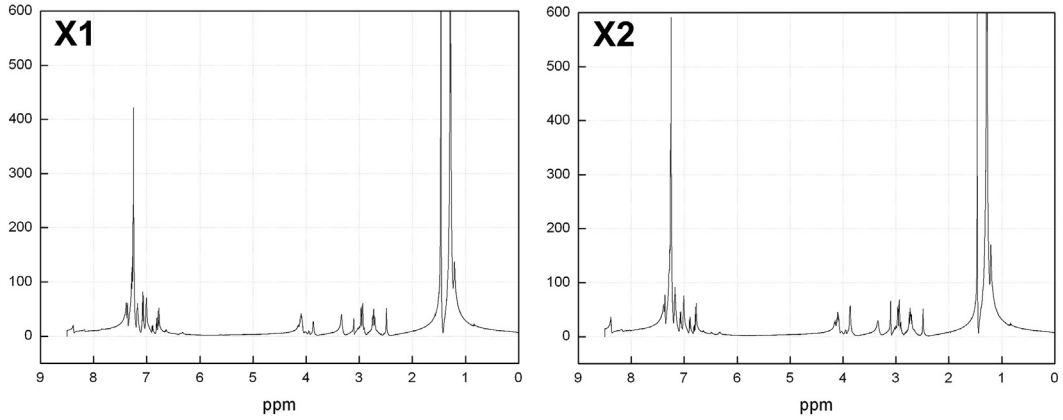

Fig. 3. ^{1}H NMR magnitude spectra of mixtures: $\mathbf{X}_1$ and $\mathbf{X}_2$.

to $\mathbf{X}_5$ were dissolved in 600 μL of DMSO-d_6. 2D COSY NMR spectra were acquired on a Bruker AV300 spectrometer, operating at 300.13 MHz and 298 K.

3.4. Software environment

The studies on experimental data reported below were executed on a personal computer running a 64-bit Windows operating system, with an Intel Core i7 920 processor at a clockspeed of 2.67 GHz and 24 GB of RAM. The MATLAB® 2011b (The MathWorks, Inc., Natick, MA) environment has been used for programming.

4. Results and discussion

Fig. 1 shows mean values (±standard deviation) of sensitivities (8) and (9) as a function of $k = 1,\ldots,10$. Results are shown for two different phases of the first component in order to demonstrate that its selection does not play a role in sensitivity analysis. Under the simulation setup described in Section 3.1, it follows that the linearity condition for model (3), implied by Eq. (10), should be $\partial|\mathbf{X}_n(\omega_t)|/\partial|\mathbf{S}_m(\omega_t)| = a_{nm} = 1$. Likewise, the linearity condition for model (5), implied by Eq. (11), should be $\partial|\mathbf{X}_n(\omega_t)|^2/\partial|\mathbf{S}_m(\omega_t)|^2 = a_{nm}^2 = 1$. It is seen that the linearity condition for model (5) holds *on average* for all values of k, while the standard deviation is increasing with k (implying that uncertainty of the outcome of the factorization is increasing with the increase of k). Implication of the sensitivity analysis of mixture model (5) is practically important. That is because, in many cases, it is reasonable to expect that only a small number, k, out of M components will coincide at each particular frequency (otherwise components will be highly similar). As opposed to mixture model (5), the linearity condition for model (3) is violated severely when k is increased, both in average and in standard deviation. In summary, when k grows, accuracy of the NMU-based

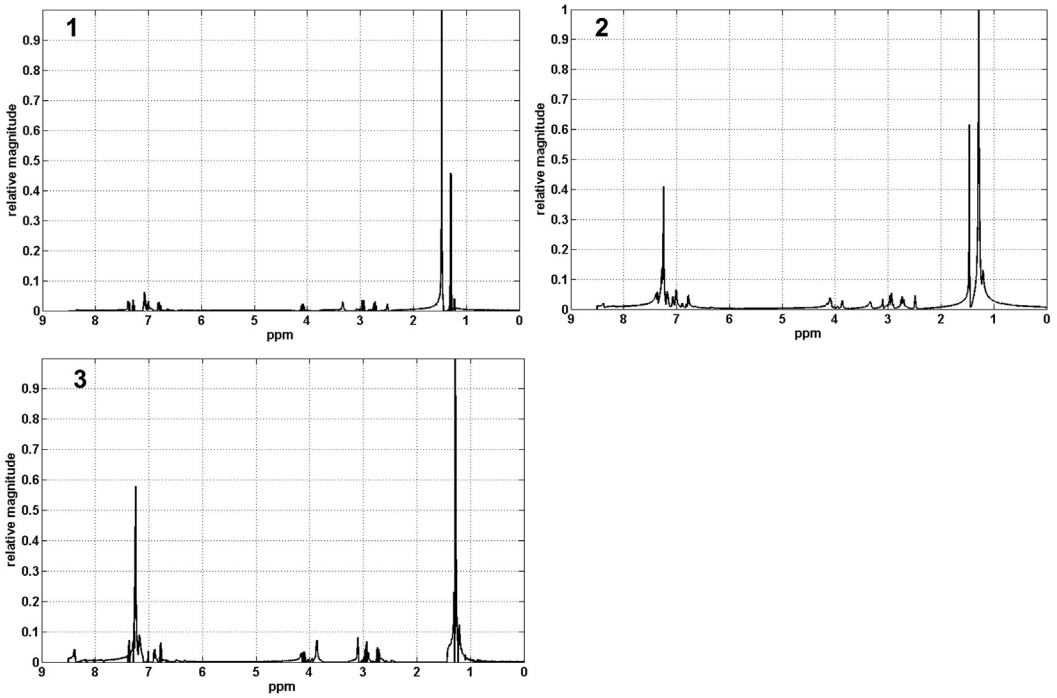

Fig. 4. ^{1}H NMR magnitude spectra of pure components **1**, **2** and **3** estimated by the NMU-S algorithm.

factorization of the mixture model (5), the NMU-S algorithm, is expected to be greater than accuracy of the NMU-based factorization of the mixture model (2)/(3), the NMU-A algorithm. That justifies use of the proposed mixture model (5) for blind extraction of analytes from mixtures of NMR spectra.

Owing to significant overlap between pure component spectra, blind separation of ^{1}H NMR spectra is considered rarely in BSS analysis. The normalized correlation coefficients between three pure component ^{1}H NMR spectra, shown in Fig. 2, were: $c_{12} = 0.4818$, $c_{13} = 0.3505$ and $c_{23} = 0.7607$. Thus, due to high correlation between component spectra, the related underdetermined BSS problem is hard.

Table 1 reports normalized correlation coefficients between pure component ^{1}H NMR spectra and ^{1}H NMR spectra of the components estimated by the SCA algorithm [34], as well as the NMU-S and NMU-A algorithms proposed herein. Also reported is the average absolute value of the error between the true correlation matrix and the correlation matrix between the estimated and true spectra:

$$\varepsilon = \frac{\sum_{i=1}^{M}\sum_{j=1}^{M}\left|c\left(\left|\mathbf{S}_i\right|,\left|\mathbf{S}_j\right|\right)-c\left(\left|\hat{\mathbf{S}}_i\right|,\left|\mathbf{S}_j\right|\right)\right|}{M^2} \tag{12}$$

such that $c(\mathbf{S}_i, \mathbf{S}_j) = \langle \mathbf{S}_i, \mathbf{S}_j \rangle / \|\mathbf{S}_i\| / \|\mathbf{S}_j\|$, where $\|\mathbf{S}_i\|$ denotes ℓ_2-norm of $\mathbf{S}_i$.

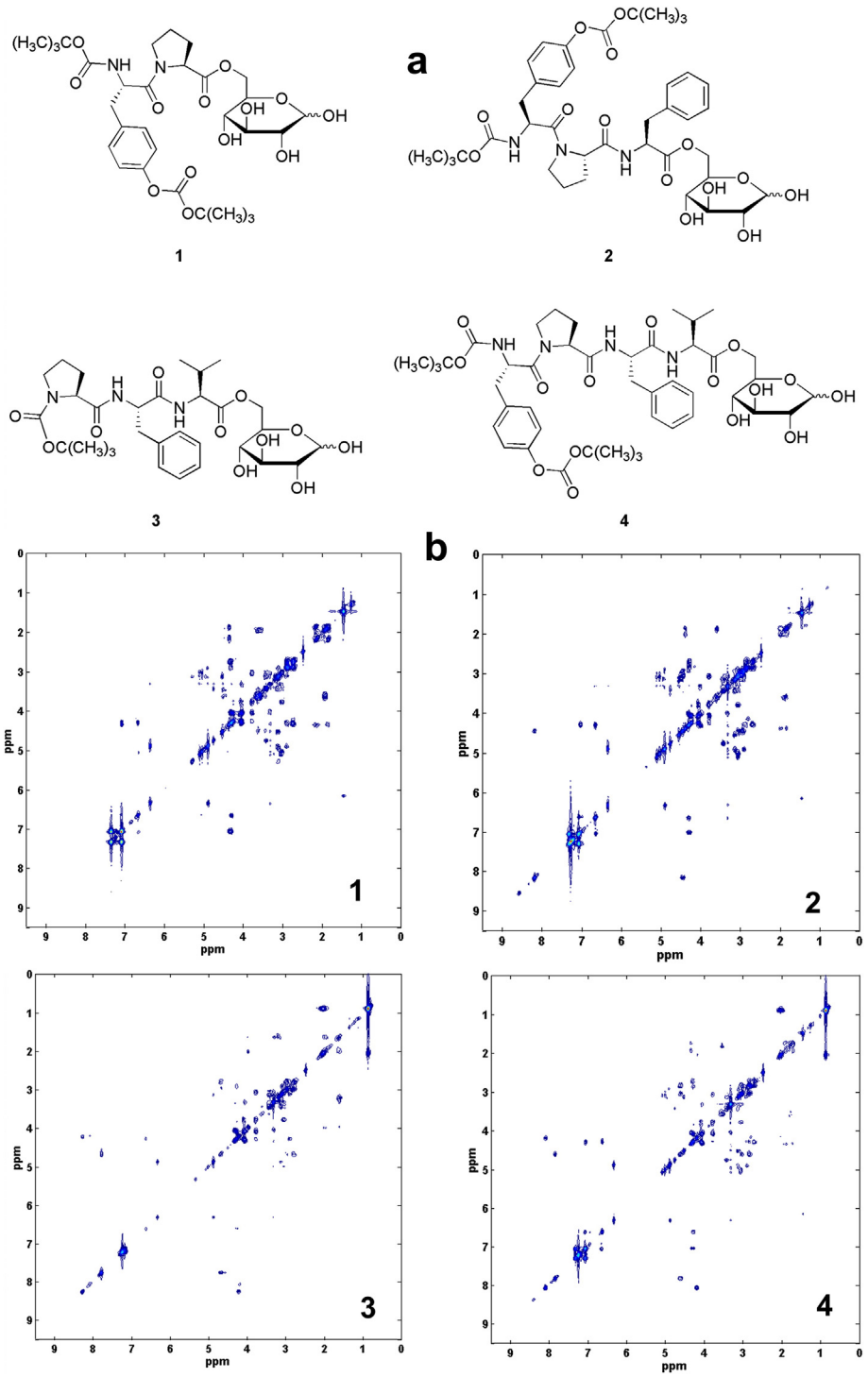

Fig. 5. COSY NMR magnitude spectra and structures of pure components **4**–**7**.

Table 2
Normalized correlation coefficients between true and estimated pure components **4**–**7** COSY NMR. The estimation error, ε, is defined in Eq. (12). The best values are in bold.

	c_{44}	c_{55}	c_{66}	c_{77}	ε
SCA	0.8468	**0.8123**	**0.8779**	0.7578	**0.1026**
NMU-S	0.8742	0.8019	0.7313	**0.8342**	0.1177
NMU-A	**0.8883**	0.6840	0.7164	0.8267	0.1251

Fig. 3 shows ^{1}H NMR magnitude spectra of the mixtures $\mathbf{X}_1$ and $\mathbf{X}_2$, while Fig. 4 shows ^{1}H NMR magnitude spectra of pure components **1, 2** and **3**, estimated by the NMU-S algorithm proposed herein.

It is observed from Table 1 that the NMU-S algorithm and SCA algorithm reported in [34] yielded virtually the same performance in extraction of three correlated pure component ^{1}H NMR spectra from two mixtures. Thus, the approximate linear mixture model (5) is experimentally grounded. In contrast, the NMU-A algorithm yielded significantly worse separation performance, implying that the linear mixture model (2), which is implicitly assumed by the NMU-A approach, is inappropriate. As demonstrated in Fig. 1, mixture model (3) moves away from linearity when the number of analytes, k, simultaneously present at some frequency grows. The NMU-S algorithm achieved similar performance as the SCA algorithm without explicitly demanding the "single component point" assumption. As discussed in [34], to identify those points, time domain NMR signals have to be transformed into the wavelet domain by selecting the appropriate wavelet function. Afterwards, the concentration matrix ought to be estimated by tuning parameter dependent data clustering. The NMU-S algorithm simplifies significantly the component extraction procedure. The normalized correlation coefficients between four pure component COSY NMR spectra, shown in Fig. 5, were: $c_{45} = 0.6333$, $c_{46} = 0.2535$, $c_{47} = 0.4998$, $c_{56} = 0.3937$, $c_{57} = 0.6078$ and $c_{67} = 0.8142$. Due to high correlation between component spectra, the related underdetermined BSS problem is demanding. Table 2 reports normalized correlation coefficients between pure component COSY NMR spectra and COSY NMR spectra of the components estimated by the SCA algorithm [35], NMU-S algorithm and NMU-A algorithm. Also reported is the average absolute value of the error between the true correlation matrix and the correlation matrix between the estimated and true spectra ε in Eq. (12). Fig. 6 shows COSY NMR magnitude spectra of the mixtures $\mathbf{X}_1$ to $\mathbf{X}_3$, while Fig. 7 shows COSY NMR magnitude spectra of the pure components **4** to **7** estimated by means of the NMU-S algorithm proposed herein. Again, the NMU-S algorithm and the SCA algorithm [35] yielded very comparable performance, even though "single component points" were not explicitly required by the NMU-S algorithm. The achieved performance confirmed practical validity of the approximate linear mixture model (5). Due to the second dimension added by COSY NMR, overlapping between the peaks is decreased. That is why performance of the NMU-A algorithm is

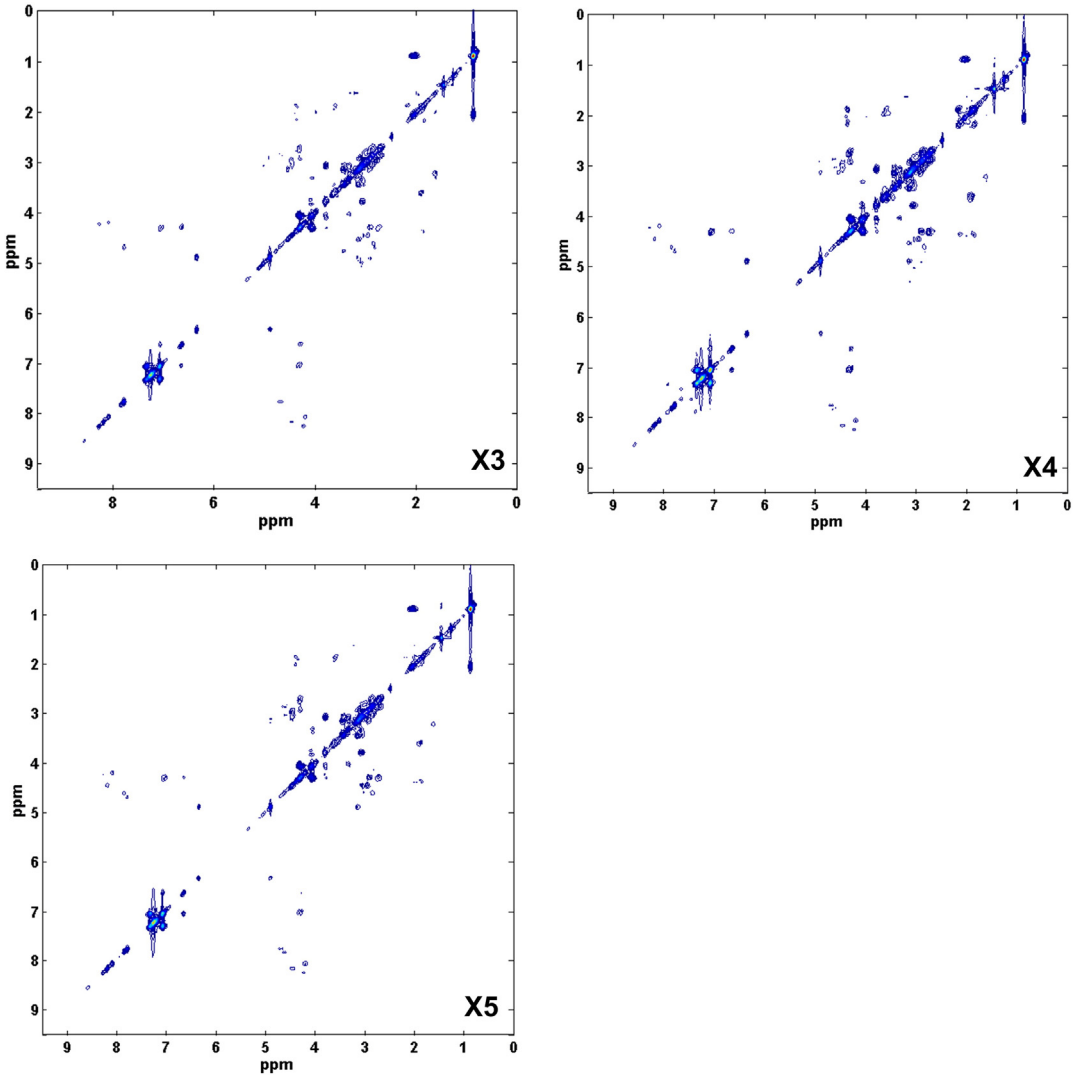

Fig. 6. COSY NMR magnitude spectra of mixtures $\mathbf{X}_3$–$\mathbf{X}_5$.

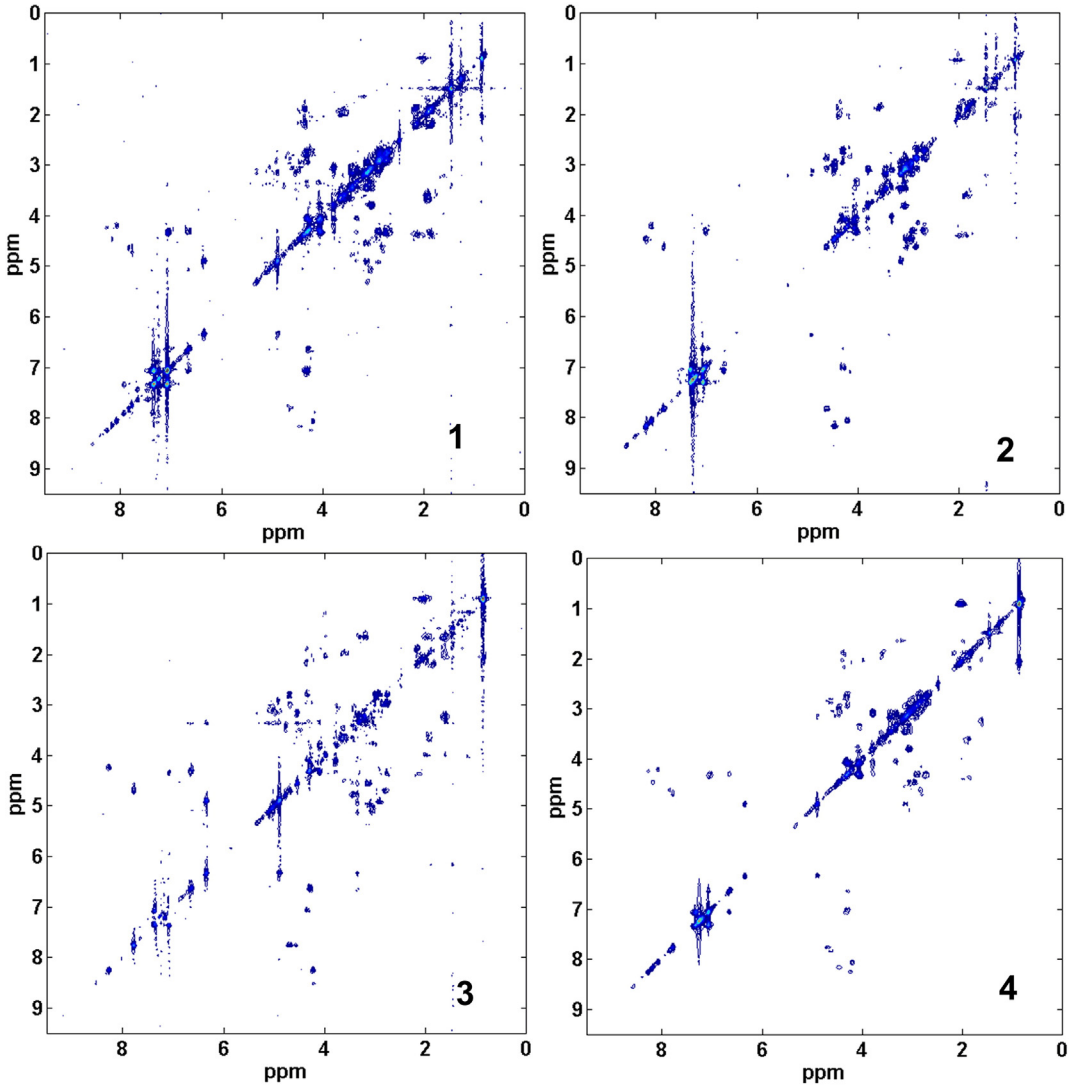

Fig. 7. COSY NMR magnitude spectra of components **4**–**7** estimated by the NMU-S algorithm.

compared more favorably than in the case of the ^{1}H NMR mixtures. It is however, in average, still worse than the performance achieved by the SCA algorithm [35] and the NMU-S algorithm. In summary, the proposed method based on sparseness constrained NMF and squared amplitude mixture model (5) enables blind extraction of correlated NMR component spectra from smaller number of mixture spectra without explicitly demanding existence of the "single analyte points", as well as without demanding *a priori* information about tuning parameters. That makes it practically relevant.

5. Conclusions

Quantitative metabolomics has shown tremendous potential for studying the nature of biological processes. However, the development of analytical tools for the analysis of complex datasets is necessary for full development of this potential. Samples of biological origin (plasma, urine, saliva or tissues) contain a large number of compounds. Because of this complexity, most state-of-the-art MCR methods fail to provide unambiguous results in NMR spectra analysis. Nevertheless, these methods are anticipated to serve as a screening or diagnostic tool in biomedical research and clinical studies. Proposed pre- and post-processing methods can enable more accurate extraction of correlated analyte NMR spectra, as well as their concentrations, from a smaller number of mixtures by using state-of-the-art sparseness-constrained NMF algorithms. By selection of the NMU algorithm and the like, the demand for *a priori* knowledge of the tuning parameters, such as the sparseness-related regularization constant or the explicit knowledge of "single analyte points", is removed. It is conjectured that the proposed method can play an important role in identification of metabolites in biomarker identification studies, and that is one of the most challenging tasks in chemical biology. In particular, it is expected that application of the proposed method on NMR spectra mapped in reproducible kernel Hilbert space (see Ref. [47]) will enable more accurate separation of pure components that are present in mixture spectra in small concentrations. It is also anticipated that the proposed method could increase efficiency of spectral library search procedures by reducing the number of false positives and negatives.

Conflict of interest

Authors declare no conflict of interest.

Acknowledgments

This work has been supported through grant 9.01/232 "Nonlinear component analysis with applications in chemometrics and pathology" funded by the Croatian Science Foundation.

References

[1] J.K. Nicholson, J. Connelly, J.V. Lindon, E. Holmes, Metabonomoics: a platform for studying drug toxicity and gene function, Nat. Rev. Drug Discov. 1 (2002) 153–161.
[2] J. Keiser, U. Duthale, J. Utzinger, Update on the diagnosis and treatment of food-bone trematode infections, Curr. Opin. Infect. Dis. 23 (2010) 513–520.
[3] D.G. Robertson, Metabonomics in toxicology: a review, Toxicol. Sci. 82 (2005) 809–822.
[4] T. Hyotylainen, Novel methods in metabolic profiling with a focus on molecular diagnostic applications, Expert. Rev. Mol. Diagn. 12 (2012) 527–538.
[5] N.R. Patel, M.J.W. McPhail, M.I.F. Shariff, H.C. Keun, S.D. Taylor-Robinson, Biofluid metabpnomics using ^{1}H NMR spectroscopy: the road to biomarker discovery in gastroenterology and hepatology, Expert Rev. Gastroenterol. Hepatol. 6 (2012) 239–251.
[6] D.S. Wishart, Metabonomics: applications to food science and nutrition research, Trends Food Sci. Technol. 19 (2008) 482–493.
[7] S. Durand, M. Sancelme, P. Besse-Hoggan, B. Combourieu, Biodegradation pathway of mesotrione: complementarities of NMR, LC–NMR and LC–MS for qualitative and quantitative metabolic profiling, Chemosphere 81 (2010) 372–380.
[8] E.M. Lenz, D.I. Wilson, Analytical strategies in metabonomics, J. Proteome Res. 6 (2007) 443–458.
[9] S.L. Robinette, R. Brüschweiler, F.C. Schroeder, A.S. Edison, NMR in metabolomics and natural products research: two sides of the same coin, Acc. Chem. Res. 45 (2012) 288–297.
[10] R.R. Forseth, F.C. Schroeder, NMR spectroscopic analysis of mixtures: from structure to function, Curr. Opin. Chem. Biol. 15 (2011) 38–47.
[11] A. Smolinksa, L. Blanchet, L.M.C. Buydens, S.S. Wijmenga, NMR and pattern recognition methods in metabolomics: from data acquisition to biomarker discovery: a review, Anal. Chim. Acta. 750 (2012) 82–97.
[12] R. Schicho, R. Shaykhutdinov, J. Ngo, A. Nazyrova, C. Schneider, R. Panaccione, G.G. Kaplan, H.J. Vogel, M. Storr, Quantitative metabolomics profiling of serum, plasma and urine by ^{1}H NMR spectroscopy discriminates between patients with inflammatory bowel disease and healthy individuals, J. Proteome Res. 11 (2012) 3344–3357.
[13] J.K. Nicholson, J.C. Lindon, Systems biology: metabonomics, Nature 455 (2008) 1054–1056.
[14] M. Abu-Farha, F. Elisma, H. Zhou, R. Tian, M.S. Asmer, D. Figeys, Proteomics: from technology developments to biological applications, Anal. Chem. 81 (2009) 4585–4599.
[15] C. Li, J. Han, Q. Huang, B. Li, Z. Zhang, C. Guo, An effective two-stage spectral library search approach based on lifting wavelet decomposition for complicated mass spectra, Chemom. Intell. Lab. Syst. 132 (2014) 75–81.
[16] B.R. Seavey, E.A. Farr, W.M. Westler, J.L. Markley, A relational database for sequence-specific protein NMR data, J. Biomol. NMR 1 (1991) 217–236.
[17] The NIST 11 Mass Spectral Library, web site http://www.sisweb.com/software/ms/nist.htm.
[18] V.A. Likić, Extraction of pure components from overlapped signals in gas chromatography–mass spectrometry (GC–MS), BioData Min. 2 (2009) 6 (11 pages).
[19] S. Kim, I. Koo, J. Jeong, S. Wu, X. Shi, X. Zhang, Compound identification using partial and semipartial correlations for gas chromatography–mass spectrometry data, Anal. Chem. 84 (2012) 6477–6487.
[20] C. Shao, W. Sun, F. Li, R. Yang, L. Hnag, Y. Gao, Oscore: a combined score to reduce false negative rates for peptide identification in tandem mass spectrometry analysis, J. Mass Spectrom. 44 (2009) 25–31.
[21] J. Razumovskaya, V. Olman, D. Xu, E.C. Uberbacher, N.C. VerBerkmoes, R.L. Hettich, Y. Xu, A computational method for assessing peptide-identification reliability in tandem mass spectrometry analysis with SEQUEST, Proteomics 4 (2004) 961–969.
[22] T. Baczek, A. Bucinski, A.R. Ivanov, R. Kaliszan, Artificial neural network analysis for evaluation of peptide MS/MS spectra in proteomics, Anal. Chem. 76 (2004) 1726–1732.
[23] B. Bracewell, Fourier Transform and its Applications, MacGraw-Hill, New York, US, 1999.
[24] V.A. Shashilov, I.K. Lednev, Advanced statistical and numerical methods for spectroscopic characterization of protein structural evaluation, Chem. Rev. 110 (2010) 5692–5712.
[25] P. Comon, C. Jutten (Eds.), Handbook of Blind Source Separation, Academic Press, Oxford, UK, 2010.
[26] I.T. Joliffe, Principal Component Analysis, Springer Series in Statistics, 2nd ed. Springer, New York, US, 2002.
[27] A. Hyvärinen, J. Karhunen, E. Oja, Independent Component Analysis, Wiley, New York, US, 2001.
[28] P. Comon, Independent component analysis, a new concept? Signal Proc. 36 (1994) 287–314.
[29] M. Zibulevsky, B.A. Pearlmutter, Blind source separation by sparse decomposition, Neural Comput. 13 (2001) 863–882.
[30] P. Georgiev, F. Theis, A. Cichocki, Sparse component analysis and blind source separation of underdetermined mixtures, IEEE Trans. Neural Networks 16 (2005) 992–996.
[31] A. Cichocki, R. Zdunek, A.H. Phan, S.I. Amari, Nonnegative Matrix and Tensor Factorizations, John Wiley, Chichester, UK, 2009.
[32] D. Nuzillard, S. Bourg, J.M. Nuzilard, Model-free analysis of mixtures by NMR using blind source separation, J. Magn. Reson. 133 (1998) 358–363.
[33] E. Visser, T.W. Lee, An information-theoretic methodology for the resolution of pure component spectra without prior information using spectroscopic measurements, Chemom. Intell. Lab. Syst. 70 (2004) 147–155.
[34] I. Kopriva, I. Jerić, V. Smrečki, Extraction of multiple pure component ^{1}H and ^{13}C NMR spectra from two mixtures: novel solution obtained by sparse component analysis-based blind decomposition, Anal. Chim. Acta. 653 (2009) 143–153.
[35] I. Kopriva, I. Jerić, Blind separation of analytes in nuclear magnetic resonance spectroscopy and mass spectrometry: sparseness-based robust multicomponent analysis, Anal. Chem. 82 (2010) 1911–1920.
[36] D.A. Snyder, F. Zhang, S.L. Robinette, L. Brüschweiler-Li, R. Brüschweiler, Non-negative matrix factorization of two-dimensional NMR spectra: application to complex mixture analysis, J. Chem. Phys. 128 (2008) 052313 (4 pages).
[37] S. Du, P. Sajda, R. Stoyanova, T. Brown, Recovery of metabolomic spectral sources using non-negative matrix factorization, Proc. of the 2005 IEEE Eng. Med. Biol. Soc. 27th Ann. Conf, 2005, pp. 1095–1098.
[38] I. Toumi, B. Torrésani, S. Caldarelli, Effective processing of pulse field gradient NMR of mixtures by blind source separation, Anal. Chem. 85 (2013) 11344–11351.
[39] W.S.B. Ouedraogo, A. Souloumiac, M. Jaïdane, C. Jutten, Non-negative blind source separation algorithm based on minimum aperture simplical cone, IEEE Trans. Signal Proc. 62 (2014) 376–389.
[40] L. Guo, A. Wiesmath, P. Sprenger, M. Garland, Development of 2D band-target entropy minimization and application to deconvolution of multicomponent 2D nuclear magnetic resonance spectra, Anal. Chem. 77 (2005) 1655–1662.
[41] W. Naanaa, J.M. Nuzillard, Blind source separation of positive and partially correlated data, Signal Proc. 85 (2005) 1711–1722.
[42] M.S. Karoui, Y. Deville, S. Hosseini, S. Ouamri, Blind spatial unmixing of multispectral images: new methods combining sparse component analysis, clustering and nonnegativity constraints, Pattern Recog. 45 (2012) 4263–4278.
[43] Y. Sun, C. Ridge, F. del Rio, A.J. Shaka, J. Xin, Postprocessing and sparse blind source separation of positive and partially overlapped data, Signal Proc. 91 (2011) 1838–1851.
[44] N. Gillis, F. Glineur, Using underapproximations for sparse nonnegative matrix factorization, Pattern Recog. 43 (2010) 1676–1687.
[45] A. Cichocki, R. Zdunek, S.I. Amari, Algorithms for nonnegative matrix factorization and 3D tensor factorization, Lect. Notes Comput. Sci 4666 (2007) 169–176.
[46] R. Peharz, F. Pernkopf, Sparse nonnegative matrix factorization with ℓ^0-constraints, Neurocomputing 80 (2012) 38–46.
[47] I. Kopriva, I. Jerić, L. Brkljačić, Nonlinear mixture-wise expansion approach to underdetermined blind separation of nonnegative dependent sources, J. Chem. 27 (2013) 189–197.
[48] P.O. Hoyer, Non-negative matrix factorization with sparseness constraints, J. Mach. Learn. Res. 5 (2004) 1457–1469.
[49] F. Jiru, Introduction to post-processing techniques, Eur. J. Radiol. 67 (2008) 202–217.
[50] The Nicolas Gillis, web site https://sites.google.com/site/nicolasgillis/code.
[51] K. Huang, N.D. Sidiropoulos, A. Swami, Non-negative matrix factorization revisited: uniqueness and algorithms for symmetric decomposition, IEEE Trans. Signal Proc. 62 (2014) 211–224.
[52] E.R. Malinowski, Factor Analysis in Chemistry, 3rd ed. John Wiley & Sons, Inc., New York, US, 2002.
[53] S.J. Kim, K. Koh, M. Lustig, S. Boyd, S. Gorinevsky, A method for large-scale l1-regularized least squares, IEEE J. Sel. Top. Signal Proc. 1 (2007) 606–617.
[54] I. Jerić, Š. Horvat, Novel ester-linked carbohydrate—peptide adducts: effect of the peptide substituent on the pathways of intramolecular reactions, Eur. J. Org. Chem. 2001 (2001) 1533–1539.