Blind Separation of Analytes in Nuclear Magnetic Resonance Spectroscopy: Improved Model for Nonnegative Matrix Factorization

Ivica Kopriva[†*] *and Ivanka Jerić*[‡]

Ruđer Bošković Institute, Bijenička cesta 54, HR-10000, Zagreb, Croatia

[†]Division of Laser and Atomic Research and Development

[‡]Division of Organic Chemistry and Biochemistry

[*]ikopriva@irb.hr; Tel.: +385-1-4571-286; Fax: +385-1-4680-104

Abstract

We introduce improved model for sparseness constrained nonnegative matrix factorization (sNMF) of amplitude mixtures nuclear magnetic resonance (NMR) spectra into greater number of component spectra. In proposed method selected sNMF algorithm is applied to the square of the amplitude of the mixtures NMR spectra instead to the amplitude spectra itself. Afterwards, the square roots of separated squares of components spectra and concentration matrix yield estimates of the true components amplitude spectra and of concentration matrix. Proposed model

remains linear in average when number of overlapping components is increasing, while model based on amplitude spectra of the mixtures moves away from the linear one when number of overlapping components is increased. That is demonstrated through conducted sensitivity analysis. Thus, proposed model improves capability of the sparse NMF algorithms to separate correlated (overlapping) components spectra from smaller number of mixtures NMR spectra. That is demonstrated on two experimental scenarios: extraction of three correlated components spectra from two ^{1}H NMR mixtures spectra and extraction of four correlated components spectra from three COSY NMR mixtures spectra. Proposed method can increase efficiency in spectral library search by reducing occurrence of false positives and false negatives. That, in turn, can yield better accuracy in biomarker identification studies which makes proposed method important for natural products research and the field of metabolic studies.

Keywords: Nuclear magnetic resonance spectroscopy, (non-)linear mixture model, blind source separation, nonnegative matrix factorization, compound identification.

1 Introduction

Metabolites, low-molecular-weight compounds, are functional endpoints of metabolism and are reflection of genetic and environmental perturbations of the system. Measurement of metabolites in biological fluids, typically urine and serum, is actually measurement of living system's responses to disease, drugs or toxins. Metabolic profiling is therefore indispensable tool in drug development [1, 2], toxicology studies [3], disease diagnosis [4, 5], food, nutrition and environmental sciences [6]. Nuclear magnetic resonance (NMR) spectroscopy is emerging as a

key technique in metabolomics in an attempt to identify and quantify individual compounds the biological fluids are composed of [8, 9, 10]. The problem is notoriously difficult owing to the presence of large number of analytes in 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 by use of ^{1}H NMR spectroscopy [12]. Since many analytes are structurally similar, their NMR spectra are highly correlated with many peaks overlapping. It is thus, the complexity of samples that limits identification of analytes, that is seen as one of the most challenging tasks in chemical biology [13]. Compound identification is often achieved by matching experimental spectra to the ones stored in the library [14, 15], for an example BioMagResBank metabolomics database [16] or, in 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 in 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 number of analytes in a mixture spectra 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 diversity of datasets is high or number of spectra from a specific group is small accuracy in analytes identification will deteriorate. Moreover, accuracy will be affected further by the overlapping of analytes 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 concentration and spectra of individual components from multicomponent mixtures spectra [24]. In particular, blind source separation (BSS) methods [25] refer to class of multivariate data analysis methods capable of blind (unsupervised) extraction of analytes from mixtures spectra, i.e. concentrations of analytes are not required to be known to the BSS algorithms. It is however clear that under stated conditions related inverse problem is severely ill-posed. To narrow-down infinite number of solutions to, possibly, essentially unique one, constraints have to be imposed on analytes 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 on analytes extraction from spectroscopic mixtures [32-39]. PCA, ICA and many NMF algorithms require that the *unknown* number of analytes is less than or equal to the number of mixtures spectra available [32, 33, 36-39]. That is also true for many "deconvolution" methods [40]. This makes them inapplicable for the analysis of multicomponent mixtures spectra such as those acquired from biological samples. Sparseness-based approaches to BSS are currently highly active research area in signal processing. Unlike PCA and ICA methods, SCA methods enable solution of an underdetermined BSS problem, i.e. extraction of more analytes than mixtures available in 1D and 2D NMR spectroscopy [34, 35]. Sparseness implies that at each frequency (in a case of NMR spectroscopy) only small number of analytes is active.

[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

However, 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 complexity of mixture grows and when, due to reasons elaborated previously, multiple analytes get 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 blind separation of positive and partially overlapped sources require that each analyte is 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, that in addition to nonnegativity also use sparseness constraint, are capable to solve 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 sparseness related regularization parameter are of practical value [44]. However, in majority of cases the NMF algorithms have been applied to extract number of components that is smaller than number of available mixtures NMR spectra [37, 38]. Herein, we demonstrate how sparseness constrained NMF ought to be applied to mixtures NMR spectra to improve quality of separation of correlated NMR components spectra. It is conjectured that 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 linearity of proposed method against number of overlapping components is compared with amplitude mixtures spectra-based model and demonstrated through sensitivity analysis. Proposed method is further compared with state-of-the-art SCA algorithms. To this end, three highly correlated ^{1}H NMR components

---

to solving system of linear equations that describes memoryless (instantaneous) system with multiple inputs (analytes) and multiple outputs (mixtures spectra).

spectra are were extracted from two mixtures [34] and four highly correlated COSY NMR components spectra were extracted from three mixtures [35].

2 Theory and method

2.1 Linear mixture model of multicomponent NMR spectra

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 amplitude decaying exponentially with some time constant, [49], linear mixture model in the absence of additive noise reads as:

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

where $\mathbf{X} \in \mathbb{C}^{N \times T} =: \left\{ \mathbf{x}_n \in \mathbb{C}^{1 \times T} \right\}_{n=1}^{N}$ represents 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". $\mathbf{A} \in \mathbb{R}_{0+}^{N \times M} =: \left\{ \mathbf{a}_m \in \mathbb{R}_{0+}^{N \times 1} \right\}_{m=1}^{M}$ represents mixture (a.k.a. concentration) matrix, whereas each column vector represents concentration profile of one of the M analytes across the N mixtures. $\mathbf{S} \in \mathbb{C}^{M \times T} =: \left\{ \mathbf{s}_m \in \mathbb{C}^{1 \times T} \right\}_{m=1}^{M}$ is a matrix with the rows

representing NMR temporal signals of the analytes present in the mixture signals $\mathbf{X}$.[2] Thereby, it is assumed that $M > N$. That leads to underdetermined BSS problem in which case it is assumed that information about concentration of analytes, stored in the mixing matrix $\mathbf{A}$, is not known to the BSS algorithm. Thus, it is expected from BSS method to estimate matrix of analytes $\mathbf{S}$ and matrix of concentrations $\mathbf{A}$ by having at disposal matrix with recorded mixture signals $\mathbf{X}$ only. However, amplitude spectra of the NMR signals, that are of the actual interest, are amplitudes of the Fourier transform of the corresponding time domain NMR signals. Due to linearity of Fourier transform it yields linear mixture model in frequency domain with the same structure as (1), whereas T time domain instants are now interpreted as T frequencies. However, NMF algorithms are inapplicable to (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. 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 linear mixture model [37]:

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

[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 frequency domain multidimensional Fourier transform is applied to multidimensional time domain NMR signals mixture-wise: $\{\mathbf{X}_n =: F(\mathbf{x}_n)\}_{n=1}^{N}$, where F stands for Fourier transform of appropriate dimension.

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

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

$$0\leq k\leq M,\ 1\leq t\leq T,\ \ 1\leq n\leq N \qquad (3)$$

where Re($\mathbf{S}_i$), resp. Im($\mathbf{S}_i$), stand for 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

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

stands for cross-terms that are explicitly dependent on k. Thus, 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:

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

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

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

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

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

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

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

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

2.2 Sparseness constrained factorization.

The underdetermined BSS problem (6) is ill-posed because matrix factorization suffers from indeterminacies: $\left|\mathbf{X}\right|^{squared} \approx \mathbf{A}^{squared}\left|\mathbf{S}\right|^{squared} = \mathbf{A}^{squared}\mathbf{D}\mathbf{D}^{-1}\left|\mathbf{S}\right|^{squared}$ 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 $\mathbf{D}=\mathbf{P}\boldsymbol{\Lambda}$, where **P** represents permutation and $\boldsymbol{\Lambda}$ represents diagonal scaling matrix. Constraints are necessary to be imposed on $\mathbf{A}^{squared}$ and $\left|\mathbf{S}\right|^{squared}$ to obtain solution of (6) unique up to permutation and scaling indeterminacies. For underdetermined BSS (uBSS) problems, of interest herein, the necessary constraint is sparseness of squares of analytes spectra stored in rows of $\left|\mathbf{S}\right|^{squared}$. Due to the character of the problem, nonnegativity constraint is imposed on $\mathbf{A}^{squared}$ and $\left|\mathbf{S}\right|^{squared}$ as well i.e. $\mathbf{A}^{squared} \geq \mathbf{0}$ and $\left|\mathbf{S}\right|^{squared} \geq \mathbf{0}$. While several methods are

available for solving sparseness constrained NMF problem (6) [44-48], in the experiments reported below we have used the nonnegative matrix under-approximation (NMU) algorithm [44] with a MATLAB code available at [50]. The NMU method performs factorization of (6) in a recursive manner extracting one component at a time. After identifying optimal rank-one solution $\left(\mathbf{a}_1^{squared}, |\mathbf{s}_1|^{squared}\right)$ 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 non-negativity of $|\mathbf{X}|^{squared}$ an underapproximation constraint is imposed on $\mathbf{A}^{squared}$ and $|\mathbf{S}|^{squared}$: $\mathbf{A}^{squared} |\mathbf{S}|^{squared} \leq |\mathbf{X}|^{squared}$. It has been proven in theorem 1 in [44] that number of non-zero entries of $\mathbf{A}^{squared}$ and $|\mathbf{S}|^{squared}$ is less than number of non-zero entries of $|\mathbf{X}|^{squared}$. That is important in light of the very recent result proved in [51], see Theorem 4 and Corollary 2, that uniqueness of some asymmetric NMF **X=WH** 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 given to the algorithm as an input. It is emphasized in [52] and recently in [38] that no criterion for determining 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 computational example related to comparative sensitivity analysis of models (3) and (5) and two experiments: blind extraction of three analytes ^{1}H NMR spectra from two mixtures and blind extraction of four analytes 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 SCA approach to blind extraction of analytes and their concentrations. The SCA approach explicitly demands observation points (not necessarily in 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 matrix of concentrations **A**. Afterwards, either linear program or least square program regularized by ℓ_1-norm (implemented by interior-point method) [53] had to be solved in frequency domain to estimate amplitude spectra of the analytes (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 proposed methodology, which applies the NMU algorithm in (6) on squares of the mixtures 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 NMU algorithm on the amplitude mixtures NMR spectra (the NMU-A) in order to demonstrate deterioration in accuracy of the estimated analytes amplitude spectra. For the purpose of completeness the experiments reported in [34, 35] are briefly described here.

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

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. to the number of analytes $0 \le k \le M$ simultaneously active at some frequency ω_t, $t=1,..,T$. Therefore, we calculate variation of $\left|\mathbf{X}_n(\omega_t)\right|$ in (3) w.r.t. $\left|\mathbf{S}_m(\omega_t)\right|$ as well as variation of $\left|\mathbf{X}_n(\omega_t)\right|^2$ in (5) w.r.t. $\left|\mathbf{S}_m(\omega_t)\right|^2$ as follows:

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

(8)

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

(9)

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

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

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

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

In simulation of eq.(8) and (9) we have assumed that component m=1 is dominantly active at frequency ω_t with amplitude $\left|\mathbf{S}_m\left(\omega_t\right)\right|=1$ and arbitrary phase $\varphi_m\left(\omega_t\right)\in\left[0,2\pi\right]$. Amplitudes and phases of other components, for $k\geq2$, 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 $\left\{a_{ni}=1\right\}_{i=1}^{k}$. That is because strength of the presence of source i=2,..,k has been regulated by random amplitude $\left|\mathbf{S}_i\left(\omega_t\right)\right|$.

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 $\mathbf{X}_3$ 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

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

4 Results and discussion

Figure 1 shows mean values (± standard deviation) of sensitivities (8) and (9) as a function of k=1,...,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 simulation setup described in section 3.1 it follows that linearity condition for model (3), and implied by eq. (10), should be $\partial\left|\mathbf{X}_n\left(\omega_t\right)\right|/\partial\left|\mathbf{S}_m\left(\omega_t\right)\right| = a_{nm} = 1$. Likewise, linearity condition for model (5) implied by

eq. (11) should be $\partial\left|\mathbf{X}_n\left(\omega_t\right)\right|^2 \Big/ \partial\left|\mathbf{S}_m\left(\omega_t\right)\right|^2 = a_{nm}^2 = 1$. It is seen that linearity condition for model (5) holds *in 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 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), 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 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 mixture of NMR spectra.

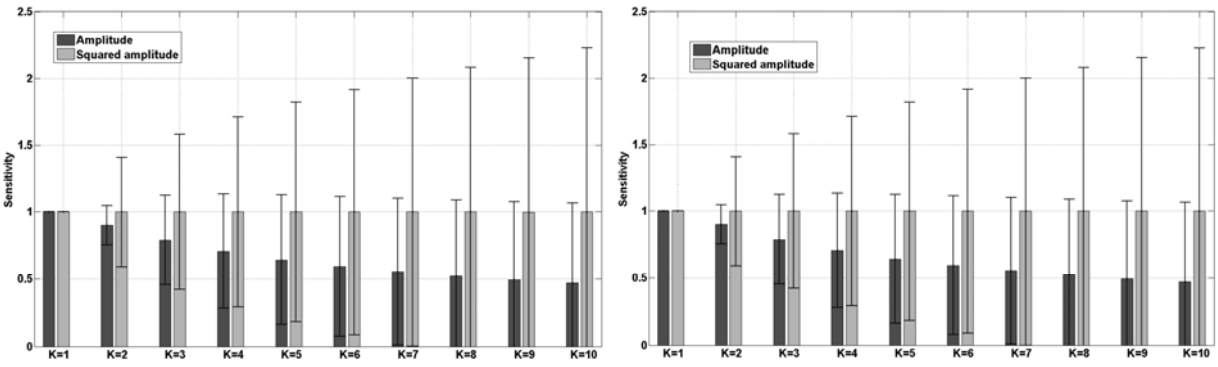

Figure 1. Sensitivities (± standard deviation) (8) and (9) of, respectively, amplitude model (3) and squared amplitude model (5) vs. number of analytes k present at some frequency ω_t. Simulation setup is described in section 3.1. Phase of component 1: left- $\varphi_1=\pi/4$, right-$\varphi_1=5\pi/7$.

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

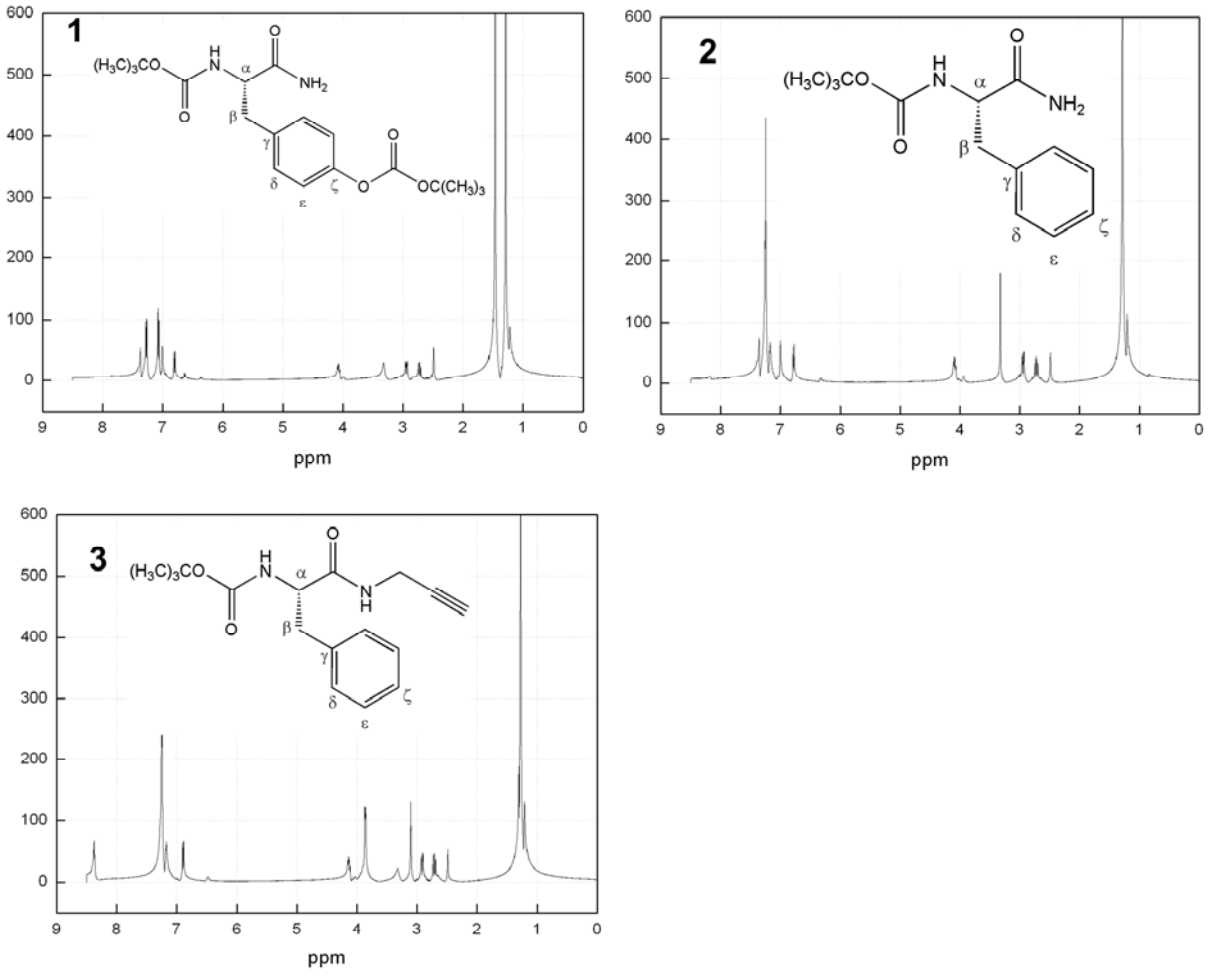

Figure 2. ^{1}H NMR magnitude spectra and structures of pure components **1**-**3**.

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

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

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

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

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

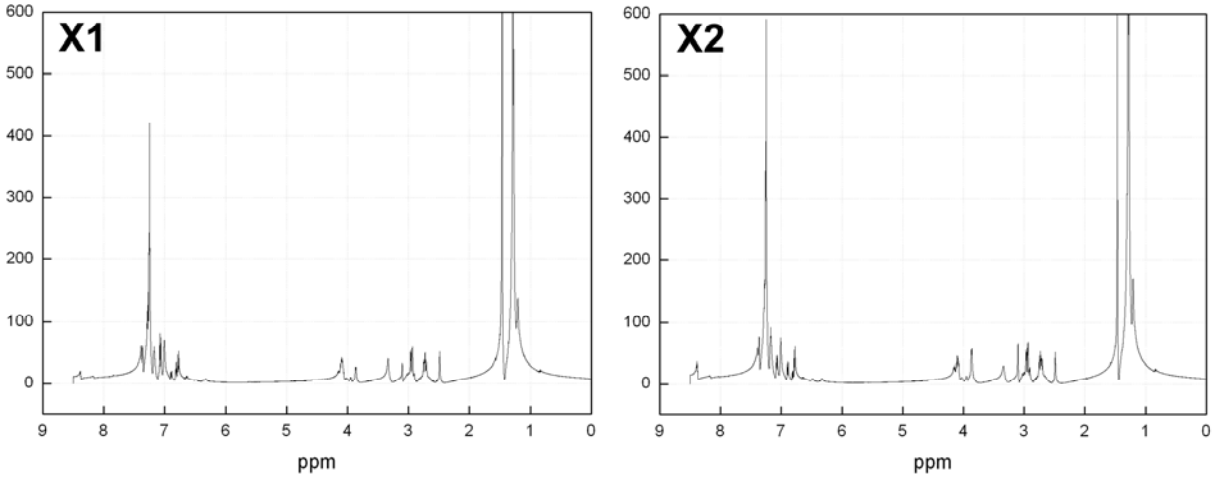

Figure 3. ^{1}H NMR magnitude spectra of mixtures:$\mathbf{X}_1$ and $\mathbf{X}_2$.

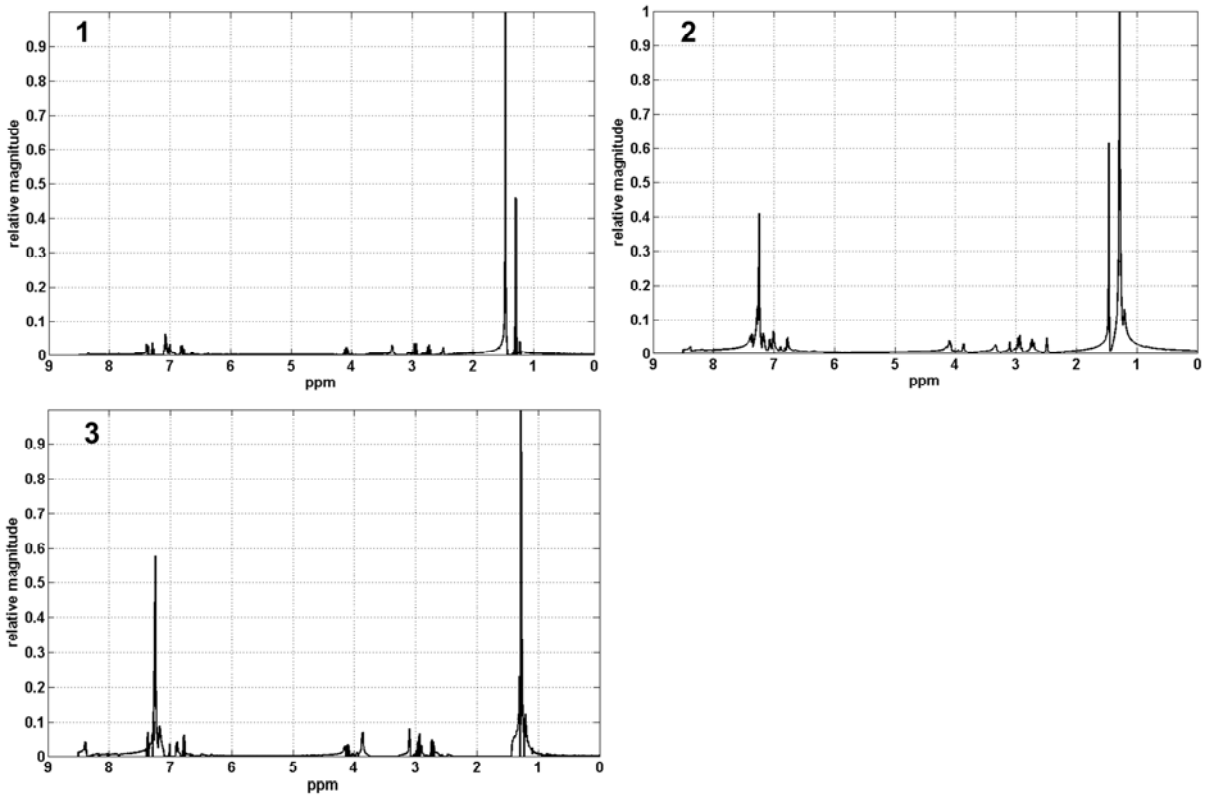

Figure 4. ^{1}H NMR magnitude spectra of pure components **1, 2** and **3** estimated by the NMU-S algorithm.

It is observed from Table 1 that the NMU-S algorithm and SCA algorithm reported in [34] yielded virtually same performance in extraction of three correlated pure components ^{1}H NMR spectra from two mixtures. Thus, approximate linear mixture model (5) is experimentally grounded. On the other side, the NMU-A algorithm yielded significantly worse separation performance implying that the linear mixture model (2), that is implicitly assumed by the NMU-A approach, is inappropriate. As demonstrated in Figure 1, mixture model (3) moves away from linearity when number of analytes k simultaneously present at some frequency grows. The NMU-S algorithm achieved similar performance as the SCA algorithm without explicitly demanding "single component point" assumption. As discussed in [34] to identify those points time domain NMR signals have to be transformed into wavelet domain by selecting appropriate wavelet function. Afterwards, concentration matrix ought to be estimated by tuning parameter dependent data clustering. NMU-S algorithm simplifies significantly components extraction procedure. Normalized correlation coefficient between four pure components COSY NMR spectra, shown in Figure 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 components spectra, related underdetermined BSS problem is demanding. Table 2 reports normalized correlation coefficients between pure components COSY NMR spectra and COSY NMR spectra of the components estimated by SCA algorithm [35], NMU-S algorithm and NMU-A algorithm. It is also reported the average absolute value of the error between true correlation matrix and correlation matrix between estimated and true spectra ε in (12). Figure 6 shows COSY NMR magnitude spectra of the

mixtures $\mathbf{X}_1$ to $\mathbf{X}_3$, while Figure 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 the SCA algorithm [35] yielded very comparable performance, even though "single component points" were not explicitly required by the NMU-S algorithm. Achieved performance confirmed practical validity of the approximate linear mixture model (5). Due to second dimension added by COSY NMR, overlapping between the peaks is decreased. That is why performance of the NMU-A algorithm is compared more favorably than in 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, proposed method based on sparseness constrained NMF and squared amplitude mixture model (5) enables blind extraction of correlated NMR components spectra from smaller number of mixtures 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.

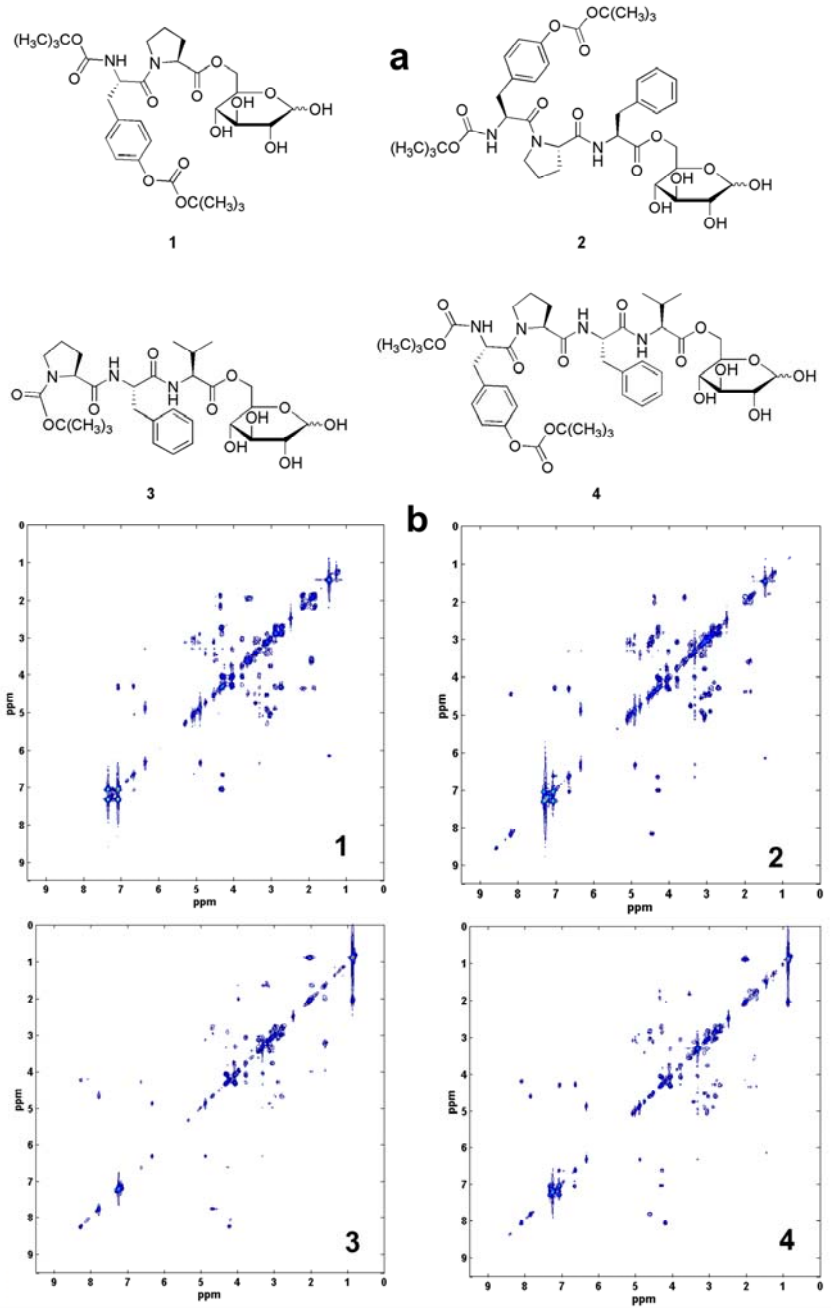

Figure 5 (color online)**.** COSY NMR magnitude spectra and structures of pure components **4**-**7**.

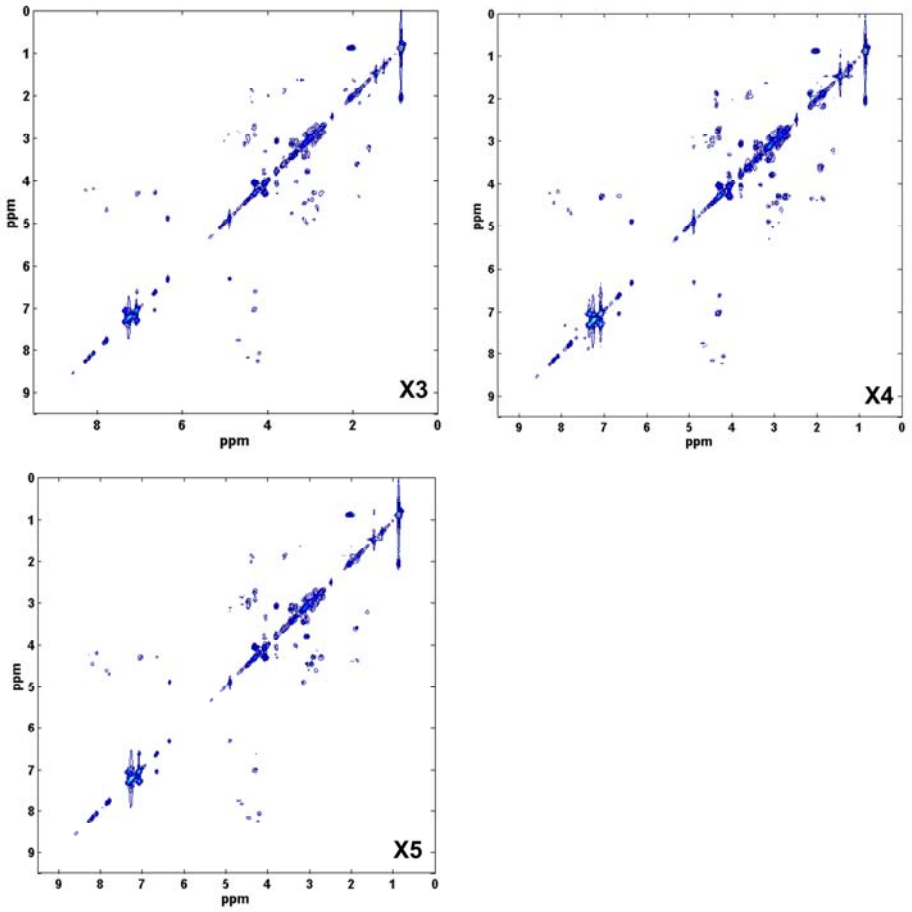

Figure 6 (color online)**.** COSY NMR magnitude spectra of mixtures $\mathbf{X}_3$-$\mathbf{X}_5$.

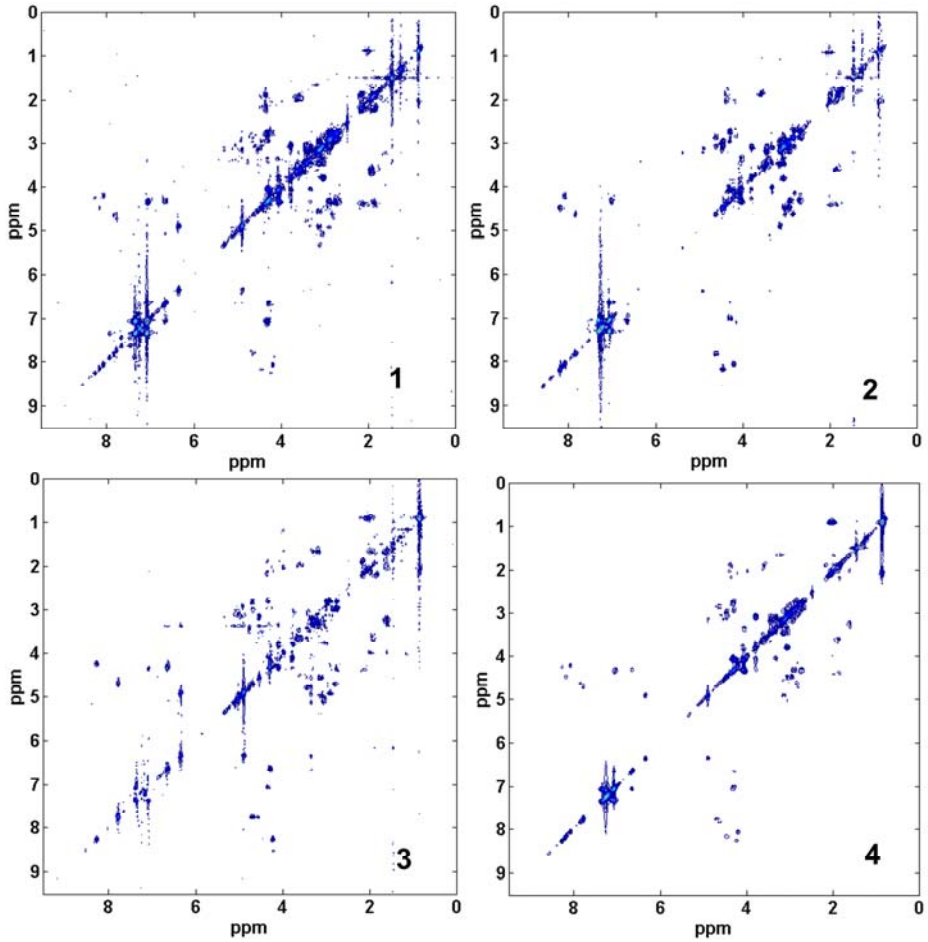

Figure 7 (color online)**.** COSY NMR magnitude spectra of components **4-7** estimated by the NMU-S algorithm.

Table 2. Normalized correlation coefficients between true and estimated pure components **4-7** COSY NMR. 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

5 Conclusions

Quantitative metabolomics has shown tremendous potential for studying nature of biological processes. However, development of analytical tools for analysis of complex datasets is what is necessary for full development of this potential. Samples of biological origin (plasma, urine, saliva or tissues) contain large number of compounds. Due to this reason most of state-of-the-art MCR methods fail to provide unambiguous results in NMR spectra analysis. Nevertheless, these methods are anticipated as a screening or diagnostic tool in biomedical research and clinical studies. Proposed pre- and post-processing method can enable more accurate extraction of correlated analytes NMR spectra as well as their concentrations form smaller number of mixtures by using state-of-the-art sparseness constrained NMF algorithms. By selection of the NMU algorithm and the like, it removes demand on *a priori* knowledge of the tuning parameters such as sparseness related regularization constant as well explicit knowledge of "single analyte points". It is conjectured that 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 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 mixtures spectra in small concentrations. It is also anticipated that proposed method could increase efficiency of spectral library search procedures by reducing number of false positives and negatives.

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 Discovery 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 pathaway of mesotrione: Complementaries of NMR, LC-NMR and LC-MS for qualitative and quantitaive 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 acquistion 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, Quantitaive Metabolomic 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, Chem. Int. Lab. Sys. 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?, Sig. 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 Net. 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. Int. 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. Cim. 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. on Sig. 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, Sig. 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, Patt. 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, Sig. 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. Not. Comp. 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, European 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. Sig. 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. Topics 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.