Research Article

Journal of MASS SPECTROMETRY

Received: 16 March 2009 Accepted: 20 July 2009 Published online in Wiley Interscience:

(www.interscience.com) DOI 10.1002/jms.1627

Multi-component analysis: blind extraction of pure components mass spectra using sparse component analysis

Ivica Kopriva[a]* and Ivanka Jerić[b]

The paper presents sparse component analysis (SCA)-based blind decomposition of the mixtures of mass spectra into pure components, wherein the number of mixtures is less than number of pure components. Standard solutions of the related blind source separation (BSS) problem that are published in the open literature require the number of mixtures to be greater than or equal to the unknown number of pure components. Specifically, we have demonstrated experimentally the capability of the SCA to blindly extract five pure components mass spectra from two mixtures only. Two approaches to SCA are tested: the first one based on ℓ_1 norm minimization implemented through linear programming and the second one implemented through multilayer hierarchical alternating least square nonnegative matrix factorization with sparseness constraints imposed on pure components spectra. In contrast to many existing blind decomposition methods no *a priori* information about the number of pure components is required. It is estimated from the mixtures using robust data clustering algorithm together with pure components concentration matrix. Proposed methodology can be implemented as a part of software packages used for the analysis of mass spectra and identification of chemical compounds.

Keywords: mass spectrometry; chemometrics; blind source separation; sparse component analysis; nonnegative matrix factorization

Introduction

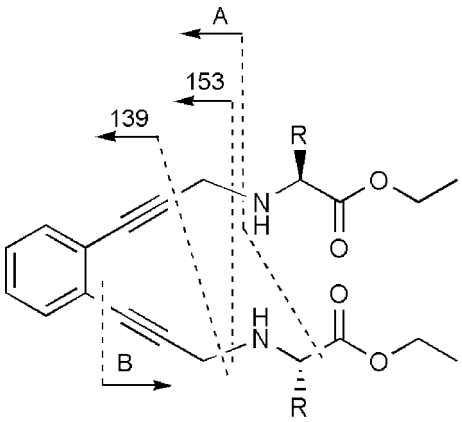
During the past decade, mass spectrometry (MS) has been rapidly developing to meet ever-increasing requirements of natural sciences. It has been successfully applied on complex biological samples with numerous analytes having diverse physico-chemical properties and being present in different concentrations.[1] MS is closely related to the study of proteins and proteosomes in post-genome times, and MS proteomics has been suggested to become an early primary screening approach to disease diagnosis.[2] Moreover, mass spectrometry is occupying a central position in the methodologies developed for determination of the metabolic state. Profiling of metabolites in biological fluids is useful in screening for disease biomarkers.[3,4] Also, MS is used in pharmaceutical industry to study metabolism of xenobiotics, since it is known that many drugs may easily undergo metabolic activation to form chemically reactive metabolites capable to modify cellular macromolecules.[5] Mass spectrometry has also been used for the determination of protein-protein interactions[6] and natural product biosynthetic processes.[7]

All these examples comprise multi-component mixtures and components of biological or pharmaceutical samples should be separated before MS detection. Efficiency of this separation is not always satisfactory; it is always time- and money consuming and can be described as a bottleneck of the particular analysis. Extraction of the pure component spectra from the mixtures of their linear combinations is therefore of great interest. Classical approach to extraction of the spectra of pure components is to match the mixture's spectra with a library of reference compounds. This approach is ineffective with the accuracy strongly dependent on the library's content of the pure component spectra and can not reflect the variation of the spectral profile due to environmental changes. Alternatives to library matching approach are blind decomposition methods, wherein pure components' spectra are extracted using mixtures spectra only. Blind approaches to pure components spectra extraction have been reported in NMR spectroscopy,[8] infrared (IR)[9–11] and near infrared (NIR) spectroscopy,[11–17] EPR spectroscopy,[18,19] mass spectrometry,[11,16,17] Raman spectroscopy[18,19] etc. In a majority of blind decomposition schemes independent component analysis (ICA)[20–22] is employed to solve related blind source separation (BSS) problem. ICA assumes that (1) pure components are statistically independent, (2) at most one is normally distributed and (3) number of mixtures is greater than or equal to the unknown number of pure components. The two requirements, i.e. to have more linearly independent mixtures than pure components and to have statistically independent pure components seem to be most critical for the success of the BSS approach to blind decomposition of the mixtures spectra into pure components spectra.[11,12,15,17] Statistical independence assumption is not very likely to be fulfilled for mass spectra because they are correlated, i.e. overlapped.[16,17] An algorithm for blind decomposition of EPR spectra has been derived in Ref. [15] minimizing contrast function that exploits sparseness among the pure components. However, this method as well as all discussed blind spectra decomposition methods require the number of mixtures to be equal to or greater than

* *Correspondence to: Ivica Kopriva, Division of Laser and Atomic Research and Development, Ruđer Bošković Institute, Bijenička cesta 54, HR-10000, Zagreb, Croatia. E-mail: ikopriva@irb.hr*

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*



1 R=H
2 R=CH_3
3 R=$CH(CH_3)_2$
4 R=$CH_2CH(CH_3)_3$
5 R=$CH_2C_6H_5$

Figure 1. Structures of compounds **1–5** (pure components **1–5**).

the *unknown* number of pure components. In a number of real world situations it is however not easy to acquire mixtures spectra with different concentrations of the pure components spectra. In this regard it is a desirable property of blind decomposition methods to solve related BSS problem with as few mixtures as possible. Addressing these issues, we have applied sparse component analysis (SCA) approach to blind decomposition of *five pure components mass spectra from two mixtures only*. The aim of the presented work is to demonstrate utility of SCA for solving multi-component related problems in mass spectrometry and to assist implementation of signal processing on experimentally obtained data. Results presented here can be considered as an important step towards further expansion and even wider application of mass spectrometry in all fields of natural sciences.

Methods and Experimental

Compounds

Pure components **1–5** (Fig. 1) used in this study belong to a class of symmetrical enediyne-bridged compounds derived from glycine, alanine, valine, leucine and phenylalanine. They were prepared for the temperature-dependent cycloaromatization studies and their synthesis and properties will be published in due course.

Two mixtures consisting of compounds **1–5** were prepared by mixing different volumes of pure components' stock solutions (1 mg/ml) to obtain the following ratios: X_1 (**1:2:3:4:5** = 6:4:3:2:1) and X_2 (**1:2:3:4:5** = 1:2:3:4:6).

Mass spectrometry measurements

Electrospray ionization mass spectrometry (ESI-MS) measurements were performed on a HPLC-MS triple quadrupole instrument equipped with an autosampler (Agilent Technologies, USA) operating in a positive ion mode. Mass spectra of pure components **1–5**, acquired in a full scan mode, were recorded by injection of 3 µl of 1 mg/ml stock solution in methanol, at a flow rate of 0.2 ml/min (mobile phase methanol-0.1% formic acid, 50:50). Mass spectra of the two mixtures X_1 and X_2 were obtained by injection of 2 µl of mixture solutions prepared as described above.

Software environment

Described SCA-based approaches to blind decomposition of mass spectra were tested using custom scripts in MATLAB programming language (version 7.1.; The MathWorks, Natick, MA). The linear programming part of the SCA algorithm has been implemented using linprog command from the Optimization toolbox. The nonnegative matrix factorization part and data clustering parts of the SCA algorithm have been implemented in MATLAB programming language by us. All programs were executed on PC running under the Windows XP operating system using Intel Core 2 Quad Processor Q6600 operating with clock speed of 2.4 GHz and 4GB of RAM installed.

Results

The positive ion mass spectra of **1–5** show molecular ions at *m/z* 357 (**1**), *m/z* 385 (**2**), *m/z* 441 (**3**), *m/z* 467 (**4**) and *m/z* 537 (**5**) accompanied by traces of the corresponding $[M+Na]^+$ ions (Fig. 2). Also, certain degree of fragmentation occurs in the ion source. Low abundant peaks at the *m/z* 139 and 153 in the spectrum of glycine-related **1** and alanine-related compound **2** can be assigned as enediyne-derived fragments shown in Fig. 1. Additionally, cleavage of two bonds (structure A, Fig. 1) gives rise to low abundant ions found in spectra of all compounds **1–5** at *m/z* 180, 194, 222, 236 and 270, respectively. In the mass spectra of compounds **3–5** with bulky side-chains (Val, Leu and Phe) dominant process is cleavage of amino acid moiety, apparent in the presence of high intensity ions at *m/z* 146, 160 and 194, and the formation of structure B (Fig. 1) with *m/z* 184, 198 and 232, respectively.

Mass spectra of compounds **1–5** were used to verify the quality of results obtained by the sparse component analysis. Mass spectra of two mixtures consisting of all five compounds, shown in Fig. 3, were used as input for SCA algorithms. To solve related underdetermined blind source separation (uBSS) problem two approaches were applied: (1) combination of robust data clustering[23] and constrained ℓ_1 norm minimization,[24–26] (2) recently developed nonnegative matrix factorization (NMF) algorithm that is known as local or hierarchical alternating least squares (HALS) NMF algorithm.[27,28] Its main property is to estimate concentration or mixing matrix globally and pure components spectra locally, wherein sparseness constraints are imposed on the pure components spectra. Unlike majority of the BSS algorithms that assume the number of pure components to be known, proposed approach estimates it from the mixtures spectra.[23] Data clustering algorithm also yields as a result estimate of the concentration or mixing matrix. Constrained ℓ_1 norm minimization is used afterwards to solve underdetermined system of linear equations in order to estimate pure components mass spectra. To implement constrained ℓ_1 norm minimization we have tested two approaches: linear programming method with equality constraints and interior point method aimed to solve ℓ_1-regularized least square problem.[29] Both approaches yielded results of basically same quality. HALS NMF algorithms with sparseness constraints were employed in multilayer form for which it has been found to improve sparseness of the solutions.[30,31] In addition to that we have used for the initial value of the data

Journal of MASS SPECTROMETRY

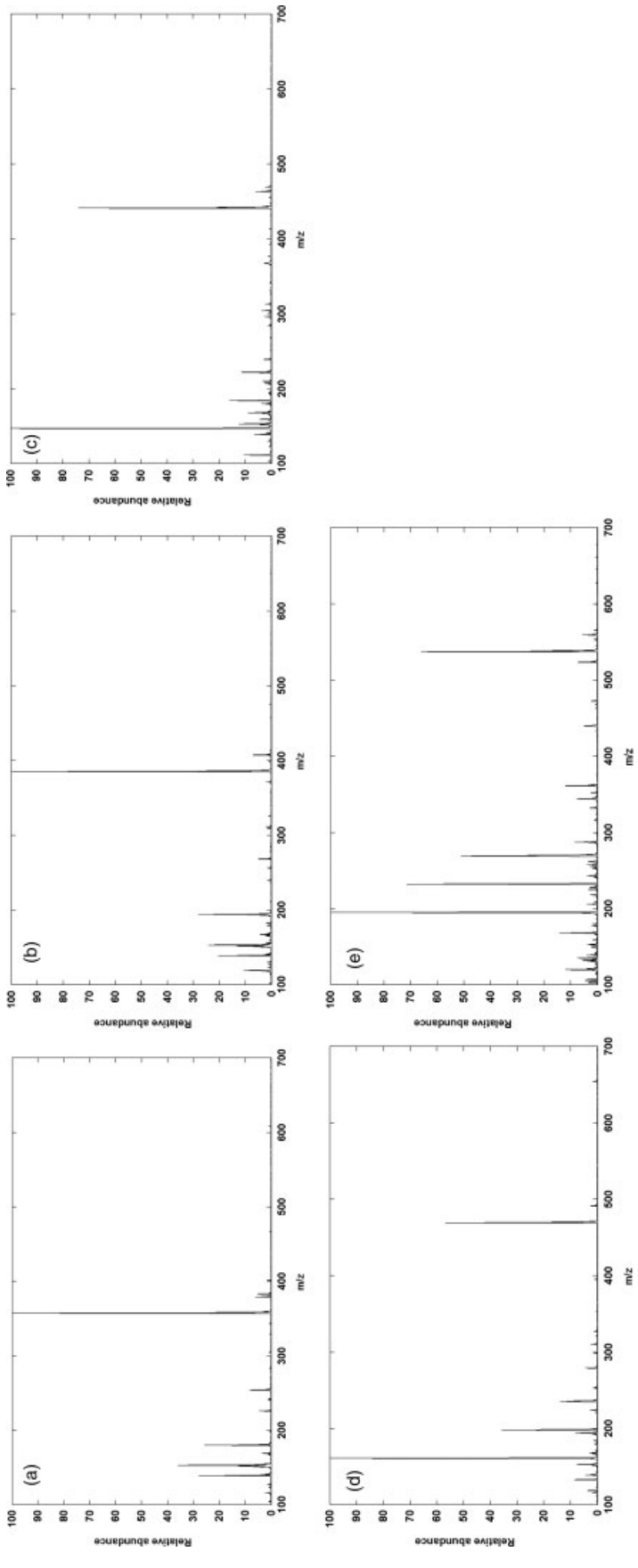


Figure 2. Pure components mass spectra.

concentration matrix in the first layer of the NMF algorithm the estimate of the concentration matrix obtained by data clustering algorithm. This additionally improved separation quality in relation to one obtained by constrained ℓ_1 norm minimization approach.

Sparse component analysis

Like many decomposition methods, the proposed approach is based on static linear mixture model

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

where $\mathbf{X} \in \mathrm{R}_{0+}^{N\times T}$ represents matrix of N measured mixtures spectra across T m/z variables, $\mathbf{A} \in \mathrm{R}_{0+}^{N\times M}$ represents the matrix of concentration profiles, also called the mixing matrix, and matrix $\mathbf{S} \in \mathrm{R}_{0+}^{M\times T}$ contains M pure components mass spectra across T m/z variables. Because of the nature of the problem all quantities in Eqn (1) are nonnegative. As already pointed out, the number of pure components M is in principle unknown although many BSS/ICA algorithms assume that it is either known in advance or can be easily estimated. This does not seem to be true in practice, especially when the BSS problem is underdetermined. Here, we shall treat M as unknown parameter that will be estimated by the clustering algorithm described in *Data clustering* section. In addition to estimating the number of pure components, used data clustering algorithm also estimates the concentration matrix. Estimate of the concentration matrix is necessary for the ℓ_1 norm minimization approach, but it is shown to be useful for the initialization of the HALS NMF approach that will be described later. In overall, the BSS problem related to blind mass spectra decomposition consists of (1) estimating the number of pure components spectra and concentration matrix by data clustering algorithm, (2) estimating the matrix of the pure components spectra **S** by ℓ_1 norm minimization algorithm, (3) as an alternative to (2) estimating the concentration matrix **A** and the matrix of the pure components spectra **S** by HALS NMF algorithm in multilayer mode. All three tasks are executed using matrix of mixtures spectra **X** only. In addition to that, we allow the number of pure components spectra M to be greater than the number of mixtures spectra N. Hence, blind mass spectra decomposition problem becomes uBSS problem.

Data clustering

In mass spectra decomposition problem considered in this paper we have assumed that pure components spectra are in average $k = M - 1$ sparse. This implies that at each m/z coordinate, on average only one pure component is active i.e. nonzero. By looking at the morphology of the mass spectra, e.g. references [9] and [10], this assumption appears to hold in practice. It allows to reduce number of mixtures to $N = 2$, hence reducing the computational complexity of the used data clustering algorithm[23] by reducing dimension of the concentration subspaces, that equals average number of active components, to 1. Then, the appropriately chosen function, e.g. Equation (3), will effectively cluster data, wherein the number of clusters corresponds to the estimate of the number of pure components M. If the number of m/z coordinates that violates $k = M - 1$ sparseness assumption is relatively large, this will influence accuracy of the estimation of the concentration matrix due to the repositioning of the cluster centers. It will not however influence to the same extent the accuracy of the estimation of the number of clusters. Thus, performance of the ℓ_1 norm minimization algorithms that require the estimate of the concentration matrix in order to proceed to the next phase and solve underdetermined system of linear equations will be affected significantly if mass spectra are not sparse enough. On the other hand, HALS NMF approach will be significantly less sensitive to the level of sparseness of the mass spectra because it only requires from the clustering algorithm the estimate of the number of pure components spectra. However, because estimate of the concentration matrix is available as the by-product of the data clustering algorithm it can be very useful in the initialization of the HALS NMF algorithms as well. This is due to the fact that simultaneous minimization of chosen cost function with respect to **A** and **S** is nonconvex problem with many local minima. Hence, quality of the decomposition highly depends on the strategy employed for selection of initial values of **A** and **S**.

Because solution of the BSS problem is generally characterized by scale indeterminacy we have assumed the unit norm constraint (in the sense of ℓ_2 norm) on the columns of the concentration matrix **A**, i.e., $\{||\mathbf{a}_m||_2 = 1\}_{m=1}^{M}$. As already pointed out, in this paper we do assume the number of mixtures to be $N = 2$. Thus, the normalized mixing vectors $\{\mathbf{a}_m\}_{m=1}^{M}$ lie in the first quadrant on the unit circle, i.e. they are parameterized as follows:

$$\mathbf{a}_m = [\cos(\varphi_m)\ \sin(\varphi_m)]^{\mathrm{T}} \quad m = 1, \ldots, M \tag{2}$$

where φ_m represents mixing angle that is confined within the interval $[0, \pi/2]$. By assuming *1*-dimensional concentration subspaces the clustering algorithm in ref. 23 is outlined by the following steps:

1. We removed all data points close to the origin for which applies $\{||\mathbf{x}(t)||_2 \leq \varepsilon\}_{t=1}^{T}$, where ε represents some predefined threshold. This corresponds to the case where pure components spectra are close to zero.
2. Normalize to unit ℓ_2 norm remaining data points $\mathbf{x}(t)$, i.e. $\{\mathbf{x}(t) \leftarrow \mathbf{x}(t)/||\mathbf{x}(t)||_2\}_{t=1}^{\overline{T}}$, where $\overline{T} \leq T$ denotes number of data points that remained after the elimination process in step 1.
3. Calculate function $f(\mathbf{a})$, where $\mathbf{a}$ is defined with Eqn (2) as follows:
$$f(\mathbf{a}) = \sum_{t=1}^{\overline{T}} \exp\left(-\frac{d^2(\mathbf{x}(t), \mathbf{a})}{2\sigma^2}\right) \tag{3}$$
where $d(\mathbf{x}(t), \mathbf{a}) = \sqrt{1 - (\mathbf{x}(t) \cdot \mathbf{a})^2}$ and $(\mathbf{x}(t) \cdot \mathbf{a})$ denotes inner product. Parameter σ in Eqn (3) is called dispersion. If set to sufficiently small value, in our experiments this turned out to be $\sigma \approx 0.065 \pm 0.01$, the value of the function $f(\mathbf{a})$ will approximately equal the number of data points close to $\mathbf{a}$. Thus, by varying mixing angles $0 \leq \varphi \leq \pi/2$ we effectively cluster data. However, it is clear that the reported value is empirical. For another set of mixtures, depending on the concentration profiles of the pure components, it can yield either overestimated or underestimated number of pure components. To obtain robust estimator we propose to decrease value of σ until the estimated number of pure components is increased for 1 or 2. False pure components will be either repeated versions of some of the true pure components or their linear combinations. Thus, they can be detected after blind extraction phase as the ones that are highly correlated with the rest of the extracted pure components.
4. Number of peaks of the function $f(\mathbf{a})$ corresponds to the estimate of the number of pure components spectra $\hat{M}$.

Journal of MASS SPECTROMETRY

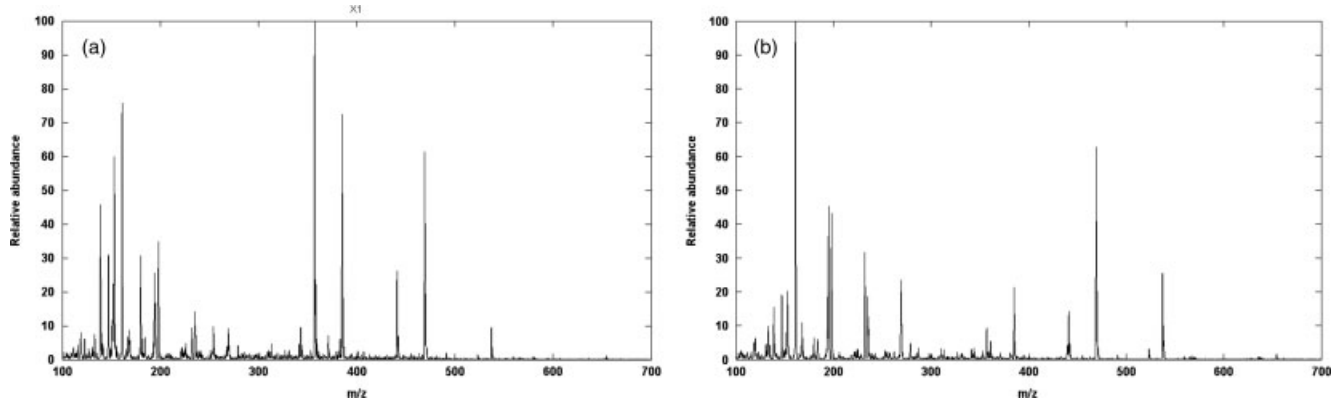


Figure 3. Mass spectra of mixtures X_1 and X_2.

Locations of the peaks correspond to the estimates of the mixing angles $\{(\hat{\varphi}_m)\}_{m=1}^{\hat{M}}$, i.e. mixing or concentration vectors $\{\hat{\mathbf{a}}_m\}_{m=1}^{\hat{M}}$, where $\hat{\mathbf{a}}_m$ is given with Eqn (2). The hat sign introduced here is used to denote estimate of the related quantity. Hence, at the end of data clustering phase estimates of the number of pure components M and concentration matrix $\mathbf{A}$ are obtained.

ℓ_1 Norm minimization

SCA enables to find a possible good approximation of the true solution to an underdetermined system of linear equations subject to sparseness constraints. When in Eqn (1) $N < M$, the null-space of $\mathbf{A}$ is nontrivial, and the inverse problem has many solutions. Therefore, an additional constraint such as sparseness between the components of the column vectors $\{\mathbf{s}(t)\}_{t=1}^{T}$ is necessary. A sparse signal is a signal whose most samples are nearly zero, and just few percent take significant values. Signal that has at least $k \leq M$ zero components is called k-sparse. Therefore, it is possible to obtain solution of the resulting uBSS problem through as ℓ_1 norm minimization,[24–26] once the number of pure components M and concentration matrix $\mathbf{A}$ are estimated using geometric concept known as data clustering.[23] The pure components extraction problem is reduced to solving resulting underdetermined system of linear equations. This last step is carried out as linear programming[24,32,33] or for example as ℓ_1 -regularized least square problem.[29,34] Linear programming solution is obtained as

$$\hat{\mathbf{s}}(t) = \underset{\mathbf{s}(t)}{\arg\min} \sum_{m=1}^{\hat{M}} s_m(t) \quad \text{subject to } \hat{\mathbf{A}}\mathbf{s}(t) = \mathbf{x}(t) \quad \forall t = 1, \ldots, T$$
$$\mathbf{s}(t) \geq \mathbf{0} \qquad (4)$$

where $\hat{\mathbf{A}}$ denotes estimate of the true mixing matrix $\mathbf{A}$ obtained by previously described data clustering algorithm. Pure component spectra are obtained from the solution of linear program (Eqn (4)) as $\hat{\mathbf{s}}(t)$. If the noise is present in blind decomposition problem more robust sparse solution for $\{\mathbf{s}(t)\}_{t=1}^{T}$ is obtained by solving ℓ_1-regularized least square problem:[29,34]

$$\hat{\mathbf{s}}(t) = \underset{\mathbf{s}(t)}{\arg\min} \frac{1}{2} \left\| \hat{\mathbf{A}}\mathbf{s}(t) - \mathbf{x}(t) \right\|_2^2 + \lambda ||\mathbf{s}(t)||_1 \quad \forall\, t = 1, \ldots, T \quad (5)$$

We have tested both linear programming method (Eqn (4)) and interior point method[29,35] used to implement Eqn (5). The two algorithms yielded results with basically similar quality, which implies that noise level in the used experimental data was low.

HALS NMF

The second approach to SCA employs NMF algorithms, wherein mixing matrix $\mathbf{A}$ and source matrix $\mathbf{S}$ are estimated simultaneously, usually through ALS methodology.[27,28,36] Majority of algorithms used for adaptive NMF are based on the alternating minimization of the squared Euclidean distance expressed by the Frobenius norm with respect to two sets of parameters $\{a_{nm}\}$ and $\{s_{mt}\}$:[27,28,36]

$$D_F(\mathbf{X}||\mathbf{AS}) = \frac{1}{2}||\mathbf{X} - \mathbf{AS}||_2^2 + \alpha_{\mathbf{S}} J_{\mathbf{S}}(\mathbf{S}) + \alpha_{\mathbf{A}} J_{\mathbf{A}}(\mathbf{A}) \qquad (6)$$

where $J_{\mathbf{S}}(\mathbf{S})$ and $J_{\mathbf{A}}(\mathbf{A})$ represent sparseness constraints imposed on $\mathbf{S}$ and $\mathbf{A}$, while $\alpha_{\mathbf{S}}$ and $\alpha_{\mathbf{A}}$ represent corresponding regularization constants. Decomposition implied by Eqn (1) that is based on minimization of the squared Euclidean distance only has many solutions. Thus, constraints are necessary to obtain solutions for $\mathbf{A}$ and $\mathbf{S}$ that are meaningful. For this purpose sparseness constraints are imposed on $\mathbf{A}$ and $\mathbf{S}$ in a majority of cases. Consequently, we shall impose sparseness constraints on rows of $\mathbf{S}$: $\{\underline{\mathbf{s}}_m\}_{m=1}^{M}$. In addition to that, to solve uBSS problem we shall employ minimization of the of the local cost functions:[27,28]

$$D_F^{(m)}\left(\mathbf{X}^{(m)} \,\|\, \mathbf{a}_m \underline{\mathbf{s}}_m\right) = \frac{1}{2}\left\|\mathbf{X}^{(m)} - \mathbf{a}_m \underline{\mathbf{s}}_m\right\|_2^2 + \alpha_{\mathbf{s}}^{(m)} \left\|\underline{\mathbf{s}}_m\right\|_1 \quad m = 1, \ldots, M \qquad (7)$$

with respect to $\{s_{mt}\}$ where

$$\mathbf{X}^{(m)} = \mathbf{X} - \sum_{j \neq m} \mathbf{a}_j \, \underline{\mathbf{s}}_j \qquad (8)$$

Constant $\alpha_{\mathbf{s}}^{(m)}$ regulates level of sparseness of the pure component mass spectra $\{\underline{\mathbf{s}}_m\}_{m=1}^{M}$. Assuming that columns of $\mathbf{A}$ are normalized to ℓ_2 unit norm, minimization of Eqn (7) with respect to $\{\underline{\mathbf{s}}_m\}_{m=1}^{M}$ yields the following learning rules:

$$\underline{\mathbf{s}}_m \longleftarrow \frac{1}{1 + \alpha_{\mathbf{s}}^{(m)}} \left[\mathbf{a}_m^{\mathrm{T}} \mathbf{X}^{(m)} - \alpha_{\mathbf{s}}^{(m)} \mathbf{1}_{1 \times T}\right]_+ \qquad (9)$$

As opposed to pure components spectra the concentration matrix is learned globally through minimization of Eqn (6) without any

constraints imposed on it. This yields the following learning rule for **A:**

$$\mathbf{A} \longleftarrow \left[\mathbf{X}\mathbf{S}^{\mathrm{T}}\left(\mathbf{S}\mathbf{S}^{\mathrm{T}} + \lambda \mathbf{I}_M\right)\right]_+ \tag{10}$$

wherein after each iteration, **A** is normalized to ℓ_2 unit column norm. In Eqn (10) $\mathbf{I}_M$ is an $M \times M$ identity matrix and $\mathbf{1}_{1\times T}$ is row vector with all entries equal to one. In Eqns (9) and (10), $[\xi]_+ = \max\{\varepsilon, \xi\}$ (e.g. $\varepsilon = 10^{-16}$) is used to prevent negative solutions for **A** and **S**. Regularization constant λ in Eqn (10) is used to improve ill-conditioning of the matrix $\mathbf{S}\mathbf{S}^{\mathrm{T}}$ and changes as a function of the iteration index k as $\lambda_k = \lambda_0 \exp(-k/\tau)$ (with $\lambda_0 = 100$ and $\tau = 0.02$ in the experiments). Additional improvement in performance of the NMF algorithms can be obtained when they are applied in the multilayer mode,[30,31] whereas sequential decomposition of the nonnegative matrices is performed as follows. In the first layer, the basic approximation decomposition is performed $\mathbf{X} \cong \mathbf{A}^{(1)}\mathbf{S}^{(1)} \in \mathrm{R}_{0+}^{N\times T}$. In the second layer result from the first layer is used to build up new input data matrix for the second layer $\mathbf{X} \leftarrow \mathbf{S}^{(1)} \in \mathrm{R}_{0+}^{M\times T}$ yielding $\mathbf{X}^{(1)} \cong \mathbf{A}^{(2)}\mathbf{S}^{(2)} \in \mathrm{R}_{0+}^{M\times T}$. After L layers the data decomposes as follows:

$$\mathbf{X} \cong \mathbf{A}^{(1)}\mathbf{A}^{(2)} \cdots \mathbf{A}^{(L)}\mathbf{S}^{(L)} \tag{11}$$

Discussion

Model validation

First, we test validity of the linear mixture model (Eqn (1)) that forms the basis for the proposed blind spectra decomposition approach. For this purpose we have tested mixtures of pure components **1** and **2** (mass spectra shown in Fig. 4(a) and (b)). Mass spectrum shown in Fig. 4(c) was created computationally by linearly combining pure components **1** and **2** in 50 : 50 ratio. Mass spectrum shown in Fig. 4(d) has been obtained experimentally, by injection of 2 µl of 50 : 50 mixture of pure components **1** and **2.** As can be clearly seen, main peaks located in the mass spectrum of each pure component are located at the same positions in both computed and recorded mixture. This implies high degree of equivalence between experimental recorded mixture and its linear mathematical model. This statement is further supported by the correlation coefficient between computed and recorded mixtures, which is 0.972. Thus, it can be concluded that linear mixture model (Eqn (1)) is valid and can be used in blind spectra decomposition methods.

Concept demonstration

Mathematically demanding problem that can be frequently encountered in everyday research is selected to test real capability of SCA approach to blind decomposition of mass spectra. Based on present knowledge and applications of blind approaches to pure components spectra extraction, our judgment was that use of only *two mixtures* consisting of as many as *five components* was a challenging task. Choosing two three-component mixtures would be far less demanding and also far away from real complexity of multi-component data analysis in mass spectrometry as well as from true capability of the SCA algorithms. Therefore, it was important to show that useful information can be obtained from a single experiment and this aim was successfully accomplished through a described approach. Number of pure components and concentration matrix are estimated from two mixtures, shown in Fig. 3, with the described clustering algorithm. When dispersion factor in Eqn (3) is set to $\sigma = 0.06$ the number of the pure components is estimated as $M = 5$. The estimated concentrations were X_1 (**1 : 2 : 3 : 4 : 5** = 86.06 : 66.54 : 50.61 : 34.91 : 18.14) and X_2 (**1 : 2 : 3 : 4 : 5** = 13.94 : 33.46 : 49.39 : 65.09 : 81.86). Based on the given concentrations of the pure components in the mixture, as described in the section *compounds*, it is easy to recalculate concentrations in terms of percentage: X_1 (**1 : 2 : 3 : 4 : 5** = 85.7 : 66.7 : 50 : 33.6 : 14.3) and X_2 (**1 : 2 : 3 : 4 : 5** = 14.3 : 33.3 : 50 : 66.7 : 85.7). We consider the agreement between true and estimated concentrations satisfactory. The corresponding clustering function given byEqn (3) is shown in Fig. 5. Figure 6 shows results obtained by data clustering algorithm (Eqn (3)) and linear programming algorithm (Eqn (4)). The degree of similarity between true pure components, shown in Fig. 2, and extracted pure components, shown in Fig. 6, expressed as normalized correlation coefficients is: 0.8836 (component **1**), 0.7673 (component **2**), 0.7160 (component **3**), 0.9726 (component **4**) and 0.9601(component **5**). These numbers can be easily illustrated by comparing Figs 2 and 6. The highest correlations found for components **4** and **5** are entirely in agreement with mass spectra (Fig. 6(d) and (e)); all fragment peaks found for pure components are present in extracted spectra with even the same relative intensities. Next, extracted mass spectrum of component **1** (Fig. 6(a)) missed some of the fragment ions (*m/z* 139 and 153), while extracted mass spectrum of component **2** (Fig. 6(b)) contains traces of component **1**, found in higher intensities of peaks at the *m/z* 139 and 153. These finding are in agreement with somewhat lower correlation coefficients. Finally, it is evident that the mass spectrum of component **3** (Fig. 6(c)) is extracted with the lowest accuracy; besides characteristic ions at the *m/z* 441 and 146, it contains also ions corresponding to the component **4** at the *m/z* 467 and 198.

Although results obtained by the linear programming algorithm were very good and components were extracted with high accuracy, we considered HALS NMF algorithm to see whether it can further improve quality of extracted pure components. This has been especially motivated by relatively poor extraction of component **3**. Learning rules (Eqns (9) and (10)) can be combined with multilayer mode of operation (Eqn (11)), constituting multilayer HALS NMF algorithm, to increase performance of the NMF algorithms. Furthermore, performance of the NMF algorithm critically depends on the strategy employed to select initial values for **A** and **S**. The reason is that cost functions (Eqns (6) and (7)) are convex with respect to **A** or **S** but not with respect to both of them. This increases chance, especially in a case of large scale problems, that NMF algorithm will be stuck in local minima yielding poor performance. Therefore, we propose the following initialization strategy to reduce these problems. For the first layer it includes number of random guess for **S** and initial value for **A** obtained as the result of data clustering algorithm. For the second and higher layers it includes random guesses for both **S** and **A**. As it is demonstrated in Fig. 7 this brings additional improvement of the quality of the pure components extraction in relation to the one obtained by linear programming algorithm. Figure 7 shows results obtained by HALS NMF algorithm, Eqns (6)–(11), with 100 layers with regularization constants $\alpha_{\mathbf{s}_p}^{(m)} = 0.5$ after 500 iterations per layer. Now, in direct comparison between Figs 2(c) and 7(c) it can be seen that pure component **3** has been extracted more accurately than by linear programming algorithm, (Fig. 6(c)). The molecular $[M+H]^+$ ion can now be assigned undoubtedly, although fragment ion at the *m/z* 198 still remains in the extracted mass spectrum. The results of other pure components are of the similar quality as those obtained

Journal of MASS SPECTROMETRY

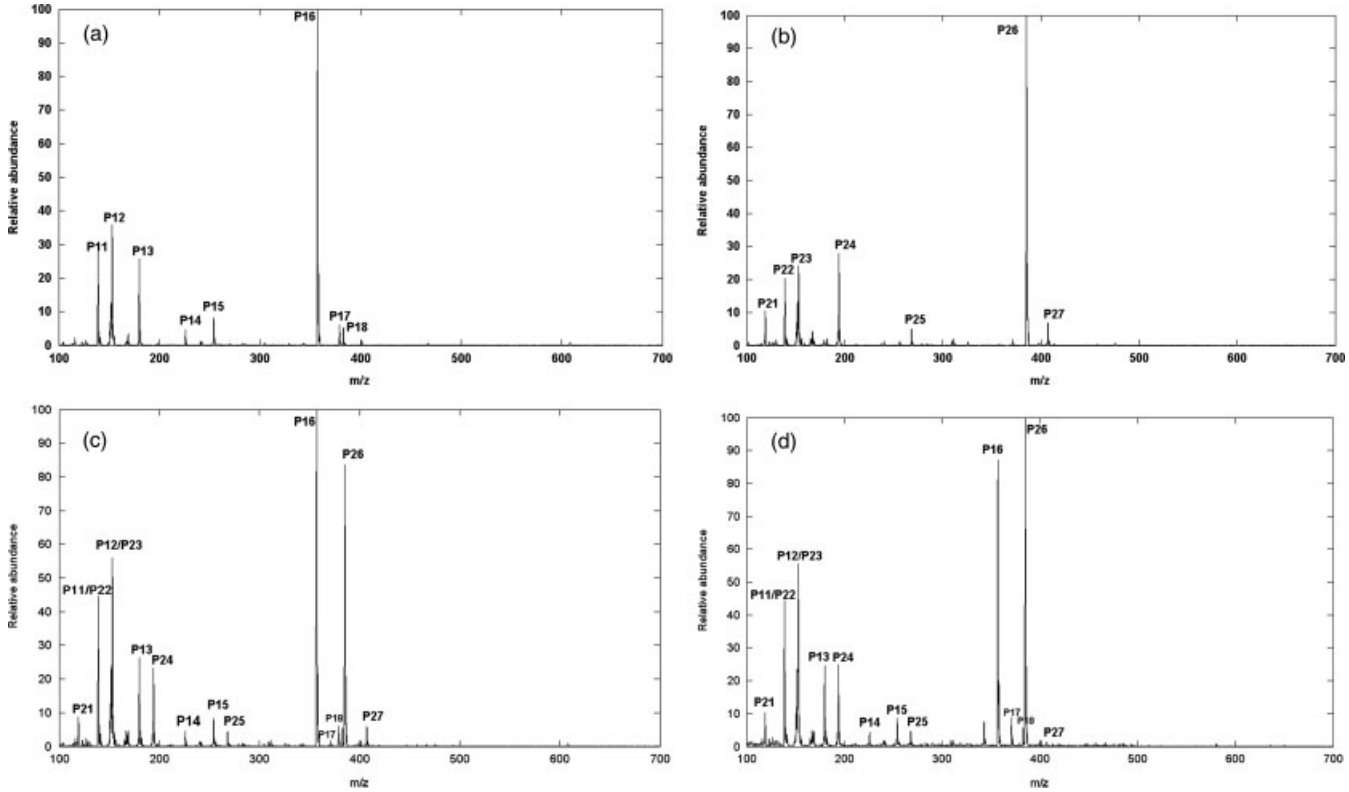


Figure 4. (a) and (b): mass spectra of pure components **1** and **2** with main peaks located; (c) mass spectrum of mixture computed as linear combination of pure components **1** and **2** in 50 : 50 ratio; (d) mass spectrum obtained experimentally from 50 : 50 mixture containing pure components **1** and **2**.

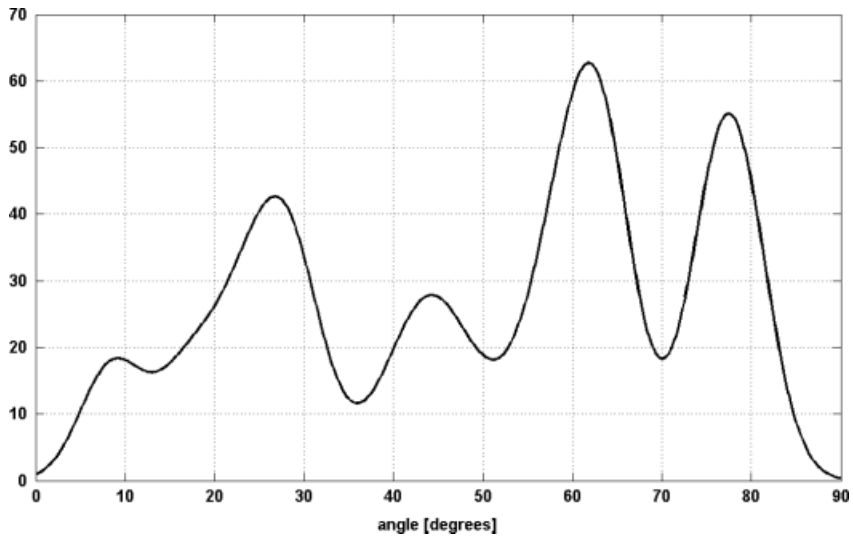


Figure 5. Clustering function, Eqn (3), for two mixtures shown in Fig. 3. Dispersion factor was set to $\sigma = 0.06$. Five peaks indicate existence of five pure components spectra in two mixtures.

by linear programming algorithm that is shown in Fig. 6. This is confirmed by the values of normalized correlation coefficients between true and estimated pure components which are: 0.9084 (component **1**), 0.7432 (component **2**), 0.7389 (component **3**), 0.9732 (component **4**) and 0.9698 (component **5**).

It is however fair to say that, due to the random guess of **S**, the HALS NMF algorithm does not converge toward the solution presented in Fig. 7 each time. Our computational studies indicate that this happens 3 to 4 times out of 10 attempts.

From an analytical point of view, these results are very important. Number of components present in the mixture can be determined without previous separation, which is a great advantage when dealing with multi-component systems. With this knowledge extracted mass spectra can be treated even if they are not extracted with absolute accuracy. By inspection of mass spectra shown in Fig. 6 or 7, it is clear that the molecular $[M+H]^+$ ions of all components can be designated, which is a necessary starting point for further structural analysis. It is also important to emphasize

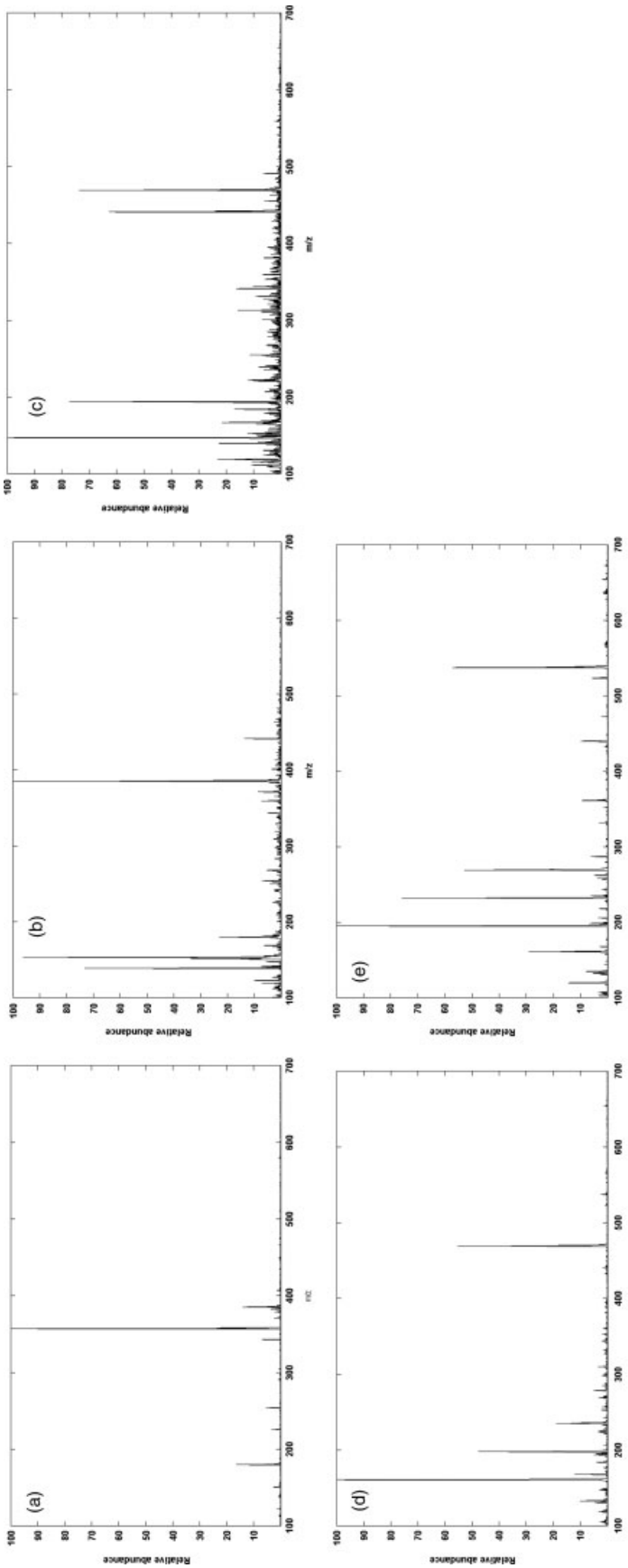


Figure 6. Mass spectra of five pure components extracted by data clustering and linear programming algorithm (Eqn(4)).

Journal of MASS SPECTROMETRY

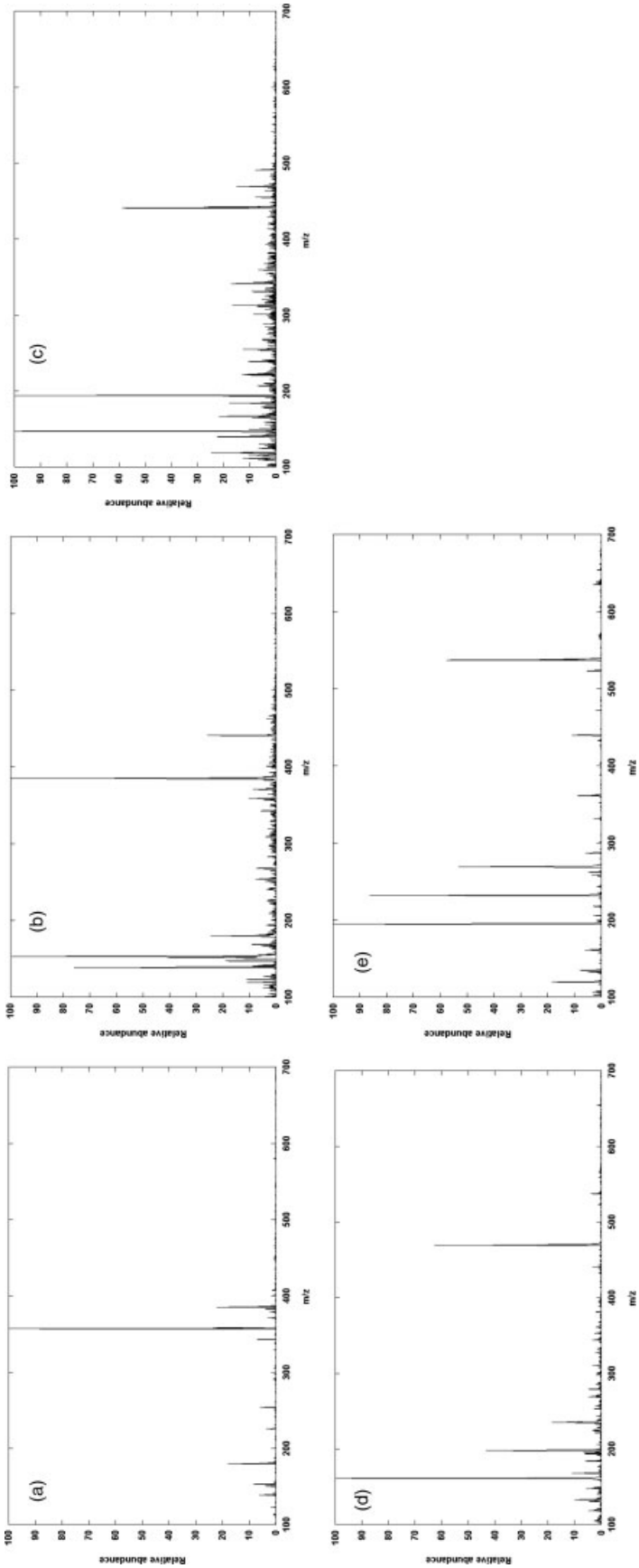


Figure 7. Mass spectra of five pure components extracted by data clustering and multilayer HALS NMF algorithm (Eqns (6)–(11)).

that accuracy of pure components extraction can be increased by increasing the number of mixtures from two to three or more. This is due to the fact that level of sparseness required between pure components to obtain unique solution of blind decomposition is $k = M-N+1$. Hence, if pure components violate $k = M - 1$, sparseness requirement number of mixtures N ought to be increased. This could be important for mixtures with even greater number of components, mixtures of similar components or those present in low concentrations (or traces). However, in this paper our goal was to demonstrate that challenging blind pure components extraction problem can be solved even when only the minimal requirements are met, i.e. having at disposal as few as $N = 2$ mixtures only.

To sum it up, the presented SCA approach brought information that are of great importance and can be used to guide further experiments and research of complex systems by mass spectrometry.

Conclusions

SCA-based approach using ℓ_1 norm minimization and multilayer HALS NMF algorithm was successfully applied for blind extraction of *more than two* pure components spectra in mass spectrometry measuring *two mixtures only*. This appears to be the first time to report such result in the chemical literature, because other blind decomposition methods require the number of mixtures to be equal to or greater than the unknown number of pure components. Unlike many existing BSS methods that assume the number of pure components to be known in advance, proposed SCA-based method estimates it by data clustering algorithm. Proposed SCA-based approach can be used as a part of software packages for the analysis of mass spectra and identification of the chemical compounds. It could be of great importance for the analysis of multi-component samples obtained from either reaction mixtures or biological sources.

Acknowledgements

The work of I. Kopriva and I. Jerić was respectively supported by the Ministry of Science, Education and Sports, Republic of Croatia under Grants 098-0982903-2558 and 098-0982933-2936. We thank Mr. Vatroslav Letfus for performing mass spectrometry measurements.

References

[1] J. C. Wright, J. S. Hubbard. Recent developments in proteome informatics for mass spectrometry analysis. *Combinatorial Chemistry and High Throughput Screening* **2009**, *12*, 194.
[2] R. A. McDonald, P. Skipp, J. Bennell, C. Potts, L. Thomas, C. D. O'Connor. Mining whole-sample mass spectrometry proteomics for biomarkers: an overview. *Expert Systems with Applications* **2009**, *36*, 5333.
[3] K. Dettmer, P. A. Aronov, B. D. Hammock. Mass spectrometry-based metabolomics. *Mass Spectrometry Reviews* **2007**, *26*, 51.
[4] K. T. Myint, K. Aoshima, S. Tanaka, T. Nakamura, Y. Oda. Quantitative profiling of polar cationic metabolites in human cerebrospinal fluid by reversed-phase nanaliquid chromatography/mass spectrometry. *Analytical Chemistry* **2009**, *81*, 1121.
[5] B. Wen, W. L. Fitch. Screening and characterization of reactive metabolites using glutathione ethyl ester in combination with Q-trap mass spectrometry. *Journal of Mass Spectrometry* **2009**, *44*, 90.
[6] M. Vermeulen, N. C. Hubner, M. Mann. High confidence determination of specific protein-protein interactions using quantitative mass spectrometry. *Current Opinion in Biotechnology* **2008**, *19*, 331.
[7] S. B. Bumpus, N. L. Kelleher. Accessing natural product biosynthetic processes by mass spectrometry. *Current Opinion in Chemical Biology* **2008**, *12*, 475.
[8] D. Nuzillard, S. Bourg, J. M. Nuzillard. Model-free analysis of mixtures by NMR using blind source separation. *Journal of Magnetic Resonance* **1998**, *133*, 358.
[9] E. Visser, T.-W. Lee. An information-theoretic methodology for the resolution of pure component spectra without prior information using spectroscopic measurements. *Chemometrics and Intelligent Laboratory Systems* **2004**, *70*, 147.
[10] G. Wang, Q. Ding, Y. Sun, L. He, X. Sun. Estimation of source infrared spectra profiles of acetylspiramycin active components from troches using kernel independent component analysis. *Spectrochimica Acta, Part A* **2008**, *70*, 571.
[11] G. Wang, Q. Ding, Z. Hou. Independent component analysis and its applications in signal processing for analytical chemistry. *Trends in Analytical Chemistry* **2008**, *27*, 368.
[12] J. Chen, X. Z. Wang. A new approach to near-infrared spectral data analysis using independent component analysis. *Journal of Chemical Information and Computer Sciences* **2001**, *41*, 992.
[13] X. Shao, W. Wang, Z. Hou, W. Cai. A new regression method based on independent component analysis. *Talanta* **2006**, *69*, 676.
[14] J. Y. Ren, C. Q. Chang, P. C. W. Fung, J. G. Shen, F. H. Y. Chan. Free radical EPR spectroscopy analysis using blind source separation. *Journal of Magnetic Resonance* **2004**, *166*, 82.
[15] C. Chang, J. Ren, P. C. Fung, Y. S. Hung, J. G. Shen, F. H. Y. Chan. Novel sparse component analysis approach to free radical EPR spectra decomposition. *Journal of Magnetic Resonance* **2005**, *175*, 242.
[16] X. Shao, G. Wang, S. Wang, Q. Su. Extraction of mass spectra and chromatographic profiles from overlapping GC/MS signal with background. *Analytical Chemistry* **2004**, *76*, 5143.
[17] G. Wang, W. Cai, X. Shao. A primary study on resolution of overlapping GC-MS signal using mean-field approach independent component analysis. *Chemometrics and Intelligent Laboratory Systems* **2006**, *82*, 137.
[18] V. A. Shashilov, M. Xu, V. V. Ermolenkov, I. K. Lednev. Latent variable analysis of Raman spectra for structural characterization of proteins. *Journal of Quantitative Spectroscopy and Radiative Transfer* **2006**, *102*, 46.
[19] H. Li, T. Adali, W. Wang, D. Emge, A. Cichocki. Non-negative matrix factorization with orthogonality constraints and its application to Raman spectroscopy. *Journal of VLSI Signal Processing* **2007**, *48*, 83.
[20] A. Hyvärinen, J. Karhunen, E. Oja. *Independent Component Analysis*. Wiley Interscience: New York, USA, **2001**.
[21] A. Cichocki, S. I. Amari. *Adaptive Blind Signal and Image Processing*. John Wiley: New York, **2002**.
[22] P. Comon. Independent component analysis - a new concept. *Journal of VLSI Signal Processing* **1994**, *36*, 287.
[23] F. M. Naini, G. H. Mohimani, M. Babaie-Zadeh, Ch Jutten. Estimating the mixing matrix in sparse component analysis (SCA) based on partial k-dimensional subspace clustering. *Neurocomputing* **2008**, *71*, 2330.
[24] Y. Li, A. Cichocki, S. Amari. Analysis of sparse representation and blind source separation. *Neural Computation* **2004**, *16*, 1193.
[25] P. Georgiev, F. Theis, A. Cichocki. Sparse component analysis and blind source separation of underdetermined mixtures. *IEEE Transactions on Neural Networks* **2005**, *16*, 992.
[26] P. Bofill, M. Zibulevsky. Underdetermined Blind Source Separation Using Sparse Representations. *Journal of VLSI Signal Processing* **2001**, *81*, 2353.
[27] A. Cichocki, R. Zdunek, S. I. Amari. Hierarchical ALS algorithms for nonnegative matrix factorization and 3D tensor factorization. *Lecture Notes in Computer Science* **2007**, *4666*, 169.
[28] A. Cichocki, A. H. Phan, R. Zdunek, L.-Q. Zhang. Flexible component analysis for sparse, smooth, nonnagative coding or representation. *Lecture Notes in Computer Science* **2008**, *4984*, 811.
[29] S. J. Kim, K. Koh, M. Lustig, S. Boyd, S. Gorinevsky. An interior-point method for large-scale ℓ_1-regularized least squares. *IEEE Journal of Selected Topics in Signal Processing* **2007**, *1*, 606.

[30] A. Cichocki, R. Zdunek. Multilayer nonnegative matrix factorization. *Electronics Letters* **2006**, *42*, 947.
[31] K. Stadlthanner, F. J. Theis, C. G. Puntonet, J. M. Górriz, A. M. Tomé, E. W. Lang. Hybridizing sparse component analysis with genetic algorithms for blind source separation. *Lecture Notes in Computer Science* **2005**, *3745*, 137.
[32] I. Takigawa, N. Kudo, J. Toyama. Performance analysis of minimum l_1-norm solutions for underdetermined source separation. *IEEE Transactions on Signal Processing* **2004**, *52*, 582.
[33] D. L. Donoho, M. Elad. Optimally sparse representation in general (non-orthogonal) dictionaries via l_1 minimization. *Proceedings of the National Academy of Sciences of United States of America* **2003**, *100*, 2197.
[34] J. A. Tropp, A. C. Gilbert. Signal Recovery from Random Measurements via Orthogonal Matching Pursuit. *IEEE Transactions on Information Theory* **2007**, *53*, 4655.
[35] http://www.stanford.edu/∼boyd/l1_ls/ [accessed: 2008].
[36] A. Cichocki, R. Zdunek, S. Amari. Nonnegative matrix and tensor factorization. *IEEE Signal Processing Magazine* **2008**, *25*, 142.

