Empirical Kernel Map Approach to Nonlinear Underdetermined Blind Separation of Sparse Nonnegative Dependent Sources: Pure Components Extraction from Nonlinear Mixtures Mass Spectra

Journal:	*Journal of Chemometrics*
Manuscript ID:	CEM-14-0009.R1
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Kopriva, Ivica; Ruđer Bo koviæ Institute, Laser and Atomic Research and Development Jerić, Ivanka; Ruđer Bošković Institute, Organic Chemistry and Biochemistry Filipović, Marko; Ruđer Bošković Institute, Laser and Atomic Research and Development Brkljačić, Lidija; Ruđer Bošković Institute, Organic Chemistry and Biochemistry
Keyword:	Nonlinear underdetermined blind source separation, Robust principal component analysis, Thresholding, Empirical kernel maps, Nonnegative matrix factorization

Empirical Kernel Map Approach to Nonlinear Underdetermined Blind Separation of Sparse Nonnegative Dependent Sources: Pure Components Extraction from Nonlinear Mixtures Mass Spectra

Ivica Kopriva[1*], Ivanka Jerić[2], Marko Filipović[1] and Lidija Brkljačić[2]

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

[1]Division of Laser and Atomic Research and Development

phone: +385-1-4571-286, fax:+385-1-4680-104

e-mail: ikopriva@irb.hr, Marko.Filipovic@irb.hr

[2]Division of Organic Chemistry and Biochemistry

e-mail: ijeric@irb.hr, Lidija.Brkljacic@irb.hr

Abstract

Nonlinear underdetermined blind separation of nonnegative dependent sources consists in decomposing set of observed nonlinearly mixed signals into greater number of original nonnegative and dependent component (source) signals. That hard problem is practically relevant for contemporary metabolic profiling of biological samples, where

sources (a.k.a. pure components or analytes) are aimed to be extracted from mass spectra of nonlinear multicomponent mixtures. This paper presents method for nonlinear underdetermined blind separation of nonnegative dependent sources that comply with sparse probabilistic model, i.e. sources are constrained to be sparse in support and amplitude. That model is validated on experimental pure components mass spectra. Under sparse prior nonlinear problem is converted into equivalent linear one comprised of original sources and their higher-, mostly second, order monomials. Influence of these monomials, that stand for error terms, is reduced by preprocessing matrix of mixtures by means of robust principal component analysis, hard-, soft- and trimmed thresholding. Preprocessed data matrices are mapped in high-dimensional reproducible kernel Hilbert space (RKHS) of functions by means of empirical kernel map. Sparseness constrained nonnegative matrix factorizations (NMF) in RKHS yield sets of separated components. They are assigned to pure components from the library using maximal correlation criterion. The methodology is exemplified on demanding numerical and experimental examples related respectively to extraction of 8 dependent components from 3 nonlinear mixtures and to extraction of 25 dependent analytes from 9 nonlinear mixtures mass spectra recorded in nonlinear chemical reaction of peptide synthesis.

Key words: Nonlinear underdetermined blind source separation, Robust principal component analysis, Thresholding, Empirical kernel maps, Nonnegative matrix factorization.

1. INTRODUCTION

Identification of pure components present in the mixture is a traditional problem in spectroscopy (nuclear magnetic resonance-, infrared, Raman) and mass spectrometry [1-4]. Identification proceeds often by matching separated components spectra with a library of reference compounds [5-7], whereas degree of correlation depends on how well pure components are separated from each other. Thereby, of interest are blind source separation (BSS) methods that use only the matrix with recorded mixtures spectra as input information [8-11]. In majority of scenarios, separation of pure components is performed by assuming that mixture spectra are linear combinations of pure components [1-4]. While linear mixture model is adequate for many scenarios, nonlinear model offers more accurate description of processes and interactions occurring in biological systems. Living organisms are best examples of complex nonlinear systems that function far from equilibrium. Internal and external stimuli (disease, drug treatment, environmental changes) cause perturbations in the system as a result of highly synchronized molecular interactions [12]. As opposed to many BSS methods developed for linear problems, the number of methods that address nonlinear BSS problem is considerably smaller, see for example chapter 14 in [11]. That number is reduced further when related nonlinear BSS problem is underdetermined, that is when number of pure components is greater than number of mixtures. That is why metabolic profiling, that aims to identify and quantify small-molecule analytes (a.k.a. pure components or sources) present in biological samples (typically urine, serum or biological tissue extract) is seen as one of the most challenging tasks in systems biology [13]. Therefore, underdetermined problem is of practical relevance.

The aim of the paper is to present method for blind separation of pure components from smaller number of multicomponent nonlinear mixtures mass spectra. Therefore, it is assumed that components are nonnegative and sparse. To this end, we address underdetermined nonlinear nonnegative BSS (uNNBSS) problem with sparse and dependent sources. As it has been discussed at great length in [4], even linear underdetermined BSS problem comprised of dependent sources is challenging with only few algorithms addressing it. There is basically no method proposed for uNNBSS problem. Herein, we propose method for uNNBSS problem that can be considered as generalization of the method developed in [4] for underdetermined linear nonnegative BSS (uLNBSS) problem comprised of dependent sources. Proposed method constrains sources to be nonnegative and comply with sparse probabilistic model [14, 15], that is sources are assumed to be sparse in support and amplitude. The model is validated on experimental mass spectrometry data and is therefore practically relevant, see section 3.2. This represents first original contribution of the paper. Under this sparse prior, nonlinear problem is approximated by a linear one comprised of original sources and their second order monomials. This follows from analytical derivations based on Taylor expansion of nonlinear mixture model (that is the vector function with vector argument) up to an arbitrary order. Analytical derivation of Taylor expansion based on Tucker model of tensor derivatives represents, arguably, second original contribution of the paper. The key contribution of the paper is reduction of influence of higher order monomials that stand for error terms. That is achieved by preprocessing matrix of mixtures by means of robust principal component analysis (RPCA) [16, 17], hard- (HT), soft- (ST) [18] and trimmed thresholding (TT) [19]. Preprocessed data matrices are mapped observation-wise in high-dimensional RKHS by means of empirical kernel map

(EKM). Thus, one uNNBSS problem is converted into four nonnegative BSS problems in RKHS with the same number of observations but increased number of mixtures. Sparseness constrained NMF is performed in RKHS to solve these nonnegative BSS problems. Thereby, components separated by NMF are assigned to pure components from the library using maximal correlation criterion.

The rest of the paper is organized as follows. Section 2 gives overview of nonlinear BSS methods and presents theory upon which proposed uNNBSS is built. Section 3 describes experiments performed on computational and experimental data. Section 4 presents and discusses results of comparative performance analysis between proposed uNNBSS and some state-of-the-art NMF algorithms. Concluding remarks are given in Section 5.

2. THEORY AND ALGORITHM

Aimed application of proposed uNNBSS method is extraction of analytes from multicomponent nonlinear mixtures of mass spectra. As emphasized in [4] mass spectrometry is chosen due to its increasing importance in clinical chemistry, safety and quality control as well as biomarker discovery and validation. As in [4, 5], we assume that library of reference mass spectra is available to evaluate quality of components extracted by the proposed method.[1] For an example the NIST and Wiley-Interscience universal spectral library [7], contains more than 800 000 mass spectra (corresponding

[1] Please note that any BSS algorithm when applied to experimental data requires some kind of expert knowledge to evaluate the separation results. Herein the library of pure components is such an "expert". The same concept is used in hyperspectral image analysis where identification of minerals proceeds by comparison of estimated endmembers with spectral profiles stored in the library, see for an example the ASTER spectral library at [20].

to more than 680 000 compounds). As opposed to [4], where linear mixture model is assumed, nonlinear model is assumed herein. Thereby, linear model is implicitly included as a special case.

From the viewpoint of uNNBSS problem with dependent sources existing algorithms for nonlinear BSS problem have at least one of the several deficiencies: (*i*) they assume that number of mixtures is equal to or greater than the unknown number of sources [21-29]; (*ii*) they do not take into account nonnegativity constraint that is present when sources are pure components mass spectra [21-32]; (*iii*) they assume that source signals are statistically independent [22-24, 27, 28-32] and, sometimes, individually correlated [28, 30, 31]. None of these assumptions holds true for the uNNBSS problem considered herein. Algorithm described in [33] is developed for uNNBSS problem composed of nonnegative sources. However, the assumption made by the algorithm is that set of observation indexes exist such that each source is present alone in at least one of these observations. That assumption seems too strong for the considered uNNBSS problem where mass spectra of structurally similar pure components are expected to overlap. That is especially the case if the resolution of the mass spectrometer is low. Algorithms [34-36] execute nonlinear nonnegative BSS by means of nonnegative matrix factorization (NMF) in reproducible kernel Hilbert space (RKHS). Nevertheless, unlike the uNNBSS method proposed herein, they do not: (*i*) enforce sparseness constraint that is shown herein to be enabling condition for solving otherwise intractable uNNBSS problem; (*ii*) reduce influence of higher order monomials of the original sources (error terms) induced by nonlinear mixing process and that is shown herein to be crucial for obtaining reasonably accurate solution of the uNNBSS problem. As it is seen in section 2.2, uNNBSS problem is converted into

equivalent uLNBSS problem with large number of sources: the original ones and their higher order monomials induced by nonlinear mixing process. Without activation of sparse probabilistic prior equivalent uLNBSS problem is intractable.

As it is seen in sections 3.1 and 3.2, proposed methodology significantly improves accuracy relative to the case when the NMF algorithm is performed on EKM-mapped matrix of mixtures data without suppression of higher order monomials. It has already been discussed in [37, 4] that performance of many NMF algorithms depends on optimal usage of parameters required to be known *a priori*, such as balance parameter that regulates influence of sparseness constraint [38], or number of overlapping components that exist in mixtures [39]. Often, these parameters are difficult to select optimally in practice. That is why the nonnegative matrix underapproximation (NMU) algorithm [40] is proposed to solve nonnegative BSS problems in RKHS. That is, it does not require *a priori* information from the user. Thus, we propose herein to combine RPCA, HT, ST and TT preprocessing transforms, EKM based nonlinear mapping with the NMU algorithm in mapping induced high-dimensional RKHS. Hence, the PTs-EKM-NMU algorithm. The PTs-EKM-NMU is exemplified on numerical and experimental problems. Nevertheless, proposed preprocessing transforms can also be used in combination with other sparseness constrained NMF algorithms. Provided that number of overlapping components can be inferred reasonably accurate, an NMF algorithm with ℓ_0-constraints (NMF_L0) [39] is a good choice.

2.1 Underdetermined nonlinear nonnegative blind source separation with dependent sources

The uNNBSS problem with dependent sources is described as:

$$\mathbf{x}_t = \mathbf{f}(\mathbf{s}_t) \qquad t = 1,...,T \tag{1}$$

where $\mathbf{x}_t \in R_{0+}^{N\times 1}$ stands for nonnegative measurement vector comprised of intensities acquired at some of T mass-to-charge (m/z) channels, $\mathbf{s}_t \in R_{0+}^{M\times 1}$ stands for unknown vector comprised of intensities of M nonnegative sources. $\mathbf{f}: R_{0+}^{M} \to R_{0+}^{N}$ is an unknown multivariate mapping such that $\mathbf{f}(\mathbf{s}_t) = \left[f_1(\mathbf{s}_t) \ldots f_N(\mathbf{s}_t) \right]^T$ and $\left\{ f_n : R_{0+}^{M} \to R_{0+} \right\}_{n=1}^{N}$. Problem (1) can be casted in the matrix framework:

$$\mathbf{X} = \mathbf{f}(\mathbf{S}) \tag{2}$$

such that $\mathbf{X} \in R_{0+}^{N\times T}$, $\mathbf{S} \in R_{0+}^{M\times T}$, where $\{\mathbf{x}_t\}_{t=1}^{T}$ and $\{\mathbf{s}_t\}_{t=1}^{T}$ are column vectors of matrices $\mathbf{X}$ and $\mathbf{S}$ respectively and $\mathbf{f}(\mathbf{S})$ implies that nonlinear mapping is performed column wise such as in (1). It is further assumed that: $\left\{ \|\mathbf{s}_t\|_0 \le L \right\}_{t=1}^{T}$ where $\|\mathbf{s}_t\|_0$ stands for ℓ_0 quasi-norm that counts number of non-zero coefficients of $\mathbf{s}_t$ and $L = \max_{t=1,...,T} \|\mathbf{s}_t\|_0$. Evidently, it applies: $L \le M$, where L denotes maximal number of sources that can be present at any coordinate t. The uNNBSS problem implies that components mass spectra, $\left\{ \mathbf{s}_m \in R_{0+}^{1\times T} \right\}_{m=1}^{M}$, ought to be inferred from mixture data matrix $\mathbf{X}$ only. In this paper the following assumptions are made on nonlinear mixture model (1)/(2):

A1) $0 \le x_{nt} \le 1$ $\forall n=1,...,N$ and $\forall t=1,...,T$,

A2) $0 \le s_{mt} \le 1 \quad \forall m=1,...,M$ and $\forall t=1,...,T$,

A3) $M > N$,

A4) Amplitude s_{mt} obeys exponential distribution on (0, 1] interval and discrete distribution at zero, see also eq.(3),

A5) Components of the vector valued function $\mathbf{f}(\mathbf{s})$: $f_n(\mathbf{s}): R_{0+}^{M\times 1} \mapsto R_{0+}$, $\forall n=1,...,N$ are differentiable up to unknown order K,

A6) $M << T$.

To avoid confusion between column and row vectors they will be indexed by lowercase letters that correspond with uppercase letters related to dimensions of the corresponding matrix. As an example $\mathbf{s}_t$ refers to the column- and $\mathbf{s}_m$ to the row vector of matrix $\mathbf{S} \in R_{0+}^{M\times T}$. Evidently, uppercase bold letters denote matrices, lowercase bold letters denote vectors and italic lowercase letters denote scalars. In order to be useful solution of the uNNBSS problem is expected to be essentially unique, that is estimated matrix of pure components (sources) $\hat{\mathbf{S}}$ and the true matrix of pure components $\mathbf{S}$ have to be related through $\hat{\mathbf{S}} = \mathbf{P}\Lambda\mathbf{S}$, where $\mathbf{P}$ and $\mathbf{\Lambda}$ stand respectively for $M\times M$ permutation and diagonal matrices. As discussed at great length in [4] even linear underdetermined BSS problem requires constraints to be imposed on sources in order to ensure essentially unique solution. Nonlinear BSS problem is more difficult. Herein, we assume that pure components $\{\mathbf{s}_m\}_{m=1}^{M}$ comply with sparse probabilistic model imposed by A4. It implies that each component will be zero at great part of its support (number of *m/z* channels T)

as well as that non-zero intensity will be distributed according to exponential distribution with small expected value. These two constraints are expected to ensure that, in probability, compared to N and M the maximal number of analytes L present at the particular m/z coordinate is small enough. However, N stands for number of biological samples available and it is expected to be small. Thus, it can virtually be impossible to satisfy above requirement. That is why, as in [4], in order to increase the number of measurements (samples) the original uNNBSS problem (1) has to be mapped into RKHS by using EKM. Before that, we need to approximate uNNBSS problem (1)/(2) by an equivalent uLNBSS problem.

2.2 Sparse probabilistic model of source signals

Taylor expansion of the nonlinear model (1) up to an arbitrary order K is derived in Supporting Information. It is shown that uNNBSS problem (1) can be represented by an equivalent uLNBSS problem, eq. (7) in Supporting Information, comprised of M original sources and $\sum_{k=2}^{K} M^{(k)}$ higher order monomials, where $M^{(k)} = \binom{M+k-1}{k}$. Thus, without further constraints uNNBSS problem (1) is computationally intractable. That is why, according to A4, we assume that sources **s** comply with sparse probabilistic model comprised of mixed state distribution [14, 15, 4]:

$$p(s_{mt}) = \rho_m \delta(s_{mt}) + (1-\rho_m)\delta^*(s_{mt}) f(s_{mt}) \quad \forall m = 1,...,M \quad \forall t = 1,...,T \qquad (3)$$

where $\delta(s_{mt})$ is an indicator function and $\delta^*(s_{mt})=1-\delta(s_{mt})$ is its complementary function, $\rho_m = \{P(\mathbf{s}_{mt}=0)\}_{t=1}^{T}$. Hence, $\{P(s_{mt}>0)=1-\rho_m\}_{t=1}^{T}$. The nonzero state of s_{mt} is distributed according to $f(s_{mt})$. We have chosen exponential distribution: $f(s_{mt})=(1/\mu_m)\exp(-s_{mt}/\mu_m)$ to model sparse distribution of the nonzero states, in which case the most probable outcomes are equal to μ_m. It has been verified in [4] that model (3) describes well mass spectra of the pure components. Herein, by using mass spectra of 25 pure components we have estimated $\hat{\rho}_m \in [0.27, 0.74]$ and $\hat{\mu}_m \in [0.0012, 0.0014]$, see section 3.2 and Figure 4 for more details.[2] Under exponential prior, probability that amplitude $s_{mt} \in [\varepsilon, \mu_m]$, for $0<\varepsilon<<1$, is 0.632. Thus, in 36.8% of the cases random realization of s_{mt} will have amplitudes greater than most probable value μ_m. For a given μ_m and given probability $p(\varepsilon<s_{mt}\leq s)$ the value of s is obtained as: $s\approx-\mu_m\ln(1-p)$. Thus, for p=0.99 and $\mu_m=1.5\times10^{-3}$ it follows $s=7\times10^{-3}$. Hence, we may approximate equivalent uLNBSS model, eq. (7) in Supporting Information, by retaining second order terms only:

$$\mathbf{X}=\mathbf{G}_{(1)}^{1}\mathbf{S}+\frac{1}{2}\mathbf{G}_{(1)}^{2}\begin{bmatrix}\mathbf{s}_1^2\\ \ldots\\ \mathbf{s}_M^2\\ \ldots.\\ \{\mathbf{s}_{m_1}\mathbf{s}_{m_2}\}_{m_1,m_2=1}^{M}\end{bmatrix}+HOT \qquad (4)$$

[2] Even though the exponential distribution has support on the $[0,\infty)$ interval, by setting μ=0.01 realizations will be contained in [0, 1] interval with a probability that is close to 1 with an error of 3.72×10^{-44}. Thus, this justifies a choice of exponential distribution to model sparse distribution of amplitudes s_{mt} on interval [0, 1].

where $\mathbf{G}^1_{(1)}$, respectively $\mathbf{G}^2_{(1)}$, stand for unfolded versions of the tensor of first, respectively second, order derivatives and *HOT* stands for higher-order terms. Contribution of third order terms in (4) is of the order $(7\text{x}10^{-3})^3=3.43\times10^{-7}$. In order to reduce *HOT* entry-wise thresholding of **X** can be performed. By neglecting fourth- and higher-order terms we have empirically arrived at the threshold value of: $\tau\in[10^{-6}, 10^{-4}]$.[3]

2.3 Suppression of higher order (error) terms

Mass spectra of 25 pure components recorded in nonlinear chemical reaction of peptide bond formation, see section 3.2 and Figures 3 and S-4 in Supporting Information, illustrate diversity of morphologies. Some have few very dominant (large) peaks (see spectra of pure components 1, 2, 8, 13, 16, 17, 18, 19, 20, 21, 22, 23, 24 and 25), some have intensities distributed on several *m/z* values, whereas intensities can be small (see spectra of pure components 3, 4, 5, 6, 7, 9, 10, 11, 12, 14 and 15). It is thus hard to propose one preprocessing (thresholding) transform for suppression of higher order terms induced by nonlinear mixing process. We, therefore, propose the combination of methods for this purpose.

[3] These threshold values can be justified by the following analysis. Due to A1 and A2 elements of **G** in (7) in Supporting Information are less than 1. In pursuing worst case analysis of third-order effects we assume that third-order derivatives coefficients in **G** are less than some value g_3. Thus, contribution of third-order terms is limited by above by $x^{(3)}=M^{(3)}g_3 s$. If mixture value x_{nt} is greater than $x^{(3)}$ then it is probably due to first and second-order terms. The threshold value evidently depends on values of $M^{(3)}$, g_3 and s. For example, assuming M=100 ($M^{(3)}$=171700), g_3=0.1 and $s=3.4\times10^{-7}$ we get $x^{(3)}=5.8\times10^{-3}$. However, that is overly pessimistic given the fact that most of the third-order cross-products will, due to sparseness, vanish. Thus, optimal threshold value is somewhere in the interval $[10^{-6}, 10^{-4}]$.

2.3.1 Robust principal component analysis

RPCA has been proposed in [16, 17] to decompose data matrix **X** into sum of two matrices: **X**=**A**+**E**. Provided that **A** is low rank matrix and **E** is sparse matrix decomposition is unique and it is obtained as a solution of the optimization problem:

$$\textit{minimize } \|\mathbf{A}\|_* + \lambda \|\mathbf{E}\|_1 \textit{ subject to}: \mathbf{A} + \mathbf{E} = \mathbf{X}. \quad (5)$$

Thereby, $\|\mathbf{A}\|_* = \sum_{i=1}^{I \le N} \sigma_i$ denotes nuclear norm (sum of singular values) and $I \le N$ is a rank of matrix **A**; $\|\mathbf{E}\|_1 = \sum_{n=1}^{N} \sum_{t=1}^{T} e_{nt}$ denotes ℓ_1-norm of **E** and $\lambda \approx 1/\sqrt{T}$ is a regularization constant. In term of equivalent uLNBSS problem (4) **A** is associated with first and second order terms and **E** is associated with *HOT*. **A** is actually represented by linear mixture model composed of $2M + M(M\text{-}1)/2$ sources and N mixtures. Since both N and $2M+M(M\text{-}1)/2$ are small compared to T rank of **A** equals $min(N, 2M + M(M\text{-}1)/2)=N$. Thus, it is low. **E** is comprised of monomials (products of the original source components) of the order three- or higher. Since by assumption A4 source components are sparse in support and amplitude their three- and higher-order products are either zero or very small. Thus, **E** is sparse. Therefore, it is justified to use RPCA decomposition of **X** in (4) to suppress higher-order terms induced by nonlinear mixing process. That yields approximation of **X**, that is **A** , with suppressed higher-order terms. In the experiments reported in Section 3 we have used accelerated proximal gradient algorithm [41] , available with a MATLAB code at [42], to solve (5).

2.3.2 Hard thresholding

Hard thresholding (HT) operator, [18], can be applied entry-wise to **X** in (4) according to: $b_{nt} = HT(x_{nt}) = \begin{cases} x_{nt} & if \quad x_{nt} \geq \tau_1 \\ 0 & if \quad x_{nt} < \tau_1 \end{cases}$, n=1,...,N, t=1,...,T and $\tau_1 \in [10^{-6}, 10^{-4}]$ stands for a threshold. HT preprocessing transform of **X** yields matrix **B** that is expected to have the same structure as **A** in (5).

2.3.3 Soft thresholding

Soft thresholding (ST) operator, [18], can be applied entry-wise to **X** in (4) according to $c_{nt} = ST(x_{nt}) = \max(0, x_{nt} - \tau_2)$, n=1,...,N, t=1,...,T and $\tau_2 \in [10^{-6}, 10^{-4}]$. ST preprocessing transform of **X** yields matrix **C** that, as **B** obtained by HT**,** is also expected to have the same structure as **A** in (5).

2.3.4 Trimmed thresholding

Trimmed thresholding (TT) operator, [19], is applied entry-wise to **X** in (4) according to: $d_{nt} = TT(x_{nt}) = \begin{cases} x_{nt} \dfrac{x_{nt}^{\alpha} - \tau_3^{\alpha}}{x_{nt}^{\alpha}} & if \quad x_{nt} \geq \tau_3 \\ 0 & if \quad x_{nt} < \tau_3 \end{cases}$, n=1,...,N, t=1,...,T and $\tau_3 \in [10^{-6}, 10^{-4}]$. α is a trade-off parameter between hard and soft thresholding. When α=1, TT equals ST. When $\alpha \to \infty$ TT is equivalent to HT. Herein, we set α=3.5 because this value yields TT to operate between ST and HT [19]. TT preprocessing transform of **X** yields matrix **D**

that, as **B** obtained by HT and **C** obtained by ST**,** is also expected to have the same structure as **A** in (5).

2.4 Empirical kernel map based nonlinear mapping of preprocessed mixture matrix

So far we have substituted uNNBSS problem (1)/(2) by four uLNBSS problems in a form of (4). While original uNNBSS problem is characterized by nonlinear multivariate mapping **f** and triplet (*N, M, L*) the uLNBSS problems are characterized by (*N*, *P*, *Q*) where $P \approx 2M + M(M-1)/2$ stands for number of dependent sources in (4) and $Q \approx 2L + L(L-1)/2$ stands for maximal number of sources at particular *m/z* coordinate. Since by assumption A3 $M>N$ it follows that $P>>N$. Thus, even with activation of sparseness constraints imposed by A4 it will be virtually impossible to ensure essentially unique solution of these uLNBSS problems. To this end, as in [4], we apply the EKM-based nonlinear mapping of uLNBSS problems represented by preprocessed mixture matrices **A**, **B**, **C** and **D** to RKHS in order to increase number of samples/mixtures from N to $D>>N$. Theory and discussion related to it has been presented in great details in section 2.2 in [4]. We therefore present it in a concise form herein. EKM Ψ of column vectors $\{\mathbf{a}_t\}_{t=1}^{T}$ in (4) with respect to a basis $\{\mathbf{v}_d\}_{d=1}^{D}$ is $\psi : R^N \to R^D$, such that: $\mathbf{a}_t \mapsto \kappa(\circ, \mathbf{a}_t)\Big|_{\{\mathbf{v}_d\}_{d=1}^{D}} = \left[\kappa(\mathbf{v}_1, \mathbf{a}_t), ..., \kappa(\mathbf{v}_D, \mathbf{a}_t)\right]^T \ \forall t = 1, ..., T$. Thereby, $\kappa(\mathbf{v}_d, \mathbf{a}_t)$ is a positive definite symmetric function. The basis $\{\mathbf{v}_d\}_{d=1}^{D}$ has to span the empirical set of patterns $\{\mathbf{a}_t\}_{t=1}^{T}$ such that $span\{\mathbf{v}_d\}_{d=1}^{D} \approx span\{\mathbf{a}_t\}_{t=1}^{T}$. In this case $span\{\phi(\mathbf{v}_d)\}_{d=1}^{D} \approx span\{\phi(\mathbf{a}_t)\}_{t=1}^{T}$, where $\{\mathbf{a}_t \mapsto \phi(\mathbf{a}_t) \in R_{0+}^{\bar{N}}\}_{t=1}^{T}$, i.e.

$\left\{\mathbf{v}_d \mapsto \phi\left(\mathbf{v}_d\right) \in R_{0+}^{\bar{N}}\right\}_{d=1}^{D}$, is in principle infinite dimensional nonlinear mapping. If $\phi\left(\mathbf{a}_t\right)=\kappa\left(\circ,\mathbf{a}_t\right)$, respectively $\phi\left(\mathbf{v}_d\right)=\kappa\left(\circ,\mathbf{v}_d\right)$, projection of $\left\{\phi\left(\mathbf{a}_t\right)\right\}_{t=1}^{T}$ onto $\left\{\phi\left(\mathbf{v}_d\right)\right\}_{d=1}^{D}$ yields in matrix form:

$$\Psi\left(\mathbf{A}\right)=\begin{bmatrix} \kappa\left(\mathbf{a}_1,\mathbf{v}_1\right) & \ldots & \kappa\left(\mathbf{a}_T,\mathbf{v}_1\right) \\ \ldots & \ldots & \ldots \\ \kappa\left(\mathbf{a}_1,\mathbf{v}_D\right) & \ldots & \kappa\left(\mathbf{a}_T,\mathbf{v}_D\right) \end{bmatrix} \tag{6}$$

Herein, as in [4], we choose $\kappa\left(\mathbf{a}_t,\mathbf{v}_d\right)=\exp\left(-\left\|\mathbf{a}_t-\mathbf{v}_d\right\|^2/\sigma^2\right)$. When assumption A1 holds we can set $\sigma^2\approx 1$. We analogously obtain EKM-mappings of matrices **B**, **C** and **D** and that respectively yields $D\times T$ matrices Ψ(**B**), Ψ(**C**) and Ψ(**D**). Likewise, as in [4], we use *k-means* data clustering algorithm to estimate basis **V** by clustering $\left\{\mathbf{a}_t\right\}_{t=1}^{T}$ in D clusters. Thereby, by setting $D=T$ clustering is unnecessary because each empirical pattern is a basis vector. That, however, comes at increased computing cost. By using sparseness assumption A4 it is shown in [4] that:

$$\Psi\left(\mathbf{A}\right)=\mathbf{Z}+\bar{\mathbf{G}}\begin{bmatrix}\mathbf{0}_{1\times T}\\ \bar{\mathbf{S}}\end{bmatrix}+HOT \tag{7}$$

where **Z** is a bias term and does not play a role in parts based decomposition that follow, $\mathbf{0}_{1\times T}$ is row vector of zeros and $\bar{\mathbf{S}} \in R_{0+}^{P\times T}$ is matrix with $P \approx 2M + M(M-1)/2$ rows that contain original source components and their second order monomials. $\bar{\mathbf{G}}$ is a matrix of appropriate dimensions. EKM-mapped matrices Ψ(**B**), Ψ(**C**) and Ψ(**D**) follow the same approximation as Ψ(**A**) in (7). It is important to emphasize that in (4) higher order (error) terms are induced by nonlinear mixing process **f**(**X**) while in (7) they are induced by nonlinear character of the EKM. That is, increase of number of mixtures from *N* to *D* in Ψ(**A**), Ψ(**B**), Ψ(**C**) and Ψ(**D**) comes at the cost of errors induced by the EKM. However, as in [4] and (4), we can again apply preprocessing transforms to suppress *HOT*. Since matrices **B**, **C** and **D** were obtained by respectively applying *HT*, *ST* and *TT* operators on **X** in (4) we apply these operators in the same order on Ψ(**B**), Ψ(**C**) and Ψ(**D**). In order to keep level of notational complexity as low as possible we keep the same notation for thresholded versions of matrices Ψ(**B**), Ψ(**C**) and Ψ(**D**). We do not apply RPCA decomposition on Ψ(**A**) because rank of it is dictated by **Z** and is equal to $min(D,T)=D$ and that is not low. The final effect of EKM-based mappings is to ensure that sparseness constrained factorization of Ψ(**A**), Ψ(**B**), Ψ(**C**) and Ψ(**D**) yields, with greater probability, more accurate solution compared to decomposition by the same method of **A**, **B**, **C** and **D**. That will be the case when the following condition holds:

$$(D/N) >> (P/M) \text{ and } (D/N) >> (Q/L). \quad (8)$$

Because $P \approx 2M+M(M-1)/2$ and $Q \approx 2L+L(L-1)/2$ condition (8) becomes: $(D/N) >> (M/2-3/2)$ and $(D/N) >> (L/2-3/2)$. Numerical problem studied in section 3 is characterized by $N=3$, $M=8$, $L=3$ and $D=T=1000$. Evidently, above condition is fulfilled.

2.5 Sparseness constrained factorization

To increase accuracy of the pure components extraction we apply sparseness constrained NMF (sNMF) in RKHS to matrices $\Psi(\mathbf{A})$, $\Psi(\mathbf{B})$, $\Psi(\mathbf{C})$ and $\Psi(\mathbf{D})$.[4] That yields four sets of separated components:

$$\left\{\overline{\mathbf{s}}_m^{\mathbf{A}}\right\}_{m=1}^{P} = sNMF\left(\Psi\left(\mathbf{A}\right)\right) \tag{9}$$

$$\left\{\overline{\mathbf{s}}_m^{\mathbf{B}}\right\}_{m=1}^{P} = sNMF\left(\Psi\left(\mathbf{B}\right)\right) \tag{10}$$

$$\left\{\overline{\mathbf{s}}_m^{\mathbf{C}}\right\}_{m=1}^{P} = sNMF\left(\Psi\left(\mathbf{C}\right)\right) \tag{11}$$

$$\left\{\overline{\mathbf{s}}_m^{\mathbf{D}}\right\}_{m=1}^{P} = sNMF\left(\Psi\left(\mathbf{D}\right)\right) \tag{12}$$

When it comes to implementation of the sNMF algorithms we use, as in [4], the NMU algorithm [40] with a MATLAB code available at [44] and the NMF_L0 algorithm [39] with a MATLAB code available at [45]. The NMF_L0 algorithm was run with the

[4]To ensure essentially unique decomposition sparseness constrained NMF algorithms have been formulated such as [38, 39, 40]. However, only very recently it is proved in [43], see Theorem 4 and Corollary 2, that uniqueness of some asymmetric NMF **S=WH** implies that each column of **W** (row of **H**) contains at least M-1 zeros, where M is nonnegative rank of **S**.

following parameter setup: reverse sparse nonnegative least square sparse coder and alternating nonnnegative least square for dictionary update stage. A main reason for preferring the NMU algorithm over other sparseness constrained NMF algorithms is that there are no regularization constants that require a tuning procedure. When performing NMU-based factorizations in (9) to (12) the unknown number of pure components P needs to be given to the algorithm as an input. As in [4] we set: $P = D = T$. That is, in order not to lose some component we prefer to extract all T rank-one factors.[5] These four sets of separated components are compared with the pure components stored in the library using normalized correlation coefficient. Each pure component is associated with the separated component by which it has the highest correlation. As a reference in the benchmark numerical study we have used solution obtained by applying the NMF_L0 algorithm to the (9) to (12). Afterwards, maximal correlation criterion has been used to assign separated components to pure components in the library. NMF_L0 is based on natural sparseness measure, the ℓ_0-pseudo-norm of the component matrix $\overline{\mathbf{S}}$, and that is known from compressed sensing theory, [47], to yield the best results when sparseness of $\overline{\mathbf{S}}$ decreases. The NMF_L0 when applied in (9) to (12) requires *a priori* information on the number of components P and number of overlapping components Q and they are related to M and L through: $P \approx 2M + M(M\text{-}1)/2$ and $Q \approx 2L + L(L\text{-}1)/2$. In numerical scenario both M and L are known while in experimental

[5] The factorization problems (9) to (12) are related to the determination of nonnegative rank of nonnegative matrix and that is defined as the smallest number of rank one matrices into which original matrix can be decomposed [46]. For some matrix $\mathbf{\Psi} \in R_{0+}^{D \times T}$ with $D \leq T$ nonnegative rank equals the smallest positive integer P for which there exists nonnegative column vectors $\{\mathbf{g}_p\}_{p=1}^{P}$ such that each column vector of $\mathbf{\Psi}$ can be represented as linear combination with nonnegative coefficients of the column vectors $\{\mathbf{g}_p\}_{p=1}^{P}$.

scenario selection of optimal (true) value of L is hard. We summarize the PTs-EKM-NMU/NMF_L0 algorithm in the Algorithm 1.

3.0 EXPERIMENTS

Studies on numerical and experimental data reported below were executed on personal computer running under Windows 64-bit operating system with 64GB of RAM using Intel Core i7-3930K processor and operating with a clock speed of 3.2 GHz. MATLAB 2012b environment has been used for programming.

3.1 Numerical study

In numerical study we simulate uNNBSS problem (2) with N=3, M=8, L=3 and T=1000. Source signals were generated according to mixed state probabilistic model (3) with exponential prior. Thereby, μ_m=1.5×10^{-3} $\forall m$=1,...,M. We have generated two scenarios with ρ_m=0.5 and ρ_m=0.8 $\forall m$=1,...,M. Values for μ_m and ρ_m are equivalent to those obtained by fitting probabilistic model (3) to experimental mass spectra of 25 pure components, see section 3.2 and Figure 4 for details. The uNNBSS problem (2) has been simulated using nonlinear mixtures:

$$f_1(\mathbf{s}) = s_1^3 + s_2^2 + \tan^{-1}(s_3) + s_4^2 + s_5^3 + s_6^3 + \tanh(s_7) + \sin(s_8)$$

$$f_2(\mathbf{s}) = \tanh(s_1) + s_2^3 + s_3^3 + \tan^{-1}(s_4) + \tanh(s_5) + \sin(s_6) + s_7^2 + s_8^2$$

$$f_3(\mathbf{s}) = \sin(s_1) + \tan^{-1}(s_2) + s_3^2 + s_4^3 + \tanh(s_5) + \sin(s_6) + s_7^3 + \tan^{-1}(s_8)$$

Nonlinear mixtures are chosen arbitrary to demonstrate capability of proposed algorithm to solve uNNBSS problem comprised of unknown nonlinear mixtures. HT, ST and TT operators used in steps 2, 3, 4 and 6 in Algorithm 1 were implemented with $\tau=10^{-5}$ and $\alpha=3.5$ has been used for TT operator. Gaussian kernel based EKM has been used with $\sigma^2=1$ and $D=T=1000$. Table 1 shows results of comparative analysis, for the case of $\rho_m=0.8$, obtained by NMU and NMF_L0 applied to uNNBSS (1)/(2); NMU and NMF_L0 applied in (9) to (12) without suppression of higher order monomials (EKM-NMU and EKM-NMF_L0); and NMU and NMF_L0 applied in (9) to (12) after RPCA, HT, ST and TT preprocessing transforms (PTs-EKM-NMU and PTs-EKM-NMF_L0). Due to sparse prior imposed on sources it was reasonable to expect that useful results can be obtained by direct factorization of uNNBSS problem (2). Results for $\rho_m=0.5$ are shown in Table S-1 in Supporting Information, while results for $\rho_m=0.8$ and $\rho_m=0.5$ as a function of Monte Carlo index are shown in Figure 1. For the value of normalized correlation coefficient between pure component and assigned separated component we evaluate performance in term of four metrics described in caption of Table 1. They are defined with respect to predefined labeling of the pure components stored in the library. The first three metrics are calculated for correctly assigned components only. That is why NMU and NMF_L0 appear to have comparable performance in term of mean and minimal correlation metrics. But they are inferior in number of separated components correlated with pure components with correlation greater than or equal to 0.6 as well as in number of (in)correctly assigned separated components (due to poor separation).

Thereby, incorrect assignment implies that two or more pure components are assigned to the same separated component. We also can see that preprocessing transforms improve performance compared with factorizations of mixtures data without preprocessing related to suppression of higher order monomials.

3.2 Experimental data on chemical reaction comprising peptide synthesis

3.2.1 Chemicals

Chemical reaction has been performed according to the following procedure: L-Leucine (200 mg, 1.52 mmol) was dissolved in 5 mL of dry dimethylformamide (DMF) and solution was cooled to 0 °C. N-methylmorpholine (NMM, 3.05 mmol, 337 µL) and isobutylchloroformate (IBCF, 3.34 mmol, 458 µL) were added. Aliquots of the reaction mixture (100µL) were withdrawn every 30 minutes (t_0-t_8), solvent was evaporated and the residue dissolved in 1mL of 0.1 % formic acid (FA) in 50 % MeOH. Aliquots (100 µL) were diluted with 400 µL of 0.1 % FA in 50 % MeOH and 10 µL were injected through autosampler on a column (Zorbax XDB C18, 3.5 µm, 4.7 mm) at the flow rate of 0.5 mL/min. Mobile phase was 0.1 % FA in water (solvent A) and 0.1 % FA in MeOH (solvent B). Gradient was applied as follows: 0 min 40 % B; 0-15 min 90 %B; 12-15min 90% B; 17.1 min 40% B; 17.1-20 min 40 %B. Figure S-2 in Supporting Information shows 9 chromatograms corresponding to reaction mixture recorded at 9 time instants (t_0-t_8) during the reaction. Mass spectra of 9 mixtures ($\mathbf{x}_1$ to $\mathbf{x}_9$), obtained by full integration of chromatograms, and mass spectra of 25 pure components ($\mathbf{s}_1$ to $\mathbf{s}_{25}$) arising during the reaction are respectively shown in Figure S-3 and Figure S-4 in

Supporting Information. Mass spectra of pure components 1, 4, 8 and 11 are also shown in Figure 3.

3.2.2. Mass spectroscopy measurements

Electrospray ionization-mass spectrometry (ESI-MS) measurements operating in a positive ion mode were performed on a HPLC-MS triple quadrupole instrument equipped with an autosampler (Agilent Technologies, Palo Alto, CA, USA). The desolvation gas temperature was 300^0C with flow rate of 8.0 L/min. The fragmentor voltage was 135 V and capillary voltage was 4.0 kV. Mass spectra were recorded in *m*/*z* segment of 10-2000. All data acquisition and processing was performed using Agilent MassHunter software. Acquired mass spectra are composed of intensities at T=9901 *m/z* coordinates.

3.2.3. Setting up an experiment

Peptides and proteins are compounds involved in numerous biological processes of key importance, like cell-cell communication, immune response, cell growth and proliferation, hormonal and enzymatic activity. They are, therefore of ever-increasing interest as tools in studies of biological systems and modulators of biological functions. Chemical synthesis of peptides involves condensation of two suitably protected parts (amino acids or peptides) in order to obtain single, desirable product. However, for the purpose of this work, a different approach was undertaken. Non-protected amino acid, L-leucin, was allowed to react under basic conditions (NMM) in the presence of IBCF

giving various products: di-, tri-, tetrapeptides as well as corresponding intermediates. Nonlinearity of the described reaction was assured based on the following: (*i*) concentration of individual components does not change linearly with time and (*ii*) as reaction proceeds, new components appear that were not present at the beginning of the reaction. Figure 2 schematically describes possible components present in the reaction mixture. It is important to note, that aim of this experiment was not to determine structure of all components, but to provide reliable experimental data on nonlinear reaction. Library of compounds required for the validation of algorithm was built by integration of each peak in the chromatogram corresponding to the mixture $\mathbf{x}_9$ and subsequent extraction of mass spectrum. During the library generation, no discrimination based on the intensity of peaks was made. Therefore, all peaks were treated as pure components.

4. RESULTS AND DISCUSSION

Inspection of pure components mass spectra shown in Figures S-4 in Supporting Information shows significant overlapping, resulting from the similarity of chemical structure of components. Pure components 1 and 2, 16 and 17 as well as 19 and 21 have normalized correlation coefficient above 0.97 and, consequently, they are impossible to be distinguished. In addition to that, pure components 5 and 7 have normalized correlation coefficient above 0.78. Thus, they are also expected to be very hard to discriminate. However, we expect from proposed PTs-EKM_NMU method to be able to discriminate the rest of the components. That is not trivial given the fact that normalized correlation coefficients for 26 combinations of pure components vary between 0.1 and

0.44. That makes the uNNBSS problem comprised of correlated pure components very hard. Correlation matrix of the pure components mass spectra, where pairs of pure components are identified with normalized correlation coefficient above 0.1, is shown in Table S-2 in the Supporting Information. As emphasized previously, it is sparseness of the pure components mass spectra in support and amplitude that is expected to enable solution of related uNNBSS. To this end, mixed state probabilistic model (3) with exponential prior on continuous distribution of the non-zero amplitude has been fitted to experimental pure components mass spectra (they are shown in Figure S-4 in Supporting Information as well as in Figure 3 for pure components 1, 4, 8 and 11). Even though these pure components are correlated with others and some (4 and 11) have small intensity they are uniquely assigned to the true pure components from the library. Figure 4 (left), also Figure S-5 in Supporting Information, shows estimated probability that value of the pure component mass spectra is zero. As can be seen 22 out of 25 pure components have zero amplitudes at 40% to 75% of their support. Figure 4 (right), also Figure S-6 in Supporting Information, shows most expected values (mean) of exponential distribution estimated by fitting exponential distribution to amplitude histograms. They were estimated for 25 pure components in the range (0, 1] within intervals of the 0.01 width. It can be seen that $\hat{\mu}_m \in [0.0012, 0.0014]$ for m=1,...,25. Figure S-7 shows probability that amplitude of the pure components mass spectra occurs in interval [0, A], such that 0.01≤A≤1. That is an average estimate over 25 pure components. It is seen that 0.01≤A≤0.08 occurs with probability 0.97. Reported results confirm that sparse probabilistic model (3) is experimentally well grounded. That is further confirmed by Figure S-8 in Supporting Information that shows estimated histograms (stars) and exponential probability density functions (squares) calculated

with the mean values from Figure 4-right. It is seen that approximation is very good. Estimated histograms vs. exponential probability density functions for pure components 1, 4, 8 and 11 are also shown in Figure 5.

Table 2 presents results of comparative performance analysis using the four metrics as in section 3.1 for NMU, EKM-NMU, PTs-EKM-NMU for D=T=9901 and PTs-EKM-NMU for D=4000. Thus, in the last case *k-means* clustering has been used to find a basis $\{\mathbf{v}_d\}_{d=1}^{4000}$ in the input space of patterns $\{\mathbf{x}_t\}_{t=1}^{9901}$. Provided that it retains accuracy, the subspace approximation is very important from computational reasons. That is because when four preprocessing transforms are combined, sparseness constrained NMF in (9) to (12) has to be performed four times. That can be done in parallel. Nevertheless, one factorization of the 9901×9901 matrix by NMU algorithm takes approximately 79 hours on above specified machine, while factorization of the 4000×9901 matrix by the same algorithm takes approximately 13.7 hours. For NMF_L0 number of overlapping components, L, has to be reported to the algorithm as input information. For Pts_EKM-NMF_L0 algorithm optimal value of L can be inferred by running NMF_L0 algorithm multiple times on problem such (9). That, however, would result in high computational costs. That is why NMF_L0 has not been used in RKHS on problems (9) to (12). It is seen from Table 2 that linear sparseness constrained matrix factorization yields poor quality of separation compared to linear factorization in the RKHS. That is especially the case with number of incorrectly assigned components and that is direct consequence of the low purity of separated components. That, indirectly, also confirms nonlinear character of the mixtures mass spectra of the desired chemical reaction. It is also seen that combination of four preprocessing transforms for

suppression of higher order monomials and sparseness constrained factorization in RKHS significantly improves quality of separation. In this regard, Figure S-9 in Supporting Information shows mass spectra of 25 separated components assigned to pure components according to maximal correlation criterion. Separated pure components 1, 4, 8 and 11 are also shown in Figure 3. Thereby, value of normalized correlation coefficient and preprocessing transform (RPCA, HT, ST or TT) that yielded best result are also reported. Due to the diversity of morphologies of mass spectra all four preprocessing transforms yielded best results at some cases. It is also important to notice that subspace approximation of proposed method with D=4000 yields results very comparable to those obtained by D=9901 but with much shorter computation time. Thus, proposed approach to pure components extraction can, when implemented on state-of-the-art multiprocessor (grid) platform, be executed in even shorter time which makes it practically relevant.

5.0 CONCLUSION

Blind source separation approach to pure components extraction is most often based on linear mixture model. That is, mixtures spectra are assumed to be the unknown weighted linear combination of pure components spectra. Herein, we have addressed problem related to extraction of pure components from nonlinear mixtures of mass spectra. Thereby, number of mixtures is assumed to be (significantly) less than number of pure components. We propose an approach that combines four preprocessing methods for suppression of higher order monomials induced by nonlinear mixing process and sparseness constrained nonnegative matrix factorization in RKHS induced

by EKM. Two practically important properties of the proposed approach are that no information about character of the nonlinear mixing process is required and that linear mixing problem is contained implicitly as a special case. It is believed that these properties make the proposed approach practically relevant for contemporary metabolic profiling of biological samples, that is pure components extraction in biomarker identification studies. Proposed approach is demonstrated on demanding numerical and experimental scenarios. In the last case, related to chemical reaction of synthesis of peptides, components separated from 9 nonlinear mixtures mass spectra are assigned uniquely to 25 the pure components from the library. On the same problem separation by linear NMF algorithms yielded 15 (NMU) and 7 (NMF_L0) incorrectly assigned components.

ACKNOWLEDGMENT

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. Nuzillard D, Bourg S, Nuzilard J M. Model-Free Analysis of Mixtures by NMR Using Blind Source Separation *J. Magn. Reson.* 1998; **133**: 358-363.

2. Visser E, Lee T W. An information-theoretic methodology for the resolution of pure component spectra without prior information using spectroscopic measurements. *Chemom. Int. Lab. Syst.* 2004; **70**: 147-155.

3. Kopriva I, Jerić I. Blind separation of analytes in nuclear magnetic resonance spectroscopy and mass spectrometry: sparseness-based robust multicomponent analysis. *Anal. Chem.* 2010; **82**: 1911-1920.

4. Kopriva I, Jerić I, Brkljačić L. Nonlinear mixture-wise expansion approach to underdetermined blind separation of nonnegative dependent sources. *J. Chemometrics* 2013; **27**: 189-197.

5. Roux A, Xu Y, Heilier J-F, Olivier M-F, Ezan E, Tabet J-C, Junot C. Annotation of the human adult urinary metabolome and metabolite identification using ultra high performance liquid chromatography coupled to a linear quadrupole ion trap-orbitrap mass spectrometer. *Anal. Chem*. 2012; **84**: 6429−6437.

6. Abu-Farha M, Elisma F, Zhou H, Tian R, Asmer M S, Figeys D. Proteomics: from technology developments to biological applications. *Anal. Chem.* 2009; **81**: 4585-4599.

7. McLafferty F W, Stauffer D A, Loh S Y, Wesdemiotis C. Unknown Identification Using Reference Mass Spectra. Quality Evaluation of Databases. *J. Am. Soc. Mass. Spectrom.* 1999; **10**: 1229-1240.

8. Hyvärinen A, Karhunen J, Oja E. *Independent Component Analysis*. John Wiley & Sons, Inc.: New York, US, 2001.

9. Cichocki A, Amari S. *Adaptive Blind Signal and Image Processing*. John Wiley: New York, 2002.

10. Cichocki A, Zdunek R, Phan A H, Amari, S I. *Nonnegative Matrix and Tensor Factorizations*. John Wiley: Chichester, UK, 2009.

11. Comon P, Jutten C (eds). *Handbook of Blind Source Separation*. Academic Press: Oxford, UK, 2010.

12. Walleczek J(ed). *Self-organized biological dynamics and non-linear control*. Cambridge University Press: Cambridge, UK. 2000

13. Nicholson J K, Lindon J C. Systems biology: Metabonomics. *Nature* 2008; **455** (7216): 1054-1056.

14. Bouthemy P, Piriou C H G, Yao J. Mixed-state auto-models and motion texture modeling. *J. Math Imaging Vision* 2006; **25**: 387-402.

15. Caifa C, Cichocki A. Estimation of Sparse Nonnegative Sources from Noisy Overcomplete Mixtures Using MAP. *Neural Comput.* 2009; **21**: 3487-3518.

16. Candès E J, Li X, Ma Y, Wright H. Robust Principal Component Analysis? *J. ACM* 2011; **58**: Article 11 (37 pages).

17. Chandrasekaran V, Sanghavi S, Paririlo P A, Wilsky A S. Rank-Sparsity Incoherence for Matrix Decomposition. *SIAM J. Opt.* 2011; **21**: 572-596.

18. Donoho D L. De-Noising by Soft-Thresholding. *IEEE Trans. Inf. Theory* 1995; **41** (3): 613-627.

19. Fang H T, Huang D S. Wavelet de-noising by means of trimmed thresholding. in: *Proc. of the 5th World Congress on Intelligent Control and Automation*, June 15-19, 2004, Hangzhou, P. R. China, pp. 1621-1624.

20. The website of the ASTER spectral library. http://speclib.jpl.nasa.gov [13 January 2014]

21. Zhang K, Chan L. Minimal Nonlinear Distortion Principle for Nonlinear Independent Component Analysis. *J. Mach. Learn. Res.* 2008; **9**: 2455-2487.

22. Levin D N. Using state space differential geometry for nonlinear blind source separation. *J. Appl. Phys.* 2008; **103**: 044906:1-12.

23. Levin D N. Performing Nonlinear Blind Source Separation With Signal Invariants. *IEEE Trans. Sig. Proc.* 2010; **58**: 2131-2140.

24. Taleb A, Jutten C. Source Separation in Post-Nonlinear Mixtures. *IEEE Trans. Sig. Proc.* 1999; **47**: 2807-2820.

25. Duarte L T, Suyama R, Rivet B, Attux R, Romano J M T, Jutten C. Blind compensation of nonlinear distortions: applications to source separation of post-nonlinear mixtures. *IEEE Trans. Sig. Proc.* 2012; **60**: 5832-5844.

26. Filho E F S, de Seixas J M, Calôba L P. Modified post-nonlinear ICA model for online neural discrimination. *Neurocomputing* 2010; **73**: 2820-2828.

27. Nguyen T V, Patra J C, Das A. A post nonlinear geometric algorithm for independent component analysis. *Digital Sig. Proc.* 2005; **15**: 276-294.

28. Ziehe A, Kawanabe M, Harmeling S, Müller K R. Blind Separation of Post-Nonlinear Mixtures Using Gaussianizing Transformations And Temporal Decorrelation. *J. Mach. Learn. Res.* 2003; **4**: 1319-1338.

29. Zhang K, Chan L W. Extended Gaussianization Method for Blind Separation of Post-Nonlinear Mixtures. *Neural Comput.* 2005; **17**: 425-452.

30. Harmeling S, Ziehe A, Kawanabe M. Kernel-Based Nonlinear Blind Source Separation. *Neural Comput.* 2003; **15**: 1089-1124.

31. Martinez D, Bray A. Nonlinear Blind Source Separation Using Kernels. *IEEE Tr. Neural Net.* 2003; **14**: 228-235.

32. Almeida L. MISEP-Linear and nonlinear ICA based on mutual information. *J. Mach. Learn. Res.* 2003; **4**: 1297-1318.

33. Vaerenbergh S V, Santamaria I A. Spectral Clustering Approach to Underdetermined Postnonlinear Blind Source Separation of Sparse Sources. *IEEE Trans. Neural Net.* 2006; **17**: 811-814.

34. Buciu I, Nikolaidis N, Pitas I. Nonnegative Matrix Factorization in Polynomial Feature Space. *IEEE Trans. Neural Net.* 2007; **19**: 1090-1100.

35. Zafeiriou S, Petrou M. Non-linear Non-negative Component Analysis. *IEEE Trans. Image Proc.* 2010; **19**: 1050-1066.

36. Pan B, Lai J, Chen W S. Nonlinear nonnegative matrix factorization based on Mercer kernel construction. *Pattern Rec.* 2011; **44**: 2800-2810.

37. Yang Z, Xiang Y, Xie S, Ding S, Rong Y. Nonnegative Blind Source Separation by Sparse Component Analysis Based on Determinant Measure. *IEEE Trans. Neural Net. and Learn. Sys.* 2012; **23** (10): 1601-1610.

38. Cichocki A, Zdunek R, Amari S I. Hierarchical ALS Algorithms for Nonnegative Matrix Factorization and 3D Tensor Factorization. *LNCS* 2007; **4666**: 169-176.

39. Peharz R, Pernkopf, F. Sparse nonnegative matrix factorization with ℓ^0-constraints. *Neurocomputing* 2012; **80**: 38-46.

40. Gillis N, Glineur F. Using underapproximations for sparse nonnegative matrix factorization. *Pattern Rec*. 2010; **43**: 1676-1687.

41. Lin Z, Ganesh A, Wright J, Wu L, Chen M, Ma Y. Fast Convex Optimization Algorithms for Exact Recovery of a Corrupted Low-Rank Matrix. *UIUC Technical Report UILU-ENG-09-2214,* August 2009.

42. The website on low-rank matrix recovery and completion via convex optimization: http: //perception.csl.illinois.edu/matrix-rank/sample_code.html [13 January 2014]

43. Huang K, Sidiropoulos N D, Swami A. Non-Negative Matrix Factorization Revisited: Uniqueness and Algorithm for Symmetric Decomposition. *IEEE Trans. Sig. Proc.* 2014; **62**: 211-224.

44. The Nicolas Gillis Website. https://sites.google.com/site/nicolasgillis/code [13 January 2014].

45. The Robert Peharz Website. http://www3.spsc.tugraz.at/people/robert-peharz [13 January 2014].

46. Cohen J E, Rothblum U C. Nonnegative Ranks, Decompositions, and Factorizations of Nonnegative Matrices. *Linear Algebra and Its Applications* 1993; **190**: 149-168.

47. Chartran R, Staneva V. Restricted isometry properties and nonconvex compressive sensing. *Inverse Problems* 2008; **24**: 035020 (14 pages).

Figure Captions

Figure 1 (color online). Numerical study. Normalized correlation coefficient vs. Monte Carlo run index between true and extracted sources by algorithms: NMF_L0 (crosses), NMU (circles) and PTs-EKM-NMU (pluses) and PTs-EKM-NMF_L0 (stars). Mean value (first row), minimal value (second row), number of values greater than or equal to 0.6 (third row), number of incorrect pairs (fourth row). Probability of state zero equal to 0.5 (left column) and 0.8 (right column).

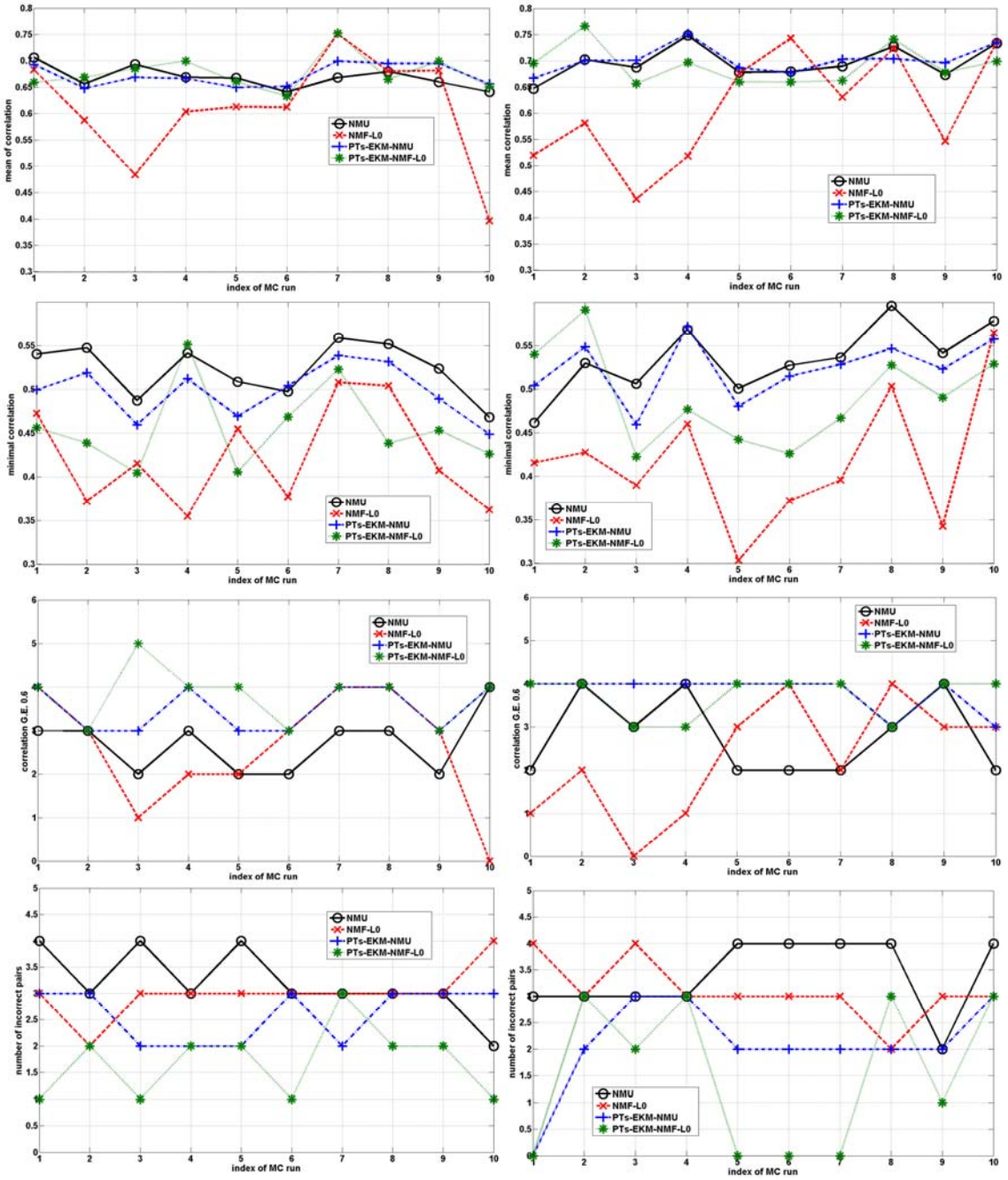


Fig. 1, Kopriva, Jerić, Filipović & Brkljačić

Figure 2. Structures of possible components present in the reaction mixture.

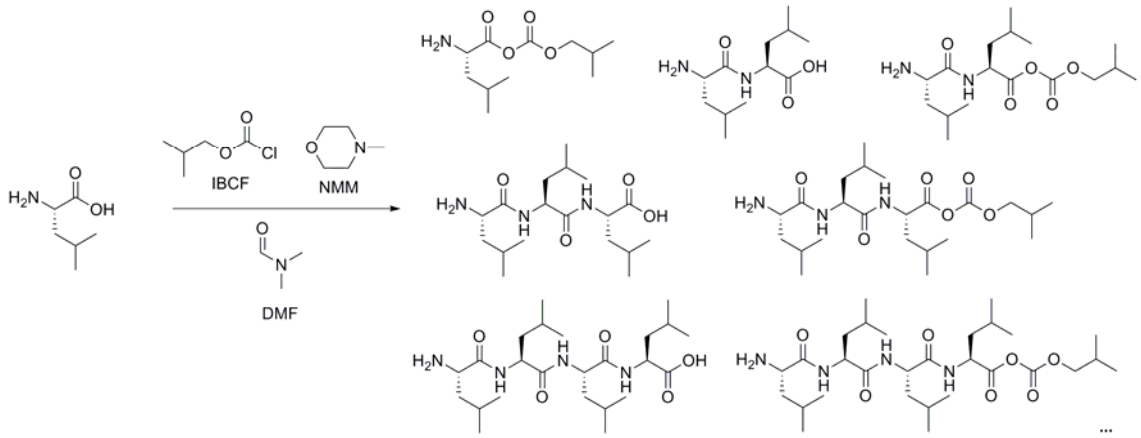


Fig. 2, Kopriva, Jerić, Filipović & Brkljačić

Figure 3. Two top rows: mass spectra of pure components $\mathbf{s}_1$, $\mathbf{s}_4$, $\mathbf{s}_8$ and $\mathbf{s}_{11}$. Two bottom rows: estimated mass spectra of pure components $\mathbf{s}_1$, $\mathbf{s}_4$, $\mathbf{s}_8$ and $\mathbf{s}_{11}$ by proposed PTs-EKM-NMU algorithm. Information on value of highest normalized correlation coefficient and associated error reduction method (RPCA, HT, ST and TT) are also displayed.

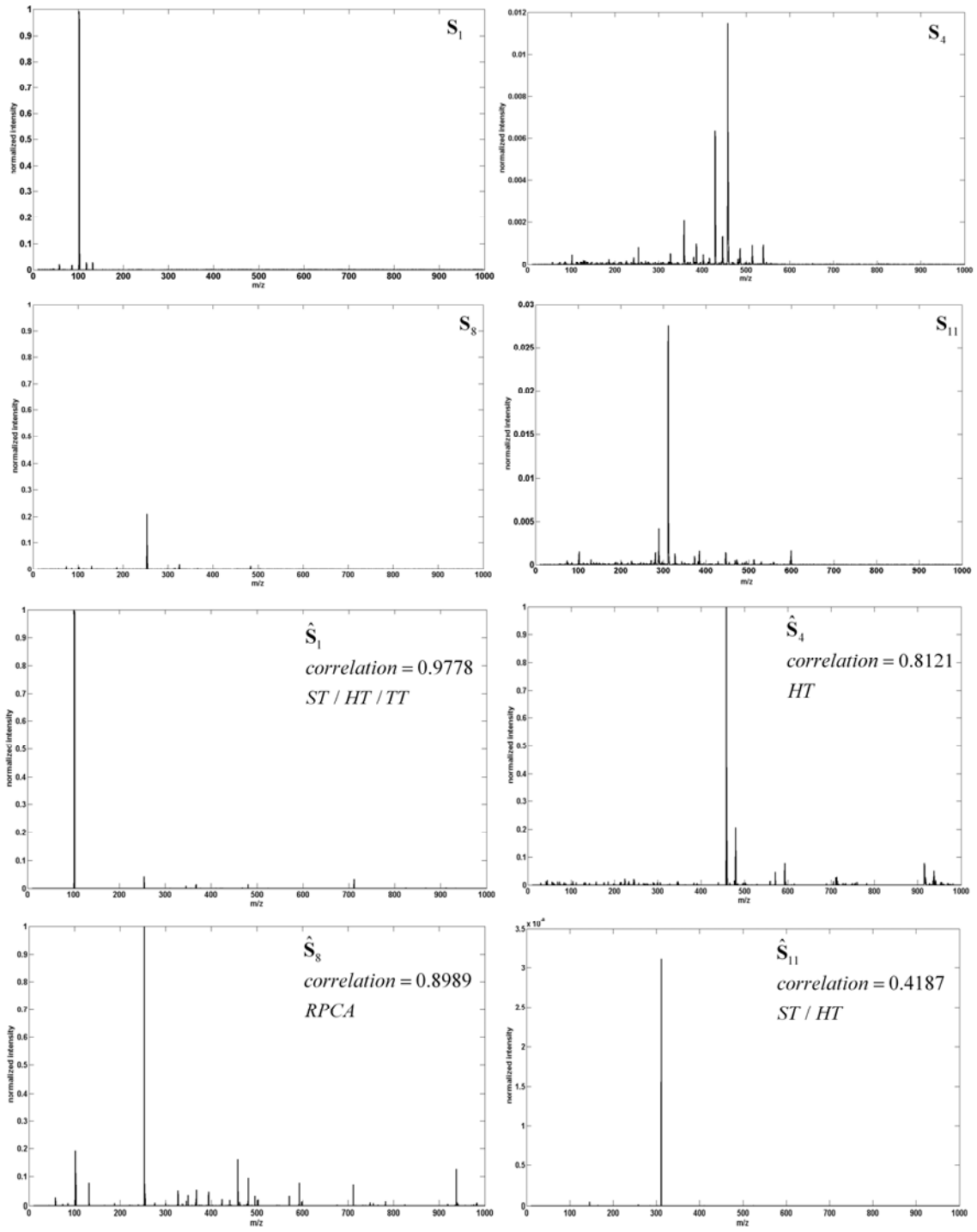


Fig. 3, Kopriva, Jerić, Filipović & Brkljačić

Figure 4. Experimental study. Left: estimated probability that value of the pure component mass spectra is zero, that is estimate of ρ_m, m=1,...,25. Right: estimates of most expected values (means) of exponential distribution obtained by fitting exponential distribution to amplitude histograms. They were estimated for 25 pure components in the range (0, 1] within intervals of the 0.01 width.

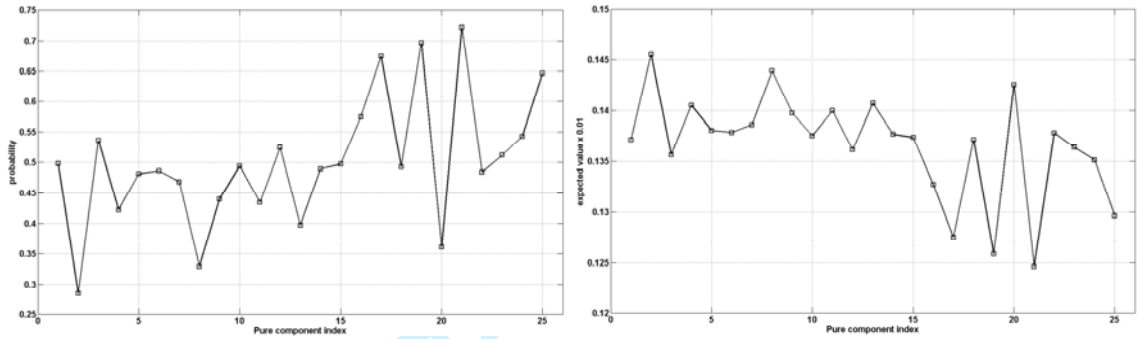


Fig. 4, Kopriva, Jerić, Filipović & Brkljačić

Figure 5. (color online). Experimental study for pure components 1, 4, 8 and 11. Estimated histograms (stars) vs. exponential probability density functions (squares), calculated with the estimates of mean values shown in Figure 4 -right, fitted to amplitude histograms.

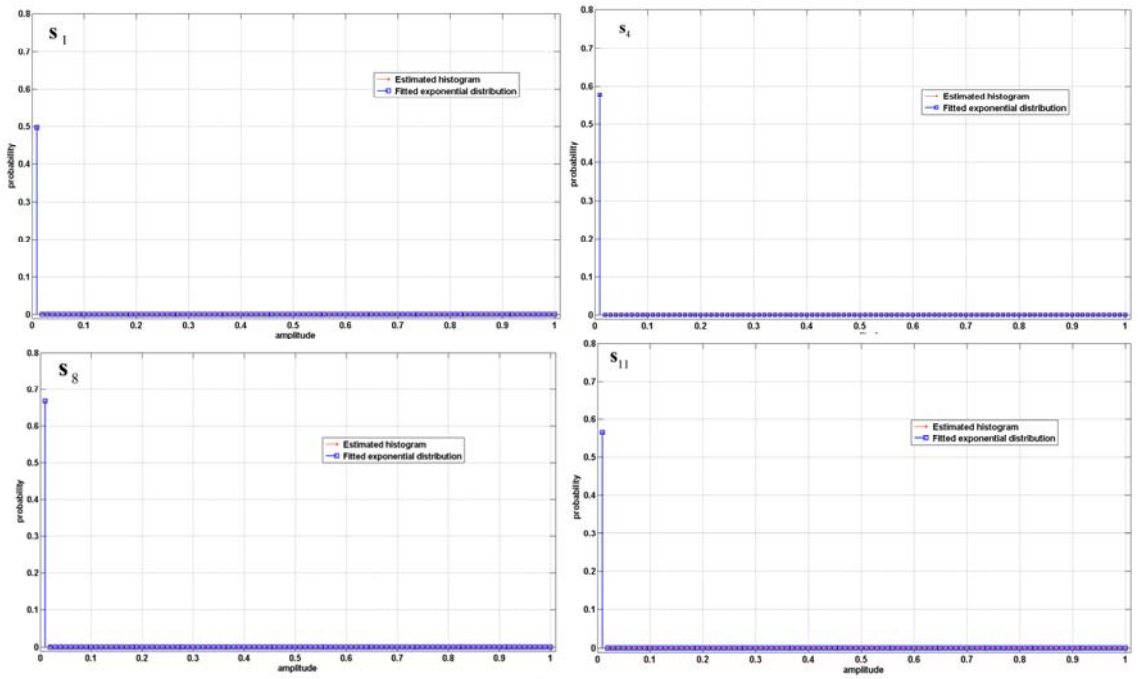


Fig. 5, Kopriva, Jerić, Filipović & Brkljačić

Table Captions

Algorithm 1. The PTs-EKM-NMF (preferably NMU) algorithm.

Required:

$\mathbf{X} \in R_{0+}^{N \times T}$. If A1) is not satisfied perform scaling

$\mathbf{X} \to \mathbf{X} / \arg\max_t \left\{ \left\| \mathbf{x}_t \right\|_1 \right\}_{t=1}^{T}$ or $\mathbf{X} \to \mathbf{X} / \arg\max_{nt} \left\{ \mathbf{X}_{nt} \right\}_{n,t=1}^{N,T}$.

1. Perform RPCA (5) on $\mathbf{X}$ in (2)/(4) with $\lambda \approx 1/\sqrt{T}$. It yields approximation $\mathbf{A}$ in (4).
2. Perform HT on $\mathbf{X}$ in (2)/(4) with $\tau_1 \in [10^{-6}, 10^{-4}]$. It yields approximation $\mathbf{B}$.
3. Perform ST on $\mathbf{X}$ in (2)/(4) with $\tau_2 \in [10^{-6}, 10^{-4}]$. It yields approximation $\mathbf{C}$.
4. Perform TT on $\mathbf{X}$ in (2)/(4) with $\tau_3 \in [10^{-6}, 10^{-4}]$ and α=3.5. It yields approximation $\mathbf{D}$.
5. Perform EKM mappings $\mathbf{A} \to \Psi(\mathbf{A})$, $\mathbf{B} \to \Psi(\mathbf{B})$, $\mathbf{C} \to \Psi(\mathbf{C})$ and $\mathbf{D} \to \Psi(\mathbf{D})$ according to (6). Use Gaussian kernel with σ^2=1.
6. Perform HT, ST and TT respectively of matrices $\Psi(\mathbf{B})$, $\Psi(\mathbf{C})$ and $\Psi(\mathbf{D})$.
7. Perform sparseness constrained factorization, preferably by NMU algorithm, of matrices $\Psi(\mathbf{A})$, $\Psi(\mathbf{B})$, $\Psi(\mathbf{C})$ and $\Psi(\mathbf{D})$ to obtain separated components $\overline{\mathbf{S}}^{\mathbf{A}}$, $\overline{\mathbf{S}}^{\mathbf{B}}$, $\overline{\mathbf{S}}^{\mathbf{C}}$ and $\overline{\mathbf{S}}^{\mathbf{D}}$.
8. Assign to pure components from the library those separated components $\overline{\mathbf{S}}^{\mathbf{A}}$, $\overline{\mathbf{S}}^{\mathbf{B}}$, $\overline{\mathbf{S}}^{\mathbf{C}}$ and $\overline{\mathbf{S}}^{\mathbf{D}}$ with highest normalized correlation coefficient.

Table 1. Comparative performance analysis of NMU, NMF_L0, EKM-NMU, EKM-NMF_L0, PTs-EKM-NMU and PTs-EKM-NMF_L0 algorithms. Probability of zero state was ρ_m=0.8. Four metrics used in comparative performance analysis were: number of associated components with normalized correlation coefficient greater than or equal to 0.6, mean value of correlation coefficient over all associated components, minimal value of correlation coefficient and number of pure components assigned incorrectly (that occurs due to poor separation). All four metrics were calculated with respect to predefined labeling of the pure components stored in the library. Incorrect assignment implies that, based on maximal correlation criterion, two or more pure components are assigned to the same separated component. Mean values and variance are reported and estimated over 10 Monte Carlo runs. The best result in each metric is in bold. The first three metrics are calculated only for correctly assigned components. That is why NMU and NMF_L0 appear to have comparable performance.

	NMU	**NMF_L0**	**EKM-NMU**	**EKM-NMF_L0**	**PTs_EKM-NMU**	**PTs-EKM-NMF_L0**
correlation G.E. 0.6	2.8±0.92	2.3±1.34	3.7±0.48	3.2±0.63	**3.8±0.42**	3.7±0.48
mean correllation	**0.70±0.03**	0.61±0.11	0.69±0.02	0.64±0.03	**0.70±0.03**	0.69±0.04
minimal correlation	**0.53±0.04**	0.42±0.08	0.51±0.03	0.45±0.04	0.52±.04	0.49±0.06
inccorect assignments	3.4±0.70	3.1±0.57	2.4±0.97	2.2±0.63	2.0±0.88	**1.5±1.43**

Table 2. Comparative performance analysis of NMU, NMF_L0, EKM-NMU, PTs-EKM-NMU (D=T=9901) and PTs-EKM-NMU (D=4000) algorithms of 9 experimental nonlinear mixtures mass spectra related to peptide synthesis. Number of pure components equals 25. Four metrics used in comparative performance analysis were: number of associated components with normalized correlation coefficient greater than or equal to 0.6, mean value of correlation coefficient over all associated components, minimal value of correlation coefficient and number of pure components assigned incorrectly (that occurs due to poor separation). The best result in each metric is in bold. The first three metrics are calculated only for correctly assigned components.

	NMU	**NMF_L0**	**EKM-NMU**	**PTs_EKM-NMU** D=T=9901	**PTs-EKM-NMU** D=4000
correlation G.E. 0.6	8	14	16	**18**	**18**
mean correlation	0.342	0.518	0.673	**0.702**	**0.708**
minimal correlation	0.038	0.039	0.267	**0.419**	0.283
inccorect assignments	15	7	**0**	**0**	1
CPU time	1.3s	40 s	78.78h	4×78h	4×13.7h

Empirical Kernel Map Approach to Nonlinear Underdetermined Blind Separation of Sparse Nonnegative Dependent Sources: Pure Components Extraction from Nonlinear Mixtures Mass Spectra

Ivica Kopriva[1*], Ivanka Jerić[2], Marko Filipović[1] and Lidija Brkljačić[2]

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

[1]Division of Laser and Atomic Research and Development

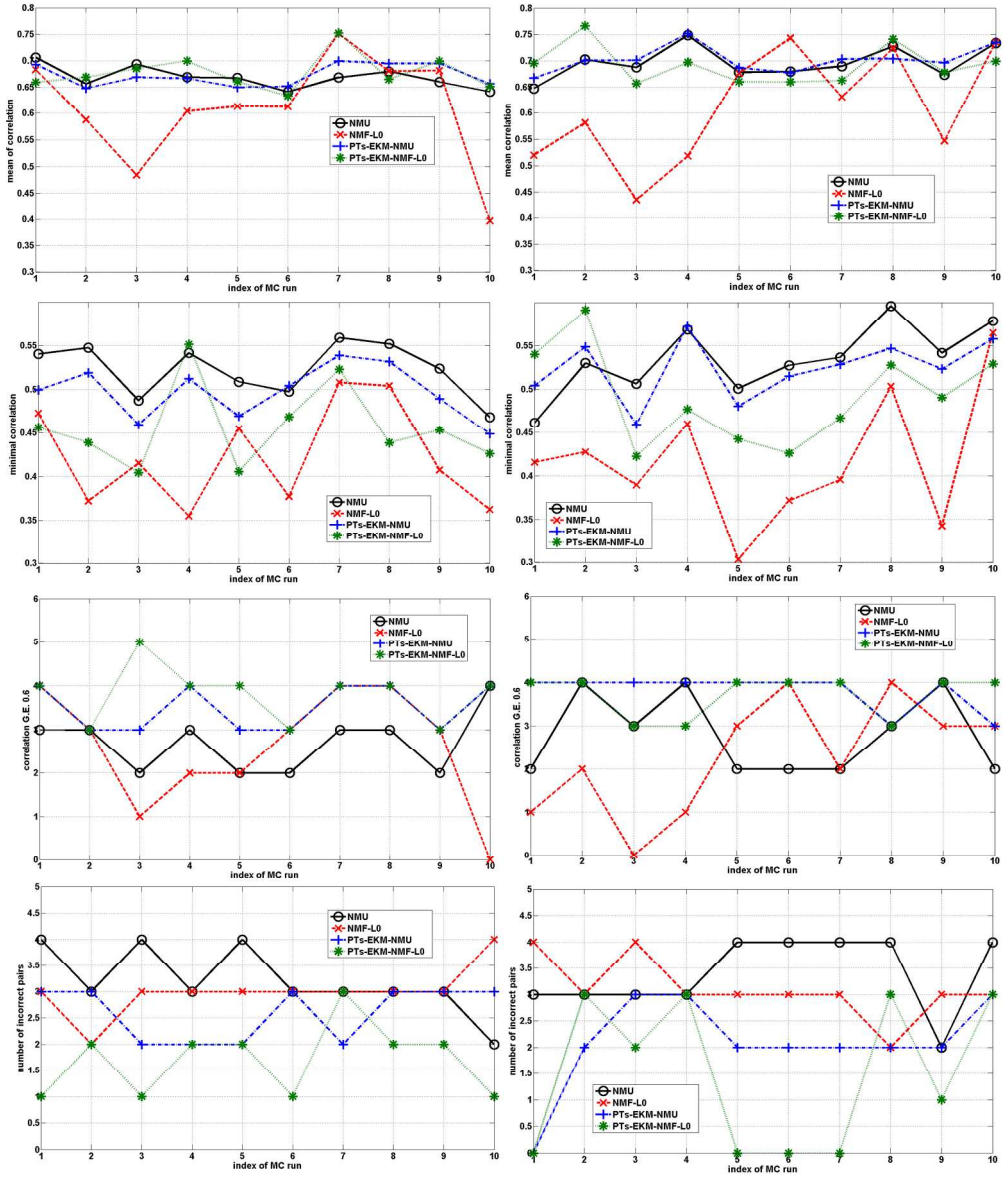
phone: +385-1-4571-286, fax:+385-1-4680-104

e-mail: ikopriva@irb.hr, Marko.Filipovic@irb.hr

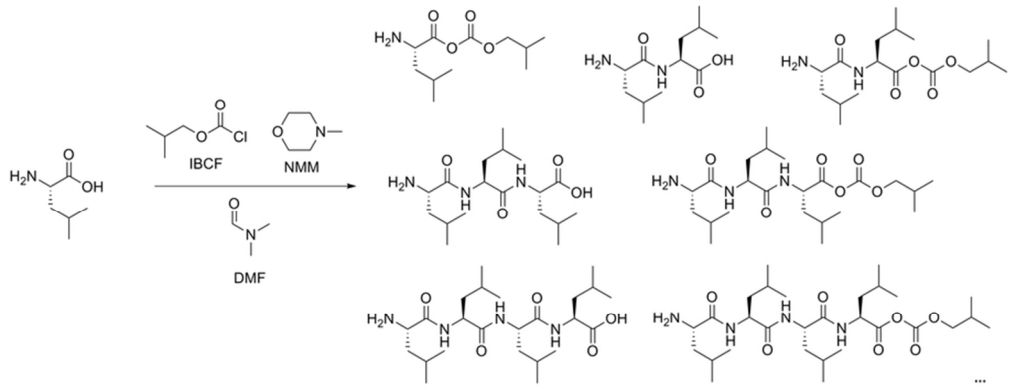
[2]Division of Organic Chemistry and Biochemistry

e-mail: ijeric@irb.hr, Lidija.Brkljacic@irb.hr

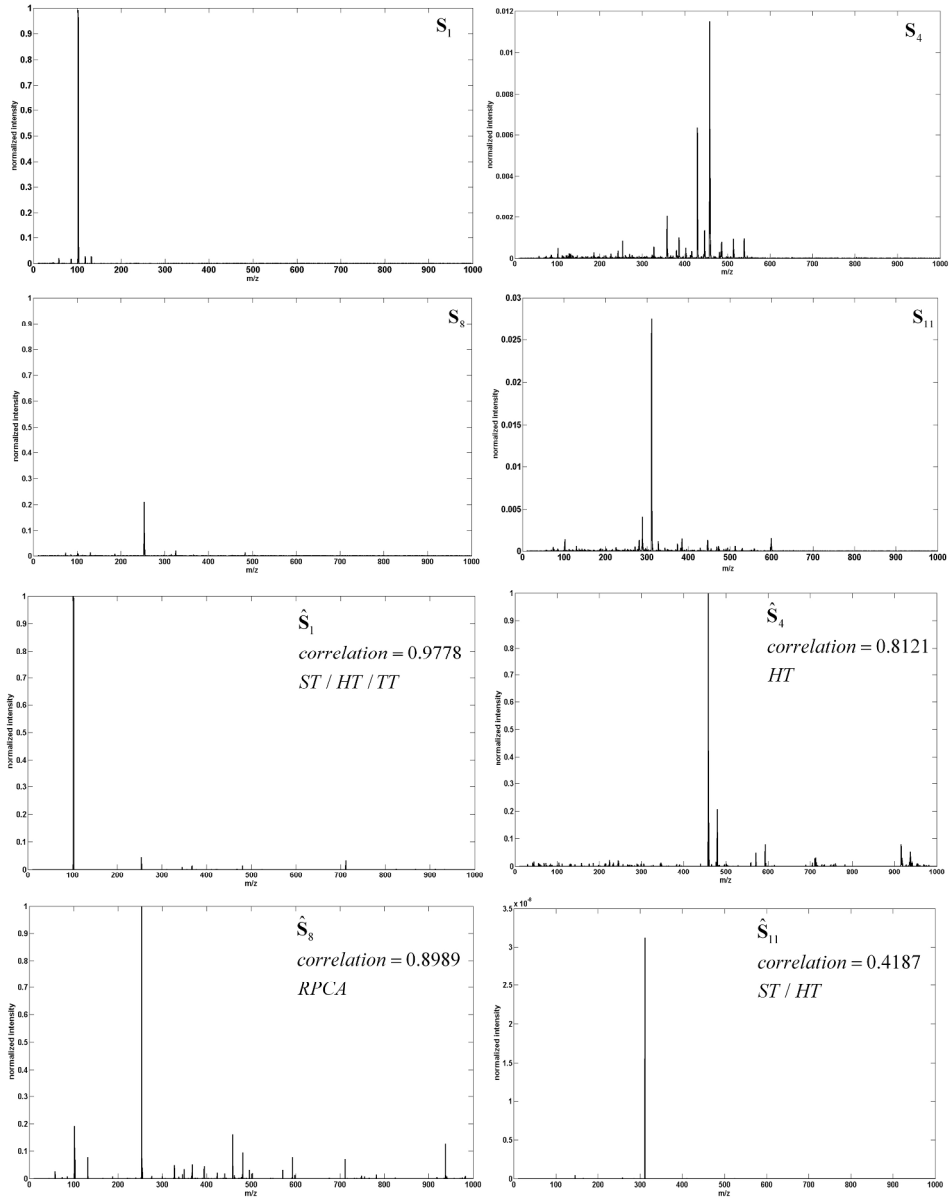
Summary abstract. A method for underdetermined nonlinear blind separation of nonnegative sparse dependent sources is proposed. It combines robust principal component analysis, hard-, soft- and trimmed thresholding to suppress higher order monomials induced by nonlinear mixing with empirical kernel map based nonlinear mapping of preprocessed mixtures data and sparseness constrained nonnegative matrix factorization (NMF) in high-dimensional mapped space. The method is aimed to extract analytes from mass spectra of nonlinear multicomponent mixtures of biological samples.



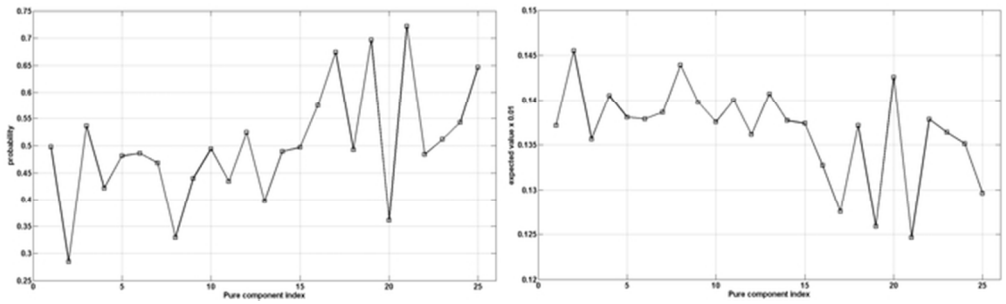
(color online). Numerical study. Normalized correlation coefficient vs. Monte Carlo run index between true and extracted sources by algorithms: NMF_L0 (crosses), NMU (circles) and PTs-EKM-NMU (pluses) and PTs-EKM-NMF_L0 (stars). Mean value (first row), minimal value (second row), number of values greater than or equal to 0.6 (third row), number of incorrect pairs (fourth row). Probability of state zero equal to 0.5 (left column) and 0.8 (right column).
205x241mm (300 x 300 DPI)



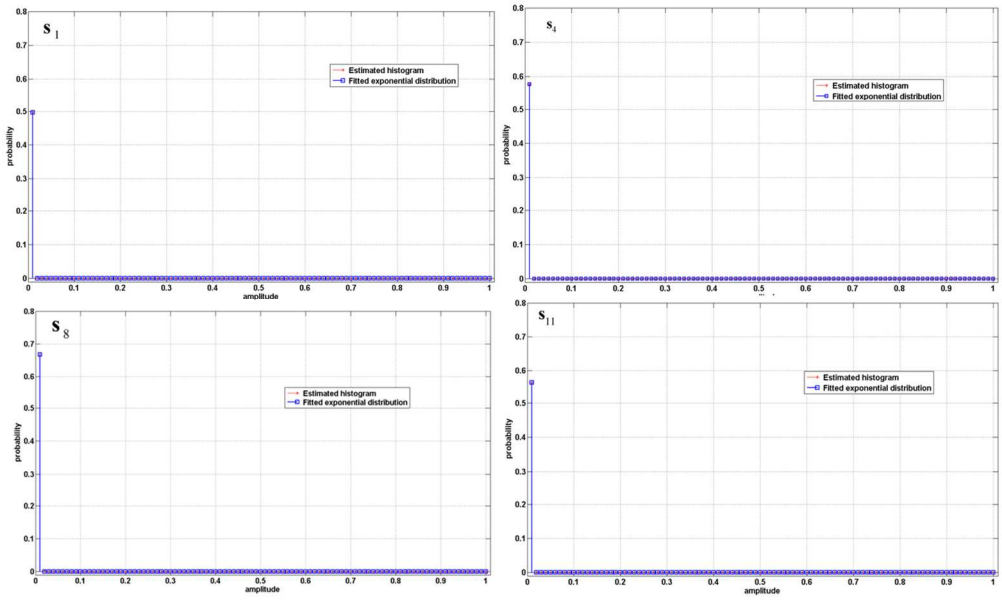
Structures of possible components present in the reaction mixture.
76x32mm (300 x 300 DPI)



Two top rows: mass spectra of pure components s1, s4, s8 and s11. Two bottom rows: estimated mass spectra of pure components s1, s4, s8 and s11 by proposed PTs-EKM-NMU algorithm. Information on value of highest normalized correlation coefficient and associated error reduction method (RPCA, HT, ST and TT) are also displayed.
256x323mm (300 x 300 DPI)



Experimental study. Left: estimated probability that value of the pure component mass spectra is zero, that is estimate of ρm, m=1,...,25. Right: estimates of most expected values (means) of exponential distribution obtained by fitting exponential distribution to amplitude histograms. They were estimated for 25 pure components in the range (0, 1] within intervals of the 0.01 width.
59x17mm (300 x 300 DPI)



(color online). Experimental study for pure components 1, 4, 8 and 11. Estimated histograms (stars) vs. exponential probability density functions (squares), calculated with the estimates of mean values shown in Figure 4 -right, fitted to amplitude histograms.
122x72mm (300 x 300 DPI)