Library-Assisted Nonlinear Blind Separation and Annotation of Pure Components from a Single 1H Nuclear Magnetic Resonance Mixture Spectra

Kopriva, Ivica; Jerić, Ivanka; Popović Hadžija, Marijana; Hadžija, Mirko; Vučić Lovrenčić, Marijana; Brkljačić, Lidija

Source / Izvornik: **Analytica Chimica Acta, 2019, 1080, 55 - 65**

Journal article, Accepted version
Rad u časopisu, Završna verzija rukopisa prihvaćena za objavljivanje (postprint)

https://doi.org/10.1016/j.aca.2019.07.004

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:264:221880

Library-Assisted Nonlinear Blind Separation and Annotation of Pure Components from a Single ^{1}H Nuclear Magnetic Resonance Mixture Spectra

Ivica Kopriva[1], Ivanka Jerić[2], Marijana Popović Hadžija[3], Mirko Hadžija[3], Marijana Vučić Lovrenčić[4] and Lidija Brkljačić[2]*

[1]Division of Electronics

[2]Division of Organic Chemistry and Biochemistry

[3]Division of Molecular Medicine

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

[4]Department of Medical Biochemistry and Laboratory Medicine

University Hospital Merkur, Zajčeva 19, HR-10000 Zagreb, Croatia

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

Abstract

Due to its capability for high-throughput screening ^{1}H nuclear magnetic resonance (NMR) spectroscopy is commonly used for metabolite research. The key problem in ^{1}H NMR spectroscopy of multicomponent mixtures is overlapping of component signals and that is increasing with the number of components, their complexity and structural similarity. It makes metabolic profiling, that is carried out through matching acquired spectra with metabolites from the library, a hard problem. Here, we propose a method for nonlinear blind separation of highly correlated components spectra from a single ^{1}H NMR mixture spectra. The method transforms a single nonlinear mixture into multiple high-dimensional reproducible kernel Hilbert Spaces (mRKHSs). Therein, highly correlated components are separated by sparseness constrained nonnegative matrix factorization in each induced RKHS. Afterwards, metabolites are identified through comparison of separated components with the library comprised of 160 pure components. Thereby, a significant number of them are expected to be related with diabetes type 2. Conceptually similar methodology for nonlinear blind separation of correlated components from two or more mixtures is presented in the Supplementary material. Single-mixture blind source separation is exemplified on: (*i*) annotation of five components spectra separated from one ^{1}H NMR model mixture spectra; (*ii*) annotation of fifty five metabolites separated from one ^{1}H NMR mixture spectra of urine of subjects with and without diabetes type 2. Arguably, it is for the first time a method for blind separation of a large number of components from a single nonlinear mixture has been proposed. Moreover, the proposed method pinpoints urinary creatine, glutamic acid and 5-hydroxyindoleacetic acid as the most prominent metabolites in samples from subjects with diabetes type 2, when compared to healthy controls.

Keywords: nonlinear blind source separation, single mixture, multiple reproducible kernel Hilbert spaces, nonnegative sparse matrix factorization, ^{1}H NMR spectroscopy, metabolic profiling.

1. Introduction

Metabolic profiling aims to identify and quantify small-molecule analytes (a.k.a. metabolites or pure components) present in complex multicomponent mixtures acquired in drug development [1, 2], toxicology studies [3], disease diagnosis [4,5], food, nutrition and environmental sciences [6-8]. Because both techniques provide structural information on chemical classes in a single analysis, metabolic profiling technologies are mainly based on nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry. NMR spectroscopy is a quantitative, non-destructive, robust and reliable technique that provides detailed information of structurally diverse metabolites. Candidates for biomarkers are obtained through matching acquired spectra with those from a library [9], such as the BioMagResBank metabolomics database [10] or Wiley ^{1}H NMR database [11]. However, because many metabolites are structurally similar, their NMR spectra are highly correlated, with many overlapping peaks [12, 8]. That makes metabolic profiling a notoriously difficult problem. That is especially true for ^{1}H NMR spectroscopy [9], which, due to its capability for high-throughput screening [13], is routinely used for metabolite biomarker research. Difficulties related to ambiguous elucidation of the chemical structures are caused by ^{1}H-^{1}H J-couplings that generate broad multiplets [14, 8]. Since many metabolites are not species dependent, that allows translation of some specific biomarkers from preclinical studies directly in clinical studies [15]. Quantitative metabolomic profiling of patients with inflammatory bowel disease characterized 44 serum, 37 plasma, and 71 urine metabolites using

^{1}H NMR spectroscopy [16]. Therefore, for the present study an in-house library comprised of 160 ^{1}H NMR spectra of pure components is built, whereas many of them are known to be present in urine samples of diabetic patients.[1]

The above discussion suggests that computational methods for multivariate analysis of complex metabolomic datasets are of utmost importance for extraction of metabolic information, sample classification and biomarker discovery, [17, 18, 12, 8]. Thus, the main motivation of this paper is: development of a method for blind separation of nonnegative correlated sources from single nonlinear mixture. Its capability to separate components from a single mixture makes it of potential clinical relevance. The method, through the use of explicit (feature map-based) and implicit (kernel-based) nonlinear transforms, maps the original single-mixture blind source separation (BSS) problem into new ones in multiple reproducible kernel Hilbert spaces (RKHSs). In so doing, the method increases significantly the number of (pseudo)mixtures, while the number of new components generated by nonlinear transforms is increased only modestly. That, in combination with the sparse distribution of amplitudes of analytes ^{1}H NMR spectra, enables approximate separation of highly correlated analytes spectra. That is performed by means of sparseness constrained nonnegative matrix factorization (sNMF) in mapping induced RKHSs. Afterwards, analytes are annotated through comparison of the separated components with the pure components from the library. Conceptually similar methodology for blind separation of correlated components from two or more nonlinear mixtures is presented in Supplementary material.

---

[1] The reason for building the in-house library was to solve problems associated with annotations of components recorded using NMR spectrometers with different magnetic field strengths [17]. That is, ^{1}H NMR spectra of the same compound recorded with different spectrometers will have different splitting (J-couplings) and line widths at the resonances. When used similarity measures, such as correlation, are not invariant to these shifts that will affect annotation accuracy.

Essential differences between the proposed method and single-mixture nonlinear BSS method [19] are: (*i*) the method [19] maps single mixture spectra onto one RKHS, while the method proposed herein maps single mixture spectra onto multiples RKHSs. As it is shown in experiments conducted on one model ^{1}H NMR mixture spectra in Section 3.1. as well as two model ^{1}H NMR mixtures spectra in Supplement Section S3.1, usage of multiple RKHSs enables separation and annotation of more pure components than it is possible from one RKHS only; (*ii*) after sparseness constrained separation in each RKHS, library of in-house recorded pure components ^{1}H NMR spectra is used for annotation of separated components. Thus, the proposed method is based upon the implicit assumption that the spectral library is rich enough to contain pure components that correspond to metabolites expected to be present in mixture spectra.

The proposed method is demonstrated based upon two experiments: (*i*) separation and annotation of five correlated components spectra from one model ^{1}H NMR mixture spectra, and (*ii*) separation and annotation of components present in ^{1}H NMR mixtures spectra of urine of subjects with and without diabetes type 2. To the best of our knowledge, this is the first demonstration of a method for the blind separation of a large number of components from single ^{1}H NMR nonlinear mixture spectra. Furthermore, the proposed method highlighted urinary creatine, glutamic acid and 5-hydroxyindoleacetic acid as the most prominent metabolites in samples from subjects with diabetes type 2, when compared to healthy controls.

The rest of the paper is organized as follows. Section 2 presents nonlinear mixture models of multicomponent ^{1}H NMR spectra, analysis of solvability conditions, nonlinear transformations of a single mixture nonlinear BSS problem as well as criteria for evaluation of separation and annotation quality. Section 3 describes experiments and materials used for comparative performance analysis of methods for nonlinear blind separation of components from model and

experimental single ^{1}H NMR spectra. Results related to separation and annotation performance of developed algorithm are presented in section 4. Section 5 discusses metabolic interpretation of the most prominent metabolites in urine of patients with diabetes type 2. Conclusion is presented in section 6.

2. Theory and methods

2.1 Related and background work

The linear mixture model (LMM) is commonly used in NMR spectroscopy [20-25]. It is the model upon which linear instantaneous BSS methods are based, [21-27]. These methods have already been applied successfully for the separation of components from various types of spectroscopic mixtures [21-25]. The majority of these algorithms require that the *unknown* number of analytes is less than or equal to the number of mixtures available. That makes them inapplicable for the analysis of complex mixtures spectra. Sparseness-based approaches to BSS are presently a highly active research area in signal processing. They enable separation of more analytes than mixtures available [23-25]. Sparseness implies that at each chemical shift coordinate only a small number of analytes is present. However, the majority of these algorithms require that each analyte is present alone at certain chemical shift region [23-27]. In case of ^{1}H NMR spectroscopy, due to the complexity of mixtures, it is impossible to satisfy this assumption. Furthermore, sparseness and nonnegativity constrained blind separation of analytes from mixtures of ^{1}H NMR spectra is additionally limited by the assumption that the ^{1}H NMR spectrum is a linear mixture of analytes spectra. That is true for chemical shifts where only one analyte is present. Otherwise, the spectrum of the mixture becomes more nonlinear when the complexity of

the mixture grows, i.e. when the number of overlapping peaks increases [28].[2] Compared with the method proposed herein, existing nonlinear BSS methods assume the availability of multiple nonlinear mixtures [29-42].

Algorithms for single-mixture BSS must first transform the single- to the pseudo multi-mixture BSS problem [43-53]. Subsequently, some existing multivariate algorithms are used to perform BSS from pseudo-mixtures. We use an approximate explicit feature map (aEFM) for observation-wise nonlinear mapping of the recorded mixture ^{1}H NMR spectra into pseudo mixtures spectra. Afterwards, pseudo mixture data are mapped observation-wise in multiple high-dimensional RKHSs using empirical kernel maps (EKM) [54]. The proposed single-mixture nonlinear BSS algorithm differs from the existing single-mixture BSS algorithms in the following aspects: (*i*) algorithms [43-53] address the linear BSS problem, while the proposed method addresses the nonlinear BSS problem, and (*ii*) the hard constraints imposed on the source signals by single-mixture BSS algorithms [43-53] do not apply to the pure component ^{1}H NMR amplitude spectra that are of interest in this study. This statement is supported through the following analysis. The method [43] assumes that the source signals have disjoint support. The method [44] partitions single-channel time series to yield a pseudo multichannel mixture, to which an independent component analysis algorithm was applied to extract the sources. The disjoint support assumption does not hold for overlapping pure components ^{1}H NMR amplitude spectra. The algorithm [45] uses empirical mode decomposition to decompose the single-channel mixture into intrinsic mode functions that represent the pseudo multichannel mixture. For

[2] The reason for formulating single mixture BSS problem in amplitude spectra domain, where mixture is nonlinear, instead of time domain, where mixture is linear, is much higher degree of overlap between components. That occurs because time domain NMR signals are not sparse. Please see [24] for the more in depth discussion of this issue. As discussed in Section 2.3, degree of the components overlap at independent variable (chemical shift in the present paper) is one of the key factors that enables (or disables) solvability of the related underdetermined BSS problem.

separation by independent component analysis algorithms, sources of interest are required to be intrinsic mode functions, what is not true for the pure components ^{1}H NMR amplitude spectra. In [46], the wavelet transform is used to generate a pseudo multichannel mixture from a single-channel version. In this way, the mother wavelet has to be non-orthogonal and has to match the shapes to the sources of interest. The Morlet wavelet was used in [46], but other non-orthogonal wavelets can be used as well. Thus, this method is applicable to separation of the specific source signals, such as vibration signals [46, 47]. Many of the single-channel BSS algorithms are derived to separate acoustic signals by factorizing a nonnegative spectrogram [48-53].

2.2 Nonlinear mixture model of multicomponent ^{1}H NMR spectra

NMR signals are intrinsically time domain harmonic signals with amplitude decaying exponentially with some time constant. Thus, LMM applies to either time domain or Fourier transform domain representations. The model in the Fourier (chemical shift) domain in the absence of additive noise reads out as:

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

where $\mathbf{X} \in \mathbb{C}^{N \times T} =: \left\{ \mathbf{X}_n = FT(\mathbf{x}_n) \in \mathbb{C}^{1 \times T} \right\}_{n=1}^{N}$ represents mixture matrix such that each row of $\mathbf{X}$ contains one complex ^{1}H NMR mixture signal and $\mathbb{C}$ stands for the set of complex numbers. $\mathbf{X}_n$ is obtained as the Fourier transform (*FT*) of the time domain equivalent $\mathbf{x}_n$, comprised of

complex values at *T* chemical shift instants, and symbol " =:" means "by definition".

$\mathbf{A} \in \mathbb{R}_{0+}^{N \times M} =: \left\{ \mathbf{a}_m \in \mathbb{R}_{0+}^{N \times 1} \right\}_{m=1}^{M}$ represents a mixture matrix, whereas each column represents a concentration profile of one of the *M* analytes across the *N* mixtures, and $\mathbb{R}_{0+}$ denotes a set of nonnegative real numbers. $\mathbf{S} \in \mathbb{C}^{M \times T} =: \left\{ \mathbf{S}_m = FT(\mathbf{s}_m) \in \mathbb{C}^{1 \times T} \right\}_{m=1}^{M}$ is a matrix with the rows representing ^{1}H NMR complex signals of the analytes present in the mixture signals **X**. However, as shown in [28], amplitudes of the NMR mixture spectra, $|\mathbf{X}| \in \mathbb{R}_{0+}^{N \times T} =: \left\{ |\mathbf{X}_n| \in \mathbb{R}_{0+}^{1 \times T} \right\}_{n=1}^{N}$, are nonlinear mixtures of the amplitudes of the components NMR spectra, $|\mathbf{S}| \in \mathbb{R}_{0+}^{M \times T} =: \left\{ |\mathbf{S}_m| \in \mathbb{R}_{0+}^{m \times T} \right\}_{m=1}^{M}$. Thus, instead of LMM (1) we assume nonlinear mixture model (NMM) for ^{1}H NMR amplitude spectra:

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

where $\mathbf{f} : \mathbb{R}_{0+}^{M} \to \mathbb{R}_{0+}^{N}$ stands for an unknown nonlinear mapping $\mathbf{f}\left(|\mathbf{S}|\right) := \left[f_1\left(|\mathbf{S}|\right) \ldots f_N\left(|\mathbf{S}|\right) \right]^T$ acting observation-wise. We also assume $\left\{ \left\| |\mathbf{S}_t| \in \mathbb{R}_{0+}^{M \times 1} \right\|_0 \le K \right\}_{t=1}^{T}$. Here $\left\| \mathbf{S}_t \right\|_0$ is indicator function that counts number of non-zero entries of $|\mathbf{S}_t|$ and *K* denotes maximal number of sources that can be present at any observation coordinate *t*. The nonlinear BSS problem (2) implies that the amplitude spectra of pure components $|\mathbf{S}|$ ought to be inferred from the mixture amplitude ^{1}H

NMR spectra $|\mathbf{X}|$ only. Since the nonlinear BSS method that will be developed herein is aimed to be used for metabolic profiling we assume:

A1) $N \geq 1$

A2) $M > N$

Thus, the nonlinear BSS problem (2) is underdetermined. In particular, we are interested in clinically most relevant case of a single mixture, i.e. $N=1$. Since the peaks in amplitude spectra are not statistically related, the pure components are treated as independent and identically distributed (i.i.d.) random variables. Hence, we propose a method for blind separation of mutually dependent but individually i.i.d. nonnegative pure components from one nonlinear mixture. To the best of our knowledge, existing methods cannot address the nonlinear BSS problem under the assumed scenario.

2.3 Solvability of underdetermined nonlinear system and sparse probabilistic model of ^{1}H NMR components spectra

We further assume the following:

A3) $0 \leq |\mathbf{S}_{mt}| < 1 \;\; \forall m = 1,...,M \;\; t = 1,...,T$,

A4) $|\mathbf{S}_{mt}|$ is i.i.d. random variable that obeys truncated exponential distribution on (0, 1] interval and discrete distribution at zero, see Eq. (3),

A5) Components of the vector-valued function $\mathbf{f}(|\mathbf{S}|) := [\, f_1(|\mathbf{S}|) \ldots f_N(|\mathbf{S}|)]^T$ are differentiable up to second-order.

Assumptions A3 to A5 are shown in [58] to be relevant for separation of pure components from nonlinear mixtures of mass spectra. They also hold for separation of pure components from amplitude ^{1}H NMR spectra, whereas A4 is confirmed below. To be useful, the solution of any BSS problem is expected to be unique up to the scaling and permutation indeterminacy. That is referred to in BSS as essential uniqueness [55]. However, even for the linear underdetermined BSS problem hard (sparseness) constraints ought to be imposed on pure components [56, 57, 58, 19] to obtain an essentially unique solution. The quality of separation depends heavily on the degree of sparseness, i.e. the value of *K*. To make the nonlinear underdetermined BSS problem tractable we assume, as in [57], that amplitudes of the source signals comply with the sparse probabilistic model [56]:

$$p(|\mathbf{S}_{mt}|) = \rho_m \delta(|\mathbf{S}_{mt}|) + (1-\rho_m)\delta^*(|\mathbf{S}_{mt}|)\, g(|\mathbf{S}_{mt}|) \;\; \forall m = 1,\ldots,M \text{ and } \forall t = 1,\ldots T \qquad (3)$$

where $\delta(|\mathbf{S}_{mt}|)$ is delta function and $\delta^*(|\mathbf{S}_{mt}|) = 1 - \delta(|\mathbf{S}_{mt}|)$ is its complementary function. $\rho_m(|\mathbf{S}_{mt}|) = P(|\mathbf{S}_{mt}| = 0)$. Hence, $P(|\mathbf{S}_{mt}| > 0 = 1 - \rho_m)$. Sparse probabilistic model (3) is justified through the following analysis. We assume that the ^{1}H NMR components spectra comply with

the sparse probabilistic model represented by the truncated exponential distribution on interval (0, 1]:

$$g\left(\left|\mathbf{S}_{mt}\right|\right)=\left(1/\mu_m\right)\exp\left(-\left|\mathbf{S}_{mt}\right|/\mu_m\right)\ \forall m=1,...,M \tag{4}$$

First, the library of 160 pure components amplitude spectra are, according to A3, scaled to [0, 1] interval. We performed a maximum likelihood based fitting, using MATLAB function `fitdist`, of the exponential distribution (4) to the experimental analytes ^{1}H NMR amplitude spectra and obtained $\hat{\mu}_m \in \left[0.001206, 0.002338\right]$, $m=1,...,160$. This result implies that, due to the small value of μ_m, there is virtually no difference on (0, 1] interval between probability density function (pdf) of the truncated exponential distribution and pdf of the exponential distribution. For the exponential prior (4) with given μ_m and given probability $p(0<\left|\mathbf{S}_{mt}\right|\leq s)$ the value of s is obtained as: $s\approx-\mu_m\ln(1-p)$. For p=0.99 and μ_m=0.002338 it follows s=0.0108. We also estimated ρ_m in (3). For each pure component spectra we counted chemical shifts where the amplitude value was at least a hundred times smaller than the maximal value in the library. The obtained number was divided by the length of the amplitude spectra. We obtained the following result: $\hat{\rho}_m \in \left[0.9581, 0.9954\right]$, $m=1,...,160$. Thus, in probability the ^{1}H NMR components spectra are sparse and will have very small values. Hence, even though ^{1}H NMR spectra are not inherently sparse, as opposed to mass spectra, they comply with the sparse probabilistic model (3). That justifies cancellation of the higher order terms (HOT) in the nonlinear transform that

follows. Under the sparse probabilistic prior (3)/(4) the nonlinear mixture model (2) simplifies to [58]:

$$\left|\mathbf{X}\right| = \mathbf{J}\left|\mathbf{S}\right| + \tfrac{1}{2}\mathbf{H}_{(1)}\begin{bmatrix} \left|\mathbf{S}_1\right|^2 \\ \cdots \\ \left|\mathbf{S}_M\right|^2 \\ \cdots \\ \left\{\left|\mathbf{S}_i\right|\left|\mathbf{S}_j\right|\right\}_{i,j=1}^{M} \end{bmatrix} + HOT = \mathbf{B}\begin{bmatrix} \left|\mathbf{S}\right| \\ \left|\mathbf{S}_1\right|^2 \\ \cdots \\ \left|\mathbf{S}_M\right|^2 \\ \cdots \\ \left\{\left|\mathbf{S}_i\right|\left|\mathbf{S}_j\right|\right\}_{i,j=1}^{M} \end{bmatrix} + HOT \tag{5}$$

where $\mathbf{J}$ stands for the Jacobian matrix, $\mathbf{H}_{(1)}$ stands for the mode-1 unfolded third-order Hessian tensor and $\mathbf{B} = \left[\mathbf{J}\ \tfrac{1}{2}\mathbf{H}_{(1)}\right]$ stands for overall mixing matrix. Since the original nonlinear problem (2) is underdetermined the equivalent linear problem (5) is even more underdetermined. That is, the problem (5) is comprised of the same number of mixtures, N, but has the $P=2M + M(M-1)/2$ dependent sources. When the degree of sources overlap in (2) is K, the degree of overlap of new sources in (5) is $Q \approx 2K + K(K-1)/2$. The uniqueness of the solution of (5) depends on the triplet (N, P, Q). For deterministic mixing matrix $\mathbf{B}$, the necessary condition for uniqueness is $N=O(Q^2)$ [59]. Thus, it becomes virtually impossible to obtain an essentially unique solution of the underdetermined nonlinear BSS problem (5) with overlapped sources. Separation quality can however, be increased through nonlinear mapping of mixture data:

$$\left\{\left|\mathbf{X}_t\right| \in \mathbb{R}_{0+}^{N\times 1} \rightarrow \phi\left(\left|\mathbf{X}_t\right|\right) \in \mathbb{R}_{0+}^{\bar{N}\times 1}\right\}_{t=1}^{T} \tag{6}$$

where EFM $\phi\left(\left|\mathbf{X}_t\right|\right)$ maps data into, in principle, infinite dimensional space. To make calculations in mapped space computationally tractable, $\phi\left(\left|\mathbf{X}\right|\right) \coloneqq \left\{\phi\left(\left|\mathbf{X}_t\right|\right)\right\}_{t=1}^{T}$ needs to be projected to a low-dimensional subspace of induced space spanned by $\phi\left(\mathbf{V}\right) \coloneqq \left\{\phi\left(\mathbf{v}_d\right)\right\}_{d=1}^{D}$. The projection known as EKM, see definition 2.15 in [54], maps data from the input space onto RKHS:

$$\Psi\left(\left|\mathbf{X}\right|, \mathbf{V}\right) = \phi\left(\mathbf{V}\right)^T \phi\left(\left|\mathbf{X}\right|\right) = \mathrm{K}\left(\left|\mathbf{X}\right|, \mathbf{V}\right) \tag{7}$$

where $\mathrm{K}\left(\left|\mathbf{X}\right|, \mathbf{V}\right) \in \mathbb{R}_{0+}^{D\times T}$ denotes Gram or kernel matrix with the elements $\left\{\kappa\left(\left|\mathbf{X}_t\right|, \mathbf{v}_d\right) = \phi\left(\mathbf{v}_d\right)^T \phi\left(\left|\mathbf{X}_t\right|\right)\right\}_{d,t=1}^{D,T}$. It is shown in [58] that under sparse probabilistic prior (3)/(4), Eq.(7) becomes:

$$\Psi\left(\left|\mathbf{X}\right|, \mathbf{V}\right) = \mathbf{G}\begin{bmatrix} \mathbf{0}_{1\times T} \\ \left|\mathbf{S}\right| \\ \left\{\left|\mathbf{S}_i\right|\left|\mathbf{S}_j\right|\right\}_{i,j=1}^{M} \end{bmatrix} + HOT \tag{8}$$

where **G** denotes a nonnegative mixing matrix of appropriate dimensions and $\mathbf{0}_{1\times T}$ stands for row vector of zeros. The uniqueness condition for system (8) becomes: $D=O(Q^2)$, [59]. For $D>>N$ the uniqueness condition can be fulfilled with greater probability than the uniqueness condition for

system (5): $N=O(Q^2)$. Thus, the role of the EKM-based mapping is to "increase the number of ^{1}H NMR mixture spectra". However, due to the Lorentzian shape of the pure NMR peak, K will be greater than it is the case with mass spectra. Therefore, the role of the EKM-based mapping is even more important than it is for the case with the mass spectra mixtures [19]. In particular, that is the main reason why nonlinear blind separation of pure components from single ^{1}H NMR mixture spectra has to be performed in multiple RKHSs, as opposed to [19] where only one RKHS was used.

2.4 Nonlinear transformation of the original single mixture nonlinear BSS problem

Algorithms for single-mixture BSS first have to transform single- to pseudo multi-mixture BSS problem [43-53]. For the single-mixture case, EFM (6) reduces to [19]:

$$\left\{ \left| \mathbf{X}_t \right| \in \mathbb{R}_{0+} \to \phi_i \left(\left| \mathbf{X}_t \right| \right) \in \mathbb{R}_{0+}^{\bar{N} \times 1} \right\}_{t=1}^{T} \quad \forall i = 1,..,I . \tag{9}$$

It is seen from (9) that single mixture data $\left| \mathbf{X}_t \right|$ is mapped onto I >1 spaces. To obtain pseudo multi-mixture data, EFM $\phi_i \left(\left| \mathbf{X}_t \right| \right)$ has to satisfy two conditions: (*i*) it has to be of finite order and (*ii*) it has to have analytic form. Hence, we provide in (10) an analytic expressions for EFM obtained by factorization of the Gaussian kernel that, with the slight abuse of notation, is given

with: $\kappa\left(\left|\mathbf{X}\left(\delta_t\right)\right|,\mathbf{v}\right)=\exp\left(-\left\|\left|\mathbf{X}\left(\delta_t\right)\right|-\mathbf{v}\right\|_2\Big/\sigma^2\right)$, where δ_t denotes chemical shift. The aEFM is obtained as:

$$\phi_i^d\left(\left|\mathbf{X}_t\right|\right)=\mathrm{e}^{-\frac{\left\|\mathbf{X}_t\right\|_2^2}{\sigma_i^2}}\left[1\quad\frac{\sqrt{2}}{\sigma_i}\left|\mathbf{X}_t\right|\quad\frac{\sqrt{2}}{\sigma_i^2}\left|\mathbf{X}_t\right|^2\quad\ldots\quad\sqrt{\frac{2^d}{d!}}\frac{1}{\sigma_i^d}\left|\mathbf{X}_t\right|^d\right]^T\quad\forall t=1,\ldots,T\quad\text{and}\quad\forall i=1,\ldots,I. \tag{10}$$

To simplify notation we have substituted in (10) the chemical shift δ_t with the index *t*. For $0\le d<\infty$ (10) represents the aEFM of order *d*. Hence, for mapping associated with RKHS induced with the Gaussian kernel instead of (9) we use:

$$\left\{\left|\mathbf{X}_t\right|\in\mathbb{R}_{0+}\rightarrow\phi_i^d\left(\left|\mathbf{X}_t\right|\right)\in\mathbb{R}_{0+}^{(d+1)\times 1}\right\}_{t=1}^{T}\quad\forall i=1,\ldots,I. \tag{11}$$

The best results reported in the experimental section 4.1 for single mixture case, as well as in Tables S-68 to S-70 in Supplementary material for two-mixtures case, were obtained for order of aEFM *d*=2. Thus, single-mixture $\left|\mathbf{X}\right|\in\mathbb{R}_{0+}^{1\times T}$ is mapped into pseudo multi-mixture according to:

$$\left|\mathbf{X}\right|\in\mathbb{R}_{0+}^{1\times T}\rightarrow\phi_i^2\left(\left|\mathbf{X}\right|\right)\in\mathbb{R}_{0+}^{3\times T}\quad\forall i=1,\ldots,I. \tag{12}$$

Before mapping $\phi_i^2\left(\left|\mathbf{X}\right|\right)$ into RKHS we need to introduce the EKM. To increase the probability of separation of highly correlated analytes from ^{1}H NMR spectra, the number of pseudo-mixture spectra in (12) has to be increased to *D>>3*. For this purpose the nonlinear mapping, known as EKM, $\Psi:\mathbb{R}_{0+}^3\to\mathbb{R}_{0+}^D$ was proposed in (7)/(8). The mapping is performed chemical shift-wise:

$$\left\{\phi_i^2\left(\left|\mathbf{X}\left(\delta_t\right)\right|\right)\in\mathbb{R}_{0+}^3\mapsto\Psi_{\kappa_i}\left(\left|\mathbf{X}\left(\delta_t\right),\mathbf{V}_i\right|\right)\in\mathbb{R}_{0+}^D\right\}_{t=1}^T\quad\forall i=1,...,I \tag{13}$$

where:

$$\Psi_{\kappa_i}\left(\left|\mathbf{X}\right|,\mathbf{V}_i\right)=\begin{bmatrix}\kappa_i\left(\left|\mathbf{X}\left(\delta_1\right)\right|,\mathbf{v}_{1(i)}\right) & ... & \kappa_i\left(\left|\mathbf{X}\left(\delta_T\right)\right|,\mathbf{v}_{1(i)}\right)\\ ... & ... & ...\\ \kappa_i\left(\left|\mathbf{X}\left(\delta_1\right)\right|,\mathbf{v}_{D(i)}\right) & ... & \kappa_i\left(\left|\mathbf{X}\left(\delta_T\right)\right|,\mathbf{v}_{D(i)}\right)\end{bmatrix}\quad\forall i=1,...,I. \tag{14}$$

Equations (9) to (14) indicate dependence of mappings $\phi_i^2\left(\left|\mathbf{X}\right|\right)$ and $\Psi_{\kappa_i}\left(\left|\mathbf{X}\right|,\mathbf{V}_i\right)$ on corresponding parameters of the chosen kernel κ_i. It is, in principle, unclear how to select both the kernel κ_i as well as the optimal value of the kernel parameters. In particular, for the Gaussian kernel that is used in the present study, it is known that the value of its variance has to be adopted to signal-to-noise-ratio (SNR) [60]. If the SNR is low, a large value of σ^2 ought to be selected

and vice versa. It is, however, hard to know the SNR value in practice. That is why, as opposed to [19], we propose herein mapping of the original ^{1}H NMR mixture spectra onto multiple RKHS (13). The role of the basis $\mathbf{V}_i \coloneqq \left\{\mathbf{v}_{d(i)} \in \mathbb{R}_{0+}^{N\times 1}\right\}_{d=1}^{D}$ is to approximately span the induced space:

$$span\left\{\phi\left(\phi_i^2\left(\mathbf{v}_{d(i)}\right)\right)\right\}_{d=1}^{D} \approx span\left\{\phi\left(\phi_i^2\left|\mathbf{X}\left(\delta_t\right)\right|\right)\right\}_{t=1}^{T} \tag{15}$$

Eq. (15) holds under the assumption:

$$span\left\{\phi_i^2\left(\mathbf{v}_{d(i)}\right)\right\}_{d=1}^{D} \approx span\left\{\phi_i^2\left(\left|\mathbf{X}\left(\delta_t\right)\right|\right)\right\}_{t=1}^{T} \tag{16}$$

The basis $\mathbf{V}_i$ is estimated from $\phi_i^2\left(|\mathbf{X}|\right)$ by the *k*-m*eans* clustering algorithm. In the experimental sections 4.1 and 4.2, we have used the *k-means* clustering algorithm, implemented with the MATLAB function *kmeans*, to cluster $\left\{\phi_i^2\left(\left|\mathbf{X}\left(\delta_t\right)\right|\right)\right\}_{t=1}^{T}$ into pre-specified number of *D* cluster centers $\left\{\mathbf{v}_{d(i)} \in \mathbb{R}_{0+}^{N\times 1}\right\}_{d=1}^{D}$ that represent the basis matrix $\mathbf{V}_i$. The sNMF algorithm is now ready to be applied to $\Psi_{\kappa_i}\left(|\mathbf{X}|, \mathbf{V}_i\right)$ in (14) to separate analytes ^{1}H NMR spectra through:

$$\left\{\left|\hat{\mathbf{S}}_{(m,i)}\right|\right\}_{m=1}^{D} = sNMF\left(\Psi_{\kappa_i}\left(\phi_i^2\left(|\mathbf{X}|\right), \mathbf{V}_i\right)\right) \quad \forall i = 1,...,I\,. \tag{17}$$

By executing sNMF on data mapped into multiple RKHSs and by combining the obtained results we can increase the probability of separating correlated ^{1}H NMR component spectra from one mixture spectra only. Regarding SNMF algorithm, we have used the nonnegative matrix underapproximation (NMU) algorithm [61] with the MATLAB code freely available [62]. The main reason for preferring the NMU algorithm over other sNMF algorithms is that there are no sparseness constraint regularization constants that need to be tuned. It is important to notice that in (17) the initial number of components to be extracted was set to *D* even though the expected number of components is smaller. That comes as a benefit of using the EKM-based mapping which alleviates the difficult problem related to *a priori* setting of the number of components to be separated. That, in general, is a hard problem in computer science with, so far, no algorithm agreed to work well on data of diverse origins. Components separated in (17) are compared with the pure components spectra stored in the library, $\left\{\left|\mathbf{S}_m\right|\right\}_{m=1}^{J}$, using normalized correlation coefficient as a similarity measure. Separated components are paired with the components from the in-house library comprised of J=160 ^{1}H NMR spectra of pure components $j^* \in \{1, ..., J\}$ according to :

$$j^* = \arg\max_{j=1,...,J} \frac{\left\langle\left|\hat{\mathbf{S}}_{(m,i)}\right|, \left|\mathbf{S}_j\right|\right\rangle}{\left\|\hat{\mathbf{S}}_{(m,i)}\right\|\left\|\mathbf{S}_j\right\|} \;\; \forall m = 1,...,D \;\; \forall i = 1,...,I\;. \tag{18}$$

Afterwards, normalized correlation coefficients:

$$c_{(m,i,j^*)} = \frac{\left\langle \left|\hat{\mathbf{S}}_{(m,i)}\right|, \left|\mathbf{S}_{j^*}\right| \right\rangle}{\left\|\hat{\mathbf{S}}_{(m,i)}\right\| \left\|\mathbf{S}_{j^*}\right\|} \quad \forall m = 1,...,D \quad \forall i = 1,...,I \tag{19}$$

are ranked in descending order. Finally, the list is refined by removing from it all the components $\left|\hat{\mathbf{S}}_{(m,i)}\right|$ paired with the same pure component $\left|\mathbf{S}_{j^*}\right|$ with the exception of the one with the largest correlation coefficient. Thus, we obtain the final list of separated components annotated to the only one most similar pure component from the library. The number of pure components J stored in the library can in general be large, for example, $J \approx 10^5$ for the Wiley ^{1}H NMR spectral library [11]. Herein, we used the in-house built library comprised of $J=160$ ^{1}H NMR spectra of pure components. The algorithm aEFM-EKM-mRKHS is summarized in Algorithm 1.

Algorithm 1. Summary of the nonlinear single-mixture BSS algorithm: aEFM-EKM-mRKHS.

Required:

$\mathbf{x} \in \mathbb{C}_{0+}^{1\times T}$, D, $\left\{\sigma_1^2,...,\sigma_I^2\right\}$ for the Gaussian kernel.

1. Execute the Fast Fourier transform on $\mathbf{x}$: $\mathbf{x} \mapsto \mathbf{X} = FFT(\mathbf{x})$. Scale $\left|\mathbf{X}\right|$ to satisfy A3.

2. Generate mappings $\left\{\phi_i^2\left(\left|\mathbf{X}\right|\right)\right\}_{i=1}^{I}$ according to (12).

3. Use the *k-means* algorithm to estimate bases $\left\{\mathbf{V}_i\right\}_{i=1}^{I}$ from

$\left\{\phi_i^2\left(|\mathbf{X}|\right)\right\}_{i=1}^{I}$ in (12).

4. Use the NMU algorithm to separate ^{1}H NMR components spectra according to (17).

5. Annotate separated analytes spectra with the pure components spectra from the library according to (18) and the succeeding paragraph.

2.5 Criteria for evaluation of the separation and annotation quality

After the separated components are annotated and ranked, the most desirable outcome is that the top *M* components on the ranking list correspond with the *M* pure components present in the mixture spectra. However, given that the large number of correlated pure components ^{1}H NMR spectra ought to be separated from the only one nonlinear mixture spectra, it is certain that the quality of separation will be limited. Consequently, some number of separated components will be annotated incorrectly. Thus, we propose four criteria to validate separation and annotation results achieved by the proposed nonlinear single mixture BSS method.

Criterion 1 (C1**)** counts the number of correctly annotated components out of *M* separated components ranked first on the list:

$$\mathrm{C1} = \#I_c \tag{20}$$

where $\#I_c$ denotes cardinality of the index set I_c comprised of correctly annotated components among first *M* ranked separated components. If the separation is perfect, all first *M* separated

components would be annotated correctly. Hence, it applies for the cardinality of the set I_c: $\#I_c \leq M$.

Criterion 2 (C2**)** stands for the penalized mean normalized correlation between the first M separated components and the pure components they are correctly annotated with:

$$\mathrm{C2} = \left(\sum_{i \in I_c} c_i \left(\left| \hat{\mathbf{S}}_i \right|, \left| \mathbf{S}_{j^*} \right| \right) \right) \Big/ M \tag{21}$$

where I_c is defined previously. When all first M ranked separated components are annotated correctly we have $\#I_c = M$ and the penalized mean correlation, C2, equals the mean correlation.

Criterion 3 (C3) stands for the penalized mean normalized correlation between all separated components and the pure components they are correctly annotated with:

$$\mathrm{C3} = \left(\sum_{i \in S_c} c_i \left(\left| \hat{\mathbf{S}}_i \right|, \left| \mathbf{S}_{j^*} \right| \right) \right) \Big/ L \tag{22}$$

where L stands for the number of all separated components. S_c denotes the index set of the correctly annotated components among all separated components. Hence, it applies for the cardinality of the set S_c: $\#S_c \leq L$. When all L separated components are annotated correctly we have $\#S_c = L$. Then, the penalized mean correlation equals the mean correlation. Thus, the difference with respect to the C2 is that in the case of the C3 the whole space of latent variables is considered. Hence, it applies C2≤C3, and C2=C3 in the case of the perfect separation. For the aEFM-EKM-mRKHS algorithm, the overall dimensionality of the induced RKHSs is $D \times I$. Hence, we want to benefit from mapping the original input mixture spectra onto multiple high-dimensional RKHSs.

Criterion 4 (C4) stands for the mean rank of correctly annotated separated components:

$$\mathrm{C4} = \left(\sum_{i=1}^{M} m_i \right) \Big/ R \quad \text{s.t.} \begin{cases} m_i \text{ equals position on ranking list,} & \text{for } 1 \leq i \leq \# I_c \\ m_i = R, & \text{for } \# I_c < i \leq M \end{cases} \tag{23}$$

where R equals dimensionality of the space of latent variables. As an example, for the aEFM-EKM-mRKHS it applies $R = D \times I$. C4 simultaneously takes into account two factors: (*i*) higher dimensionality of induced space increases the probability that all separated components will be annotated correctly; (*ii*) it penalizes annotated components with the large ranking indices in the latent space as well as those components that are not annotated at all. Thus, if separation is perfect and all first M ranked components are annotated correctly the value of C4 will be (very)

small, i.e. $\lim_{D\to\infty} C4 = 0$. Since with the increase of dimension of induced space the probability of both correct and incorrect annotation is increasing, the C4 is sensitive to (in)correct annotation related to dimension of induced spaces.

3.0 Experiment and materials

3.1 Recording of ^{1}H NMR spectra of 160 pure components

We have recorded the in-house library comprised of ^{1}H NMR spectra of 160 pure components expected to correspond with the small organic molecules present particularly in urine samples of patients with diabetes type 2. Among pure components there were six pairs with amplitude spectra correlated above 0.9, eight pairs with correlation above 0.8, twelve pairs with correlation above 0.7, twenty-two pairs with correlation above 0.6, thirty-four with correlation above 0.5 and fifty-nine with correlation above 0.4. Thus, the spectral library contains many structurally similar components. Because of that, annotation is expected to be incorrect when separation quality is modest or poor. The library content is presented in Table S-3 of the Supplementary material. All measurements were performed on a Bruker AVANCE 600 MHz spectrometer, operating at 298 K. Samples were dissolved in 700 µL phosphate buffer (100 mM, pH 7.2 prepared with D_2O) prior to NMR measurement. 3-(Trimethylsilyl)-1-propanesulfonic acid sodium salt was used as an internal standard. Water suppression using excitation sculpting with gradients was applied [63]. ^{1}H spectra at a spectral width of 6.700 Hz with 16K data points and a digital resolution of

0.41 Hz per point were measured with 64 scans (time delay 2 sec, acquisition time 1.22 sec, pulse with 90).

3.2 ^{1}H NMR spectroscopy measurements of two model mixtures

To validate method proposed for nonlinear BSS single mixture problem, two mixtures of five pure components were prepared in the laboratory. The compounds 4-aminoantipyrine ($\mathbf{S}_1$), 4-aminobutyric acid ($\mathbf{S}_2$), allantoin ($\mathbf{S}_3$), cholic acid ($\mathbf{S}_4$) and naphtoic acid ($\mathbf{S}_5$) 30 mg of each, were mixed together. From the resulting crude mixture, 2 samples of 10 mg were taken and their NMR spectra were recorded for the pure components. Mixture $\mathbf{X}_2$ was used for validation of the single-mixture nonlinear BSS method. Mixtures $\mathbf{X}_1$ and $\mathbf{X}_2$ were used for validation of the nonlinear BSS methods for separation of pure components from two mixtures. These methods as well as corresponding validation results are presented in the Supplementary material.

3.3 Urine samples collection, preparation and ^{1}H NMR spectroscopy measurements

Urine aliquots were obtained from residual routine samples from 33 unrelated patients with diabetes type 2 (age range: 30 – 84 years; 17 males). Urine samples were collected in the morning, during the regular outpatient checkup in the clinical laboratory affiliated to the tertiary-level diabetes clinic. Patients were categorized and treated according to the current World Health Organization (WHO) recommendations at the University Clinic Vuk Vrhovac, Zagreb, that is the WHO collaborating center for diabetes. The study protocol was approved by the institutional Ethics Committee and patients gave their written consent for usage of their residual samples. The group of control subjects included 30 healthy, unrelated consenting adult volunteers, matched for

age and sex to diabetic subjects. For each of them the glucose level was measured before taking of urine and they were all normoglycemic. All study subjects were Caucasians. Morning urine samples were stored at -20^{0}C until clean-up procedure that is performed by C18 SampliQ Solid Phase Extraction (SPE) (Agilent Technologies, USA). C18 polymer sorbents were first conditioned by passing MeOH (3x5 mL) and then equilibrated by passing QH_2O (3x5 mL). Each urine sample (3x5 mL) was loaded into the column and a fraction was collected after cleaning in separate tubes. All the steps were performed at a flow rate of 1 mL min^{-1}. Thereafter, samples were frozen by immersion in liquid nitrogen followed by evaporation in the vacuum chamber of a freeze dryer to dryness (under controlled temperature and reduced pressure). 10 mg of each dry sample was further used for spectroscopic analysis. The NMR spectra of urine samples were recorded as described for the pure components. The single-mixture method aEFM-EKM-mRKHS was applied to 33 ^{1}H NMR mixtures spectra of urine obtained from diabetic patients and 30 ^{1}H NMR mixtures spectra of urine collected from healthy controls.

3.4 Software environment

All the experiments were executed on a PC running under a 64-bits Windows 10 operating system with 256 GB of RAM using Intel Xeon CPU E5-2650 v4 2 processors and operating with a clock speed of 2.2 GHz. All codes are run using MATLAB 2017a environment.

4. Results

4.1 Blind separation and annotation of five correlated amplitude ^{1}H NMR component spectra from one model ^{1}H NMR mixture spectrum

As mentioned in section 3.2, mixture $\mathbf{X}_2$ was used for validation of the single-mixture nonlinear BSS methods. RKHSs induced with the Gaussian kernels with variances $\sigma_i^2 \in \{1.0, 0.5, 0.1, 0.05\}$ were used to evaluate the single-mixture BSS algorithm. In addition to the aEFM-EKM-mRKHS algorithm, described in Algorithm 1, we also compared its single RKHS version (aEFM-EKM-sRKHS), where RKHS was generated with the Gaussian kernel with the variance $\sigma_1^2 = 1$. In addition to the single RKHS method based on the kernel *k*-means [64] estimation of the basis matrix **V**, the aEFM-EKM-sRKHS-V_{RKHS}, was also used for comparison. The RKHS was also generated with the Gaussian kernel with the variance $\sigma_1^2 = 1$. There was not enough diversity in the EKMs (13) generated with smaller variances $\sigma^2<1$, to estimate basis **V** using the kernel *k*-means algorithm. Thus, it was not possible to formulate multiple RKHSs version of the algorithm with basis matrices estimated by the kernel *k-means* clustering. Table 1 summarizes separation and annotation results for dimension of induced RKHSs D=2000 in terms of the criteria C1 to C4. Furthermore, additional information on computation time, correlation coefficients and ranking of annotated components are also presented in Table 1. Corresponding results for dimensions D=100 and D=1000 are presented in Tables S-1 and S-2 in the Supplementary material. Only the dimension of the induced RKHSs equal to D=2000 enabled detection of three (out of five) pure components with all three BSS algorithms. Thereby, the aEFM-EKM-mRKHS has the best performance. In the agreement with the "no free lunch theorem" this algorithm has the highest computational complexity. It can also be seen from Table 1 that separated and annotated components are mostly not placed at the top of the ranking list. Thus, in the real world scenario related to the separation and annotation of metabolites from single ^{1}H NMR mixture of a biological sample such as urine, an interpretation of the list of ranked annotated components by a domain expert will be necessary.

Table 1. Separation and annotation results from one ^{1}H NMR mixture for dimension of induced RKHSs D=2000.

	aEFM-EKM-mRKHS	aEFM-EKM-sRKHS	aEFM-EKM-sRKHS-V_{RKHS}
C1	1	1	1
C2	0.119	0.119	0.119
C3	0.229	0.223	0.242
C4	2.005	2.454	2.163
Ranks and correlations of correctly annotated components 1 to 5	28: 0.2327 1: 0.5925 NOT FOUND NOT FOUND 13: 0.3198	634: 0.2072 1: 0.5931 1985: 0.0355 NOT FOUND 288: 0.2810	276: 0.2594 1: 0.5934 NOT FOUND NOT FOUND 48: 0.3572
CPU time	42 915 s	24 261 s	26 583 s

4.2 Blind separation and annotation of correlated amplitude ^1H NMR component spectra from one ^1H NMR mixture spectrum of urine of diabetic and non-diabetic subjects

We applied the aEFM-EKM-mRKHS algorithm to separate and annotate components present in the single ^{1}H NMR spectra of 33 urine samples of patients with diabetes type 2 and 30 urine samples of healthy controls. Based on a discussion in section 4.1, the dimension of individual induced RKHSs was selected to be D=2000. We provide in Tables S-4 to S-66 in the Supplementary materials results of separation and annotation of 55 metabolites, expected to be related to diabetes type 2, obtained by means of the aEFM-EKM-mRKHS algorithm from each individual ^{1}H NMR spectra. Summarized results for all 55 metabolites are presented in Table S-67 in the Supplementary materials. The most prominent metabolites in samples from diabetic subjects, when compared to healthy controls, were urinary creatine, glutamic acid and 5-hydroxyindoleacetic acid. Table 2 presents aggregate separation and annotation related performance measures: number of times detected, mean and median ranks in the latent space comprised of 160 pure components (size of the library) as well as mean and median correlation between separated and annotated pure components. It is seen from the correlation values that related nonlinear single mixture BSS problem is very hard. Nevertheless, metabolites such as urinary creatine, glutamic acid and 5-hydroxyindoleacetic acid are detected in practically all the spectra and are more distinguished in spectra of urine of patients with diabetes type 2. To support this statement we also present in Table 2 results of statistical significance analysis. Analysis was performed for the correlations coefficients between creatine, glutamic acid and 5-hydroxyindoleacetic acid pure spectra and the corresponding component separated from diabetic and control groups. Since, the three metabolites were not detected in all the samples the length of compared vectors was determined by the number of times each metabolite was detected in the

control group. We performed one sided *anova* test, implemented in the MATLAB function `anova1`. The null hypothesis of the one sided *anova* test is that both vectors of correlation coefficients are drawn from the populations with the same mean. Thus, smaller probability (the *p* value) for the specific metabolite implies statistically more significant metabolite-related difference between diabetic and control groups. As seen in Table 2, the most prominent metabolite in this regard is glutamic acid. However, given the size of the test sample (number of patients tested), the *p*-values for creatine and 5-hydroxyindoleacetic acid are also reasonably small and emphasize difference between the diabetic and control groups.

Table 2. Separation and annotation performance of metabolites urinary creatine, glutamic acid and 5-hydroxyindoleacetic acid. Metabolites were extracted by the aEFM-EKM-mRKHS algorithm from ^{1}H NMR spectra of 33 urine samples of patients with diabetes type 2 and 30 urine samples of non-diabetic subjects. *p*-values were estimated for sequences of correlation coefficients corresponding with diabetic and control groups.

Metabolite	**Number of times detected**		**Mean rank / Median rank**		**Mean correlation / Median correlation**		
	33 diabetic patients	**30 control subjects**	**Diabetic patients**	**Control subjects**	**Diabetic patients**	**Control subjects**	***p*-value of one sided ANOVA**
creatine	32	30	10.78 / 7	21.37 / 13	0.319 / 0.318	0.287 / 0.286	0.242
glutamic acid	32	29	20.93 / 8	31.45 /21	0.274 / 0.299	0.224 / 0.211	0.041
5-hydroxyindoleacetic acid	32	30	31.5 / 15.5	39.17 / 28	0.226 / 0.260	0.194 / 0.193	0.118

5. Discussion

Metabolomic studies of diabetes and metabolic syndrome, using both targeted and non-targeted approach by either mass spectrometry or ^{1}H NMR spectroscopy, so far demonstrated the significant association of plasma branched chain amino acids: isoleucine, leucine and valine, as well as two aromatic amino acids: tyrosine and phenylalanine with the development of type 2 diabetes [65]. Furthermore, lipidomic-oriented studies identified plasma glycine, lysophosphatidylcholine 18:2 and acetylcarnitine as predictors of prediabetes and type 2 diabetes, [66]. Several studies reported on the associations between various phospholipids, hexoses and metabolites generated from oxidative damage, such as 2-aminoadipic acid, with the incident diabetes [67, 68]. These plasma metabolites were linked to the organ-specific processes and pathways involved in the pathogenesis of the type 2 diabetes [65]. Urinary metabolic profiling in diabetes is less prominent. That is partly because of the complexity of matrix, containing approximately 3100 so far identified metabolites [69], and partly because of the limitations of the current methodology, both analytical and computational, in separation of the signals generated by structurally similar molecules. The nonlinear single-mixture BSS method proposed herein was able to distinguish 3 metabolites involved in diverse pathways relevant for diabetes pathogenesis: urinary creatine, glutamic acid and 5-hydroxyindoleacetic acid. Glutamic acid, in the form of its monosodium salt is a well-established neurotransmitter responsible for the synaptic plasticity. It has been hypothesized that abnormal glutamate homeostasis might contribute to diabetes pathogenesis by direct and indirect mechanisms mediating a progressive loss of insulin-producing pancreatic β-cells [70]. Recent study provided evidence on an increased plasma glutamate level in diabetic patients and mice, as well as β-cell lines following short-term

exposure to high glucose *in vitro.* Enzymatic degradation of glutamate was able to normalize insulin secretion [71]. A toxic effect of an excess of glutamate in retinal cells was proposed as one of the mechanisms involved in the pathogenesis of diabetic retinopathy [72]. Thus, it seems that elevated level of glutamate plays a significant role in diabetes pathology. Urinary 5-hydroxyindoleacetate (5-HIAA) is an established indicator of serotonin levels and is routinely used as a laboratory test for carcinoid tumor diagnosis. Serotonin, synthesized by tryptophane hydroxylation in the brainstem serves as a neurotransmitter involved in regulation of multiple physiological functions of the brain, such as behavior and learning, as well as appetite and glucose homeostasis. However, peripherally produced serotonin serves as a hormone, which is involved in the regulation of function of the organs involved in the metabolic homeostasis at both glucose and lipid level [73]. A process called serotonylation was identified as an important modulating mechanism of the insulin production and secretion within the β-cells [74]. It was reported that a high level of plasma 5-HIAA in the stage of metabolic syndrome indicates a deranged serotonin metabolism with a presumed significant role for the development of cardiovascular complications via serotonin-mediated enhanced platelet aggregation and vasoconstriction [75]. Furthermore, regarding diabetes, it was recently proposed that an increased plasma 5-HIAA level in diabetic patients may play a role in the pathogenesis of microvascular complications [76]. The accumulated body of evidence pinpoints serotonin as a potential therapeutic target for type 2 diabetes and obesity [77]. Creatine (N-methyl-N-guanyxlglycine) is an essential guanidine compound widely distributed throughout human cells, which is equally provided by dietary sources and endogenous synthesis from arginine and glycine [78]. Phosphorylated creatine serves as the major endogenous phosphagenic substrate necessary for ATP synthesis within pathway catalyzed by creatine kinase. Creatine depletion, either acquired or inherited, seems to affect a variety of organs, with muscle and brain being the most interesting

targets [79, 80]. Despite a pronounced popularity, presumed improvement of muscle mass and athletic performance by the oral supplementation of creatine remained ambiguous. However, the widespread use of creatine for fitness purposes demonstrated its safety in healthy adults [81]. It was recently proposed that creatine deficiency, due to the aging-related reduction of muscular mass, may be responsible for age-related neurodegenerative diseases. Thus, creatine supplementation has emerged as an interesting treatment approach for a variety of geriatric disorders [82]. Pleiotropic effects of creatine seem to go beyond the creatine-kinase system of energy metabolism and involve various metabolic pathways, including glucose homeostasis [83]. Studies carried out in newly-diagnosed patients with the type 2 diabetes demonstrated that short-term oral ingestion of creatine elicited a reduction of plasma glucose which was equal to the effects obtained by two common oral antihyperglycaemic agents: sulfonylurea [84] and metformin [85]. As evidenced in the recent meta-analysis [86], longer-term supplementation of creatine yielded indeterminate results regarding glycemic control, but creatine supplementation could be regarded as an adjuvant nutritional therapy with hypoglycemic effects, particularly when used in combination with exercise. *In vitro* studies revealed that creatine was able to improve glucose-stimulate insulin release [87], as well as to facilitate translocation of muscular glucose transporter GLUT4 [88]. More recent research showed that AMPK signaling may be implicated in the GLUT4 effects of creatine supplementation on glucose uptake in the type 2 diabetes [90]. However, the mechanism(s) involved in the glucoregulatory action of creatine is far from being elucidated. Results of the present study indicate that urinary creatine secretion was significantly more pronounced in diabetic patients than healthy controls, which is a novel finding. Considering the evidence collected so far on the role of creatine on the glucose homeostasis, it could be speculated that type 2 diabetes may be associated with a disturbed utilization of creatine associated with an increased renal loss, possibly due to glomerular hyperfiltration, which is

commonly associated with diabetes [90]. In the pilot study conducted herein it is estimated that, based on one sided ANOVA test, glutamic acid, 5-HIAA and creatine discriminate between diabetic group and healthy control group with the *p*-values respectively equal to 0.0406, 0.1176 and 0.2419.

6. Conclusions

Blind separation of structurally similar (overlapping) components from a small number of their nonlinear mixtures is a hard inverse problem. It becomes notoriously difficult when only a single mixture is available. Yet, separation of structurally similar components from a single nonlinear mixture (metabolic profiling) is of potentially high clinical relevance. Driven by this motivation, this paper presented a method for the nonlinear blind separation and annotation of components present in single ^{1}H NMR amplitude mixture spectra. In addition to model (laboratory prepared) mixture, the method was tested on separation and annotation of metabolites present in urinary samples collected from patients with diabetes type 2 and healthy controls. The ability of the proposed method to identify metabolite-related differences between the groups, albeit in the very early pilot-stage, revealed an interesting and novel pattern of metabolic components within various pathways, which are known to be influenced by diabetes. In particular, the method pinpointed urinary creatine, glutamic acid and 5-hydroxyindoleacetic acid as the most prominent metabolites in samples from diabetic subjects, when compared to healthy controls. Since the presented study is at a pilot stage, our results do not allow any metabolic interpretation. However, our method was able to differentiate diabetic from non-diabetic subjects by identifying

potentially relevant metabolites depicting pathways relevant for diabetes pathology. Further studies are needed to validate this method in terms of obtaining reproducible and clinically relevant results.

Conflict of interest

Authors declare no conflict of interest.

Acknowledgments

The work performed has been supported through grant IP-2016-06-5235 "Structured decompositions of empirical data for computationally-assisted diagnosis of disease" funded by the Croatian Science Foundation. The first author thanks Gary McGuire for language editing of the final version of the manuscript.

Appendix A. Supplementary material

Supplementary material associated with this article can be found, in the online version, at http://

References

[1] J. K. Nicholson, J. Connelly, J.C. Lindon, E. Holmes, Metabonomics: a platform for studying drug toxicity and gene function, Nat. Rev. Drug Discovery 1 (2002) 153-161. https://doi.org/10.1038/nrd728.

[2] J. Keiser, U. Duthale, J. Utzinger, Update on the diagnosis and treatment of food-bone trematode infections, Curr. Opin. Infect. Dis. 23 (2010) 513-520. https://doi.org/10.1097/QCO.0b013e32833de06a.

[3] D. G. Robertson, Metabonomics in Toxicology: A Review, Toxicol. Sci. 82 (2005) 809-822. https://doi.org/10.1093/toxsci/kfi102.

[4] T. Hyotylainen, Novel methods in metabolic profiling with a focus on molecular diagnostic applications, Expert Rev. Mol. Diagn. 12 (2012) 527-538. https://doi.org/10.1586/erm.12.33.

[5] N.R. Patel, M.J.W. McPhail, M.I.F. Shariff, H.C. Keun, S.D. Taylor-Robinson, Biofluid metabpnomics using ^{1}H NMR spectroscopy: the road to biomarker discovery in gastroenterology and hepatology, Expert Rev. Gastroenterol. Hepatol. 6 (2012) 239-251. https://doi.org/10.1586/egh.12.1.

[6] D. S. Wishart, Metabonomics: applications to food science and nutrition research, Trends Food Sci. Technol. 19 (2008) 482-493. https://doi.org/10.1016/j.tifs.2008.03.003.

[7] S. Durand, M. Sancelme, P. Besse-Hoggan, B. Combourieu, Biodegradation pathaway of mesotrione: Complementaries of NMR, LC-NMR and LC-MS for qualitative and quantitaive metabolic profiling, Chemosphere 81 (2010) 372-380. https://doi.org/10.1016/j.chemosphere.2010.07.017.

[8] P.-M.Nguyen, C. Lyathaud, O. Vitrac, A Two-Scale Pursuit Method for the Tailored Identification and Quantification of Unknown Polymer Additives and Contaminants by ^{1}H NMR, Indust. & Eng. Chem. Res. 54 (2015) 2667-2681. https://doi.org/10.1021/ie503592z.

[9] T. Gebregiworgis, R. Powers, Application of NMR metabolomics to search for human disease biomarkers, Combinatorial Chemistry & High Throughput Screening 15 (2012) 595-610. https://doi.org/10.2174/138620712802650522.

[10] B. R. Seavey, E. A. Farr, W. M. Westler, J. L. Markley, A relational database for sequence-specific protein NMR data, J. Biomol. NMR. 1 (1991) 217-236. https://doi.org/10.1007/BF01875516.

[11] SpecInfo on the Internet NMR, https://application.wiley-vch.de/stmdata/specinfo.php (accessed April 29 2019).

[12] I. Toumi, S. Caldarelli, B. Torrésani, A review of blind source separation in NMR spectroscopy, Progress in Nuc. Mag. Res. Spec. 81 (2014) 37-64. https://doi.org/10.1016/j.pnmrs.2014.06.002.

[13] N. M. Jukarainen, S. -P. Korhonen, M. P. Laakso, M. A. Korolainen, M. Niemitz, P. P. Soininen, K. Tuppurainen, J. Vepsäläinen, T. Pirttilä, R. Laatikainen, Quantification of ^{1}H NMR spectra of human cerebrospinal fluid: a protocol based on constrained total-line-shape analysis, Metabolomics 1 (2008) 150-160. https://doi.org/10.1007/s11306-008-0106-6.

[14] G. F. Pauli, B. U. Jaki, D. C. Lankin, Quantitative 1H NMR: development and potential of a method for natural products analysis, J. Nat. Prod. 68 (2004) 133-149. https://doi.org/10.1021/np0497301.

[15] A. Smolinksa, L. Blanchet, L. M. C. Buydens, S. S. Wijmenga, NMR and pattern recognition methods in metabolomics: From data acquistion to biomarker discovery: A review, Anal. Chim. Acta. 750 (2012) 82-97. https://doi.org/10.1016/j.aca.2012.05.049.

[16] R. Schicho, R. Shaykhutdinov, J. Ngo, A. Nazyrova, C. Schneider, R. Panaccione, G. G. Kaplan, H. J. Vogel, M. Storr, Quantitaive Metabolomic Profiling of Serum, Plasma and Urine by ^{1}H NMR Spectroscopy Discriminates between Patients with Inflammatory Bowel Disease and Healthy Individuals, J. Proteome Res. 11 (2012) 3344-3357. https://doi.org/10.1021/pr300139q.

[17] N. MacKinnon, B. S. Somashekar, P. Tripathi, W. G. Thekkelnaycke, M. R. Arul, M. Chinnaiyan, A. Ramamoorthy, MetabolID: A graphical user interface package for assignment of ^{1}H NMR spectra of bodyfluids and tissues, J. Magn Reson. 226 (2013) 93-99. https://doi.org/10.1016/j.jmr.2012.11.008.

[18] G. A. Gowda Nagana, S. Zhang, H. Gu, V. Asiago, S. Narasimhamurty, D. Raftery, Metabolomics-based methods for early disease diagnostics, Expert Rev. Mol. Diag. 8 (2008) 617-633. https://doi.org/10.1586/14737159.8.5.617.

[19] I. Kopriva, I. Jerić, L. Brkljačić, Explicit-Implicit Mapping Approach to Nonlinear Blind Separation of Sparse Nonnegative Dependent Sources from a Single-Mixture: Pure Components Extraction from Nonlinear Mixture Mass Spectr, J. Chemometrics 29 (2015) 615-626. https://doi.org/10.1002/cem.2745.

[20] V. A. Shashilov, I. K. Lednev, Advanced Statistical and Numerical Methods for Spectroscopic Characterization of Protein Structural Evaluation, Chem. Rev. 110 (2010) 5692-5712. https://doi.org/10.1021/cr900152h.

[21] D. Nuzillard, S. Bourg, J. M. Nuzilard, Model-Free Analysis of Mixtures by NMR Using Blind Source Separation, J. Magn. Res. 133 (1998) 358-363. https://doi.org/10.1006/jmre.1998.1481.

[22] E. Visser, T. W. Lee, An information-theoretic methodology for the resolution of pure component spectra without prior information using spectroscopic measurements, Chemom. Int. Lab. Syst. 70 (2004) 147-155. https://doi.org/10.1016/j.chemolab.2003.11.003.

[23] I. Kopriva, I. Jerić, Multi-component Analysis: Blind Extraction of Pure Components Mass Spectra using Sparse Component Analysis, J. Mass Spectrom. 44 (2009) 1378-1388. https://doi.org/10.1002/jms.1627.

[24] I. Kopriva, I. Jerić, Blind Separation of Analytes in Nuclear Magnetic Resonance Spectroscopy and Mass Spectrometry: Sparseness-Based Robust Multicomponent Analysis, Anal. Chem. 82 (2010) 1911-1920. https://doi.org/10.1021/ac902640y.

[25] I. Kopriva, I. Jerić, V. Smrečki, Extraction of multiple pure component ^{1}H and ^{13}C NMR spectra from two mixtures: Novel solution obtained by sparse component analysis-based blind decomposition, Anal. Cim. Acta 653 (2009) 143-153. https://doi.org/10.1016/j.aca.2009.09.019.

[26] W. Naanaa, J. M. Nuzillard, Blind source separation of positive and partially correlated data, Sig. Proc. 85 (2005) 1711-1722. https://doi.org/10.1016/j.sigpro.2005.03.006.

[27] M. S. Karoui, Y. Deville, S. Hosseini, S. Ouamri, Blind spatial unmixing of multispectral images: New methods combining sparse component analysis, clustering and nonnegativity constraints, Patt. Recog. 45 (2012) 4263-4278. https://doi.org/10.1016/j.patcog.2012.05.008.

[28] I. Kopriva, I. Jerić, Blind Separation of Analytes in Nuclear Magnetic Resonance Spectroscopy: Improved Model for Nonnegative Matrix Factorization, Chem. Int. Lab. Syst. 137 (2014) 47-56. https://doi.org/10.1016/j.chemolab.2014.06.004.

[29] K. Zhang, L. Chan, Minimal Nonlinear Distortion Principle for Nonlinear Independent Component Analysis, J. Mach. Learn. Res. 9 (2008) 2455-248.

[30] Levin, D. N., 2008. Using state space differential geometry for nonlinear blind source separation, J. Appl. Phys. 103, 044906. https://doi.org/10.1063/1.2826943.

[31] D. N. Levin, Performing Nonlinear Blind Source Separation with Signal Invariants, IEEE Trans. Sig. Proc. 58 (2010) 2132-2140. https://doi.org/10.1109/TSP.2009.2034916.

[32] A. Taleb, C, Jutten, Source Separation in Post-Nonlinear Mixtures, IEEE Trans. Sig. Proc. 47 (1999) 2807-2820. https://doi.org/10.1109/78.790661.

[33] L. T. Duarte, R. Suyama, B. Rivet, R. Attux, J. M. T. Romano, C. Jutten, Blind compensation of nonlinear distortions: applications to source separation of post-nonlinear mixtures, IEEE Trans. Sig. Proc. 60 (2012) 5832-5844. https://doi.org/10.1109/TSP.2012.2208953.

[34] E. F. S. Filho, J. M. de Seixas, L. P. Calôba, Modified post-nonlinear ICA model for online neural discrimination, Neurocomputing 73 (2010) 2820-2828. https://doi.org/10.1016/j.neucom.2010.03.025.

[35] V. T. Nguyen, J. C. Patra, A. Das, A post nonlinear geometric algorithm for independent component analysis, Digital Sig. Proc. 15 (2005) 276-294. https://doi.org/10.1016/j.dsp.2004.12.006.

[36] A. Ziehe, M. Kawanabe, S. Harmeling, K. R. Müller, Blind Separation of Post-Nonlinear Mixtures Using Gaussianizing Transformations And Temporal Decorrelation, J. Mach. Learn. Res. 4 (2003) 1319-1338.

[37] K. Zhang, L. W. Chan, Extended Gaussianization Method for Blind Separation of Post-Nonlinear Mixtures, Neural Comput. 17 (2005) 425-452. https://doi.org/10.1162/0899766053011500.

[38] B. Ehsandoust, M. Babaie-Zadeh, B. Rivet, C. Jutten, Blind Source Separation in Nonlinear Mixtures: Separability and a Basic Algorithm, IEEE Trans. Sig. Proc. 65 (2017) 4339-4352. https://doi.org/10.1109/TSP.2017.2708025.

[39] S. Harmeling, A. Ziehe, M. Kawanabe, Kernel-Based Nonlinear Blind Source Separation, Neural Comput. 15 (2003) 1089-1124. https://doi.org/10.1162/089976603765202677.

[40] D. Martinez, A. Bray, Nonlinear Blind Source Separation Using Kernels, IEEE Tr. Neural Net. 14 (2003) 228-235. https://doi.org/10.1109/TNN.2002.806624.

[41] H. G. Yu, G. M. Huang, J. Gao, Nonlinear Blind Source Separation Using Kernel Multi-set Cannonical Correlation Analysis, Int. J. Comp. Net. Inf. Sec. 1 (2010) 1-8.

[42] L. Almeida, MISEP-Linear and nonlinear ICA based on mutual information, J. Mach. Learn. Res. 4 (2003) 1297-1318.

[43] M. E. Davies, C. I. James, Source separation using single channel ICA, Sig. Proc. 87 (2007) 1819-1832. https://doi.org/10.1016/j.sigpro.2007.01.011.

[44] Y. C. Ouyang H. M. Chen, J. W. Chai, C. C. C. Chen, S. K. Poon C. W. Yang, S. K. Lee, C. I. Chang, Band Expansion-Based Over-Complete Independent Component Analysis for Multispectral Processing of Magnetic Resonance Image, IEEE Trans. Biomed. Eng. 55 (2008) 1666-1677. https://doi.org/10.1109/TBME.2008.919107.

[45] B. Mijović, M. De Vos, I. Gligorijević J. Taelman , S. Van Huffel, Source Separation from Single-Channel Recordings by Combining Empirical Mode Decomposition and Independent Component Analysis, IEEE Trans. on Biomed. Eng. 57 (2010) 2188-2196. https://doi.org/10.1109/TBME.2010.2051440.

[46] J. Lin, A. Zhang, Fault feature separation using wavelet-ICA filter, NDT&E International 38 (2005) 421-427. https://doi.org/10.1016/j.ndteint.2004.11.005.

[47] Q. He, S. Su, R. Du, Separating mixed multi-component signal with an application in mechanical watch movement, Dig. Sig. Proc. 18 (2008) 1013-1028. https://doi.org/10.1016/j.dsp.2008.04.009.

[48] D. Gunawan, D. Sen, Iterative Phase Estimation for the Synthesis of Separated Sources from Single-Channel Mixtures, IEEE Sig. Proc. Let. 17 (2010) 421-424. https://doi.org/10.1109/LSP.2010.2042530.

[49] R. M. Parry, I. Essa I, Phase-Aware Non-negative Spectrogram Factorization, Lect. Notes Comp. Sci. 4666 (2007) 536-543. https://doi.org/10.1007/978-3-540-74494-8_67.

[50] B. Gao, W. L. Woo, B. W. K. Ling, Machine Learning Source Separation Using Maximum A Posteriori Nonnegative Matrix Factorization, IEEE Trans. on Cybernetics 44 (2014) 1169-1179. https://doi.org/10.1109/TCYB.2013.2281332.

[51] G. J. Jang, T. W. Lee, A Maximum Likelihood Approach to Single-channel Source Separation, J. Machine Learn. Res. 4 (2003) 1365-1392.

[52] S. T. Roweis, One microphone source separation, Advance in Neural Information Processing Systems 13 (2000) 793-799.

[53] E. M. Grai, H. Erdogan, Source separation using regularized NMF with MMSE estimates under GMM priors with online learning from uncertainties, Dig. Sig. Proc. 29 (2014) 20-34. https://doi.org/10.1016/j.dsp.2014.02.018.

[54] B. Schölkopf, A. Smola, Learning with kernels. MIT Press, Cambridge, MA, US, 2002.

[55] P. Comon, C. Jutted (Eds), Handbook of Blind Source Separation, Academic Press, Oxford, UK, 2010.

[56] C. Caifa, A. Cichocki, Estimation of Sparse Nonnegative Sources from Noisy Overcomplete Mixtures Using MAP, Neural Comput. 21 (2009) 3487-3518. https://doi.org/10.1162/neco.2009.08-08-846.

[57] I. Kopriva, I. Jerić, L. Brkljačić, Nonlinear mixture-wise expansion approach to underdetermined blind separation of nonnegative dependent sources, J. Chemometrics 27 (2013) 189-197. https://doi.org/10.1002/cem.2512.

[58] I. Kopriva, I. Jerić, M. Filipović, L. Brkljačić, Empirical Kernel Map Approach to Nonlinear Underdetermined Blind Separation of Sparse Nonnegative Dependent Sources: Pure Components Extraction from Nonlinear Mixtures Mass Spectra, J. Chemometrics 28 (2014) 704-715. https://doi.org/10.1002/cem.2635.

[59] R. A. DeVore, Deterministic constructions of compressed sensing matrices, J. Complexity 23 (2007) 918-925. https://doi.org/10.1016/j.jco.2007.04.002.

[60] Kopriva, I., Popović Hadžija, M., Hadžija, M., Aralica, G., 2015. Unsupervised segmentation of low-contrast multichannel images: discrimination of tissue components in microscopic image of unstained specimen. Scientific Reports 5, 11576. https://doi.org/10.1038/srep11576.

[61] N. Gillis, F. Glineur, Using underapproximations for sparse nonnegative matrix factorization, Pattern Recog. 43 (2010) 1676-1687. https://doi.org/10.1016/j.patcog.2009.11.013.

[62] The Nicolas Gillis web site: https://sites.google.com/site/nicolasgillis/code (accessed April 29 2019).

[63] T. L. Hwang, A. J. Shaka, Water Suppression That Works. Excitation Sculpting Using Arbitrary Wave-Forms and Pulsed-Field Gradients, J. Magn. Res. Series A 112 (1995) 275-279. https://doi.org/10.1006/jmra.1995.1047.

[64] R. Chitta, R. Jin, T. C. Havens, A. K. Jain, Approximate Kernel *k*-means: Solution to Large Scale Kernel Clustering, Proceedings of the 17th ACM SIGKDD conference on Knowledge Discovery and Data mining 2011, pp. 895–903, https://doi.org/10.1145/2020408.2020558.

[65] K. Suhre, Metabolic profiling in diabetes, J Endocrinol 221 (2014) R75–85. https://doi.org/10.1530/JOE-14-0024.

[66] R. Wang-Sattler, Z. Yu, C. Herder, A. C. Messias, A. Floegel, Y. He, et al., Novel biomarkers for pre-diabetes identified by metabolomics, Mol Syst Biol 8 (2012) 615, https://doi.org/10.1038/msb.2012.43.

[67] A. Floegel, A. von Ruesten, D. Drogan, M. B Schulze, C. Prehn, J. Adamski, et al., Variation of serum metabolites related to habitual diet: a targeted metabolomic approach in EPIC-Potsdam, Eur J Clin Nutr 67 (2013) 1100–1108. https://doi.org/10.1038/ejcn.2013.147.

[68] T. J. Wang, D. Ngo, N. Psychogios, A. Dejam, M. G. Larson, R. S. Vasan, et al., 2-Aminoadipic acid is a biomarker for diabetes risk, J. Clin. Invest. 123 (2013) 4309–4317. https://doi.org/0.1172/JCI64801.

[69] Bouatra, S., Aziat, A., Mandal, R., Guo, A. C., Wilson, M. R., Knox, C., et al., 2013. The Human Urine Metabolome 8, e73076. https://doi.org/10.1371/journal.pone.0073076.

[70] A. M. Davalli, C. Perego, F. B. Folli, The potential role of glutamate in the current diabetes epidemic, Acta Diabetol 49 (2012) 167–183. https://doi.org/10.1007/s00592-011-0364-z.

[71] Huang, X. T., Li, C., Peng, X. P., Guo, J., Yue, S. J., Liu, W., et al., 2017. An excessive increase in glutamate contributes to glucose-toxicity in β-cells via activation of pancreatic NMDA receptors in rodent diabetes. Sci Rep 7. https://doi.org/10.1038/srep44120.

[72] K. Gowda, W. J. Zinnantif, K. F. LaNoue, The influence of diabetes on glutamate metabolism in retinas, J Neurochem 117 (2011) 309-320. https://doi.org/10.1111/j.1471-4159.2011.07206.x.

[73] R. El-Merahbi, M. Löffler, A. Mayer, G. Sumara, The roles of peripheral serotonin in metabolic homeostasis, FEBS Lett 15 (2015) 1728–1734. https://doi.org/10.1016/j.febslet.2015.05.054.

[74] Paulmann, N., Grohmann, M., Voigt, J. P.,Bert, B., Vowinckel, J., Bader, M., Skelin, M., Jevsek, M., Fink, H., Rupnik, M., Walther, D. J., 2009. Intracellular Serotonin Modulates Insulin Secretion from Pancreatic β-Cells by Protein Serotonylation. PLoS Biol 7, e1000229. https://doi.org/ 10.1371/journal.pbio.1000229.

[75] M. Fukui, M. Tanaka, H. Toda, M. Asano, M. Yamazaki, G. Hasegawa, S. Imai, N. Nakamura, High plasma 5-hydroxyindole-3-acetic acid concentrations in subjects with metabolic syndrome, Diabetes Care 5 (2012) 163-167. https://doi.org/10.2337/dc11-1619.

[76] J. Saito, E. Suzuki, Y. Tajima, K. Takami, Y. Horikawa, J. Takeda, Increased plasma serotonin metabolite 5-hydroxyindole acetic acid concentrations are associated with impaired systolic and late diastolic forward flows during cardiac cycle and elevated resistive index at popliteal artery and renal insufficiency in type 2 diabetic patients with microalbuminuria, Endocr J 63 (2016) 69-76. https://doi.org/10.1507/endocrj.EJ15-0343.

[77] C. M. Oh, S, Park, H. Kim, Serotonin as a New Therapeutic Target for Diabetes Mellitus and Obesity, 40 (2016) 89-98. https://doi.org/10.4093/dmj.2016.40.2.89.

[78] M. Joncquel-Chevalier Curt, P. M. Voicu, M. Fontaine, A. F. Dessein, N. Porchet, K. Mention-Mulliez, D. Dobbelaere, G. Soto-Ares, D. Cheillan, J. Vamecq, Creatine biosynthesis and transport in health and disease, Biochimie 119 (2018) 146-165. https://doi.org/10.1016/j.biochi.2015.10.022.

[79] C. I. Nabuurs, C. U. Choe, A. Veltien, H. E. Kan, L. J. C. van Loon, R. J. T. Rodenburg, J. Matscjke, B. Wieringa, G. J. Kemp, D. Isbrandt, A. Heerschap, Disturbed energy metabolism and muscular dystrophy caused by pure creatine deficiency are reversible by creatine intake, J Physiol 591 (2013) 571–592. https://doi.org/ 10.1113/jphysiol.2012.241760.

[80] L. Hanna-El-Daher, O. Braissant, Creatine synthesis and exchanges between brain cells: What can be learned from human creatine deficiencies and various experimental models ?, Amino Acids 48 (2016) 1877–1895. https://doi.org/ 10.1007/s00726-016-2189-0.

[81] J. Butts, B. Jacobs, M. Silvis, Creatine Use in Sports. Sports Health 10 (2018) 31-34. https://doi.org/10.1177/1941738117737248.

[82] R. N. Smith, A. S. Agharkar, E. B. Gonzales, A review of creatine supplementation in age-related diseases: more than a supplement for athletes, F1000Research 3 (2014) 222. https://doi.org/10.12688/f1000research.5218.1.

[83] T. Wallimann, M. Tokarska-Schlattner, U. Schlattner, The creatine kinase system and pleiotropic effects of creatine, Amino Acids 40 (2011) 1271-1296. https://doi.org/10.1007/s00726-011-0877-3.

[84] B. Ročić, A. Znaor, P. Ročić, D. Weber, M. Vučić Lovrenčić, Comparison of antihyperglycemic effects of creatine and glibenclamide in type II diabetic patients, Wien Med Wochenschr 161 (2011) 519-523. https://doi.ord/ 10.1007/s10354-011-0905-7.

[85] Rocic, B., Bajuk, N. B., Rocic, P., Weber, D. S., Boras, J., Lovrencic, M. V., 2009. Comparison of antihyperglycemic effects of creatine and metformin in type II diabetic patients. Clin Invest Med 32,E322. https://doi.org/10.25011/cim.v32i6.10669.

[86] C. L. Pinto, P. B. Botelho, G. D. Pimentel, P. L. Campos-Ferraz, J. F. Mota, Creatine supplementation and glycemic control: a systematic review, Amino Acids 48 (2016) 2103-2129. https://doi.org/10.1007/s00726-016-2277-1.

[87] B. Ročić, M. Lovrenčić, M. Poje, S. Ashcroft, Effect of Creatine on the Pancreatic ß-Cell, Exp Clin Endocrinol Diabetes 115 (2007) 29-32. https://doi.org/10.1055/s-2007-949591.

[88] B. Op ’t Eijnde, B. Ursø, E. A. Richter, P. L. Greenhaff, P. Hespel, Effect of oral creatine supplementation on human muscle GLUT4 protein content after immobilization, Diabetes 50 (2001) 18-23. https://doi.org/10.2337/diabetes.50.1.18.

[89] C. R. R. Alves, J. C. Ferreira, M. A. de Siqueira-Filho, C. R. Carvalho, A. H. Lancha, B. Gualano, Creatine-induced glucose uptake in type 2 diabetes: a role for AMPK-α?, Amino Acids 43 (2012) 1803-1807. https://doi.org/10.1007/s00726-012-1246-6.

[90] G. Jerums, E. Premaratne, S. Panagiotopoulos, R. J. MacIsaac, The clinical significance of hyperfiltration in diabetes, Diabetologia 53 (2010) 2093-2104. https://doi.org/10.1007/s00125010-1794-9.

Tables Captions

Algorithm 1. Summary of the single-mixture nonlinear BSS algorithm aEFM-EKM-mRKHS.

Table 1. Separation and annotation results from one ^{1}H NMR mixture for dimension of induced RKHSs D=2000.

Table 2. Separation and annotation performance of metabolites urinary creatine, glutamic acid and 5-hydroxyindoleacetic acid. Metabolites were extracted by the aEFM-EKM-mRKHS algorithm from ^{1}H NMR spectra of 33 urine samples of patients with diabetes type 2 and 30 urine samples of non-diabetic subjects. *p*-values were estimated for for sequences of correlation coefficients corresponding with diabetic and control groups.