Blind Separation of Analytes in Nuclear Magnetic Resonance Spectroscopy and Mass Spectrometry: Sparseness-Based Robust Multicomponent Analysis

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

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

Metabolic profiling of biological samples involves nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry coupled with powerful statistical tools for complex data analysis. Here, we report a robust, sparseness-based method for the blind separation of analytes from mixtures recorded in spectroscopic and spectrometric measurements. The advantage of the proposed method in comparison to alternative blind decomposition schemes is that it is capable of estimating the number of analytes, their concentrations, and the analytes themselves from available mixtures only. The number of analytes can be less than, equal to, or greater than the number of mixtures. The method is exemplified on blind extraction of four analytes from three mixtures in 2D NMR spectroscopy and five analytes from two mixtures in mass spectrometry. The proposed methodology is of widespread significance for natural products research and the field of metabolic studies, whereupon mixtures represent samples isolated from biological fluids or tissue extracts.

Current achievements and progress in the field of systems biology and functional genomics depend sensitively on the level of development of associated analytical techniques.[1] Metabolic profiling of biological fluids, cells, and tissues provides insight into physiological processes, where it addresses multiple aims. These include disease diagnostics, xenobiotic toxicity, and nutrition- and environmental-influenced responses of living systems.[2] Information-rich techniques such as NMR spectroscopy and mass spectrometry (MS) represent powerful diagnostic tools for metabolomic and metabonomic studies, particularly through the identification and quantification of chemical entities directly correlated with certain disorder or disease (biomarkers).[3] One of the main disadvantages of ^{1}H NMR spectroscopy is signal overlapping, which increases with the number of components, their complexity, and/or similarity. This shortcoming can be significantly reduced by spreading to the second dimension. While 2D NMR spectroscopy is commonly used for the structure elucidation of biomacromolecules, there are limited examples of its application in metabolic analysis. 2D homonuclear and heteronuclear NMR spectroscopy was applied to the studies of central nervous system and muscles[4] but recently also to analyze healthy and cancerous tissues.[5] Despite significant improvement in many aspects, through isotopic labeling and chemoselective tagging,[6] 2D NMR spectra are still challenged by limited resolution. Thus, a high level of data complexity generated in metabolic studies requires adequate data analysis. A multivariate data analysis methodology capable of blind extraction of a single component (analyte) spectrum out of a mixture would significantly improve and accelerate metabolic fingerprinting, biomarker searches, and natural products analysis. Known as blind source separation (BSS), it has been reported previously in NMR,[7] infrared (IR),[8] electronic paramagnetic resonance (EPR),[9] and Raman[10] spectroscopy as well as mass spectrometry (MS).[11] However, in all these examples, algorithms of independent components analysis (ICA)[12] were used. These techniques assume that components are statistically independent, and their number is less than or equal to the number of mixtures available. When the spectra of different analytes overlap significantly, the statistical independence assumption is only partially fulfilled,[11] causing ICA to fail. Moreover, ICA cannot solve BSS problems characterized by more components than available mixtures. When mixtures represent samples of biological fluids or plant or tissue extracts with a few hundred analytes, overlapping of

* To whom correspondence should be addressed. E-mail: ikopriva@irb.hr. Phone: +385-1-4571-286. Fax: +385-1-4680-104.

[†] Division of Laser and Atomic Research and Development.

[‡] Division of Organic Chemistry and Biochemistry.

[1] van der Greef, J.; Stroobant, P.; van der Heijden, R. *Curr. Opin. Chem. Biol.* **2004**, *8*, 559–565.

[2] Ellis, D. E.; Dunn, W. B.; Griffin, J. L.; Allwood, J. W.; Goodacre, R. *Pharmacogenomics* **2007**, *8*, 1243–1266.

[3] Lindon, J. C.; Nicholson, J. K. *Ann. Rev. Anal. Chem.* **2008**, *1*, 45–69.

[4] Méric, P.; Autret, G.; Doan, B. T.; Gillet, B.; Sébrié, C.; Beloeil, J.-C. *Magn. Reson. Mater. Phys. Biol. Med.* **2004**, *17*, 317–338.

[5] Thomas, M. A.; Lange, T.; Velan, S. S.; Nagarajan, R.; Raman, S.; Gomez, A.; Margolis, D.; Swart, S.; Raylman, R. R.; Schulte, R. F.; Boesiger, P. *Magn. Reson. Mater. Phys. Biol. Med.* **2008**, *21*, 443–458.

[6] Ye, T.; Mo, H.; Shanaiah, N.; Gowda, G. A. N.; Zhang, S.; Raftery, D. *Anal. Chem.* **2009**, *81*, 4882–4888.

[7] Nuzillard, D.; Bourg, S.; Nuzillard, J. M. *J. Magn. Reson.* **1998**, *133*, 358–363.

[8] Visser, E.; Lee, T. W. *Chemom. Intell. Lab. Syst.* **2004**, *70*, 147–155.

[9] Ren, J. Y.; Chang, C. Q.; Fung, P. C. W.; Shen, J. G.; Chan, F. H. Y. *J. Magn. Reson.* **2004**, *166*, 82–91.

[10] Shashilov, V. A.; Xu, M.; Ermolenkov, V. V.; Lednev, I. K. *J. Quant. Spectrosc. Radiat. Transfer* **2006**, *102*, 46–61.

[11] Shao, X.; Wang, G.; Wang, S.; Su, Q. *Anal. Chem.* **2004**, *76*, 5143–5148.

[12] Cichocki, A.; Amari, S. I. *Adaptive Blind Signal and Image Processing*; John Wiley: New York, 2002.

10.1021/ac902640y © 2010 American Chemical Society
Published on Web 02/04/2010

resonant peaks is a common phenomenon.[3] Moreover, the number of components present and their concentrations are not known in advance. This adversely affects the accuracy of the ICA-based blind extraction of analytes. The same comment applies to the band target entropy method (BTEM) applied to the analysis of multicomponent 2D NMR spectra.[13] These lead to the underdetermined blind source separation (uBSS) scenario,[12,14,15] where an unknown number of components ought to be extracted, having only a smaller number of mixtures spectra at ones disposal, whereas the number of mixtures ought to be greater than one. We have recently demonstrated blind extraction of analytes from a smaller number of mixtures in Fourier transform-infrared (FT-IR) spectroscopy,[16] mass spectrometry,[17] and ^{1}H and ^{13}C NMR spectroscopy,[18] exploiting sparseness between the components in some representation domain. However, the mutual sparseness assumption is severely violated when the number of components, their complexity, or their similarity increases, leaving us unable to deal with biologically relevant problems. Here, we have proposed and verified an approach toward the solution of a problem considered within the chemometrics community "too ill-posed and thus, unsolvable."[13] The method relies on an assumption that components are mutually sparse (do not overlap) at a small number of points only. Thus, it is well founded to expect it to be successful in blind extraction of analytes from a small number of complex mixtures. In combination with daily improvements of analytical tools, this approach could potentially yield a viable method for biomarker identification and extraction from biological samples. The method is exemplified on blind extraction of analytes from the mixtures recorded in 2D NMR spectroscopy and mass spectrometry. Since it is derived to solve general-purpose BSS problems, its applications clearly extend beyond the blind extraction of analytes. As but one application example of great importance in systems biology, we point to the reconstruction of transcription factors in gene regulating networks.[3,19,20]

THEORY AND ALGORITHM

Linear Mixture Model. Blind extraction of the analytes is based upon the linear mixture model 1[7−11,16−18]

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

where $\mathbf{X} \in C^{I_n \times I_1 I_2 \ldots I_{n-1}}$ represents a matrix of in the general case, complex data. The I_n rows contain mixtures measured by some $(n - 1)$-dimensional spectroscopic modality; $\mathbf{A} \in R_{0+}^{I_n \times J}$ is an unknown nonnegative real matrix of concentration profiles of the unknown number of J analytes; and $\mathbf{S} \in C^{J \times I_1 I_2 \ldots I_{n-1}}$ is a matrix of (potentially complex) data, the J rows of which contain analytes. The linear mixtures model 1 is verified in ref 17 to be a valid description of the mixture of analytes mass spectra. The adopted notation can be used to represent mixtures measured either from one-dimensional or from multidimensional spectroscopic or spectrometric modalities, which might be necessary if the complexity of analytes is very high and/or their number is very large. According to the standard notation adopted for use in multiway analysis,[21] mixtures recorded by $(n - 1)$-dimensional modalities actually form the n-dimensional tensor: $\underline{\mathbf{X}} \in C^{I_1 \times I_2 \times \ldots \times I_n}$. The two-dimensional representation $\mathbf{X}$ adopted by the model 1 is obtained from a tensor $\underline{\mathbf{X}}$ through a mapping process known as n-mode flattening, matricization, or unfolding. To solve the BSS problem associated with the blind extraction of analytes, the mixtures data 1 will often have to be transformed into a new representation domain by means of some linear transform T:

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

Examples of such linear transform are wavelet or Fourier transforms.[18] The transform T is applied to $\mathbf{X}$ row-wise. If mixtures are recorded by higher-dimensional spectroscopic or spectrometric modality (2D NMR, for example), a higher-dimensional transform T is applied to each mixture before it is mapped to its one-dimensional counterpart. For example, the blind extraction of four analytes from three mixtures of 2D NMR spectra presented in the Results and Discussion has been carried out by transforming each mixture to a 2D wavelet domain to identify the matrix of concentrations and then to a 2D Fourier domain to identify the spectrum of the analytes. The BSS concept for extraction of analytes requires a complex signal format in order to detect samples of single-component activity. This format is not supported for some modalities, such as FT-IR spectroscopy or mass spectrometry. In such cases we propose a complex equivalent of real data $\mathbf{X}$ that is obtained through the analytic representation:[22]

$$\tilde{\mathbf{X}} = \mathbf{X} + jH(\mathbf{X}) \tag{3}$$

where $j = \sqrt{-1}$ denotes an imaginary unit and H denotes the Hilbert transform that is applied to $\mathbf{X}$ row-wise. Since a complex format, such as eq 3, is only necessary to detect points (indicies) of the single analyte activity in the chosen basis, any transform that yields a complex signal format can be used for this purpose as well. We have used the analytic representation eq 3 in the experiment, reported in the Results and Discussion, related to the blind extraction of five analytes from two mixtures of MS data.

Sparse Representations and Single-Component-Points. The matrix factorization $\mathbf{X} = \mathbf{AS}$ assumed by the linear mixture model 1 suffers from indeterminacies because $\mathbf{ATT}^{-1}\mathbf{S} = \mathbf{X}$ for any invertible $\mathbf{T}$, i.e., it implies that infinitely many ($\mathbf{A}$,$\mathbf{S}$) pairs can give rise to $\mathbf{X}$. The meaningful solution of the factorization of $\mathbf{X}$ is characterized with $\mathbf{T} = \mathbf{P\Lambda}$, where $\mathbf{P}$ is the permutation matrix and $\mathbf{\Lambda}$ is a diagonal matrix. These standard blind decomposition indeterminacies are obtained by imposing

[13] Guo, L.; Wiesmath, A.; Sprenger, P.; Garland, M. *Anal. Chem.* **2005**, *77*, 1655–1662.

[14] Bofill, P.; Zibulevsky, M. *Signal Process.* **2001**, *81*, 2353–2362.

[15] Georgiev, P.; Theis, F.; Cichocki, A. *IEEE Trans. Neural Net.* **2005**, *16*, 992–996.

[16] Kopriva, I.; Jerić, I.; Cichocki, A. *Chemom. Intell. Lab. Syst.* **2009**, *97*, 170–178.

[17] Kopriva, I.; Jerić, I. *J. Mass Spectrom.* **2009**, *44*, 1378–1388.

[18] Kopriva, I.; Jerić, I.; Smrečki, V. *Anal. Chim. Acta* **2009**, *653*, 143–153.

[19] Kitano, H. *Science* **2002**, *295*, 1662–1664.

[20] Liao, J. C.; Boscolo, R.; Yang, Y.-L.; Tran, L. M.; Sabatti, C.; Roychowdhury, V. P. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 15522–15527.

[21] Kiers, H. A. L. *J. Chemom.* **2000**, *14*, 105–122.

[22] Gabor, D. *Trans. Inst. Electr. Eng.* **1946**, *93*, 429–456.

statistical independence constraints on **S** when ICA[12,23] is used to solve related BSS problems.[7−11] As discussed previously, the ICA-related requirements for analytes are not met when mixtures represent complex systems. They can contain many analytes and it is therefore very likely that $J > I_n$. Therefore, a sparseness-based solution of the BSS problem eq 1 is proposed. It is said that the n-dimensional signal $\mathbf{y}$ is k-sparse in basis T if it is represented by $k \ll n$ coefficients, i.e., it is of special interest to look for the basis T where only a few entries of the vector of coefficients $T(\mathbf{y})$ are nonzero. In relation to the BSS problem associated with model 1, we comment that the sparseness request applies to the J-dimensional column vectors $\mathbf{s}_i$ of **S** or $T(\mathbf{S})$, $i \in \{1, ..., I_1I_2I_{n-1}\}$, while rows of **S** or $T(\mathbf{S})$ correspond to the analytes or their transformations. However, it is clear that if the row vectors of **S** or $T(\mathbf{S})$ are sparse, the column vectors in the corresponding representation will be sparse as well. In the absence of noise, if the column vectors of either **S** or $T(\mathbf{S})$ are $k = I_n - 1$ sparse, i.e., have $J - I_n + 1$ zero components, a unique solution of the underdetermined BSS (uBSS) problem, characterized with $J > I_n$, can be obtained.[15] Provided that the column vectors of either **S** or $T(\mathbf{S})$ are $k = 1$ sparse, a unique solution of the uBSS problem can be obtained, even from $I_n = 2$ mixtures only. However, for some signals such as those arising in NMR or FT-IR spectroscopy, it is very hard or even impossible to find a basis T where complex samples will be $k = 1$ sparse. Thus, instead of looking for the basis T that will yield $k = 1$ sparse representation of analytes at all sample points (in the most general case there are $I_1I_2...I_{n-1}$ sample points), we are interested in a representation T that will provide us with only P sample points where analytes are $k = 1$ sparse such that $J \leq P \ll I_1I_2...I_{n-1}$. Since $J \ll I_1I_2...I_{n-1}$, it ought to be possible to find such a small amount of points even when analytes exhibit a high degree of mutual similarity/complexity or their number is large. This belief is based on two facts: (i) the existence of a basis such as the wavelet basis with multiple degrees of freedom that provides signal representation at various resolution levels and different types of wavelet function and (ii) the number of detected points of single analyte activity is also governed by the choice of angular threshold $\Delta\theta$ in the direction-based criterion (eq 4), defined below. Thus, for the situation when analytes are highly complex or their number is large, the threshold $\Delta\theta$ can be increased, slightly compromising accuracy of the method through detection not of components of single analyte activity but single analyte dominance. Yet, if the complexity of the analytes is very high or their number is large, it might be necessary, for example, to use higher than 2D NMR spectroscopy. The use of points of single-component activity in the BSS has been exploited in the DUET algorithm in ref 24 for the separation of speech signals, wherein it has been assumed that at each point in the time-frequency plane only one speech signal is active. In our approach, we rely on the geometric concept of direction to detect points where single analytes are present. This detection criterion was proposed in ref 25. It requires complex representation of signals and was originally applied in the Fourier basis. The criterion is based on the notion that the real and imaginary parts of the complex vector of mixtures point either in the same or in opposite directions at the sample points of single analyte activity. This is based on the following reasoning. Let us denote by $\mathbf{x}_i$ the complex column vector of either the mixtures data **X** (eq 1) or the transformed mixtures data $T(\mathbf{X})$ (eq 2) at the sample index i. At the point i, where only one analyte is active, it applies for the vector of mixtures: $\mathbf{x}_i = \mathbf{a}_j s_{ij}$, where $\mathbf{a}_j$ is the vector of concentrations of the jth analyte across the mixtures and s_{ij} is the jth analyte that is active at point i. Since the vector of concentrations $\mathbf{a}_j$ is real, the real and imaginary parts of vector $\mathbf{x}_i$ must point in the same direction when the real and imaginary parts of s_{ij} have the same sign or, in opposite directions, when the real and imaginary parts of s_{ij} have different signs. Thus, the sample point i belongs to the set of single analyte points (SAPs) provided that the following criterion is satisfied

$$\left| \frac{R\{\mathbf{x}_i\}^{\mathrm{T}} I\{\mathbf{x}_i\}}{||R\{\mathbf{x}_i\}||\,||I\{\mathbf{x}_i\}||} \right| \geq \cos(\Delta\theta) \quad i \in \{1, ..., I_1I_2I_{n-1}\} \tag{4}$$

where $R\{\mathbf{x}_i\}$ and $I\{\mathbf{x}_i\}$ denote the real and imaginary part of $\mathbf{x}_i$, respectively. "T" denotes the transpose operation, $||R\{\mathbf{x}_i\}||$ and $||I\{\mathbf{x}_i\}||$ denote the ℓ_2-norms of $R\{\mathbf{x}_i\}$ and $I\{\mathbf{x}_i\}$, and $\Delta\theta$ denotes the angular displacement from directions of either 0 or π radians. Equation 4 follows from the definition of the inner product $R\{\mathbf{x}_i\}^{\mathrm{T}} I\{\mathbf{x}_i\} = ||R\{\mathbf{x}_i\}||\,||I\{\mathbf{x}_i\}|| \cos(\Delta\theta)$. At single analyte points, $\Delta\theta = 0$ and the inequality sign in eq 4 is replaced by an equality sign. Evidently, the smaller $\Delta\theta$ is, the smaller will be the number of candidates for identified as SAPs. However, the accuracy of the estimation of the number of analytes J and the concentration matrix **A** will be greater. In this regard, when mixtures **X** in eq 1 represent NMR signals, we propose the use of the wavelet rather than the Fourier basis to detect SAPs. If either the complexity of the analytes or their number is too great so that the chance of detecting sufficient SAPs is reduced (or zero), $\Delta\theta$ can be increased. This, in part, will affect the accuracy of the estimation of the concentration matrix **A** due to the fact that chance is increased that, instead of SAPs, we are detecting points where some of the analytes are dominant. This is important for not losing information about analytes that appear only on the diagonal in 2D NMR spectra and, therefore, are more likely to be dominant at a certain number of points rather than a single one.

Data Clustering-Based Estimation of the Number of Analytes and the Matrix of Concentration Profiles. A satisfactorily identified set of the SAPs enables the accurate estimation of the number of analytes J and the matrix of concentration profiles **A**. This is due to the fact that analytes in this set are $k = 1$ sparse and this condition, in the absence of noise, guarantees that the estimation of **A** is unique up to the permutation and scale.[15,25,26] At SAPs the following relation holds

$$\mathbf{x}_i = \mathbf{a}_j s_{j,i}\, j \in \{1, ..., J\}, \quad i \in \{1, ..., I_1I_2I_{n-1}\} \tag{5}$$

i.e., samples in the mixtures, which are column vectors of data matrix **X**, coincide with some of the columns of **A**. Thus, **A** can be estimated from **X** employing some of the available data-

(23) Comon, P. *Signal Process.* **1994**, *36*, 287–314.

(24) Jourjine, A.; Rickard, S.; Yilmaz, O. *Proc. Int. Conf. Acoust., Speech, Sig. Process.* **2000**, *5*, 2985–2988.

(25) Reju, V. G.; Koh, S. N.; Soon, I. Y. *Signal Process.* **2009**, *89*, 1762–1773.

(26) Naini, F. M.; Mohimani, G. H.; Babaie-Zadeh, M.; Jutten, Ch. *Neurocomputing* **2008**, *71*, 2330–2343.

clustering algorithms.[26,27] However, in many BSS algorithms it is assumed that the number of analytes J is either known or can be estimated easily. This does not seem to be true in practice at all, especially when the BSS problem is underdetermined and mixtures represent samples of biological fluids or tissue extracts.[3] Generally speaking, the estimation of the number of analytes is a complex issue known in computer science as the intrinsic dimensionality problem.[28] A few related methods are described in refs 29−31. However, they all assume $J \geq I_n$. Thus, they are not applicable to the uBSS problem that is of central interest here. To estimate the number of analytes for an identified set of SAPs, we propose to use the clustering function:[16−18]

$$f(\mathbf{a}) = \sum_{i=1}^{P} \exp\left(-\frac{d^2(\mathbf{x}_i, \mathbf{a})}{2\sigma^2}\right) \tag{6}$$

where d denotes the distance calculated as $d(\mathbf{x}_i,\mathbf{a}) = [1 - (\mathbf{x}_i \cdot \mathbf{a})^2]^{1/2}$ and $(\mathbf{x}_i \cdot \mathbf{a})$ denotes the inner or dot product. $\mathbf{a}$ represents the mixing vector in a two-dimensional subspace that is parametrized as

$$\mathbf{a} = [\cos(\varphi)\ \sin(\varphi)]^{\mathrm{T}} \tag{7}$$

where φ represents the mixing angle that is confined in the interval $[0, \pi/2]$ due to the non-negativity of the mixing coefficients (they represent concentration profiles of the analytes). Parameter σ defines the resolving power of the function $f(\mathbf{a})$. When σ is set to a sufficiently small value (in reported experiments this turned out to be $\sigma \approx 0.05$), the value of the function $f(\mathbf{a})$ will approximately equal the number of data points close to $\mathbf{a}$. The number of peaks of the function $f(\mathbf{a})$ in the interval $[0, \pi/2]$ corresponds to the estimate of the number of analytes J present in the mixtures. The selection of a two-dimensional subspace out of an I_n-dimensional mixture space greatly simplifies the computational complexity of the estimation process due to the fact that an $(I_n - 1)$ dimensional search in the space of mixing angles is reduced to a one-dimensional search. The reduction to the two-dimensional subspace is enabled by the fact that each analyte is present in some concentration in each of the I_n mixtures available. It is clear that the value of σ reported above is empirical. For another set of mixtures it can yield a different value for J. To obtain a robust estimator of the number of analytes J, we have proposed in refs 16−18 to decrease the value of σ until the estimated number of analytes is increased by 1 or 2. False analytes will be either a repeated version of some of the true analytes or their linear combinations. Thus, they can be detected after blind extraction as the ones that are highly correlated with the rest of the extracted analytes. It is also clear that if the concentration profiles of the analytes are very similar it will be increasingly more difficult to discriminate them. In such a case, the solution might be to evaluate the clustering function in 3D or even higher-dimensional space, because this will decrease the probability that different analytes have the same concentration profiles across an increased number of mixtures. This however adds to the computational complexity of the algorithm due to the fact that the one-dimensional search in the domain of mixing angles is replaced by a search in a higher-dimensional space. After the number of analytes J is estimated, the matrix of concentration profiles $\mathbf{A}$ is estimated on the same set of SAPs employing some of data clustering methods.[27] In the subsequent experimental setup, hierarchical and k-means clustering, implemented through the *clusterdata* and *kmeans* commands from MATLAB's Statistical toolbox, have been used for this purpose.

Estimation of Analytes in Over-, Even- and Under-Determined Scenarios. When the estimated number of analytes J is less than or equal to the number of mixtures I_n, the resulting BSS problem is, respectively, over- or even-determined. Analytes can be estimated through the simple matrix pseudoinverse:

$$\mathbf{S} = \mathbf{A}^{\dagger}\mathbf{X} \tag{8a}$$

or

$$T(\mathbf{S}) = \mathbf{A}^{\dagger}T(\mathbf{X}) \tag{8b}$$

where $\mathbf{A}^{\dagger}$ denotes the Moore−Penrose pseudoinverse of $\mathbf{A}$. Whether eq 8a or 8b is employed depends on the type of the spectroscopic modality that is used. If NMR spectroscopy is used, it is customary to estimate analytes in the Fourier domain, in which case eq 8b is preferred with T representing the Fourier transform (note that for NMR data, $\mathbf{A}$ is identified in the wavelet domain). If mass spectrometry or FT-IR spectroscopy are used, it is customary to estimate analytes in the recording domain, eq 8a. For reasons of clarity, we emphasize again that the transform T is applied to $\mathbf{X}$ row-wise. If mixtures are recorded by higher-dimensional spectroscopic or spectrometric modality (2D NMR, for example) a higher-dimensional transform T is applied to each mixture before it is mapped to its one-dimensional counterpart. The accuracy of the pseudoinverse approach (eq 8a/8b) with $\mathbf{A}$ identified on a set of SAPs greatly outperforms the one obtained by ICA, as reported in ref 25. When the number of analytes J is greater than the number of mixtures I_n, the resulting BSS problem is underdetermined. In such a case, the inverse problem has many solutions and the simple pseudoinverse approach (eq 8a/8b) can no longer be applied. Provided that either $\mathbf{s}_i$ or $T(\mathbf{s}_i)$ are $k = I_n - 1$ sparse, i.e., have $J - I_n + 1$ zero components, it is possible to obtain the solution of the resulting uBSS problem through ℓ_1 -norm minimization,[14,15] once the number of analytes J and concentration matrix $\mathbf{A}$ are estimated. The analyte extraction problem is then reduced to solving the resulting underdetermined system of linear equations that is carried out as linear programming[14,32,33] or the ℓ_1 -regularized least-squares problem.[34,35] Provided that the concentration

[27] Gan, G.; Ma, Ch.; Wu, J. *Data Clustering-Theory, Algorithms and Applications*; SIAM: Philadelphia, PA, 2007.

[28] Fukunaga, K.; Olsen, D. R. *IEEE Trans. Comput.* **1971**, *C-20*, 176–183.

[29] Malinowski, E. R. *Anal. Chem.* **1977**, *49*, 612–617.

[30] Levina, E.; Wagman, A. S.; Callender, A. F.; Mandair, G. S.; Morris, M. D. *J. Chemom.* **2007**, *21*, 24–34.

[31] Westad, F.; Kermit, M. *Anal. Chim. Acta* **2003**, *490*, 341–354.

[32] Takigawa, I.; Kudo, N.; Toyama, J. *IEEE Trans. Signal Process.* **2004**, *52*, 582–591.

[33] Donoho, D. L.; Elad, M. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 2197–2202.

[34] Kim, S. J.; Koh, K.; Lustig, M.; Boyd, S.; Gorinevsky, S. *IEEE J. Sel. Topics Signal Proc.* **2007**, *1*, 606–617.

[35] Tropp, J. A.; Gilbert, A. C. *IEEE Trans. Inf. Theory* **2007**, *53*, 4655–4666.

matrix $\mathbf{A}$ is estimated accurately, the result in ref 32 states that the minimum of the ℓ_1 -norm yields an accurate solution of the uBSS problem even if analytes are I_n-sparse, i.e., have $J - I_n$ zero components. It means that I_n analytes can coexist at each sample point. When multiple analytes occupy each sample point of a generally complex mixture, we, respectively, notice the relation between the real and imaginary parts of $\mathbf{x}_i$ as $R\{\mathbf{x}_i\} = \mathbf{A}R\{\mathbf{s}_i\}$ and $I\{\mathbf{x}_i\} = \mathbf{A}I\{\mathbf{s}_i\}$, $i \in \{1, ..., I_1 I_2 I_{n-1}\}$. Written in matrix formulation it reads as

$$\begin{bmatrix} R\{\mathbf{x}_i\} \\ I\{\mathbf{x}_i\} \end{bmatrix} = \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{A} \end{bmatrix} \begin{bmatrix} R\{\mathbf{s}_i\} \\ I\{\mathbf{s}_i\} \end{bmatrix} \tag{9a}$$

or

$$\bar{\mathbf{x}}_i = \bar{\mathbf{A}}\bar{\mathbf{s}}_i \tag{9b}$$

where in eq 9a $\mathbf{0}$ is the matrix with the same dimensions as $\mathbf{A}$ and all entries are equal to 0. We introduce dummy variables $\mathbf{u},\mathbf{v} \geq \mathbf{0}$ such that $\bar{\mathbf{s}}_i = \mathbf{u} - \mathbf{v}$. Assuming that

$$\mathbf{z} = \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix}$$

and

$$\bar{\bar{\mathbf{A}}} = [\bar{\mathbf{A}} \; - \bar{\mathbf{A}}]$$

the linear programming based solution with equality constrains is obtained as

$$\hat{\mathbf{z}}_i = \arg\min_{\mathbf{z}_i} \sum_{j=1}^{2J} z_{j,i} \text{ subject to } \bar{\bar{\mathbf{A}}}\mathbf{z}_i = \bar{\mathbf{x}}_i \; \forall i = 1, ..., I_1 I_2 I_{n-1}$$
$$\mathbf{z}_i \geq \mathbf{0} \tag{10}$$

Linear programming (eq 10) favors the solution with the minimal ℓ_1 -norm. With high probability, this is the sparsest solution of eq 9b.[33−35] Hence, if analytes satisfy the desired degree of mutual sparseness, the $k \leq I_n$ solution of eq 10 will successfully recover them. Analytes are obtained from the solution of the linear program (eq 10) as $\bar{\mathbf{s}}_i$, where $\mathbf{u}$ is obtained from the upper half of $\hat{\mathbf{z}}_i$ and $\mathbf{v}$ is obtained from the lower half of $\hat{\mathbf{z}}_i$. The real part of $\mathbf{s}_i$ is obtained from the upper half of $\bar{\mathbf{s}}_i$ while the imaginary part of $\mathbf{s}_i$ is obtained from the lower half of $\bar{\mathbf{s}}_i$. If noise is present in the uBSS problem, a more robust solution for $\hat{\mathbf{z}}_i$ (thus also $\mathbf{s}_i$) is obtained by solving the ℓ_1 -regularized least-squares problem:[34]

$$\hat{\mathbf{z}}_i = \arg\min_{\mathbf{z}_i} \frac{1}{2}\|\bar{\bar{\mathbf{A}}}\mathbf{z}_i - \bar{\mathbf{x}}_i\|_2^2 + \lambda\|\mathbf{z}_i\|_1 \quad \forall i = 1, ..., I_1 I_2 I_{n-1} \tag{11}$$

Solution of eq 11 minimizes the ℓ_2 -norm of the error between data $\bar{\mathbf{x}}_i$ and its model $\bar{\bar{\mathbf{A}}}\mathbf{z}_i$, trading the degree of error for the degree of sparseness of the solution. The degree of compromise is balanced by the value of the regularization factor λ. There are other methods developed over the past few years for solving underdetermined systems of linear equations. Most notable are methods that minimize the ℓ_p -norm ($0 < p \leq 1$) of the solution coefficients (analytes), such as the iterative recursive least-squares (IRLS) algorithm,[36] methods that optimize the null-space of the concentration matrix $\mathbf{A}$,[37] and methods that work with a smooth approximation of the ℓ_0 -quasi norm of the solution coefficients.[38] We have checked the IRLS algorithm and the smoothed ℓ_0 -quasi norm algorithm on the experimental problem considered below. These methods did not bring any improvement relative to the performance achieved by the interior point method employed to solve the ℓ_1 -regularized least-squares problem (eq 11) or the linear programming method employed to solve eq 10.

EXPERIMENTAL SECTION

NMR Measurements. We used 6-*O*-(*N*,*O*-bis-*tert*-butyloxycarbonyl-L-tyrosyl-L-prolyl)-D-glucopyranose (**1**), 6-*O*-(*N*,*O*-bis-*tert*-butyloxycarbonyl-L-tyrosyl-L-prolyl-L-phenylalanyl)-D-glucopyranose (**2**), 6-*O*-(*N*-*tert*-butyloxycarbonyl-L-prolyl-L-phenylalanyl-L-valyl)-D-glucopyranose (**3**), and 6-*O*-(*N*,*O*-bis-*tert*-butyloxycarbonyl-L-tyrosyl-L-prolyl-L-phenylalanyl-L-valyl)-D-glucopyranose (**4**)[39] to prepare three mixtures with different ratios of **1**−**4**: $\mathbf{X_1}$ (**1**/**2**/**3**/**4** = 1.1:1.7:2.7:1), $\mathbf{X_2}$ (**1**/**2**/**3**/**4** = 2.5:1.7:1.3:1), and $\mathbf{X_3}$ (**1**/**2**/**3**/**4** = 1:4:2.7:2.2). To test the ability of the ICA-based approach, which requires the number of mixtures to be equal or greater than the number of analytes, the fourth mixture $\mathbf{X_4}$ (**1**/**2**/**3**/**4** = 3.2:1:2.3:3.5) has been prepared and treated as described above. Compounds **1**−**4** and mixtures $\mathbf{X_1}$−$\mathbf{X_4}$ were dissolved in 600 μL of DMSO-d_6 and NMR spectra recorded with a Bruker AV300 spectrometer, operating at 300.13 MHz and 298 K. The ^{1}H−^{1}H correlation spectroscopy (COSY) spectra were obtained in the magnitude mode with 2048 points in the F_2 dimension and 512 increments in the F_1 dimension. Each increment was obtained with 4 scans and a spectral width of 6173 Hz. The resolution was 3.01 and 6.02 Hz per point in the F_1 and F_2 dimensions, respectively.

Mass Spectrometry Measurements. The compounds used for the analysis and procedures regarding MS measurements are described in ref 17.

Software Environment. The BSS method described was tested on the decomposition of 2D COSY NMR spectra and mass spectra using custom scripts in the MATLAB programming language (version 7.1.; The MathWorks, Natick, MA). The data clustering part of the SCA algorithm was implemented using the *clusterdata* and *kmeans* commands from the Statistics toolbox. The *clusterdata* command was used with the following set of parameters: distance, cosine; linkage, complete; maxclus, J, where J represents number of analytes estimated previously from the peaks of the clustering function (eqs 6/7). The linear programming part of the SCA algorithm was implemented using the *linprog* command from the Optimization toolbox and the interior point method.[34,40] The two-dimensional wavelet transform was implemented using the *swt2* command from the Wavelet toolbox. All programs were executed on a PC running under the Windows

[36] Cahrtrand, R.; Staneva, V. *Inverse Problems* **2008**, *24*, 035020 (14 pages).
[37] Kim, S. G.; Yo, Ch. D. *IEEE Trans. Signal Process.* **2009**, *57*, 2604–2614.
[38] Mohimani, H.; Babaie-Zadeh, M.; Jutten, C. *IEEE Trans. Signal Process.* **2009**, *57*, 289–301.
[39] Jerić, I.; Horvat, Š. *Eur. J. Org. Chem.* **2001**, 1533–1539.
[40] http://www.stanford.edu/~boyd/l1_ls/.

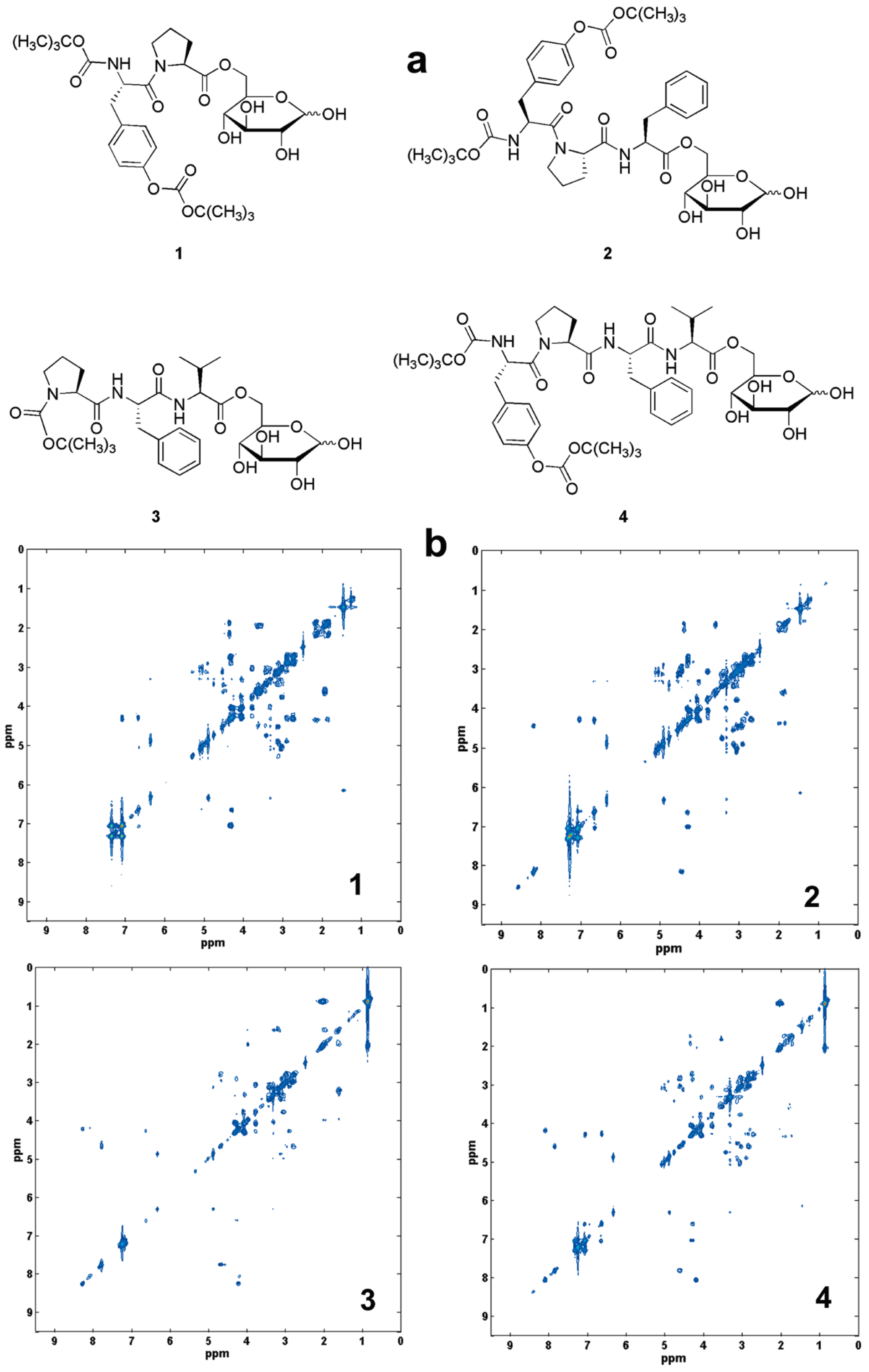

Figure 1. (a) Structures of glycopeptides**1**−**4**; (b) COSY NMR spectra of pure analytes **1**−**4**.

XP operating system using an Intel Core 2 Quad Processor Q6600 operating with a clock speed of 2.4 GHz and 4 GB of RAM installed.

RESULTS AND DISCUSSION

Setting up an Experiment. To demonstrate the efficiency of the proposed multivariate data analysis method, a "control" experiment was set up. Since we are targeting *complex mixtures*, it was important to choose a group of compounds that will comply with the complexity requirement. It was also important to have known and well-characterized compounds to verify the accuracy of estimation. Among many options, we have selected the glycopeptides **1**−**4**,[39] where the N-terminally protected dipeptide (Tyr-Pro, **1**), tripeptides (Tyr-Pro-Phe, **2**) and (Pro-Phe-Val, **3**), and tetrapeptide (Tyr-Pro-Phe-Val, **4**) are linked to the C-6 group of D-glucose (Figure 1a). Crude compounds **1**−**4** were mixed to obtain three mixtures with different concentrations of components (see Experimental Section for details).

As seen from their structures, compounds **1**−**4** are structurally analogous, and consequently their spectral profiles are, to a large extent, similar (Figure 1b). Additionally, the presence of a reducing sugar gives rise to both α- and β-pyranose forms in the solution, while the presence of the proline residue causes *cis*−*trans* isomerization of the X-Pro peptide bond. All together, accurate assignment of all resonances requires 2D NMR measurements. Even so, COSY spectra obtained from mixtures consisted of compounds **1**−**4** (Figure 2) showed overlap, which undoubtedly hampered assignment. Thus, the proposed mixture model passed the complexity requirement and seemed adequate for the data analysis.

Blind Extraction of Four Analytes from Three Mixtures in 2D NMR Spectroscopy. Figures 1−4 and Table 1 demonstrate the experimental blind extraction of four pure-component COSY spectra from three mixtures by means of the described sparseness-based multivariate data analysis method. The COSY spectra of compounds **1**−**4** are presented in Figure 1b, while Figure 2 shows the COSY spectra of the three mixtures. The structural similarity of the selected compounds accounts for the complexity and overlap in the sugar resonance region, as well as in the amino-acid amide and side-chain resonance areas. A convenient way to quantify this overlapping is to calculate normalized correlation coefficients between the spectra of the analytes **1**−**4** (Table 1a). It is clear that compound **1** is highly correlated with (similar to) compound **2** (0.5509). Compound **2** is additionally correlated with **4** (0.5120), while compounds **3** and **4** are highly mutually correlated, with a coefficient of 0.7965. Clearly, these correlation coefficients reflect structural and spectral similarities between the studied compounds and allow simplified numerical analysis of often complex NMR spectra.

Clustering functions described by the eqs 6 and 7, are shown in the mixing angle domain in Figure 3 for three two-dimensional subspaces $\mathbf{X}_1\mathbf{X}_2$, $\mathbf{X}_1\mathbf{X}_3$, and $\mathbf{X}_2\mathbf{X}_3$, i.e., all combinations of two mixtures were used for the estimation of the number of analytes present in the mixtures. The clustering functions were calculated on a set of 203 SAPs (eq 5) detected in the symmlet 8 wavelet domain using direction based criterion (eq 4) with the angular displacement set to $\Delta\theta = 1°$. The value of the dispersion factor σ in eq 5 has, respectively, been set to 0.04, 0.06, and 0.05. The meaning of 203 SAPs is that only one of components **1**−**4** was active at only 203 out of 65 536 points available. The four peaks in clustering functions suggest the existence of four analytes in the mixtures. The small variation of the dispersion factors parameter confirms the statement that any two-dimensional mixtures subspace can be used for the estimation of the number of analytes.

The spectra of the pure components estimated from three mixtures $\mathbf{X}_1$−$\mathbf{X}_3$ (Figure 2) are shown in Figure 4. Since the concentration matrix is estimated accurately on a subset of SAPs, the ℓ_1-regularized least-squares method, eq 11, yielded good estimates of the analytes spectra, even when two components occupy the same frequency. The similarity between the spectra of pure and estimated analytes is quantified in Table 1b, where normalized correlation coefficients between the true and estimated analytes spectra are shown. The closer these numbers are to those in Table 1a, the better is the extraction of the components from the mixtures. Inspection of the data shows that all four components were successfully separated from three mixtures; even highly correlated components **3** and **4** are assigned reliably.

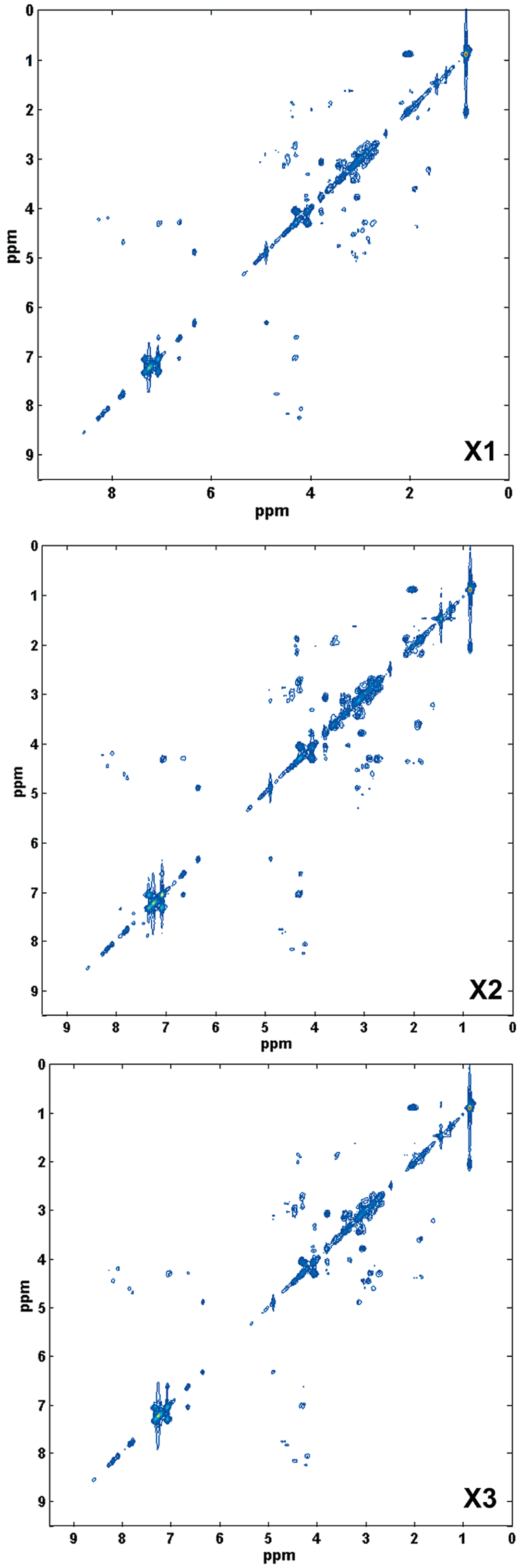

Figure 2. COSY NMR spectra of three mixtures $\mathbf{X}_1$−$\mathbf{X}_3$.

To demonstrate the importance of the wavelet basis for providing sparse representation of the NMR signals, we have

Table 1. Normalized Correlation Coefficients for (a) Pure Analytes 1−4; (b) Analytes 1−4 Estimated on 203 SAPs Detected in Symmlet 8 Wavelet Domain; (c) Analytes 1−4 Estimated on 23 SAPs Detected in Fourier Domain; (d) Analytes 1−4 Estimated by Means of JADE ICA Algorithm from Four Mixtures[a]

entry		An_1	An_2	An_3	An_4
a	An_1	**1**	0.5509	0.1394	0.3730
	An_2	0.5509	**1**	0.3051	0.5120
	An_3	0.1394	0.3051	**1**	0.7965
	An_4	0.3730	0.5120	0.7965	**1**
b	$\hat{A}n_1$	**0.8931**	0.4753	0.2638	0.4132
	$\hat{A}n_2$	0.5634	**0.8579**	0.2795	0.5366
	$\hat{A}n_3$	0.1945	0.5048	**0.8990**	0.7953
	$\hat{A}n_4$	0.4386	0.6124	0.8060	**0.8381**
c	$\hat{A}n_1$	**0.8924**	0.6009	0.2754	0.4602
	$\hat{A}n_2$	0.5482	**0.8469**	0.3107	0.5695
	$\hat{A}n_3$	0.0931	0.4101	**0.8432**	0.7249
	$\hat{A}n_4$	0.3108	0.3411	**0.8236**	0.7331
d	$\hat{A}n_1$	0.7189	0.7090	0.6805	**0.7939**
	$\hat{A}n_2$	0.6873	0.7571	0.6524	**0.7790**
	$\hat{A}n_3$	0.6606	0.7325	0.7142	**0.8177**
	$\hat{A}n_4$	0.6322	0.7232	0.7474	**0.8342**

[a] A significant degree of correlation between spectra of true analytes caused failure of the ICA-based extraction of analytes, part d. An_1−An_4 pure analytes **1**−**4**; $\hat{A}n_1$−$\hat{A}n_4$ estimated analytes **1**−**4**.

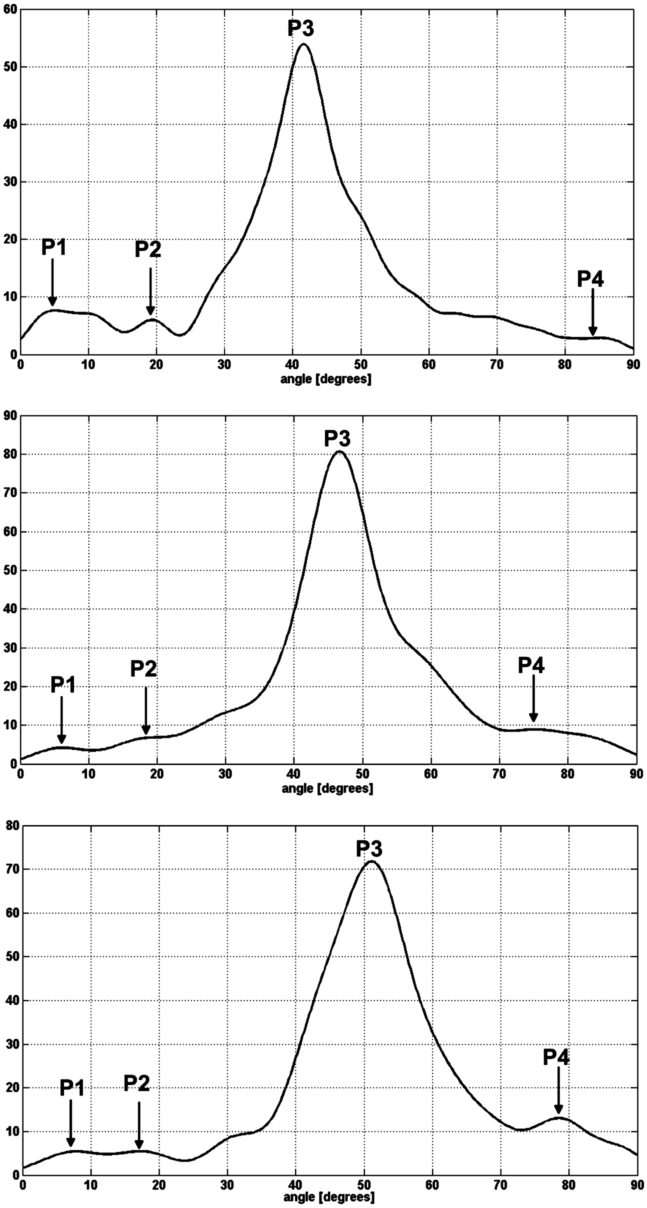

Figure 3. Clustering functions calculated on 203 SAPs in the wavelet domain for three two-dimensional mixture subspaces: $\mathbf{X}_1\mathbf{X}_2$, $\mathbf{X}_1\mathbf{X}_3$, and $\mathbf{X}_2\mathbf{X}_3$. Positions of the four peaks P1−P4 in each function are marked.

estimated the set of SAPs in the Fourier domain with the angular displacement criterion set to $\Delta\theta = 2°$, i.e., 2 times greater than in the case of the wavelet basis. However, only 23 SAPs were detected in this case and the estimation of the matrix of concentration profiles was less accurate. Consequently, the ℓ_1-regularized least-squares method failed to provide good estimates of the analytes spectra. This is quantified in Table 1c, where normalized correlation coefficients between pure analytes spectra and spectra of the estimated analytes in Fourier domain are shown. While the accuracy of the estimation of components **1**−**3** generally follows that in Table 1b, component **4** is estimated incorrectly. This is a consequence of a high degree of similarity (correlation factor) in combination with a low number of SAPs detected in the Fourier domain. Therefore, the importance of the wavelet basis for providing sparse representation of the NMR signals is clearly verified.

Finally, significant degrees of correlation between spectra of the pure analytes would cause the ICA-based approach to fail even if the number of mixtures would be equal to the number of analytes. This is due to the fact that significant correlation between spectra of the pure analytes violates the statistical independence assumption required by ICA. This has been demonstrated by using the JADE ICA algorithm[41] to separate the same four analytes but from four mixtures, whereupon normalized correlation coefficients between pure and estimated analytes spectra are shown in Table 1d.

As discussed previously, only a few methods or algorithms have been developed for the extraction of analytes from multicomponent spectral data without any known a priori information. The majority of these blind decomposition methods require the number of mixtures to be greater than or equal to the, in principle unknown, number of pure components. Some of these methods, like the band-target entropy minimization (BTEM), have been applied on the extraction of components in 2D NMR (COSY and heteronuclear single quantum coherence (HSQC)) spectroscopy from multicomponent mixtures.[13] However, seven mixtures were used for the reconstruction of three pure components of simple structure. Moreover, as stated by the authors, the BTEM approach is inapplicable when the number of experimentally measured spectra is less than the number of observed components. Here, as well as in recent publications,[16−18] we have demonstrated that the sparseness-based approach successfully estimates pure components when the number of available mixtures is less than the unknown number of components.

Blind Extraction of Five Analytes from Two Mixtures in Mass Spectrometry. The proposed sparseness-based multivariate data analysis method for blind analyte extraction relies on the

[41] Cardoso, J. F.; Soulomiac, A. *Proc. IEE F.* **1993**, *140*, 362–370.

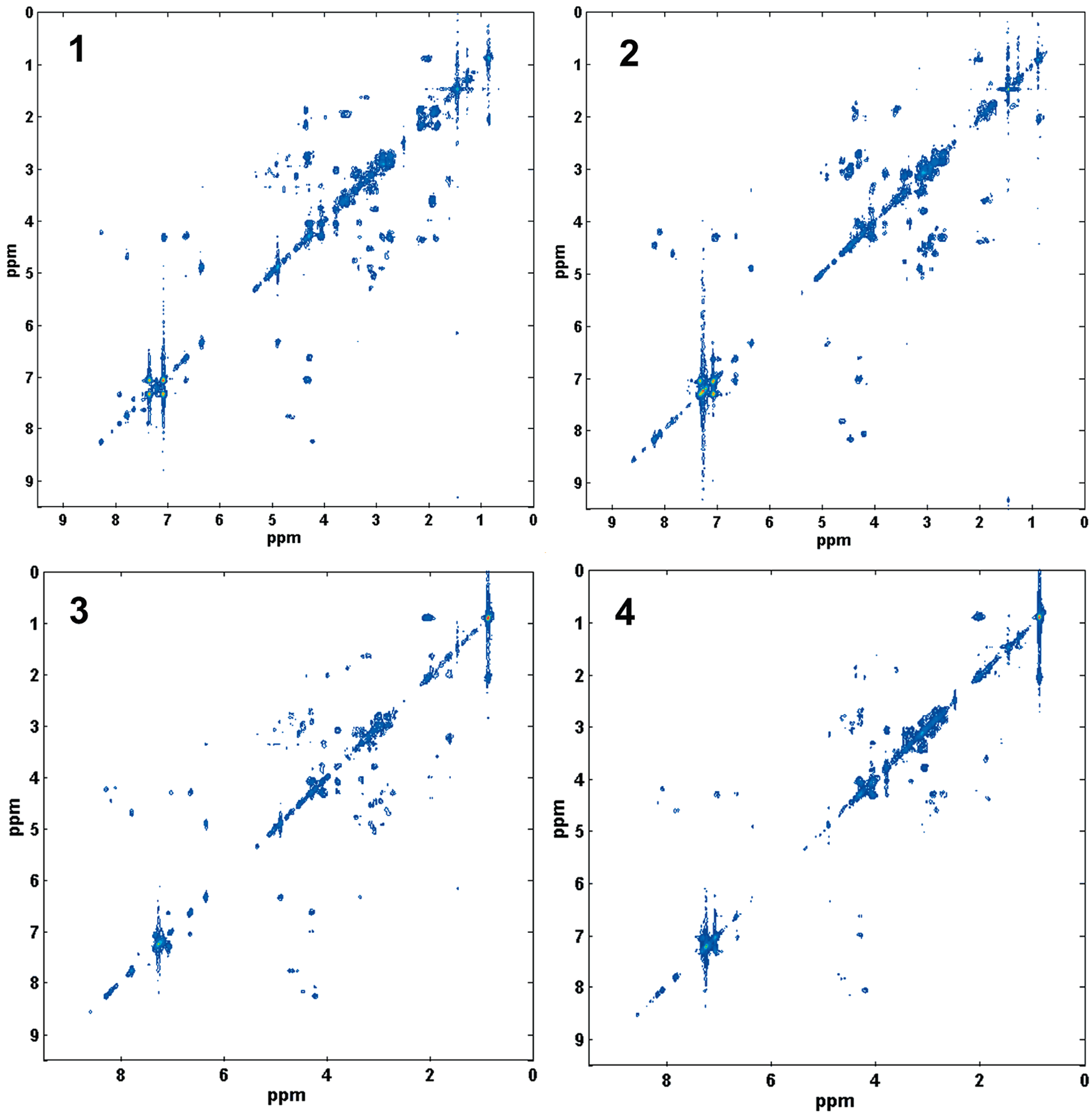

Figure 4. COSY NMR spectra of estimated analytes **1**−**4**.

detection of a large enough set of SAPs in a suitably chosen basis, using direction-based criterion 4. As discussed previously, the direction based criterion (eq 4) requires complex signals. To circumvent this difficulty for the case of real signals, arising, for example, in mass spectrometry or FT-IR spectroscopy, we have proposed the use of the analytic representation (eq 3) of the real signals to detect the positions of the SAPs. We have recently described blind extraction of five pure components mass spectra from only two mixtures by means of sparse component analysis.[17] The structures of the pure components, their mass spectra, and the mass spectra of the two mixtures are available as Supporting Information (Figures S-1, S-2, and S-3, respectively). The same data set was used to validate the sparseness-based multivariate data analysis method proposed herein. With the angular displacement criterion set to $\Delta\theta = 2°$, 290 SAPs were detected using the analytical representation (eq 3). The clustering function (Figure S-4 in the Supporting Information) showed five peaks corresponding to five analytes present in the mixtures. The estimated mass spectra are presented in Figure S-5 in the Supporting Information and are consistent with the results already obtained in ref 17. This is due to the fact that the mass spectra of the analytes were weakly correlated (see Table S-1 in the Supporting Information). However, this validates an approach for the detection of SAPs in the case of real signals, which is based on the use of an analytic representation (eq 3).

This result should be considered in the wider context of the utility of mass spectrometry for metabolic profiling. Chromatographic separation of analytes present in mixtures prior to MS analysis is a standard procedure but suffers from some drawbacks. Different samples (mixtures) require different separation techniques (column packages, mobile phases), and determining optimal conditions for the separation is usually a time- and resource-consuming process.[3] On the other hand, direct infusion of the complex sample into the mass spectrometer is generally not applicable, owing to the ionization suppression and the formation of adducts in the ion source.[1] There are a few successful examples that are, however, limited to the analysis of plant extracts.[1] The presented multivariate data analysis method based on the detection of SAPs can reduce the need for the accurate separation prior to MS analysis and represents an innovative approach for the metabolic analysis based on mass spectrometry. Furthermore, we plan to test this approach on less "controlled" and more biologically relevant experiments to determine possibilities and limitations of the presented multivariate data analysis method.

CONCLUSIONS

We developed and demonstrated a sparseness-based method for blind estimation of analytes exhibiting a high level of complexity and structural similarity, whereupon their number is greater than the number of mixtures available. The method relies on the realistic assumption about the existence of a representation domain or basis where a small number of data sample points can be found at which analytes do not overlap. Although of general importance, the method was developed to solve an important problem in metabolic studies: the blind extraction of analytes from a possibly smaller number of mixtures of NMR or mass spectra. We exemplified the method through the estimation of four analytes from three mixtures in 2D NMR spectroscopy and five analytes from two mixtures in mass spectrometry. The advantages of the proposed sparseness-based approach over the presently used multivariate data analysis methods are expected to be of greatest significance in applications such as metabolic profiling of biological fluids and tissues in search for new biomarkers, analysis of plant and microbial extracts in seeking new biologically active compounds, and the reconstruction of transcription factors in gene regulating networks.

ACKNOWLEDGMENT

A patent is pending under the number PCT/HR2009/00028. The work of I. Kopriva and I. Jerić was, respectively, supported by the Ministry of Science, Education and Sports, Republic of Croatia, under Grants 098-0982903-2558 and 098-0982933-2936. David Smith's help in proofreading the manuscript is also gratefully acknowledged.

SUPPORTING INFORMATION AVAILABLE

Additional information as noted in text. This material is available free of charge via the Internet at http://pubs.acs.org.

Received for review November 18, 2009. Accepted January 22, 2010.

AC902640Y