

A Blind Source Separation Method for Nearly Degenerate Mixtures and Its Applications to NMR Spectroscopy

Yuanchang Sun ^{*}and Jack Xin^{*}

Abstract

In this paper, we develop a novel blind source separation (BSS) method for nonnegative and correlated data, particularly for the nearly degenerate data. The motivation lies in nuclear magnetic resonance (NMR) spectroscopy, where a multiple mixture NMR spectra are recorded to identify chemical compounds with similar structures (degeneracy).

There have been a number of successful approaches for solving BSS problems by exploiting the nature of source signals. For instance, independent component analysis (ICA) is used to separate statistically independent (orthogonal) source signals. However, signal orthogonality is not guaranteed in many real-world problems. This new BSS method developed here deals with nonorthogonal signals. The independence assumption is replaced by a condition which requires dominant interval(s) (DI) from each of source signals over others. Additionally, the mixing matrix is assumed to be nearly singular. The method first estimates the mixing matrix by exploiting geometry in data clustering. Due to the degeneracy of the data, a small deviation in the estimation may introduce errors (spurious peaks of negative values in most cases) in the output. To resolve this challenging problem and improve robustness of the separation, methods are developed in two aspects. One technique is to find a better estimation of the mixing matrix by allowing a constrained perturbation to the clustering output, and it can be achieved by a quadratic programming. The other is to seek sparse source signals by exploiting the DI condition, and it solves an ℓ_1 optimization. We present numerical results of NMR data to show the performance and reliability of the method in the applications arising in NMR spectroscopy.

^{*}Department of Mathematics, University of California at Irvine, Irvine, CA 92697, USA.

1 Introduction

Blind source separation (BSS) is a major area of research in signal and image processing. It aims at recovering source signals from their mixtures without detailed knowledge of the mixing process. Applications of BSS include signal analysis and processing of speech, image, and biomedical signals, especially, signal extraction, enhancement, denoising, model reduction and classification problems [8]. The goal of this paper is to study new BSS methods for nearly degenerate data arising from Nuclear Magnetic Resonance (NMR) spectroscopy. The BSS problem is defined by the following matrix model

$$X = AS, \text{ with } A_{ij} \geq 0, S_{ij} \geq 0, \quad (1.1)$$

where $X \in \mathbb{R}^{m \times p}$, $A \in \mathbb{R}^{m \times n}$, $S \in \mathbb{R}^{n \times p}$. Rows of X represents the measured mixed signals, rows of S are the source signals. The X, S are sampled functions of an acquisition variable which may be time, frequency, position, wavenumber, etc, depending on the underlying physical process. Hence there are p samples in the measurements. The objective of BSS is to solve for A and S given X . In the context of NMR spectroscopy, the mixing coefficients are not typically measured. This is where BSS techniques become useful. The problem is also known as nonnegative matrix factorization (NMF [18]). Similar to factorizing a composite number ($48 = 6 * 8 = 8 * 6 = 4 * 12 = 12 * 4 = 2 * 24 = 24 * 2 = 3 * 16 = 16 * 3$), there are permutation and scaling ambiguities in solutions to BSS. For any permutation matrix P and invertible diagonal matrix Λ , $(APA, \Lambda^{-1}P^{-1}S)$ is another pair equivalent to the solution (A, S) , since

$$X = AS = (APA)(\Lambda^{-1}P^{-1}S). \quad (1.2)$$

Various BSS methods have been proposed relying on *priori* knowledge of source signals such as spatio-temporal decorrelation, statistical independence, sparseness, nonnegativity, etc, [7, 8, 12, 16, 19, 20, 21, 25, 30, 31, 32, 33]. Recently there have been considerable interests for solving nonnegative BSS problems, which emerge in computer tomography, biomedical image processing, NMR spectroscopy [2, 3, 14, 17, 18, 23, 25, 26, 27, 28, 29, 30, 31, 32, 33, 35]. This work is originated from analytic chemistry, in particular, NMR spectroscopy. Applications include identification of organic compounds, metabolic fingerprinting, disease diagnosis, and drug design. As chemical mixtures abound in human organs, for example, blood, urine, and metabolites in brain and muscles. Each compound has a unique spectral fingerprint defined by the number, intensity and locations of its NMR peaks. In drug design, structural information must be isolated from spectra that also contain the target molecule, side products, and impurities.

The different spectra come from Fourier transform of NMR measurement of absorbance of radio frequency radiation by receptive nuclear spins of the same mixture sample at different time segments when exposed to high magnetic fields. The NMR spectra are nonnegative. Besides, NMR spectra of different chemical compounds are usually not independent, especially as compounds (component molecules) have similar functional groups, the peaks overlap in the composite NMR spectra making it difficult to identify the compounds involved. ICA-type approaches recover independent source

signals and thus are unable to separate NMR source spectra. New methods need to be invented to handle this class of data. Recently nonnegative BSS has been attracted considerable attention in NMR spectroscopy [1, 17, 25, 28, 29, 31, 32, 33, 34, 36, 37]. For example, Naanaa and Nuzillard (NN) proposed a nonnegative BSS method in [25] based on a strict local sparseness assumption of the source signals. The NN assumption (NNA) requires the source signals to be strictly non-overlapping at some locations of acquisition variable (e.g., frequency). In other words, each source signal must have a stand-alone peak where other sources are strictly zero there. Such a strict sparseness condition leads to a dramatic mathematical simplification of a general nonnegative matrix factorization problem (1.1) which is non-convex. Geometrically speaking, the problem of finding the mixing matrix A reduces to the identification of a minimal cone containing the columns of mixture matrix X . The latter can be achieved by linear programming. In fact, NN's sparseness assumption and the geometric construction of columns of A were known in the 1990's [2, 35] in the problem of blind hyper-spectral unmixing, where the same mathematical model (1.1) is used. The analogue of NN's assumption is called pixel purity assumption [6]. The resulting geometric (cone) method is the so called N-findr [35], and is now a benchmark in hyperspectral unmixing. NN's method can be viewed as an application of N-findr to NMR data. It is possible that measured NMR data may not strictly satisfy NN's sparseness conditions, which introduces spurious peaks in the results. Postprocessing methods have been developed to address the resulting errors. Such a study has been performed recently in case of (over)-determined mixtures [30] where it is found that larger peaks in the signals are more reliable and can be used to minimize errors due to lack of strict sparseness.

In this paper, we consider how to separate the data if NN assumption is not satisfied. We are concerned with the regime where source signals do not have stand-alone peaks yet one source signal dominates others over certain intervals of acquisition variable. In other words, a dominant interval(s) condition (DI) is required for source signals. This is a reasonable condition for many NMR spectra. For example, the DI condition holds well in the NMR data which motivated us. The data is produced by the so-called DOSY (diffusion ordered spectroscopy) experiment where a physical sample of mixed chemical compounds in solvent (water) is prepared. DOSY tries to distinguish the chemicals based on variation in their diffusion rates. However, DOSY fails to separate them if the compounds have similar chemical functional groups (i.e., they have similar diffusion rates). In this application, the diffusion rates of the chemicals serve as the mixing coefficients. This presents an additional mathematical challenge due to the near singularity of the mixing matrix. Separating these degenerate data is intractable to the convex cone methods, thus we are prompted to develop new approaches. Examination the DI condition reveals a great deal about the geometry of the mixtures. Actually, the scattered plot of columns of X must contain several clusters of points, and these clusters are centered at columns of A . Hence, the problem of finding A boils down to the identification of the clusters, and it can be accomplished by data clustering, for example, K-means. Although the data clustering in general produces a fairly good estimate of the mixing matrix, its output deviates from the true solution due to the presence of the noise, initial guess of the clustering algorithm, and so on. In the case of nearly singular mixing matrix, a small perturbation can lead to large errors in the source recovery (e.g., spurious peaks). To overcome this

difficulty and improve robustness of the separation, we propose two different methods. One is to find a better estimation of mixing matrix by allowing a constrained perturbation to the clustering output, and it is achieved by a quadratic programming. The intention is to move the estimation closer to the true solution. The other is to seek sparse source signals by exploiting the DI condition. An ℓ_1 optimization problem is formulated for recovering the source signals.

The paper is outlined as follows; In section 2, we shall review the essentials of NN approach, then we propose a new condition on the source signals motivated by NMR spectroscopy data. In section 3, we introduce the method. In section 4, we further illustrate our method with numerical examples including the processing of an experimental DOSY NMR data set. Section 5 is the conclusion. We shall use the following notations throughout the paper. The notation A^j stands for the j -th column of matrix A , S^j for the j -th column of matrix S , X^j the j -th column of matrix X . While S_j and X_j are the j -th rows of matrix S and X , or the j -th source and mixture, respectively.

This work was partially supported by NSF-ADT grant DMS-0911277 and NSF grant DMS-0712881. The authors thank Professor A.J. Shaka and Dr. Hasan Celik for helpful discussions and their experimental NMR data.

2 The method

In the paper, we shall consider the determined case ($m = n$), although the results can be easily extended to the over-determined case ($m > n$). Consider the linear model (1.1) where each column in X represents data collected at a particular value of the acquisition variable, and each row represents a mixture spectrum. In this section, we shall first discuss the briefs of NN method, then introduce the new source conditions and the method.

2.1 NN approach

In [25], Naanaa and Nuzillard (NN) presented an efficient sparse BSS method and its mathematical analysis for nonnegative and partially orthogonal signals such as NMR spectra. Consider the (over)-determined regime where the number of mixtures is no less than that of sources ($m \geq n$), and the mixing matrix A is full rank. In simple terms, NN's key sparseness assumption (referred to as NNA below) on source signals is that each source has a stand-alone peak at some location of the acquisition variable where the other sources are identically zero. More precisely, the source matrix $S \geq 0$ is assumed to satisfy the following condition

Assumption (NNA). : For each $i \in \{1, 2, \dots, n\}$ there exists an $j_i \in \{1, 2, \dots, p\}$ such that $s_{i,j_i} > 0$ and $s_{k,j_i} = 0$ ($k = 1, \dots, i - 1, i + 1, \dots, n$).

Eq. (1.1) can be rewritten in terms of columns as

$$X^j = \sum_{k=1}^n s_{k,j} A^k \quad j = 1, \dots, p, \quad (2.1)$$

where X^j denote the j th column of X , and A^k the k th column of A . Assumption NNA implies that $X^{j_i} = s_{i,j_i} A^i$ $i = 1, \dots, n$ or $A^i = \frac{1}{s_{i,j_i}} X^{j_i}$. Hence Eq. (2.1) is rewritten as

$$X^j = \sum_{i=1}^n \frac{s_{i,j}}{s_{i,j_i}} X^{j_i}, \quad (2.2)$$

which says that every column of X is a nonnegative linear combination of the columns of \hat{A} . Here $\hat{A} = [X^{j_1}, \dots, X^{j_n}]$ is the submatrix of X consisting of n columns each of which is collinear to a particular column of A . It should be noted that j_i ($i = 1, \dots, n$) are not known and have to be computed. Once all the j_i s are found, an estimation of the mixing matrix is obtained. The identification of \hat{A} 's columns is equivalent to identifying a convex cone of a finite collection of vectors. The cone encloses the data columns in matrix X , and is the smallest of such cones. Such a minimal enclosing convex cone can be found by linear programming methods. Mathematically, the following constrained equations are formulated for the identification of \hat{A} ,

$$\sum_{j=1, j \neq k}^p X^j \lambda_j = X^k, \lambda_j \geq 0 \quad k = 1, \dots, p. \quad (2.3)$$

Then any column X^k will be a column of \hat{A} if and only if the constrained equation (2.3) is inconsistent. However, if noises are present, the following optimization problems are suggested to estimate the mixing matrix

$$\begin{aligned} \text{minimize score} &= \left\| \sum_{j=1, j \neq k}^p X^j \lambda_j - X^k \right\|_2, k = 1, \dots, p \\ \text{subject to} & \lambda_j \geq 0. \end{aligned}$$

A score is associated with each column. A column with a low score is unlikely to be a column of \hat{A} because this column is roughly a nonnegative linear combination of the other columns of X . On the other hand, a high score means that the corresponding column is far from being a nonnegative linear combination of other columns. Practically, the n columns from X with highest scores are selected to form \hat{A} , the mixing matrix. The Moore-Penrose inverse \hat{A}^+ of \hat{A} is then computed and an estimate to S is obtained: $\hat{S} = \hat{A}^+ X$. NN method proves to be both accurate and efficient if NNA condition holds. However, if the condition is not satisfied, errors and artifacts may be introduced because the true mixing matrix is no longer the smallest enclosing convex cone of columns of the data matrix.

Recently, the authors have developed postprocessing techniques on how to improve NN results with abundance of mixture data, and how to improve mixing matrix estimation with major peak based corrections [30]. The work in [30] actually considered a relaxed NNA (rNNA) condition

Assumption (rNNA). : For each $i \in \{1, 2, \dots, n\}$ there exists an $j_i \in \{1, 2, \dots, p\}$ such that $s_{i,j_i} > 0$ and $s_{k,j_i} = \epsilon_k$ ($k = 1, \dots, i-1, i+1, \dots, n$), where $\epsilon_k \ll s_{i,j_i}$.

Simply said, each source signal has a dominant peak at acquisition position where the other sources are allowed to be nonzero. NNA condition recovers if all $\epsilon_k = 0$. The rNNA is more realistic and robust than the ideal NNA for real-world NMR data [25].

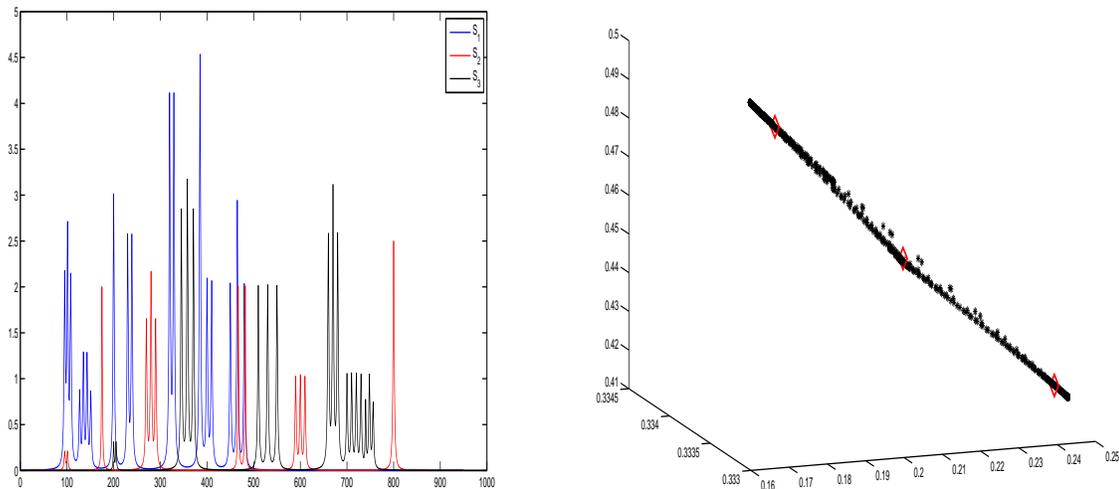


Figure 1: three source signals with dominant intervals (left panel); the geometry of the mixture matrix (right panel). The centers (red diamond) of three clusters are detected by k-means.

2.2 Source assumption and mixing matrix

Motivated by the DOSY NMR spectra, we propose here a different relaxed NN condition on the source signals. Note that the rows S_1, S_2, \dots, S_n of S are the source signals, and they are required to satisfy the following condition: For $i = 1, 2, 3, \dots, n$, source signal S_i is required to have dominant interval(s) over $S_n, \dots, S_{i+1}, S_{i-1}, \dots, S_2, S_1$, while S_i is allowed to overlap with other signals at the rest of the acquisition region. More formally, it implies that source matrix S satisfies the following condition

Assumption. For each $k \in \{1, 2, 3, \dots, n\}$, there is a set $\mathcal{I}_k \subset \{1, 2, \dots, p\}$ such that for each $l \in \mathcal{I}_k$ $s_{kl} \gg s_{jl}, j = 1, 2, \dots, k-1, k+1, \dots, n$.

We shall call this dominant interval condition, or DI condition. Fig. 1 is an idealized example of three DI source signals. In addition to the DI source condition, the mixing matrix is required to be near singular. The motivation is the similar diffusion rates of the chemicals with similar structure. This poses a mathematical challenge to invert a near singular matrix, since a small error in the recovered mixing matrix might lead to a considerable deviation in the source recovery. Among the singularly mixed signals (or degenerate data), in this paper we shall consider the following two types: 1) columns of the mixing matrix are parallel; 2) one column of the mixing matrix is a nonnegative linear combination of others. Case 1 is motivated by NMR of the chemicals with similar diffusion rate. We shall call this condition parallel column condition, or PCC. Case 2 can also be encountered in NMR spectroscopy of chemicals, and we shall call it one column degenerate condition, or OCDC. Please note that both PCC and OCDC should be considered to hold approximately in real-world data.

2.3 Our approach

2.3.1 Data clustering

Now suppose we have a set of nearly degenerate signals from DI sources. We require that compared to the size of dominant interval(s) in the acquisition region, the source signals overlapping region is much smaller. In fact, this is a reasonable assumption for the NMR data which motivates us. More importantly, this requirement enables the success of the clustering method. Next, we shall estimate the columns of mixing matrix A by data clustering. The dominant interval(s) from each of the source signals implies that there is a region where the source S_i dominates others. More precisely, there are columns of X such that

$$X^k = s_{i,k}A^i + \sum_{j=1}^{n-1} o_{j,k}A^j, \quad (2.4)$$

where $s_{i,k}$ dominate $o_{i,k}$ ($i = 1, \dots, n-1$), i.e., $s_{i,k} \gg o_{i,k}$. The identification of A^i ($i = 1, \dots, n$) is equivalent to finding a cluster formed by these X^k 's in \mathbb{R}^n . As illustrated in the geometry plot of X in Fig. 1, three clusters are formed. Many clustering techniques are available for locating these clusters, for example, k-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. We shall use k-means analysis in this paper because it is computationally fast, and easy to implement. Consider an example of three DI source signals with OCDC mixing matrix condition, the three centers are shown in Fig. 1. For real-world data, we show an example of NMR spectra of quinine, geraniol, and camphor mixture in Fig. 2. The clusters in the middle implies that OCDC condition hold well for this data. Apparently, NN method (and other convex cone methods) would fail to separate the source signals due to the degeneracy of the mixing matrix. It might be able to identify two columns of A as the two edges, it by no means can locate A 's degenerate column. For the PCC degenerate case, clustering is also able to deliver a good estimation, even when the data is contaminated by noise. We show the results in Fig. 3 where the three clusters are very close due to the PCC degeneracy. NN solution would deviate considerably from the true solution. For the data we tested, clustering techniques like k-means works well when the condition number of the mixing matrix is up to 10^8 . Though the solutions of mixing matrix by clustering methods are rather good estimation to the true solution, small deviations from the true ones will introduce large errors in the source recovery ($S = \text{inverse}(A) X$). Next we propose two approaches to overcome this difficulty. Both approaches need to solve optimization problems. The first one intends to improve the source recovery by seeking a better mixing matrix, while the second approach reduces the spurious peaks by imposing sparsity constraint on the sources.

2.3.2 Better inverse of the mixing matrix

Suppose the estimation of the mixing matrix by clustering is \hat{A} . Then the source recovery can be obtained $\hat{S} = \text{inverse}(\hat{A}) X$. As discussed above, errors in \hat{S} could be introduced even by a small perturbation in \hat{A} to the ground truth. Negative spurious peaks are produced in most cases, see the Fig. 5 where the negative peaks on the left

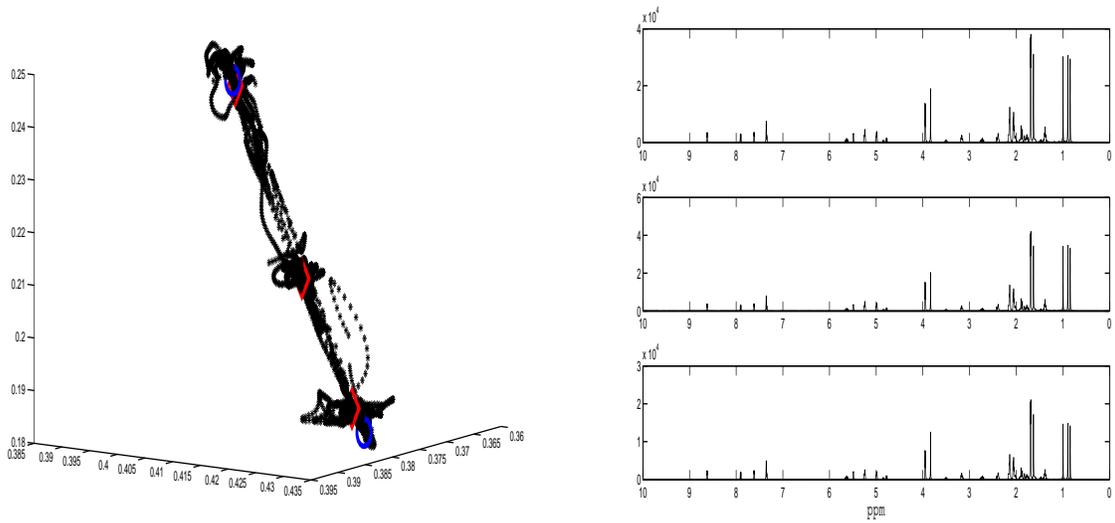


Figure 2: Real data example: three columns of A are identified as the three center points (in red diamond) attracting most points in scatter plots of the columns of X (left), and the three rows of X (right). NN method identifies two columns of A as the points in the blue circle.

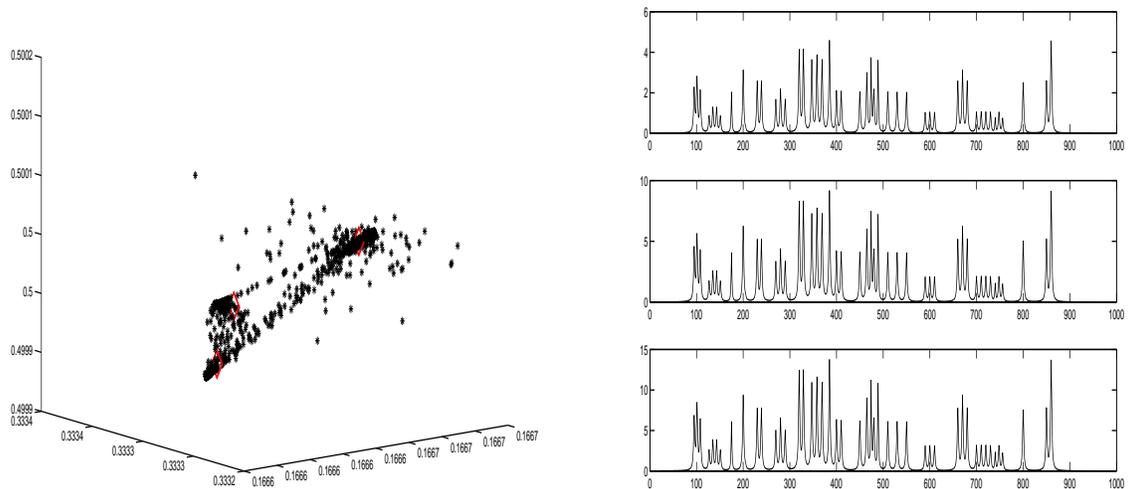


Figure 3: Example of PCC case: three columns of A are identified as the three center points (in red diamond) attracting most points in scatter plots of the columns of X (left), and the three rows of X (right).

plot actually can be viewed as bleed through from another source. Clearly, a better estimation of mixing matrix is required to reduce these spurious peaks. Instead of looking for a better mixing matrix, we propose to solve the following optimization problem for a better inverse of the matrix,

$$\min_B \frac{1}{2} \|I - \hat{A}B\|_2^2 \quad \text{subject to } BX \geq 0, \quad (2.5)$$

where $I \in \mathbb{R}^{m \times m}$ is the identity matrix. The constraint $BX \geq 0$ is used to reduce the negative values introduced in the source recovery. (2.5) is a linearly constrained quadratic program and it can be solved by a variety of methods including interior point, gradient projection, active sets, etc. In this paper, interior point algorithm is used. Once the minimizer B^* is obtained, we solve for the sources by $S = B^* X$.

2.3.3 Sparser source signals

The method proposed above works well for mixing matrix whose condition number is up to 10^8 . If the mixing matrix is much more ill-conditioned, the problem (1.1) becomes under-determined. It appears that solving the equation exactly for S is hopeless even an accurate A is provided. However, a meaningful solution is possible if the actual source signals are structurally compressible, meaning that they essentially depend on a low number of degrees of freedom. Although the source signals (rows of S) are not sparse, the columns of S possess sparsity due to the dominant intervals condition. Hence, we seek the sparsest solution for each column S^i of S as

$$\min \|S^i\|_0 \quad \text{subject to } \hat{A}S^i = X^i, \quad S^i \geq 0. \quad (2.6)$$

Here $\|\cdot\|_0$ (0-norm) represents the number of nonzeros. Because of the non-convexity of the 0-norm, we minimize the ℓ_1 -norm:

$$\min \|S^i\|_1 \quad \text{subject to } AS^i = X^i, \quad S^i \geq 0, \quad (2.7)$$

which is a linear program [11] because S^i is non-negative. The fact that data may in general contain noise suggest solving the following unconstrained optimization problem,

$$\min_{S^i \geq 0} \mu \|S^i\|_1 + \frac{1}{2} \|X^i - AS^i\|_2^2, \quad (2.8)$$

for which Bregman iterative method [15, 38] with a proper projection onto non-negative convex subset can be used to obtain a solution. Under certain conditions of matrix A , it is known [5, 39] that solution of ℓ_1 -minimization (2.8) gives the exact recovery of sufficiently sparse signal, or solution to (2.6), [5, 39]. Though our numerical results support the equivalence of ℓ_1 and ℓ_0 minimizations, the mixing matrix A does not satisfy the existing sufficient conditions [5, 39].

3 Numerical experiments

In this section, we report the numerical examples solved by the method. We compute three examples. The data of the first two examples are synthetic, while the third

example uses real NMR data. In the first example, two sources are to be separated from two mixtures. The mixtures are constructed from two real NMR source signal by simulating the linear model (1.1). The two columns of mixing matrix are nearly parallel, and its condition number is about 1.25×10^8 . The true mixing matrix A , its estimation \hat{A} via clustering, and the improved estimate A_p by solving (2.5) are (for ease of comparison, the first rows of \hat{A}, A_p are scaled to be same as that of A)

$$A = \begin{pmatrix} 0.894427190999916 & 0.894427182055644 \\ 0.447213595499958 & 0.447213613388501 \end{pmatrix},$$

$$\hat{A} = \begin{pmatrix} 0.894427190999916 & 0.894427182055644 \\ \mathbf{0.447213596792237} & \mathbf{0.447213596447341} \end{pmatrix},$$

$$A_p = \begin{pmatrix} 0.894427190999916 & 0.894427182055644 \\ \mathbf{0.447213595582167} & \mathbf{0.447213613388502} \end{pmatrix}.$$

Clearly A_p is a better estimate. The mixtures are plotted in Fig. 4, and the results are presented in Fig. 5.

In the second example, three sources are to be separated from three mixtures. The mixing matrix satisfies the OCDC condition, i.e., one of its columns is a nonnegative linear combination of the other two. To test the robustness of the method, we added Gaussian noise (SNR = 60 dB) to the data. The mixtures and their geometric structure are plotted in Fig. 6. First the data clustering was used to obtain an estimation of the mixing matrix, then an ℓ_1 optimization problem is solved to retrieve the sources. The results are shown in Fig. 7. It can be seen that the recovered sources agree well with the ground truth.

For the third example, we provide a set of real data to test our method. The data is produced by diffusion ordered spectroscopy (DOSY) which is an NMR spectroscopy technique used by chemists for mixture separation [22]. However, the three compounds used in the experiment (quinine, geraniol, and camphor) have similar chemical functional groups (i.e. there is overlapping in their NMR spectra) [24], for which DOSY fails to separate them. It is known that each of the three sources has dominant interval(s) over others in its NMR spectrum. This can also be verified from the three isolated clusters formed in their mixed NMR spectra (see the geometry of their mixtures in Fig. 8). Here we separate three sources from three mixtures. Fig. 8 plots the mixtures (rows of X) and their geometry (columns of X) where three clusters of points can be spotted. Then the columns of A are identified as the center points of three clusters. The solutions are presented in Fig. 9, the results are satisfactory comparing with the ground truth. As a comparison, the source signals recovered by NN [25] is shown in Fig. 10 where $S = \text{inverse}(A) X$, here the inverse is Moore-Penrose (the least squares sense) pseudo-inverse which produces some negative (erroneous) peaks in S .

4 Conclusion

This paper presented novel methods to retrieve source signals from the nearly degenerate mixtures. The motivation comes from NMR spectroscopy of chemical compounds with similar diffusion rates. Inspired by the NMR structure of these chemicals, we

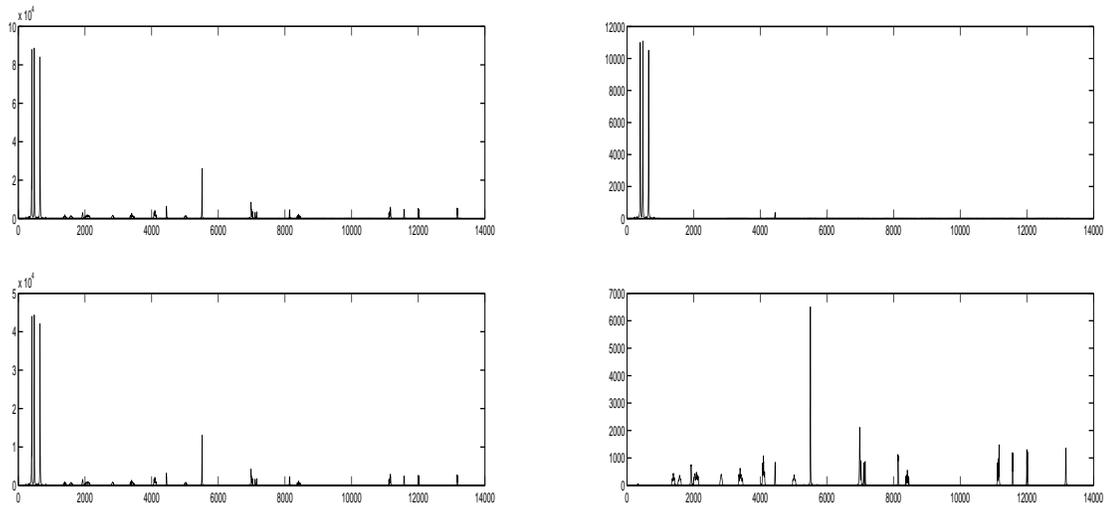


Figure 4: recovered sources by clustering (left column) and the ground truth (right column).

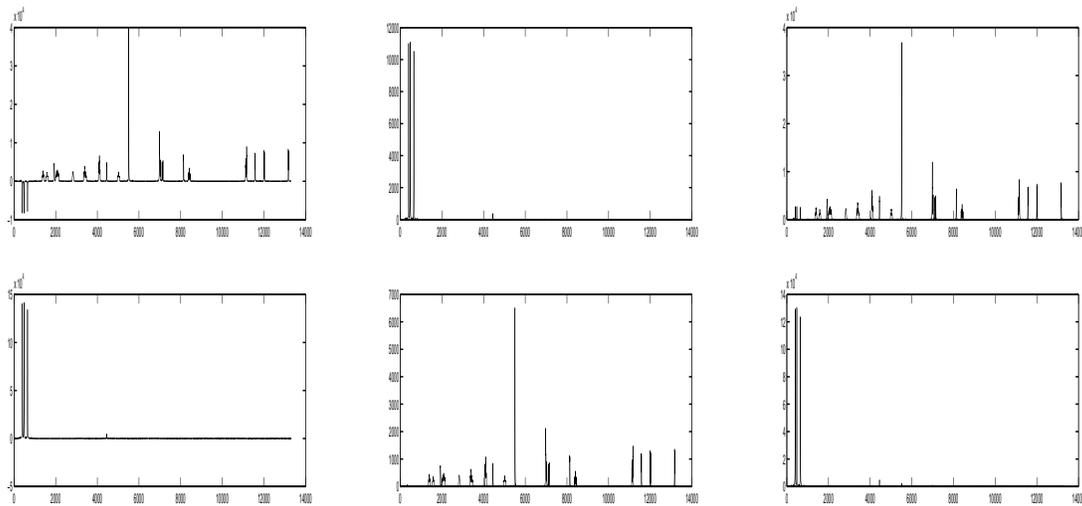


Figure 5: recovered sources by clustering (left column), the ground truth (middle column), and the improved results by a better estimate of mixing matrix (right column).

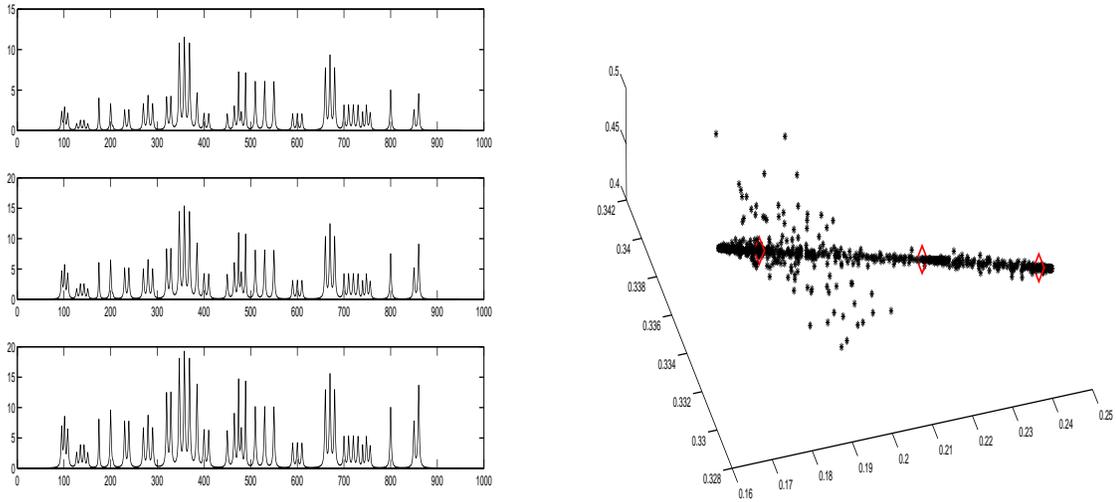


Figure 6: three mixtures (left column) and their scattered plot (right column).

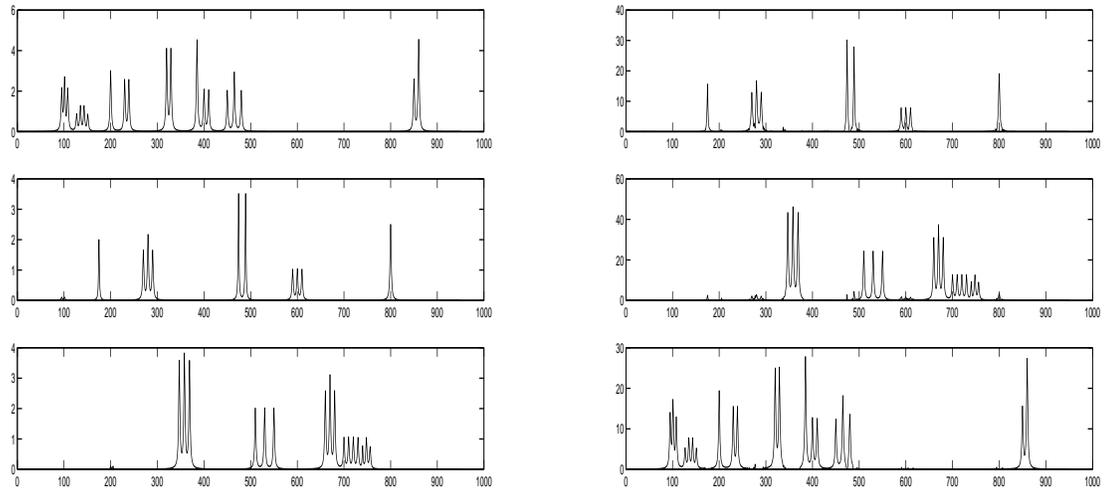


Figure 7: three sources (left column) and their recovery by clustering and ℓ_1 minimization (right column).

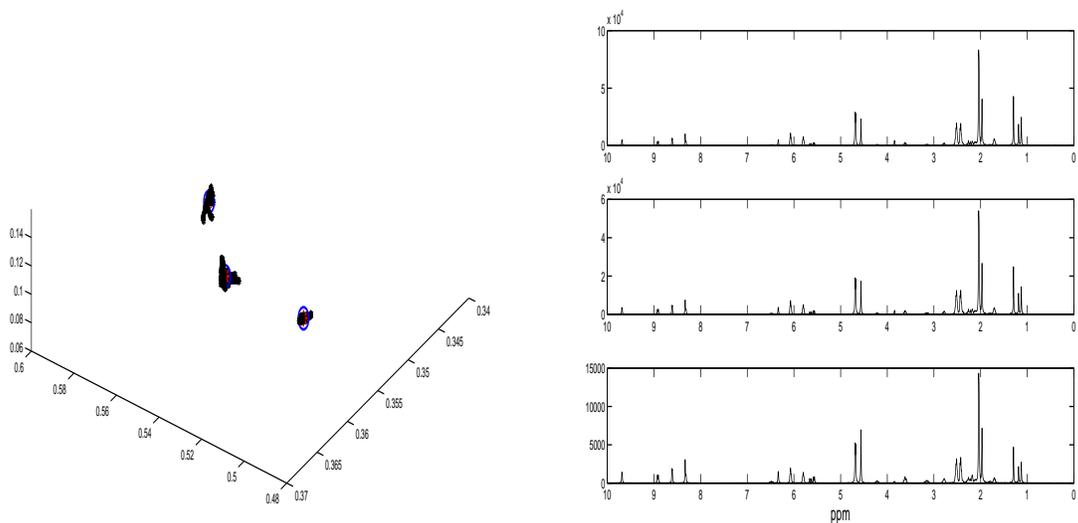


Figure 8: three columns of A are identified as the three center points in blue circles attracting most points in scatter plots of the columns of X (left), and the three rows of X (right).

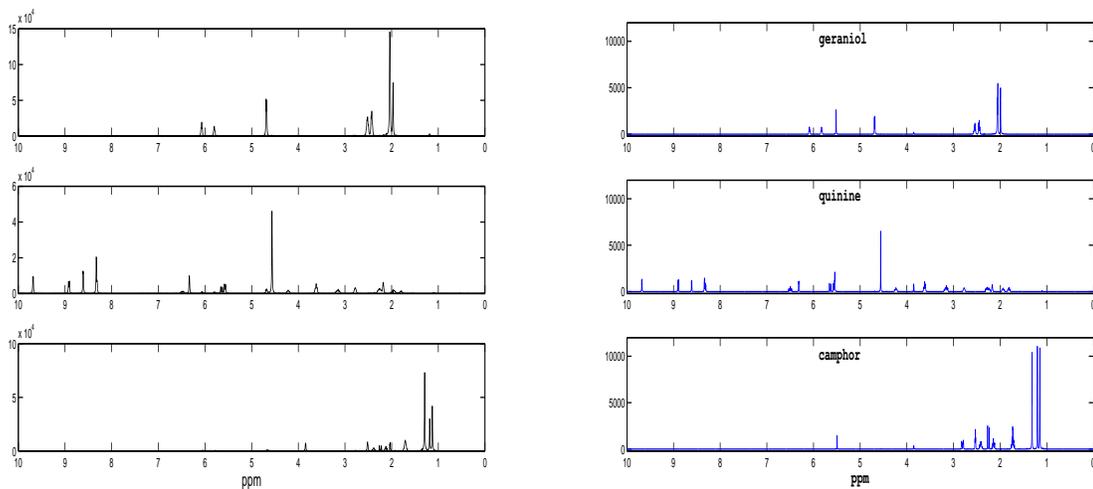


Figure 9: the recovered source signals by nonnegative ℓ_1 (left) and the ground truth (right).

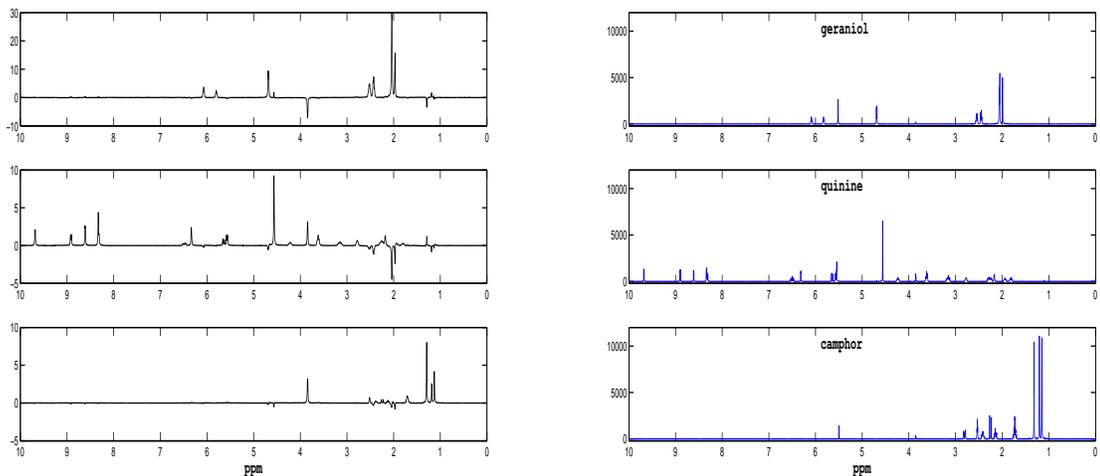


Figure 10: the recovered source signals using NN method (left) and the ground truth (right).

propose a viable source condition which requires dominant interval(s) from each source signal over the others. This condition is well suited for many real-life signals. Besides, the nearly degenerate mixtures are assumed to be generated from the following two types of mixing processes: 1) all the columns of the mixing matrix are parallel; 2) One column of the mixing matrix is the nonnegative linear combination of others. We first use data clustering to identify the mixing matrix, then we develop two approaches to improve source signals' recovery. The first approach minimizes a constrained quadratic program for a better mixing matrix, while the second method seeks the sparsest solution for each column of the source matrix by solving an ℓ_1 optimization. Numerical results on NMR spectra data show satisfactory performance of our method and offer promise towards understanding and detecting complex chemical spectra. Though the methods are motivated by the NMR spectroscopy, the underlying ideas may be generalized to different data sets in other applications.

References

- [1] R. Barton, J. Nicholson, P. Elliot, and E. Holmes, *High-throughput 1H NMR-based metabolic analysis of human serum and urine for large-scale epidemiological studies: validation study*, Int. J. Epidemiol. 37(2008)(suppl 1)pp. i31–i40.
- [2] J. Boardman, *Automated spectral unmixing of AVIRIS data using convex geometry concepts*, in Summaries of the IV Annual JPL Airborne Geoscience Workshop, JPL Pub. 93-26, Vol. 1, 1993, pp 11-14.
- [3] P. Bofla and M. Zibulevsky, *Underdetermined blind source separation using sparse representations*, Signal Processing, 81 (2001) pp. 2353–2362.
- [4] J. Cai, S. Osher, and Z. Shen, *Linearized Bregman iterations for compressed sensing*, UCLA CAM Report, vol. 08, no. 06, 2008.

- [5] E. Candés, J. Romberg, and T. Tao, *Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information*. IEEE Trans. Inform. Theory, 52 (2006) pp. 489–509.
- [6] C-I Chang, ed., “Hyperspectral Data Exploitation: Theory and Applications”, Wiley-Interscience, 2007.
- [7] S. Choi, A. Cichocki, H. Park, and S. Lee, *Blind source separation and independent component analysis: A review*, Neural Inform. Process. Lett. Rev., 6 (2005), pp. 1–57.
- [8] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*, John Wiley and Sons, New York, 2005.
- [9] P. Comon, *Independent component analysis—a new concept?*, Signal Processing, 36 (1994) pp. 287–314.
- [10] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications*, Academic Press, 2010.
- [11] D. Donoho and J. Tanner, *Sparse nonnegative solutions of underdetermined linear equations by linear programming*, Proc Natl Acad Sci USA, 102 (2005) pp. 9446–9451.
- [12] I. Drori, *Fast ℓ_1 minimization by iterative thresholding for multidimensional NMR spectroscopy*, EURASIP Journal on Advances in Signal Processing, 2007, No. 23, pp 1-10 (doi:10.1155/2007/20248).
- [13] R. Ernst, G. Bodenhausen, and A. Wokaun, *Principles of Nuclear Magnetic Resonance in One and Two Dimensions*, Oxford University Press, 1987.
- [14] P. Georgiev, F. Theis, and A. Cichocki, *Sparse component analysis and blind source separation of underdetermined mixtures*, IEEE Transactions on Neural Networks, 16(4) (2005) pp. 992–996.
- [15] Z. Guo and S. Osher, *Template matching via ℓ_1 minimization and its application to hyperspectral target detection*, Tech. Rep. 09-103, UCLA, www.math.ucla.edu/applied/cam/, 2009.
- [16] A. Hyvärinen, J. Karhunen and E. Oja, *Independent Component Analysis*, John Wiley and Sons, New York, 2001.
- [17] I. Koprivaa, I. Jerić, and V. Smrečki, *Extraction of multiple pure component 1H and ^{13}C NMR spectra from two mixtures: Novel solution obtained by sparse component analysis-based blind decomposition*, Analytica Chimica Acta, 653 (2009) pp. 143–153.
- [18] D. D. Lee and H. S. Seung, *Learning of the parts of objects by non-negative matrix factorization*, Nature, 401 (1999) pp. 788–791.

- [19] J. Liu, J. Xin, Y-Y Qi, *A Dynamic Algorithm for Blind Separation of Convolutional Sound Mixtures*, Neurocomputing, 72(2008), pp 521-532.
- [20] J. Liu, J. Xin, Y-Y Qi, *A Soft-Constrained Dynamic Iterative Method of Blind Source Separation*, SIAM J. Multiscale Modeling Simulations, Vol. 7, No. 4, pp 1795-1810, 2009.
- [21] J. Liu, J. Xin, Y-Y Qi, F-G Zeng, *A Time Domain Algorithm for Blind Separation of Convolutional Sound Mixtures and L-1 Constrained Minimization of Cross Correlations*, Comm. Math Sci, Vol. 7, No. 1, 2009, pp 109-128.
- [22] G. Morris, *In Encyclopedia of Nuclear Magnetic Resonance*, D. Grant, R. Harris, Eds, John Wiley & Sons, 2002.
- [23] S. Moussaouia, H. Hauksdóttir, F. Schmidt, C. Jutten, J. Chanussot, D. Brie, S. Douté, and J.A. Benediktsson, *On the decomposition of Mars hyperspectral data by ICA and Bayesian positive source separation*, Neurocomputing, 71(2008), pp 2194–2208.
- [24] M. Nilsson, M. Connel, A. Davies, and G. Morris, *Biexponential Fitting of Diffusion-Ordered NMR Data: Practicalities and Limitations*, Analytical Chemistry, 78 (2006), pp. 3040–3045.
- [25] W. Naanaa and J.-M. Nuzillard, *Blind source separation of positive and partially correlated data*, Signal Processing 85 (9) (2005), pp. 1711–1722.
- [26] D. Nuzillard, S. Bourgb and J.-M. Nuzillard, *Model-Free analysis of mixtures by NMR using blind source separation*, J. Magn. Reson. 133 (1998) pp. 358–363.
- [27] M. Plumbley, *Conditions for non-negative independent component analysis*, IEEE Signal Processing Letters, 9 (2002) pp. 177–180.
- [28] M. Plumbley, *Algorithms for nonnegative independent component analysis*, IEEE Transactions on Neural Networks, 4(3) (2003) pp. 534–543.
- [29] K. Stadlthanner, A. Tom, F. Theis, W. Gronwald, H.-R. Kalbitzer, and E. Lang, *On the use of independent analysis to remove water artifacts of 2D NMR Protein Spectra*, In Proc. Bioeng'2003 (2003).
- [30] Y. Sun, C. Ridge, F. del Rio, A.J. Shaka and J. Xin, *Postprocessing and Sparse Blind Source Separation of Positive and Partially Overlapped Data*, Signal Processing 91 (2011) pp. 1838–1851
- [31] Y. Sun and J. Xin, *Unique Solvability of Under-Determined Sparse Blind Source Separation of Nonnegative and Partially Overlapped Data*, IASTED International Conference on Signal and Image Processing, 710-017, August 23–25, 2010, Hawaii, USA.
- [32] Y. Sun and J. Xin, *Nonnegative Sparse Blind Source Separation for NMR Spectroscopy by Data Clustering, Model Reduction, and ℓ_1 Minimization*, preprint.

- [33] Y. Sun and J. Xin, *A Recursive Sparse Blind Source Separation Method and its Application to Correlated Data in NMR Spectroscopy of Biofluids*, J. Sci Computing, to appear.
- [34] C. Vitols and A. Weljie, *Identifying and Quantifying Metabolites in Blood Serum and Plasma*, Chenomx Inc., 2006.
- [35] M.E. Winter, *N-findr: an algorithm for fast autonomous spectral endmember determination in hyperspectral data*, in Proc. of the SPIE, vol. 3753, 1999, pp 266-275.
- [36] W. Wu, M. Daszykowski, B. Walczak, B.C. Sweatman, S. Connor, J. Haselden, D. Crowther, R. Gill, M. Lutz, *Peak alignment of urine NMR spectra using fuzzy warping*, J. Chem. Inf. Model., 46(2006) pp. 863–875.
- [37] W. Yang, Y. Wang, Q. Zhou, and H. Tang, *Analysis of human urine metabolites using SPE and NMR spectroscopy*, Sci China Ser B-Chem, 51(2008) pp. 218–225.
- [38] W. Yin, S. Osher, D. Goldfarb, J. Darbon, *Bregman iterative algorithm for ℓ_1 -minimization with applications to compressive sensing*, SIAM J. Image Sci, 1(143), pp 143-168, 2008.
- [39] Y. Zhang, *Theory of compressive sensing via L_1 -Minimization: A Non-RIP analysis and extensions*, Technical report, 2009, Rice University.