Proceedings of the 2005 IEEE
Engineering in Medicine and Biology 27th Annual Conference
Shanghai, China, September 1-4, 2005

Recovery of Metabolomic Spectral Sources using Non-negative Matrix Factorization

Shuyan Du[1], Paul Sajda[1], Radka Stoyanova[2], Truman. R. Brown[1]
[1] Department of Biomedical Engineering, Columbia University, New York, NY, USA 10027
[2] Fox Chase Cancer Center, Philadelphia, PA, USA 19111

Abstract- **^{1}H magnetic resonance spectra (MRS) of biofluids contain rich biochemical information about the metabolic status of an organism. Through the application of pattern recognition and classification algorithms, such data have been shown to provide information for disease diagnosis as well as the effects of potential therapeutics. In this paper we describe a novel approach, using non-negative matrix factorization (NMF), for rapidly identifying metabolically meaningful spectral patterns in ^{1}H MRS. We show that the intensities of these identified spectral patterns can be related to the onset of, and recovery from, toxicity in both a time-related and dose-related fashion. These patterns can be seen as a new type of biomarker for the biological effect under study. We demonstrate, using k-means clustering, that the recovered patterns can be used to characterize the metabolic status of the animal during the experiment.**

I. Introduction

Metabolomics [1] quantitatively measures the dynamic metabolic response of living systems to pathophysiological stimuli or genetic modification. Metabolomic analysis of biofluids based on high-resolution MRS and chemometric methods are valuable in characterizing the biochemical response to toxicity [2]. Interpretation of high-resolution ^{1}H biofluid NMR spectra dataset is challenging, specifically for traditional peak-quantifying techniques: a typical dataset consists of at least tens of highly correlated spectra, with thousands of partially overlapping peaks arising from hundreds of endogenous molecules. This has created the need for approaches that can analyze the entire dataset simultaneously for discriminating between different combinations of metabolites, including their dynamic changes.

Previous work by our group has demonstrated that sophisticated multivariate analysis techniques can be used to analyze large, high dimensional datasets [3]. However, some of these techniques, for example Bayesian spectral decomposition (BSD) [4], though capable of identifying a series of spectral features inherent in the observed MRS, are computationally intensive and can require hours to obtain the final solutions.

In this paper we consider the utility of a very fast algorithm called non-negative matrix factorization (NMF) [5] to blindly recover metabolomic spectral sources. Earlier work has demonstrated that NMF can be successfully applied to chemical shift imaging (CSI) of brain (both ^{31}P and ^{1}H) given the validity of a spatial linear mixing model [6-8]. In addition, the NMF algorithm also leads to realistic recovery of source spectra and, compared to BSD, can achieve almost identical fidelity of spectral recovery while requiring only 10^{-4} computational time. We apply the NMF algorithm to simultaneously recover source spectral patterns and their magnitudes from a MRS dataset. We then use k-means clustering [9] to demonstrate how the recovered patterns can be used to characterize the status of the animal during the experiment.

II. Methods

A. Linear mixture problem

Since the amplitudes of metabolites in the recovered sources are proportional to their concentration/abundance in the samples, we seek to model the observations $\mathbf{X}$ as a linear mixture with additive noise,

$$\mathbf{X} = \mathbf{AS} + \mathbf{N} \quad , \tag{1}$$

where the rows of $\mathbf{S}$ are constituent spectral patterns and the columns of $\mathbf{A}$ are the magnitudes of each constituent spectral pattern contributing to the individual observed spectra. The mixing matrix $\mathbf{A}$ has M columns (one for each constituent spectrum) and N rows (one for each sample). $\mathbf{X}$ and $\mathbf{S}$ have L columns (one for each resonance frequency). Given this model, the challenge is to estimate $\mathbf{A}$ and $\mathbf{S}$ simultaneously, given only $\mathbf{X}$.

Since we interpret $\mathbf{A}$ as magnitudes, we can assume its elements to be non-negative. In addition, since the constituent spectra, $\mathbf{S}$, represent amplitudes of resonances, in theory the smallest resonance amplitude is zero, corresponding to the absence of signal at that frequency. The factorization of Equation 1 is therefore constrained by,

$$\mathbf{A} \geq \mathbf{0}, \ \mathbf{S} \geq \mathbf{0} \quad . \tag{2}$$

B. Non-negative matrix factorization

The NMF algorithm has been previously described [5-8]. The basic idea is to construct a gradient descent over an objective function that optimizes $\mathbf{A}$ and $\mathbf{S}$, and by appropriately choosing gradient step sizes, to convert the additive update rules to multiplicative ones. With the objective function,

$$F = \min_{\mathbf{A},\mathbf{S}} \left\| \mathbf{X} - \mathbf{AS} \right\|^2 \quad , \tag{3}$$

the update rules of NMF are constructed as,

$$A_{i,m} \leftarrow A_{i,m} \frac{\left(\mathbf{XS}^{\mathrm{T}}\right)_{i,m}}{\left(\mathbf{ASS}^{\mathrm{T}}\right)_{i,m}}$$
$$S_{m,\lambda} \leftarrow S_{m,\lambda} \frac{\left(\mathbf{A}^{\mathrm{T}}\mathbf{X}\right)_{m,\lambda}}{\left(\mathbf{A}^{\mathrm{T}}\mathbf{AS}\right)_{m,\lambda}} \tag{4}$$

Thus with observations and initializations both non-negative, recovery of $\mathbf{A}$ and $\mathbf{S}$ can be guaranteed non-negative. With $\mathbf{N}$ modeled as Gaussian noise (a reasonable assumption for metabolomic MRS data), it is equivalent to formulating the problem of recovering $\mathbf{A}$ and $\mathbf{S}$ as a maximum likelihood (ML) estimation [7].

Important to consider is the dimensionality of the matrices, namely choosing M, the number of sources to recover, since the factorization in Equation 1 includes an explicit subspace reduction from a N dimensional space into a constrained M dimensional space. Such a compression or "bottleneck" has been shown useful in having the subspace capture statistical regularities in the data [10][11]. In this paper we use principal component analysis (PCA) [12][13] to estimate M.

C. Classification of animal status by k-means clustering

K-means clustering [9] is a simple clustering method that partitions the input data into exactly k clusters. The solution that k-means clustering reaches often depends on the initialization, potentially leading to local minima. However this can be overcome by taking the best solution from multiple random starts. In this paper, we apply k-means clustering to the magnitudes of the recovered patterns, using the correlations between points as the distance metric and 100 random starts to find a "globally" optimal solution. The cluster results are used to characterize the status of the animal(s) during the experiment—either characterizing as "normal" or "abnormal", as a function of time. The "silhouette value" [14] is used to determine whether $k = 2$ is an appropriate estimate of the number of clusters. The silhouette value for each point is a measure of how similar that point is to points in its own cluster compared to points in other clusters. Its value ranges from +1, indicating points that are very distant from neighboring clusters, through 0, indicating points that are not distinctly in one cluster or another, to -1, indicating points that are likely assigned to the wrong cluster. It is defined as

$$sil(i) = (\min_k b(i,k) - a(i)) / \max(a(i), \min_k b(i,k)) \ , \tag{5}$$

where $a(i)$ is the average distance from the i^{th} point to the other points in its cluster, and $b(i,k)$ is the average distance from the i^{th} point to points in another cluster k.

III. Results

We apply NMF to ^{1}H NMR spectra of urine from Han Wistar rats in a hydrazine experiment [15]. Samples were collected from control rats and those treated with three different doses of hydrazine (75, 90, 120 mg/kg) over a period of 150 hours. Preprocessing, including normalization of the data, has been described elsewhere [3].

A. NMF source spectra recovery

PCA analysis is applied to estimate the number of sources M. The scree plot of the first ten normalized principal components (PCs) from the spectral region between 2.20 and 3.63 ppm is shown in Figure 1(a). Figure 1(b) shows the first four PCs and Figure 1(c) their corresponding coefficients. From the PCA analysis we can see $M = 4$ is a reasonable estimate of the number of underlying sources, with the first four PCs account for more than 95% of the variance in the dataset. Solutions for $M = 4$ are presented in Figure 2. We use the following convergence rule,

$$(\chi^{(k)} - \chi^{(k+1)}) / \chi^{(k)} < 10^{-8} \quad , \tag{6}$$

where $\chi = \left\| \mathbf{X} - \mathbf{AS} \right\|$, and the NMF algorithm requires about 300 seconds (Intel Pentium4 1.2GHz) to obtain the recovered spectral sources. The magnitudes in each dose-group, as a function of time, are presented in Figure 2(a), with the identified spectral patterns in Figure 2(b). NMF was run 100 times (100 independent initializations) with Figure 2(b) showing the mean results (solid lines) together with ± 2 standard deviation (std) (dash lines). The small variance demonstrates the robustness and fidelity of the NMF in spectral pattern recovery.

Clear is the association of the four spectral patterns with the hydrazine treatment. In control rats, the first (filled diamonds, ◆) and second (filled upper-triangle,▲) spectral sources maintain almost a constant high level while the third (inverted-triangle,▽) and fourth (open circle,○) are very low. Thus the first spectral source (Krebs cycle intermediates: citrate and succinate) and second spectral source (2-oxoglutarate) are related to the "normal" patterns, while the third and fourth (2-aminoadipic acid, taurine and creatine) - to hydrazine. Indeed, in the treated animals, the "normal" patterns decrease in response to hydrazine and recover after 36 hours, while the other two exhibit reciprocal behaviors during the course of the experiment. The data from the 120mg/kg dose indicates no sign of recovery at 56h, at which point the animal was sacrificed [15].

B. K-means clustering of samples

We apply k-means clustering to the amplitudes in the matrix $\mathbf{A}$ to classify the metabolic status ("normal" vs. "abnormal") of the rats as a function of time. The silhouette value is measured to check the estimate of the number of clusters. K-means clustering was run 100 times, each with a different random initialization, to avoid local minima. The final "globally" optimal results for clustering the samples

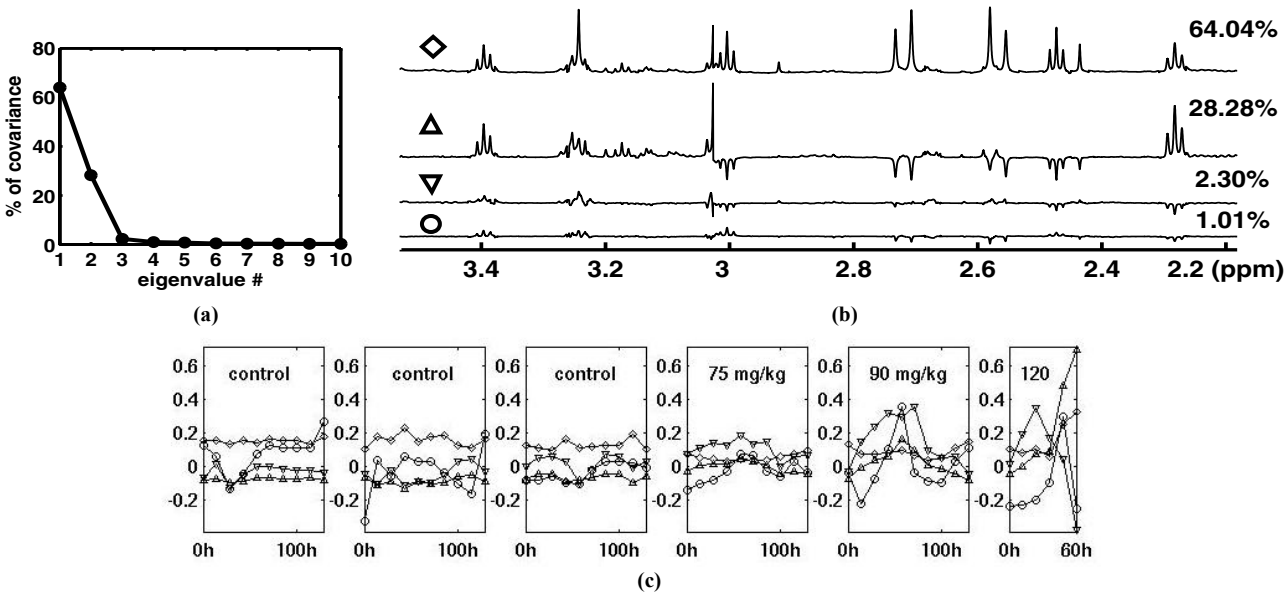

Figure 1. PCA analysis on the dataset. (a) A scree plot of the first ten normalized eigenvalues from the spectral region between 2.20 and 3.63 ppm. (b) The first four principal components (PCs). (c) Coefficients of the first four PCs with markers corresponding to those in (b).

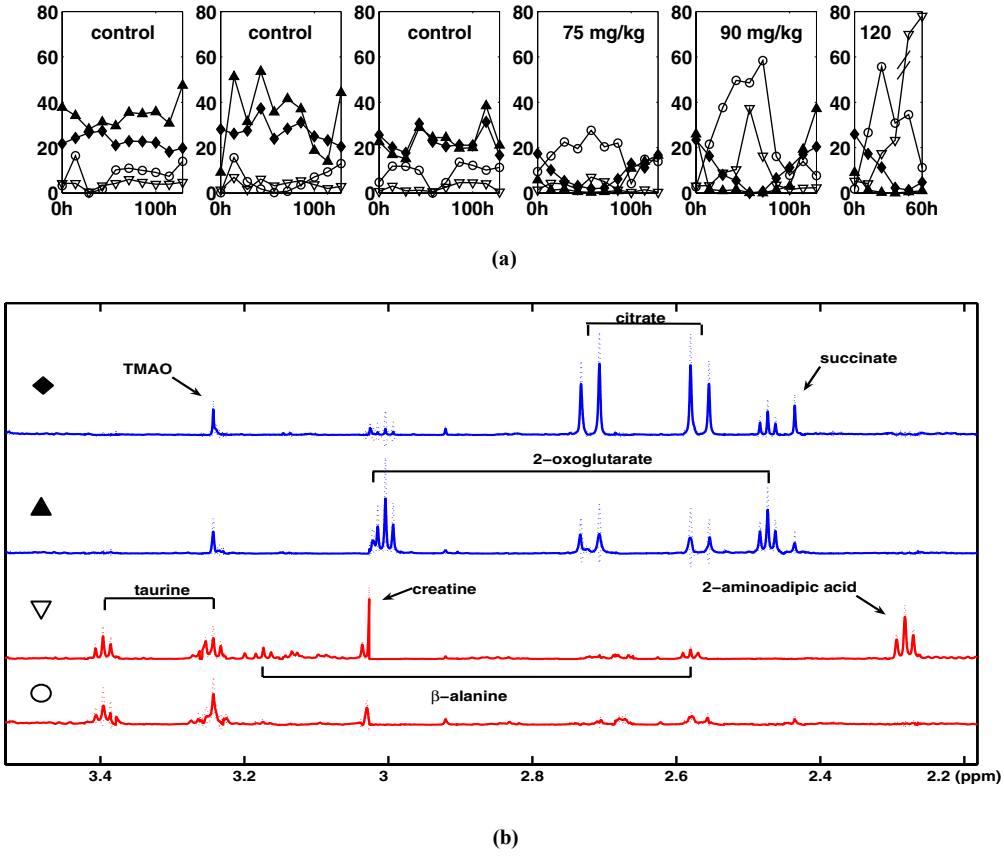

Figure 2. NMF recovery results with $M = 4$. (a) Magnitude of the recovered spectral patterns obtained by NMF for $M = 4$ in the rats as function of time. The filled diamonds (◆) and filled upper-triangles (▲) are associated with split normal patterns, and the inverted-triangles (▽) and open circles (○) are associated with aberrant patterns. (b) Shown are the corresponding spectral patterns marked by the corresponding markers in (a), where the solid lines are the mean results and the dash lines are (mean $\pm$2std). The spectra associated with the "normal" patterns are plot in blue while the ones associated with the "aberrant" patterns are in red.

into two clusters, "normal" and "abnormal", are shown in Figure 3(a), from which we can see that the control rats are clearly separated from those that are treated. Both the initial measurements (0h), taken prior to hydrazine administration, and the later data points (after 104h) for the treated rats are correctly assigned to the "normal" cluster. These samples have NMR spectra very similar to those from untreated animals and in fact correspond to time points when the manifested toxic effect of hydrazine is almost minimized by biologic recovery. The mean silhouette value of $k = 2$ is 0.82, which shows it is a reasonable estimate of the number of clusters. Figure 3(b) shows the classification results using the coefficients of the first four PCs. Clearly, these results are less realistic compared to Figure 3(a), since some of the time points for the control rats are classified into "abnormal". We see that a source recovery method, which imposes physically realistic constraints, can improve classification.

IV. Conclusion

In this paper we demonstrate the potential of the NMF algorithm, when applied in combination with correct and rigorous preprocessing of the data, to identify underlying constituent spectral patterns within a large, high-dimensional and complex MRS dataset. We connect the recovered sources, quantitatively, with the biological end-point measurements, as well as use them to accurately characterize the metabolic status of the animals. This approach shows promise for understanding complex metabolic responses from disease, pharmaceuticals, and toxins. NMF is shown to be very fast, thus enabling a rapid assessment of the metabolomic spectra with respect to known biochemical pathways and the basic toxicological processes.

Acknowledgment

This research was supported by an NSF CAREER Award BES-0133804 and NIH R33DK070301. We thank John C. Lindon and Jeremy K. Nicholson for the hydrazine/urine data.

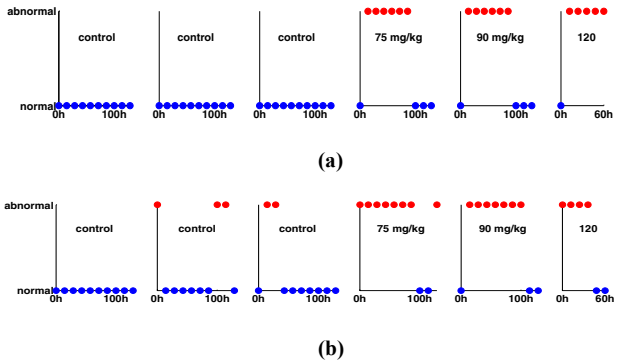

Figure 3. K-means cluster analysis applied to the amplitudes of the NMF patterns for $M = 4$ and the first four PCs. (a) K-means clustering on amplitudes of NMF patterns for $M = 4$. Samples are clustered into 2 groups: "normal" and "abnormal". The samples corresponding to the control rats and the ones collected before hydrazine administration, as well as after more than 104 hours after hydrazine administration for the treated rats are assigned into the "normal" cluster, and the other samples collected in the experiment are correctly assigned into the "abnormal" cluster. (b) K-means clustering on the first four PCs. Comparing (a) and (b) it is clear that patterns recovered by NMF greatly improve the characterization of animal status during the experiments, as evidenced in (b) by misclassification of some of the time points for controls.

References

[1] J. K. Nicholson, J. C. Lindon, E. Holmes, Metabonomics: understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data, *Xenobiotica*, 29(11), 1181-1189, 1999.

[2] H. C. Keun, T. M. Ebbels, H. Antti, M. E. Bollard, O. Beckonert, G. Schlotterbeck, H. Senn, U. Niederhauser, E. Holmes, J. C. Lindon, J. K. Nicholson, Analytical reproducibility in ^{1}H NMR-based metabonomic urinalysis, *Chem. Res. Toxicol*., 15(11), 1380-1386, 2002

[3] R. Stoyanova, J. K. Nicholson, J. C. Lindon, T. R. Brown, Sample classification based on Bayesian spectral decomposition of metabonomic NMR data sets, *Anal. Chem*., 76(13), 3666-3674, 2004

[4] M. F. Ochs, R. Stoyanova, F. Arias-Mendoza, T. R. Brown, A new method for spectral decomposition using a bilinear Bayesian approach, *J. Magn. Reson*., 137(1), 161-176, 1999

[5] D. D. Lee, H. S. Seung, Algorithm for non-negative matrix factorization, *Adv. Neural. Info. Proc. Sys. 13, MIT Press*, 556-562, 2001

[6] P. Sajda, S. Du, T. R. Brown, L. C. Parra, R. Stoyanova, Recovery of constituent spectra in 3D chemical shift imaging using non-negative matrix factorization, *Proc. Int. Conf. on Independent Component Analysis and Signal Separation (ICA2003)*, 71-76, Nara, Japan, 2003

[7] P. Sajda, S. Du, T. R. Brown, R. Stoyanova, D. Shungu, X. Mao, L. Parra, Non-negative matrix factorization for rapid recovery of constituent spectra in magnetic resonance chemical shift imaging of the brain, *IEEE Trans. Med. Imag*., 23(12), 1453-1465, 2004

[8] S. Du, X. Mao, D. Shungu, P. Sajda, Blind recovery of biochemical markers of brain cancer in MRSI, *Proc. SPIE Medical Imaging Symposium*, 5370-5378, San Diego, CA, USA, 2004

[9] C. M. Bishop, *Neural networks for pattern recognition*, England: Oxford University Press, Oxford, 1995

[10] N. Tishby, F. Pereira, W. Bialek, The information bottleneck method, *in Proc. of 37th Annual Allerton Conference on Communication, Control and Computing*, 368-377, 1999

[11] N. Friedman, O. Mosenzon, N. Slonim, N. Tishby, Multivariate information bottleneck, *in Proc. of 17th Conf. on Uncertainty in Artificial Intelligence (UAI)*, 152-161, 2001

[12] I. T. Jolliffe, *Principal Component Analysis*, New York, Springer Verlag, 1986

[13] M. G. Kendall, *Multivariate Analysis*, Charles Griffin Co., London, 1980

[14] L. Kaufman, P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*, John Wiley, 1990

[15] A. W. Nicholls, E. Holmes, J. C. Lindon, J. P. Shockcor, R. D. Farrant, J. N. Haselden, S. J. Damment, C. J. Waterfield, J. K. Nicholson, Metabonomic investigations into hydrazine toxicity in the rat, *Chem. Res. Toxic.*, 14(8), 975-987, 2001