Bayesian Inference and Maximum Entropy Methods in Science and Engineering, São Paulo, Brazil, 2008

Characterization of Interstellar Organic Molecules

Deniz Gençağa[a] , Duane F. Carbon[b], Kevin H. Knuth[a]

[a]*University at Albany, Department of Physics, Albany, NY, USA..*
[b]*NASA Ames Research Center, NASA Advanced Supercomputing Division, Moffett Field, CA, USA.*

Abstract. Understanding the origins of life has been one of the greatest dreams throughout history. It is now known that star-forming regions contain complex organic molecules, known as Polycyclic Aromatic Hydrocarbons (PAHs), each of which has particular infrared spectral characteristics. By understanding which PAH species are found in specific star-forming regions, we can better understand the biochemistry that takes place in interstellar clouds. Identifying and classifying PAHs is not an easy task: we can only observe a single linear superposition of PAH spectra at any given astrophysical site, with the PAH species perhaps numbering in the hundreds or even thousands. This is a challenging source separation problem since we have only one observation composed of numerous mixed sources. However, it is made easier with the help of a library of hundreds of PAH spectra. In order to separate PAH molecules from their mixture, we need to identify the specific species and their unique concentrations that would provide the given mixture. We develop a Bayesian approach for this problem where sources are separated from their mixture by Metropolis Hastings algorithm. Separated PAH concentrations are provided with their error bars, illustrating the uncertainties involved in the estimation process. The approach is demonstrated on synthetic spectral mixtures where the template data are taken from the Infrared Space Observatory. Performance of the method is tested for different noise levels.

Keywords: Bayesian Source Separation, Spectral Estimation, Astrophysics, Astrobiology.
PACS: 02.50.Tt, 02.50.Ga, 02.50.Ng, 96.55.+z

INTRODUCTION

In space, interstellar mediums (ISM) contain abundant amounts of large, complex organic molecules known as PAHs. These are mainly composed of many Carbon and Hydrogen atoms and can be found in neutral and ionic forms. Sometimes, they also involve Deuterium and Nitrogen atoms, as well [1]. These molecules are thought to have formed after supernovae explosions. In star-forming regions, the ultraviolet light of star excites these molecules and causes them to emit radiation in the infrared spectrum. Since each of these molecules has a unique vibration mode, they also possess unique emission spectra [2]. That is why; finding specific PAH molecules will give us information regarding the biochemical composition of a particular astrophysical site of interest.

However, we are only capable of observing a mixture of these species in the infrared spectrum range. Thus, finding the hidden PAHs from their linear superposition leads us to the source separation problem. In literature, a very limited research has been done to solve this problem. Although fitting data by hand has been tried [1], a satisfactory separation could not be achieved. Our ultimate goal is to

handle this formidable problem by developing a Bayesian method. The nature of the problem, on the other hand, causes serious complications: From a source separation point of view, the problem is highly challenging since we have only one measurement and hundreds to thousands of sources. We overcome this problem by using the library of spectral templates of the PAH molecules provided by our collaborators at NASA Ames Research Center. Dust radiation and atomic emissions also contribute to the observed mixture [3], making the problem even harder. These additional contaminations can be modeled by a Planck blackbody and a mixture of Gaussians, respectively [3]. In this work, we focus on the Bayesian separation of the PAH species. Although Non-negative Least Squares (NNLS) method has been used satisfactorily for this purpose [3], it is not capable of providing the uncertainties in the estimations. This problem is avoided by the Bayesian approach developed here. Each PAH molecule within the PAH library is modeled by a concentration parameter indicating the degree to which a particular species contribute to the mixture. Concentration of each PAH species within the library is inferred by using the Metropolis-Hastings algorithm with its associated error bar.

This paper is organized as follows: Next section presents problem statement followed by the description of the Bayesian methodology. Results are demonstrated in Section 4.

PROBLEM STATEMENT

Identifying PAH molecules is of utmost importance since we know which species are already present in our environment, where life has originated. Thus, finding similar PAHs elsewhere could provide us with an invaluable information regarding where to look for signs of life. In order to identify these molecules, we need to look at the infrared spectrum which includes the characteristic signatures of each PAH species. As an example, two of these species are illustrated below along with their spectra:

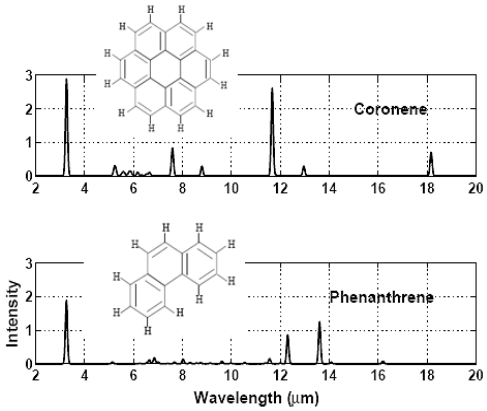

FIGURE 1. An example of two PAHs and their spectra.

PAHs are very stable, large and flat molecules of carbon and hydrogen. Each carbon has three neighboring atoms. Typically, all PAHs have emission lines near 3.3, 6.2, 7.7, 8.6, 11.2, and 15-20 microns. Separating these molecules from the spectral

mixture is a very *challenging* problem: A significant amount of PAH species possess tiny spectral flux at similar wavelengths. That is why; one PAH could easily be confused with another, having similar spectral characteristics.

Below, we present our mathematical forward model to describe the spectral measurement:

$$F(\lambda) = \sum_{i=1}^{N} c_i s_i(\lambda) + \phi(\lambda) \tag{1}$$

where $F(.)$ denotes the measurement. c_i, s_i are used for the concentration and spectral flux of the i^{th} source, respectively. The additive noise is shown by $\phi(\lambda)$ at a specific wavelength of λ.

Our goal is to infer the concentration parameters, c_i, given the data $F(.)$ and templates $s_i(\lambda)$ for $i = 1,2,...,N$ PAHs. In order to be able to deal with the challenging difficulties of this problem, we prefer using an informed Bayesian source separation methodology [4] rather than a blind one where we can exploit the prior information that we possess. Therefore, our methodology can be summarized as follows, by the well known Bayesian formula:

$$P(\mathbf{c} \mid D, I) = P(\mathbf{c} \mid I) \frac{P(D \mid \mathbf{c}, I)}{P(D \mid I)} \tag{2}$$

where $\mathbf{c}$ denotes the model parameter vector, i.e. $\mathbf{c} = [c_1, c_2, ..., c_N]$ (concentrations), D represents data and I denotes the prior information. In order to infer the concentration parameters, the posterior probability, $P(\mathbf{c} \mid D, I)$, is estimated by shaping our prior belief, $P(\mathbf{c} \mid I)$, with the observed data using the likelihood, $P(D \mid \mathbf{c}, I)$. We incorporate our prior belief by the selection of the prior probability and the spectrum model depicted by (1). Without loss of generality, the noise component in (1) could be modeled by a zero-mean Gaussian distribution, $\mathrm{N}(0, \sigma^2)$, where σ denotes the unknown standard deviation. This selection leads to the following likelihood function:

$$P(D \mid \mathbf{c}, I) = \left(2\pi\sigma^2\right)^{-N/2} \exp\left\{ -\sum_{\lambda} \frac{\left(F(\lambda) - D(\lambda)\right)^2}{2\sigma^2} \right\} \tag{3}$$

where $D(\lambda)$ and $F(\lambda)$ denote the measured flux (data) and the modeled spectral flux, respectively. Since we do not know the value of the standard deviation, we can integrate (3) over all possible values of σ using a Jeffrey's prior and obtain the following Student-t distribution for the likelihood function [5]:

$$P(D \mid M, I) = \left[\sum_{\lambda} \left(F(\lambda) - D(\lambda)\right)^2 \right]^{-N/2} \tag{4}$$

We incorporate our prior information on the concentration parameters by assigning a uniform distribution in (2) as shown below:

$$P(c_i \mid I) = \frac{1}{c_{\max} - c_{\min}}, \qquad i = 1,2,3,...,N \tag{5}$$

In order to estimate the posterior distribution given by (2), Metropolis-Hastings algorithm is utilized as described in the next section.

THE PROPOSED METHOD

In order to estimate the posterior probability of the concentration parameters, we propose using a Bayesian search and optimization scheme utilizing the Metropolis-Hastings algorithm. Metropolis-Hastings is one of widely used Markov Chain Monte Carlo (MCMC) methods where the objective is to draw independent, identically distributed (i.i.d) samples from the posterior distribution [6]. To accomplish this goal, a Markov chain is generated in such a way that its samples are asymptotically distributed according to the desired distribution, namely $P(\mathbf{c} \mid D, I)$. Once we get samples from the desired distribution, we can also obtain its statistical summaries such as the mean and error bars of the related parameters.

To construct a Markov chain, a new sample is generated from the proposal distribution which is located at the current value of the parameter, $\mathbf{c}^{(t)}$. This iterative sampling is represented by $\mathbf{c}^* \sim q(\mathbf{c})$ where $q(.)$ denotes the proposal distribution and $\mathbf{c}^*$ represents the candidate sample. Having drawn a new sample from the proposal distribution, acceptance ratio is calculated as shown below:

$$\rho = \frac{p(\mathbf{c}^*)q(\mathbf{c}^{(t+1)};\mathbf{c}^*)}{p(\mathbf{c}^*)q(\mathbf{c}^*;\mathbf{c}^{(t+1)})} \tag{6}$$

where $\rho, p(.)$ denote the acceptance ratio and the desired distribution, respectively. Here, $q(y;x)$ denotes the value of the proposal distribution evaluated at *y* and located at *x*. If $\rho \geq 1$, $\mathbf{c}^*$ is accepted to the Markov chain: $\tilde{\mathbf{C}} = \{..., \mathbf{c}^{(t-1)}, \mathbf{c}^{(t)}, \mathbf{c}^*\}$, i.e. $\mathbf{c}^{(t+1)} = \mathbf{c}^*$. If $\rho < 1$, then $\mathbf{c}^*$ is accepted with probability ρ. If it is rejected, then the Markov chain proceeds by $\mathbf{c}^{(t+1)} = \mathbf{c}^{(t)}$, i.e. $\tilde{\mathbf{C}} = \{..., \mathbf{c}^{(t-1)}, \mathbf{c}^{(t)}, \mathbf{c}^{(t)}\}$. The reader is referred to [7] for further details on MCMC methods.

A pseudocode of the methodology is given below to demonstrate each step in the algorithm explicitly.

TABLE 1. Bayesian methodology

1. Draw initial samples from the prior distribution of the concentration parameters:

$$c_i^* \sim q(\mathbf{c}) \text{ where } q(c_i) = P(c_i \mid I) = \frac{1}{c_{\max} - c_{\min}} \text{ for } i = 1,2,3,...,N$$

2. Calculate the initial likelihood value:

$$\log(L) = -\left(K/2\right)\log\left\{\sum_{k=1}^{K}(F(\lambda_k) - D(\lambda_k))^2\right\} \text{ where } F(\lambda) = \sum_{i=1}^{N} c_i s_i(\lambda) + \phi(\lambda)$$

3. FOR t = 1 TO T (number of iterations)

FOR i = 1 TO N (number of components)

SET number of accepts = 0 (A=0), SET mean value: $m = 0$, SET mean squared value $m^2 = 0$

FOR r = 1 TO R

Draw new samples from the proposal distribution:

$$c_i^* = c_i + \mu_i x \quad \text{where } x \sim N(0,1) \text{ , } i = 1,2,3,...,N$$

Verify that each $c_{\min} < c_i < c_{\max}$

Calculate the likelihood of the proposed samples:

$$\log(\tilde{L}) = -\left(K/2\right)\log\left\{\sum_{k=1}^{K}(\tilde{F}(\lambda_k) - D(\lambda_k))^2\right\} \text{ where } \tilde{F}(\lambda) = \sum_{i=1}^{N} c_i^* s_i(\lambda) + \phi(\lambda)$$

Accept new samples with probability ρ and augment the chain :

$$\rho = \min\left\{0, \frac{p(\mathbf{c}*)q(\mathbf{c}^{(t+1)};\mathbf{c}*)}{p(\mathbf{c}*)q(\mathbf{c}*;\mathbf{c}^{(t+1)})}\right\}$$

$$\tilde{\mathbf{C}} = \left\{...,\mathbf{c}^{(t-1)},\mathbf{c}^{(t)},\mathbf{c}*\right\}$$

SET $m_i = m_i + \tilde{\mathbf{C}}(i)$ and $m_i^2 = m_i^2 + \tilde{\mathbf{C}}(i)^2$, $i = 1,2,3,...,N$

INCREMENT NUMBER OF ACCEPTS BY 1: A = A + 1

END

$\overline{m}_i = m_i/R$ and $\sigma_i = \sqrt{\left(m_i^2 - \overline{m}_i^2\right)}$

CALCULATE THE ACCEPTANCE RATE: A = A/R

ADJUST THE STEP-SIZE VALUE:

IF A<0.1 , $\mu_i = \mu_i/10$

ELSEIF A<0.34, IF $\mu_i > 0.1\sigma$, $\mu_i = \mu_i/2$

ELSEIF A>0.67, IF $\mu_i < |\overline{m}_i|$, $\mu_i = 1.1\mu_i$

END

END

EXPERIMENTS

In this section, we demonstrate our method on synthetic spectral mixture data where the templates are taken from the ISO. We mix 47 PAH species from a template of 187 species with random concentrations varying between 0 and 3000. First, the performance of the method is examined without the additive noise, i.e. $\phi(\lambda)=0$ is taken in (1). The logarithm of the likelihood of the true solution is calculated to be $2.864\text{x}10^5$. For this data set, we use the Metropolis-Hastings method starting from 10 random concentration vectors and run it for 10000 iterations. The burnin period is chosen to be 9800 as a result of our observations. In Fig. 2, the propagation of one of the 10 samples is illustrated. Using the sample values at the steady-state, mean value of each concentration parameter is shown in Fig. 3. along with its error bar.

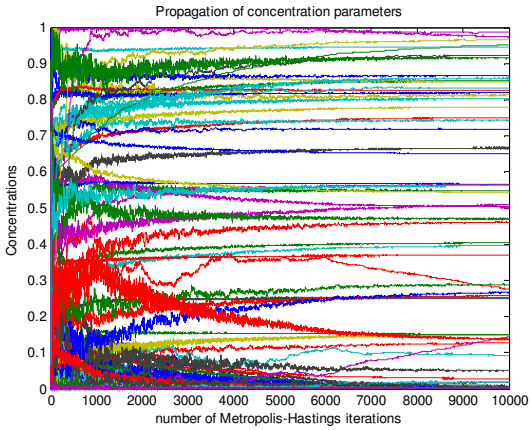

FIGURE 2. Propagation of the concentration parameter estimates vs. the number of iterations

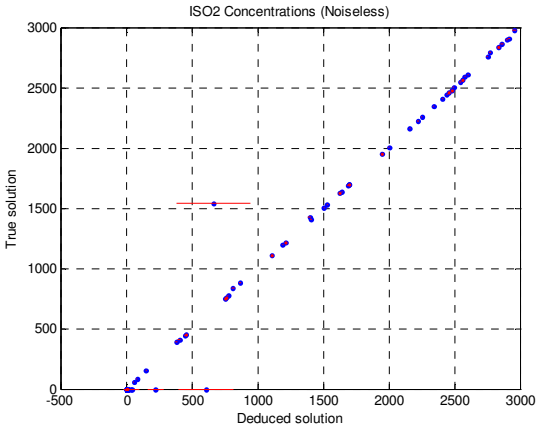

FIGURE 3. Proposed method: Deduced vs. true solutions of the concentration parameters for the noiseless ISO2 mixture

Above, deduced concentrations are plotted vs. true values. The error-bar of each estimated parameter is shown by a line located on the corresponding mean estimate illustrated by the dot. It is observed that except for three outliers, almost every parameter lies within one standard deviation of the true value providing a 45-degree line. The mean solution has a log-likelihood value of -59608 with a Euclidean distance of $d=1093.6$ from the true solution in the 187 dimensional space. Despite the three outliers, reconstructed spectrum fits perfectly with the true spectrum of the ISO2 data as shown in Fig. 4. In order to compare these results, we use NNLS technique [8] to estimate the concentration parameters. This algorithm is run starting from 200 different points in the 187 dimensional space. The quality of the estimation is illustrated by the scatter plot of the deduced concentrations vs. true values in Fig. 5. Similar to Fig. 3, NNLS method provides an almost perfect estimation with the mean solution having a log-likelihood of -66767 within a distance of 1275.4 from the true solution. In order to test the performance of the proposed method under different noise levels, three simulation results are demonstrated where the noise power is taken to be

1/1000,1/100, and 1/10 of the signal power. The scatter plots of the deduced vs. true concentrations are illustrated in Figs. 6a-6c for three situations. Spectral reconstruction is illustrated in Fig. 7 for the most noisy case among three.

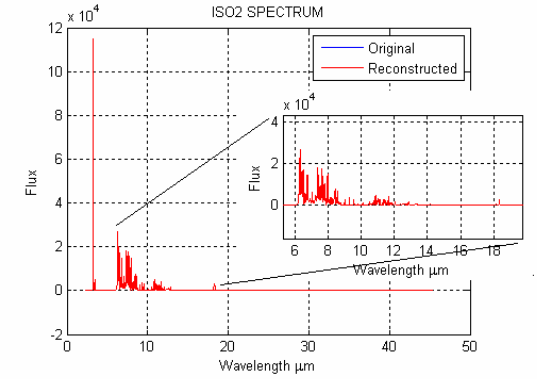

FIGURE 4. Original and reconstructed spectra for the ISO2 data under no noise

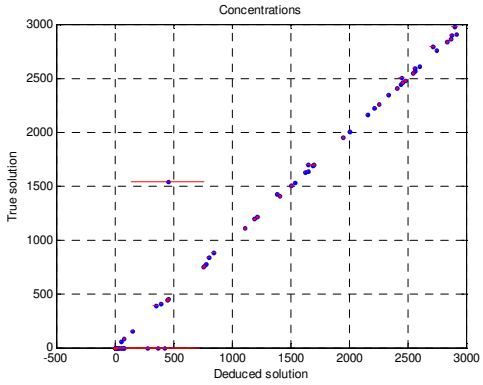

FIGURE 5. NNLS: Deduced vs. true solutions of the concentration parameters for the noiseless ISO2 mixture

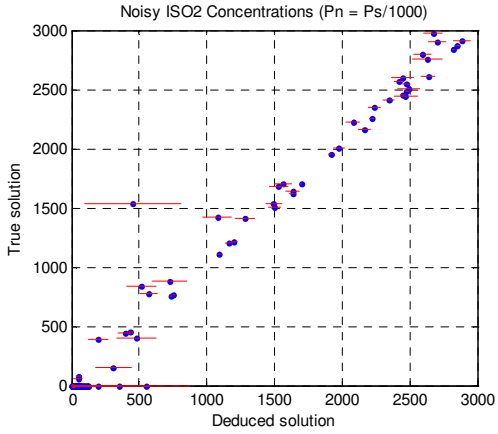

FIGURE 6a. Noise level Pn = Ps/1000

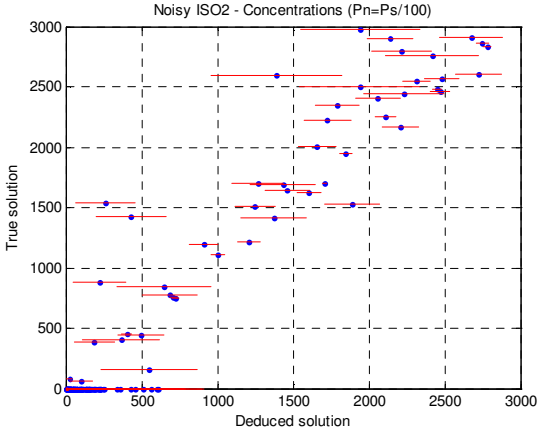

FIGURE 6b. Noise level Pn = Ps/100

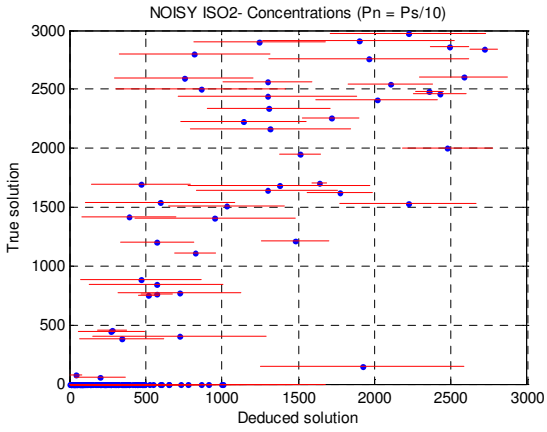

FIGURE 6c. Noise level Pn = Ps/10

Deduced vs. true concentration parameters

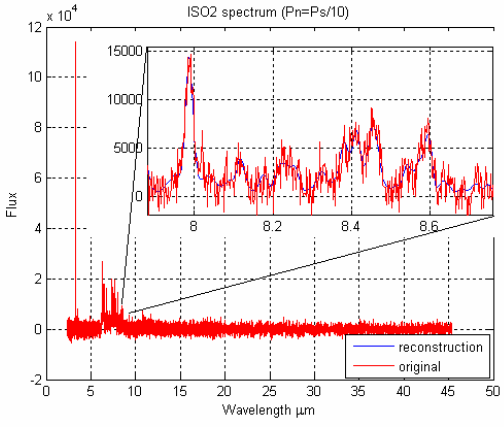

FIGURE 7. Original and reconstructed spectra of ISO2 under noise: Pn = Ps/10

CONCLUSIONS

A Bayesian methodology is presented to identify the PAH molecules from their mixtures enabling us to estimate the posterior probability distributions of the concentration parameters. This allows us to summarize our inference with their error-bars and provides the most honest solution about the problem without being constrained to a local optima. Simulation results demonstrate that the estimations lie within one standard deviation of the true solution, providing promising solutions for the future applications where the number of PAHs will be increased. Having the error-bars, we will have the flexibility to express our uncertainty in the estimations unlike frequentist approaches such as NNLS. Thus, it will enable us to deal with this formidable problem by letting us express our uncertainty in the estimations done by our prior models and it will also allow us to change these models as we learn more from the problem.

REFERENCES

1. L.J. Allamandola, D.M. Hudgins, S.A. Sandford, "Modeling the unidentified infrared emission with combinations of polycyclic aromatic hydrocarbons," *ApJ*, 511, L115-119, 1999.
2. L.J. Allamandola, A.G.G.M. Tielens, J.R. Barker, "Polycyclic aromatic hydrocarbons and the unidentified infrared emission bands: Auto exhaust along the Milky Way!" *Astrophys. J. Letters*, 290, L25, 1985.
3. K. H. Knuth, M. K. Tse, J. Choinsky, H. Maunu, D. F. Carbon,, "Bayesian source separation applied to identifying complex organic molecules in space," *IEEE/SP 14th Workshop on Statistical Signal Processing*, Aug. 2007, pp. 346 – 350.
4. K.H. Knuth, "Informed source separation: A Bayesian tutorial," In: B. Sankur , E. Çetin, M. Tekalp , E. Kuruoğlu (eds.), *Proceedings of the 13th European Signal Processing Conference (EUSIPCO 2005)*, Antalya, Turkey, 2005.
5. D.S. Sivia, J. Skilling "Data Analysis: A Bayesian Tutorial", 2nd Ed. Oxford University Press, Oxford, 2006.
6. N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller, "Equations of state calculations by fast computing machines," Journal of Chemical Physics, 21, pp. 1087-1091, 1953.
7. D. McKay, "Information Theory, Inference and Learning Algorithms," Cambridge University Press, 2003.
8. Lawson, C.L. and R.J. Hanson, "Solving Least Squares Problems," Prentice-Hall, 1974, Chapter 23, p. 161.