SOURCE SEPARATION ALGORITHMS FOR THE ANALYSIS OF HYPER-SPECTRAL OBSERVATIONS OF VERY SMALL INTERSTELLAR DUST PARTICLES

O. Berné[1], C. Joblin[2], A. Tielens[3], Y. Deville[4], M. Puigt[4], R. Guidara[4], S. Hosseini[4], G. Mulas[5], J. Cami[6]

[1] Centro de Astrobiologia CSIC/INTA Madrid - Spain
[2]CESR - Université de Toulouse, CNRS - France
[3]Leinden University - Netherlands
[4]LATT - Université de Toulouse, CNRS - France
[5]OAC - INAF - Italy
[6]University of Western Ontario - Canada

ABSTRACT

The mid-infrared (mid-IR) spectrum of our galaxy is dominated by continuum and band emission due to carbonaceous very small dust particles amongst which are polycyclic aromatic hydrocarbon (PAH) molecules. Because they absorb the UV photons of massive stars and re-emit this energy in the infrared, IR spectro-imaging of extended interstellar (or circumstellar) regions is a powerful tool to diagnose the nature of these particles together with the local physical conditions. In this paper, we review how the applications of blind / bayesian source separation (BSS) methods applied to mid-IR hyperspectral data can help analyzing these data. We then discuss, in the light of simulations in progress, how BSS methods could be used to identify specific PAH molecules in the interstellar medium when applied to the hyper-spectral data of forthcoming IR telescopes (Herschel, SOFIA, SPICA).

Index Terms— Source Separation Methods, Infrared Spectroscopy, Interstellar Medium

1. INTRODUCTION

The interstellar medium (ISM) is the space that lies between the stars. It contains gas and dust particles that are major actors in the chemical and physical evolution of galaxies. In the recent past, space missions have unveiled for many regions of the ISM the presence of a set of bands in the mid-infrared (3-14 μm, mid-IR) attributed to the emission of very small dust particles amongst which Polycyclic Aromatic Hydrocarbons (PAHs, [1]). PAHs are large carbonaceous molecules composed of benzenoid rings surrounded by hydrogen atoms. However, the correspondence between the observed mid-IR bands and the exact population of particles, or specific molecules at the origin of this emission has not yet been established clearly. In this paper, we review the recent progresses that have been made in this field using source separation algorithms. In section 2, we present how the application of different Blind Signal Separation (BSS) methods to Spitzer hyperspectral cubes has enabled us to extract the genuine spectrum of different populations of mid-IR emitters. This is a first step in their identification, however, the ultimate goal remains to spectroscopically identify a *specific* PAH molecule. In section 3 we present how signal separation algorithms applied to far-IR spectrosopy can help on this topic.

2. BLIND SIGNAL SEPARATION FOR THE IDENTIFICATION OF VERY SMALL DUST PARTICLE POPULATIONS BASED ON MID-IR HYPESPECTRAL DATA

2.1. Applying BSS to mid-IR spectral cubes

The shape of the band emission observed in the mid-IR and attributed to PAHs varies significantly from one astrophysical environment to the other, probably due to the fact that the observed spectra are mixtures of emission from several populations of PAHs and small particles. Hyperspectral images of photodissociation regions (regions of the ISM where dust is processed by the UV photons of massive stars, hereafter PDR) obtained with the Spitzer Space Telescope (Spitzer) exhibit strong spectral variations (Fig. 1). We mathematically refer to these data as $C(p_x, p_y, \lambda)$, where p_x, p_y and λ are the spatial positions and wavelength indexes respectively. We suggest that the spectral variation observed in $C(p_x, p_y, \lambda)$ are due to linear instantaneous mixtures of a few *source* spectra (representative of chemical populations) with different mixing coefficients. In the frame of this hypothesis, the use of source separation algorithms enables to obtain an estimation of the source spectra.

Blind Signal Separation is commonly used to restore a set of unknown "source" signals from a set of observed signals which are mixtures of these source signals, with unknown mixture parameters [2]. It has e.g. been applied in acoustics

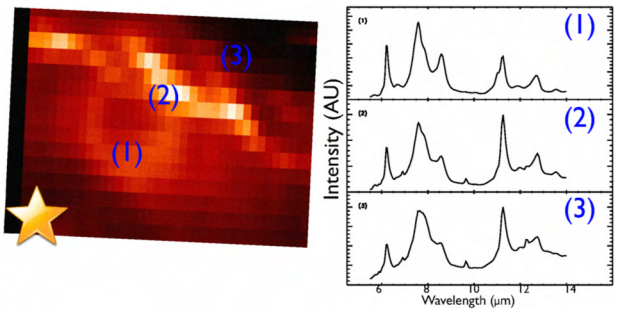

Fig. 1. The NGC 7023 North PDR mid-IR hyperspectral cube. An image of this cube at 7.7 μm (left) and spectra for the three indicated positions are shown. The position of the illuminating star is schematically represented.

for unmixing recordings, or in the biomedical field for separating mixed electromagnetic signals produced by the brain. BSS is most often achieved using Independent Component Analysis (ICA) methods. An alternative class of methods for achieving BSS is non negative matrix factorization (NMF), which was introduced in [4] and then extended by a few authors. The other main class of methods concerns the algorithms based on sparsity (a sparse signal is often equal to zero but has a non-zero variance, e.g. voice signals are sparse) of the sources [2]. The simplest version of the BSS problem concerns so-called "linear instantaneous" mixtures. It is modeled as follows:

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

where X is an $m \times n$ matrix containing n samples of m observed signals, A is an $m \times r$ mixing matrix and S is an $r \times n$ matrix containing n samples of r source signals. The observed signal samples are considered to be linear combinations of the source signal samples (with the same sample index). It is assumed that $r \leq m$ in most investigations, including this paper. The objective of BSS algorithms is then to recover the source matrix S and/or the mixing matrix A from X, up to BSS indeterminacies.

The correspondence between the generic BSS data model (1) and the 3-dimensional spectral cube $C(p_x, p_y, \lambda)$ to be analyzed in the present paper may be defined as follows. In this paper, the sample index is associated to the wavelength λ, and each observed signal consists of the overall spectrum recorded for a cube pixel (p_x, p_y). Each one of these signals defines a row of the matrix X in Eq. (1). Moreover, each observed spectrum is a linear combination of *source spectra*, which are respectively associated to each of the (unknown) types of PAH populations and nanoparticles that contribute to the recorded spectral cube. Therefore, the recorded spectra may here be expressed according to (1), with unknown combination coefficients in A, unknown source spectra in S and an unknown number r of source spectra. We have used four BSS methods in order to recover the A and S matrix from the observed spectra placed in the rows of X, i.e. NMF (based on positivity of the sources), FASTICA and Markovian non-stationnary (based on independence of the source) and LI-TIFROM (based on sparsity of the sources). We briefly detail these methods hereafter.

We used NMF as presented in [5]. The matrix of observed spectra X is approximated using WH, where W and H are non-negative matrices, with the same dimensions as in (1). This approximation is optimized by adapting the matrices W and H using the algorithm of [5] in order to minimize the divergence between X and WH. We implemented the algorithm with Matlab. After convergence, W and H provide an estimation of A and S respectively. Convergence is reached after about 1000 iterations (which takes less than one minute with a 3.2 GHz processor).

We used FASTICA in the deflation version [3] in which each estimated source is extracted one after the other and subtracted from the observations until all sources are extracted. The advantage of this FASTICA method is that it is not necessary to fix, before running the algorithm, the number r of sources that we want to extract, as it is for NMF (see the method used to overcome this issue in [6]). The extraction of the sources takes less than one minute using *FastICA* coded with Matlab, and with a 3.2 GHz processor.

We used the Markovian non-stationnary method in [8]. Based on a maximum likelihood approach, this method supposes that the sources are first-order Markovian processes. Moreover, the non-stationarity of the sources is taken into account using a blocking approach. The algorithm is implemented in Matlab code and the separation takes less than three minutes with a 1.16 GHz processor.

We used the symmetrical version of LI-TIFROM[1] based on time-frequency (TF) ratios of observations [7]. It first finds *analysis zones* (each of them is a set of adjacent TF windows) where a source is *visible* (i.e. occurs alone) and then estimates in each of these zones the column of the mixing matrix associated to the visible source. Lastly, sources are recovered in the original analysis domain by inverting the mixing matrix. For Spitzer spectra, LI-TIFROM provides a complete separation with "well" selected TF parameters (in less than one second with a 1.4 GHz processor).

2.2. Results

In a selection of several (about 10) PDRs, we were able to extract two source spectra from each hyperspectral cubes: one carrying broad bands and a continuum and a second with narrower, intense bands and no continuum (Spectra 1 and 2 given as examples for the Ced 201 PDR in Fig.2). Spectrum 1 was attributed to very small carbonaceous grains (VSGs), possibly

[1] The Matlab code of this method is accessible at http://www.ast.obs-mip.fr/bss-softwares.

PAH clusters [9], while spectrum 2 was attributed to a population of free PAH molecules [6, 10]. In the case of NGC 7023, we were able to find 3 spectra using NMF (Fig. 3), including two PAH spectra with different band intensity ratios, one is attibuted to neutral PAHs (PAH^0) and the other to ionized PAHs (PAH^+).

Using the extracted spectra of VSG, PAH^0, and PAH^+ we can reconstruct the spatial distribution of each population in the spectral cube by fitting a linear combination of the three extracted spectra to the observations. We perform this using a least mean square fit and obtain the map in Fig.3. We find that the three populations emit in very different regions. It appears that there is an evolution from the population of VSGs to PAH^0 and then PAH^+ while approaching the star. This reveals chemical processing by the UV stellar radiation. Conversely, this means that each population can thus be used as a tracer of the local UV field.

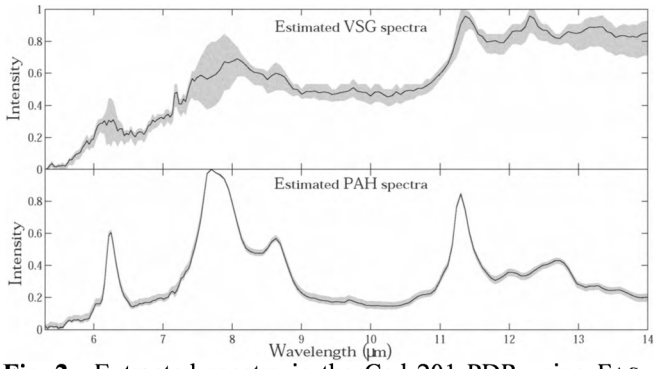

Fig. 2. Extracted spectra in the Ced 201 PDR using FastICA, Lee and Seung's NMF, TIFROM and Markovian methods. Lines are mean values while the gray areas represent the envelope of solutions.

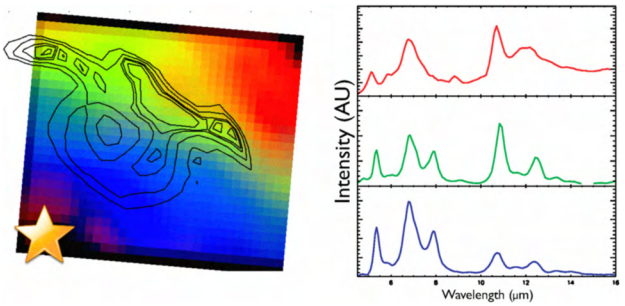

Fig. 3. *Left:* The reconstructed distribution maps for each population, VSGs in red, PAH^0 in green and PAH^+ in blue. *Right:* The three extracted spectra in NGC 7023 North (see Fig. 1) using Lee and Seungs NMF, with same color scheme as for distribution map i.e. VSGs in red, PAH^0 in green and PAH^+ in blue.

3. TOWARDS THE IDENTIFICATION OF SPECIFIC PAH MOLECULES IN THE FAR-IR

The mid-IR wavelength domain appears rather inappropriate to perform the distinction between specific molecules since the emission bands modes of local chemical bonds. However, emission features are also present in the far-IR [11] that are much more specific of individual species. Therefore a detailed analysis of the far-IR spectrum ($\sim$100 to 500 μm) is required to progress in the identification of interstellar PAHs. Forthcoming instruments onboard airborne (SOFIA) or space missions (Herschel that will be launched in 2009, SPICA a future planned Japanese mission) will allow to perform spectro-imagery in the far-IR range at increasing sensitivities from Herschel to SPICA. We therefore propose to investigate, how, in the frame of these future observations in the far-IR, we can identify individual PAH species.

We make the assumption that the PAH mixture is made of small and large species, in proportions that depend on the environment, the very large PAHs being more resistant and thus able to survive in harsh irradiation conditions. We retrieved fundamental data from quantum chemical calculations [12, 13] for 10 small PAHs ($N_C\sim$20-40 carbon atoms) and 2 large PAHs called Grand-PAHs, in their neutral and ionized forms (24 different spectra in total). We then ran the photophysical model of [11] in the UV irradiation conditions of the Red Rectangle nebula to provide the emission spectra for all these molecules in this specific environment. We then mix the four spectra of Grand-PAHs randomly and add noise to it. Noise here consists in white gaussian noise and a contribution from small PAHs at a level of 50% in abundance compared to the total Grand-PAH abundance. We generate a sample of 100 "observations" using this technique (see Fig. 4).

We used FastICA, NMF and Bayesian Positive source separation (BPSS [14]) in order to try to recover at least one of the source spectrum. The three methods give good preliminary results. For now we have restricted our analysis to the 150-350 μm wavelength domain which was found to be the most appropriate. Indeed, we are able to recover the spectrum of $C_{66}H_{20}^+$ quite efficiently (see e.g. for NMF in Fig. 4) even though some remains of $C_{66}H_{20}$ are seen.

4. CONCLUSION AND PROSPECTIVES

Source separation algorithms have proven their efficiency and unveiled previously unidentified populations of interstellar macromolecules and nanoparticles. This first step, based on the mid-IR hyperspectral mapping was limited by the level of information provided at these wavelengths. As suggested in this paper, the application of BSS methods to far-IR hyperspectral imaging is highly promising and will possibly enable to identify the first specific PAH molecule in space. In order to be prepared when real data will be available with Herschel, SOFIA, and possibly SPICA, modelling and sim-

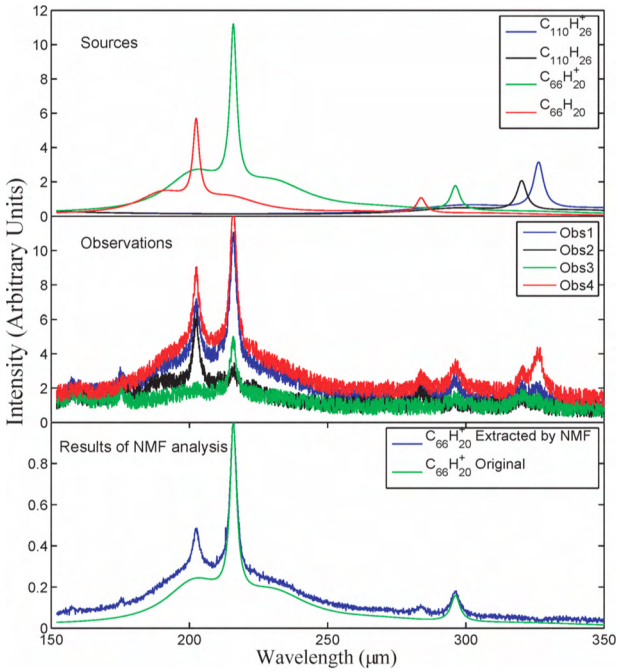

Fig. 4. The four calculated Grand-PAH emission spectra ("Sources"), an example of 4 simulated observations (labelled "Observations"), and the NMF extracted and original spectrum of $C_{66}H^+_{20}$ ("Results of NMF analysis")

ulation should not be neglected. In particular, a missing part in the preliminary effort presented here, is an efficient and realistic model of instrumental noise. We will add this to our simulations in the near future.

5. REFERENCES

[1] A. G. G. M. Tielens, "Interstellar Polycyclic Aromatic Hydrocarbon Molecules," *Annual Review of Astronomy and Astrophysics*, vol. 46, pp. 289–337, Sept. 2008.

[2] P. D. O'Grady, B. A. Pearlmutter, and S. T. Rickard, "Survey of Sparse and Non-Sparse Methods in Source Separation," *International Journal of Imaging Systems and Technology, Special Issue: Blind Source Separation and Deconvolution in Imaging and Image Processing*, 2005.

[3] A. Hyvarinen, "A Fast Robust Fixed Point Algorithm for Independant Component Analysis," *IEEE Transactions on Neural Networks*, vol. 10, pp. 626–934, 1999.

[4] D. D. Lee and H. S. Seung, "Learnig the parts of onjects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, Oct. 1999.

[5] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *NIPS*, MIT press, Ed., 2001, vol. 13, p. 556.

[6] O. Berné, C. Joblin, Y. Deville, J. D. Smith, M. Rapacioli, J. P. Bernard, J. Thomas, W. Reach, and A. Abergel, "Analysis of the emission of very small dust particles from Spitzer spectro-imagery data using blind signal separation methods," *Astronomy and Astrophysics*, vol. 469, pp. 575–586, July 2007.

[7] Y. Deville, M. Puigt, and B. Albouy, "Time-frequency blind signal separation: extended methods, performance evaluation for speech sources," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN 2004)*, Budapest, Hungary, July 25–29, 2004, vol. 1, pp. 255–260.

[8] R. Guidara, S. Hosseini, and Y. Deville, "Blind separation of Non-stationary Markovian Sources Using an Equivariant Newtion-Raphson Algorithm," *IEEE Signal processing Letters*, vol. 16, pp. 426–429, May 2009.

[9] M. Rapacioli, F. Calvo, C. Joblin, P. Parneix, D. Toublanc, and F. Spiegelman, "Formation and destruction of polycyclic aromatic hydrocarbon clusters in the interstellar medium," *Astronomy and Astrophysics*, vol. 460, pp. 519–531, Dec. 2006.

[10] M. Rapacioli, C. Joblin, and P. Boissel, "Spectroscopy of polycyclic aromatic hydrocarbons and very small grains in photodissociation regions," *Astronomy and Astrophysics*, vol. 429, pp. 193–204, Jan. 2005.

[11] G. Mulas, G. Malloci, C. Joblin, and D. Toublanc, "A general model for the identification of specific PAHs in the far-IR," *Astronomy and Astrophysics*, vol. 460, pp. 93–104, Dec. 2006.

[12] G. Malloci, C. Joblin, and G. Mulas, "On-line database of the spectral properties of polycyclic aromatic hydrocarbons," *Chemical Physics*, vol. 332, pp. 353–359, Feb. 2007.

[13] C. W. Bauschlicher, Jr., E. Peeters, and L. J. Allamandola, "The infrared spectra of very large, compact, highly symmetric, polycyclic aromatic hydrocarbons (PAHs)," *ArXiv e-prints*, vol. 802, Feb. 2008.

[14] S. Moussaoui, D. Brie, A. Mohammad-Djafari, and C. Carteret, "Separation of non-negative mixture of non-negative sources using a bayesian approach and mcmc sampling," *Signal Processing, IEEE Transactions on*, vol. 54, no. 11, pp. 4133–4145, Nov. 2006.