Estimating Background Spectra

M. K. Tse\textsuperscript{a}, J. Choinsky\textsuperscript{a,b}, D. F. Carbon\textsuperscript{c}, and K. H. Knuth\textsuperscript{a,d}

\textit{a. Univ. at Albany, Dept of Physics, Albany NY USA}
\textit{b. Univ. at Albany, Dept of Computer Science, Albany NY USA}
\textit{c. NASA Ames Research Center, NASA, Moffett Field CA USA}
\textit{d. Univ. at Albany, Dept of Informatics, Albany NY USA}
http://knuthlab.rit.albany.edu/

Abstract. All measurements consist of a mixture of the signal of interest and additional signals called the background. Here we focus on the problem of measuring infrared spectra emitted by interstellar clouds. The signals of interest are infrared emissions from polycyclic aromatic hydrocarbons (PAHs), which are a class of complex organic molecules. The PAH emissions are characterized by emission bands near 3.3, 6.2, 7.7, 8.6, 11.2, and 15-20 microns. The background consists of a host of associated spectral signals which, in the simplest case, can include emissions from multiple Planck blackbodies as well as broadband and narrowband emissions. To analyze the PAH spectra we must accurately assess this background. To do this, we have developed a Bayesian algorithm based on nested sampling (Skilling 2005, Sivia & Skilling 2006). The spectral model consists of a mixture of Planck functions and Gaussians. We demonstrate this algorithm on both synthetic data and infrared spectra recorded from interstellar clouds. The result shows that the algorithm can accurately identify and remove simple backgrounds. In future work, we plan to incorporate mixtures of PAH spectra and more complex models for the background so that the algorithm will simultaneously estimate both the signals of interest and the background.

Keywords: background, spectrum, nested sampling, MCMC, Planck blackbody, mixture of Gaussians, astrophysics, astrobiology

PACS: 98.70.Lt, 98.70.Vc

1. INTRODUCTION

In astrophysics, all observed spectra consist of a mixture of the signal of interest and additional signals collectively called the background. The background can be decomposed into emission from unknown sources and known sources. Here we focus on the problem of analyzing infrared spectra emitted by interstellar clouds (Figure 1). Infrared spectra in regions of star formation have strong emission from a class of benzene-based molecules known as Polycyclic Aromatic Hydrocarbons (PAHs). The observed spectra are the combined emission of numerous PAH species, both neutral and ionized, in addition to emission from atomic species plus dust emission and absorption at multiple temperatures. While characterizing these factors promises to enhance the level of understanding of star-forming regions, identifying individual PAHs in a spectrum that consists of a mixture of a large number of PAH species is a difficult problem in itself.

The definition of background depends on both the application and the interests of the researcher; one person’s background is another person’s signal. In this paper, the
signals of interest are the PAH emissions, which are characterized by bands near 3.3, 6.2, 7.7, 8.6, 11.2 and 15-20 microns. There are thousands of PAH species, each with their unique emission spectrum (Figure 2 shows some examples of PAHs and their spectra). However, there are additional signals present in the recorded infrared spectrum. There are dust clouds along the line-of-sight that both radiate and absorb radiation. Each of the clouds has distinct physical conditions. In addition, these are unknown sources which cannot be described phenomenologically. For this reason, estimating the background is a difficult problem.

In this paper, Nested Sampling (Skilling 2005; Sivia & Skilling 2007) will be used to characterize the background spectrum. We will demonstrate how the algorithm can isolate PAHs from the other sources, such as dust emission. Furthermore, we will use the evidence calculations to evaluate whether the contributions of what are interpreted as unknown sources are warranted by the data.

2. MODELING SPECTRA

2.1 Blackbody Radiation

Interstellar clouds of dust in star-forming regions are heated by the radiation emitted by young stars (Figure 1). The particles in the clouds re-radiate this energy as infrared radiation with a spectrum dictated by the physical conditions in the cloud. In the very
simplest approximation, the cloud radiation can be represented by a Planck blackbody function at some temperature $T$:

$$
Planck (\lambda, T) = \frac{\lambda_{\text{max}}^3}{\lambda^2} \frac{\exp\left(\frac{hc}{\lambda_{\text{max}} kT}\right) - 1}{\exp\left(\frac{hc}{\lambda kT}\right) - 1}
$$

where $k$ is Boltzmann’s constant, $h$ is Planck’s constant, $c$ is the speed of light, and $\lambda_{\text{max}}$ is the wavelength where the blackbody spectral energy peaks.

We employ three distinct types of source spectra models: a dictionary of atomic and PAH spectra, blackbody radiators describing dust radiation and a 'non-parametric' mixture of Gaussians describing the unknown source radiation. To isolate the signals emitted from PAHs, the background must be first identified and separated from the linear superposition of the various PAH species. The mixture of Gaussians and the blackbody radiators are used to model the background.

### 2.2 Modeling Unknown Signals

The recorded infrared spectra usually contain a certain level of broad spectral features. These features cannot be attributed to any known source. To model these unknown sources, we utilize a mixture of Gaussians. Since unknown spectral sources are not a focus of this project, the mixture of Gaussians model is sufficient to represent these background features to a first approximation.

$$
\sum_i N(\mu_i, \sigma_i) = \sum_i \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(\frac{x - \mu_i}{2\sigma_i^2}\right)
$$
where $\sum_i N(\mu_i, \sigma_i)$ is the mixture of Gaussian, $i$ is the index of Gaussians, $\mu_i$ is the mean of the $i$-th Gaussian and $\sigma_i$ is the standard deviation of the $i$-th Gaussian.

2.3 PAH Spectra

A Polycyclic Aromatic Hydrocarbon (PAH) is an assembly of hexagonally-shaped carbon rings of the simplest aromatic molecule. Three examples of PAHs are shown in Figure 2. Each PAH has distinct spectral features, which arise from the vibrational states unique to the structure of each particular molecule. Ionization and element substitution can modify PAH wavelengths and relative intensities. In order to identify the characterized PAH emissions, our partners at NASA Ames Research Center have compiled a library of over a thousand PAH spectra based on laboratory measurements and quantum mechanical simulations. Figure 2 (right) shows spectra for three PAH species from the NASA Ames PAH library.

2.4 Combined Spectral Model

The infrared spectra we are working with are the combined emission of the PAH molecular species, the Planck radiation and unknown sources.

$$F(\lambda) = \sum_k Planck(\lambda, T_k) + \sum_i N(\lambda; \mu_i, \sigma_i) + \sum_j PAH_j (\lambda)$$  \hspace{1cm} (3)

where $F(\lambda)$ is the modeled spectral flux at wavelength $\lambda$. The first term represents a sum of blackbodies, the second term is a mixture of Gaussians representing the unknown sources, and the third term represents the PAH contributions. With this parameterized model of the observed spectra, we can begin to make inferences about the contributions from blackbody radiation, unknown sources, or PAH spectra. Since we have constrained the $\sigma_i$ to have values above some minimum value substantially larger than the widths of the atomic features, we have not explicitly modeled the atomic emission in the current study. As a consequence of this approximation, the atomic features are treated here as though they were “PAHs”. In later work, we will add a mixture of very sharp Gaussians and a catalog of wavelengths for known atomic emission species.

3. NESTED SAMPLING

We employ a new technique called Nested Sampling (Skilling 2005, Sivia & Skilling 2006). The main goal of Nested Sampling is to numerically integrate the posterior and obtain the evidence. In future work it is expected that this will be invaluable as we test multiple spectral theories. Nested sampling works by sampling
from the prior probability within a likelihood constraint. The likelihood constraint defines an implicit boundary over which the algorithm shall not cross.

The algorithm maintains a set of samples, each of which describes a hypothesized spectral model. The likelihood of each sample is computed and the sample with the least likelihood is flagged and discarded. To maintain the set of samples, a new sample is generated, by duplicating one of the valid samples, and varied by changing its parameter values slightly several times until it is decorrelated from the initial sample. Each sample is assigned a weight based on the product of the estimated prior probability shrinkage ratio and the likelihood. The end result is a set of samples with weights that denote their contribution to the integral of the posterior probability.

We work with uniform priors, so that the algorithm samples uniformly from the hypothesis space. The likelihood of each sample is calculated by using the Student t-distribution, which is proportional to the sum of squared differences between observed spectrum $\Phi(\lambda)$ and model spectrum $F(\lambda)$

$$\log L = -(N/2) \log \left( \sum_{\lambda} (\Phi(\lambda) - F(\lambda))^2 \right)$$

where $N$ is the number of measured spectral fluxes.

There are several advantages to nested sampling. First, the algorithm naturally favors a minimal number of Gaussians and Planck blackbodies, which serves to prevent over-fitting the observed spectrum. More importantly, from the list of samples, one can compute mean quantities and uncertainties of the model parameter values as well as the evidence of the model itself.
4. RESULTS

We illustrate the technique using sample infrared emissions recorded from the Orion Bar which is within the Orion Nebula (Figure 1). The data used here were collected by the Infrared Space Observatory and processed by Sloan et al. (2003). The nested sampling algorithm worked with 200 objects and ran for 10,000 iterations. The results show (Figures 3 and 4) that nested sampling can begin to isolate the background from the PAH plus atomic emission. However, there are some difficulties. For instance, the dip at 15 microns and the structure beyond 38 microns suggest that the actual dust emission is more complex than can be modeled by Gaussians and blackbody functions. In addition, the current procedure tends to give excessive weight to errors in the fit at the longer wavelengths. This leads to the unfortunate outcome that the background is not reasonably defined in the shorter wavelength regions where important PAH emission occurs. This can be readily seen in right panel of Figure 4 where we show an expanded view of the 2.3 – 15 micron region from Figure 4. Clearly work needs to be done to better fit the background in this crucial region. We are currently working on refining the fitting procedure so as to give more weight to the PAH regions.

Our modeling approach does provide additional information. For example, the Planck term indicates that there are possibly (36% chance) two blackbody radiators: one with a temperature of 61.043 +/- 0.004K and the other in the range of about 20K. This type of additional quantitative information may prove valuable in understanding star-forming regions.

FIGURE 4. The left panel displays the signal of interest obtained from the Orion Nebula after removing the background. We observe several remaining difficulties due to the negative spectral intensities. The right panel focuses in on the PAH spectra in the range of 2.3-15 microns.
5. CONCLUSION

Separating a mixed spectrum is a formidable problem, especially when the possibility exists for so many chemical species. This problem is made even more difficult by the fact that we can only record one spectrum, and that there may be multiple clouds each with different physical conditions along the line-of-sight.

Despite the difficulty of the task, we aim to infer as much as possible about the astrochemical environment. To this end, the nested sampling algorithm with its ability to estimate evidence, which is critical for model testing, is central to our methodology. Here prior information is our greatest asset and we can incorporate a great deal into the form of the likelihood function (4), which depends on the spectral model we have adopted (3). Our future work will be to improve the ability of the algorithm to identify the background signal, as well as to identify and characterize individual PAHs by relying on the NASA Ames PAH library of over a thousand PAH spectra.

ACKNOWLEDGEMENTS

This research was supported by NASA AISR NNH05ZDA001N. Analysis based on observations with the Infrared Space Observatory, a European Space Agency (ESA) project with instruments funded by ESA Member States (especially the Principal Investigator countries: France, Germany, the Netherlands, and the United Kingdom) and with the participation of the Institute of Space and Astronautical Science (ISAS) and the National Aeronautics and Space Administration (NASA).

REFERENCES