Draft version June 28, 2016
Preprint typeset using LaTeX style emulateapj v. 12/16/11

[AUTOMATIC IDENTIFICATION OF SPECTRAL LINES THROUGH SPECTRUM RECONSTRUCTION]

Andrés Riveros [1], Karim Pichara [1,5,2], Pavlos Protopapas [2], Diego Mardones [3], and Mauricio Araya [4]
[1] Computer Science Department, Pontifícia Universidad Católica de Chile, Santiago, Chile
[2] Institute for Applied Computational Science, Harvard University, Cambridge, MA, USA
[3] Astronomy Department, Universidad de Chile, Santiago, Chile
[4] Computer Science Department, Universidad Técnica Federico Santa María, Valparaíso, Chile and
[5] Millennium Institute of Astrophysics, Chile
Draft version June 28, 2016

ABSTRACT

Astronomy is facing new challenges on how to analyze big data and therefore, how to search or predict events/patterns of interest. New observations in previously unexplored wavelength regions will be available from instruments such as the Atacama Large Millimeter Array (ALMA). Given this growing amount of high spectral resolution data, any non-automatic analysis would be an effort beyond human's capacity. Currently, classifying emission lines means to decide if a particular emission line belongs to a specific isotope. This classification is mainly done by comparing them with known isotopes emission lines. An automatic line-classification algorithm would dramatically reduce human efforts to analyze spectral data, allowing astronomers to focus their efforts in deeper analysis.

In this work, we propose an algorithm that uses a sparse model to represent the spectra and automatically classify emission lines. We use spectral line databases to determine a set of basis vectors that represent the presence of theoretical emission lines. Then, to classify lines in a given spectrum, the difference between the spectrum and a linear combination of the determined basis vectors is minimized. The model's output correspond to a probability vector representing the distribution of the prediction over a set of possible isotopes. We test our algorithm with experimental data from Splatalogue and simulated data from the ASYDO project. The results of the analysis show that the algorithm is able to identify emission lines with an accuracy of 90% when no blending nor hyperfine cases are present. As wavelength separation decreases (equal or less than 1 MHz), accuracy goes down to 82%.

The source code of the algorithm, synthetic data cubes used and list of suggested identifications are publicly available [a].

Keywords: spectral lines: emission lines; technique: spectroscopy; method: data analysis

1. INTRODUCTION

Modern astronomical observatories have brought increasing amounts of data over the last few years. Radiotelescopes's sensors have improved with higher resolution and wider wavelength ranges sensors. Instruments like the Atacama Pathfinder Experiment (APEX) (Gusten et al. 2006), the Sub-millimeter Array (SMA) (Ho et al. 2004), Heterodyne Instrument for the Far Infrared (HIFI) (de Graauw et al. 2004), Stratospheric Observatory For Infrared Astronomy (SOFIA) (Becklin 2006) and the Atacama Large Millimeter Array (ALMA) , provide higher resolution and details from submillimeter regions that will make this region very attractive for spectroscopy (Schilke et al. 2001; Müller et al. 2005; Schilke et al. 2011).

Imminent data growth and higher resolution will allow an analysis at a much more detailed level. This makes it impractical for astronomers to process and analyze all the data in a traditional, pedestrian way (Schilke et al. 2011; Skoda et al. 2014).

The traditional spectra analysis to identify emission lines involves to search similar lines characteristics, such as wavelength, intensity and the presence or absence of other observed lines for each possible isotope (Sharpee et al. 2003). Using both intuition and experience, astronomers estimate each possible presence of peaks to relate them to known molecules of interest (Schilke et al. 2011).

This process is very time-consuming and, consequently, it would be of a great help to have an automatic tool that contributes to the analysis. Several approaches to this problem have been proposed, including models that simulate molecular spectra (Schilke et al. 2001; Comito et al. 2005; Maret et al. 2010; Caux et al. 2015; Vastel et al. 2015), models that fit synthetic spectra with observed ones (Pequignot 1996; Walsh et al. 2003), and heuristic analysis of simultaneous presence of lines (Sharpee et al. 2003).

In this work, we propose an automatic line-classification algorithm to support astronomers in spectral data analysis. Our approach is by no means an exhaustive solution, but a way to reduce scientists's efforts in pre-classification of lines.

The proposed algorithm does not rely on any physical model about spectra, it just learns a suitable spectra representation from data. This representation allows the model to find line-detection patterns that constitute the key of automatic classification.

We work with three-dimensional velocity data cubes, i.e., intensity as a function of both position and velocity, or it's equivalent frequency, as shown in figure 1 (Eguchi 2013). The algorithm takes as input the observed spectra from a data cube, and gives as output a list of candidate emission lines present along the spectra.

[a] https://github.com/ChileanVirtualObservatory/DISPLAY

The algorithm has two general steps: i) detect a list of candidate frequency ranges; ii) confirm the candidate frequency ranges that belong to known isotopes. Step i) uses an heuristic that compares intensity differences along the spectra. Each consecutive pair of frequencies is evaluated by looking for intensity differences greater than a threshold parameter. This threshold parameter is given by the 3-σ criterion (three standard deviations over the random noise value on the data cube) (Sharpee et al. 2003). Therefore, a candidate emission line must meet two conditions: a) Its intensity has to surpass its predecessor/successor's intensity in more than 3-σ [1]. b) Its intensity must be above the 3-σ value. Step ii) uses a set of criteria to discern both the existence of a line together with its specific isotope.

This step is based on signal reconstruction, relying on a sparse representation. This representation makes use of a convenient set of basis vectors so that each one represents the presence of a theoretical emission line.

The convenient basis vectors set to reconstruct a spectrum is obtained by minimizing the difference between the spectrum and a linear combination of basis vectors. The minimization returns the magnitudes of the coefficients in the linear combination, where non necessary basis vectors are automatically discarded (their coefficients are zero). Each coefficient is associated to a specific known isotope, hence their resulting values represent a degree of match between a given isotope and the emission lines present in the spectrum. Normalizing the coefficient values result in a probability vector, representing a distribution of possible isotopes matching a given emission line. High entropy in a resulting distribution would show a low degree of certainty in the classification, while low entropy suggest that the model is very confident about its classification.

Spectral data can drastically vary between different sample measurements for the same object. Variability depends on physical properties such as internal/external rotation, observation angle and direction, and the specific radio-telescope used. This variability can affect both intensity and frequency of observed lines, the presence of some rotational sequence members and relative temperatures among emission lines (Howley et al. 2005). An example of the variability effect applied to NH_3 rotational sequence can be seen in figure 2.

High dimensionality and collinearity of spectral domain introduce problems for prediction models that rely on data, leading to over-fitting and degradation of prediction accuracy (Howley et al. 2005). The use of sparse coding makes intuitive sense, as those characteristics are a typical scenario in spectral data applications (Wright et al. 2010; Xiang et al. 2011). Specifically, at ALMA wavelength measure ranges, i. e., between 84 GHz and 950 GHz, spectral lines separation allow to consider this measures as sparse data.

This paper is structured as follows: In section 2, theoretical background is introduced both for spectroscopy theory and sparse coding modelling. In section 3, an overview of previous works is presented. Later, in section 4, the problem scope and the data origin are described. Then, section 5 details the proposed algorithm in greater depth. Finally, in section 6, results are shown and discussed, followed by the conclusion of this work in section 7.

[1] the predecessor/successor's may be an absorption line

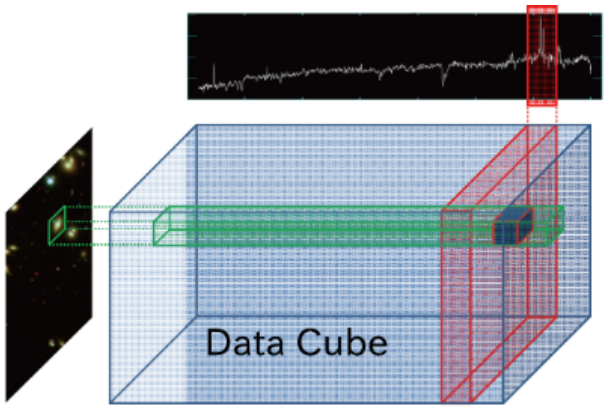

Figure 1. The schematic illustration of a data cube of ALMA, with two spatial dimensions and a frequency or wavelength one (Eguchi 2013).

2. BACKGROUND

2.1. *Spectroscopy*

Spectroscopy is a technique that enables the analysis of interaction between matter and light (Smith 2005). This analysis provides information on chemical structures and physical forms that can be used to identify substances from the characteristic spectral patterns. These patterns appear as light intensity peaks observed along the spectra, known as emission lines (Struve 1989). Each line has a different frequency depending on the molecule and energy level associated to it (Smith 2005).

Detection of emission lines and subsequent association with a molecule's isotope allow to know stellar objects's molecular structure. The combination of emission lines for each object generates an unique fingerprint. This allows to identify similar objects observing the similarities between observed spectra and theoretical known behavior of molecules (Howley et al. 2005).

A traditional process to identify lines consist on mapping observed frequencies to theoretical ones. Unfortunately, observed and theoretical intensities do not coincide. For instance, two spectral lines that are close in the frequency space are hard to dissociate, mainly because they appear as lines with double peaks or blended into one single line (Cernicharo et al. 2013; Smith et al. 2015).

Internal factors such as the temperature of objects, travel speed, type of astronomical object and its belonging to interstellar or intergalactic space (Sembach et al. 2001) cause variability in the observed spectra. However, the presence of lines within objects of similar composition are in general similar. Lines present from the same isotope at different state energy levels, and thus, different frequencies, reinforces the hypothesis that an isotope is present. This analysis of the presence and co-presence of spectral lines from the same isotope is known as rotational spectroscopy (Schilke et al. 2001).

There exists a relationships among intensity lines when homogeneous temperature and origin are assumed, i. e., local thermodynamic equilibrium (LTE) exists. For dif-

ferent stellar objects, and for different spectra measured within the same object, intensities of spectral lines with the same frequency may vary (Madden et al. 2005). However, relationships between intensities of rotational sequences for the same object should be consistent but not necessarily linear (Nummelin et al. 2000; Smith et al. 2015), as can be seen in figure 2.

Noise and lack of sensitivity in the measurements can either modify the frequency of a spectral line or produce false positives (Nummelin et al. 1998). This makes mandatory to involve astronomers to perform the identifications (Schilke et al. 2001).

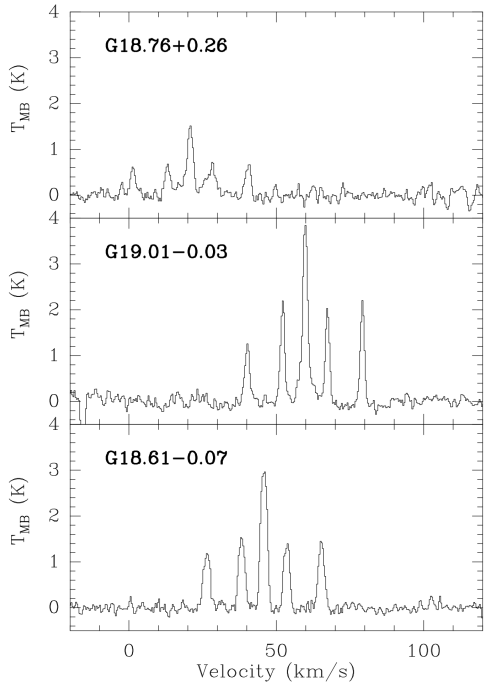

Figure 2. Example spectra in a NH_3 transition for three sources. We can see variations in velocity and intensity besides the similar structure (Schuller et al. 2009).

2.2. *Sparse coding*

Sparse coding consist in modelling signals through linear combinations of basis vectors under sparsity constraints, in order to have a sparse representation of input signals (Mairal et al. 2009; Bristow et al. 2013). Sparse coding aims to represent or recover a signal through a reconstruction procedure. We can divide this process in two main steps: i) find a set of fundamental components able to represent any possible signal by linearly combining the components; ii) for a given signal, determine the coefficients of the linear combination that minimize the difference between the signal and the linear combination of the components. Fundamental components set is known as *dictionary*, where each component is called a *word* (Mairal et al. 2009).

Formally, let $s = [s_1, s_2, \ldots, s_n]$ be a given signal, where $s_i \in [0,1] \ \forall i \in [1,\ldots,n]$. Let $D = \{w_1, w_2, \ldots, w_d\}$ be the dictionary, where $w_i = [w_{i1}, \ldots, w_{in}]$ and $w_{ik} \in [0,1] \ \forall i \in [1,\ldots,d] \ \forall k \in [1,\ldots,n]$. Let $\alpha = [\alpha_1, \ldots, \alpha_d]$ be the set of coefficients such that $s = \sum_{i=1}^{d} \alpha_i \ w_i$. Equation 1 corresponds to the optimization problem that has to be solved in order to determine the sparse coding coefficients.

$$\begin{aligned} &\text{Minimize}_\alpha ||s - \sum_{i=1}^{d} \alpha_i \ w_i||_2^2 \\ &\text{Subject to: } ||\alpha||_1 \leq \lambda \\ &\qquad\qquad \alpha \geq 0 \end{aligned} \tag{1}$$

where $||()||_k$ is the L-k norm, and λ is sparsity-inducing constant.

The solution to this problem is known as positive basis pursuit (Chen et al. 2001) or positive lasso regression (Efron et al. 2004). The optimization solution solved in closed form is very expensive in terms of processing (Mairal et al. 2009), instead, we use the iterative algorithm presented in Turlach et al. (2005); Mairal (2013).

There are two constraints applied to alpha values: i) the lambda sparsity-inducing constraint allows the algorithm to select a convenient set of basis vectors so that the number of non-zero values minimized; ii) the positive formulation of the problem allow us to give a meaningful use to the found set of alpha values; both are detailed in section 5.

In sparse coding, there exist two types of possible dictionaries to create the basis vector set: i) Previously defined one, where a set is selected accord to the nature of signal domain. ii) Automatically learned one, where methods such as clustering or another generalization searching are used (Mairal et al. 2009). For wavelength domains, predefined dictionaries give satisfactory results (Mallat 2009).

3. RELATED WORK

In the last years, detection of spectral lines has been following traditional methods, limited to manual analysis of data. This analysis primarily involves the estimation of frequencies associated to peaks in spectra, together with the mapping of those peaks to theoretical frequencies. After determining the presence of the lines at different frequencies, they associate those lines to certain isotope energy states. This non-automated process lacks of scalability, making impossible to apply this method for large databases (Schilke et al. 2001).

Spectra classification is found in other areas such as classification of substances, determination of raw material purity or even detection of skin cancer (Sigurdsson et al. 2004). Supervised machine learning classifiers have been proposed to separate different types of substances within spectra (Howley et al. 2005), but they are not designed to identify individual lines.

Several methods specialize in the individual detection of lines along the spectra. For example, EMILI software identifies spectral lines considering three features: i) wavelength agreement with observed line, ii) expected flux from relative computed intensities and iii) co-presence of other confirming lines. It assign numerical values to each criteria and calculates a score with them, both for observed lines and candidate theoretical lines. Then, probabilities are calculated for each candidate line. The near its score, the higher its probability

(Sharpee et al. 2003).

To fit functions to shape the optical depth of lines is a very common technique still widely used. Gaussian functions (Fuller and Myers 1993; Nummelin et al. 2000) or top-hat functions (Smith et al. 2015) are adjusted through the estimation of both the full width maximum height ($fwmh$) and peak intensities for line profiling. Then, a residual baseline offset is set to differentiate lines from signal artifacts.

Tools that fits more complex functions to spectra are XCLASS [2], CASSIS [3] and WEEDS [4]. These tools fit different functions along the spectra to estimate several parameters of observed lines. They build a line list fitting of all the transitions of an isotope through two steps: i) the fit of the line, ii) the fit of the baseline. For fitting of lines, CASSIS determines optimal parameter functions that allows to simultaneously fit all peaks along spectra. Gauss, Lorentz, Sinc and Voight are examples of functions for fitting lines. Then, in step ii), weaker lines are discarded fitting a baseline function, which can be sinusoidal or polynomial functions (Caux et al. 2015; Vastel et al. 2015)

These tools make use of catalogs of spectral lines that contains the theoretical frequencies of all known lines for each isotope. Those catalogs are publicly available, and are constructed by the data of (JPL [5], CDMS [6], Toyama [7]). Unfortunately, a lot of human effort is needed just to identify spectral lines in this way as (Schilke et al. 2011) states.

Furthermore, chemical and physical models takes into account the source structure through a modelling, using complex simulations to reproduce stars formation and later, spectral lines. These simulations involve two main steps: i) 3D chemical models and ii) radioactive transfer models (Schilke et al. 2011). In step i), the structure of object is modeled using molecular abundance, which is used either from provided values or from chemical models. An example of programs to get molecular abundance is RATRAN (Hogerheijde and van der Tak 2000). In step ii), the structure temperature of cores are estimated using Monte Carlo. Both **radme-3d** [8] and LIME (Brinch and Hogerheijde 2010) use this sampling simulation assuming LTE approximation. The first estimates object temperature, while the latter receives it as a parameter. An analysis of line shapes and temperature modeling allow to assign them to known isotope lines.

Regarding signal reconstruction models, approaches that compares observed spectra with synthetic modelling have been proposed in (Pequignot 1996; Walsh et al. 2003). These models have a rigorous treatment of blending and have the flexibility to deal with wavelength uncertainty from databases (Sharpee et al. 2003). Techniques described above are in general not scalable, do not rely on automatic processes or are based on complex theoretical underlying models.

[2] https://www.astro.uni-koeln.de/projects/schilke/XCLASS
[3] http://cassis.cesr.fr
[4] https://www.iram.fr/IRAMFR/GILDAS
[5] http://spec.jpl.nasa.gov
[6] http://www.astro.uni-koeln.de/cdms
[7] http://www.sci.u-toyama.ac.jp/phys/4ken/atla
[8] http://www.ita.uni-heidelberg.de/~dullemond/software/radmc-3d/

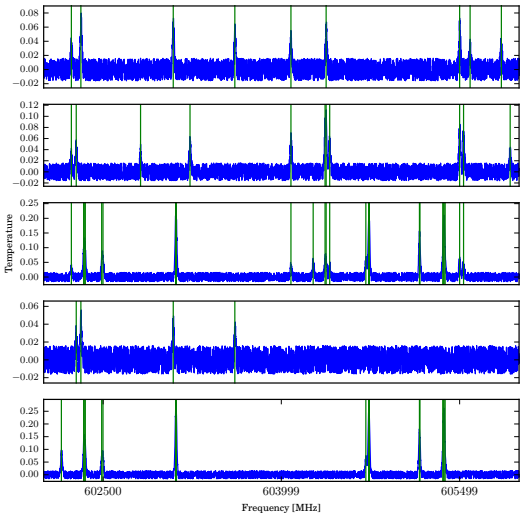

Figure 3. Example of ASYDO simulated spectra. Green vertical lines correspond to theoretical isotope lines used for the simulation.

4. DATA

4.1. *Synthetic data*

The solution proposed does not rely on underlying physics or chemistry, so it need enough data to find line-detection patterns. At this time, available spectral data from ALMA is not enough, hence the use of synthetic data is necessary.

In the next section, the tools to get synthetic data are introduced and also, its use in this project.

4.2. *Splatalogue. Catalogue of theoretical lines*

Splatalogue is the most up-to-date and complete spectral line database. It consists in a catalog of experimental lines that gives a list of all known frequency for isotopes at their different known transitions states (Remijan and Markwick-Kemper 2008). It aims to contain all known emission line data currently archived from labs all over the world - Jet Propulsion Laboratory (JPL), The Cologne Database for Molecular Spectroscopy (CDMS) (Müller et al. 2005), Lovas National Institute of Standards and Technology (NIST), among others sources (Remijan 2010).

Also, it allows to filter and search for spectral lines by isotope and wavelength ranges. This tool is important for this work, since it is used to create the dictionary.

4.3. *ASYDO. Synthetic data*

The Astronomical Synthetic Data Observations (ASYDO) package [9] is used to simulate ALMA-like data. The simulation generates a training set to develop and test identification accuracy.

ASYDO can create fits files containing simulated hypothetical stellar objects using the next parameters:

- **Isolist** : subset isotope list to generate a cube
- **Parameters**

[9] https://github.com/ChileanVirtualObservatory/ASYDO

- **freq** : spectral center (MHz)
- **spe_res** : spectral resolution (MHz)
- **spe_bw** : spectral bandwidth (MHz)
- **(fwhm, α -skew)**: skew-normal distribution parameters (MHz, parameter)

Skew-normal function gives form to spectral lines, in which $fwhm$ is full width at half maximum, and $\alpha - skew$ is its kurtosis parameter. If $\alpha - skew = 0$, it degenerates to a Gaussian function, if $\alpha - skew < 0$, it is left-biased and $\alpha - skew > 0$, a right bias.

We assume object movement redshift as known and corrected. A previous step is necessary to identify a set of stronger lines in the spectra and determine velocity shift Sharpee et al. (2003). Eliminating general redshift just left two error margins for observed frequencies: noise and internal redshift given by rotation and internal movements.

Each band has different noise because of both radio-telescope sensitivity at each band, and nature of spectra signals at different wavelengths.

The width of each line depends on skew-normal parameter $fwhm$. For testing purposes, we modified the width in a 4 MHz range with a modification of ASYDO, incorporating this randomness to use different width for each spectral line. The parameters we use for simulations are: both **alpha** and **delta** as 0 *degrees*, **spe_res** of 1 MHz, **spe_bw** as 4000 MHz and **(fwhm, α -skew)** as $\mathbf{(8, 0)}$.

5. ALGORITHM

Our algorithm has two main steps: i) spectra pre-processing, which also involves the creation and recalibration of the dictionary, covered in section 5.1. ii) optimization of equation 1, which allow us later to predict emission lines present along the spectra, viewed at detail in section 5.2. An overview of the process is illustrated in figure 4.

5.1. *Pre-processing*

At first, a dimensional pixel from a data cube is selected to analyze its wavelength range, as shown in figure 4(a). Then, a normalization and filtering of spectra is performed. Savitzky-Golay filter is applied to reduce white noise influence along the signal(Howley et al. 2005). Normalization does not have any effects in the sparse coding solution, however, it is applied for convenient purposes, as we will explain later. Figure 4(b) illustrates the output after the preprocessing step.

5.1.1. *Delta Dirac function*

In this stage, we perform the following steps: i) select all theoretical lines for all isotopes present in range of measurement. ii) create delta Dirac vectors for each theoretical line previously selected. Delta direct functions allow to determine a representative and meaningful dictionary for this problem. One word is defined for each theoretical frequency known in spectra wavelength range. This formulation allows to represent each theoretical frequency range with a specific word in the dictionary.

Let $\boldsymbol{D} = \{\boldsymbol{w_1}, \boldsymbol{w_2}, \ldots, \boldsymbol{w_d}\}$ be the dictionary, where $\boldsymbol{w_i} = [\boldsymbol{w_{i1}}, \ldots, \boldsymbol{w_{in}}]$ and $\boldsymbol{w_{ik}} \in [\mathbf{0}, \mathbf{1}]$ $\forall \boldsymbol{i} \in [\mathbf{1}, \ldots, \boldsymbol{d}]$ $\forall \boldsymbol{k} \in [\mathbf{1}, \ldots, \boldsymbol{n}]$. Let $\boldsymbol{F} = \{\boldsymbol{f_1}, \boldsymbol{f_2}, \ldots, \boldsymbol{f_n}\}$ be the set of frequencies at the range of measure. The value of each element of $\boldsymbol{w_i}$ is given by the function:

$$\boldsymbol{w_{in}} = \begin{cases} \mathbf{1}, & \text{if } \boldsymbol{f_i} \text{ is the theoretical frequency of isotope } \boldsymbol{i} \\ \mathbf{0}, & \text{otherwise} \end{cases} \tag{2}$$

Figure 4(c) lists all the theoretical frequencies combined along the spectra in range of measure.

Hyperfine lines are a particular case, in which two close theoretical lines that belong to same isotope are present. In general they are both present as one wider line (see figure 5).

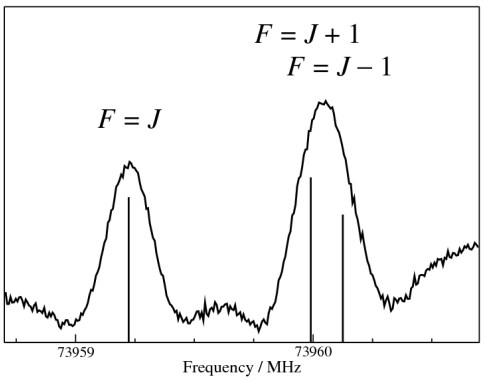

Figure 5. Rotational spectrum of vinyl cyanide transition of $H_2^{13}C{=}CHCN$ showing hyperfine unseparable structures (Müller et al. 2008).

To deal with hyperfine lines, we combine their delta Dirac functions and merge them into a single word. Sensitivity of data determines when two hyperfine lines should merge as one. As we use 1 MHz for sensitivity, two words are merged if they are closer that 1 MHz and belong to the same isotope.

A problem with the use of Dirac Delta functions is that observed lines must be at the precise observed frequencies to be used. If a difference between a delta Dirac function and its observed candidate line exists, it is impossible for the sparse coding optimization to use the theoretical word to reconstruct the shifted observed frequency. This makes necessary to adjust the previous words, so that a soft matching can be possible. If a word is not at the exact frequency than a candidate line's frequency, but near it, the word still can be used, but with a loss of confidence.

Two steps are necessary to make this adjustment to the dictionary: i) detect all the candidate lines along the observed spectra, ii) expand each word in the recalibration step.

5.1.2. *Candidate emission lines*

We use an heuristic to pre-define frequency ranges at which further steps evaluate the presence or non-presence of emission lines. A peak detection function is ran to select these ranges associated to possible lines along the spectra. We call these ranges candidate emission lines.

We make use of a threshold given by the 3-$\boldsymbol{\sigma}$ criterion. An empty pixel is selected from a spectra of the data cube in which the observed object is not present, so that just background and noise is observable. Then, we compute the mean and standard deviation of the noise. With this,

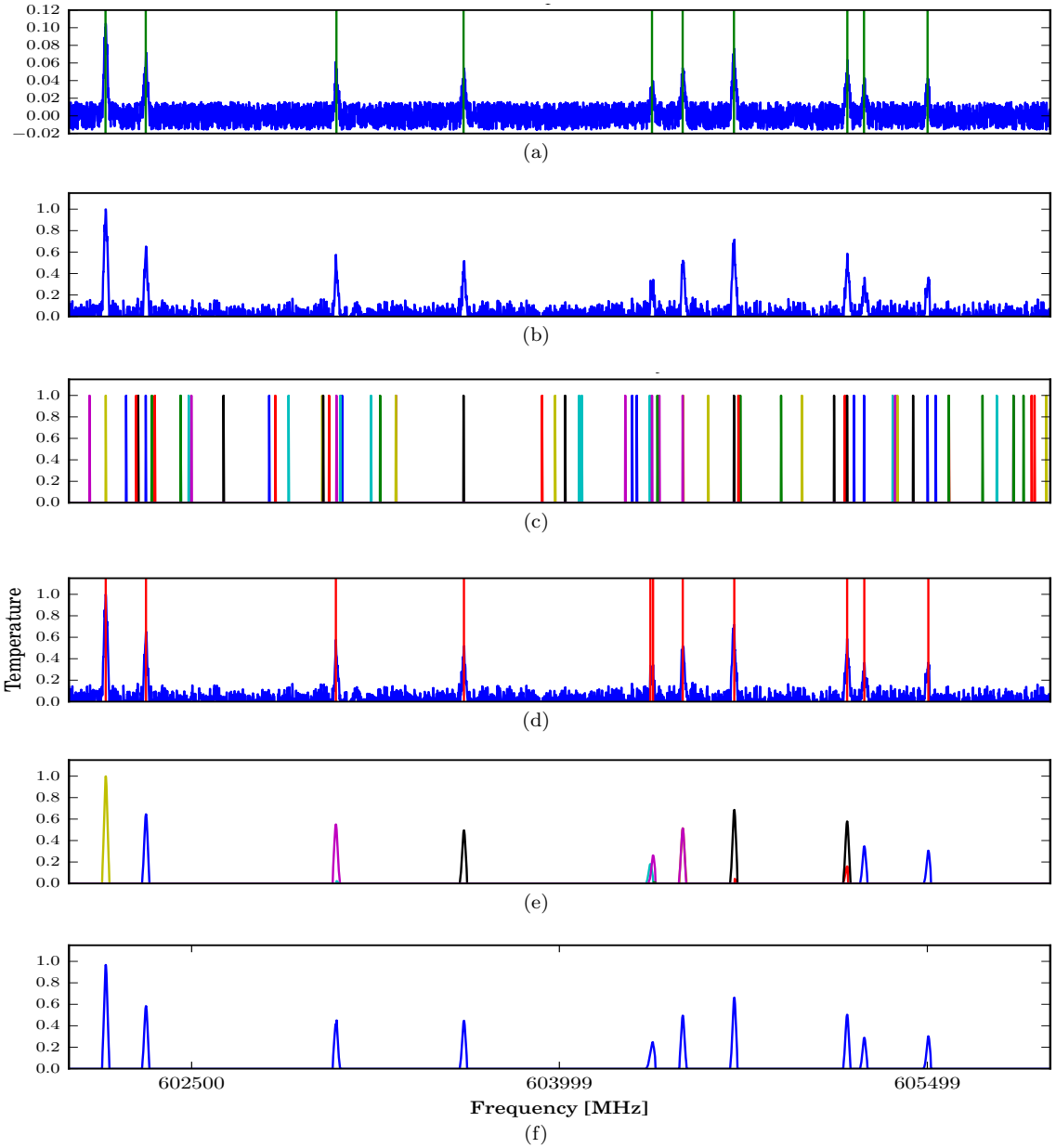

Figure 4. Spectra example though the whole process. Each step: 1. Pre-processing: (a) Read raw data cube, (b) Filter and normalize, (c) Determine dictionary, (d) Detected candidate lines, (e) Recalibrate the dictionary. 2. (f) Reconstruction of the signal.

we search for intensity differences between each consecutive pair of frequencies, and when the difference between the temperature of a frequency and the previous temperature is higher that the threshold, the frequency from higher temperature is saved as candidate line.

Following this idea, an iterative process is performed. All the peaks are detected from the original spectra and the frequency of the higher intensity is saved as a candidate emission line. Then, a Gaussian function is fitted at the detected frequency and subtracted from the signal. The process is repeated until the higher intensity of the detected peaks is less than the 3-$\boldsymbol{\sigma}$ threshold. At the end of the process, a list of candidate lines is determined, as can be seen in figure 4(d).

5.1.3. *Recalibration*

At the end of pre-processing step, the dictionary passes for a step of recalibration, where each word is expanded to a range from the initial Dirac delta function. We use an exponential kernel function that assign values to each word depending on the distance between theoretical frequencies of the words and their nearest candidate lines. Word's expansion allows to associate probabilities to matches, which are also used to combine several words at certain frequencies and to replicate blended lines.

In recalibration step, we introduce the use of candidate line's temperature to weight words according to the intensity of the nearest candidate lines. This reflects that smaller candidate lines are less probable to be emission lines as they get closer in intensity to the threshold. The

final value of a word is given by equation 4.

Let $\boldsymbol{s} = [\boldsymbol{s_1}, \boldsymbol{s_2}, \ldots, \boldsymbol{s_n}]$ be a normalized signal, where $\boldsymbol{s_i} \in [\boldsymbol{0}, \boldsymbol{1}]\ \forall \boldsymbol{i} \in [\boldsymbol{1}, \ldots, \boldsymbol{n}]$. Let $\boldsymbol{D} = \{\boldsymbol{w_1}, \boldsymbol{w_2}, \ldots, \boldsymbol{w_d}\}$ be the dictionary, where $\boldsymbol{w_i} = [\boldsymbol{w_{i1}}, \ldots, \boldsymbol{w_{in}}]$ and $\boldsymbol{w_{ik}} \in [\boldsymbol{0}, \boldsymbol{1}]\ \forall \boldsymbol{i} \in [\boldsymbol{1}, \ldots, \boldsymbol{d}]\ \forall \boldsymbol{k} \in [\boldsymbol{1}, \ldots, \boldsymbol{n}]$. Let $\boldsymbol{F} = \{\boldsymbol{f_1}, \boldsymbol{f_2}, \ldots, \boldsymbol{f_n}\}$ be the set of frequencies at the range of measure. Function $\boldsymbol{c(f)}$ is defined as $\boldsymbol{c(f_i)} = \boldsymbol{s_i}, \forall \boldsymbol{i} \in [\boldsymbol{1}, \ldots, \boldsymbol{n}]$, i. e., signal's intensity at frequency $\boldsymbol{f_i}$. $\boldsymbol{g}$ is defined in 3 as the closer candidate line's frequency to a given frequency $\boldsymbol{f_i}$, such as for a word $\boldsymbol{w_{ki}}$:

$$\boldsymbol{g} = \text{argmin}_{\boldsymbol{f}} ||\boldsymbol{f_i} - \boldsymbol{f}|| \tag{3}$$

and a word's expansion is given by equation 4

$$\boldsymbol{w_{ki}} = \boldsymbol{c(g)} \cdot \frac{\boldsymbol{1}}{\sqrt{\boldsymbol{2\pi\sigma}}} \boldsymbol{e}^{-\frac{1}{2}(\frac{||\boldsymbol{f_i} - \boldsymbol{g}||)^2}{\boldsymbol{\sigma}}} \tag{4}$$

For simplicity sake, this case uses $\boldsymbol{\sigma}$ value as 1 that works well in the ALMA domain we analyzed.

The final representation of the dictionary can be seen in figure 4(e). The majority of the words are expanded to low values and only words closer to candidate lines have appreciable values.

5.2. *Prediction*

The optimization of equation 1 gives a set of convenient alpha values to reconstruct the observed spectra at ranges of interest. After the reconstruction, at each frequency along the reconstructed spectra, a subset of used words can be obtained. Alpha values different than zero are used to assign possible isotopes to each detected line.

Sparse coding select the minimal amount of alpha values different than zero, so that combined reconstruct the normalized signal. This amount of non-zero values is restricted by the Lambda sparsity-inducing parameter, which is experimentally determined as the number of detected candidate lines. This makes sparse coding to use a similar number of words as candidate lines were detected.

An important restriction must be applied to the alpha values. At the convenient solution of the optimization formulation, non-zero values must be positive to be able to detect emission lines, preventing the use of both negative and positive words. If not, the word's meaning as presence of emission lines would be lost, resulting in over fitting and false positive predictions.

5.2.1. *Probability of prediction*

At candidate line's frequencies, all non-zero alphas that are near to those frequency are used to determine a probability list. The superposition of words is used to deal with blending or false double peaks cases.

Spectra normalization is a convenient convention to give a meaning to alpha values scooped at range $(\boldsymbol{0}, \boldsymbol{1})$. If its value is near to 1, its word is used unscaled, and it has an higher probability to be describing candidate lines. If an alpha value is closer to 0 or has a value higher than 1, to make use of its word is harder for the optimization. A symmetric convention allows to assign the same importance to alpha values bellow and over 1. Let $\boldsymbol{\alpha} = [\boldsymbol{\alpha_1}, \ldots, \boldsymbol{\alpha_d}]\ \forall \boldsymbol{i} \in [\boldsymbol{1}, \ldots, \boldsymbol{d}]$ be the set of coefficient values for each theoretical isotope state.

$$\boldsymbol{\alpha_k^*} = \begin{cases} \boldsymbol{\alpha_k}, & \text{if } \boldsymbol{\alpha_k} \leq \boldsymbol{1} \\ \boldsymbol{1/\alpha_k}, & \text{if } \boldsymbol{\alpha_k} > \boldsymbol{1} \end{cases} \tag{5}$$

Alphas at each frequency, and its subsequent normalization (by the sum of all alphas used in that frequency), give a probability distribution over possible isotopes. The probability of presence for theoretical line $\boldsymbol{k}$, at a given frequency $\boldsymbol{i}$, for $\boldsymbol{D}$ isotope states, is given by 6

$$\boldsymbol{P_{ik}} = \frac{\boldsymbol{\alpha_k^*}}{\sum_{\boldsymbol{j=1}}^{\boldsymbol{d}} \boldsymbol{\alpha_j^*}} \tag{6}$$

Finally, the use of multiple adjacent pixels in the same cube allows to get a more reliable prediction, excluding false positives from the prediction. This is done by multiplying the probabilities of isotopes presence for each analyzed spectra. Let $\boldsymbol{x} \in [\boldsymbol{n}, \ldots, \boldsymbol{n} + \boldsymbol{m}]$, $\boldsymbol{y} \in [\boldsymbol{v}, \ldots, \boldsymbol{v} + \boldsymbol{w}]$,

$$\boldsymbol{P_{ik}} = \prod_{(\boldsymbol{n},\boldsymbol{v})}^{(\boldsymbol{n+m},\boldsymbol{v+w})} \boldsymbol{P_{ik}(x, y)}) \tag{7}$$

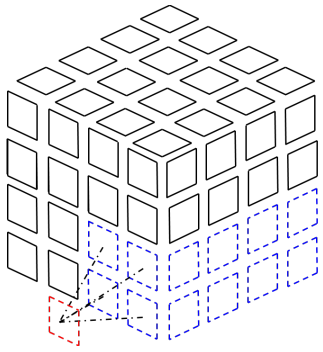

Figure 6. The schematic convergence of predictions from a data cube, where 4 spectras (blue pixels in frequency deep) are analyzed independently, and their results merged into one (red).

The algorithm pseudo-code summary can be seen at 1.

Data: ***data_cube***, ***isotopes_set***
Result: ***probability_predictions***
get parameters from ***data_cube***;
get input spectra from pixel of ***data_cube***;
compute ***threshold*** from noise;
get ***Dictionary*** from ***isotopes_set***;
initialize ***candidate_set*** as [( )];
detect ***max_set*** from spectra;
$\boldsymbol{max_{freq}} = \boldsymbol{max(max\_set)}$;
while $(\boldsymbol{max_{freq}} > \boldsymbol{threshold})$ **do**
  (***candidate_set***).append($\boldsymbol{max_{freq}}$);
  subtract Gaussian function at $\boldsymbol{max_{freq}}$;
  detect ***max_set*** from residual spectra;
  $\boldsymbol{max_{freq}} = \boldsymbol{max(max\_set)}$;
end
recalibrate ***Dictionary*** from ***candidate_set***;
get ***alphas*** from solving sparse coding;
compute ***probability_predictions*** from ***alphas***;
return ***probability_predictions***;

Algorithm 1: Proposed algorithm

5.3. *Training/Test set*

For testing purposes, each lines present in the simulated spectras are stored as data cube meta-data, allowing us to evaluate the predictive model and determine metrics to validate predictions.

For this, and given the ASYDO simulator capabilities, data cube specifications are separated in two main parameters: i) signal to noise ratio, using different ALMA bands ii) fixed/variable width of lines, For i), the differences lie in both signal to noise ratio and spectral line density. ALMA bands 7 and 9 are selected for experiments, and the test consists in 50 cubes, in which half of them are run for different subsets of isotopes present at both ranges 602 - 606 $\boldsymbol{GHz}$ (ALMA band 9) and 275 - 277 $\boldsymbol{GHz}$ (ALMA band 7). For ii), 50 test run on the same bands, but using variable line width. Each line width is assigned independently in a range of variation of $\boldsymbol{(-2,+2)MHz}$ from original width.

6. EXPERIMENTAL RESULTS

In this section, we present and analyze the experimental results. The idea behind these tests is to simulate data cubes using a known isotope list and then to retrieve of as much elements of the known list as possible. For each band, the list of all theoretical isotopes in that range are searched in Splatalogue, and a subset of them is selected randomly to simulate data cubes.

6.1. *Measure of accuracy*

To evaluate identification performance, a measure of test accuracy must be design such that we can evaluate percentage of matches between selected words and present lines in simulations. We use as measures precision, recall and f-score in view of its intuitive ability to explain performance differentiating accuracy from true/false positives/negatives (Perry et al. 1955). F-score values in table 1 show an overview of obtained results.

Moreover, confusion matrices in figure 7 allow to visually analyze the classifier performance. The matrices, that corresponds to first 20 experiments at ALMA band 9, show that predictions become less certain at darker zones. Confusion matrices tends to be higher than wider because of the greater number of false positive predictions over actual isotope lines. Indicators of performance precision, recall and f-score are calculated from their respective confusion matrices. The precision/recall curves are shown at figures 8 and 9 for ALMA bands 9 and 7 respectively.

6.2. *Prediction results*

Overall results give a f-score above 90% when the results are filtered for cases in which the lines present in the simulation are higher than 1 MHz. One might expect higher results in such that cases, but the fact that present lines are not closer does not necessarily simplifies the task. Theoretical lines keep being very close for some cases and an error margin is expected.

When all results are included, an overall f-score of 82% is reached, showing that the idea behind this approach is suitable to solve the problem. In next sections, we will address differences between each group of results.

ALMA Band/Width	Fixed	Variable
Band 7	85.80 %	84.79 %
Band 9	82.05 %	78.17 %

Table 1
The f-score for different noise level (bands) and width of the lines.

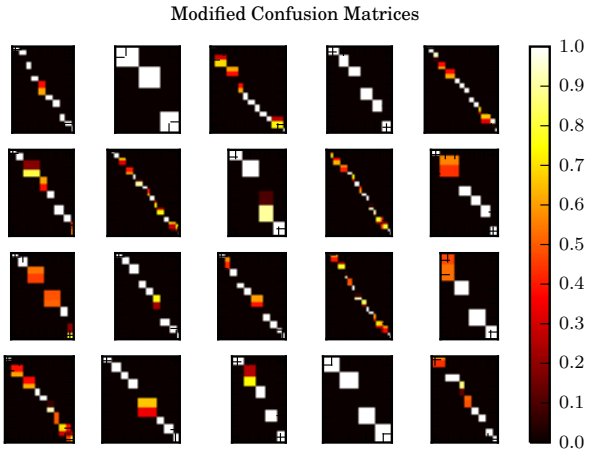

Figure 7. Modified confusion matrices for 20 experiments for data cubes with fixed line width at band 9. Predictions tend to be on the diagonal because of the algorithm preference for closer theoretical line's frequencies.

6.3. *Signal to noise effect in predictions*

There exists noticeable differences between results for different noise levels. For band 9, an overall of 80% shows that both the higher noise and density affect prediction accurately. On the other hand, band 7 reaches an overall of 85%, although there are not appreciable differences between the measures distribution as can be seen in both figures 10 and 12.

Figures 8 and 9 show an intuition of exchange ratio between true positive and false negatives. Precision/recall curve at band 7 has a better trade-off as its slope is smaller, and this is reflected in better prediction of true positives without increasing false positives.

6.4. *Variable width effect in predictions*

To test more realistic cases, where line width variates randomly, there is a small difference of almost 1% for band 7. Not so at band 9, where a difference of 4% shows that higher density of lines is affected by the randomness of lines's width. Also, figures 10 and 12 shows the differences between accuracy measures for both cases, being the fixed width focused in a smaller range than variable width results.

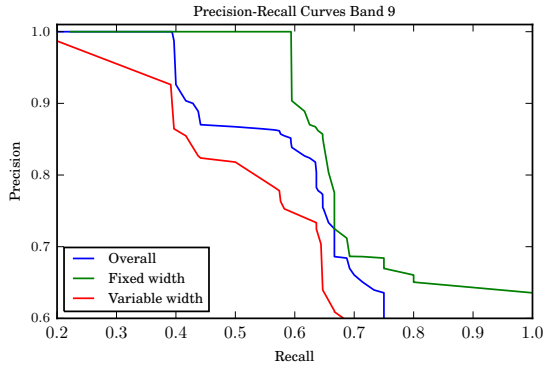

Figure 8. The measures of accuracy, precision and recall obtained for fixed line width cubes, variable line width and overall results for ALMA band 9.

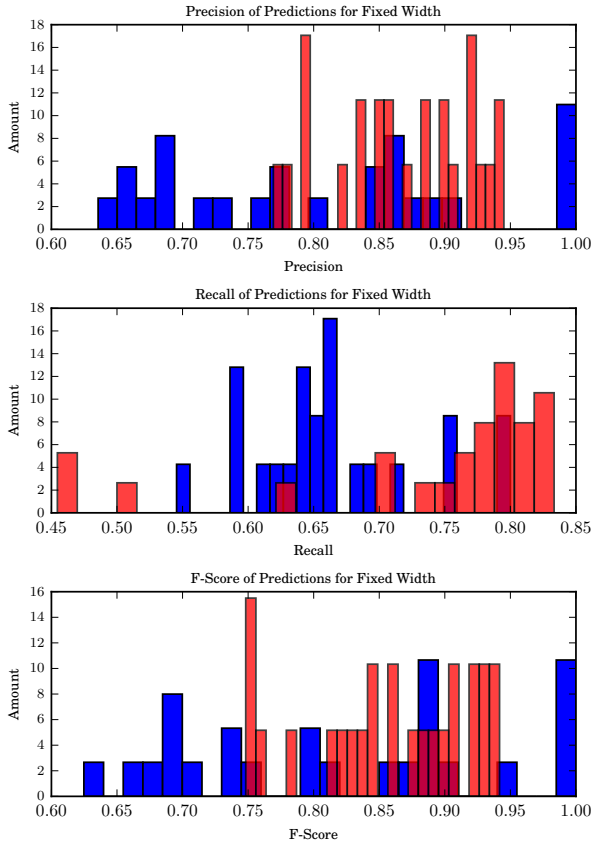

Figure 10. Histograms of results obtained for the test performed for precision, recall and f-score for fixed line width (red) vs variable line width (blue) in Band 9.

6.5. *Complex cases*

We focus our analysis on complex cases and show examples of how the algorithm handle them.

For blending cases, the algorithm gives a probability distribution of potential overlapped lines. In general, when blending exists, one of the predicted lines losses certainty, as showed in figure 11.

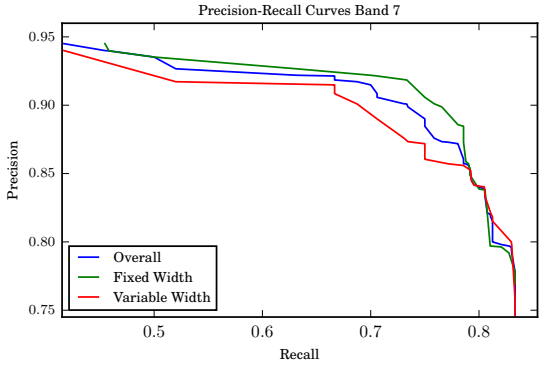

Figure 9. The measures of accuracy, precision and recall obtained for fixed line width cubes, variable line width and overall results for ALMA band 7.

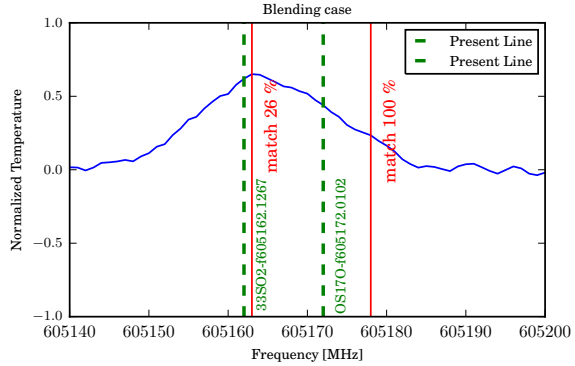

Figure 11. Blending case

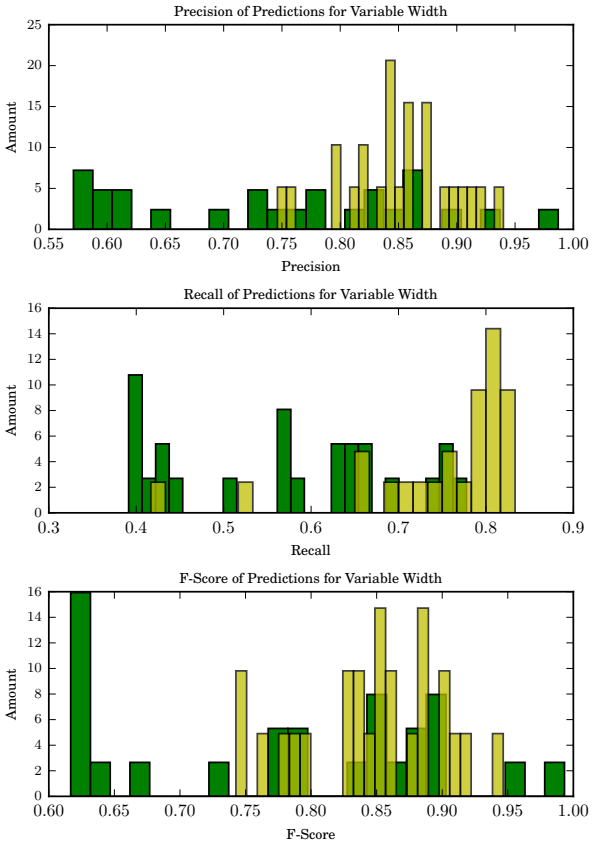

Figure 12. Histograms of results obtained for the test performed for precision, recall and f-score for fixed line width (yellow) vs variable line width (green) in Band 7.

False double peaks product of artifacts are handled by the algorithm and it determines the correct lines among false peaks, as showed in figure 13.

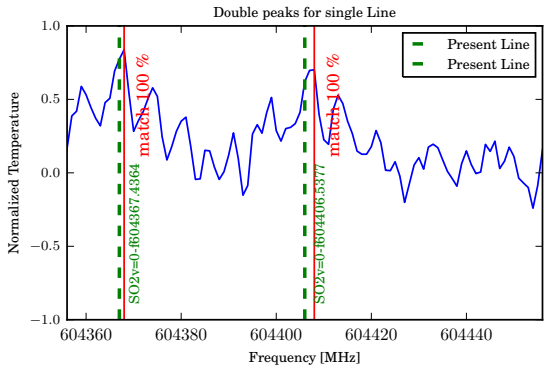

Figure 13. Double peaks for single Line

6.6. *Real Data*

Additionally to synthetic results, real data from ALMA is used to test the algorithm's behavior. The experiments consist in the analysis of product cubes associated to molecules observed for each cube. The used cubes are product of a deconvolution using CASA task CLEAN, as seen at Higuchi et al..

File	Predicted frequencies
$13CH3CN19-18$	-
$CH3OH7-6$	338486.337 and 338486.337, 338583.195, 338639.939
CS_v1_7-6	-
$SO2$	-
$SO2-28\_2\_26-28\_1\_27$	-

Table 2
Predicted frequencies associated to the molecules of interest for each clean cube for ALMA project #2011.0.00419.S.

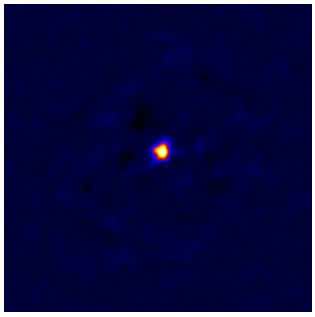

Figure 14. Example of used data cubes, slice of $IRAS16547-4247\_Jet\_CH3OH7-6.clean$ from project ALMA project #2011.0.00419.S.

In this case, the high density of theoretical frequencies plays a major role in the difficulty of the prediction. With synthetically data, a subset of isotopes were selected, but with real data, all the Splatalogue lines were include to test the process in a real world scenario.

The algorithm gives a list and a probability of presence in each cube. In table 2 we present only predicted frequencies associated to the isotopes of interest. The algorithm is able to predict lines for $CH_3OH_7(7-6)$, but the algorithm is not able to predict other isotopes of interest.

The process for file $CH_3OH_7(7-6)$ can be seen in figure 15. The amount of theoretical frequencies that must be filtered in order to make a prediction can be seen explicitly in the visual representation of the process.

7. CONCLUSION

Our approach to identify emission lines is the reconstruction of an input signal. The combination of representative basis vectors allows us to predict the presence of isotopes lines along the wavelength spectra. The set of coefficients used to reconstruct the signal give us an idea of the presence of each isotope line.

This process can be summarized as two main hypothesis: i) a set of basis vectors representing theoretical lines allows to reconstruct an input spectra, and ii) the used combination of basis vectors gives useful information to identify the presence of emission lines along the spectra.

Results has shown support for the hypothesis, but leaves room for improvements that increases its possibilities given the near arrival of new real data. This data will allow to future investigators to train models and capture actively the patterns in data, its correlations and hidden latent variables.

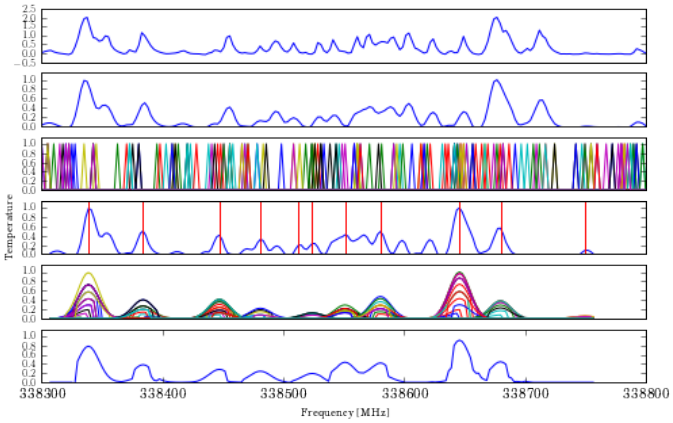

Figure 15. Data cube example following the steps previously explained.

Sparse coding technique allows to identify isotope lines even when blending is present. This gives a notion of the molecular composition of the astronomical object and allows astronomers to focus on complex cases.

A major issue in the algorithm elaboration is the lack of information about relative intensities relationships and the co-presence dependence for lines of the same isotope. That makes the algorithm to try to find each isotope line independently. Future extensibility from real data can be: i) the inclusion of relationship between temperatures of lines belonging to the same isotope, and ii) to learn the dependence of co-presence of lines, not just for the same isotopes, but for all molecules. On that line, theoretical lines belonging to unknown molecules are an interesting case to cover. The possible relationships between unidentified lines and known molecules could be used as a way to assign unknown lines to an isotope.

The solution proposed resulted in a first approach to solve this problem. Real data will give to researchers new tools to analyze and develop more complex models to make use of patterns that simulations do not allow us to use. Certainly, future work can make use of a big amount of data available with the forward of ALMA project to apply more complex word representations and signal reconstruction models.

ACKNOWLEDGMENTS

This investigation is supported by Vicerrectora de Investigación (VRI) from Pontificia Universidad Católica de Chile. Institute of Applied Computer Science at Harvard University.

This work is funded by project FONDEF D11I1060: Development of an astro-informatics platform for management and intelligent analysis of large-scale data, a collaborative project between several Chilean universities in order to create a Chilean virtual observatory, called Observatorio Virtual Chileno (ChiVO) [10].

This paper makes use of the following ALMA data: ADS/JAO.ALMA#2011.0.00419.S. ALMA is a partnership of ESO (representing its member states), NSF (USA) and NINS (Japan), together with NRC (Canada), NSC and ASIAA (Taiwan), and KASI (Republic of Korea), in cooperation with the Republic of Chile. The Joint ALMA Observatory is operated by ESO, AUI/NRAO and NAOJ.

[10] http://www.chivo.cl

REFERENCES

E. E. Becklin. Stratospheric observatory for infrared astronomy (SOFIA). In *Astrochemistry: Recent Successes and Current Challenges*, volume 1 of *Proceedings of the International Astronomical Union*, pages 323–324, 2006. doi:10.1017/S1743921306007332. URL http://journals.cambridge.org/article_S1743921306007332.

C. Brinch and M. R. Hogerheijde. LIME - a flexible, non-LTE line excitation and radiation transfer method for millimeter and far-infrared wavelengths. *Astronomy & Astrophysics*, 523, 2010. ISSN 0004-6361, 1432-0746. doi:10.1051/0004-6361/201015333. URL http://arxiv.org/abs/1008.1492.

H. Bristow, A. Eriksson, and S. Lucey. Fast convolutional sparse coding. In *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. doi:10.1109/CVPR.2013.57.

E. Caux, S. Bottinelli, C. Vastel, and J. M. Glorian. CASSIS, a software package to analyse high spectral resolution observations. volume 280, page 120P, 2015. ISBN 1743-9221. URL http://adsabs.harvard.edu/abs/2011IAUS..280P.120C.

J. Cernicharo, B. Tercero, A. Fuente, J. L. Domenech, M. Cueto, E. Carrasco, V. J. Herrero, I. Tanarro, N. Marcelino, E. Roueff, M. Gerin, and J. Pearson. Detection of the ammonium ion in space. *Astrophysical Journal Letters*, 2013. URL http://arxiv.org/abs/1306.3364.

S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001. ISSN 0036-1445. doi:10.1137/S003614450037906X. URL http://epubs.siam.org/doi/abs/10.1137/S003614450037906X.

C. Comito, P. Schilke, T. G. Phillips, D. C. Lis, F. Motte, and D. Mehringer. A molecular line survey of orion KL in the 350 micron band. *The Astrophysical Journal Supplement Series*, 156:127–167, 2005. ISSN 0067-0049. doi:10.1086/425996. URL http://adsabs.harvard.edu/abs/2005ApJS..156..127C.

T. de Graauw, E. Caux, R. Gusten, W. Jellema, W. Luinge, J. Pearson, T. Phillips, R. Schieder, J. Stutzki, K. Wafelbakker, Nick Whyborn, and K. Wildeman. The herschel-heterodyne instrument for the far-infrared (HIFI). In *Conference Digest of the 2004 Joint 29th International Conference on Infrared and Millimeter Waves, 2004 and 12th International Conference on Terahertz Electronics, 2004.*, pages 579–580, 2004. doi:10.1109/ICIMW.2004.1422223.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004. ISSN 0090-5364, 2168-8966. doi:10.1214/009053604000000067. URL http://projecteuclid.org/euclid.aos/1083178935.

Satoshi Eguchi. "superluminal" FITS file processing on multiprocessors: Zero time endian conversion technique. *Publications of the Astronomical Society of the Pacific*, 125 (927):565–579, 2013. ISSN 00046280, 15383873. doi:10.1086/671105. URL http://arxiv.org/abs/1304.5302.

G. A. Fuller and P. C. Myers. Thermal material in dense cores: A new narrow-line probe and technique of temperature determination. *The Astrophysical Journal*, 418:273, 1993. ISSN 0004-637X, 1538-4357. doi:10.1086/173389. URL http://adsabs.harvard.edu/doi/10.1086/173389.

R. Gusten, R. S. Booth, C. Cesarsky, K. M. Menten, C. Agurto, M. Anciaux, F. Azagra, V. Belitsky, A. Belloche, P. Bergman, C. De Breuck, C. Comito, M. Dumke, C. Duran, W. Esch, J. Fluxa, A. Greve, H. Hafok, W. Hupl, L. Helldner, A. Henseler, S. Heyminck, L. E. Johansson, C. Kasemann, B. Klein, A. Korn, E. Kreysa, R. Kurz, I. Lapkin, S. Leurini, D. Lis, A. Lundgren, F. Mac-Auliffe, M. Martinez, J. Melnick, D. Morris, D. Muders, L. A. Nyman, M. Olberg, R. Olivares, M. Pantaleev, N. Patel, K. Pausch, S. D. Philipp, S. Philipps, T. K. Sridharan, E. Polehampton, V. Reveret, C. Risacher, M. Roa, P. Sauer, P. Schilke, J. Santana, G. Schneider, J. Sepulveda, G. Siringo, J. Spyromilio, K.-H. Stenvers, F. van der Tak, D. Torres, L. Vanzi, V. Vassilev, A. Weiss, K. Willmeroth, A. Wunsch, and F. Wyrowski. APEX: the atacama pathfinder EXperiment. volume 6267, pages 626714–626714–26, 2006. doi:10.1117/12.670798. URL http://dx.doi.org/10.1117/12.670798.

Aya E. Higuchi, Kazuya Saigo, James O. Chibueze, Patricio Sanhueza, Shigehisa Takakuwa, and Guido Garay. IRAS 165474247: A new candidate of a protocluster unveiled with ALMA. 798(2):L33. ISSN 2041-8205. doi:10.1088/2041-8205/798/2/L33. URL http://stacks.iop.org/2041-8205/798/i=2/a=L33.

Paul T. P. Ho, James M. Moran, and Kwok Yung Lo. The submillimeter array. *The Astrophysical Journal*, 616(1), 2004. ISSN 0004-637X, 1538-4357. doi:10.1086/423245. URL http://arxiv.org/abs/astro-ph/0406352.

M. R. Hogerheijde and F. F. S. van der Tak. An accelerated monte carlo method to solve two-dimensional radiative transfer and molecular excitation. with applications to axisymmetric models of star formation. *Astronomy and Astrophysics*, 362: 697–710, 2000. ISSN 0004-6361.

T. Howley, M. G. Madden, M. O Connell, and A. G. Ryder. The effect of principal component analysis on machine learning accuracy with high dimensional spectral data. 2005. URL http://aran.library.nuigalway.ie/xmlui/handle/10379/194.

M. G. Madden, M. N. Leger, A. G. Ryder, T. Howley, and M. O Connell. Classification of a target analyte in solid mixtures using principal component analysis, support vector machines and raman spectroscopy. 2005. URL http://aran.library.nuigalway.ie/xmlui/handle/10379/192.

J. Mairal. Optimization with first-order surrogate functions. *arXiv:1305.3120 [cs, math, stat]*, 2013. URL http://arxiv.org/abs/1305.3120.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *arXiv:0908.0050 [cs, math, stat]*, 2009. URL http://arxiv.org/abs/0908.0050.

S. Mallat. A wavelet tour of signal processing (third edition). Academic Press, Boston, second edition edition, 2009. ISBN 978-0-12-374370-1.

S. Maret, P. Hily-Blant, J. Pety, S. Bardeau, and E. Reynier. Weeds: a CLASS extension for the analysis of millimeter and sub-millimeter spectral surveys. *arXiv:1012.1747 [astro-ph]*, 2010. URL http://arxiv.org/abs/1012.1747.

H. S. P. Müller, F. Schlder, J. Stutzki, and G. Winnewisser. The cologne database for molecular spectroscopy, CDMS: a useful tool for astronomers and spectroscopists. *Journal of Molecular Structure*, 742(1):215–227, 2005. ISSN 0022-2860. doi:10.1016/j.molstruc.2005.01.027. URL http://www.sciencedirect.com/science/article/pii/S0022286005000888.

H. S. P. Müller, A. Belloche, K. M. Menten, C. Comito, and P. Schilke. Rotational spectroscopy of isotopic vinyl cyanide, h2cchcn, in the laboratory and in space. *Journal of Molecular Spectroscopy*, 251(1):319–325, 2008. ISSN 0022-2852. doi:10.1016/j.jms.2008.03.016. URL http://www.sciencedirect.com/science/article/pii/S0022285208001252.

A. Nummelin, P. Bergman, Hjalmarson, P. Friberg, W. M. Irvine, T. J. Millar, M. Ohishi, and S. Saito. A three-position spectral line survey of sagittarius b2 between 218 and 263 GHz. i. the observational data. *The Astrophysical Journal Supplement Series*, 117(2):427, 1998. ISSN 0067-0049. doi:10.1086/313126. URL http://iopscience.iop.org/0067-0049/117/2/427.

A. Nummelin, P. Bergman, Hjalmarson, P. Friberg, W. M. Irvine, T. J. Millar, M. Ohishi, and S. Saito. A three-position spectral line survey of sagittarius b2 between 218 and 263 GHz. II. data analysis. *The Astrophysical Journal Supplement Series*, 128(1):213, 2000. ISSN 0067-0049. doi:10.1086/313376. URL http://iopscience.iop.org/0067-0049/128/1/213.

D. Pequignot. Deep spectroscopy of gaseous nebulae. *Physica Scripta Volume T*, 65:137–143, 1996. ISSN 0281-1847. doi:10.1088/0031-8949/1996/T65/019. URL http://adsabs.harvard.edu/abs/1996PhST...65..137P.

James W. Perry, Allen Kent, and Madeline M. Berry. Machine literature searching x. machine language; factors underlying its design and development. *American Documentation*, 6(4): 242–254, 1955. ISSN 1936-6108. doi:10.1002/asi.5090060411. URL http://dx.doi.org/10.1002/asi.5090060411.

A. J. Remijan. Splatalogue - motivation, current status, future plans. volume 215, page 568, 2010. URL http://adsabs.harvard.edu/abs/2010AAS...21547905R.

A. J. Remijan and A. J. Markwick-Kemper. SPLATALOGUE: DATABASE FOR ASTRONOMICAL SPECTROSCOPY. 2008. ISSN http://hdl.handle.net/1811/33544. URL http://hdl.handle.net/1811/33544.

P. Schilke, D. J. Benford, T. R. Hunter, D. C. Lis, and T. G. Phillips. A line survey of orion-KL from 607 to 725 GHz. *The Astrophysical Journal Supplement Series*, 132(2):281, 2001. ISSN 0067-0049. doi:10.1086/318951. URL http://iopscience.iop.org/0067-0049/132/2/281.

P. Schilke, R. Rolffs, and C. Comito. Analysis tools for spectral surveys. In *The Molecular Universe*, volume 7 of *Proceedings of the International Astronomical Union*, pages 440–448, 2011. doi:10.1017/S174392131102518X. URL http://journals.cambridge.org/article_S174392131102518X.

F. Schuller, K. M. Menten, Y. Contreras, F. Wyrowski, P. Schilke, L. Bronfman, T. Henning, C. M. Walmsley, H. Beuther, S. Bontemps, R. Cesaroni, L. Deharveng, G. Garay, F. Herpin, B. Lefloch, H. Linz, D. Mardones, V. Minier, S. Molinari, F. Motte, L.-A. Nyman, V. Reveret, C. Risacher, D. Russeil, N. Schneider, L. Testi, T. Troost, T. Vasyunina, M. Wienen, A. Zavagno, A. Kovacs, E. Kreysa, G. Siringo, and A. Weiss. ATLASGAL - the APEX telescope large area survey of the galaxy at 870 microns. *Astronomy and Astrophysics*, 504(2):415–427, 2009. ISSN 0004-6361, 1432-0746. doi:10.1051/0004-6361/200811568. URL http://arxiv.org/abs/0903.1369.

K. R. Sembach, J. C. Howk, B. D. Savage, J. M. Shull, and W. R. Oegerle. Far ultraviolet spectroscopy of the intergalactic and interstellar absorption toward 3c 273. *The Astrophysical Journal*, 561(2):573–599, 2001. ISSN 0004-637X, 1538-4357. doi:10.1086/323408. URL http://arxiv.org/abs/astro-ph/0108047.

B. Sharpee, R. Williams, J. A. Baldwin, and P. A. M. van Hoof. Introducing EMILI: Computer aided emission line identification. *The Astrophysical Journal Supplement Series*, 149(1):157–187, 2003. ISSN 0067-0049, 1538-4365. doi:10.1086/378321. URL http://arxiv.org/abs/astro-ph/0307053.

S. Sigurdsson, P.A. Philipsen, L.K. Hansen, J. Larsen, M. Gniadecka, and H.C. Wulf. Detection of skin cancer by classification of raman spectra. *IEEE Transactions on Biomedical Engineering*, 51(10):1784–1793, 2004. ISSN 0018-9294. doi:10.1109/TBME.2004.831538.

P. Skoda, Peter W. D., N., M. Castro, D. Andresic, and T. Jenness. Spectroscopic analysis in the virtual observatory environment with SPLAT-VO. *Astronomy and Computing*, 7-8, 2014. ISSN 22131337. doi:10.1016/j.ascom.2014.06.001. URL http://arxiv.org/abs/1407.1765.

C. L. Smith, A. A. Zijlstra, and G. A. Fuller. A molecular line survey of a sample of AGB stars and planetary nebulae. *arXiv:1508.05014 [astro-ph]*, 2015. URL http://arxiv.org/abs/1508.05014.

Dent G. Smith, E. Modern raman spectroscopy: A practical approach, 2005. URL http://www.wiley.com/WileyCDA/WileyTitle/productCd-0471497940.html.

W. S. Struve. *Fundamentals of Molecular Spectroscopy*. Wiley-Interscience, 1 edition edition, 1989. ISBN 9780471854241.

B. A. Turlach, W. N. Venables, and S. J. Wright. Simultaneous variable selection. *Technometrics*, 47(3):349–363, 2005. ISSN 0040-1706. doi:10.1198/004017005000000139. URL http://dx.doi.org/10.1198/004017005000000139.

C. Vastel, S. Bottinelli, E. Caux, J.-M. Glorian, and M. Boiziot. CASSIS: a tool to visualize and analyse instrumental and synthetic spectra. 2015. URL http://adsabs.harvard.edu/abs/2015sf2a.conf..313V.

J. R. Walsh, D. Pquignot, C. Morisset, P. J. Storey, B. Sharpee, J. Baldwin, P. A. M. van Hoof, and R. E. Williams. A deep UV-blue planetary nebula template spectrum from NGC 7027. volume 209, page 337, 2003. ISBN 1743-9221. URL http://adsabs.harvard.edu/abs/2003IAUS..209..337W.

J. Wright, Yi Ma, J. Mairal, G. Sapiro, T.S. Huang, and Shuicheng Yan. Sparse representation for computer vision and pattern recognition. *Proceedings of the IEEE*, 98(6):1031–1044, 2010. ISSN 0018-9219. doi:10.1109/JPROC.2010.2044470.

Zhen J. Xiang, Hao Xu, and Peter J Ramadge. Learning sparse representations of high dimensional data on large scale dictionaries. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 900–908. Curran Associates, Inc., 2011. URL http://papers.nips.cc/paper/4400-learning-sparse-representations-of-high-dimensional-data-on-la pdf.