PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE

SCHOOL OF ENGINEERING

AUTOMATIC IDENTIFICATION OF SPECTRAL LINES THROUGH SPECTRUM RECONSTRUCTION

ANDRÉS ANTONIO RIVEROS MOYA

Thesis submitted to the Office of Research and Graduate Studies

in partial fulfillment of the requirements for the degree of

Master of Science in Engineering

Advisor:

KARIM PICHARA BAKSAI

Santiago de Chile, July 2016

PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE

SCHOOL OF ENGINEERING

AUTOMATIC IDENTIFICATION OF SPECTRAL LINES THROUGH SPECTRUM RECONSTRUCTION

ANDRÉS ANTONIO RIVEROS MOYA

Members of the Committee:

KARIM PICHARA BAKSAI

DIEGO MARDONES

MAURICIO ARAYA

PAVLOS PROTOPAPAS

CRISTIÁN TEJOS

Thesis submitted to the Office of Research and Graduate Studies

in partial fulfillment of the requirements for the degree of

Master of Science in Engineering

Santiago de Chile, July 2016

ACKNOWLEDGEMENTS

Acknowledgments

This investigation is supported by Vicerrectora de Investigación (VRI) from Pontificia Universidad Católica de Chile, the Institute of Applied Computer Science at Harvard University, and CONICYT projects ICHAA 79120008 and Basal FB 0821.

This work was funded by project FONDEF D11I1060, a collaborative project between several Chilean universities in order to create a Chilean virtual observatory, called Observatorio Virtual Chileno (ChiVO) *.

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.

*http://www.chivo.cl

TABLE OF CONTENTS

Acknowledgements v

Acknowledgments v

LIST OF FIGURES viii

LIST OF TABLES x

Abstract xi

Resumen xiii

1. Introduction 1

1.1. Introduction 1

2. Background Theory 5

2.1. Background 5

2.1.1. Spectroscopy 5

2.1.2. Sparse coding 6

3. Related Work 9

3.1. Related work 9

4. Methodology 12

4.1. Data 12

4.1.1. Synthetic data 12

4.1.2. ASYDO. Synthetic data 12

4.2. Algorithm 14
4.2.1. Pre-processing 16
4.2.2. Prediction 20
4.2.3. Training/Test set 25
5. Experimental Setup and Results 26
5.1. Experimental results 26
5.1.1. Measure of accuracy 26
5.1.2. Prediction results 27
5.1.3. Signal to noise effect in predictions 27
5.1.4. Variable width effect in predictions 28
5.1.5. Complex cases 30
5.1.6. Real Data 33
6. Conclusions 36
6.1. Conclusion 36
References 38

LIST OF FIGURES

1.1 NH_3 (1, 1) transition from three different sources (Schuller et al., 2009). . . . 3

4.1 Example of ASYDO simulated spectra. Green vertical lines correspond to theoretical isotope lines used for the simulation. 14

4.2 Example though the whole process: 1. Pre-processing: (a) Raw data cube, (b) Filtered and normalized, (c) Dictionary, (d) Detected candidate lines, (e) Dictionary recalibration. 2. (f) Signal reconstruction. 15

4.3 Rotational spectrum of vinyl cyanide transition of $H_2^{13}C$=CHCN showing hyperfine unseparable structures (Müller et al., 2008). 17

4.4 The schematic convergence of predictions. Four spectra (blue pixels along frequency dimension) independently analyzed, their prediction merged. 23

5.1 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. 28

5.2 The measures of accuracy, precision and recall obtained for fixed line width cubes, variable line width and overall results for ALMA band 9. 29

5.3 The measures of accuracy, precision and recall obtained for fixed line width cubes, variable line width and overall results for ALMA band 7. 29

5.4 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. . . . 30

5.5 Blending case 31

5.6 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. . 32

5.7 Double peaks for single Line 33

5.8 Example of used data cubes, slice of $IRAS16547-4247\_Jet\_CH3OH7-6.clean$ from ALMA project #2011.0.00419.S. 34

5.9 Data cube example following the steps previously explained. 35

LIST OF TABLES

5.1 The f-score for different noise level (bands) and width of the lines. 27

5.2 Predicted frequencies associated to the molecules of interest for each clean cube for ALMA project #2011.0.00419.S. 35

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 90% accuracy when no

blending nor hyperfine cases are present. As wavelength separation decreases (equal or less than 1 MHz), accuracy goes down to 82%.

Algorithm source code, synthetic data and list of suggested identifications are publicly available [†].

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

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

RESUMEN

La astronomía enfrenta nuevos desafíos en cuanto a como analizar $big\ data$, y por lo tanto, como buscar o predecir eventos/patrones de interés. Nuevas observaciones en regiones de longitudes de onda previamente inexploradas estarán disponibles gracias a instrumentos como el Atacama Large Millimeter Array (ALMA). Dada esta creciente cantidad de datos de alta resolución espectral, cualquier análisis no automatizado constituiría un esfuerzo maś allá de la capacidad humana. Actualmente, la clasificación de lines de emisión significa decidir si un línea de emisión pertenece a un isótopo específico. Esta clasificacion es principalmente hecha comparando las líneas observadas con líneas de emisión de isótopos conocidas. Un algoritmo de clasificación automático reduciría dramáticamente los esfuerzos humanos para analizar datos espectrales, permitiendo a los astrónomos enfocar sus esfuerzos en análisis más profundos.

En este trabajo, proponemos un algoritmo que utiliza un modelo $sparse$ para representar el espectro y automáticamente clasificar líneas de emisión. Para esto utilizamos una base de datos de líneas espectrales para determinar un set de vectores base que represente la presencia de líneas de emisión teóricas. Luego, para clasificar líneas en un espectro dado, se minimiza la diferencia entre el espectro y una combinación lineal de los vectores base determinados.

El output del modelo corresponde a un vector de probabilidad que representa la distribución de la predicción sobre un set de posibles isótopos. Realizamos pruebas de nuestro algoritmo con datos experimentales de Splatalogue y datos simulados del proyecto ASYDO. El resultado del análisis muestra que el algoritmo es capaz de identificar líneas de emisión con una precisión del 90% cuando ni blending ni casos hiperfinos están presentes. En tanto que la separación de longitud de onda entre líneas decrece (menor o igual que 1 MHz) la precisión baja a un 82%.

El código fuente del algoritmo, los datos sintéticos y la lista de identificaciones sugerida están públicamente disponibles [‡].

Keywords: líneas espectrales: líneas de emisión; técnicas: espectroscopía; método: análisis de datos

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

1. INTRODUCTION

1.1. Introduction

Modern astronomical observatories have brought increasing amounts of data over the last few years. Radio-telescope sensitivity has 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) (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 sub-millimeter 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 way (Schilke et al., 2011; Skoda et al., 2014).

The traditional spectra analysis to identify emission lines involves searching for 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 spectrum (Schilke et al., 2001; Comito et al., 2005; Maret et al., 2010; Caux et al., 2015; Vastel et al., 2015), models that fit synthetic spectrum using observed ones (Pequignot, 1996; Walsh et al., 2003), and heuristic analysis of simultaneous presence of lines (Sharpee et al., 2003).

The techniques described above are in general not scalable, do not rely on automatic processes or are based on complex theoretical underlying models.

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 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) compares intensity differences along the spectra and evalues them by looking for intensity differences greater than a given σ threshold (Sharpee et al., 2003). 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 modeling.

Spectral data can drastically vary between different sample measurements for the same object. 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 1.1.

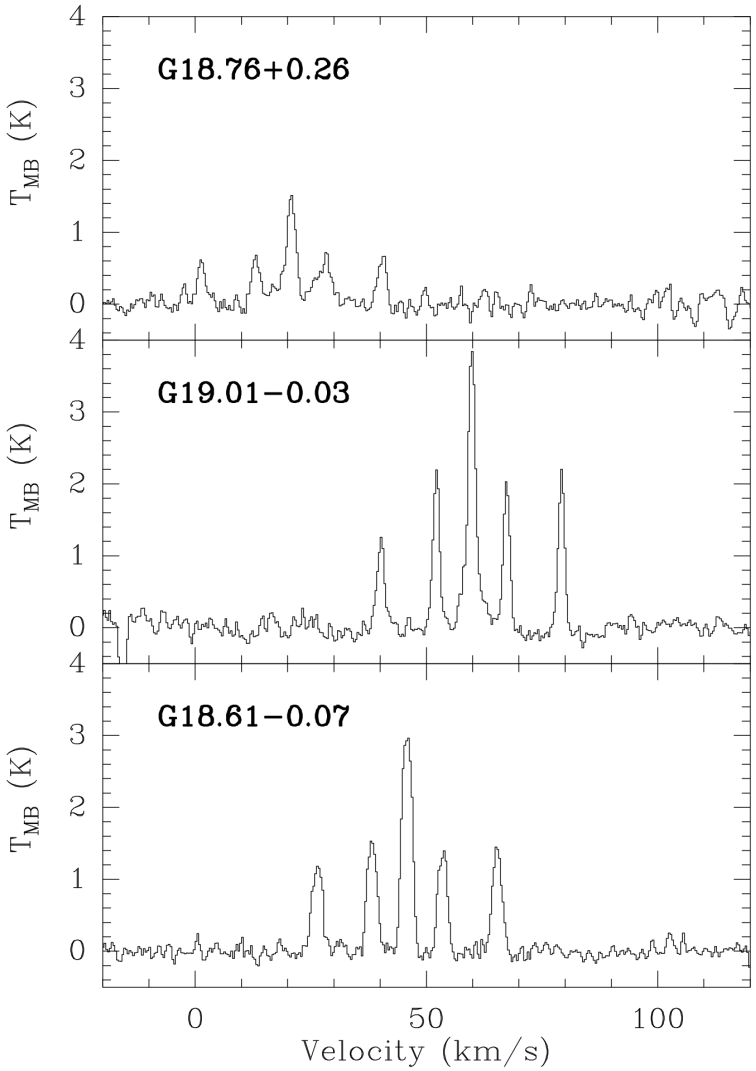

FIGURE 1.1. NH_3 (1, 1) transition from three different sources (Schuller et al., 2009).

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: Section 2 introduces the theoretical background. Section 3 presents an overview of previous works. Section 4.1 describes both problem scope and data origin. Section 4.2 covers the proposed algorithm. Section 5 shows the results and its discussed. Section 6 exposes the conclusion of this work.

2. BACKGROUND THEORY

2.1. Background

2.1.1. Spectroscopy

Spectroscopy is a technique that enables the analysis of interaction between matter and light (D. G. Smith E., 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 spectrum, known as emission lines (Struve, 1989). Each line has a different frequency depending on the molecule and energy level associated to it (D. G. Smith E., 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).

The traditional process of mapping observed frequencies to experimental ones has the difficulty associated to differences in both intensities and frequency. For instance, two closer lines in frequency space are hard to dissociate, mainly, because they appear as double peaks or they are blended into one single line (Cernicharo et al., 2013; C. L. Smith et al., 2015).

Internal factors such as the temperature of objects, velocity gradients, 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 different 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; C. L. Smith et al., 2015), as can be seen in figure 1.1.

2.1.2. Sparse coding

Sparse coding consist in modeling signals through linear combinations of basis vectors under sparsity constraints, in order to have a sparse representation of input signals (Mairal et al., 2009b; 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., 2009b).

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 2.1 corresponds to the optimization problem that has to be solved in order to determine the sparse coding coefficients.

$$\begin{aligned} &\text{Minimize}_{\alpha} ||s - \sum_{i=1}^{d} \alpha_i\ w_i||_2^2 \\ &\text{Subject to: } ||\alpha||_1 \leq \lambda \\ &\alpha \geq 0 \end{aligned} \tag{2.1}$$

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., 2009b). Instead, we use the iterative algorithm presented in Turlach et al. (2005); Mairal (2013). The implementation used in this work is the SPAMS package * (Mairal et al., 2009b,a).

*http://spams-devel.gforge.inria.fr/

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 4.2.

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., 2009b). For wavelength domains, predefined dictionaries give satisfactory results (Mallat, 2009).

3. RELATED WORK

3.1. Related work

In the last years, automatic detection of spectral lines has been developing from different paradigms. Spectral classification is found in 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 a spectrum. 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).

Fitting functions to shape the optical depth of lines is a very common technique that is still widely used. Gaussian functions (Fuller & Myers, 1993; Nummelin et al., 2000) or top-hat functions (C. L. 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.

XCLASS, CASSIS and WEEDS also fit functions to spectra. They build a line list fitting of all the transitions of an isotope through two steps: i) lines fit, ii) baseline fit. For lines fitting, CASSIS determines optimal parameter functions to simultaneously fit all peaks along the spectrum. Gauss, Lorentz, Sinc and Voigt are examples of functions for fitting lines. Then, in step ii), a sinusoidal or polynomial baseline function is used to fit the noise (Caux et al., 2015; Vastel et al., 2015).

Chemical and physical models takes into account the source structure 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. In step ii), the structure temperature of cores are estimated using Monte Carlo. Both **radmc-3d** and LIME (Brinch & Hogerheijde, 2010) use this sampling simulation assuming LTE approximation. An analysis of line shapes and temperature modeling allow to assign them to known isotope lines. These models have a rigorous treatment of blending and have the flexibility to deal with wavelength uncertainty from databases (Sharpee et al., 2003). Approaches that compares observed spectra with synthetic modelings have been proposed in (Pequignot, 1996; Walsh et al., 2003).

These tools make use of catalogs of experimental lines constructed mainly by the data from Jet Propulsion Laboratory (JPL), The Cologne Database for Molecular Spectroscopy (CDMS) (Müller et al., 2005), Toyama and Lovas National Institute of Standards

and Technology (NIST). These catalogs are compiled into Splatalogue, the most up-to-date and complete spectral line database (Remijan & Markwick-Kemper, 2008; Remijan, 2010).

4. METHODOLOGY

4.1. Data

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

4.1.2. ASYDO. Synthetic data

The Astronomical Synthetic Data Observations (ASYDO) package * is used to simulate ALMA-like data. The simulation generates a set to develop and test identification accuracy (Araya et al., 2015).

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

- **Isolist** : subset isotope list to generate a cube
- **Parameters**
  - **freq** : spectral center (MHz)
  - **spe_res** : spectral resolution (MHz)
  - **spe_bw** : spectral bandwidth (MHz)

*https://github.com/ChileanVirtualObservatory/ASYDO

– **(fwhm**, $\boldsymbol{\alpha}$ **-skew**): skew-normal distribution parameters (MHz, parameter)

Also, for each Isolist, a spatial form and a temperature need to be given, allowing to produce complex structures by combining simple ones with different parameters.

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 velocity gradients.

The intensity of each line is obtained by a simplified version of the *detection equation* of Stahler and Palla (Stahler & Palla, 2004), that considers the temperature of the transition from Splatalogue, the temperature of the object, relative intensities of the isotopes, and an adjustable random intensity associated to each molecule.

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: **spe_res** of 1 MHz, **spe_bw** as 4000 MHz and **(fwhm**, $\boldsymbol{\alpha}$ **-skew**) as $\mathbf{(8, 0)}$.

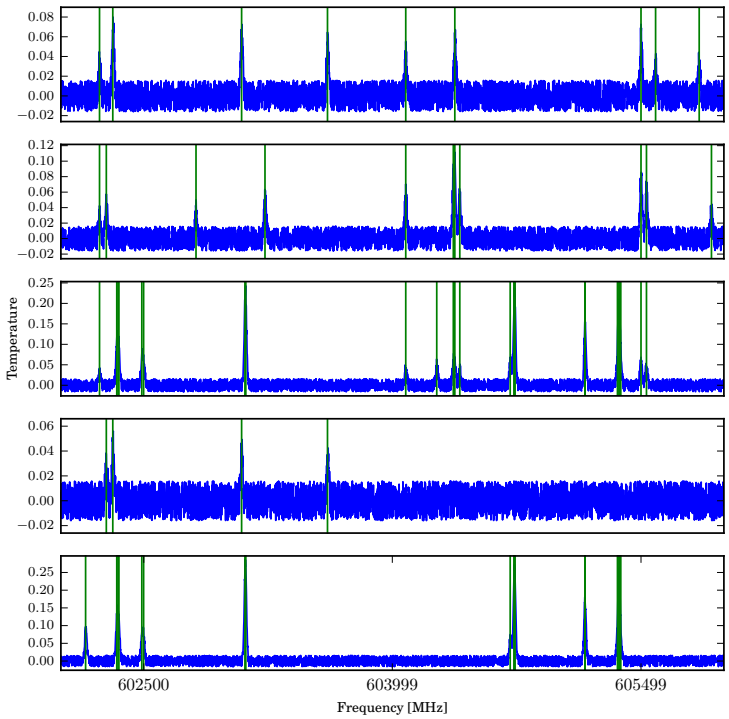

FIGURE 4.1. Example of ASYDO simulated spectra. Green vertical lines correspond to theoretical isotope lines used for the simulation.

The objective of ASYDO is not to produce realistic emission lines, but a large number of spectra that mimics real data and its diversity. The simplistic astrophysical model behind ASYDO will surely bias the classification towards incorrect results for some cases, so we only use this package for testing and as a proof of concept meanwhile enough training data becomes available.

4.2. Algorithm

Our algorithm has two main steps: i) spectrum pre-processing, which also involves the creation and recalibration of the dictionary, covered in section 4.2.1. ii) optimization

of equation 2.1, which allow us later to predict emission lines present along the spectrum, viewed at detail in section 4.2.2. An overview of the process is illustrated in figure 4.2.

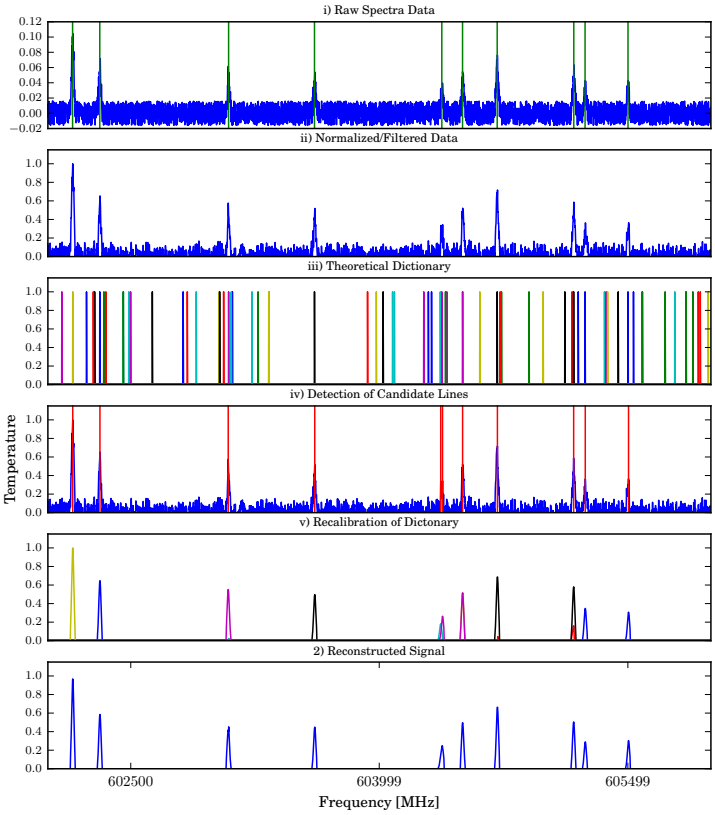

FIGURE 4.2. Example though the whole process: 1. Pre-processing: (a) Raw data cube, (b) Filtered and normalized, (c) Dictionary, (d) Detected candidate lines, (e) Dictionary recalibration. 2. (f) Signal reconstruction.

4.2.1. Pre-processing

At first, a dimensional pixel from a data cube is selected to analyze its wavelength range, as shown in figure 4.2 (i). Then, a normalization and filtering of spectrum 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 4.2 (ii) illustrates the output after the preprocessing step.

4.2.1.1. 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 Dirac 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:

$$\boldsymbol{w_{in}} = \begin{cases} \mathbf{1}, & \text{if } \boldsymbol{f_i} \text{ is the theoretical frequency of isotope } \boldsymbol{i} \\ \mathbf{0}, & \text{otherwise} \end{cases} \tag{4.1}$$

Figure 4.2 (iii) 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 4.3).

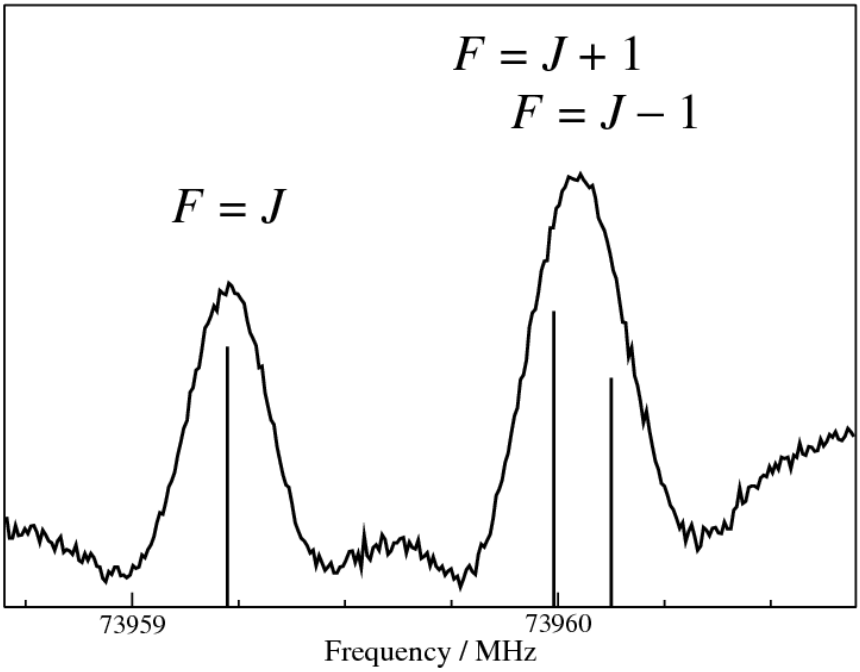

FIGURE 4.3. Rotational spectrum of vinyl cyanide transition of $H_2^{13}C$=CHCN showing hyperfine unseparable structures (Müller et al., 2008).

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.

4.2.1.2. 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-σ 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, 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 4.2 (iv).

4.2.1.3. 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 4.3.

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 $\boldsymbol{c}(\boldsymbol{f})$ is defined as $\boldsymbol{c}(\boldsymbol{f_i}) =$

$s_i, \forall i \in [1, \ldots, n]$, i. e., signal's intensity at frequency f_i. g is defined in 4.2 as the closer candidate line's frequency to a given frequency f_i, such as for a word w_{ki}:

$$g = \text{argmin}_f ||f_i - f|| \tag{4.2}$$

and a word's expansion is given by equation 4.3

$$w_{ki} = c(g) \cdot \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}(\frac{||f_i - g||)^2}{\sigma}} \tag{4.3}$$

For simplicity sake, this case uses σ value as 1 that works well in the ALMA domain we analyzed.

The final representation of the dictionary can be seen in figure 4.2 (v). The majority of the words are expanded to low values and only words closer to candidate lines have appreciable values.

4.2.2. Prediction

The optimization of equation 2.1 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 spectrum, 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.

4.2.2.1. 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 $(0, 1)$. 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 $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_d]$ $\forall i \in [1, \ldots, d]$ be the set of coefficient values for each theoretical isotope state.

$$\alpha_k^* = \begin{cases} \alpha_k, & \text{if } \alpha_k \leq 1 \\ 1/\alpha_k, & \text{if } \alpha_k > 1 \end{cases} \tag{4.4}$$

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 4.5

$$\boldsymbol{P_{ik}} = \frac{\boldsymbol{\alpha_k^*}}{\sum_{\boldsymbol{j=1}}^{\boldsymbol{d}} \boldsymbol{\alpha_j^*}} \tag{4.5}$$

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 in a given dimensional range. Let $\boldsymbol{x} \in [\boldsymbol{n}, \ldots, \boldsymbol{n+m}]$, $\boldsymbol{y} \in [\boldsymbol{v}, \ldots, \boldsymbol{v+w}]$,

$$\boldsymbol{P_{ik}} = \prod_{(\boldsymbol{n},\boldsymbol{v})}^{(\boldsymbol{n+m},\boldsymbol{v+w})} \boldsymbol{P_{ik}(x,y)}) \tag{4.6}$$

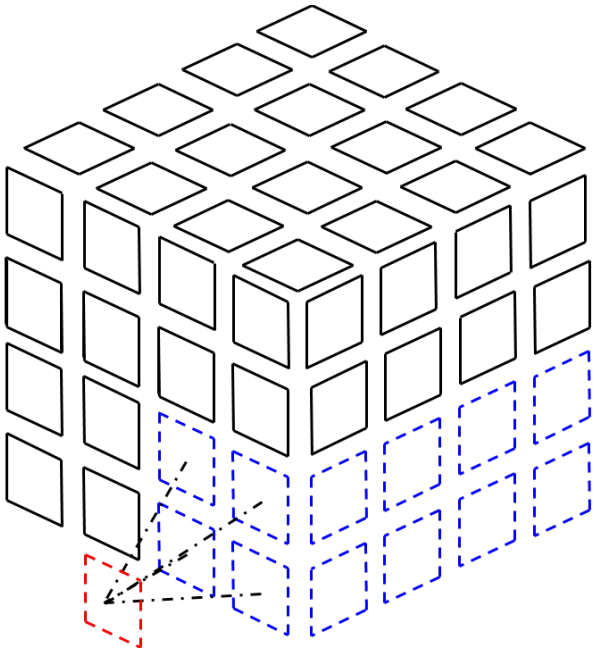

FIGURE 4.4. The schematic convergence of predictions. Four spectra (blue pixels along frequency dimension) independently analyzed, their prediction merged.

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 spectrum from a pixel of $data\_cube$;

compute $threshold$ from noise;

get $Dictionary$ from $isotopes\_set$;

initialize $candidate\_set$ as [( )];

detect max_set from the spectrum;

$max_{freq} = max(max\_set)$;

while $(max_{freq} > threshold)$ **do**
    ($candidate\_set$).append(max_{freq});

    subtract Gaussian function at max_{freq};

    detect max_set from residual spectrum;

    $max_{freq} = max(max\_set)$;

end

recalibrate $Dictionary$ from $candidate\_set$;

get $alphas$ from solving sparse coding;

compute $probability\_predictions$ from $alphas$;

return $probability\_predictions$;

Algorithm 1: Proposed algorithm

4.2.3. Training/Test set

For testing purposes, each line present in the synthetic spectra is stored as cube metadata, 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 $(\mathbf{-2}, \mathbf{+2})\boldsymbol{MHz}$ from original width.

5. EXPERIMENTAL RESULTS

5.1. 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.

5.1.1. Measure of accuracy

To evaluate identification performance, a measure of test accuracy must be designed 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 5.1 show an overview of obtained results.

Moreover, confusion matrices in figure 5.1 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 5.2 and 5.3 for ALMA bands 9 and 7 respectively.

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

TABLE 5.1. The f-score for different noise level (bands) and width of the lines.

5.1.2. 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.

5.1.3. 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 5.4 and 5.6.

Figures 5.2 and 5.3 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.

5.1.4. 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 5.4 and 5.6 shows the differences between accuracy measures for both cases, being the fixed width focused in a smaller range than variable width results.

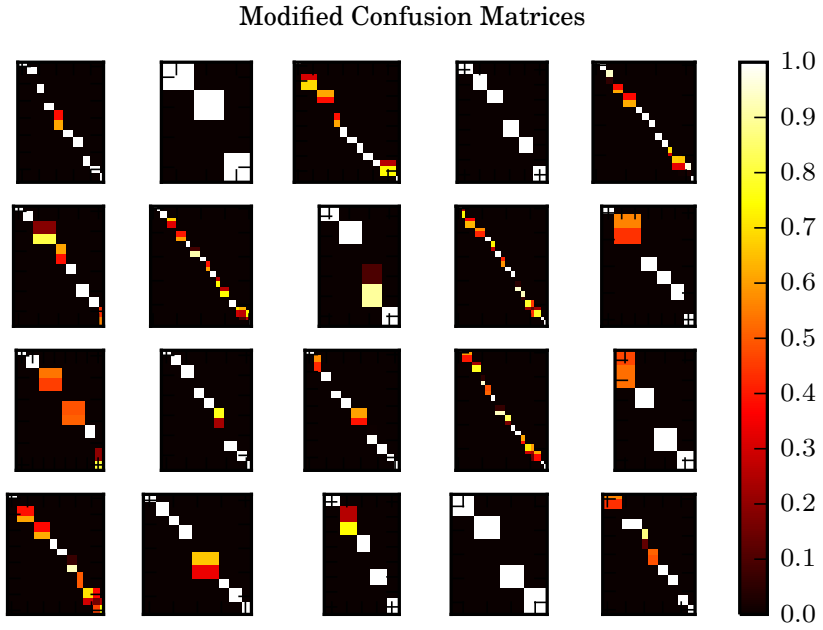

FIGURE 5.1. Modified confusion matrices for 20 experiments for data cubes with fixed line width at band 9. Predictions tend to be on the diagonal because of the algorithm preference for closer theoretical line's frequencies.

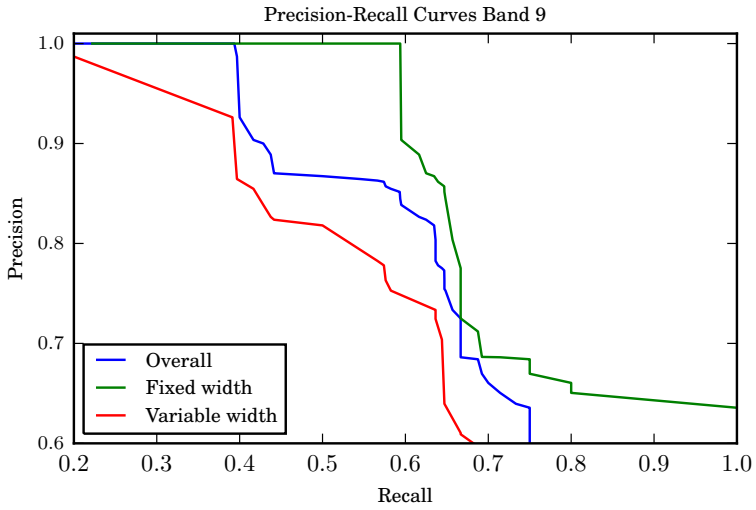

FIGURE 5.2. The measures of accuracy, precision and recall obtained for fixed line width cubes, variable line width and overall results for ALMA band 9.

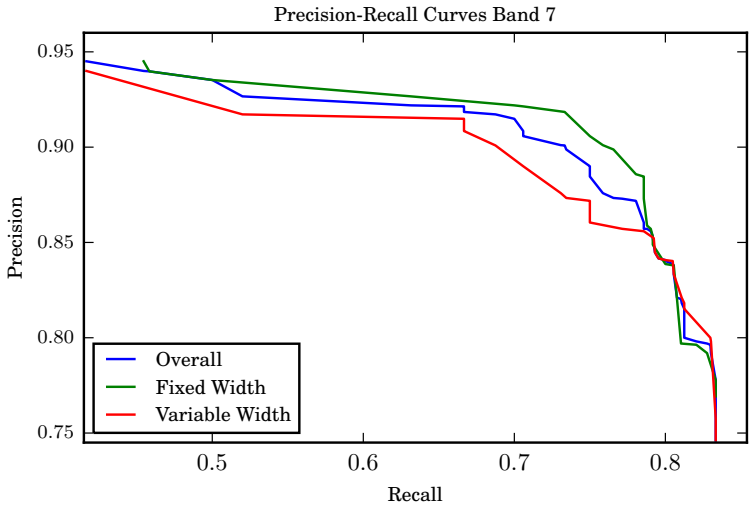

FIGURE 5.3. The measures of accuracy, precision and recall obtained for fixed line width cubes, variable line width and overall results for ALMA band 7.

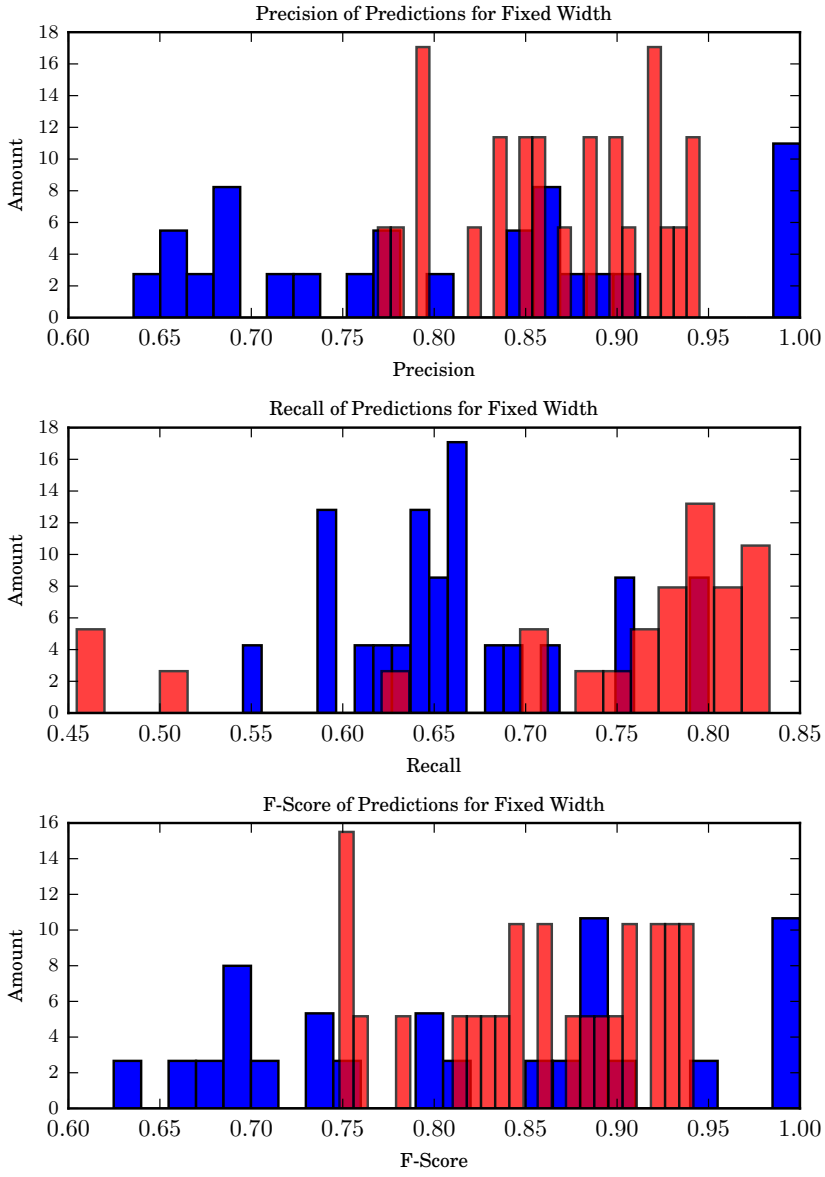

FIGURE 5.4. Histograms of results obtained for the test performed for precision, recall and f-score for fixed line width (red) vs variable line width (blue) in Band 9.

5.1.5. 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 5.5.

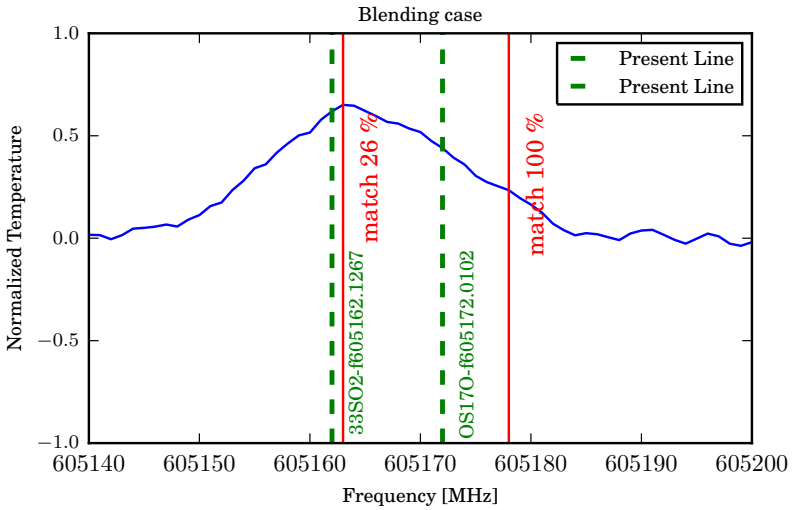

FIGURE 5.5. Blending case

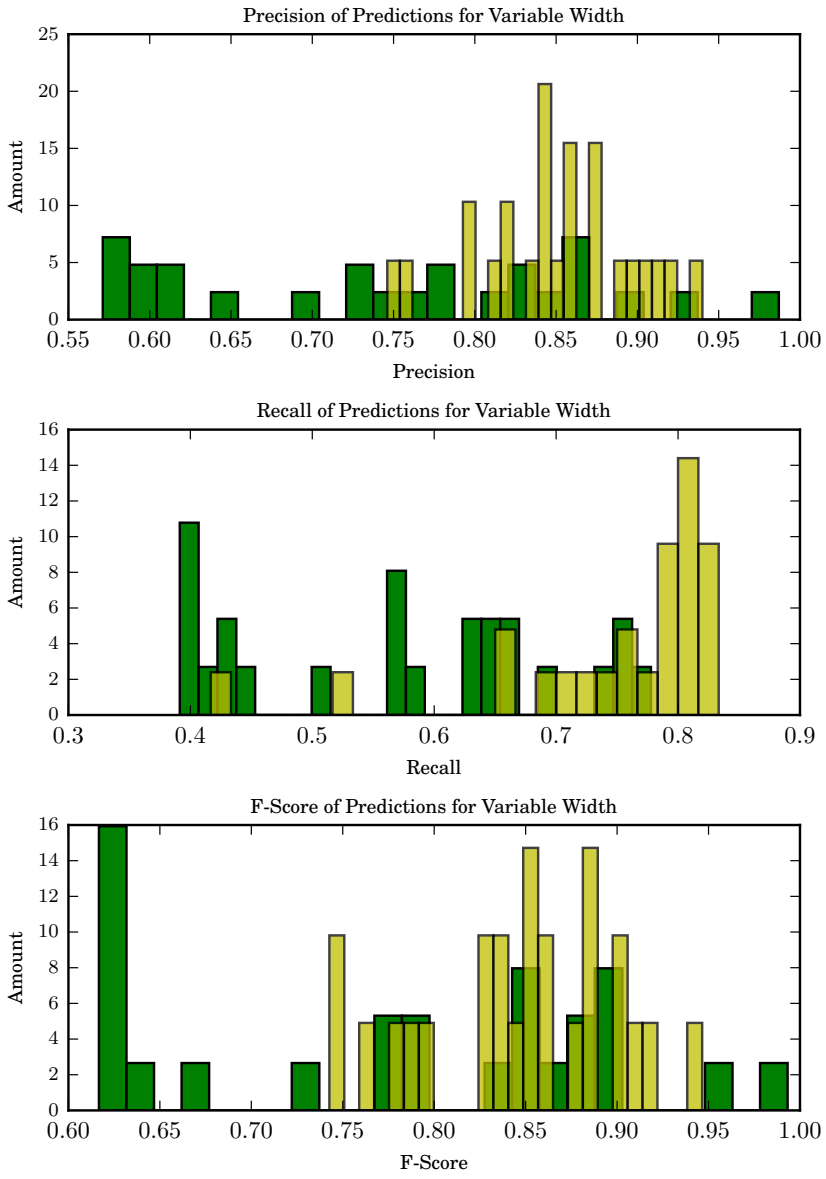

FIGURE 5.6. Histograms of results obtained for the test performed for precision, recall and f-score for fixed line width (yellow) vs variable line width (green) in Band 7.

False double peaks product of artifacts are handled by the algorithm and it determines the correct lines among false peaks, as showed in figure 5.7.

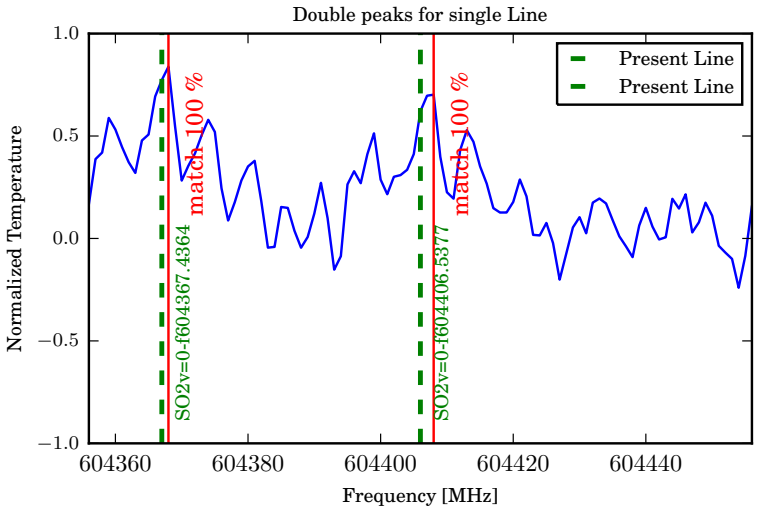

FIGURE 5.7. Double peaks for single Line

5.1.6. 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. (2015).

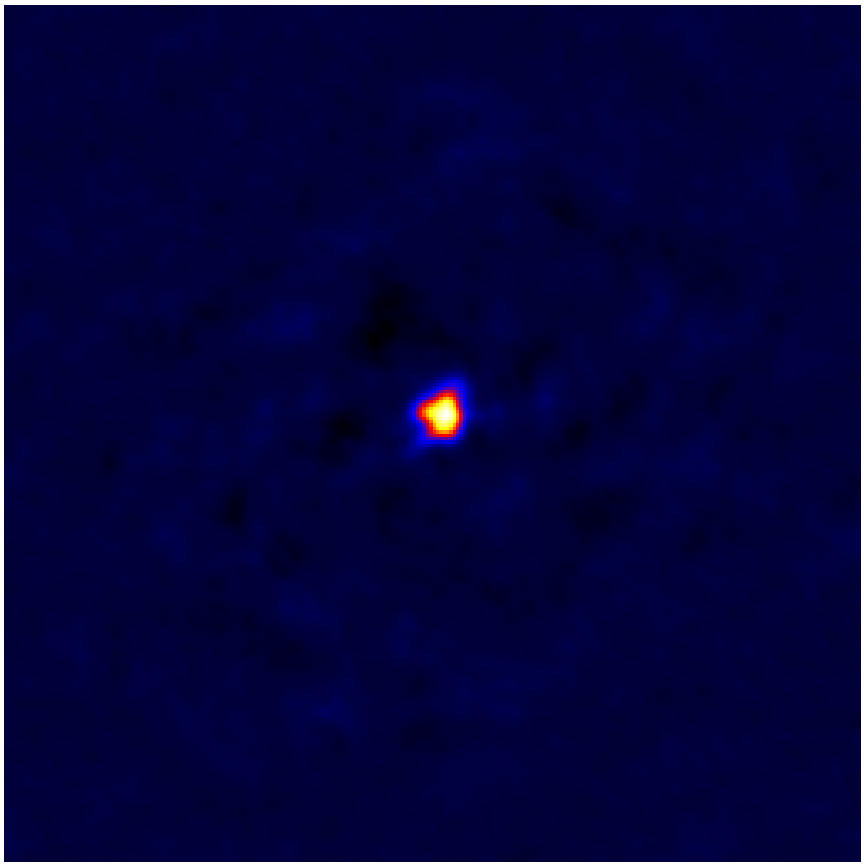

FIGURE 5.8. Example of used data cubes, slice of $\boldsymbol{IRAS16547-4247\_Jet\_CH3OH7-6.clean}$ from ALMA project #2011.0.00419.S.

In this case, the high density of theoretical frequencies plays a major role in the difficulty of the prediction. With synthetically generated 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 5.2 we present only predicted frequencies associated to the isotopes of interest. The algorithm is able to predict lines for $\boldsymbol{CH_3OH_7(7-6)}$, but the algorithm is not able to predict other isotopes of interest.

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

The process for file $CH_3OH_7(7-6)$ can be seen in figure above. 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.

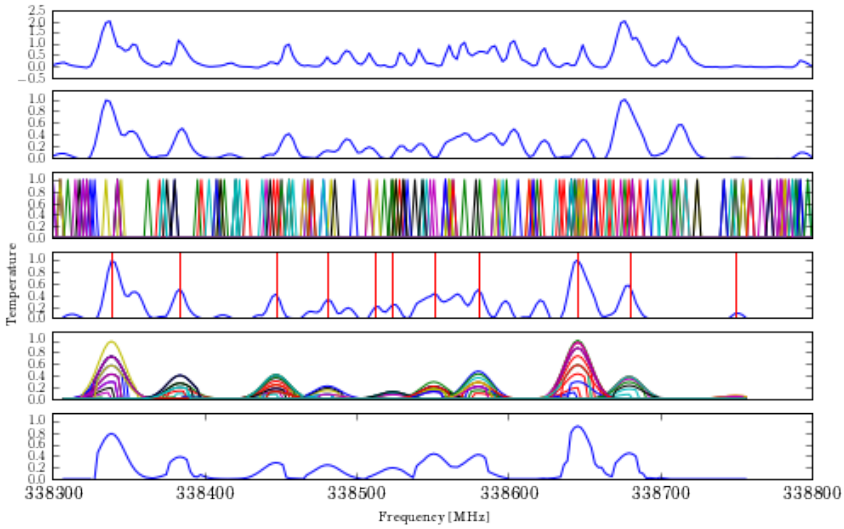

FIGURE 5.9. Data cube example following the steps previously explained.

6. CONCLUSIONS

6.1. 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 spectrum. 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 spectrum, and ii) the used combination of basis vectors gives useful information to identify the presence of emission lines along the spectrum.

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.

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.

References

Araya, M., Solar, M., Mardones, D., & Hochfrber, T. (2015). Exorcising the ghost in the machine: Synthetic spectral data cubes for assessing big data algorithms. In (Vol. 495, p. 57).

Becklin, E. E. (2006). Stratospheric observatory for infrared astronomy (SOFIA). In *Astrochemistry: Recent successes and current challenges* (Vol. 1, pp. 323–324).

Brinch, C., & Hogerheijde, M. R. (2010). LIME - a flexible, non-LTE line excitation and radiation transfer method for millimeter and far-infrared wavelengths. *Astronomy & Astrophysics*, *523*.

Bristow, H., Eriksson, A., & Lucey, S. (2013). Fast convolutional sparse coding. In *2013 IEEE conference on computer vision and pattern recognition (CVPR).*

Caux, E., Bottinelli, S., Vastel, C., & Glorian, J. M. (2015). CASSIS, a software package to analyse high spectral resolution observations. In (Vol. 280, p. 120P).

Cernicharo, J., Tercero, B., Fuente, A., Domenech, J. L., Cueto, M., Carrasco, E., et al. (2013). Detection of the ammonium ion in space. *Astrophysical Journal Letters*.

Chen, S., Donoho, D., & Saunders, M. (2001). Atomic decomposition by basis pursuit. *SIAM Review*, *43*(1), 129–159.

Comito, C., Schilke, P., Phillips, T. G., Lis, D. C., Motte, F., & Mehringer, D. (2005). A molecular line survey of orion KL in the 350 micron band. *The Astrophysical Journal Supplement Series*, *156*, 127–167.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, *32*(2), 407–499.

Fuller, G. A., & Myers, P. C. (1993). Thermal material in dense cores: A new narrow-line probe and technique of temperature determination. *The Astrophysical Journal*, *418*, 273.

Graauw, T. de, Caux, E., Gusten, R., Jellema, W., Luinge, W., Pearson, J., et al. (2004). 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* (pp. 579–580).

Gusten, R., Booth, R. S., Cesarsky, C., Menten, K. M., Agurto, C., Anciaux, M., et al. (2006). APEX: the atacama pathfinder EXperiment. In (Vol. 6267, pp. 626714–626714–26).

Higuchi, A. E., Saigo, K., Chibueze, J. O., Sanhueza, P., Takakuwa, S., & Garay, G. (2015). IRAS 165474247: A new candidate of a protocluster unveiled with ALMA. *798*(2), L33.

Ho, P. T. P., Moran, J. M., & Lo, K. Y.(2004). The submillimeter array. *The Astrophysical Journal*, *616*(1).

Howley, T., Madden, M. G., O Connell, M., & Ryder, A. G.(2005). The effect of principal component analysis on machine learning accuracy with high dimensional spectral data.

Madden, M. G., Leger, M. N., Ryder, A. G., Howley, T., & O Connell, M.(2005). Classification of a target analyte in solid mixtures using principal component analysis, support vector machines and raman spectroscopy.

Mairal, J. (2013). Optimization with first-order surrogate functions. *arXiv:1305.3120 [cs, math, stat]*.

Mairal, J., Bach, F., Ponce, J., & Sapiro, G. (2009a). Online dictionary learning for sparse coding. In *Proceedings of the 26th annual international conference on machine learning* (pp. 689–696). ACM.

Mairal, J., Bach, F., Ponce, J., & Sapiro, G. (2009b). Online learning for matrix factorization and sparse coding. *arXiv:0908.0050 [cs, math, stat]*.

Mallat, S. (2009). A wavelet tour of signal processing (third edition). In (Second Edition ed.). Boston: Academic Press.

Maret, S., Hily-Blant, P., Pety, J., Bardeau, S., & Reynier, E. (2010). Weeds: a CLASS extension for the analysis of millimeter and sub-millimeter spectral surveys. *arXiv:1012.1747 [astro-ph]*.

Müller, H. S. P., Belloche, A., Menten, K. M., Comito, C., & Schilke, P. (2008). Rotational spectroscopy of isotopic vinyl cyanide, h2cchcn, in the laboratory and in space. *Journal of Molecular Spectroscopy*, *251*(1), 319–325.

Müller, H. S. P., Schlder, F., Stutzki, J., & Winnewisser, G. (2005). The cologne database for molecular spectroscopy, CDMS: a useful tool for astronomers and spectroscopists. *Journal of Molecular Structure*, *742*(1), 215–227.

Nummelin, A., Bergman, P., Hjalmarson, ., Friberg, P., Irvine, W. M., Millar, T. J., et al. (2000). 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.

Pequignot, D. (1996). Deep spectroscopy of gaseous nebulae. *Physica Scripta Volume T*, *65*, 137–143.

Perry, J. W., Kent, A., & Berry, M. M. (1955). Machine literature searching x. machine language; factors underlying its design and development. *American Documentation*, *6*(4), 242–254.

Remijan, A. J. (2010). Splatalogue - motivation, current status, future plans. In (Vol. 215, p. 568).

Remijan, A. J., & Markwick-Kemper, A. J. (2008). SPLATALOGUE: DATABASE FOR ASTRONOMICAL SPECTROSCOPY.

Schilke, P., Benford, D. J., Hunter, T. R., Lis, D. C., & Phillips, T. G. (2001). A line survey of orion-KL from 607 to 725 GHz. *The Astrophysical Journal Supplement Series*, *132*(2), 281.

Schilke, P., Rolffs, R., & Comito, C. (2011). Analysis tools for spectral surveys. In *The molecular universe* (Vol. 7, pp. 440–448).

Schuller, F., Menten, K. M., Contreras, Y., Wyrowski, F., Schilke, P., Bronfman, L., et al. (2009). ATLASGAL - the APEX telescope large area survey of the galaxy at 870 microns. *Astronomy and Astrophysics*, *504*(2), 415–427.

Sembach, K. R., Howk, J. C., Savage, B. D., Shull, J. M., & Oegerle, W. R. (2001). Far ultraviolet spectroscopy of the intergalactic and interstellar absorption toward 3c 273. *The Astrophysical Journal*, *561*(2), 573–599.

Sharpee, B., Williams, R., Baldwin, J. A., & Hoof, P. A. M. van. (2003). Introducing EMILI: Computer aided emission line identification. *The Astrophysical Journal Supplement Series*, *149*(1), 157–187.

Sigurdsson, S., Philipsen, P., Hansen, L., Larsen, J., Gniadecka, M., & Wulf, H. (2004). Detection of skin cancer by classification of raman spectra. *IEEE Transactions on Biomedical Engineering*, *51*(10), 1784–1793.

Skoda, P., D., P. W., N., Castro, M., Andresic, D., & Jenness, T. (2014). Spectroscopic analysis in the virtual observatory environment with SPLAT-VO. *Astronomy and Computing*, *7-8*.

Smith, C. L., Zijlstra, A. A., & Fuller, G. A. (2015). A molecular line survey of a sample of AGB stars and planetary nebulae. *arXiv:1508.05014 [astro-ph]*.

Smith, D. G., E. (2005). *Modern raman spectroscopy: A practical approach.*

Stahler, S. W., & Palla, F. (2004). *The formation of stars.* Wiley-VCH Verlag GmbH.

Struve, W. S. (1989). *Fundamentals of molecular spectroscopy* (1 edition ed.). Wiley-Interscience.

Turlach, B. A., Venables, W. N., & Wright, S. J. (2005). Simultaneous variable selection. *Technometrics*, *47*(3), 349–363.

Vastel, C., Bottinelli, S., Caux, E., Glorian, J.-M., & Boiziot, M. (2015). CASSIS: a tool to visualize and analyse instrumental and synthetic spectra.

Walsh, J. R., Pquignot, D., Morisset, C., Storey, P. J., Sharpee, B., Baldwin, J., et al. (2003). A deep UV-blue planetary nebula template spectrum from NGC 7027. In (Vol. 209, p. 337).

Wright, J., Ma, Y., Mairal, J., Sapiro, G., Huang, T., & Yan, S. (2010). Sparse representation for computer vision and pattern recognition. *Proceedings of the IEEE*, *98*(6), 1031–1044.

Xiang, Z. J., Xu, H., & Ramadge, P. J. (2011). Learning sparse representations of high dimensional data on large scale dictionaries. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 24* (pp. 900–908). Curran Associates, Inc.