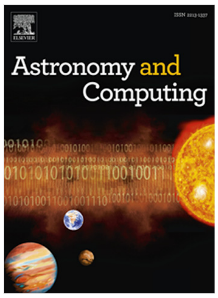
Accepted Manuscript



Unsupervised learning of structure in spectroscopic cubes

M. Araya, M. Mendoza, M. Solar, D. Mardones, A. Bayo

PII: S2213-1337(17)30023-9
DOI: https://doi.org/10.1016/j.ascom.2018.06.001
Reference: ASCOM 230

To appear in: *Astronomy and Computing*

Received date : 22 February 2017
Accepted date : 4 June 2018

Please cite this article as: Araya M., Mendoza M., Solar M., Mardones D., Bayo A., Unsupervised learning of structure in spectroscopic cubes. *Astronomy and Computing* (2018), https://doi.org/10.1016/j.ascom.2018.06.001

Unsupervised Learning of Structure in Spectroscopic Cubes

Mauricio Araya[a,*], Marcelo Mendoza[a], Mauricio Solar[a], Diego Mardones[b], Amelia Bayo[c]

[a]*Departamento de Informática, Universidad Técnica Federico Santa María, Casilla 110-V, Valparaíso, Chile*
[b]*Departamento de Astronomía, Universidad de Chile, Casilla 36-D, Santiago, Chile*
[c]*Instituto de Física y Astronomía, Universidad de Valparaíso, Casilla 5030, Valparaíso , Chile*

Abstract

We consider the problem of analyzing the structure of spectroscopic cubes using unsupervised machine learning techniques. We propose representing the target's signal as an homogeneous set of volumes through an iterative algorithm that separates the structured emission from the background while not overestimating the flux. Besides verifying some basic theoretical properties, the algorithm is designed to be tuned by domain experts, because its parameters have meaningful values in the astronomical context. Nevertheless, we propose an heuristic to automatically estimate the signal-to-noise ratio parameter of the algorithm directly from data. The resulting light-weighted set of samples ($\leq 1\%$ compared to the original data) offer several advantages. For instance, it is statistically correct and computationally inexpensive to apply well-established techniques of the pattern recognition and machine learning domains; such as clustering and dimensionality reduction algorithms. We use ALMA science verification data to validate our method, and present examples of the operations that can be performed by using the proposed representation. Even though this approach is focused on providing faster and better analysis tools for the end-user astronomer, it also opens the possibility of content-aware data discovery by applying our algorithm to big data.

Keywords: Astronomical Imaging, Image Analysis, Homogeneous Representations, Machine Learning

1. Introduction

Even though there is a large body of work regarding 2D image analysis, most of the techniques do not directly scale up to more dimensions. There are specific-purpose algorithms in astronomy to deal with 3D data, such as clump finding algorithms for spectroscopic data cubes (Williams et al., 1994; Stutzki and Guesten, 1990; Berry, 2015), yet the current state of the practice requires a huge effort in terms of storage space, computational time and actually human-machine interaction to generate useful products for astronomers (McMullin et al., 2007). Moreover, data growth in sensitivity and resolution with each new instrument, so the next-generation of projects in Astronomy will produce several terabytes of data every night (Ivezic et al., 2009; Dewdney et al., 2008), making impossible to perform analysis without automatically reducing its dimensionality.

Machine learning, and other advanced statistical methods, have been a source of success for astronomers (Vanderplas et al., 2012; Richards et al., 2011; Gibson et al., 2012): learning and inference are powerful tools to represent data in a compact way that allows us to make automatic decisions. Machine learning methods for classification, model-based regression, clustering and feature selection (Bishop, 2007), often rely on samples being independent and identically distributed (i.i.d.), which is not the case for the pixels of an image. Therefore, applying state-of-the-art machine learning techniques usually involves adapting the method or preprocessing the data to comply with this assumption.

We propose representing spectroscopic data cubes as a compact and homogeneous set of volumes, that can be treated directly as samples of the underlying signal of interest to achieve two goals:

- reduce the size of the cube representation to limit computational and memory resources needed to perform astronomical analysis, and
- comply with the i.i.d. assumption allowing astronomers to use machine learning and statistical analysis tools that are based on this assumption.

In this article we focus on spectroscopic data cubes, and specifically on interferometric synthesized spectral cubes on the millimeter/sub-millimeter range, but this work could be straightforwardly applied to 2D images. Moreover, new acquisition techniques are now producing even higher dimensional data with axes such as polarization, spatial depth or time series. For instance, the data that is currently generated by the ALMA observatory (Testi et al., 2010) are actually 4D data hypercubes. Again, our algorithm can work with these higher dimensions without major changes.

The outline of the article is the following. In Section 2 we describe the problem of analyzing spectroscopic cubes by reviewing current approaches to do this. Section 3 presents an homogeneous compact representation for spectroscopic cubes, while Section 4 presents the experimental results of computing this representation. In Section 5 we show how to use our representation for data analysis. Finally, we conclude in Section 6 giving remarks and discussing future work.

*mauricio.araya@usm.cl