Design of a machine learning model for the detection of young planets

September 21, 2021

Abstract

Protoplanetary discs present substructures, such as axisymmetric regions of luminosity depletion (gaps), that can be explained by the presence of forming planets. Features of these objects can be inferred from their observation and analysis. A remarkable example is the estimation of the planetary mass from the gaps morphology. The approaches currently used, empirical formulae or numerical simulations, are both limited in precision or applicability. In this thesis we propose a machine learning approach: using a neural network to infer this information from disk images with the requirement of the least amount of physical features not directly observable. Possible future developments of such models require data for the train and test phases. We design and build a database for this purpose collecting data obtained from numerical simulations and providing an easy-to-use interface for the implementation of machine learning models using TensorFlow libraries.

Contents

Chapter 1

Introduction

Chapter 2

Protoplanetary discs

During its formation, a star accrete its mass drawing matter from the surrounding structures of gas and dust. This matter is usually organized in discs which evolve along with the star. At some point in the star formation process, after $\sim 10^5$ yr, when most of the disc matter has accreted into the star the remaining disc is called protoplanetary disc. At this stage the disc temperature and emission is mainly due to the irradiation it receives from the star. In these astronomical objects planet formation takes place: solid fragments called planetesimals form from disc matter and start to accrete their mass with different mechanisms. The study of protoplanetary discs is thus of primary importance in the detection and characterization of young planets allowing the development and test of planet formation theories.

2.1 Structural properties

Protoplanetary discs form in the context of star formation. This process takes place in specific regions known as "star-forming regions" filled with the interstellar medium: a mixture of gases, mainly hydrogen and helium, enriched with some heavier elements. The gravitational collapse of denser regions gives birth to stars and eventually, after $\sim 10^5$ yr, leads the formation of protoplanetary discs. These regions had been widely studied and classified. Many features of protoplanetary discs are intimately linked to the star formation process and to the environment they generate from.

The disc structure appears as simply a consequence of angular momentum conservation. We will always consider axisymmetric discs. In the following sections we will use cylindrical coordinates defining the frame of reference in figure 1, to discuss disc properties.

Most of the discs were observed at distances of about 150 pc with a typical diameter of 100 a.u. meaning that they span approximately 1 arcsec of the sky.

Two main components can be distinguished according to their physical state: gas and solids. The solid component consists in dust and debris of different dimensions, going from micrometers to few meters, which build up about the 1% of the total disc mass. Despite being a fraction of the disc the solid components are actually the easier to observe and measure due to some features which will be further discussed in section

Dust and solid fragments are embedded in a gaseous medium which provide most of the disc mass. The most abundant molecule is H_2 which is challenging to observe due to its lack of a dipole moment. Measures related to less abundant molecules, such as HD or CO, provide insights into the properties of the gas component.

The overall mass of protoplanetary discs is measured to account for some jupyter masses. Estimates of these quantities can provide upper limits to the masses of forming planets.

Gas and dust temperature is strictly related to sundry factors both in the dynamic and radiative emission. Its value changes with the radial and vertical distance to the star going from hundreds of K to approximately 20K. The interstellar medium, the background of observations, has observable features compatible with a temperature of about 10K.

2.2 Disc dynamics and evolution

The disc evolution is governed and properly described by fluidodynamics. (+forze gravità) Some assumptions need to be made in order to acquire a predictive model of practical use. The first one is called "thin-disc approximation" which consists in assuming that the radial distance is greater than the vertical typical length scale H, thus requiring $H/R \ll 1$. This quantity, called aspect-ratio, has been measured showing a value of ~ 0.1 which justify the approximation. This assumption allows the study of disc properties integrating the equations along the vertical direction Self-gravitation of the disc will be neglected. The stability condition of the disc against self-gravity can be written in the form

$$\frac{M_{disk}}{M_{star}} \lesssim \frac{H}{R} \tag{2.1}$$

This condition is well satisfied in the late epochs when protoplanetary discs are studied.

Keeping in mind these assumptions, I am going to further describe how the fluidodynamic description is applied to model the disc structure and dynamics. First, I am going to focus on the gas component which is modelled as an ideal gas. The results obtained in the study of the gas account for most of the macroscopical disc features due to its relative abundance with respect to the solid component.

The vertical structure of the gas is determined by a steady-state solution of the hydrodynamical equations and the Poisson equation that accounts for the gravitational potential. The assumptions stated above allow a great simplification of this problem leading to the vertical density profile

(2.2)

The equation above offers a quantitative definition of the aspect-ratio H, previously introduced as the typical height length scale: it is the standard deviation of the Gaussian describing the vertical density profile. It also provides the relation

$$H = \frac{c_s}{\Omega_K} \tag{2.3}$$

Where c_s is the sound speed defined as $c_s^2 = \frac{\mathrm{d}P}{\mathrm{d}\rho} = \frac{k_B T}{\mu m_p}$ while Ω_K is the Keplerian angular velocity.

In the simplest disc models shear viscosity is taken into account with the introduction of the new parameter α called "Shakura-Seneyev viscosity" which gathers all the ignorance about viscous processes. This parameter is introduced through the following reasoning. In an ideal gas $\nu = \frac{1}{3} c_s \lambda$, with λ indicating the mean free path of particles in the fluid. In protoplanetary discs we have to consider also the turbulent regime. Dimensional arguments lead to the assumption of $\nu_T \sim v_T \lambda_T$. In this equation v_T indicates a typical velocity of turbulent motions which should satisfy $v_T \lesssim c_s$, upper velocities would lead to shocks thermalizing the turbulent motion. The λ_T factor represents the typical length scale in the turbulent regime which, assuming isotropic turbulence, can not be greater than H, the disc heigth.

The dust component is modelled as a pressure less fluid with grains of different dimensions coupled with the gas medium. The strength of the coupling is expressed by the Stokes number St. Two drag forces come into play depending on the grain size. If $s \lesssim \lambda$ (with λ indicating the mean free path of molecules within the disk), the drag force is called Epstein drag. In this regime, which is usually the most relevant for most particle sizes, the drag is caused by the difference in the frequency of collisions with the gas molecules between the front and back size of the grain as a consequence of its motion in the gas medium. Once particles reach sizes much larger than the molecular mean free path they begin to experience a force of different nature called Stokes drag.

2.3 Observations

Here I am going to explain how discs are observed presenting the different observational primers for the gas and the dust component. I am going to present some links between structural and observational properties (such as $\lambda \sim s/2\pi$).

I am also going to explain which is the best image resolution currently achievable.

2.4 Planet formation

Here I am going to discuss how planets are formed within these discs. I will explain:

— how they accrete their mass
— the forces they experience and thus the radial drift
— the substructures they form in the disc
— typical mass values and their relation with substructures (qualitatively)

2.5 Gaps

Here I am going to:

— explain how planets are not the only thing that generates gaps
— present the radial profile of a gap in dust and gas densities
— explain the differences between gap structures in gas and dust
— explain how depth and width are defined

Chapter 3

State-of-the-art investigative techniques

3.1 Addressed questions

Here I am going to explain which features are commonly extrapolated from discs images and why they are important. Then I am going to state that in the following paragraphs I will mainly focus on methods for the determination of embedded planets' masses.

3.2 Empirical formulae

In this section I am going to explain that many disc features can be analytically linked using simple linear or power laws.

3.2.1 Planet mass and gap width

I am going to focus on the link between planet's mass and gap width, providing the 'Lodato' and 'Kanagawa' models.

3.2.2 Strengths and limitations

Here I am going to discuss the strengths and limitations of the analytical approach. From this section it should be clear why numerical simulations are preferred.

3.3 Numerical approach

In this section I am going to explain how disc features can be inferred from simulations of the entire disk. I am going to present the possibilities in the choices of the simulating software. I am then going to focus on a specific choice and discuss the main steps of the simulation workflow.

This section plays a double purpose: it presents the current approach for the study of protoplanetary discs and explains how the images used to build the database were generated.

3.3.1 Hydrodynamical simulations

Here I am going to give some background about phantom and the type of data it generates.

3.3.2 Radiative transfer

Here I am going to discuss the software used for radiative transfer: MCFOST. I will explain why this step is performed and the meaning of the results obtained.

3.3.3 Generation of synthetic images

Here I am going to discuss the different methods that can be used to simulate the limitations of observing instruments. I will further explore some key features of pymcfost.

3.3.4 Strengths and limitations

Here I am going to discuss the strengths and limitations of numerical simulations in the context of disc analysis. From this subsection the need for a faster method should emerge.

Chapter 4

Machine learning and neural networks

Very brief introduction to machine learning. (Birth and definition)

4.1 Neural networks

Basic idea behind neural networks.

4.1.1 The perceptron

Here I am going to present the perceptron model explaining:
— how outputs are generated from inputs
— what are the weights tuned during the training process
— what is the activation function

4.1.2 Architecture and types

In this subsection I am going to explain how perceptrons are organized within a neural network. The concept of layer (and hidden layer) will be explained. After acknowledging the existence of many types of neural networks I am going to focus on the Feedforward model.

Then I am going to discuss the possibility to improve a feedforward neural network with convolutional layers. I will explain how they work, what they are designed for and how they can be exploited.

4.1.3 Training

Here I am going to explain the key steps of the training process. I will explain how it works, what algorithm can be used and the concepts of loss functions and metrics.

4.2 Strengths and limitations

In this section I am going to discuss the strengths and limitations of machine learning techniques, focusing on aspects with direct relevance to this thesis.

4.2.1 The universal approximation theorem

Here I am going to discuss the flexibility of neural networks and the theoretical framework that proves their potential.

4.2.2 Hyperparameters

Here I am going to write about hyperparameters. I am going to list them providing basics explanations about how their value can affect the model. They will be presented as both a strength and a limitation.

This subsection should highlight the importance of carefully tuning the hyperparameters.

I am going to cite the existence of algorithm for doing this job automatically and more efficiently than by simple trial and error.

4.2.3 Training data

Here I am going to discuss the importance of having a large dataset in the implementation of a machine learning model. I am going to weight pro and cons of this data driven approach.

4.2.4 Overlearning and underlearning

Here I am going to discuss overlearning and underlearning. I am going to:

— define them

— explain their causes

— provide a method for their detection

— discuss the solutions (early stopping, vary the number of trainable parameters, ...)

4.2.5 Computational complexity

Here I am going to discuss the computational complexity of machine learning algorithms in comparison with numerical simulations.

I have to highlight that the most resource requiring part is the training process. The aim is thus to obtain a trained neural network that can be deployed and used for the study of a wide range of different disc images without the need to re-train it.

4.3 Machine learning and protoplanetary discs

In this section I am going to develop the idea of applying machine learning methods to the study of protoplanetary discs.

4.3.1 The proposed approach

Here I am going to discuss the approach we want to propose. I am going to give some details about the suggested architecture for the neural network and what we expect to be able to predict with the trained model.

I am going to think about possible scenarios which can take advantage from this approach (ex. large surveys with lots of disc images: the neural network could quickly provide measures of their physical properties (?))

4.3.2 Previous attempts in literature

Here I am going to discuss the Sayantan Auddy and Min-Kai Lin's paper citing their results and explaining the main differences with the approach we propose.

Chapter 5

Dataset design

In this chapter I am going to unfold the main part of my work: the design and implementation of the dataset. Here I am going to recall the key features of a good dataset for machine learning.

5.1 The data

In this section I am going to present what the dataset is made of: fits images and the list of parameters included in the fits files and in data.js. I am going to briefly explain why they were included and their possible use.

I am also going to present images showing the data distribution over the parameters space, discussing which of them are properly explored and which not.

5.2 Structure and interface

Here I am going to discuss the structure we designed for the db and the interface provided to access the data.

5.3 Supporting scripts

Here I am going to write about the scripts I wrote which allow the user to handle the database and preprocess the results coming from MCFOST simulations.

5.4 Expanding the dataset

Here I will explain how the tools provided allow to easily expand the dataset. Then I am going to give future perspectives on how the database could be improved.

Chapter 6

Proof of concept

Here I am going to present an example of model trained to predict planet's mass from disc images. From the different models I tested (and I will test) I am going to report here the one which gives the best result. I am also going to cite the python libraries used to implement this model (TensorFlow).

6.1 Adopted model

Here I am going to explain the model used: number of layers, type of neural network, activation functions, optimizer, metrics.

6.2 Data pre-processing

Here I will explain how I chose the data used during the training process. This section should highlight the versatility of the dataset showing the possibility to split the data according to our needs.

6.3 Results

Here I will discuss the results obtained.

Chapter 7

Conclusions

7.1 Conclusion

7.2 Future perspectives

Acknowledgements

Bibliography