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

July 26, 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

Very brief introduction (1/2 sentences) about what a protoplanetary disc is, how they generate, where we can find them and why they are studied.

2.1 Structural properties

Here I am going to discuss some key properties of protoplanetary discs and give gross estimates of their typical values. I am going to discuss:

— what discs are made of
— absolute and relative masses of gas and dust components
— dimension and distance
— their age
— temperature

2.2 Disc dynamic and evolution

Here I am going to explain how the dynamic of gas and dust is modelled.

I'm going to provide the equations describing the vertical structure, explain the meaning of the aspect ratio and the viscous forces at play.

I will also explain the model describing the interaction between the gas and solid components (Epstein force, stokes number).

Finally, I am going to cite other forces and effects which play a role in disc dynamic, such as magnetorotational instability, turbulence, winds, photoevaporation, ...

2.3 Observations

Here I am going to explain how discs are observed, present 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 prove 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 tune 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 realization 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 hydrodynamical 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 the 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 have chosen 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