PHY - AstroStatistics : Statistical Inference in Astrophysics/Cosmology

Instructor: Dr. Muntazir Abidi

Course Credits: 3.0

Table of contents

1 Syllabus

Instructor	Dr. Muntazir Abidi (muntazir.mehdi@sse.habib.edu.pk)
Office Hours	TBD
Lecture Timings	TBD
Pre-requisite	Calculus I and Linear Algebra
Meets requirements for	Phys, CS, Maths, EE
TAs	TBD

1.1 Overview

The known universe contains a lots of stars, galaxies and other objects. To a rough estimate there are 100s of billions starts in our galaxy and there are approximately almost 100 billion galaxies exist in the known universe. Modern cosmological and astrophysical observations are gathering enormous amount of data which requires researchers to come up with novel ways to interpret those data and find interesting information. Using the statistical tools people can measure the reliability of their measurements, quantify uncertainties in their theoretical models and find interesting patterns in observational data [1].

1.1.1 Big Problems in Cosmology

Over the last few decades, there have been tremendous theoretical and observational advancements in cosmology that have greatly improved our current understanding of the Universe. However, there are still many fundamental questions that need to be answered. For instance, why is the universe accelerating? What is the nature of Dark Energy and Dark Matter that dominate the Universe? What is the physical origin of inflation which provided the seeds

[1] https://www.cfa.harvard.edu/research/topic/astrostatistics

Fig. 1.1 The galaxy cluster as seen by NASA's Hubble Space Telescope. This cluster was identified by the Sloan Digital Sky Survey (SDSS). Credit: ESA/Hubble & NASA

for the formation of large scale structure (LSS)? Is Einstein theory of general relativity a correct description of gravity or does it need to be modified on large scales? The current and future cosmological surveys have huge potential to answer these questions because of the enormous data we expect to get from these several cosmological observations.

1.2 Universe in Data

1.2.1 Cosmological Surveys: Nature of the Data Sets

There are two types of large-scale structure (LSS) observations in cosmology. One is called Photometric surveys and other is called the spectroscopic surveys. In Photometric survey we take the images of galaxies. These images contains information about the 3D clustering of galaxies in 2D projected images. The colors and the brightness of these beautiful images which contain interesting information that scientists need. On the other hand, spectroscopic surveys detect light coming from the far away galaxies from different part of the sky, from the information of the redshift in those light rays one can infer how far those galaxies

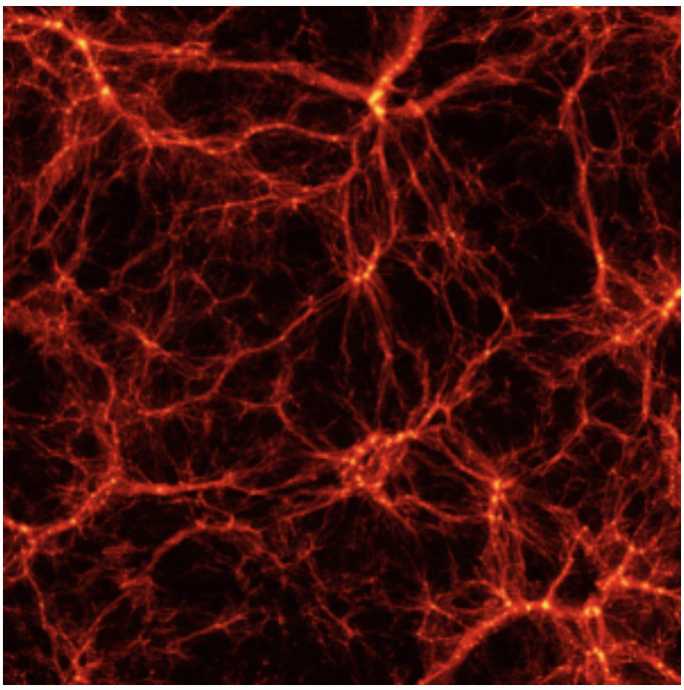

Fig. 1.2 Snapshot of the dark matter clustering of the Universe obtained from Quijote Simulations [2]

are from us. This mapping provides a 3D information of galaxies. Researcher have been developing powerful statistical tools to handle this enormous amount of data, and extract the useful information about the Universe as well as any unknown underlying physics.

1.2.2 How to test theory with observations

We only have ONE Universe. The experiments are expensive and we can only do the observation of the Universe once. how can we test whether what we are observing is actually what we would expect to observe or does it contain all noise. Extracting signal from noise and validate theoretical models to underlying physics is a big challenge. One way to overcome this challenge is to test theories and statistics on computer simulations before applying them directly to observational data. Power computer simulations have been developed which we can use to test our models. One example is a recent state of the art simulations called Quijote Simulations [3]. These simulations have been obtained using 35 Million CPU hours. It contain 8.5 trillions of particles at a single redshift and billions of dark matter halos and voids. To give an idea 1 Petabytes of Data is available publicly from these simulations. Good news is we can apply statistical methods on these simulations and can also use them to train Machine Learning models for cosmology. In cosmology, machine

[3] https://quijote-simulations.readthedocs.io/en/latest/features.html

Fig. 1.3 Astrostatistics as an interdisciplinary intersection of statistics and astrophysics.

learning models have started to play an extremely important role to extract cosmological information. The idea is we can train machine learning models on powerful cosmological simulations (such as Quijote Sims) and test on the data sets from cosmological survey to get precise measurements of the cosmological parameters.

1.2.3 About Astrostatistics

This course is about applied statistical methods necessary to interpret complex data sets in cosmology/astrophysics, as well as how machine learning can be applied to these complex data sets. Astrostatistics is an interdisciplinary intersection of statistics and astrophysics Fig. 1.3. There is no theory of astrostatistics, it is application-driven, applied field, where we will be focusing on real problems. using the tools in this course we will be able to understand how to interpret and analyse large and complex astronomical data sets. We will think about where do the data come from? how to build a statistical model? What assumptions do we make, whether implicit or explicit? Are these assumptions reasonable? How do we pick the right model given a data set? How about if we have different data sets or too many models, how to choose the right one, or the one more closer to the reality.

We will start with basics probability theory and statistics, and will cover some advanced topics such as Bayesian inference, sampling methods, Gaussian processes and model selection methods. We will also cover basic machine learning methods, regression models, classification methods, and neural nets. Examples will be motivated by case studies across cosmology and astrophysics.

1.2.4 Examples of Cosmological Surveys

Some famous examples of cosmological surveys are given below:

1. Plank Satellite : measurements of cosmic microwave background (CMB) radiations, finished in 2018

2. Euclid Satellite by European Space Agency

3. Sloan Digital Sky Survey by NASA

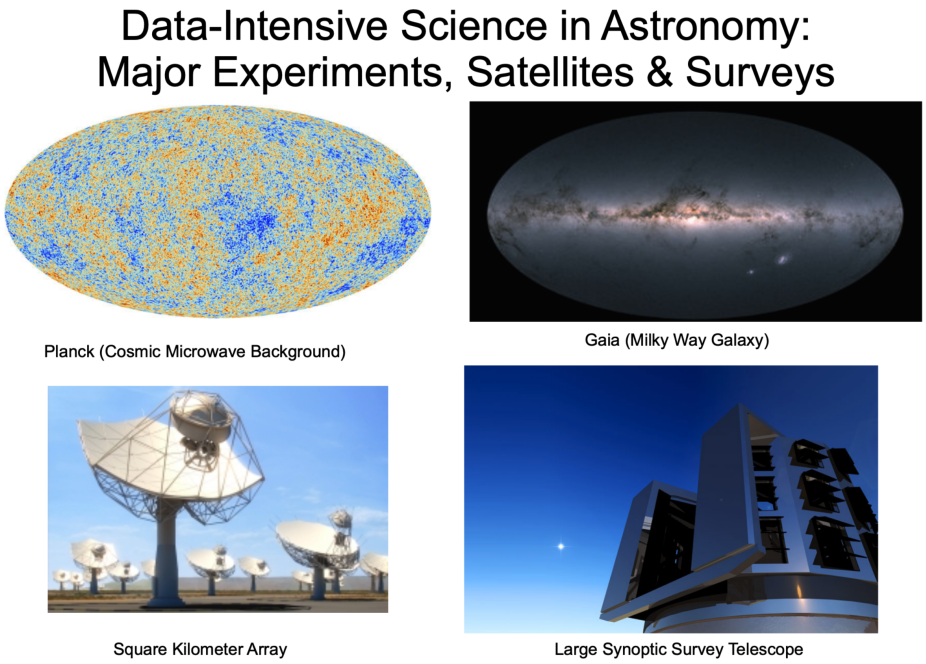

Fig. 1.4 Cosmological Survey examples

4. Large Synoptic Survey Telescope (LSST)
5. Square Kilometer Array (SKA), radio telescope
6. Gaia by European Space Agency to provide a precise three-dimensional map for our Milky Way galaxy.

1.3 Learning Outcomes and Objectives

1.3.1 Aim

I have seen a lot of enthusiasm from students about astrophysics. However, due to lack of formal education is astrophysics and unavailability of relevant courses in Pakistan these students can not pursue research in these fields. The aim of this course is to teach students important statistical tools which, if they like in future, can be used to do research in astrophysics and cosmology.

Another aim is to make this course beneficial for those interested in astrophysics as well as in statistics. The goal is to make you think critically about the data set, rather than blindly applying black-box models. Understanding statistical methods are crucial if you want to understand the results properly. Astrophysics is an observational field, billions of dollars

have been spent on building experiments and we can only do the observations once from these experiments. Our goal is to be able to be able to extract as much scientific information possible from the complex data sets.

Using the tools developed in this course provide a strong premise to be able to understand how to answer some deep questions about the Universe.

1.3.2 Learning Outcomes

By the end of this course students should be able to understand basic statistics and probability. How to interpret cosmological data and find patterns. basic machine learning methods which have interesting applications in cosmology/astrophysics. They should also be able to solve theoretical problems as well as solve problems using computers (such as Python). Introductory Python sessions will be given during the lectures. Although the case studies will be from cosmology, the emphasis through out the course will be skills, more important problem solving skills. Using these useful skills they can be able to solve similar problems in other data-intensive fields.

The list of important are given below:

1. Mathematical Modelling
2. Statistical Modelling
3. Understanding complex data sets
4. Problem solving
5. Basic understanding of Cosmology / Astrophysics
6. Quantitative Reasoning
7. Scientific Methods

1.4 Format and Assessment

1.4.1 Delivery

The format of these lecture will be alternative synchronous and pre-recorded lectures. Every week, there will be interactive online lectures where new concepts and derivations will be discussed. I will upload pre-recorded videos for several important concepts covered in the course. Live programming sessions will be held to familiarize students with Python. There will be theory lectures, problem solving lectures, and training sessions where students will

Grading Scale		
Letter Grade	**GPA Points**	**Percentage**
A+	4.00	[95, 100]
A	4.00	[90, 95)
A-	3.67	[85, 90)
B+	3.33	[80, 85)
B	3.00	[75, 80)
B-	2.67	[70, 75)
C+	2.33	[67, 70)
C	2.00	[63, 67)
C-	1.67	[60, 63)
F	0.00	[0, 60)

Fig. 1.5 Grading scale

use computers to solve problems. Problem solving sessions will be interactive and credits will be given to students for active participation.

1.4.2 Technological Requirements

Computer is needed for online lectures with Zoom installed. For python sessions, python 3.++ version should be installed, as well Jupyter notebook.

1.4.3 Assessment

The following forms of assessments will be used in this course.

1. **Introductory Discussions:** One-time discussion to introduce each other and also discuss about the course.

2. **Example Sheets:** Example Sheets will be uploaded 4 or 5 (exact no tbd.) These will contain theoretical problems to solve - 10% of the final grades.

3. **Training sheets:** Besides example sheet, training sheets will be given, these will contain technical problems to solve using computers. Solutions will be discussed during lectures. 10% of the final grades.

4. **Quizzes:** Will be conducting quizzes online, total 4, will be announced 2 weeks earlier. 5% of the final grades.

5. **Mid Term:** there will be only one midterm, will be conducted online.20% of the final grades.

6. **Final:** There will a final exam at the end of the course 30% of the final grades.

7. **Course project:** There might be a course projects and will be required to submit a report. I will decide later about it and if it will be included students will be informed at the beginning of the course 20% of the final grades.

8. **Student Participation:** I encourage students to actively participate. Credits will be given - 5% of the final grades.

1.4.4 Grading

Knowledge and skills assessed: All assessment tasks will be assessed against the learning outcomes outlined above, specifically, the ability to apply the concepts discussed in class to solving simple and complex problems and the ability to communicate mathematical arguments in a coherent and logical manner.

Assessment criteria: The criteria for marking all assessment tasks will focus on correct working and appropriate reasoning in solutions and not just on correctness of answers. Indeed, a wrong attempt producing an answer coinciding with the correct answer will be discarded. On the other hand, no mistake will be penalized twice and a logical output of a mistake will be graded.

Grading scale is given in Fig 1.5.

1.5 Content of the course

1.5.1 Modules

Following Topics will be covered:

1. Quick Introduction to Cosmology
2. Probability and random variables
3. Descriptive statistics
4. Distribution functions
5. Central Limit Theorem
6. Entropy and Inference
7. Classical Statistical Inference

- Maximum Likelihood Estimation (MLE)
- Goodness of fit and model selection

8. Bayesian Statistical Infernece
   - Introduction to the Bayesian Method
   - Bayesian Priors
   - Bayesian model selection
   - Numerical Methods for complex problems (MCMC)
9. Marginalisation and Fisher forecasts
10. Regression and model fitting
11. Classification
12. Neural Networks and Gaussian Processes
13. Introduction to Scientific Computing with Python

1.5.2 Readings and Textbooks

Following books will be used to cover the lectures. Specific readings will be provide.

1. Information Theory, Inference, and Learning Algorithms by David J.C. MacKay : http://www.inference.org.uk/itprnn/book.pdf
2. Pattern Recognition and Machine Learning by Christopher Bishop: https://www.microsoft.com/en-us/research/people/cmbishop/prml-book%20pdf/
3. 21st Century Statistical and Computational Challenges in Astrophysics : https://arxiv.org/pdf/2005.13025.pdf
4. Statistical methods in cosmology by Licia Verde : https://arxiv.org/pdf/0911.3105.pdf
5. Statistics, Data Mining and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data by Zeljko Ivezic, Andrew J. Connolly, Jacob T. VanderPlas : https://www.astroml.org/_downloads/1270ba6a44f327517e5e322a1ba0845e/DMbookTOC.pdf
6. More advanced level : https://github.com/3raserback/PartIII-Astrostatistics-2019

Useful YouTube Channel to learn Statistics: https://www.youtube.com/c/joshstarmer

Python: https://www.python.org

Jupyter Notebook download: https://jupyter.org

1.6 Logistical Policies

1.6.1 Academic Integrity

Each student in this course is expected to abide by the Habib University Student Honor Code of Academic Integrity. Any work submitted by a student in this course for academic credit will be the student's own work. There is zero tolerance for plagiarism. Every case will be reported to the conduct office and you'll get a zero on that particular test or assignment. Scholastic dishonesty shall be considered a serious violation of these rules and regulations and is sub- ject to strict disciplinary action as prescribed by Habib University regulations and policies. Scholastic dishonesty includes, but is not limited to, cheating on exams, plagiarism on assignments, and collusion.

Plagiarism: Plagiarism is the act of taking the work created by another person or entity and presenting it as one's own for the purpose of personal gain or of obtaining academic credit. As per University policy, plagiarism includes the submission of or incorporation of the work of others without acknowledging its provenance or giving due credit according to established academic practices. This includes the submission of material that has been appropriated, bought, received as a gift, downloaded, or obtained by any other means. Students must not, unless they have been granted permission from all faculty members concerned, submit the same assignment or project for academic credit for different courses.

Cheating: The term cheating shall refer to the use of or obtaining of unauthorized information in order to obtain personal benefit or academic credit.

Collusion: Collusion is the act of providing unauthorized assistance to one or more person or of not taking the appropriate precautions against doing so. All violations of academic integrity will also be immediately reported to the Student Conduct Office.

You are encouraged to study together and to discuss information and concepts covered in lecture and the sections with other students. You can give "consulting" help to or receive "consulting" help from such students. However, this permissible cooperation should never involve one student having possession of a copy of all or part of work done by someone else, in the form of an e-mail, an e-mail attachment file, a diskette, or a hard copy. Should copying occur, the student who copied work from another student and the student who gave

material to be copied will both be in violation of the Student Code of Conduct. During examinations, you must do your own work. Talking or discussion is not permitted during the examinations, nor may you compare papers, copy from others, or collaborate in any way. Any collaborative behavior during the examinations will result in failure of the exam, and may lead to failure of the course and University disciplinary action. Penalty for violation of this Code can also be extended to include failure of the course and University disciplinary action.

1.6.2 Students with disability

In compliance with the Habib University policy and equal access laws, we are available to discuss appropriate academic accommodations that may be required for students with disabilities. Requests for academic accommodations are to be made during the first two weeks of the semester, except for unusual circumstances, so arrangements can be made. Students are encouraged to register with the Office of Academic Performance to verify their eligibility for appropriate accommodations.

1.6.3 Inclusivity Statement

We understand that our members represent a rich variety of backgrounds and perspectives. Habib University is committed to providing an atmosphere for learning that respects diversity. While working together to build this community we ask all members to share their unique experiences, values and beliefs, be open to the views of others and value each other's opinions and communicate in a respectful manner. We also ask them appreciate the opportunity that we have to learn from each other in this community and use this opportunity together to discuss ways in which we can create an inclusive environment in this course and across the Habib community. It is important to keep confidential discussions that the community has of a personal (or professional) nature.

Ref: [4]

[4]This section is taken from Linear Algebra Fall Course 2020

=====