Large Scale Inference & Multiple Testing

SUMMARY.

This METEOR provides an introduction to advanced statistical concepts and methods for the analysis of large data sets in Astrophysics, with a focus on detection. Today, astrophysical research routinely deals with large data sets (Gaia for instance). But data sets composed of millions of images or of times series pose intrinsic methodologic problems. Simply applying in parallel one million times the same detection process used for one image does not work, because this either leads to an explosion of false positives or to procedures with vanishing detection power. Specific statistical methods derived in the last decades to cope with this phenomenon will be studied. METEOR projects implementing such methods will be taken in a wide range of topics, from cosmological studies to exoplanets detection, with data from MUSE and SPHERE instruments @ VLT, or space mission like Gaia, KEPLER and JWST.

OBJECTIVES

- One main objective of this METEOR is to train the students to learn in autonomy, to identify what can help them to progress, to identify and correct their errors, to define a project related to a problem of their interest and to solve it.
- The students will understand and practice general methodological tools in multiple testing and statistics.
- The students will learn to identify specific problems posed by the analysis of large astrophysical data sets and to pose these problems in statistical terms.
- The students will learn to find solutions for specific detection problems found in Astrophysics and to evaluate their performances.
- The students will learn to implement their solutions in the form of Python programs.

PREREQUISITES

Preparatory MAUCA courses:

- Statistical methods
- General Astrophysics
- Numerical methods
- Signal & Image processing

Interested students are encouraged to contact the supervisor and previous students who have followed this METEOR (or the METEOR 'Detection of Exoplanets', with the same supervisor) to have an idea of how the METEOR works.

THEORY

by D. Mary

This METEOR contains five main parts:

1. Chapter 1 deals with Empirical Bayes estimation, a powerful methodology, with works by Stein (in estimation) and Robbins (detection). We will see that empirical Bayes estimation blurs the frontier between estimation and detection and between frequentist and Bayesian approaches.
2. Chapter 2 introduces the problem of large-scale hypothesis testing, which is a framework not envisioned in the classical frequentist detection theory of Neyman, Pearson and Fisher. The important concept of False Discovery Rate and its connection with Empirical Bayes will be discussed.
3. Chapter 3 turns to specific significance testing algorithms aimed at controlling the false alarm rate (or familywise error rate, FWER)

when making N simultaneous tests.

4. The limits of the classical methods of Chapter 3 when N is large (millions or billions) lead to other procedures, aimed at controlling different types of errors, like the False Discovery Rate (FDR). A very important detection algorithm controlling the FDR while being more powerful than classical approaches controlling the FWER is called the Benjamini-Hochberg procedure. This approach of the late 90's has brought a genuine 'revolution' in the thinking and practice of sciences involving detection tests.

5. The last Chapter turns to approaches coupling modern machine learning methods (like neural networks) and multiple testing. In particular we will study robust detection methods (inspired from the problem of detecting galaxies in MUSE data) with theoretical guarantees on the FDR even when the noise distribution is not well known.

APPLICATIONS

by D. Mary

- With the help of the supervisor, students will define a small research project according to their personal interest. Following the students' interests, the detection techniques learned in the theoretical part will be applied in this project to data from MUSE and SPHERE instruments @ VLT, or space mission like Gaia, KEPLER and JWST.
- The METEOR provides an intensive training to Python for the numerical exercises of the Theoretical part and for the project.

MAIN PROGRESSION STEPS

- During the whole duration of the METEOR, each student has a personal channel on Discord allowing easy connection with the supervisor outside the scheduled meeting slots. A general channel serves also as a forum for general infos/questions/hints.
- First half of the period (possibly more): the students learn theory. They are requested to work on the lecture notes on their own, with regular discussions planned with the supervisors to answer their questions. They do the theoretical and numerical exercises proposed in the lecture notes and they post the solutions on the fly on their personal channel. As in the Statistical Methods lecture, each chapter has its own 'Friendly Quiz' and 'Noted Quiz'.
- @ mid METEOR, the students identify a point of the lecture they are mostly interested in and define the framework of their project: they choose their favorite astrophysical targets (detection of exoplanets, asteroids, galaxies,...) with the corresponding data sets. They define the statistical problem to be studied and the algorithms that should be implemented.The supervisor helps the student to ensure that the project's objectives are relevant and reachable.
- Second half of the period : the students work on their research project.
- Last week : last results and preparation of the final oral presentation.

EVALUATION

- Average mark of 4 quizs, one mark for the numerical exercises, one mark for the final written exams (2h) on the theoretical part. The average of the three marks provides the mark 'Theory' (30% of the total mark).
- The mark for the 'Project part' is the average of 6 marks (autonomy; interaction; initiative; efficiency; progression (final project status); critical thinking).
- Final evaluation during the global oral presentation (40% of the total mark).

BIBLIOGRAPHY & RESSOURCES

- On-line lecture notes, slides,, homeworks, criteria evaluation grid, data, solution codes.
- *Large Scale Inference*, Cambridge University Press, 2013
- *Statistics, Data Mining & Machine Learining in Astronomy*, Princeton Series in Modern Observational Astronomy, Second Edition, 2020
- *Modern Statistical Methods for Astronomy*, Cambridge University Press, 2012
- *Computer Age Large Scale Inference*, Cambridge University Press, 2019

CONTACT

☎ +33492076384 (D. Mary)
✉ david.mary@oca.eu