Data Science and Information Theory (ED127 course, 2025)

Florent Leclercq[a)]
CNRS & Sorbonne Université, UMR 7095, Institut d'Astrophysique de Paris, 98 bis boulevard Arago, F-75014 Paris, France

(Dated: 5 November 2024)

PROGRAMME

This will be a 6-day course/workshop on data science methods and information-theoretic tools for data analysis, aimed principally at first- and second-year PhD students. Sessions will be 09:45–12:30 and 14:00–17:45 on 31 March, 1, 2, 7, 8 and 9 April 2025. All sessions will be held in the computer room (salle de TP 35-37, ground floor) of the IAP.

A preliminary schedule is as follows:

Day 1: Probability theory and signal processing

- Historical and conceptual introduction
- Probability theory: inference, prediction, priors, maximum entropy principle
- Bayesian signal processing: Gaussian random fields, Wiener filtering, signal de-noising and de-blending

Day 2: Monte Carlo techniques

- Monte Carlo integration
- Markov Chain Monte Carlo, the Metropolis-Hastings algorithm
- Slice sampling, Gibbs sampling, Hamiltonian Monte Carlo

Days 3 and 4: Advanced Bayesian topics

- Model comparison, model averaging
- Bayesian decision theory and Bayesian experimental design
- Bayesian hierarchical models
- Fisher information and forecasts
- Simulation-based inference/implicit likelihood inference
- Caveats!

Day 5: Information Theory

- The noisy binary symmetric channel
- Shannon's theorem
- Measures of information and information-theoretic experimental design
- Thermodynamics and inference

Day 6: Machine Learning Theory

- History of Machine Learning
- Statistical learning theory
- Learning via optimisation

The format will be interactive, featuring examples and practical applications throughout. Presentation of new material will be interspersed with exercises, as suitable. There is enough room in the schedule to respond to your specific needs– the material will be adapted to your interests.

The Institut d'Astrophysique de Paris has a Code of Conduct. By registering to this course, you are agreeing to abide by it.

[a)] Electronic mail: florent.leclercq@iap.fr; https://www.florent-leclercq.eu/

Schedule	Monday 31 March	Tuesday 1 April	Wednesday 2 April	Monday 7 April	Tuesday 8 April	Wednesday 9 April
09:45-11:00	Overview	Monte Carlo	Model Comparison	Fisher Information	Information Theory	Machine Learning Theory
11:00-11:15	Break	Break	Break	Break	Break	Break
11:15-12:30	Probability Theory	Monte Carlo	Bayesian Decision Theory	Bayesian vs Frequentist statistics	Information Theory	Machine Learning Theory
12:30-14:00	Lunch	Lunch	Lunch	Lunch	Lunch	Lunch
14:00-15:15	Bayesian Signal Processing	Monte Carlo	Bayesian Hierarchical Models	Implicit Likelihood Inference	Information Theory	Machine Learning
15:15-15:30	Break	Break	Break	Break	Break	Break
15:30-17:45	Bayesian Signal Processing	Monte Carlo	Data science for your research project	Caveats	Thermodynamics and Inference	Machine Learning
Colour code	Lecture	Hands-on session	Discussion or Q/A session			

TRAINING OBJECTIVES

At some stage, most researchers will need to conduct some form of data analysis, ranging from basic line-fitting and parameter estimation to complex, computationally intensive sampling for model selection on large datasets. Anecdotal evidence indicates that many doctoral researchers are not well-equipped for such tasks, often using correct approaches improperly or selecting unsuitable statistical tools. The purpose of this course is to build a solid understanding of principled data analysis and provide practical experience in applying appropriate methods to data.

The training objectives are to provide participants with a comprehensive understanding and practical experience in key areas of data science and information theory. They will develop a solid understanding of probability theory and apply Bayesian methods for signal processing, inference, prediction, model comparison, decision-making, and experimental design. Participants will gain the ability to build and implement Bayesian models effectively, including Bayesian hierarchical models. They will receive both theoretical and hands-on experience with numerical techniques such as Markov Chain Monte Carlo and simulation-based inference for data analysis with implicit likelihoods. Through the exploration of information theory, participants will understand the principles of data transmission and data compression, and leverage information measures for designing experiments. Additionally, they will acquire the foundations of statistical learning theory, enabling them to grasp the fundamental concepts underlying many popular machine learning algorithms.

TEACHING METHODS

The course plan combines lectures with hands-on computational work. It will concentrate on setting down firm foundations of principled Bayesian data analysis, but a feature of the workshop will be a substantial element of hands-on classes where participants will learn how to apply the ideas in practice. Two slots for discussion or Q&A sessions are also scheduled.

PREREQUISITES TO PARTICIPATE IN THE TRAINING

Basic mathematics and experience with at least one programming language are prerequisites.

We expect all participants to bring their own laptop, and to do a simple computational exercise in advance (in whatever language suits, preferentially python) to ensure they have appropriate software in place before the workshop starts. The preliminary exercise can be found at `https://cloud.aquila-consortium.org/s/Leclercq_Supernovae_Exercise`. Participants are asked to read section I–III there and code the answers to section IV. Interested readers can go on with sections V–VIII, which use notions that will be introduced during the course.

LEARNING OUTCOMES

At the end of the course, participants should be able to (non-exhaustive list):

- Express stochastic problems in terms of fundamental probability and Bayes' theorem,

- Formulate data analysis problems in a principled statistical framework, and be capable of executing some methods of solution,
- Understand the concepts of probability theory, inference, priors, posteriors, marginalisation, parameter inference, model selection, sampling, and apply them to real data,
- Code and apply a simple Markov Chain Monte Carlo sampler to physical data,
- Apply principles of information theory to understand data compression and forecast experimental results,
- Understand the basics of statistical learning theory, the link between learning and optimisation problems, and the concept of a loss function.

By the end of this course, participants will be equipped with a strong foundational and practical understanding of these crucial aspects, empowering them to perform principled data analysis and make informed methodological choices. They will be prepared to apply these methods confidently to real-world data challenges.

LECTURER

Florent Leclercq is a CNRS researcher at the Institut d'Astrophysique de Paris. As an interdisciplinary specialist in cosmology, data analysis and machine learning, his current work focuses on the development of principled statistical methods for numerical cosmology.