Universitat Pompeu Fabra Barcelona

Research Seminar: Data Science
David Rossell

2023-2024 Academic Year
Master of Research in Economics, Finance and Management

1. Teaching guide

- **Introduction**

This research seminar is part of a two course series which, jointly with "Topics in Data Science", discusses a variety of research-oriented topics related to data analysis. This research seminar focuses on laying out the foundations, placing emphasis on regression problems with many parameters and the basics of the Bayesian framework. The "Topics in Data Science" course follows up by overviewing more flexible models (semi-parametric regression, tree-based, deep learning) and latent variable models (e.g. text data analysis models) and discussing specific applications, such as principles of causal inference with many covariates.

The motivation for both courses is common. Statistical and Machine learning techniques are having a deep effect on many disciplines, including Economics. Due to the increasing number of applications using data science with techniques popular in Economics many World-leading institutions incorporated data science in their PhD programs. As examples see the talk by Novel prize Chris Sims, the set of lectures in the NBER Summer Institute by Guido Imbens and Susan Athey (in particular lecture 3, www.nber.org/econometrics_minicourse_2015/), the Big Data and Machine Learning syllabus at the Harvard Economics PhD program, the Chicago Booth PhD courses on Bayesian inference, Big Data and Machine Learning, or the Bocconi PhD courses in text analysis, Econometrics of networks and causal analysis.

Data analysis methods are evolving to cope with increasingly challenging problems, a case in point being high-dimensional situations where one considers a large number of parameters or models. For instance, one may have a regression or factor model where the number of covariates far exceeds the sample size, or may want to simplify the interpretation of complex data via clustering, text or latent variable analysis, or may want to predict an outcome using flexible algorithms. Engaging effectively in such research, either from a methodological or applied perspective, requires one to understand, and when needed modify or extend, such methodology. Just as importantly, they need to communicate such ideas effectively to a potentially non-expert audience.

The goal of this research seminar is to introduce students to some foundations behind these methods, with a certain emphasis on the Bayesian framework and penalized likelihood methods methods, expose them to and discuss research literature, and practice the skills needed for applying and presenting novel research. The learning outcomes are an improved familiarity with selected research topics in Statistics that are relevant for Data Science, at a level sufficient to critically appraise, modify and apply novel methods, and improved oral and written presentation skills. The course also intends to provide students with applied data analysis skills useful for their MRes thesis and subsequent research work.

Pre-requisites: the course is designed for MRes students who are familiar with basic Statistical inference, specifically linear regression and maximum likelihood estimation. Basic R programming skills are beneficial, though examples and links to learning resources will be provided for students who are not familiar with R.

- **Teaching methodology**

The course will be delivered in a combination of regular lectures, computer-based seminars where students get hands-on experience with the taught data analysis methods, and presentations by students (on published manuscripts chosen by the students and on the final project). The idea is to give students flexibility to deepen their understanding on the topics they find more interesting / relevant for their own research.

- **Assessment and Grading System**

Students will be asked to orally present 1 research paper of their choice (40% of the final mark), some selected exercises from the seminar sessions (10% of final mark) and a written report (50% of the final mark). This project will be decided by the students but must be pre-approved by the lecturer, and should involve the application, critical assessment or extension of the research methods seen in class. The content can be theoretical, empirical, a practical application or a combination of the former. Students who also take the "Topics in Data Science" course can submit a single joint project for both courses, specific instructions on the format will be given in class.

- **Contents**

Foundations. We review classical results from maximum likelihood estimation and computational methods such as the bootstrap, and we introduce the basic Bayesian paradigm for statistical inference, its use for model selection, parameter estimation and prediction, and standard computational tools such as Gibbs or Metropolis-Hastings. We learn programming skills in R and the probabilistic programming software STAN.

Foundations of variable selection in high-dimensional regression. We review the fundamental penalized likelihood and Bayesian frameworks for linear regression models with a large number of variables, and for treatment effect estimation in such settings. We shall discuss the relative merits of current strategies such as LASSO, adaptive LASSO and related penalties to help decouple variable selection from prediction and various Bayesian strategies to achieve good performance in high dimensions. We will discuss theoretical and practical considerations and computational strategies.

References. The books below provide a good introduction to a substantial part of the topics covered in this course (and many others), however we shall complement them with a number of additional selected research manuscripts.

- Andrew Gelman, John B. Carlin, Hals S. Stern, David B. Dunson, Aki Vehtari, Donald B. Rubin. Bayesian Data Analysis (3rd edition). CRC Press, 2013.
- Trevor Hastie, Robert Tibshirani, Jerome Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.
- Trevor Hastie, Robert Tishirani, Martin Wainwright. Statistical learning with sparsity. The LASSO and its generalizations. CRC press.
- Mahlet Tadesse, Marina Vannucci, M. (Eds.). (2021). Handbook of Bayesian variable selection.
- Sara van de Geer, Peter Bühlman. Statistics for high-dimensional data: methods, theory and applications. Springer, 2001.