UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

Inferenza Bayesiana

2425-2-F8203B042-F8203B042M

Learning objectives

The course enables the student to learn analytical and inferential procedures under the Bayesian paradigm. The Bayesian methods are illustrated according to an integrated approach with classical statistical inference.

Knowledge and understanding

The student is introduced to the main Bayesian statistical models for analysing data with different types of response variables. The relevant assumptions underlying the theory are also illustrated by considering conjugate models and estimation algorithms to obtain the a posteriori distribution through simulations. The student also learns how to deal with longitudinal data and some statistical models that take repeated measures into account. Data analysis is conducted both using R software and the RMarkdown environment, which allows reproducible documents containing code, results and comments, and using specific procedures for Bayesian analysis with SAS software. The applications cover real and simulated data from different fields of reference for the course of study. The student is also encouraged to provide a critical evaluation of the results obtained from the empirical analyses.

Ability to apply knowledge and understanding

The course provides skills in the use of Bayesian models for applications to relevant case studies in the following fields: biostatistics, epidemiology, medicine, biology, environment, genetics and public health Through R and RStudio, students learn how to organically set up statistical reasoning by analysing data and writing reports that illustrate the code, analyses and results. Through the use of SAS software, students learn to estimate complex Bayesian models through simulations and to set up the inputs required by the estimation algorithms. Theory is complemented by practical applications. The course enables students to acquire a solid theoretical foundation and the ability to apply the proposed statistical models to real data. The student is also encouraged to critically evaluate the results obtained from the empirical analyses.

The teaching is fundamental for the subsequent university course as it provides the essential concepts for the development of Bayesian methods in both the theoretical and applied fields for the target job contexts (biostatistics/statistics/demography and related) of students on the Biostatistics degree course.

Contents

Introduction to Bayesian inference and Bayes' rule. Methods of model specification and prior distributions.

Determination of the posterior distribution by exact methods.

Conjugate families: Gaussian, Poisson-gamma, beta-binomial, multinomial-Dirichlet.

Introduction to Bayesian non-parametric inference.

Methods to summarize the posterior distribution: credibility intervals and intervals with the highest posterior density.

Introduction to stochastic Markov processes, random walk.

Markov chain models for longitudinal data and latent Markov models without and with covariates.

Introduction to the Markov Chain Monte Carlo Methods: Metropolis-Hastings algorithm and Gibbs sampler.

Diagnostic tools to assess convergence of the MCMC procedure.

R environment and RStudio interface using, in particular, the following libraries: probBayes, learnBayes, LMest, LaplaceDemon.

RMarkdown will be employed to produce reproducible documents and to integrate code and output within the knitr library. SAS software with proc MCMC.

Detailed program

The Bayesian paradigm is introduced and compared with the frequentist approach, including Bayes’rule, and the total probability rule. A short introduction to Bayesian non-parametric methods is provided, and the notions of exchangeability and De Finetti’s theorem are explained. The beta-binomial model and the other conjugate families such as Gaussian, Poisson-gamma, beta-binomial, and multinomial-Dirichlet, are introduced. The choice of the prior distribution is considered. Bayesian inference is compared with the classical approach. Methods to draw conclusions from the posterior distribution include Bayesian interval estimation, credible intervals, and intervals with the highest posterior density. The prediction context is also explored along with the empirical Bayes estimation.
Several examples of the application of Bayesian models in biostatistics, using real and simulated data concerning epidemiology, drug epidemiology, medicine, biology, ecology, and environmental sciences support the theory.

An introduction to stochastic processes within the Markov random field is proposed. The properties and features of the Markov chains are illustrated and explained using simulations. The random walk process is also described. Markov chain models for longitudinal data are explained, and the latent Markov models without and with covariates are introduced both from a theoretical and applied perspective.
Markov Chain Monte Carlo (MCMC) algorithms are explained with a focus on Metropolis-Hastings and Gibbs sampling algorithms. Diagnostic evaluations of the convergence are considered.

Some time is devoted to explaining the theory by imparting the flavor of the applications using observed data arising from different fields. The examples are developed within the statistical environment R, RStudio, and RMarkdown to create reproducible documents. The SAS software is proposed to perform analyses to estimate Bayesian linear and logistic models using PROC MCMC. During exercises, students are encouraged, also through

collaborative learning, to develop reproducible documents concerning critical comments on the results of the analyses.

Prerequisites

The student is encouraged to know the content of the following courses: Statistics, Probability, and Statistical Inference and Statistical Models II.

Teaching methods

Classroom lectures consists of a theoretical part and applications carried out using R and SAS software. Many practical examples based on real and simulated are proposed, enabling students to learn data analysis and Bayesian modeling using R through the RMarkdown interface and SAS software. They are also encouraged to engage in collaborative learning and interact with their peers and finalize the required steps of the analysis. Weekly summarizing exercises are assigned, which involve applying the proposed models to real or simulated data. During the course, with the help of R in the RStudio environment and the RMarkdown interface, students learn to create reproducible documents. The scheduled hours of traditional teaching are 30, and those of interactive teaching are 24, including lesson concerning exercises.

Assessment methods

The following methods of verifying learning apply to both students attending and non-attending lectures in presence. The examination is written with open questions and an optional oral part is possible; there are no intermediate tests. The written exam has a duration of around an hour and half and takes place in the computer lab. During the exam, open theory questions must be answered, and exercises must be solved based on the topics covered during the course. The theory questions assess the understanding of the theoretical concepts taught during the course. The empirical analyses must be conducted using the R environment, Rstudio, RMarkdown and SAS allowing verification of the ability to understand the problem and resolve it by applying advanced statistical models to real or simulated data. Students must also elaborate on reports in which the procedure is described, and the results are illustrated. The examination is open book, and students can consult all the material as well as the R code provided during the lectures. The student passes the test with a mark of at least 18/30.

Textbooks and Reading Materials

The teaching material consists mainly of handouts prepared by the teacher. These cover theory, applications, exercise and solutions developed with R software. All the files are available on the course page of the university's e-learning platform. In addition, the teacher publishes the following material at the end of each lesson: slides, R and SAS code, exercises, datasets, and solutions to some of the exercises. Previous exam texts are also published on the same page.

The main references are listed in the bibliography of the handouts, some of which are the following and are available in the university library, also in ebook format:

Albert, J. (2009). Bayesian computation with R. Springer Science & Business Media.

Albert, J., Hu, J. (2019). Probability and Bayesian modeling. Chapman and Hall/CRC.

Bartolucci, F., Farcomeni, A., Pennoni, F. (2013). Latent Markov Models for longitudinal data, Chapman and Hall/CRC, Boca Raton.

Migon, H. S., Gamerman, D., Louzada, F. (2014). Statistical inference: an integrated approach. Chapman & Hall.

Pennoni, F. (2024). Dispensa di Inferenza Bayesiana: Teoria e Applicazioni con R e SAS. Dipartimento di Statistica e Metodi Quantitativi, Università degli Studi di Milano-Bicocca.

Robert, C., Casella, G. (2004). Monte Carlo Statistical Methods (second edition). Springer–Verlag, New York. Dipak, D. K., Ghosh, S. K., Mallick, B. K. (2000). Generalized linear models: A Bayesian perspective. CRC press.

SAS/STAT PROC MCMC, User's guide, SAS Institute, 2012.

R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Semester

Semester I, cycle II, November 2024-January 2025

Teaching language

The course is delivered in Italian. Erasmus students may use the teaching material in English and request the teacher to conduct the examination in English.

Sustainable Development Goals

GOOD HEALTH AND WELL-BEING