UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

COURSE SYLLABUS

Bayesian Inference

2223-2-F8203B042-F8203B042M

Learning objectives

The course provides knowledge of the basic and advanced statistical principles under the Bayesian paradigm. The Bayesian methods are illustrated according to an integrated approach with classical statistical inference.

Knowledge and understanding

The course allows the student to learn:

- the Bayes' rule and the use of probability to update the information provided from the observed data;
- Bayesian inferential methods: priors,
  computation of the likelihood and posterior distributions;
- Monte Carlo methods to simulate the posterior distribution;
- calculus of the predictive distribution for features predictions and model checking;
- Markov Chain Monte Carlo algorithms: Metropolis-Hastings and Gibbs sampler;
- Bayesian inference and prediction for the multiple linear and logistic regression models;
- latent Markov models for the analysis of longitudinal data.

Ability to apply knowledge and understanding

The course allows the student:

- to apply the Bayesian methodology by using data available in the field of biostatistics, epidemiology, medicine, biology, environmental impact, genetics, and public health;
- to apply suitable statistical models when repeated observations for the same unit are available over time;
- to apply model-based classification methods;
- to implement suitable R and SAS code to carry out the analyses;
- to provide reproducible documents with code, results, and comments.

The theory is supported by practical applications based on real and simulated data. R is used within the Rstudio

interface, and Rmarkdown, SAS is also proposed so that the student gets a deep knowledge of two program languages.
This course provides the main concepts of Bayesian inference, an essential statistical method in the theoretical and data analysis fields for the job contexts (biostatistics/statistics/demography and related) of students in the Biostatistics degree program. The course is found to be essential for the subsequent courses.

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-Dirichelet.
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
introduction to the latent Markov models with covariates.
Introduction to the Markov Chain Monte Carlo Methods: Metropolis-Hastings algorithm and Gibbs sampler.
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 along with the Bayes‘ rule and the total probability rule. A short introduction to the Bayesian non-parametric methods is provided and the notions of exchangeability and De Finetti’s theorem are explained. The beta-binomial model is introduced along with the other conjugate families Gaussian, Poisson-gamma, beta-binomial, and multinomial-Dirichlet distributions. Choice of the prior distribution is considered. Inference is compared with that of the classical approach. Methods to draw conclusions from the posterior distribution: Bayesian interval estimation, credible intervals, and intervals with the highest posterior density. The prediction context is also explored along with the empirical Bayes estimation.
Theory is supported by several examples of the application of Bayesian models in biostatistics through real and simulated data concerning epidemiology, drug epidemiology, medicine and biology, ecology and environmental sciences.
An introduction to the stochastic processes within the Markov random field is proposed. Properties and features of the Markov chains are illustrated and explained with the use of simulations. The random walk process is also described.
Markov chain models for longitudinal data are explained, and the latent Markov models with covariates are introduced both from a theoretical and applied perspective.
Markov Chain Monte Carlo (MCMC) algorithms are provided 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, RMarkdown to make reproducible documents. The SAS software is proposed to perform the analyses to estimate Bayesian linear and logistic models with PROC MCMC. During the exercises, the student is encouraged, also through cooperative learning, to develop reproducible documents also 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

The lectures are held in the lab since the theoretical part is placed side-by-side with the computer's applications using R and SAS software. Many practical examples based on real and simulated data referred to different contexts are proposed to the students so that they can learn to analyze data and estimate Bayesian models with R through the RMarkdown interface and SAS softwares. The student is also encouraged to develop cooperative learning to interact with each other and finalize the required steps of the analysis. Exercises are carried out to report in a written form the results by adding critical comments and creating reproducible documents.

Assessment methods

The following methods of verifying learning apply to both students attending and non-attending lectures in presence. The examination is in written form with open questions and with optional oral; there are no intermediate tests. The written exam has a maximum total duration of two hours and takes place in the computer lab. During the examination, open theory questions must be answered, and exercises must be solved in the light of the theoretical topics developed during the course. The theory questions allow verifying the learning of the theoretical concepts taught during the course. The empirical analyses are conducted using the R environment, Rstudio, and RMarkdown and SAS and allow verifying the ability to understand the problem and its resolution by applying advanced statistical models to real or simulated data and the elaboration of reports in which the procedure is described, and the results are illustrated. The examination is open book and students can consult the R code used during the course. The student passes the test with a mark of at least 18/30.

Textbooks and Reading Materials

The main teaching material consists of theory and application handouts prepared by the teacher. They are made available on the page of the e-learning platform of the university dedicated to the course at the end of each lecture. The teaching material includes the slides, the calculation programs, the exercises, the datasets, the solutions of the exercises, and some examination tests referred to previous examinations.

The primary reference texts are listed in the bibliography of the handouts; among others, the following are noted. Some of these are also available in ebook.

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. (2022). 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 (2022). 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 2022-January 2023

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