Politecnico di Milano – Scuola di Ingegneria Industriale e dell'Informazione
Academic Year 2020/2021 - FIRST semester
Course code 052499 - BAYESIAN STATISTICS - 10 ECTS credits
Laurea Magistrale di ING MATEMATICA
Master of Science in Mathematical Engineering - LEONARDO Campus

Instructor

Professor Alessandra Guglielmi *e-mail*: alessandra.guglielmi@polimi.it
Dip. di Matematica, Edificio La Nave (II floor), Leonardo Campus.
Office hours: alternatively, by appointment on MS TEAMS.

Course Webpage

The course webpage is at the Beep platform: `https://beep.metid.polimi.it/it`
Check it regularly to find course announcements, solutions of exam tests, exam results, etc.

Teaching Assistant

Dott. Riccardo Corradin *e-mail*: riccardo.corradin@unimib.it

Schedule & Location

Wednesday: 14.15 - 16.15, room 2.1.1 (ex N.1.1)
Thursday: 10.15 - 12.15 room 25.S.3 (ex D.0.4)
Friday: 10.15 -12.15 and 12.30-14.15 (TA class), both in 26.13 (ex L.26.13).

Textbooks

- Jackman S. (2009). *Bayesian analysis for the social sciences.* Wiley, New York. An e-book edition can be borrowed for one year from the Polimi e-library via Adobe Digital Editions.
- Christensen R., Johnson, W., Branscum, A. Hanson, T.E (2011). *Bayesian ideas and data analysis.* CRC Press, Boca Raton (USA). An e-book edition can be partially downloaded from the Polimi e-library.
- Müller, P., Quintana, F.A., Jara, A., Hanson, T. (2015). *Bayesian Nonparametric Data Analysis.* Springer.

Software

R, JAGS, Stan, WinBUGS/OpenBUGS.

Further references

1. Gelman, A., Carlin, J. B., Rubin, D. B., Vehtari, A., Dunson, D. B., Stern, H. S. (2013). *Bayesian data analysis, Third Edition.* CRC Press.

2. Hoff, P.(2009). *A first course in Bayesian Statistical Methods.* Springer, New York.

3. Lunn D., Jackson C., Best N., Thomas A., Spiegelhalter D. (2013). *The BUGS book.* CRC Press.

4. Robert C. (2007). *The Bayesian Choice*, Second Edition. Springer.

Course Content

1. **Basics of Bayesian inference:** Likelihood principle, prior and posterior distribution. Bayes' Theorem for dominated models. Posterior summary values. Interpretation of scientific inference via the Bayesian approach. Simple univariate Bayesian models. The three main inferential problem: point estimation, hypothesis testing, interval estimation: comparison between the frequentist and the Bayesian approach. Prior distributions. The choice of a prior distribution, noninformative priors, conjugate priors and their mixtures, semi-conjugate priors. Robustness. Exchangeability and de Finetti's representation theorem for exchangeable sequences. Implications of de Finetti's theorem on the Bayesian approach. Predictive inference. Asymptotic results on the posterior distribution.

2. **Simulation methods for Bayesian Statistics:** Some results on the theory on general state space Markov chains. Markov chain Monte Carlo methods. Gibbs sampler and Metropolis-Hastings algorithms for computing posterior inference. Implementing Markov chain Monte Carlo: software for Markov chain Monte Carlo (BUGS/JAGS from R), assessing convergence and run-length.

3. **Bayesian hierarchical models:** Construction of hierarchical priors. Model fitting, predictions and sensitivity analysis.

4. **Goodness-of-fit and model choice:** AIC, BIC, DIC, Predictive Bayesian tail probabilities, log-pseudo marginal likelihood (LPML), WAIC; prior and posterior distributions for the model index, maximum a posteriori (MAP) model selection.

5. **Bayesian linear models and generalized linear models:** Hierarchical linear models with random effects. Parameter estimation and covariate selection.

6. **Bayesian survival analysis / reliability with censored data:** Regression models: accelerated failure time and proportional hazards models.

7. **Bayesian nonparametric models:** The Dirichlet process and generalizations. Bayesian nonparametric mixture models, with application to density estimation and clustering.

8. **Introduction to models for longitudinal data, time series, spatial data (point-referenced and areal data).**

Testing and Evaluation Policy

The exam consists of two parts: a written exam and an oral exam. It is compulsory to register for the written exam, which is scheduled twice in January-February, twice in July, and one in September.

The oral exam consists in the illustration of the project analyzing a statistical problem with data, using Bayesian tools and models. This evaluation policy is in agreement with the 2-credits of innovative teaching (out of 10 credits of the whole course). As a general rule, the projects will be developed in teams. The

evaluation of the projects will be based on two partial presentations given by the teams during the course (20 Nov 2020; 7 Jan 2021), both 10 minutes long, and a conclusive presentation (a whole day in 15-19 Feb 2021, last week of the break between semesters). During the conclusive presentation, each students team must hand over a project report (10-20 pages); the project illustration should be presented using slides on a laptop in not more than 15 minutes (for each team).
Participation to these three appointments is mandatory.

To obtain a positive final mark, the student should pass with a grade greater or equal than 18/30 both parts of the exam. Final grading: 50% written exam, 50% oral (project) presentation.

Lesson log

16/09/20: Course presentation and rules for evaluation.
A brief introduction to Bayesian Statistics; R example: Bayesian inference for a proportion. *Hoff: Chapter 1, slides.*
A Bayesian hierarchical model for multilevel data. R example, US math score dataset: parameters estimation. *Hoff: Chapter 8, slides.*

17/09/20: A Bayesian hierarchical model for multilevel data: R example, US math score dataset, school comparison.
Review of conditional probability and expectation; exercises.

18/09/20: The Radon-Nikodym Theorem. Dominated parametric models and likelihood. Bayes Theorem for dominated models with proof. *Notes on Beep.*
Informal definition of a conjugate prior. Example: beta-Bernoulli model: computation of posterior distribution, hyperparameters update. *Jackman: Sect 2.1.*

X **18/09/20 (TA):** Distributions of functions of a random vector. The Gamma distribution: definition of the gamma function and its properties, n-th moment and special cases (exponential and chi-squared distributions). Properties: the distribution of the sum of n gamma random variables with the same rate parameter and the independence among the sum and the normalization of two gamma random variables. Beta distribution: definition, calculation of the moment of order n, special cases. Dirichlet distribution: multivariate extension of the beta. Constructive definition via transformation of a vector of gamma r.v.s. Support of the Dirichlet distribution. Property (with proof) of neutrality and calculation of the density of the marginals. Property of aggregation. The Bernoulli, Binomial and Multinomial distributions: definitions and fundamental quantities.

23/09/20: Bayesian estimator as those minimizing the posterior expected loss function. Posterior credible regions, credible intervals, HPD credible regions. Bayesian hypothesis testing: posterior odds, Bayes factor (when the *dimension* of Θ_0 and Θ_1 is the same). *Jackman: Chapter 1.*
Predictive distributions, example. *Jackman: Chapter 1.*

25/09/20: Predictive distributions, example. *Jackman: Chapter 1.*
Example: Gaussian model for iid r.v.'s with known variance, Gaussian prior for the mean, computation of posterior and marginal densities, predictive distributions. *Jackman: Sect 2.4 e Appendix C.*

X **25/09/20 (TA):** Bayesian point estimation: definition of loss function and of Bayesian estimator. Calculation of the estimator for quadratic losses (univariate and multivariate), linear losses and 0-1 losses (discrete and continuous cases). Conjugacy of the Multinomial-Dirichlet model with an example.

30/09/20: Exchangeability of data, de Finetti's representation theorem for binary variables. Re-interpretation of the Bayesian approach. De Finetti's representation theorem for general random variables, Bayesian non-parametric approach. *Jackman: Chapter 1*; *slides.*

01/10/20: The choice of the prior. The exponential family, the conjugate distribution to the exponential family and the corresponding posterior. Posterior mean of the mean population when the likelihood belong to the natural exponential family and the prior is conjugate, as the convex linear combination of the *prior guess* and the empirical mean. Examples: Gaussian likelihood (with known variance) and Gaussian conjugate prior; Bernoulli likelihood and Beta conjugate prior. Mixtures of conjugate families are still conjugate. "Non-informative" priors: its meaning, pros and cons. The Jeffreys prior: definition, invariance under one-to-one reparameterization. *Slides.*
Details on the projects (oral exam).

07/10/20: Examples of Jeffreys' prior: Gaussian likelihood (with known variance) and and Jeffreys prior for the mean θ.
Merging of the priors. Asymptotic normality of the posterior density. *Slides.*
The inverse-gamma distribution. Bayesian model: iid Gaussian sample with unknown mean and variance, conjugate prior (normal-InvGamma). Computation of the posterior distribution. Computation of the (prior) marginal density of the mean μ. *Jackman: Sect 2.4 and Appendix C.*

X **08/10/20 (TA):** Example of Normal/Normal-Inverse-Gamma models: elicitation of the hyperparameters. Example of Beta-Bernoulli model (Jackman 2.3): conjugacy, posterior distribution, marginal distribution, choice of the hyperparameters (equivalent sample principle), predictive $n+1$, predictive $n+m$. Poisson-Gamma model: conjugacy, posterior distribution, choice of the hyperparameters, marginal for one observation, predictive distribution.

09/10/20: Unidimensional t-distributions; predictive distribution of the normal model with unknown mean μ and variance σ^2. Posterior distribution with Jeffreys' prior for μ and σ. *Jackman: Sect 2.4 and Appendix C.*
Monte Carlo methods: estimators, Monte Carlo standard errors. Review of Strong Law of Large Numbers and Central Limit Theorem. *Jackman: Chapter 3. Hoff: Chapter 4.*

X **09/10/20 (TA):** Introduction to hypothesis testing and Bayes factor. Hypothesis testing when Θ_0 has different dimension of Θ_1 (for instance, testing a punctual hypothesis $\theta = \theta_0$ versus $\theta \neq \theta_0$): mixture of a Dirac distribution and a diffuse distribution. Marginal distribution of the data when the prior is a mixture of a Dirac distribution and a diffuse distribution. Calculation of BF. Three examples of BF. Introduction to the Jeffreys' prior.

14/10/20: Monte Carlo methods: compositional method. An R example on MC methods from Hoff's textbook. *Jackman: Chapter 3. Hoff: Chapter 4.*

X **15/10/20 (TA):** Examples of Jeffreys' prior. Multivariate Normal, multivariate t-Student and Inverse-Wishart distribution (definitions). The multivariate Normal model with unknown mean and covariance. The Normal-Inverse-Wishart prior: specification, posterior distribution for the location and scale parameters $(\underline{\mu}, \Sigma)$, marginal distribution of $\underline{\mu}$ and predictive distribution. Written exam examples.

X **16/10/20 (TA):** How to generate a sample from the uniform distribution on $(0, 1)$: basic principles of the pseudo random number generators. Written exam examples.

16/10/20: Introduction to MCMC. Some theory on general state space Markov chains: irreducibility, invariant distribution, recurrence, Harris-recurrence and aperiodicity; reversibility. Law of large numbers for MCs: convergence of the ergodic means. Ergodic Theorem for aperiodic, irreducible, Harris-recurrent MCs (convergence in total variation). Geometric and uniform ergodicity; Central Limit Theorem for the MCMC error. *Slides* and *Jackman: Chapter 4.*
Accept-reject algorithm.

21/10/20: Introduction to Metropolis-Hastings algorithm: the target distribution is invariant. *Slides* and *Jackman: Chapter 5.*

R example: random walk Metropolis-Hastings algorithm to simulate from a standard Gaussian distribution. R example: random walk Metropolis-Hastings for a bivariate posterior density (from Jim Albert's textbook), output analysis as a function of the scale in the proposal density and as a function of the initial point of the MC.

22/10/20: Gibbs sampler algorithm: two-dimensional and multi-dimensional cases. *Slides, Jackman: Chapter 5.*
Monitoring the convergence and the mixing of the chain (R script).

23/10/20: Example of a Gibbs sampling for a bivariate posterior density: R script from Ch 6 in Hoff's textbook.
Homoscedastic linear model: parameter (=regression parameters + data-variance parameter) estimation under 1) the conjugate prior when the variance is known and 2) both parameters are unknown. *Jackman (2009), Sect. 2.5 and Appendix C* and *slides.*

X **23/10/20 (TA):** Method of the inverse transform for sampling from any probability distribution whose CDF is known. Examples: generation from an exponential distribution, from an absolutely continuous distribution truncated on some subset of its support, from a $Gamma(m, b)$ where m is integer and from a discrete distribution. The Box-Muller method to sample from the standard Gaussian distribution $N(0, 1)$ and extension to the multivariate case. Acceptance rejection method: basic principle and the algorithm.

28/10/20: Homoscedastic linear model: parameter (=regression parameters + data-variance parameter) estimation and predictive distributions under 3) Zellner g prior; 4) reference prior. *Jackman (2009), Sect. 2.5 and Appendix C* and *slides.*
Linear models: an example using R + JAGS (Example 2.15 from the textbook by Jackman, 2009). An introduction to the use of JAGS.

X **29/10/20 (TA):** Proof of the acceptance/rejection method. Example for simulating from a $Gamma(\alpha, 1)$, with $\alpha > 1$. How to choose the proposal in order to optimize the efficiency of the algorithm. Importance sampling method: basic principle and the algorithm. Example for evaluating the tail of a Weibull distribution. Optimal choice for the proposal distribution.

30/10/20: Credible band for a regression line (R script).
Predictive goodness-of-fit (*replicated data*): Bayesian p-values, predictive Bayesian residuals. R example from *Albert (2007)*. Predictive goodness-of-fit according to cross-validation: computation of CPOs (*conditional predictive ordinate*) and LPML (*log-pseudo marginal likelihood*). *Christensen et al.(2011), Section 4.8-4.9.*

X **30/10/20 (TA):** Metropolis-Hastings: basic principle and the algorithm. Two special cases, the independent Metropolis-Hastings and the random walk Metropolis-Hastings. Adaptive Metropolis-Hastings. Three examples: Cauchy likelihood and Gaussian priors, a mixture of two Gaussian distributions (bimodal) and the multivariate t-Student distribution (with an adaptive strategy).

X **04/11/20 (TA):** Gibbs sampler: basic principle and the algorithm. Joint distribution written in terms of conditional distributions. The Gibbs sampler as a particular case of Metropolis-Hastings. An example: changing point detection for counting data. The augmented Gibbs sampler for missing data. Example on a model with bivariate Gaussian likelihood and inverse Wishart prior on the variance covariance matrix. Hamiltonian Monte Carlo: basic principle and the algorithm.

X **05/11/20 (TA - online):** An example of Hamiltonian Monte Carlo. Introduction to STAN: basic concepts, the algorithm behind, the workflow, structure of the models, variables and primitive types, basic syntax. Two examples in R with the frogs data: a hierarchical model and a regression model. Model comparison.

12/11/20 (online): Comparison among 2 or more models: Bayes factor, BIC (*Bayesian Information Criterion*), AIC (*Aikaike Information Criterion*), DIC (*Deviance Information Criterion*). Model choice: prior and posterior distributions for the model index, maximum a posteriori (MAP) model selection. Model averaging. *Christensen et al.(2011), Section 4.8-4.9.* and *slides.*
Generalized linear models: random component, linear predictor, link function; examples: Gaussian, Poisson, Bernoulli GLMs.

13/11/20 (online): GLMs for binary responses: latent variable representation. Gibbs sampler for probit models: full-conditionals. R example (from J. Albert, Sect 10.3). *Jackman (2009), Chapter 8.*
Introduction to the use of the R package MCMCpack.

✗ **13/11/20 (TA - online):** Variable selection and shrinkage. An introduction to the model choice problem. From the model choice to the variable selection problem. The spike and slab priors: basic principle and specification. Stochastic search variable selection. Criteria for selecting covariates: HPD, MPM, HS. Brief introduction to regularization methods. The bayesian lasso and the bayesian elastic net. Example in R: covariate selection with the REACH data.

18/11/20 (online): Introduction to ANOVA models. Linear (and Generalized) mixed effects models; *fixed* and *random-effects parameters. Jackman (2009), Ch. 7* and *slides.*
R example on linear models with mixed effects: *Hoff (2009), Ch. 11*

✗ **19/11/20 (TA - online):** GLMs in Bayesian statistics. Differences between mixed models and mixture models, with the notion of partial exchangeability. Introduction to GLMMs: specification of the model and identifiability issues. Two examples: random effect for the intercept (red grouses data), and random effects for intercept and regression coefficients (ants data). Examples of models for time series data.

20/11/20 (online) 3h : First meeting on the project

20/11/20 (TA - online) 3h: First meeting on the project

25/11/20 (online): Models for longitudinal data as GLMMs. *Slides.*
Censored data: *left-, rigth-, interval-censored.* Likelihood for (right) censored data. Parametric distributions for lifetimes and "typical" priors: exponential, Weibull, log-normal, gamma. *Christensen et al. (2011), Ch 12.*

26/11/20 (online): Regression models in survival analysis: AFT model with Gaussian error, logistic or Gumbel error; survival function and median lifetimes for Gumbel errors, i.e. Weibull-distributed lifetimes *Christensen et al. (2011), Ch 13.*
AFT-Weibull model: a example (larynx cancer data) with JAGS and rjags with right censored data.

27/11/20 (online): AFT-Weibull model: a example (larynx cancer data) with JAGS and rjags with right censored data. Introduction to Cox's Proportional Hazards models.
CAR distributions for spatial areal data. GLMM with CAR priors for the random effects. *Slides*

✗ **27/11/20 (TA - online):** Survival models: basic concepts and main quantities. Censored data, right censored data and randomly censored data. Different parametric models. Likelihood for censored data. Example without covariates: exponential model. Example with covariates: Weibull model with fixed effect and random intercepts.

02/12/20 (online): R examples with the package CARbayes.
Introduction to Bayesian nonparametric statistics. Connection to exchangeability. Definition and existence of a Dirichlet process, the meaning of the measure parameter. *Notes, Phadia (2013).*

03/12/20 (online): Dirichlet processes: conjugacy property (posterior mean parameter as a convex linear combination of the prior mean and the empirical measure), stick-breaking construction. Dirichlet measure selects discrete trajectories, but it has *full support.* Weak convergence of Dirichlet processes as a function of the total mass parameter. *Notes, Müller et al.(2015)* .

04/12/20 (online): R example: simulation of trajectories.
Joint marginal distribution of a sample from the Dirichlet process (generalized Pólya urn) *Ghosal-van der Vaart (2017), Müller* et al. *(2015).* A sample from DP yields a prior on the partition of the *sample label set* (random partition); this ia a BNP model for *data clustering*; distribution of the number of unique values in a sample from a Dirichlet process. R example.

X **04/12/20 (TA - online):** Pólya urn model with two colors: application of the de Finetti representation theorem to define the underlying bayesian model (beta-bernoulli). Dirichlet process: an introduction, moments, posterior distribution, self-similarity. Chinese restaurant process. Pólya Urn scheme, connection with the CRP. Stick-breaking representation. Sampling from the posterior distribution: a sampling strategy based on the conditional law of each observation, and a sampling strategy based on the truncation of the stick-breaking representation. Example in R.

09/12/20: Density estimation with Dirichlet Process Mixture (DPM) models. *Ghosal-van der Vaart (2017), Müller et al. (2015).*
Model-based clustering via BNP: BNP mixture models, GLMM models with random effects which are a sample prom a discrete r.p.m. with an example in JAGS. *Slides.*

X **10/12/20 (TA - online):** Partitions, random partitions and exchangeable partitions. The EPPF and its properties. EPPS induced by a Dirichlet process. Mixture models: from finite to infinite number of components and the relationship with model based clustering, definition and main quantities. Dirichlet process mixture models. Two sampling strategies for Dirichlet process mixtures: marginal approach and truncated stick-breaking sampler. Example in R.

X **11/12/20 (TA - online - 2h):** Random partition estimation: basic idea, general framework. 0-1 loss function. Binder loss function, the estimation of the partition that minimize the Binder loss function as the one minimizing the distance to the posterior similarity matrix. Example (simple case) and an example in R.

16/12/20: *innovative teaching*: three short seminars on Bayesian nonparametric models by Andrea Cremaschi, Francesco Denti, Riccardo Peli

- Bayesian learning: confronto tra media(θ) e moda(θ) in posterior & prior (p. 4)
- Shrinkage effect?: confronto tra frequentist e bayesian approach (p. 9)
- Hierarchical models?