Politecnico di Milano – Scuola di Ingegneria Industriale e dell'Informazione

Academic Year 2023/2024 - FIRST semester

Course code 052499 - BAYESIAN STATISTICS - 10 ECTS credits

Master of Science in Mathematical Engineering - LEONARDO Campus

The project !

In order to implement the team (6 students) project, the students may use all the datasets and models that they like, provided that the project will be sound from a statistical and Bayesian point of view. The students private initiative in finding the dataset and the statistical problem to work on is strongly encouraged; however here there is a list of research project proposals.

List of available projects with main tutors

In blue the name of the students (in team) working on the project
Tutors in red

R1 **Learning block structures in Gaussian graphical models for spectrometric data analysis**

Goal: Probabilistic graphical modeling is a powerful tool for capturing conditional dependencies among normally distributed variables. Each node in the graph represents a variable, and the absence of an edge between nodes implies conditional independence given all others.

In previous works, we presented an application to spectrometric data to investigate relationships among the substances within a compound by observing its spectrum. The goal was achieved by coupling smoothing techniques with a Gaussian graphical model on basis expansion coefficients, hence simultaneously smoothing the data and providing an estimate of their conditional independence structure. As is common in many real world applications, evidence showed that the adjacency matrix that describes the underlying graph has a block structure, i.e., can be divided into blocks where inter-blocks dependence is much weaker than intra-block dependence. This would be equivalent to cluster the variables into disjoint groups. Nevertheless, in previous works such block structure has been either neglected or assumed as known, whereas our objective is to learn it directly from the data.

With this goal in mind, in last year's project we successfully proposed a new prior for Gaussian graphical model, enabling the learning of the underlying clustering. The goal of this project is to combine this new prior within our model for spectrometric data analysis.

Tutors: Alessandro Colombi (UniMiB)

R2 **Comparing priors over binary matrices within latent feature models framework**

Goal: Unsupervised learning seeks to uncover the underlying (latent) structure responsible for generating observed data. In the popular mixture models, each data point is assigned to a latent class, which is associated with a distribution over observable properties. In contrast, latent feature models represent each object as having multiple features. Specifically, each object can be represented as a binary vector, with entries indicating the presence or absence of each feature, and the assumption is that each feature contributes, via its associated weight, to the generation of the data point. Various probabilistic models for binary vectors have been discussed in the literature, and these can be combined with a prior on feature weights to produce continuous representations. This project mainly focuses on

the comparison between two alternative priors for binary vectors: the parametric Beta-Bernoulli and the non-parametric Beta Process. The Beta Process can be defined in terms of a stochastic process known as the Indian Buffet Process, by analogy to the Chinese Restaurant Process used in Dirichlet process mixture models. Within the framework of simple linear-Gaussian latent feature models, the goal is to investigate the impact of these two distinct priors on the produced inferences, discussing the limitations possibly experienced by the Beta-Bernoulli. The computational efficiency of the two MCMC algorithms might also be examined.

Software: C/C++.

Tutor: Lorenzo Ghilotti (UniMiB)

R3 **Integrating Bayesian Optimization and Barrier Methods in Python**

Bayesian Optimization (BO) is a category of model-based iterative algorithms for minimizing any generic function. It does not require derivative information, nor any other major assumption on the target function. At each round, it maximizes a utility function (called the "acquisition function"), instead of attempting to optimize the target function itself. For these reasons, BO is often used for optimizing black-box functions, that is, ones for which the closed-form expression is not available. Moreover, it is a sample-efficient technique, being able to reach convergence after a small amount of function evaluations. BO can also be extended to several other classes of problems, such as constrained optimization scenarios.

Goal: to write a Python code for a new acquisition function, the "expected barrier", for constrained BO. This function aims to combine locally efficient numerical methods with globally efficient statistical methods, and uses barrier functions from numerical programming.
Expertise in Python coding is required.

Main references

Frazier (2018). A Tutorial on Bayesian Optimization. arXiv:1807.02811

Brochu, Cora, De Freitas (2010). A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning. arXiv:1012.2599

Pourmohamad, Lee (2022). Bayesian Optimization via Barrier Functions. JCGS

Tutor: Bruno Guindani (DEIB, Polimi)

R4 **Mixture models with contaminated data**

Species sampling models have become relevant in many research fields over the past decades. Remarkable examples are observational ecological studies, in which the object of interest is the observed species, microbiological studies, where one might be interested in describing the composition of the bacteria population in a given subject or spot, and genetic studies, in whose the interest is on the composition of genes. Many models have been proposed in the literature to describe a sample of different species, but they generally lack in terms of flexibility when sample contamination occurs. These contaminations should be understood as a miss-reported specie or a rare gene mutation. The project aims to study state-of-the-art models to account for sample contaminations within a species framework and to compare different specifications of the models with and without contaminations.

Expertise in R or Python coding is required.

Tutor: Riccardo Corradin (University of Nottingham, UK)

R5 **Bayesian mixed effect models for functional data with wearable applications**

In this project, we will study different formalisms for specifying and inferring Bayesian multilevel functional principal component analysis decomposition and multilevel functional regression techniques. The practical motivation we consider is functional data arising from wearable devices such as Google Watch which are used to monitor Parkinson's disease patients in free-living over prolonged periods of time. The multilevel functional principal component analysis can be used to derive a decomposition of the observed data into within- and between-subject variation. The project will involve the exploration of different mixed effect constructions for capturing the joint clustering structure across individuals and subjects via adopting a hierarchical Dirichlet process prior to extending the vanilla formulation of the multilevel functional principal component analysis.

Expertise in R or Python coding is required.

Tutor: Yordan Raykov & Riccardo Corradin (University of Nottingham, UK)

R6 **Bayesian models for registration of functional data**

Tutors: Raffaele Argiento, Alessia Pini (UniBG, UniCatt)

See the slides by Argiento & Pini

R7 **Bayesian inference for nested graphical models**

Graphical models provide an effective tool to investigate dependencies among variables in a multivariate setting. Typically, the underlying graphical structure is unknown; accordingly, it must be learned from the availabledata. Basic approaches to structure learning rely on the assumption of i.i.d. observations. However, this assumption can be limiting in many real scenarios since it potentially ignores possible heterogeneity in the sample induced by an underlying clustering structure of the statistical units.

Goal: In this project, we will consider a dataset of patients affected by leukemia with multivariate observations corresponding to protein expression levels. Subjects are divided into distinct groups according to known disease subtypes. The scope of the project is to set up a statistical model to infer dependence relations between variables while considering the available information relative to the leukemia subtype.

The ultimate goal will be to develop a Bayesian nonparametric nested mixture model for multivariate data. This will induce a two-level clustering structure (across subtypes and across subjects, respectively) while reliably estimating the network of dependencies across proteins within every subpopulation.

Tutors: Federico Castelletti, Francesco Denti (UniCatt)

R8 **Variational Inference for Dirichlet Process Mixtures and Beyond**

Posterior inference for Bayesian model is usually based on Markov chain Monte Carlo. MCMC leads to asymptotically exact answers, but for a finite running time the results can be sub-optimal. Inefficiency caused by poor choice of proposal distributions, strange geometry of high-dimensional posteriors, etc... Instead, Variational Inference seeks an approximate answer very fast. The main idea of VI is to find q^* realizing

$$\arg\min_{q \in \mathcal{Q}} D(q, \pi(\theta \mid Y))$$

where $\mathcal{Q}$ is a class of simple distribution and $D(\cdot, \cdot)$ is a suitable distance.

Goal: Start by considering Dirichlet process mixtures of Gaussian distributions and implement the algorithm of Blei and Jordan (2006) in Python and JAX. Then consider one or both of the possible extensions:

(a) Move from Dirichlet process mixtures to the general class of Normalized Completely Random Measures [More maths+research focused]

(b) Consider feature allocation models based on the Beta process prior [More computational focus]

Main References

Blei, D. M., and Jordan, M. I. (2006). Variational inference for Dirichlet process mixtures. *Bayesian Analysis*

Doshi, F. et al. (2009). Variational inference for the Indian buffet process. *Artificial Intelligence and Statistics.*

Lijoi and Prünster (2010). Models Beyond the Dirichlet Process. In *Bayesian nonparametrics*

Tutor: Mario Beraha

R9 **Ozone pollution in the Po valley**

We consider **Hourly concentrations** of ozone (O_3), recorded by ARPA Lombardia monitoring network. Data were collected from 51 stations across Lombardy during **2010–2023**. For each year, we are interested in studying only a specific period (**May–Oct**) heavily affected by O_3 pollution. For each monitoring station we want to study the trend of two main quantities:

- Number of days in each month with *at least* one hour overcoming the threshold of $180 mg/m^3$;
- Number of days in each month with *at least* one N-tuple overcoming the threshold of $120 mg/m^3$.

We want to define a **spatio–temporal Bayesian model** to describe the behavior of these values over the years. Specifically we need to:

- Include **meteorological factors** as influential covariates
- Model the **temporal trend** of the time series
- Model the **spatial correlation** between monitoring sites.

Goal: compare the two quantities of interest to identify possible discrepancies. We will use ***Stan*** for MCMC implementation.

Tutor: Michela Frigeri

Tutors of the applied projects: Alessandro Carminati and Michela Frigeri

A1 Trend levels of ozone in the Po valley (dataset of Ozone levels), models for georeferenced time series of averages (over weeks) of the ozone level (e.g. see Sahu, Gelfand, Holland (2007), JASA).

A2 Clustering weekly data of one year of PM10 (plus covariates - see AGRIMONIA project), models & R packages: drpm (Page, Quintana, Dahl (2022) "Dependent Modeling of Temporal Sequences of Random Partitions", JCGS, R-package on Github) and ppmSuite (various models implemented, also PPMs) on `https://cran.r-project.org/web/packages/ppmSuite/index.html`

A3 Clustering weakly data of one year of PM2.5 (plus covariates - see AGRIMONIA project), models & R packages: drpm (Page, Quintana, Dahl (2022) "Dependent Modeling of Temporal Sequences of Random Partitions", JCGS, R-package on Github) and ppmSuite (various models implemented, also PPMs) on `https://cran.r-project.org/web/packages/ppmSuite/index.html`

A4 Clustering hourly data of one-two weeks of ammonia (plus covariates, also meteo - dataset from ARPA), models & R packages: drpm (Page, Quintana, Dahl (2022) "Dependent Modeling of Temporal Sequences of Random Partitions", JCGS, R-package on Github), and ...

A5 NOx data in Lombardy (ARPA data), georeferenced time series 2016-2023, montly averages. Two suspected change points (Sept 2019, Jan 2021). Causual inference to understand if there is an impact. `http://google.github.io/CausalImpact/CausalImpact.html`, Brodersen et al (2015) AOAS

A6 Levoglucosano and other hydrocarbons into two sites (Milano and Schivenoglia): muliple time series of various hydrocarbons in each site, clustering of the hydrocarbons in each site, time series models and clustering of the pollutants through the clustering of the parameters.

A7 ACI vehicle fleet, annual data from 2002-2022, Change points analysis for multiple time series (one time series for each province of Lombardia), then causal inference to understand if the estimated change point had an effect. R package bcp `https://cran.r-project.org/web/packages/bcp/bcp.pdf`

Dataset not user-friendly

A8 **BART in action: applications with the BART `R` package**

Bayesian Additive Regression Trees (BART) is a nonparametric Bayesian regression model that has gained widespread popularity in recent years. It approximates the unknown regression function with a sum of decision trees. A regularization prior avoids overfitting by constraining the trees to be weak learners that explain only part of the result. The BART `R` package by Sparparani *et al.* (2021) implements this model for continuous, binary, categorical, and time-to-event outcomes.

Goal: apply each model presented in Sparparani et al. (2021) to a suitable real-world application, using the BART R package. Group members are supposed to find the proper dataset by themselves.

Main References

Sparapani, R., Spanbauer, C., and McCulloch, R. (2021). Nonparametric Machine Learning and Efficient Computation with Bayesian Additive Regression Trees: The BART R Package. *Journal of Statistical Software*, 97(1), 1-66. `https://doi.org/10.18637/jss.v097.i01`

Sparapani, R. A., Logan, B. R., McCulloch, R.E., Laud, P.W. (2016). Nonparametric Survival Analysis Using Bayesian Additive Regression Trees (BART). *Statistics in Medicine*, 35(16), 2741-2753. `https://doi.org/10.1002/sim.6893`

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *Annals of Applied Statistics*, 4(1), 266-298. `https://doi.org/10.1214/09-AOAS285`

A9 **Bayesian Additive Regression Trees in spatial data analysis**

Bayesian Additive Regression Trees (BART) is a nonparametric Bayesian regression model that has gained widespread popularity in recent years. It approximates the unknown regression function with a sum of decision trees. A regularization prior avoids overfitting by constraining the trees to be weak learners that explain only part of the result. Kim (2022) presents a novel version of BART for spatial data analysis tailored to a setting with sparse spatial observations. It also provides the `R` code for the execution of the MCMC algorithm to sample from the posterior distribution of this model.

Goal: apply the model in Kim (2022), using the R code there for MCMC, for the analysis and prediction of air pollution data in Lombardy (AGRIMONIA data).

Main References

Kim, C. (2022). Bayesian additive regression trees in spatial data analysis with sparse observations. *Journal of Statistical Computation and Simulation*, 92(15), 3275-3300. https://doi.org/10.1080/00949655.2022.2102633

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *Annals of Applied Statistics*, 4(1), 266-298. https://doi.org/10.1214/09-AOAS285

A10 Vehicles entering the gates for area C in Milano, data at https://dati.comune.milano.it/: according to the available data, fit spatio-temporal regression models to this dataset.

List of useful sites

Here it is a list of sites where you can find datasets:

Our World in Data: https://ourworldindata.org/

UK Data Service: https://ukdataservice.ac.uk/

U.S. Government's open data: https://data.gov

Kaggle: http://www.kaggle.com

AIRBNB: http://insideairbnb.com/get-the-data.html

UCI machine learning: http://archive.ics.uci.edu/ml/datasets.html

Immigration in the US: see the report here: https://www.stoltzmaniac.com/us-immigration-enforcement-part-1/?utm_campaign=github_readme, where you can find a link to the dataset

European Institute for Gender Equality: https://eige.europa.eu/

US Census Bureau: https://www.census.gov/programs-surveys/acs/data/experimental-data/2020-1-year-pums.html

Comune di Milano: http://dati.comune.milano.it

Regione Lombardia: https://www.dati.lombardia.it

ISTAT: http://www.istat.it/it/prodotti/banche-dati

EUROSTAT: https://ec.europa.eu/eurostat/databrowser/explore/all/all_themes?lang=en&subtheme=migr&display=list&sort=category

NASA: http://www.nasa.gov/open/data.html

Tutor

Matteo Gianella (Polimi): matteo.gianella@polimi.it

People

Bruno Guindani (Polimi): bruno.guindani@polimi.it

Mario Beraha (UniTO): mario.beraha@unito.it

Raffaele Argiento (UniBG): raffaele.argiento@unibg.it

Alessandro Colombi(UniMiB): a.colombi10@campus.unimib.it

Lorenzo Ghilotti(UniMiB): l.ghilotti@campus.unimib.it

Francesco Denti (UniCatt): francesco.denti@unicatt.it

Riccardo Corradin (UNottingham): riccardo.corradin@nottingham.ac.uk

Yordan Raykov (UNottingham): yordan.raykov@nottingham.ac.uk

Michela Frigeri (Polimi): michela.frigeri@polimi.it

Alessandro Carminati (Polimi): alessandro.carminati@polimi.it