The Project - Bayesian Statistics

Alessandra Guglielmi

in collaboration with colleagues

29 September 2023

POLITECNICO
MILANO 1863

The project - *oral* exam

Final grading: 40% written test, 60% project

Oral exam: presentation of the work done within a *data analysis project* via the Bayesian approach (Bayesian models and tools)

The project - *oral* exam

Final grading: 40% written test, 60% project

Oral exam: presentation of the work done within a *data analysis project* via the Bayesian approach (Bayesian models and tools)

- **Team project: 6 students**
  I see now 126 enrolled students, but I don't expect much more than 100 students to *actively* participate

at most 20 groups

The project - *oral* exam

Final grading: 40% written test, 60% project

Oral exam: presentation of the work done within a *data analysis project* via the Bayesian approach (Bayesian models and tools)

- **Team project: 6 students**
  I see now 126 enrolled students, but I don't expect much more than 100 students to *actively* participate

  at most 20 groups

- **One intermediate revision + one conclusive presentation**:
  - 23 and 24 Nov 2023 (8-9 ? groups on 23 nov), 10 minutes long (+ discussion/comments by tutors, TA and the instructor)
  - one single day between 12 and 16 Feb 2024 - morning & afternoon, last week of the winter break, 15 minutes for each team

The project

- **Report** (10-20 pages, with an Appendix with further plots) the same day of the final presentation; write the report with great accuracy, I will check on it if needed!

The project

- **Report** (10-20 pages, with an Appendix with further plots) the same day of the final presentation; write the report with great accuracy, I will check on it if needed!
- **Code produced for the project should be uploaded on a GitHub repository** associated to the project before the final presentation; TA and tutors will give the precise deadline

The project

- **Report** (10-20 pages, with an Appendix with further plots) the same day of the final presentation; write the report with great accuracy, I will check on it if needed!
- **Code produced for the project should be uploaded on a GitHub repository** associated to the project before the final presentation; TA and tutors will give the precise deadline
- **Presentation in English, slides in Latex** - both presentations must be uploaded on WeBeep the day before

The project

- the grade achieved for the project is valid for the whole academic year

The project

- the grade achieved for the project is valid for the whole academic year
- **It is a project!**: you find (or look at the list of available projects) a statistical problem and *solve* it using Bayesian models and tools. As such there is *no unique solution*, as it would be the case for a work project. However the focus is not resolving the statistical problem as it is, but solving it using Bayesian statistics. The project should not be too easy or too difficult!

The revision

- Consider the intermediate revision as a *revision*. I will only evaluate your commitment/engagement in the design of the project.

The revision

- Consider the intermediate revision as a *revision*. I will only evaluate your commitment/engagement in the design of the project.
- Creativity and independence (autonomy) are encouraged and increase the value of the project! However take seriously TA's, tutors' and my advice during the revisions

The revision

- Consider the intermediate revision as a *revision*. I will only evaluate your commitment/engagement in the design of the project.
- Creativity and independence (autonomy) are encouraged and increase the value of the project! However take seriously TA's, tutors' and my advice during the revisions
-

The revision

- Consider the intermediate revision as a *revision*. I will only evaluate your commitment/engagement in the design of the project.
- Creativity and independence (autonomy) are encouraged and increase the value of the project! However take seriously TA's, tutors' and my advice during the revisions
-
- All group members have to illustrate part of the presentation at least once. At least presence of the (3) members who present the project is mandatory.

POLITECNICO
MILANO 1863

The revision

- Consider the intermediate revision as a *revision*. I will only evaluate your commitment/engagement in the design of the project.
- Creativity and independence (autonomy) are encouraged and increase the value of the project! However take seriously TA's, tutors' and my advice during the revisions
-
- All group members have to illustrate part of the presentation at least once. At least presence of the (3) members who present the project is mandatory.
- Remember that you can do an individual project and make a presentation to me just after you pass the written exam.

The revision

- Consider the intermediate revision as a *revision*. I will only evaluate your commitment/engagement in the design of the project.
- Creativity and independence (autonomy) are encouraged and increase the value of the project! However take seriously TA's, tutors' and my advice during the revisions
-
- All group members have to illustrate part of the presentation at least once. At least presence of the (3) members who present the project is mandatory.
- Remember that you can do an individual project and make a presentation to me just after you pass the written exam.
- The night before each revision, including the final presentation, the teams upload their presentation pdf files (all_lastnames.pdf) in a homework directory in WeBeep

Score of the project - innovative evaluation policy

The score of the team projects I give is typically the same for all team members, unless the *research tutors* suggest other shares

I will allow small changes in the individual score of the project (up to +2 points), if the team asks for them, given that the total team score remains fixed!

Ex: the team agrees that member X deserves $z + 2$, where z is the score assigned by me to each member of the team, and allows member X:

- to borrow 2 points from member Y, whose score is updated to $z - 2$
- or to borrow 1 points from member Y, whose score is updated to $z - 1$, and 1 points from member T, whose score is updated to $z - 1$

Looking for a capo-classe

This is a call for two capo-classe (two class presidents) students who will help us to manage the *demand for projects*, sharing with me an Excel file (on OneDrive) and a latex file (on Overleaf)

Capo-classe: Luca Maci, `luca.maci@mail.polimi.it`
Deputy capo-classe: ?

Research & applied projects I

We offer

- a limited number n ($n < 20$) of *research projects*, i.e., with a tutor who guides the team.
- $20(?) - n$ *applied projects*.

I will update the list in a pdf (on WeBeep) next week. The list might broaden until the end of October.

Procedure to choose the project in the list:

- Each group expresses **at most 3 preferences** from the list of projects (of any type) on a MicroSoft Form which will be shared with all students. Deadline to fill the form: 24th October.
  The form will be make available only a few days before the deadline, but after the 13th October. As soon as the form is available to students, I will put an "announcement" on WeBeep.
- Within 3 days, tutors make a decision (if needed) on the group assigned to the project.

Research & applied projects II

- The MS Form should be filled with this information (provided by the group): scores of all group members for the following exams: ARF (Real and Complex Analysis), ACP (ALGORITHMS AND PARALLEL COMPUTING), Applied Statistics, PACS (ADVANCED PROGRAMMING FOR SCIENTIFIC COMPUTING) yes/no, Erasmus (ingoing or outgoing) student status yes/no.
- Tutors may ask to meet the groups (on video) before making a decision.

POLITECNICO
MILANO 1863

Research & applied projects

- The intermediate revision is mainly meant to check if all the groups have been assigned to a project and if they all have enough clear in mind how to proceed.
- No problem in changing the project after the first revision.
- Tutors for the *applied projects*: Alessandro Carminati and Michela Frigeri

Research projects have typically clear research questions to be answered. They are extremely demanding in terms of work.

Applied projects imply more standard statistical questions, but posterior inference obtained must me interpreted. They are described by a dataset, some statistical questions, and a paper to serve as a guideline for the development of the project itself. They are indeed demanding in terms of work.

The choice of the project

After you have chosen the project, by Sunday 5 November, send to Luca Maci, the class president student:

- Title of the project and number in the pdf list of *research and applied projects* if applies
- Names and e-mail of each member of the team

The *class president* will share a **Latex file** containing

- Title of the project, with the corresponding numbering in the list (if the project is taken from the list)
- Names and e-mail of each member of the team

to me (I have a template).

Learning block structures in Gaussian graphical models for spectrometric data analysis

Tutor: Alessandro Colombi

Abstract: 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.

Comparing priors over binary matrices within latent feature models framework

Tutor: Lorenzo Ghilotti

Abstract: 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.

Expertise in C/C++ needed

Bayesian Optimization

Title: Integrating Bayesian Optimization and Barrier Methods in Python

Tutor: Bruno Guindani

Abstract: 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. The goal of this project is to write a Python implementation 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.

Bruno will give a short seminar on BO on 13th Oct in class.

POLITECNICO
MILANO 1863

Variational Inference for Dirichlet Process Mixtures and Beyond

Tutor: Mario Beraha

- 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...
- Variational Inference seeks an approximate answer very fast
- Main idea: find q^* realizing

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

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

Variational Inference for Dirichlet Process Mixtures and Beyond

Aim of the project:
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:

1. Move from Dirichlet process mixtures to the general class of Normalized Completely Random Measures [More maths+research focused]
2. 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*

Mixture models with contaminated data

Tutor: Riccardo Corradin, University of Nottingham (UK)

Abstract: Mixture models are a flexible and powerful family of models quite useful for performing density estimation and clustering. They are constructed by considering a weighted average of different distributions, belonging to the same family, where the components differ from each other by their values of the indexing parameters. In the last decades, this family of models has been successfully used in literature for various studies and applications. However real data are quite often subjected to the presence of outliers and contamination. Recently an extension of these models appeared in the literature, where a mixture model is contaminated with an exogenous term accounting for the presence of outliers and contaminants. The project aims to investigate the influence of the contaminants on the quality of the estimates, both for density estimation and clustering. Further, once the model has been validated, the purpose of the project is to identify contaminant observations in astronomical studies.

Email: `riccardo.corradin@nottingham.ac.uk`

Bayesian mixed effect models for functional data with wearable applications

Tutors: Yordan Raykov & Riccardo Corradin

Abstract: 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.

Email: `yordan.raykov@nottingham.ac.uk`

POLITECNICO
MILANO 1863

Bayesian inference for nested graphical models

Tutors: Francesco Denti & Federico Castelletti (UniCattolica)

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.

See the slides

Bayesian models for registration of functional data

Tutors: Raffaele Argiento & Alessia Pini

See the slides

POLITECNICO
MILANO 1863

Ozone pollution in the Po valley

Tutor: Michela Frigeri

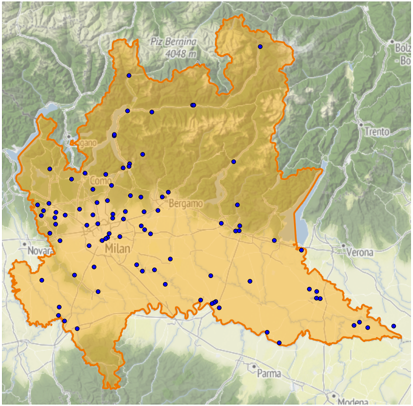


Figure: ARPA fixed monitoring stations

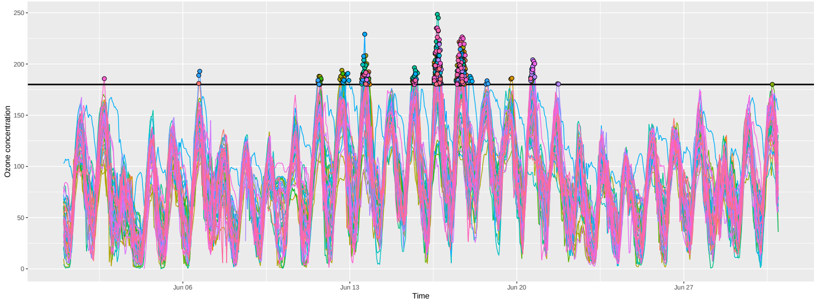
- **Hourly concentrations** of ozone (O_3) recorded by ARPA Lombardia monitoring network.
- Data collected by **51 stations** across Lombardy during **2010–2023**.
- For each year, we are interested in studying only a specific period (**May–Oct**) heavily affected by O_3 pollution.
- **Weather conditions** must be considered since they heavily affect the air pollution level.

Ozone pollution in the Po valley

For each 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 $120 mg/m^3$.

N-tuple := average on a sliding window of N consecutive hours.

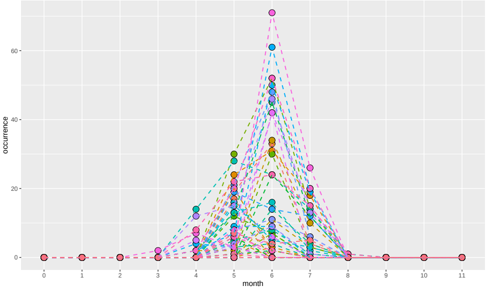


POLITECNICO
MILANO 1863

Ozone pollution in the Po valley

We want to define a **spatio–temporal Bayesian model** to catch 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.



The final goal is to **compare** the two quantities of interest to identify possible discrepancies. We will use ***Stan*** for MCMC implementation.

Applied projects I

1. 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).
2. 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`
3. 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`

POLITECNICO
MILANO 1863

Applied projects II

4. 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 ...
5. 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
6. 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.
7. 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

Applied projects III

8. BART model for ?
   Chipman, Hugh A., Edward I. George, and Robert E. McCulloch. "BART: Bayesian additive regression trees." (2010): 266-298.
9. Vehicles entering the gates for area C in Milano, data at
   `https://dati.comune.milano.it/`

Data to use for the project

Feel free to propose **your own project**, provided that your project is sound (and not too difficult or too easy) from a statistical and Bayesian point of view

Useful websites:

- Comune di Milano: `http://dati.comune.milano.it`
- Our World in Data: `https://ourworldindata.org/`
- UK Data Service: `https://ukdataservice.ac.uk/`
- U.S. Government's open data: `https://data.gov`
- 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/`

Data websites (continued)

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

Always check for confidentiality policies!

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