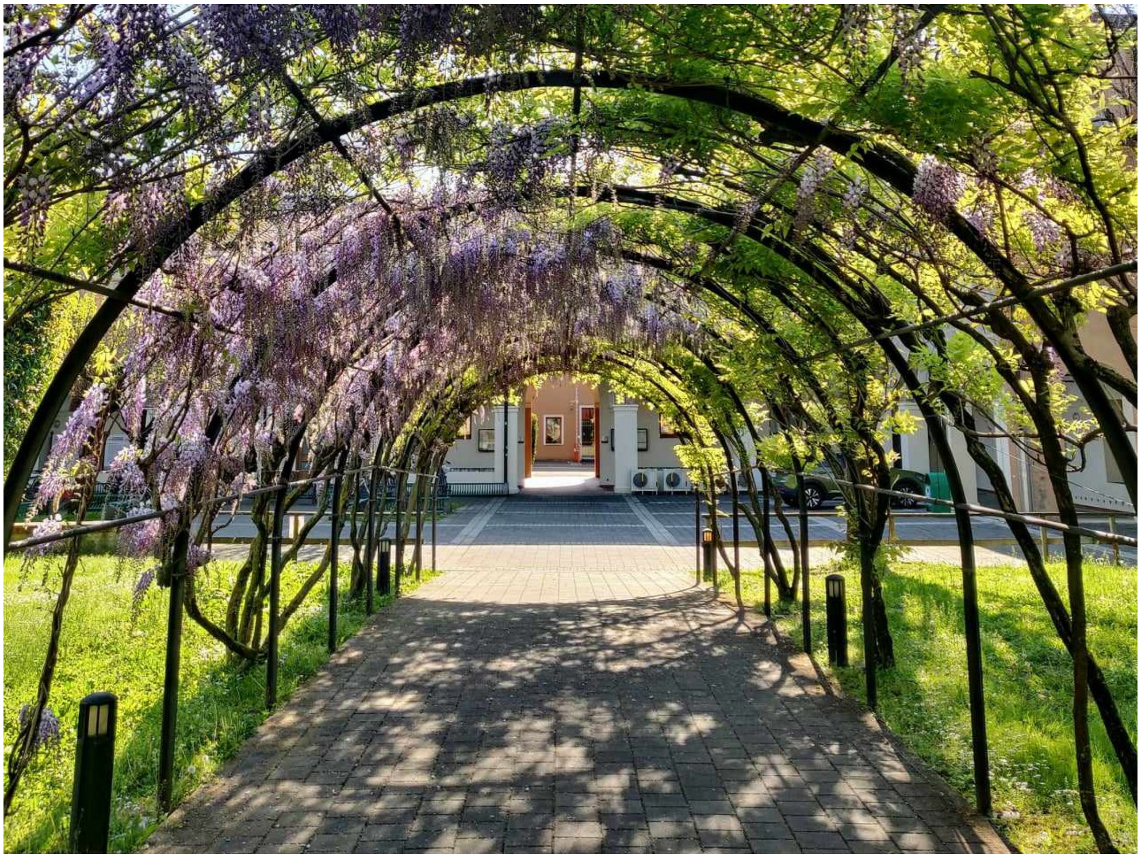

SISBAYES 2025 Workshop

Department of Statistical Sciences
University of Padova

4-5 September 2025

ABSTRACT BOOKLET

Scientific Committee

Raffaele Argiento (chair)
Isadora Antoniano-Villalobos
Veronica Ballerini
Federico Camerlenghi
Antonio Canale
Federico Castelletti
Leonardo Egidi
Brunero Liseo

Organizing Committee

Antonio Canale (chair)
Emanuele Aliverti
Francesco Denti
Stefano Rizzelli

With the support of

Università degli Studi di Padova

Dipartimento di Scienze Statistiche

Università degli Studi di Bergamo | Dipartimento di Scienze Economiche

Endorsed by

ISBA

SIS Società Italiana di Statistica

Contents

Fundational sessions

From Statistical to Causal Models

Guido Consonni
Università Cattolica del Sacro Cuore

The remarkable success of modern Machine Learning (ML) has been driven by access to large datasets, expressive models such as neural networks, and high-performance computing. At its core, this progress has relied heavily on statistical learning, which provides a coherent theoretical foundation. Many of the most celebrated achievements in ML and Artificial Intelligence (AI) are grounded in exploiting statistical dependencies under the assumption that data are independent and identically distributed. However, current challenges in AI—such as robustness to distributional shifts, transferability, and generalization—highlight the limitations of purely statistical approaches. In this context, causal models offer a principled alternative framework. In this talk, I will introduce the key ideas behind causal modeling, focusing on both graphical and structural approaches. I will explain how causal models can be used to evaluate the effects of interventions and answer counterfactual queries. I will also cover the dual challenges of causal discovery (learning causal structure from data) and causal reasoning (answering causal questions once the structure is known). Finally, I will highlight recent contributions from my own research in Bayesian causal inference.

Bayesian predictive-based uncertainty quantification

Sonia Petrone
University Bocconi

We are all familiar, at least since Breiman's provocative paper in Statistical Science (2001), with the "two cultures" - classic statistical inference versus algorithmic prediction. Bayesian statistics has prediction in its foundations and may naturally combine both cultures. In the talk we will take a Bayesian predictive approach, where one directly reasons on prediction of future observations - bypassing, in principle, models and parameters, or possibly using them implicitly. This line of research, which has a long tradition, is receiving renewed interest, also as a way to provide Bayesian uncertainty quantification for predictive algorithms - computationally convenient approximations of exact Bayesian solutions, or black-box predictive engines. I will review basic concepts and recent results, and highlight potentialities and open issues, such as calibrating the predictive rule in light of 'predictive efficiency' and good predictive and inferential properties.

Keynote sessions

A Bayesian theory for estimation of biodiversity

Tommaso Rigon
Università di Milano Bicocca

Statistical inference on biodiversity has a rich history going back to RA Fisher. An influential ecological theory suggests the existence of a fundamental biodiversity number, denoted alpha, which coincides with the precision parameter of a Dirichlet process. Motivated by this theory, we develop Bayesian nonparametric methods for statistical inference on biodiversity, building on the literature on Gibbs-type priors. We argue that sigma-diversity is the most natural extension of the fundamental biodiversity number and discuss strategies for its estimation. Furthermore, we develop novel theory and methods starting with an Aldous-Pitman process, which serves as the building block for any Gibbs-type prior with a square-root growth rate. We propose a modeling framework that accommodates the hierarchical structure of Linnean taxonomy, offering a more refined approach to quantifying biodiversity. The analysis of a large and comprehensive dataset on Amazon tree flora provides a motivating application.

Bayesian models for misreported counts data: theoretical and applied issues.

Serena Arima
University of Salento

Data quality underpins the effective use of information in all data-driven processes. A quality issue typical of surveillance, notification, and other official registers is the misreporting of events. Failing to account for missing cases may lead to severe under/overestimation. When dealing with sanitary data, missing cases may impact vital statistics, affecting the incidence and prevalence of diseases and morbidity and mortality rates in turn, thus jeopardizing the aims of a National Health System (NHS). Defective reporting may occur either because not all cases seek healthcare (under-ascertainment) or due to the inadequacy to detect symptomatic cases that have sought medical advice. We propose a nonparametric compound Poisson model for underreported count data that introduces a latent clustering structure for the reporting probabilities in observation groups. Estimates of the latter are based on model parameters exploiting experts' opinions a proxy for the reporting process. The proposed model is used to estimate the prevalence of the Chronic Kidney Disease (CKD) in Apulia, Italy, based on a unique statistical database covering information on m = 258 municipalities obtained by integrating multi-source register information. Accurate prevalence estimates are required for monitoring, surveillance, and management purposes; yet, counts are deemed to be considerably underreported, especially in some areas of Apulia, one of the most deprived and heterogeneous regions in Italy. We extend the aforementioned idea to count data that might be affected by two sources of measurement error: measurement error leading to extra counts that are not supposed to be counted (add-in or overreporting) and measurement error causing the loss of counts that should have been counted (leave-out or underreporting). We extend the compound Poisson model by incorporating both sources of error: The proposed model is applied to investigate the distribution of the counts of wildfire occurrences in the municipal areas of two neighboring Italian regions for the 2021 summer season, which may be affected by both underreporting and double-counting when multiple information sources are employed. Fire counts are obtained by processing MODIS satellite data, while several socio-economic and environmental-driven potential risk factors are also considered in the model formulation. Data from multiple sources with different spatial support are processed to comply with the municipal units. Results suggest the appropriateness of the approach and provide some insights on the features of wildfire occurrences.

Invited Session: Model Based Clustering

Model-Based Clustering of Categorical Data Based on the Hamming Distance

Lucia Paci
Università Cattolica Milano

A model-based approach is developed for clustering categorical data with no natural ordering. The proposed method exploits the Hamming distance to define a family of probability mass functions to model the data. The elements of this family are then considered as kernels of a finite mixture model with an unknown number of components. Conjugate Bayesian inference has been derived for the parameters of the Hamming distribution model. The mixture is framed in a Bayesian nonparametric setting, and a transdimensional blocked Gibbs sampler is developed to provide full Bayesian inference on the number of clusters, their structure, and the group-specific parameters, facilitating the computation with respect to customary reversible jump algorithms. The proposed model encompasses a parsimonious latent class model as a special case when the number of components is fixed. Model performances are assessed via a simulation study and reference datasets, showing improvements in clustering recovery over existing approaches.

Dependent Dirichlet-Multinomial Processes with Random Number of Components

Andrea Cremaschi
IE University

Over the past two decades, Bayesian nonparametrics has expanded to include flexible dependent prior distributions for mixture models, extending beyond univariate species sampling processes to effectively capture dependencies in grouped data under partial exchangeability. While most research has focused on nonparametric priors with almost surely infinite support points, much less attention has been given to almost surely finite-dimensional dependent mixture models under partial exchangeability, despite their strong theoretical properties and promising performance in the exchangeable case. In this work, we explore alternative strategies for defining a multivariate extension of the finite Dirichlet-Multinomial process and its counterpart incorporating a prior on the number of components. Specifically, we introduce a class of flexible dependent Dirichlet-Multinomial processes based on Generalised Wishart unnormalised weights. We analyse their theoretical properties and demonstrate that, unlike existing alternatives, the proposed prior can achieve any desired level of dependence for any fixed number of components. Additionally, our approach allows for efficient posterior computation without the need for costly variable augmentation schemes. We also illustrate the practical advantages of our model through extensive simulation studies and an application to sex-specific gene expression differentiation in the human brain, showcasing its flexibility and computational efficiency in capturing complex dependence structures.

A Bayesian nonparametric approach to discriminant analysis

Bernardo Nipoti
Università degli Studi di Milano-Bicocca

We introduce a Bayesian nonparametric framework to improve classical discriminant analysis, particularly in scenarios with sparse data. Our method provides a flexible approach that encompasses both linear and quadratic discriminant analysis as special cases. The key innovation lies in allowing information sharing across groups to improve the estimation of group-specific covariance matrices. This is accomplished through a scale-only nonparametric mixture model defined on the space of positive definite matrices. The use of a conjugate nonparametric prior ensures tractability and remarkable ease of implementation. Applications to both simulated and real datasets demonstrate the adaptability and effectiveness of the proposed methodology.

BNP4BNP: Bayesian Nonparametric Models for Biomarkers and Neuronal Patterns

Francesco Denti
Università degli Studi di Padova

This talk discusses two Bayesian nonparametric models designed to address the challenges posed by large, complex biological datasets generated by state-of-the-art imaging technologies. The first application focuses on a MALDI-mass spectrometry imaging dataset, which quantifies the abundance of numerous specific molecules across multiple locations within a biological tissue sample. We propose a model for nested, separate exchangeable data, inducing a biclustering to simultaneously group spatial locations with similar abundance profiles and molecules with similar expression patterns. A hidden Markov random field prior is incorporated to enable precise image segmentation, ensuring that clusters correspond to biologically meaningful regions. The resulting biclustering structure facilitates the detection of molecular activation patterns and provides interpretable segmentations of the analyzed image. The second model addresses calcium imaging data, which record the activity of individual neurons over time in freely moving animals. We develop a spatiotemporal mixture model to identify co-activating neurons, detecting groups of cells with a similar firing activity over time. This is achieved through a multivariate time-series framework that detects spikes in calcium traces and captures recurring temporal patterns. We employ a Dependent Dirichlet Process to incorporate information on the anatomical proximity between neurons. Simultaneously, spikes' amplitudes are segmented using a Dirichlet process, providing clusters based on their magnitudes.

Invited Session: Prior elicitation for complex problems

Objective Priors for Measures of Evidence

Laura Ventura
Università degli Studi di Padova

To test a precise (or sharp) null hypothesis on a scalar parameter of interest, Bayesian measures of evidence are the e-value and the Bayesian Discrepancy Measure. In the framework of these measures of evidence, for a parameter of interest we discuss the role of objective matching priors, which are based on higher-order asymptotic expansions. Connections of the e-value and the Bayesian Discrepancy Measure with frequentist inference are highlighted when using these objective matching priors.

Closed-Form Power and Sample Size Calculations for Bayes Factors

Samuel Pawel
University of Zurich

Determining an appropriate sample size is a critical element of study design, and the method used to determine it should be consistent with the planned analysis. When the planned analysis involves Bayes factor hypothesis testing, the sample size is usually desired to ensure a sufficiently high probability of obtaining a Bayes factor indicating compelling evidence for a hypothesis, given that the hypothesis is true. In practice, Bayes factor sample size determination is typically performed using computationally intensive Monte Carlo simulation. Here, we summarize alternative approaches that enable sample size determination without simulation. We show how, under approximate normality assumptions, sample sizes can be determined numerically, and provide the R package bfpwr for this purpose. Additionally, we identify conditions under which sample sizes can even be determined in closed-form, resulting in novel, easy-to-use formulas that also help foster intuition, enable asymptotic analysis, and can also be used for hybrid Bayesian/likelihoodist design. Furthermore, we show how power and sample size can be computed without simulation for more complex analysis priors, such as Jeffreys-Zellner-Siow priors or non-local normal moment priors. Case studies from medicine and psychology illustrate how researchers can use our methods to design informative yet cost-efficient studies. [Pawel, S., Held, L. (2025). Closed-Form Power and Sample Size Calculations for Bayes Factors. The American Statistician, ¡https://doi.org/10.1080/00031305.2025.2467919¿]

Filtering procedures for dynamic multinomial probit models

Augusto Fasano
Università degli Studi di Torino

The multinomial probit constitutes a widely-used model for categorical data in many applications, especially in the econometrics and discrete-choice literature. The computational challenges encountered when fitting this model still motivate ongoing research both from the frequentist and Bayesian view-points. In this contribution, we consider a dynamic formulation based on a state-space model where at each time one observes a sample from a multinomial probit with time-specific parameter value. Dependence across time is then induced by the Markovian dynamics of the latent parameter. We show that the filtering and predictive distribution of the latent parameter belong to the unified skew-normal family, developing an associated i..i.d. sampler to approximate quantities of interest via Monte Carlo. Motivated by the computational bottlenecks of the sampler encountered already for moderate sample sizes, we also develop approximate methods for online inference based on assumed density filtering and expectation propagation. This gives more scalable, yet accurate, algorithms for online inference about the latent state and prediction of future observations. Results are shown oversimulated data and a real dataset regarding reservations made from a list of results from an online booking platform.

Invited Session: Bayesian methods for ecological applications and beyond

Taxonomic and covariate-dependent feature allocation models

Federica Stolf
Duke University

Indian Buffet Processes (IBPs) are widely used Bayesian nonparametric models designed for binary latent feature matrices with a potentially infinite number of columns. In biodiversity studies, where features correspond to observed species, this approach is particularly powerful as it allows for the inclusion of an ever-growing number of species. However, current IBP-based models rely on unrealistic assumptions that limit their applicability to ecological data. In particular, they assume that data are exchangeable, ignoring habitat-specific characteristics. Additionally, they model species only at a single taxonomic level, discarding the full taxonomic structure of the data. To address these limitations, we propose a taxonomic and covariate-dependent extension of the IBP that leverages the hierarchical tree structure inherent in the data and incorporates heterogeneity across sampling sites. We discuss theoretical properties of the proposed modeling paradigm and implement efficient algorithms for posterior computation. Crucially, our framework also enables to clearly define the so-called beta-diversity, i.e. the taxonomic heterogeneity of species across different sampling regions, under a coherent and elegant probabilistic framework.

A Bayesian approach to capture-recapture models with misidentification

Andrea Tancredi
Sapienza Università di Roma

The absence of identification errors is a fundamental prerequisite in capture-recapture for population size estimation and species abundance estimation. However, such errors have been reported and studied in both contexts. The most common case is the failure to recognize a previously detected entity, i.e., a false negative record linkage error. This results in artificious entities, sometimes referred to as "ghosts", which typically constitute spurious singletons. We present a Bayesian parametric approach to the problem, which is applicable when data are summarized as number of captures. We develop a Markov chain Monte Carlo algorithm to estimate the proposed model and illustrate the performance of our approach on some datasets available in the microbial diversity literature.

Multivariate species sampling models

Beatrice Franzolini
Bocconi University

Species sampling models provide a structural framework for understanding random discrete distributions in the context of exchangeable observations, as they correspond to the class of exchangeable partitions. However, they do not account for more general structures, such as those that naturally arise in the context of heterogeneous data collected from related sources or under distinct experimental conditions. For instance, this is the case in problems involving the sequential sampling of species across multiple sites, where often the goal is to maximize the number of species discoveries. To address this limitation, we introduce multivariate species sampling models (mSSMs), a general class of models characterized by their partially exchangeable partition probability function. These models encompass most existing Bayesian nonparametric approaches for partially exchangeable data and help elucidate their core distributional properties and the learning mechanisms they induce. Specifically, mSSMs facilitate the study of general properties of random partitions under partial exchangeability assumptions. We demonstrate that the dependence structure is determined by the probability of ties occurring across groups, with independence across sources corresponding to a zero probability of such ties. Furthermore, mSSMs admit three equivalent representations: in terms of the law of the vector of random probability measures, the partially exchangeable random partition, or the sequence of observations. We provide three characterization theorems describing the laws governing these objects. The results presented offer a comprehensive understanding of the dependence structures induced by a wide range of nonparametric models under partial exchangeability assumptions and beyond.

Invited Session: Challenging Posteriors

Full Bayesian Reinforcement Learning via LF-IBIS

Cecilia Viscardi
Università di Salerno

Reinforcement Learning (RL) is a widely used class of methods to support decision-making. Traditional RL focuses on maximizing cumulative rewards based on the feedback an agent receives while interacting with an environment. These methods often suffer from data scarcity — i.e., a lack of knowledge about the dynamics of the environment — as data comes from costly interactions that occur gradually over time. Bayesian Reinforcement Learning (BRL) mitigates this problem by incorporating prior knowledge about the environment and updating it as data is collected. However, this approach requires explicit formalization of the environment's behaviour through a likelihood function, which is often unavailable in real-world scenarios. We propose a full Bayesian Reinforcement Learning (fBRL) strategy aimed at evaluating the posterior distributions for both environment parameters and optimal policies. Our approach is based on Likelihood-free Iterated Batch Importance Sampling (LF-IBIS), a novel algorithm that combines Approximate Bayesian Computation with Iterated Batch Importance Sampling. This hybrid sampling scheme overcomes the definition of an explicit likelihood function and allows an online updating of the posterior distributions as data coming from new interactions becomes available. Finally, we test the effectiveness of our proposal addressing the problem of Response-Adaptive Randomization in clinical trials.

Concentration of discrepancy-based approximate Bayesian computation via Rademacher complexity

Sirio Legramanti
Università d Bergamo

There has been an increasing interest on summary-free solutions for approximate Bayesian computation (ABC) that replace distances among summaries with discrepancies between the empirical distributions of the observed data and the synthetic samples generated under the proposed parameter values. The success of these strategies has motivated theoretical studies on the limiting properties of the induced posteriors. However, there is still the lack of a theoretical framework for summary-free ABC that (i) is unified, instead of discrepancy-specific, (ii) does not necessarily require to constrain the analysis to data generating processes and statistical models meeting specific regularity conditions, but rather facilitates the derivation of limiting properties that hold uniformly, and (iii) relies on verifiable assumptions that provide more explicit concentration bounds clarifying which factors govern the limiting behavior of the ABC posterior. We address this gap via a novel theoretical framework that introduces the concept of Rademacher complexity in the analysis of the limiting properties for discrepancy-based ABC posteriors, including in non-i.i.d. and misspecified settings. This yields a unified theory that relies on constructive arguments and provides more informative asymptotic results and uniform concentration bounds, even in those settings not covered by current studies. These key advancements are obtained by relating the asymptotic properties of summary-free ABC posteriors to the behavior of the Rademacher complexity associated with the chosen discrepancy within the family of integral probability semimetrics (IPS). The IPS class extends summary-based distances, and also includes the widely implemented Wasserstein distance and maximum mean discrepancy (MMD), among others. As clarified in specialized theoretical analyses of popular IPS discrepancies and via illustrative simulations, this new perspective improves the understanding of summary-free ABC.

Sampling on constrained spaces

Elena Bortolato
Universitat Pompeu Fabra

Sampling probability distributions on submanifolds is a relevant task in various problems in statistics, especially for dealing with models defined upon constraints. Specialized Markov Chain Monte Carlo (MCMC) methods were originally designed to address similar challenges in statistical physics and were recently adopted in various statistical contexts, as hypotheses testing, ABC, overparameterized models, Generalized fiducial inference. However, these MCMC methods have proven to be adaptable to a variety of sampling problems in Bayesian statistics which are not naturally defined on submanifolds and offer promising avenues for efficiently sampling probability distributions.

Invited Session: Bayesian Causal Inference

Multivariate Causal Effect: a Bayesian Regression Factor Model

Dafne Zorzetto
Brown University

The impact of wildfire smoke on air quality is a growing concern, contributing to air pollution through a complex mixture of chemical species with important implications for public health. Although previous studies have focused mainly on its association with total particulate matter ($\mathrm{PM}_{2.5}$), the causal relationship between wildfire smoke and the chemical composition of $\mathrm{PM}_{2.5}$ remains largely unexplored. Exposure to these chemical mixtures plays a critical role in shaping public health, but capturing their relationships requires advanced statistical methods capable of modeling the complex dependencies among chemical species. To fill this gap, we propose a Bayesian causal regression factor model that estimates the multivariate causal effects of wildfire smoke on the concentration of 27 chemical species in $\mathrm{PM}_{2.5}$ across the United States. Our approach introduces two key innovations: (i) a causal inference framework for multivariate potential outcomes, and (ii) a novel Bayesian factor model that employs a probit stick-breaking process as prior for treatment-specific factor scores. By focusing on factor scores, our method addresses the missing data challenge common to causal inference and enables a flexible, data-driven characterization of the latent factor structure, which is crucial to capture the complex correlation between multivariate outcomes. Through Monte Carlo simulations, we show the accuracy of the model in estimating the causal effects in multivariate outcomes and characterizing the treatment-specific latent structure. Finally, we apply our method to US air quality data, estimating the causal effect of wildfire smoke on 27 chemical species in $\mathrm{PM}_{2.5}$, providing a deeper understanding of their interdependencies.

The I-MAP Parameterization of Gaussian DAG Models

Alessandro Mascaro
Universitat Pompeu Fabra

We introduce a novel parameterization of Gaussian Directed Acyclic Graph (DAG) models, called the I-MAP parameterization. Unlike the conventional parameterization through the LDL decomposition of the precision matrix, the I-MAP parameterization always includes a single parameter corresponding to a pre-specified causal effect of interest. This feature proves especially valuable in Bayesian inference of causal effects when the DAG is unknown and must be learned from data. By incorporating this single parameter, our parameterization enables full control over the prior on the causal effect and ensures its consistent specification across all the DAGs that may be considered in a model averaging strategy. We illustrate the utility of the I-MAP parameterizations for Bayesian hypothesis testing of causal effects, and show how it enables consistent prior specification in two-step procedures that first derive the posterior distribution over DAGs and then produce Bayesian point estimates of causal effects under different loss functions.

Evaluating causal effects on time-to-event outcomes in an RCT in oncology with treatment discontinuation

Fabrizia Mealli
European University Institute

In clinical trials, patients may discontinue treatments prematurely, breaking the initial randomization. The ICH E9(R1) Addendum provides guidelines for handling such "intercurrent events;" the right strategy to adopt depends on the questions of interest. Our study is motivated by a randomized controlled trial (RCT) in oncology, where patients assigned the investigational treatment may discontinue it due to adverse events. We propose adopting a principal stratum strategy and decomposing the overall ITT effect into principal causal effects for groups of patients defined by their potential discontinuation behaviour. We first show how to implement a principal stratum strategy to assess causal effects on a survival outcome in the presence of continuous time treatment discontinuation, its advantages, and the conclusions one can draw. Our strategy deals with the time-to-event intermediate variable that may not be defined for patients who would not discontinue; moreover, discontinuation time and the primary endpoint are subject to censoring. We employ a flexible model-based Bayesian approach to tackle these complexities, providing easily interpretable results. We apply this Bayesian principal stratification framework to analyze synthetic data of the motivating oncology trial. We simulate data under different assumptions that reflect real scenarios where patients' behaviour depends on critical baseline covariates. Supported by a simulation study, we shed light on the role of covariates in this framework: beyond making structural and parametric assumptions more credible, they lead to more precise inference and can be used to characterize patients' discontinuation behaviour, which could help inform clinical practice and future protocols.

Invited Session: Bayesian graphical models

Inference of multiple high-dimensional networks with the Graphical Horseshoe prior

Claudio Busatto
Università degli Studi di Padova

We develop a novel full-Bayesian approach for multiple correlated precision matrices, called multiple Graphical Horseshoe (mGHS). The proposed approach relies on a novel multivariate shrinkage prior based on the Horseshoe prior that borrows strength and shares sparsity patterns across groups, improving posterior edge selection when the precision matrices are similar. On the other hand, there is no loss of performance when the groups are independent. Moreover, mGHS provides a similarity matrix estimate, useful for understanding network similarities across groups. We implement an efficient Metropolis-within-Gibbs for posterior inference; specifically, local variance parameters are updated via a novel and efficient modified rejection sampling algorithm that samples from a three-parameter Gamma distribution. The method scales well with respect to the number of variables and provides one of the fastest full-Bayesian approaches for the estimation of multiple precision matrices. Finally, edge selection is performed with a novel approach based on model cuts. We empirically demonstrate that mGHS outperforms competing approaches through both simulation studies and the application to a bike-sharing and a genomic dataset.

Bayesian inference of multiple Ising models for heterogeneous public opinion survey networks

Alejandra Avalos-Pacheco
Johannes Kepler Universität Linz

Public opinion studies show that relationships between opinions shift based on respondent characteristics. Understanding these complexities requires methods that account for heterogeneity across groups. We adopt a class of multiple Ising models that use graphs to analyse how external factors—such as time spent online or generational differences—shape joint dependence relationships between opinions. A Bayesian methodology is proposed based on a Markov Random Field prior, allowing information sharing across groups to encourage common edges when supported by data. A spike-and-slab prior induces sparsity and identifies shared graph structures across subgroups. Specifically, we develop two Bayesian approaches for inferring multiple Ising models, focusing on model selection: (i) a Fully Bayesian method for low-dimensional graphs using conjugate priors and exact likelihood and (ii) an Approximate Bayesian method for high-dimensional graphs based on a quasi-likelihood approach, avoiding computational intractability. These methods are applied to two US public opinion studies: one examining how time spent online affects confidence in political institutions, and another exploring intergenerational differences in opinions on public spending. Our results balance identifying significant edges (both shared and group-specific) with sparsity while quantifying uncertainty, ultimately revealing how external factors shape public opinion dynamics.

Learning block structures in Gaussian graphical models for spectrometric data analysis

Alessandro Colombi
Università degli Studi di Milano-Bicocca

Probabilistic graphical modeling serves as a robust framework for capturing the conditional dependencies among variables that follow a Gaussian distribution. Within such models, each node represents a variable, and the absence of an edge between nodes indicates conditional independence given all other variables. Previous studies have applied this methodology to spectrometric data analysis, aiming at discovering the relationships among substances within a compound by analyzing their spectra. Such a goal has been achieved by coupling smoothing techniques for functional data analysis with a Bayesian Gaussian graphical model on basis expansion coefficients, hence simultaneously smoothing the data and providing an estimate of their conditional independence structure. Empirical evidence from real-world applications has shown that the adjacency matrix describing the underlying graph often presents a block structure. This implies a natural clustering of variables into disjoint groups. In this work, a new prior for Gaussian graphical models is introduced to learn the underlying clustering structure of the nodes. The method builds upon stochastic block models while accounting for the natural ordering of the nodes. The model is employed to analyze fruit purees and discover groups of portions of their spectra.

Posters

Double covariate mean-covariance factor regression

Davide Agnoletto
Duke University

We propose a novel class of models for dimensionality reduction and structured shrinkage in Gaussian multivariate outcome linear regression models, with covariate-dependent mean and covariance. A key motivation is to treat the different variables as non-exchangeable a priori leveraging on meta-covariates, which are features of the variables instead of the samples. Latent factor regression provides a popular approach for dimensionality reduction in related contexts, but current methods favor shrinkage of the mean and off-diagonal elements of the covariance to zero. We propose a shared subspace double covariate latent factor regression framework, which incorporates information from meta-features and sample-specific covariates through a regression for the factor loadings matrix. This accommodates changes in covariance with covariates while using meta-features to inform shrinkage. Taking a Bayesian approach to inference, we develop an efficient Gibbs sampler that can adaptively infer the relevant structure to shrink towards. We motivate this approach with applications to studying variation with covariates in environmental exposures.

Control Variate-based Stochastic Sampling from the Probability Simplex

Francesco Barile
University of Milano-Bicocca

This paper presents a control variate-based Markov chain Monte Carlo algorithm for efficient sampling from the probability simplex, with a focus on applications in large-scale Bayesian models such as latent Dirichlet allocation. Standard Markov chain Monte Carlo methods, particularly those based on Langevin diffusions, suffer from significant discretization errors near the boundaries of the simplex, which are exacerbated in sparse data settings. To address this issue, we propose an improved approach based on the stochastic Cox-Ingersoll-Ross process, which eliminates discretization errors and enables exact transition densities. Our key contribution is the integration of control variates, which significantly reduces the variance of the stochastic gradient estimator in the Cox-Ingersoll-Ross process, thereby enhancing the accuracy and computational efficiency of the algorithm. We provide a theoretical analysis showing the variance reduction achieved by the control variates approach and demonstrate the practical advantages of our method in data subsampling settings. Empirical results on large datasets show that the proposed method outperforms existing approaches in both accuracy and scalability.

A Bayesian nonparametric approach to multiview clustering

Giulio Beltramin
Politecnico di Milano

Multiview data consist of various types of information about the same subjects, and multiview clustering aims to infer separate but interdependent clustering structures across these views. The challenge lies in defining models that can range from completely dependent partitions, where the clusters are identical across views, to independent partitions that treat each view separately. Taking inspiration from a recent work of Dombowsky and Dunson, we introduce a Bayesian nonparametric hierarchical model for multiview data, relying on the Pitman-Yor process. We propose a novel Chinese restaurant metaphor that facilitates the development of a sampling scheme to address Bayesian inference. The performance of our model is tested on different simulated scenarios that illustrate various dependence structures among the view-specific partitions.

Modelling the Dynamics of Intrinsically Disordered Proteins in Equilibrium Using Hierarchical Infinite Dirichlet Mixtures of Exponential Family Random Graph Models

Frances Beresford
University of California, Irvine

Intrinsically disordered proteins (IDPs) are proteins that generally do not have a stable fold - they tend not to settle into a fixed structure. IDPs can be represented using protein structure networks (PSNs) which summarise the topological structure between different types of groups of atoms within the protein - such networks can then be used to ascertain the susceptibility to aggregation of different protein variants. This modelling problem is an ideal candidate for Dirichlet Process Infinite Hierarchical Exponential Family Random Graph Models (DP-ERGMs) because we have a complex mixture of IDPs that are in different conformations, and estimating the factors that influence the different configuration is an ongoing research area of high scientific value. This application brings multiple statistical challenges including finding an effective set of base measures which may be influenced by the constraints due to the chain structure of the proteins and is especially pertinent given the final goal of density estimation. Further, it is necessary to scale up the DP-ERGM framework relative to previous use cases to take advantage of larger data sets. Here DP-ERGMs will be used to model the equilibrium behaviour of the protein Abeta(1-40) using PSN representations.

Beta-Liouville priors for Multinomial mixture models

Massimo Bilancia
University of Bari Aldo Moro

We present a variant of the Multinomial mixture model tailored to the unsupervised classification of short text data. While the Multinomial probability vector is traditionally assigned a Dirichlet prior distribution, this work explores an alternative formulation based on the Beta-Liouville distribution, which offers a more flexible correlation structure than the Dirichlet. We examine the theoretical properties of the Beta-Liouville distribution, with particular focus on its conjugacy with the Multinomial likelihood. This property enables the derivation of update equations for a CAVI (Coordinate Ascent Variational Inference) algorithm, facilitating approximate posterior inference of the model parameters. In addition, we introduce a stochastic variant of the CAVI algorithm to enhance scalability. We also demonstrate effective strategies for selecting the Beta-Liouville hyperparameters.

Revisiting self-normalized importance sampling: new methods and diagnostics

Nicola Branchini
University of Edinburgh

Importance sampling (IS) can often be implemented only with normalized weights, yielding the popular self-normalized IS (SNIS) estimator. However, proposal distributions are often learned and evaluated using criteria designed for the unnormalized IS (UIS) estimator. We aim to present a unified perspective on recent methodological advances in understanding and improving SNIS. We propose and compare two new frameworks for adaptive importance sampling (AIS) methods tailored to SNIS. Our first framework exploits the view of SNIS as a ratio of two UIS estimators, coupling two separate AIS samplers in a joint distribution selected to minimize asymptotic variance. Our second framework instead proposes the first MCMC-driven AIS sampler directly targeting the (often overlooked) optimal SNIS proposal. We also establish a close connection between the optimal SNIS proposal and so-called subtractive mixture models (SMMs), where negative coefficients are possible - motivating the study of the properties of the first IS estimators using SMMs. Finally, we propose new Monte Carlo diagnostics specifically for SNIS. They extend existing diagnostics for numerator and denominator by incorporating their statistical dependence, drawing on different notions of tail dependence from multivariate extreme value theory.

A non-informative prior for nonparametric inference on time series

Ylenia Francesca Buttigliero
Università di Torino

We consider data collected at discrete time points from a hidden Markov model whose latent signal evolves in continuous time. Our goal is to make inference and assess uncertainty in the intervals where data have not been observed, without imposing parametric assumptions. Recent work showed that Fleming-Viot (FV) processes, which are diffusions over the space of atomic probability measures, induce tractable nonparametric priors for the signal in a nonparametric hidden Markov model setting, and identified the marginal posterior distributions given the data. Our goal is to consider a non-informative limit of such formulation, which would extend the Bayesian bootstrap to time-dependent observations. In this scenario, we aim at characterizing the smoothing distributions, i.e. the distribution of the signal given past and future data, at any time point, and at devising a fast computational procedure which easily allows for multiple observation times and large sample sizes. Here we present preliminary results on the above goals which consider finite-dimensional projections of the signal onto measurable partitions of the sampling space, whereby the analysis can exploit the properties of Dirichlet distributions and Wright-Fisher diffusions. We also leverage on some asymptotic approximations for the implementation, justified by the literature of the underlying models, to improve the execution time and avoid known numerical problems. Our results describe the distribution of potential data under these assumptions at time points where no data are available, and allow for a quantification of the uncertainty on further data extraction. The adopted setting does impose restraining assumptions on the data generating model and extends the finite-dimensional Bayesian bootstrap to temporally correlated data.

Hierarchical Shot Noise Cox Process Mixture Models

Alessandro Carminati
Politecnico di Milano

We present the Hierarchical Shot Noise Cox Process mixtures, a novel Bayesian nonparametric mixture model to cluster partially exchangeable data. In this context, observations are grouped, and we aim at clustering them across groups. Hierarchical mixture models in the existing literature, e.g., the Hierarchical Dirichlet process mixtures, obtain across-group clustering as the union of within-group clusters related to mixture components with the same parameters' values. This type of across-group information sharing can be limiting in real-world applications: for instance, two mixture components belonging to different groups might have slightly different parameters' values, though representing the same across-group cluster. To overcome these limits, our model achieves across-group clustering from the union of within-group clusters related to mixture components with similar parameters' values, where the notion of similarity is encoded in a kernel probability function. We show a priori and a posteriori properties of our model, and we present Markov chain Monte Carlo algorithms for posterior inference. We illustrate our model through applications in real-world scenarios.

Bayesian Multivariate Longitudinal Modeling of Metabolic Syndrome in Blood Donors

Simone Colombara
Politecnico di Milano

Metabolic syndrome (MetS) is a cluster of conditions that significantly increases the risk of cardiovascular disease and type 2 diabetes. In collaboration with AVIS Lambrate, we analyze a rich longitudinal dataset of over 4,000 Italian blood donors to investigate the joint progression of key MetS markers—waist circumference, triglycerides, HDL cholesterol, glucose, and systolic blood pressure. We develop a Bayesian multivariate mixed-effects model that accounts for repeated measurements and the complex dependency structure among the target variables. The model incorporates a wide range of demographic, lifestyle, and clinical covariates. Posterior inference identifies key physiological and behavioral predictors of MetS, such as body weight, age, and physical activity, while highlighting the leading role of waist circumference in the syndrome's progression. This work lays the foundation for a two-stage predictive tool aimed at supporting early detection of MetS and guiding personalized health interventions. Future extensions will target other conditions relevant to blood donation eligibility, such as liver diseases. Our results support AVIS's mission to improve donor health and optimize long-term donation practices through statistical modeling and data-driven risk assessment.

Autocompound Random Measures

Riccardo Corradin
University of Milano-Bicocca

We introduce a class of time-dependent nonparametric models, suited for scenarios involving populations observed at distinct discrete times. Starting with an ancestral random measure, it is possible to define a sequence of time-specific random measures by acting on their intensity functions, in the spirit of compound random measures. The resulting family of models exhibits desirable properties, including mathematical tractability, simple expressions for its main summaries, and a closed-form representation of the joint posterior distribution at distinct observed times. Such a model can be then normalized and used as a building block for dynamic population studies, defining tractable species sampling models that evolve over time, or convoluted with a kernel function to obtain time-dependent mixture models.

Bayesian latent factor models for causal inference

Lea Anna Cozzucoli
University of Trieste

Accurate causal estimation from observational data requires properly accounting for both observed and unobserved heterogeneity. We introduce a Bayesian factor model that represents each outcome as a combination of measured covariates and latent factors capturing unobserved confounders, enabling robust recovery of unit specific treatment effects. By situating the model in a fully Bayesian framework and leveraging efficient Markov chain Monte Carlo algorithms, we deliver coherent uncertainty quantification for all parameters and for the causal effect. The approach is extends to multiple outcomes and yields both individual and average treatment effect estimates, even in small-sample settings. Simulation studies confirm that the proposed method achieves proper interval coverage and gives robust estimates for the causal effect, offering a flexible tool for reliable causal inference across diverse observational contexts.

Dependent Dirichlet processes via thinning

Laura D'Angelo
University of Milan-Bicocca

Analyzing data from multiple sources often requires models that balance the ability to share information across samples with the flexibility to capture their heterogeneity. In this work, we introduce a novel framework for modeling a collection of dependent Dirichlet processes using a thinning mechanism. The proposed approach modifies the Dirichlet process's stick-breaking representation by randomly discarding the beta random variables involved in the construction. The result is a collection of dependent random distributions based on a common set of atoms and different weights. The simplicity of the formulation allows for the characterization of several theoretical properties and the derivation of efficient computational methods for posterior inference. An application to simulated data and the Collaborative Perinatal Project data illustrates the model's ability to flexibly estimate the group-specific densities and uncover a meaningful partition of the observations, providing valuable insights into the underlying structure.

Optimal filtering for 2-parameter Poisson-Dirichlet diffusions

Marco Dalla Pria
Università di Torino

Evolving systems with unlabelled components arise in many fields, whenever the components are not associated to categories but only to normalized values based on grouping single elements in an underlying structure, like a large population or a network. In such scenarios, data arise as unlabelled partitions or Young diagrams, whose likelihood is an intractable function symmetric in the categories.

Assuming the system dynamics follow a two-parameter Poisson-Dirichlet diffusion, we show how to obtain exact distributions for the system states when the number of categories is unknown, conditional on noisy data given by Young diagrams collected at discrete times. After developing some necessary results on the coagulation of random partitions, we describe how to perform online and offline inference for such system when data become available over time or when they come in a batch, without resorting to MCMC or SMC. We also describe the predictive distributions for drawing further Young diagrams at arbitrary times, which are certain finite mixtures of Ewens-Pitman sampling formulae. After discussing the practical implementation and illustrating the inferential procedures with numerical experiments, we apply the introduced methodology to the dynamic estimation of the heterozygosity in live social interactions.

Bayesian estimation of extreme sub-hourly precipitation with increasing temperature

Matteo Darienzo
Department of Civil, Environmental, and Architectural Engineering, University of Padova, Padova, Italy

Extreme sub-hourly precipitation can lead to flash floods and other natural disasters. Improving the estimation of the exceedance probability of these events is particularly important in a changing climate. A recent study has proposed a non-asymptotic and non-stationary statistical method (called TENAX), with a physically based dependence from the temperature as covariate. This method includes all the peaks of independent ordinary rainfall events, not only a small sample of the extremes (e.g., annual maxima) and establishes dependence between the two parameters of the Weibull distribution with the near-surface air temperature (as suggested by Clausius–Clapeyron relation and several studies). The parameters were originally estimated with the maximum likelihood method, which did not allow a proper quantification of the uncertainties. Here, we implement the TENAX model within a Bayesian framework with a MCMC Metropolis algorithm. Physics-informed priors are specified on the parameters. Convergence of the several thousands MCMC simulations is assessed by computing classical tests. We quantify the uncertainty on the return levels by computing the credibility interval at 90% from all obtained spaghettis. Preliminary results on several stations in Switzerland show consistency of the past and future return levels with the previous TENAX, and with other estimates based on a non-asymptotic method (SMEV, both in its classic and time-dependent implementations). Perspectives of this work include testing other MCMC algorithms, such as adaptive and multi-block algorithms to reach a user-defined acceptance rate, and accounting for the observational uncertainties in the Inference.

A Bayesian nonparametric approach to the multi-armed bandit problem in traits allocation models

Claudio Del Sole
Università degli Studi di Milano - Bicocca

In traits allocation models, each observation may display different features and may have different levels of belonging for each feature. For example, multiple individuals from various species may be observed within a time or space window. When data are collected from multiple populations, the same feature may appear in two different observations both within and across populations. We consider the problem of maximizing the total number of observed features by sequentially selecting the population from which the next observation is collected. This task can be framed as a multi-armed bandit problem, with reward given by the number of newly discovered features. We develop a Bayesian nonparametric approach relying on hierarchical gamma processes, which promotes borrowing of information among populations by establishing a common set of features. The tractable posterior characterization allows to implement a Thompson sampling strategy to balance exploration and exploitation. This approach is compared with the simpler strategy that selects the population with highest posterior estimate for the number of new features. For further comparison, we also derive a novel frequentist estimator and implement the correspond UCB strategy. The performances of the proposed algorithms are assessed via simulation studies, and compared on a real dataset containing tree species counts from different plots at different locations in Japan.

Bayesian target discovery from categorical networks

Laura Ferrini
Università degli Studi di Milano-Bicocca

A primary goal in biology is represented by target discovery, namely the process of identifying genes that are either affected by drugs, or responsible of disease traits, from the available data. In general, both drug administration and disease occurrence can be conceived as exogenous interventions that are responsible of modifications in the data generating mechanism. Importantly, genes are organized into pathways representing their biological interactions. These can be effectively represented through Directed Acyclic Graphs (DAGs), which provide a useful framework for target discovery purposes. Specifically, disease-target discovery consists of identifying those nodes (genes) on which an intervention (disease) has produced some modifications. Since genetic variants correspond to levels of categorical variables, we base our framework on categorical DAGs, and develop a Bayesian methodology for target discovery which fully accounts for the statistical uncertainty around both targets and DAGs. Posterior inference of such parameters is carried out through a Markov chain Monte Carlo scheme. We establish theoretical guarantees relative to target discovery and show in simulation studies the advantages of our approach with respect to alternative methods. Our methodology is applied to a disease-target discovery problem for patients affected by cystic fibrosis.

A Bayesian hierarchical model for air pollution source apportionment

Michela Frigeri
Politecnico di Milano

Environmental pollutants often consist of complex mixtures of constituents from multiple sources. Understanding the contributions of each source to a pollution mixture is crucial for identifying key polluters and developing effective strategies to reduce pollution levels. While data on individual components of these mixtures are available, identifying their exact sources can be challenging, as many pollutants originate from different sources. Source apportionment techniques aim to relate emissions from various sources to observed air pollution concentrations at specific locations and times. In this study, we introduce a new Bayesian nonparametric method to identify major pollution sources in a given geographical region with multiple receptors. Bayesian approaches for source apportionment typically assume a fixed number of sources with fixed structures that remain constant across space, simplifying the posterior estimation process. However, these assumptions are often too rigid and may not accurately reflect the complexities of real-world air pollution. Our model addresses these limitations by allowing variability in both the number and composition of pollution sources across different pollutant receptors. Using measurements of pollutant concentrations over time from various monitoring sites, we leverage information about the spatial distribution, local characteristics, and temporal patterns of pollutants to develop a Bayesian hierarchical model for source apportionment. We model pollutant concentrations as functional data and incorporate Bayesian functional warping to improve model flexibility, while also accounting for local features specific to each monitoring station. Additionally, our approach provides a Bayesian nonparametric estimate of the number of pollution sources, modeling the unknown source profiles as latent factors in our hierarchical model. We illustrate the capabilities of our model through a realistic simulation study, showing its accuracy and applicability. Finally, we apply the model to measurements of particulate matter (PM) concentrations and their constituents collected from multiple monitoring sites across California (USA), showing its practical use in source identification and analysis of air pollution.

Addressing Phase Discrepancies in Functional Data: A Bayesian Approach for Accurate Alignment and Smoothing

Jacopo Gardella
Università di Milano-Bicocca

In many real-world scenarios, functional data display significant variability in both amplitude and phase-particularly in biomechanical applications such as knee flexion angle analysis. In our motivating dataset, timing discrepancies across curves and substantial inter-subject differences pose challenges for alignment without distorting key individual characteristics. We propose a Bayesian alignment model that addresses these issues by eliminating phase variability while preserving the amplitude and unique features of each curve. The model's flexible smoothing component ensures minimal distortion during alignment, and a dedicated group-level parameter allows it to naturally accommodate hierarchical structures. A novel prior on the warping functions guarantees the validity of the resulting transformations without the need for ad hoc constraints. We demonstrate the effectiveness of our approach on the knee flexion dataset, where it achieves robust alignment and smoothing performance even in the presence of complex group structures and high inter-curve variability.

Inference via mixture predictive distributions

Samuele Garelli
University of Bologna

A novel paradigm for Bayesian inference, which bypasses the usual likelihood-prior scheme, has been recently introduced by Fong, Holmes and Walker. The idea is to reconstruct the unobserved part of the population by sampling from a sequence of predictive distributions and then estimate the parameter of interest as a function of the observed and imputed data. Such mechanism is called predictive resampling. To guarantee reliable inference, predictives need to have a good fit on the initial dataset and converge to a limit distribution that preserves the information provided by the observed sample.

A new class of predictive distributions applicable in this context is proposed. We start by partitioning the observed data with model-based clustering and setting the first predictive as a mixture law whose components have the same mean and covariance matrix as the observed clusters. Then, the next predictives are obtained through a mechanism inspired by Polya urn models, which updates only the mean and covariance matrix of the component from which observations are drawn. The initial clustering ensures these predictives fit well the observed data. Moreover, we prove they converge almost surely in total variation to a mixture whose random parameters (i.e. weights, means and covariance matrices) have mean equal to the parameters of the initial predictive. This ensures predictive resampling generates uncertainty but no new information (which is desired, since the only available information about the true population is contained in the observed sample). Such approach is quite flexible, in that theoretical results hold for any mixture whose components have density with respect to Lebesgue measure.

The methodology is applied to simulated and real data and satisfying results are obtained in posterior estimation of parameters (e.g. mean, variance, skewness, kurtosis, quantiles and covariance) and in regression on non-linear and heteroscedastic data.

A Bayesian Multi-Study Model for Sparse and Dense Latent Structure Discovery

Leonardo Genesin
University of Padua

Latent structures in high-dimensional data can vary widely in sparsity and scope—ranging from dense factors that exhibit in all observations, as in traditional Factor Models, to localized patterns associated with only subsets of variables and samples, as in Biclustering methods. The growing availability of data from multiple related studies offers a powerful opportunity to identify such patterns more reliably, distinguishing replicable signals shared across studies from spurious, study-specific noise. We propose a flexible Bayesian Multi-Study Spike-and-Slab model that jointly analyzes multiple datasets to uncover a broad range of latent structures, allowing each to be either dense or sparse, and either shared among studies or study-specific. Strengths and limitations of the model are discussed via simulations and an application to single-cell transcriptomic studies.

Bayesian nonparametric processes for clustering count data with potentially unobserved traits

Lorenzo Ghilotti
University of Milano-Bicocca

Count data arise in diverse domains such as criminology, where investigations track criminals at meetings and the information is encoded through binary data, and other fields like ecology. A key goal may often be to cluster a sample of subjects, e.g., criminals, based on their count records for the observed traits, e.g., meetings. While clustering the counts of the observed traits is intuitive, such models are often misspecified. Indeed, in a variety of settings, it is realistic to assume that a larger set of traits exists, of which only a subset is observed. In the criminal context, it is plausible that investigations missed some of the actual meetings. We propose a novel Bayesian nonparametric methodology for clustering that accounts for potentially unobserved traits. Our approach embeds trait allocation models into a mixture model. The random mixing measure is almost surely supported on a countable set of realizations from finite Completely Random Measures with shared common atoms. We analytically show that our methodology yields fewer clusters than its natural naive competitor. Assuming a known clustering structure, we derive the closed-form expression for the marginal distribution of the data, we characterize the posterior distribution of the de Finetti vector of random measures and the predictive structure for new subjects. This fully-Bayesian analysis enables the development of a fast, efficient marginal algorithm and allows for trait-by-trait comparisons across clusters. Finally, we illustrate our methodology by analyzing a criminal dataset documenting the attendance of 'Ndrangheta affiliates at meetings.

Bayesian nonparametric boundary detection for income areal data

Matteo Gianella
Politecnico di Milano

Recent discussions on the future of metropolitan cities underscore the pivotal role of (social) equity, driven by demographic and economic trends. More equal policies can foster and contribute to a city's economic success and social stability. In this work, we focus on identifying metropolitan areas with distinct economic and social levels in the greater Los Angeles area, one of the most diverse yet unequal areas in the United States. Utilising American Community Survey data, we propose a Bayesian model for boundary detection based on areal income distributions. The model identifies areas with significant income disparities, offering actionable insights for policymakers to address social and economic inequalities. We have multiple observations (i.e., personal income of survey respondents) for each area, and our approach, formalised as a Bayesian structural learning framework, models areal densities through mixtures of finite mixtures. We address boundary detection by identifying boundaries for which the associated geographically contiguous areal densities are estimated as being very different without resorting to dissimilarity metrics or covariates. Efficient posterior computation is facilitated by a transdimensional Markov Chain Monte Carlo sampler. The methodology is validated via extensive simulations and applied to the income data in the greater Los Angeles area. We identify several boundaries in the income distributions, which can be explained "ex-post" in terms of the percentage of the population without health insurance, though not in terms of the total number of crimes, showing the usefulness of such an analysis to policymakers.

Bayesian Markov-Switching Model for Regional Heat Wave Detection

Vincenzo Gioia
University of Trieste

Summer temperatures fluctuate dynamically around a seasonal component and may exhibit prolonged periods of extremely high temperatures, commonly referred to as heat waves. The classification of a period as an heat wave depends on the specific definition adopted and it could be based on quantile-based thresholds. In this work, we address both heat wave detection and temperature dynamics modelling by leveraging a multi-state Bayesian Markov-switching regression model accounting for seasonality, long-term trends (via a large-scale climate index), and time-varying transition probabilities. By increasing the number of latent states, the model effectively isolates periods of high and extremely high temperatures, enabling the identification of heat wave episodes. We analyse the maximum daily temperatures across six geographically close but morphologically different weather stations of the Italian region of Friuli Venezia Giulia. We provide insights into regional variability in heat wave characteristics that reflect differences and similarities in the morphological conformation of the territory.

Mixture priors for replication studies

Roberto Macrì Demartino
University of Trieste

The credibility of scientific research relies heavily on the replicability of its findings. However, in recent years, an increasing number of published results have failed to replicate, raising concerns about a “replication crisis” in several fields. As a result, the scientific community has increasingly emphasised the importance of replication studies, even though establishing replication success remains challenging. Analysing replication studies essentially involves using historical data from the original study. Given the inherent nature of sequential information updating, Bayesian methods are a natural choice. In particular, an intuitive strategy is to incorporate historical information by using a prior distribution based on the data from the original study for the analysis of the replication data. We propose a novel Bayesian approach that employs mixture priors. Specifically, our method uses the mixture of the posterior distribution from the original study with a non-informative prior to evaluate the replication study. The mixture weight determines the extent to which the original and the replication data are pooled. This mixture prior approach offers a flexible alternative to existing methods–such as hierarchical models and power priors–for assessing replication success. We explore two strategies for setting the mixture weight parameter. The first strategy fixes the weight at a specific value, for instance, based on expert knowledge or an empirical Bayes estimate, and then assesses the sensitivity of this choice using a tipping point analysis. The second strategy introduces uncertainty by assigning a prior distribution to the mixture weight parameter. Furthermore, within these frameworks, Bayes factors can be used for formal hypothesis testing, such as evaluating the presence or absence of an effect and to assess how closely the replications align with the original study. We analyse the asymptotic behaviour of the marginal posterior distribution for the weight parameter as the Bayes factor tends to zero or infinity. Moreover, we examine how the Bayes factor associated with the effect size behaves as the replication study’s standard error approaches zero. We demonstrate the applicability of our method using data from a communication science experiment titled “Moral Credentialing”. Our findings suggest that mixture priors are a valuable and intuitive alternative to other Bayesian methods for analysing replication studies. Furthermore, we provide the free and open-source R package repmix, which implements the proposed methodology.

Optimal Bayesian designs, observational data, and utility functions: a new proposal

Nedka Dechkova Nikiforova
Department of Statistics Computer Science Applications "G.Parenti", University of Florence

In this presentation, we deal with an innovative approach for building Bayesian optimal designs by defining new utility functions specifically for the reliability field, also exploiting an existing dataset. The reliability dataset considered here relates to a soldering process in electronic engineering and involves a comparison of different alloy surface finish combinations, taking into account various batches, component types, and pin sizes. The proposed procedure enables to obtain optimal Bayesian designs using utility functions when observational data are available. Particularly in cases where trials are costly to conduct, the procedure enables final reliability modelling results to be achieved without incurring additional expenses. Moreover, we propose a tailored Weibull Bayesian model designed to address the hierarchical structure of our reliability data. Optimal designs are obtained through Markov Chain Monte Carlo (MCMC) simulations, using two utility functions that explicitly account for associated costs. A sensitivity analysis was also conducted to assess the robustness of the results under different prior settings. The outcomes are highly satisfactory, demonstrating that the proposed procedure can be effectively applied to similar engineering contexts.

Enhancing Interpretability in Constrained Latent Class Models: a novel approach

Virginia Murru
University of Padova

Constrained latent class models (CLCMs) offer a powerful framework for analysing multivariate categorical data in an unsupervised manner, where a typical dataset consists of units answering a set of survey questions, often referred to as items. CLCMs consider a set of latent features that each unit may or may not possess, with each specific combination of these features defining the unit's group membership. The response probability for each item depends on the subject's class assignment; however, different groups can share identical response probabilities for certain items through a selection matrix. While these relaxations enhance the model's ability to uncover latent structures, the number of classes increases rapidly as the number of latent features grows, a situation further complicated by the shared parameter structure. As a result, the response probabilities for different groups become difficult to interpret and compare, compromising an important advantage that characterises traditional LCMs. In this work, we develop a variant of CLCM that preserves interpretability while retaining the flexibility needed to identify complex patterns, achieved by restricting the number of possible clusters via a stick-breaking process.

Cumulative Shrinkage Processes for Bayesian Multiresolution Functional Regression

Andrea Ongarato
University of Padova

Modern regression problems require flexible methods to handle data with varying complexity across different regions of the support. Traditional approaches often fail to address this heterogeneity effectively. We propose a Bayesian multi-resolution functional regression model using basis expansions combined with Spike-and-Slab priors for adaptive regularization. In particular, we propose a modification of the widely used cumulative shrinkage process (CUSP) that automatically determines the appropriate model complexity at different resolution levels. The model represents functions through infinite basis expansions, with shrinkage probabilities increasing for more complex bases. Building on the CUSP, we address a limitation of the standard stick-breaking processes used for cumulative shrinkage: they prevent the complete elimination of unnecessary layers. Our modification allows exact shrinkage to zero for irrelevant components. Implemented via an efficient Gibbs sampler, the method successfully identifies regions of differing complexity while selecting optimal basis functions. Applications demonstrate its effectiveness in function estimation and signal decomposition. Our framework also shows promise for broader Bayesian nonparametric applications, such as mixture model component selection. This approach particularly benefits problems with spatiotemporal covariates, where local adaptation is crucial. Indeed, by combining multi-resolution modeling with rigorous Bayesian regularization, we achieve both flexibility and interpretability.

Posterior Computation for the Dirichlet-Laplace Prior

Paolo Onorati
University of Padova

Modern statistical problems routinely deal with high-dimensional parameter spaces, where the number of parameters far exceeds the sample size, typically involving structural assumptions of sparsity. While penalized optimization methods focus on sparse signal recovery, Bayesian approaches provide a probabilistic framework to formally quantify uncertainty through shrinkage priors. Among these, the Dirichlet-Laplace prior has attained recognition for its theoretical guarantees and wide applicability. This article identifies a critical yet overlooked issue in the implementation of Gibbs sampling algorithms for such priors. We demonstrate that ambiguities in the presentation of key algorithmic steps, while mathematically coherent, have led to widespread implementation inaccuracies that fail to target the intended posterior distribution—a target endowed with rigorous asymptotic guarantees. Using the normal-means problem as canonical example, we clarify these implementation pitfalls and their practical consequences and propose corrected and more efficient sampling procedures.

The Gnedin model for biodiversity estimation

Anna Petranzan
University of Milano-Bicocca

Species sampling problems have been extensively studied in ecology and biology, focusing on challenges such as quantifying species richness and estimating the diversity of rare species. Discrete random probability measures serve as fundamental tools for addressing these issues. Among Bayesian nonparametric priors, the broad family of Gibbs-type priors stands out, offering a generalization of well-known processes such as the Dirichlet and Pitman-Yor processes. This work investigates a specific Gibbs-type prior introduced by Gnedin (2010), which assumes a finite but random number of distinct species within the population. This unique characteristic enables natural Bayesian estimation of population size. We explore both the theoretical and empirical aspects of Gnedin's model, presenting new theoretical results and developing practical procedures for its estimation in the context of species discovery. Notably, while the model has been the subject of extensive theoretical study, empirical demonstrations of its effectiveness are, to our knowledge, lacking. Additionally, we examine the role of hyperparameters and propose strategies for their elicitation.

Bayesian Transfer Learning for Artificially Intelligent Geospatial Systems: A Predictive Stacking Approach

Luca Presicce
University of Milano-Bicocca

Building artificially intelligent geospatial systems require rapid delivery of spatial data analysis at massive scales with minimal human intervention. Depending upon their intended use, data analysis may also entail model assessment and uncertainty quantification. This article devises transfer learning frameworks for deployment in artificially intelligent systems, where a massive data set is split into smaller data sets that stream into the analytical framework to propagate learning and assimilate inference for the entire data set. Specifically, we introduce Bayesian predictive stacking for multivariate spatial data and demonstrate rapid and automated analysis of massive data sets. Furthermore, inference is delivered without human intervention without excessively demanding hardware settings. We illustrate the effectiveness of our approach through extensive simulation experiments and in producing inference from massive datasets on sea surface temperatures and vegetation index that are indistinguishable from traditional (and more expensive) statistical approaches.

Modeling Spatio-Temporal Dynamics of Obesity in Italian Regions via Bayesian Beta Regresion

Luciano Rota
Università Milano Bicocca

We investigate the spatio-temporal dynamics of obesity rates across Italian regions from 2010 to 2022, aiming to identify spatial and temporal trends and assess potential heterogeneities. We implement a Bayesian hierarchical Beta regression model to analyze regional obesity rates, integrating spatial, temporal, and gender effects alongside various exogenous predictors. The model leverages the Stochastic Search Variable Selection (SSVS) technique to identify significant predictors supported by the data. The analysis reveals both regional heterogeneity and dependence in obesity rates over the study period, emphasizing the importance of considering gender and spatial correlation in explaining its dynamics over time. In fact, the incorporation of structured spatial and temporal random effects captures the complexities of regional variations over time. These random effects, along with gender, emerge as the primary determinants of obesity prevalence across Italian regions, while the role of exogenous covariates is found to be minimal at the regional level. While socioeconomic and lifestyle factors remain fundamental at a micro-level, the findings demonstrate that the integration of spatial and temporal structures is critical for capturing macro-level obesity variations.

Topic-Informed Dynamic Mixture Model for Occupational Heterogeneity in Health Risk Behaviors

Lorenzo Schiavon

Ca' Foscari University of Venice

Behavioral risk factors—smoking, poor nutrition, alcohol misuse, and physical inactivity (SNAP)—are leading contributors to chronic diseases and healthcare costs worldwide. Their prevalence is shaped not only by socio-demographic characteristics but also by broader contextual influences such as occupational environments. In this study, we leverage data from PASSI, the Italian health and behavioral surveillance system, to model SNAP behaviors through a Bayesian framework that integrates textual information on occupations. Specifically, we use Structural Topic Modeling (STM) to cluster free-text job descriptions into latent occupational groups, which inform mixture weights in a multivariate ordered probit model. Covariate effects are allowed to vary across occupational clusters and evolve over time. To enhance interpretability and variable selection, we impose non-local spike-and-slab priors on regression coefficients. An online learning algorithm based on sequential Monte Carlo enables efficient updating as new data becomes available. Simulation studies validate the effectiveness of the proposed method. This dynamic, scalable, and interpretable approach reveals how occupational contexts modulate the impact of socio-demographic factors on health behaviors, providing valuable insights for targeted public health interventions.

Bayesian hierarchical modelling for multisite replication studies

Goar Shaboian

Università Cattolica del Sacro Cuore

Multisite replication designs - where multiple labs independently investigate the same phenomenon - are becoming increasingly common in empirical research. However, existing statistical methods often struggle to handle site-specific variability in effect sizes. To address this, we propose a Bayesian hierarchical partition model that captures both between-laboratory heterogeneity and within-study sampling variability. To support evaluation of replicability we propose novel metrics tailored to the multisite setting. Finally, we demonstrate the utility of our method using data from the ManyLabs replication project in psychological science.

Bayesian Mapping of Mortality Clusters

Andrea Sottosanti

University of Padova

Disease mapping analyses the distribution of several disease outcomes within a territory. Primary goals include identifying areas with unexpected changes in mortality rates, studying the relation among multiple diseases, and dividing the analysed territory into clusters based on the observed levels of disease incidence or mortality. In this work, we focus on detecting spatial mortality clusters, that occur when neighbouring areas within a territory exhibit similar mortality levels due to one or more diseases. When multiple causes of death are examined together, it is relevant to identify not only the spatial boundaries of the clusters but also the diseases that lead to their formation. However, existing methods in literature struggle to address this dual problem effectively and simultaneously. To overcome these limitations, we introduce PERLA, a multivariate Bayesian model that clusters areas in a territory according to the observed mortality rates of multiple causes of death, also exploiting the information of external covariates. Our model incorporates the spatial structure of the data directly into the clustering probabilities by leveraging the stick-breaking formulation of the multinomial distribution. Additionally, it exploits suitable global-local shrinkage priors to ensure that the detection of clusters is driven by concrete differences across mortality levels while excluding spurious differences. We propose an MCMC algorithm for posterior inference that consists of closed-form Gibbs sampling moves for nearly every model parameter, without requiring complex tuning operations. This work is primarily motivated by a case study on the territory of a local unit within the Italian public healthcare system, known as ULSS6 Euganea. To demonstrate the flexibility and effectiveness of our methodology, we also validate PERLA with a series of simulation experiments and an extensive case study on mortality levels in U.S. counties.

Bayesian Inference for Multivariate Ordinal Data with Partially Ambiguous Items

Mattia Stival
Ca' Foscari University of Venice

In survey-based research, multiple items are often used to measure latent constructs of interest. However, not all questions equally capture the intended dimension, particularly when items are subject to misinterpretation or varying respondent understanding. This can undermine the performance of traditional latent class or item response models. We propose a Bayesian mixture model for ordinal responses that explicitly accounts for multiple interpretation modes in survey items. Our approach distinguishes between well-interpreted questions and those prone to ambiguity. For the latter, we model responses through a finite mixture of components: one aligned with the latent trait of interest and others capturing alternative interpretations or noise. Covariates are incorporated both in the ordinal response model (via item-specific linear predictors) and in the mixture weights, allowing us to model how interpretation probabilities vary across demographic and spatial contexts. We apply the model to data from the Italian PASSI surveillance system, focusing on perceived environmental quality. This framework enhances interpretability and robustness in the analysis of ordinal latent constructs, particularly in heterogeneous populations.

Advances in Bayesian hidden Markov models with intractable normalizing functions

Daniele Tancini
University of Perugia

Spatial and spatio-temporal hidden Markov models are extremely difficult to estimate because their latent joint distributions are available only in trivial cases. These latent distributions are typically replaced with pseudo-distributions, which could affect the estimation results, especially in the presence of strong dependencies between the latent variables. In this work, we show how inference can be carried out in a Bayesian framework using an exchange algorithm, which eliminates the need to calculate the entire distribution of the latent variables. In addition, we propose a method to approximate the marginal likelihood of hidden Markov models with intractable normalizing functions. The new approach is based on the reciprocal importance sampling combined with the exchange algorithm. The marginal likelihood can be approximated from the output of Markov Chain Monte Carlo algorithms, using only the unnormalized posterior densities from the sampled parameter values, without requiring simulations beyond the main posterior sampling.

Bayesian Multivariate Density Regression with Coordinate-Wise Predictor Selection

Giovanni Toto
University of Padova

We propose a flexible Bayesian approach for estimating the joint density of a multivariate random variable of interest in the presence of categorical covariates. Leveraging a Gaussian copula model, our method effectively characterizes the correlation across different dimensions of the response and captures the complex behavior of the marginal distributions. The latter are modeled using mixtures of truncated normals with atoms shared across dimensions and mixture probabilities that depend on covariates through a tensor factorization framework that allows for the identification of dimension-specific sets of the most influential covariates. Dimension-specific random partition models on the covariate levels replace mode matrices providing a more flexible approach to obtain aggregated levels exhibiting similar effects on the response. Additionally, to handle high-dimensional settings with many covariates, we introduce an MCMC algorithm that scales with the number of aggregated levels rather than the original levels, significantly reducing memory requirements and improving computational efficiency. We demonstrate the method's numerical performance through simulation experiments and its practical applicability through the analysis of NHANES dietary data.

Bayesian estimation of intensity duration frequency curves

Mehwish Zaman
University of Padova

Intensity duration frequency (IDF) curves are an essential tool for characterizing the frequency of extreme rainfall events and assessing flood risk. These curves are used to describe the expected frequency of extreme rainfall measured at different durations. These curves are subject to adequate shape constraints to ensure coherent estimates of exceedance probabilities across durations. Most of the existing methods used to estimate IDF curves assume that rainfall accumulations over different durations are independent of each other, but this assumption may not always be valid. This problem is addressed in our work by proposing dependence models for the construction of IDF curves in a Bayesian parametric framework. This allows us to exploit the prior knowledge on the parameters of the IDF models and to provide a more direct assessment of the uncertainty in the estimation of higher quantiles. Our approach is based on a first-order Markov assumption, that incorporates the dependence between consecutive rainfall durations through the use of bivariate extreme models, while the marginal distributions are duration-dependent Generalized Extreme Value (d-GEV) distributions. A simulation study is presented to account for the performance of the Markov GEV model's ability to estimate IDF curves. The proposed model is applied to annual maximum intensity data over relevant rainfall durations, with an example from Veneto, Italy