

BNP 14

14th International Conference on Bayesian Nonparametrics

Book of Abstracts



UCLA, Los Angeles, June 23-27, 2025

Scientific Committee

Alessandra Guglielmi (Chair), *Politecnico di Milano (Italy)*

Jim Griffin, *UCL (UK)*

Alejandro Jara, *Pontificia Universidad Católica (Chile)*

Xenia Miscouridou, *University of Cyprus (Cyprus)*

Garritt Page, *Brigham Young University (US)*

Weining Shen, *UC Irvine (US)*

Erik Sudderth, *UC Irvine (US)*

Botond Szabo, *Bocconi University (Italy)*

Sinead Williamson, *Apple Machine Learning Research (US)*

Local Organizing Committee

Michele Guindani (*UCLA, chair*)

Falco Joannes Bargagli Stoffi (*Harvard University & UCLA*)

Veronica Berrocal (*UCI*)

Andrew Holbrook (*UCLA*)

Wesley Johnson (*UCI*)

Oscar Hernan Madrid Padilla (*UCLA*)

George Michailidis (*UCLA*)

Weining Shen (*UCI*)

Marc Suchard (*UCLA*)

Erik Sudderth (*UCI*)

Donatello Telesca (*UCLA*)

Jason Xu (*Duke University & UCLA*)

Thanks to our Sponsors

BNP 14

Platinum sponsor



U.S. National Science Foundation

Golden sponsors



European Research Council
Established by the European Commission

ERC GRANT 817257

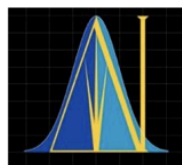


Google DeepMind

Silver sponsors



BYU
BRIGHAM YOUNG
UNIVERSITY
Department of Statistics



Data Ingenuities

 The University of Texas at Austin
Department of Statistics
and Data Sciences
College of Natural Sciences

Duke | Statistical Science

 Fielding School of Public Health
Biostatistics

 College | Physical Sciences
Statistics & Data Science

1. Invited Talks

1.1. Sergios Agapiou

Institution: University of Cyprus

Department: Department of Mathematics and Statistics

Email: agapiou.sergios@ucy.ac.cy

Website: <https://www.mas.ucy.ac.cy/sagapi01/>

Talk title:

Heavy-tailed Bayesian nonparametric adaptation, direct and inverse problems

Abstract:

We will consider Bayesian nonparametric direct and inverse problems and we will be interested in studying the asymptotic performance of the posterior in the infinitely informative data limit, in terms of rates of contraction. We will be especially interested in priors which are adaptive to the smoothness of the unknown function, at least in the direct problem. In the last decade, certain hierarchical and empirical Bayes procedures based on Gaussian process priors, have been shown to achieve adaptation to smoothness. These procedures involve the tuning of a hyperparameter of the Gaussian prior. We will present a recently developed strategy for adaptation to smoothness based on heavy-tailed priors. We will first illustrate it in the direct setting, showing in particular that adaptive rates of contraction in the minimax sense (up to logarithmic factors) are achieved without tuning of any hyperparameters. We will then focus on inverse problems settings, under a generic local Lipschitz condition on the forward map. We will show that the studied heavy-tailed priors retain their attractive adaptation properties (at least) for forward problems, again without requiring the tuning of any hyperparameters. Numerical simulations corroborating the theory will be provided.

Keywords: rates of contraction; adaptation; inverse problems; heavy-tailed priors

Co-authors: Ismael Castillo

1.2. Randolph Altmeyer

Institution: Imperial College

Email: r.altmeyer@imperial.ac.uk

Talk title:

Bayesian nonparametrics for semi-linear SPDEs

Abstract:

We consider the nonparametric estimation of the reaction function in a semi-linear stochastic partial differential equation (SPDE) from observing a trajectory of the solution continuously over a finite time interval. Given a Gaussian process prior, we derive posterior contraction rates in a novel asymptotic regime: the spatial domain grows while the time horizon remains fixed. In this setting, the solution of the SPDE converges to a stationary process and is spatially ergodic. This allows for proving concentration inequalities of functionals along spatial averages of the solution. The proofs rely on the Clark-Ocone formula from Malliavin Calculus and precise bounds on the marginal densities of the SPDE.

The posterior contraction rates are minimax-optimal. We further prove a nonparametric Bernstein-von Mises Theorem for the posterior distribution.

Keywords: SPDE; spatial ergodicity; Bernstein-von-Mises; contraction rate; Gaussian process prior

Co-authors: Sascha Gaudlitz

1.3. Julyan Arbel

Institution: Inria Grenoble

Email: julyan.arbel@inria.fr

Website: <https://www.julyanarbel.com/>

Talk title:

Bayesian deep learning, overview and challenges

Abstract:

Bayesian deep learning is appealing as it combines the coherence and natural uncertainty quantification of the Bayesian paradigm together with the expressivity and compositional flexibility of deep neural networks. It has its roots in pioneering work by Radford Neal and David Mackay in the 1990s on Bayesian neural networks. Its strengths lie in principled uncertainty quantification, improved data efficiency, and adaptability, making it impactful in safety-critical fields like healthcare and autonomous systems. In this talk I will provide an overview of Bayesian deep learning and discuss some of the key challenges the field faces in addressing modern machine learning problems.

Keywords: Bayesian neural networks, Gaussian processes

Co-authors: Mariia Vladimirova, Vincent Fortuin, Kosta Pitas, Stephane Girard

1.4. Raffaele Argiento

Institution: University of Bergamo

Department: Economics

Email: raffaele.argiento@unibg.it

Talk title:

Reducing Dimensionality in Covariate-Informed Random Partition Models

Abstract:

Covariate-informed models for clustering, such as the Product Partition Model with covariates (PPMx), have proven effective in incorporating auxiliary information to improve clustering performance. However, in high-dimensional settings, selecting a low-dimensional subset of covariates becomes a central challenge: including irrelevant variables can distort clustering by overpowering the response signal and leading to overly granular or uninterpretable partitions. In this work, we introduce a novel approach based on Covariate Subset Selection (CSS). Although originally developed within computer science, CSS has recently been shown to admit a statistical formulation in which no specific assumptions are required for the selected covariates, while a multivariate regression model

is specified for the remaining covariates, conditional on the selected subset. Building on this foundation, we assign a Bayesian mixture model to the joint distribution of the response and the selected covariates. We show that the resulting model corresponds to a PPMx that automatically performs CSS as part of the clustering mechanism. This unified framework offers several advantages: it is probabilistically coherent, computationally efficient, and retains desirable properties such as Kolmogorov consistency, all while avoiding the complexities associated with reversible jump MCMC schemes. Simulation studies and real-data applications demonstrate the robustness and practical effectiveness of the proposed approach.

Keywords: Covariate Subset Selection (CSS); Product Partition Models (PPMx); Variable selection; Mixture modeling

Co-authors: Rosario Barone; Maria De Iorio; Michele Guindani

1.5. Sudipto Banerjee

Institution: University of California, Los Angeles

Email: sudipto@ucla.edu

Talk title:

Multivariate Gaussian Processes with Graphs

Abstract:

Multivariate Gaussian process models are widely used in spatial modeling for multiple dependent variables, where we wish to account for dependence over spatial locations and among the variables being studied. Such models require us to specify a cross-covariance kernel to ensure that the multivariate Gaussian process is well-defined. Directed acyclic graphs have played a significant role in constructing Gaussian processes for scaling inference to massive spatial datasets in recent times, but the use of undirected graphs in spatial process models is a relatively recent phenomenon. Recent scientific applications have posed modeling questions where the associations among the outcomes are modeled by an undirected graph that posits conditional dependence of a node only on adjacent nodes given all other nodes. The problem here is to construct a multivariate model where the nodes are spatial processes and the edge between two nodes signifies conditional dependence of a process on another given the rest. I discuss how to formulate Bayesian hierarchical models in this context using Graphical Gaussian Processes (GGPs) while also drawing some parallels with traditional Markov random fields used in finite-dimensional inference.

1.6. A. Felipe Barrientos

Institution: Florida State University

Department: Department of Statistics

Email: abarrientos@fsu.edu

Website: <https://anfebar.github.io/>

Talk title:

Bayesian Differential Privacy: Inference with Unknown Sample Size

Abstract:

We develop both theory and algorithms to analyze privatized data in unbounded differential privacy (DP), where even the sample size is considered a sensitive quantity that requires privacy protection. We show that the distance between the sampling distributions under unbounded DP and bounded DP goes to zero as the sample size n goes to infinity, provided that the noise used to privatize n is at an appropriate rate. We also establish that ABC-type posterior distributions converge under similar assumptions. To facilitate valid, finite-sample Bayesian inference on privatized data in the unbounded DP setting, we propose a reversible jump MCMC algorithm, extending the data augmentation MCMC of Ju et al. (2022). While the primary focus will be on Bayesian Differential Privacy within parametric models, the talk will also explore potential extensions to nonparametric models, particularly for the modeling of compositional data.

Keywords: Differential privacy; Unbounded differential privacy; Reversible jump MCMC

Co-authors: Jordan Awan and Nianqiao Ju

1.7. Anirban Bhattacharya

Institution: Texas A&M University

Department: Statistics

Email: anirbanb@stat.tamu.edu

Talk title:

Consistent DAG selection for Bayesian Causal Discovery under non-Gaussian errors

Abstract:

We consider the problem of learning the underlying causal structure between a set of variables, which are assumed to follow a linear acyclic structural equation model (SEM) with the associated errors being independent and allowed to be non-Gaussian. A Bayesian hierarchical model is proposed with the objective to identify the true data-generating directed acyclic graph (DAG) structure where the nodes and edges represent the variables and the directed causal effects, respectively. We develop model selection consistency results, backed by simulation studies.

Keywords: Bayesian; Model selection; Semiparametric; Structural Equation Model

Co-authors: Anamitra Chaudhuri, Yang Ni

1.8. Tamara Broderick

Institution: MIT

Email: tbroderick@mit.edu

Website: <https://tamarabroderick.com>

Talk title:

Double trouble: Predicting new variant counts across two heterogeneous populations

Abstract:

Collecting genomics data across multiple heterogeneous populations (e.g., across different cancer types) has the potential to improve our understanding of disease. Despite sequencing advances, though, resources often remain a constraint when gathering data. So it would be useful for experimental design if experimenters with access to a pilot study could predict the number of new variants they might expect to find in a follow-up study: both the number of new variants shared between the populations and the total across the populations. While many authors have developed prediction methods for the single-population case, we show that these predictions can fare poorly across multiple populations that are heterogeneous. We prove that, surprisingly, a natural extension of a state-of-the-art single-population predictor to multiple populations fails for fundamental reasons. We provide the first predictor for the number of new shared variants and new total variants that can handle heterogeneity in multiple populations. We show that our proposed method works well empirically using real cancer and population genetics data.

Keywords: beta process; genetic variants; power law; Poisson point process

Co-authors: Yunyi Shen; Lorenzo Masoero; Joshua G. Schraiber

1.9. Fan Bu

Institution: University of Michigan

Department: Department of Biostatistics

Email: fbu@umich.edu

Website: <https://fanbu1995.github.io/>

Talk title:

Hierarchical Bayesian Framework for Evidence Synthesis across Federated Data Networks

Abstract:

Evidence synthesis, or meta-analysis, is crucial for combining results from multi-site studies, especially when subject-level data cannot be shared. Traditional meta-analysis approaches relying on mixed effects models with normality assumptions are often inadequate when single sites have small sample sizes or rare outcome events. In addition, bias can arise due to systematic error when using observational data sources. To address both challenges, we propose a Bayesian evidence synthesis framework that performs likelihood-based meta-analysis and meanwhile corrects for bias. We introduce a Bayesian hierarchical model that jointly analyzes profile likelihoods instead of point estimates, which well accommodates small sample sizes and rare outcomes. Bias correction is achieved by learning an empirical bias distribution from a large number of negative control outcomes that are not associated with exposures of interest. We employ a hierarchical Dirichlet Process prior to flexibly characterize bias globally and locally while borrowing information across data sources. We implement a Hamiltonian Monte Carlo algorithm to efficiently

carry out posterior inference. Through a series of simulation studies and a case study on comparing treatments across a multi-national health data network, we demonstrate that our approach can provide more precise estimates and coherent results for improved interpretability.

Keywords: Meta-analysis; Federated learning; Health data science; Bayesian nonparametrics

Co-authors: Weixiong Hua

1.10. Federico Camerlenghi

Institution: University of Milano-Bicocca

Department: Department of Economics, Management and Statistics

Email: federico.camerlenghi@unimib.it

Website: <https://sites.google.com/unimib.it/camerlenghi-federico/>

Talk title:

Discrete random structures for feature sampling models: theory and applications

Abstract:

Species sampling models have a well-established history in Bayesian nonparametrics. In this context, a sample is typically a random draw from a population of units, referred to as species. Feature sampling models extend this framework by allowing each individual to belong to multiple species, now referred to as features. The introduction of the Indian buffet process has spurred interest in this class of models within the Bayesian community, also driven by a wide range of applications, including matrix factorization, ecology, and genomics. In this talk, we present a unified Bayesian framework for studying feature sampling models, leveraging discrete random structures. We highlight several examples: from classical approaches to the most recent contributions. In particular, we discuss priors based on the Poisson process and its transformations, as well as random structures that capture interactions, such as repulsion or attraction, among features and their associated weights. In addition, we explore the use of the proposed priors in different applied domains: in ecology, for biodiversity estimation; in cancer genomics, to analyze rare deleterious variants and design future genetic studies; in spatial statistics, for mapping trees in a forest. Finally, we provide a principled approach to prior elicitation based on predictive characterizations of the proposed models.

Keywords: Feature sampling models; Species sampling models; Indian buffet process; Discrete random structures; predictive inference

Co-authors: The talk is based on a series of joint works (published or in progress) with the following co-authors: M. Beraha, T. Broderick, L. Ghilotti, L. Masoero, T. Rigon, S. Favaro

1.11. Antonio Canale

Institution: University of Padova

Department: Department of Statistical Sciences

Email: antonio.canale@unipd.it

Website: <https://tonycanale.github.io>

Talk title:

Adaptive Partition Factor Analysis

Abstract:

Factor Analysis has traditionally been utilized across diverse disciplines to extrapolate latent traits that influence the behavior of multivariate observed variables. Historically, the focus has been on analyzing data from a single study, neglecting the potential study-specific variations present in data from multiple studies. Multi-study factor analysis has emerged as a recent methodological advancement that addresses this gap by distinguishing between latent traits shared across studies and study-specific components arising from artifactual or population-specific sources of variation. In this talk we extend the current methodologies by introducing novel shrinkage priors for the latent factors, thereby accommodating a broader spectrum of scenarios ranging from the absence of study-specific latent factors to models in which factors pertain only to small subgroups nested within or shared between the studies. We establish rigorous identifiability conditions for factor loadings and develop computationally efficient Gibbs sampling procedures for posterior inference. The proposed method is applied to a bird species co-occurrence dataset.

Keywords: factor analysis; neural networks; Shrinkage priors

Co-authors: Elena Bortolato

1.12. Lorenzo Cappello

Institution: Universitat Pompeu Fabra

Email: lorenzo.cappello@upf.edu

Talk title:

New (and old) predictive schemes with a.c.i.d. sequences

Abstract: There is a growing interest in procedures for Bayesian inference that bypass the need to specify a model and prior but rely on a predictive rule that describes how we learn on future observations given the available one. At the heart of the idea is a bootstrap-type scheme that allows us to move from the realm of prediction to inference. A key question is which conditions the predictive rule needs to satisfy to produce valid inference. In this work, we substantially relax previous assumptions building on a generalization of martingales, opening up the possibility of employing a much wider range of predictive rules that were previously ruled out. These include "old" ideas in Statistics and Learning Theory, such as kernel estimators, and more novel ones, such as the parametric Bayesian bootstrap or copula-based algorithms. Our aim is not to advocate in favor of one predictive rule versus the other but rather to showcase the benefits of working with this larger class of predictive rules.

Keywords: predictive inference; martingales; kernel methods, bootstrap

Co-authors: Marco Battiston (Lancaster University)

1.13. Ismael Castillo

Institution: Sorbonne University

Department: LPSM

Email: ismael.castillo@upmc.fr

Website: <https://perso.lpsm.paris/~castillo/>

Talk title:

Heavy-tailed priors for nonparametric adaptation

Abstract:

We introduce a Bayesian approach to derive adaptation to structural parameters such as smoothness, intrinsic dimensions or architecture, based on a prior distribution defined from two ingredients: a sequence of heavy-tailed distributions and a sequence of deterministic scaling parameters. Computationally, an advantage of this approach is that it does not require an explicit model selection step in the prior distribution. In particular, when sampling from the posterior, there is no need to sample from smoothness or structure hyperparameters. We discuss a number of applications for which we provide (near-) optimal contraction rates of the corresponding (possibly tempered-) posterior distributions, including nonparametric regression with series priors, ReLU deep neural networks, density estimation, classification, inverse problems and Besov classes. We also discuss trade-offs between tails of the prior and adaptation properties. Implementation of the method is considered and shown to lead to competitive results relative to state-of-the-art methods in a number of nonparametric and signal processing settings.

Keywords: Posterior contraction rates, heavy-tailed priors, adaptation, Bayesian deep neural networks

Co-authors: Sergios Agapiou, Paul Egels

1.14. Marta Catalano

Institution: Luiss University

Email: mcatalano@luiss.it

Talk title:

Measuring Partial Exchangeability with Reproducing Kernel Hilbert Spaces

Abstract:

In Bayesian multilevel models, the data are structured in interconnected groups, and their posteriors borrow information from one another due to prior dependence between latent parameters. However, little is known about the behavior of the dependence a posteriori. In this work, we develop a general framework for measuring partial exchangeability for parametric and nonparametric models, both a priori and a posteriori. We define an index that detects exchangeability, is invariant by reparametrization, can be estimated through samples, and, crucially, is well-suited for posteriors. We achieve these properties through the use of Reproducing Kernel Hilbert Spaces, which map any random probability to a random object on a Hilbert space. This leads to many convenient properties and tractable expressions, especially a priori and under mixing. We apply our general framework to

investigate the dependence a posteriori for the hierarchical Dirichlet process, retrieving a parametric convergence rate under very mild assumptions on the data.

Keywords: Bayesian Nonparametrics; Correlation; Hierarchical Dirichlet Process; Multilevel model; Reproducing Kernel Hilbert Space.

Co-authors: Hugo Lavenant, Francesco Mascari.

1.15. Noirrit Kiran Chandra

Institution: The University of Texas at Dallas

Department: Mathematical Sciences

Email: noirrit.chandra@utdallas.edu

Website: <https://sites.google.com/view/noirritchandra>

Talk title:

Bayesian Semiparametric Functional Mixed Effects Drift-Diffusion Models for Cognitive Control Leveraging Stop-signal Tasks

Abstract:

Cognitive control refers to a set of dynamic processes in the brain that encompasses stimulus-driven response mechanisms, goal-oriented decision-making, and the capacity to learn and adapt based on past decisions and their outcomes. Understanding the developmental progression of cognitive control in children and youth is crucial for identifying critical growth periods and ensuring optimal developmental outcomes. The Stop-signal task (SST) is a widely used paradigm to assess cognitive control in humans, where participants engage in a series of tasks and their behavioral metrics are recorded. Wiener drift-diffusion models (DDMs) are widely used to analyze behavioral data from SSTs as several cognitive psychology studies validated the interpretability of DDM parameters with respect to task conditions. We propose a comprehensive class of Bayesian semiparametric functional mixed drift-diffusion models for the behavioral metrics of SST. Cognitive control is inherently dynamic, unfolding across successive trials during task performance. Our proposed advancements in DDMs aim to incorporate these insights by accounting for the sustained and transient natures of proactive and reactive controls, respectively, incorporating performance monitoring as quantifiable functions, reflecting the dynamic interplay between the control mechanisms, and considering across-subject variability via nonparametric random effects. Our approach provides insights into how cognitive control evolves during task performance and how external factors influence these processes, critical aspects often overlooked in traditional cognitive research.

Keywords: DDM; functional data; mixed-effects model; neuroscience; Wiener process;

Co-authors: Farabi Raihan Shuvo, Stacie Warren

1.16. Shounak Chattopadhyay

Institution: University of California, Los Angeles

Email: shounak.chattopadhyay@ucla.edu

Talk title:

Efficient sampling in Bayesian nonparametric mixture models with soft multinomial approximations

Abstract:

Bayesian nonparametric (BNP) clustering approaches provide an elegant yet flexible framework to carry out model-based clustering. Posterior computation in such models typically involves Markov chain Monte Carlo (MCMC) algorithms such as Gibbs sampling, iteratively sampling the cluster membership variables and the model hyperparameters. Although straightforward to implement, Gibbs sampling can have poor computational scalability, particularly when the discrete cluster allocations exhibit mutual dependence. As an alternative, model-based soft-multinomial clustering is developed via a continuous approximation to the discrete multinomial variable signifying cluster membership. The use of soft-multinomials allows the implementation of general-purpose MCMC algorithms such as Hamiltonian Monte Carlo (HMC) to jointly sample the cluster memberships, model parameters, and associated hyperparameters. Theoretical results are provided detailing this approximation. We demonstrate substantial computational improvements using soft-multinomial clustering with HMC for posterior sampling over standard posterior sampling approaches in a variety of simulation examples and real-world data applications.

Keywords: Bayes; Nonparametrics; Mixture models; Clustering; Hamiltonian Monte Carlo

Co-authors: Marc A. Suchard

1.17. Riccardo Corradin

Institution: University of Milan-Bicocca

Email: riccardo.corradin@unimib.it

Talk title:

Autocompound Random Measures

Abstract:

In statistical analysis, it is common to observe quantities that depend on specific observational times. We introduce a class of time-dependent nonparametric models, suited for scenarios involving populations observed at distinct discrete times. Starting with an ancestral measure, it is possible to derive a time-specific random measure by acting on its intensity function, 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 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.

Keywords: compound random measures; time-dependent random measures; population dynamics;

Co-authors: Fabrizio Leisen

1.18. Andrea Cremaschi

Institution: IE University

Department: School of Science and Technology

Email: andrea.cremaschi@ie.edu

Website: <https://andcre87.github.io/andreacremaschi.github.io/>

Talk title:

Dependent Dirichlet-Multinomial Processes with Random Number of Components

Abstract:

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 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.

Keywords: Borrowing of information, Dependent processes, Clustering, Mixture of finite mixture, Partial exchangeability

Co-authors: Beatrice Franzolini

1.19. David Dahl

Institution: Brigham Young University

Department: Statistics

Email: dahl@stat.byu.edu

Website: <https://dahl.byu.edu>

Talk title:

Exchangeable Partition Distribution with Direct Prior Specification and an Entropy Parameter

Abstract:

We propose a unified framework for constructing exchangeable partition distributions that enables modelers to directly encode prior information on both the number of clusters and the distribution of cluster sizes, encompassing classical models such as the Chinese Restaurant Process (CRP) and the Hierarchical Uniform Partition (HUP) as special cases; the probability mass function is built in three hierarchical stages: first, the modeler specifies a marginal distribution for the number of clusters based on prior knowledge or expert opinion, or alternatively chooses one that matches the distribution implied by the CRP; second, given the number of clusters, we define the conditional distribution of cluster sizes, a flexible component that may reflect the behavior of the CRP or the HUP, and incorporate a novel tilt parameter that adjusts the relative probabilities of different cluster size configurations according to their cluster entropy, such that when the tilt parameter is zero, our model reduces to either the CRP or the HUP, negative tilt values favor configurations with lower entropy (i.e., more imbalanced cluster sizes), and positive tilt values favor configurations with higher entropy (i.e., more uniform cluster sizes), with the extreme limits inducing partitions of minimum or maximum entropy, respectively; and third, conditional on both the number of clusters and their sizes, a uniform distribution over the partitions is assumed, ensuring overall exchangeability. This framework offers enhanced flexibility by enabling direct elicitation of prior information on the number of clusters, the distribution of cluster sizes, and preferences regarding cluster entropy, thereby providing an intuitive and adaptable tool for model-based clustering in a variety of applications.

Keywords: Exchangeable Random Partitions; Chinese Restaurant Process; Hierarchical Uniform Partition; Cluster Entropy; Model-Based Clustering

Co-authors: Richard L. Warr (Brigham Young University); Talmage Hilton (Brigham Young University)

1.20. Michael Daniels

Institution: University of Florida

Department: Department of Statistics

Email: mdaniels@stat.ufl.edu

Talk title:

A new sparsity inducing prior for modelling longitudinal data in BART with application to G-computation and causal inference

Abstract:

We introduce a novel sparsity-inducing Dirichlet prior for modelling longitudinal data using BART. The prior exploits the temporal nature of conditioning on current and previous values of covariates and allows for simple posterior computations. We demonstrate the usefulness of the new BART prior in simulations, and compare its performance to other Bayesian decision tree ensemble approaches. The prior is motivated by a study of long-term memory effects of an incremental blood pressure intervention in a mortal cohort.

Keywords: BART; Causal inference; G-computation; longitudinal data

Co-authors: Maria Josefsson, Umea University (Sweden)

1.21. Pierpaolo De Blasi

Institution: University of Torino and Collegio Carlo Alberto

Email: pierpaolo.deblasi@unito.it

Website: <https://sites.google.com/a/carloalberto.org/pdeblasi/>

Talk title:

Speeding up the ordered allocation sampler

Abstract:

The ordered allocation sampler is a Gibbs sampler designed to explore the posterior distribution in nonparametric mixture models. It encompasses both infinite mixtures and finite mixtures with random number of components, and it has been shown to possess mixing properties that pair well with collapsed, or marginal, samplers that integrate out the mixing distribution. The main advantage is that it adapts to mixing priors that do not enjoy tractable predictive structures needed for the implementation of marginal sampling methods. Thus it is as widely applicable as other conditional samplers while enjoying better algorithmic performances. In this paper we provide a modification of the ordered allocation sampler that enhances its performances in a substantial way while easing its implementation. In addition, exploiting the similarity with marginal samplers, we are able to adapt to the new version of the sampler the split-merge moves of Jain and Neal. Simulation studies confirm these findings.

Keywords: Gibbs sampling; mixture models; ordered allocation sampler; species sampling models; split-merge moves

Co-authors: Maria F. Gil-Leyva and Fidel Selva, UNAM Mexico

1.22. Maria De Iorio

Institution: National University of Singapore

Email: mdi@nus.edu.sg

Talk title:

Latent Structure Learning in Multi-View Data

Abstract:

Mixture models are essential tools for analysing heterogeneous populations. From a Bayesian nonparametric perspective, we introduce a novel class of priors, the Normalised Independent Point Process, and develop both marginal and conditional algorithms for finite mixture models with a random number of components. By employing an auxiliary variable MCMC approach, we efficiently address the challenges posed by intractable posterior distributions, creating a flexible and extensible framework for Bayesian modelling. In this talk, we present key methodological advances for the analysis of multiview data, where complex dependencies arise across multiple data domains. To

address these challenges, we introduce a probabilistic framework for conditional partial exchangeability, specifically designed for multiview and longitudinal settings. This framework enables the construction of flexible random partitions that vary across features, allowing for nuanced modelling of dependencies within and between data domains. Moreover, the framework can be extended to link subject-level partitions across datasets via an underlying shared latent structure, promoting information sharing and facilitating robust, interpretable inference. By jointly modelling within-subject dependencies and marginal relationships across views, the approach enhances both the flexibility and interpretability of the resulting clustering structures. We also describe scalable Bayesian solutions, such as Bayesian Distance Clustering, which strike a balance between computational efficiency and probabilistic rigour. These methods are particularly effective in handling large, high-dimensional datasets, offering accurate predictive performance while maintaining scalability. We validate these methods through extensive simulations and applications to real-world datasets, demonstrating their effectiveness across a variety of contexts. These advancements provide a robust and versatile toolkit for addressing modern data analysis challenges, particularly in the context of multiview and longitudinal data.

1.23. Francesco Denti

Institution: University of Padua

Department: Department of Statistical Sciences

Email: francesco.denti@unipd.it

Website: <https://fradenti.github.io/>

Talk title:

BNP4BNP: Bayesian Nonparametric Models for Biomarkers and Neuronal Patterns

Abstract:

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, quantifying 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.

Keywords: Nested partitions; Biclustering; Spatial priors; Probit stick breaking; Potts model

Co-authors: C. Balocchi, A. Canale, G. Capitoli, L. D'Angelo, V. Denti, M. Guindani

1.24. Sameer Deshpande

Institution: University of Wisconsin, Madison

Department: Statistics

Email: sameer.deshpande@wisc.edu

Website: <https://skdeshpande91.github.io>

Talk title:

Building faster and more expressive BART models

Abstract:

Bayesian Additive Regression Trees (BART) is an easy-to-use and highly effective nonparametric regression model that approximates unknown functions with a sum of binary regression trees (i.e., piecewise-constant step functions) that one-hot encode categorical predictors. Consequently, BART is fundamentally limited in its ability to borrow strength across multiple levels of a categorical predictor, exploit structural relationships between multiple categorical predictors (e.g., adjacency and nesting), and to estimate smooth functions. I will describe several extensions of BART based on new priors for decision rules and leaf outputs that overcome these limitations. A recurring theme across these extensions is the use of simple Metropolis-Hastings proposals that permit linear-time regression tree updates, thereby facilitating much more flexible BART models without incurring much additional computational burden.

Keywords: Regression trees; network inference; machine learning; MCMC

Co-authors: None

1.25. David Dunson

Institution: Duke University

Email: dunson@duke.edu

Talk title:

Bayesian deep latent class regression

Abstract:

High-dimensional categorical data arise in diverse scientific domains and are often accompanied by covariates. Latent class regression models are routinely used in such settings, reducing dimensionality by assuming conditional independence of the categorical variables given a single latent class that depends on covariates through a logistic regression model. However, such methods become unreliable as the dimensionality increases. To address this, we propose a flexible family of deep latent class models. Our model satisfies key theoretical properties, including identifiability and posterior consistency, and we establish a Bayes oracle clustering property that ensures robustness against the curse of dimensionality. We develop efficient posterior computation methods,

validate them through simulation studies, and apply our model to joint species distribution modeling in ecology.

Keywords: Clustering; Discrete mixture; High-dimensional; Latent variables; Neural network

Co-authors: Yuren Zhou, Duke University; Yuqi Gu, Columbia University

1.26. Stefano Favaro

Institution: University of Torino and Collegio Carlo Alberto

Email: stefano.favaro@unito.it

Website: <https://www.carloalberto.org/person/stefano-favaro/>

Talk title:

A smoothed-Bayesian approach to frequency recovery from sketched data

Abstract:

We introduce a novel statistical perspective on a long-standing problem at the intersection of computer science and information theory: recovering the empirical frequency of a symbol in a large discrete dataset using only a compressed representation, or sketch, obtained via random hashing. Departing from traditional algorithmic approaches, recent works have proposed Bayesian nonparametric (BNP) methods that can provide more informative frequency estimates by leveraging modeling assumptions on the distribution of the sketched data. Here, we propose a smoothed-Bayesian method, inspired by existing BNP approaches but designed in a frequentist framework to overcome the computational limitations of the BNP approaches when dealing with large-scale data from realistic distributions, including those with power-law tail behaviors. For sketches obtained with a single hash function, our approach is supported by frequentist guarantees, including unbiasedness and optimality under a squared error loss function within a class of linear estimators. For sketches with multiple hash functions, we introduce an approach based on multi-view learning to construct computationally efficient frequency estimators.

Keywords: Frequency recovery; multi-view learning; smoothed-Bayesian estimation; normalized random measure; random hashing;

Co-authors: Mario Beraha (University of Milano - Bicocca); Emanuele Dolera (University of Pavia); Matteo Sesia (University of Southern California)

1.27. Michael Minyi Zhang

Institution: University of Hong Kong

Email: mzhang18@hku.hk

Talk title:

A Deep Bayesian Nonparametric Framework for Robust Mutual Information Estimation

Abstract:

Mutual Information (MI) is a crucial measure for capturing dependencies between variables, yet exact computation is challenging in high dimensions with intractable likelihoods, impacting accuracy and robustness. One idea is to use an auxiliary neural network to train an MI estimator; however, methods based on the empirical distribution function (EDF) can introduce sharp fluctuations in the MI loss due to poor out-of-sample performance, thereby destabilizing convergence. We present a Bayesian nonparametric (BNP) solution for training an MI estimator by constructing the MI loss with a finite representation of the Dirichlet process posterior to incorporate regularization in the training process. With this regularization, the MI loss integrates both prior knowledge and empirical data to reduce the loss sensitivity to fluctuations and outliers in the sample data, particularly in small sample settings like mini-batches. This approach addresses the challenge of balancing accuracy and low variance by effectively redu

Keywords: Dirichlet process, Information theory, Generative models, Bayesian nonparametric learning

Co-authors: Forough Fazeli Asl,, Bei Jiang, Linglong Kong

1.28. Edwin Fong

Institution: University of Hong Kong

Email: chefong@hku.hk

Talk title:

Martingale Posteriors for Structured Data

Abstract:

The martingale posterior is a framework for Bayesian inference where posterior uncertainty is generated through predictive imputation. As a result, the Bayesian model can be directly specified using a sequence of predictive distributions, bypassing the need for explicit likelihood and prior specifications. Extending this framework beyond the i.i.d. setting however presents challenges, particularly in eliciting predictive distributions that account for the inherent structure of the data. This talk outlines recent advancements in constructing nonparametric martingale posteriors for more complex settings, including quantile regression, dependent density estimation, and hierarchical models.

Keywords: Martingale posteriors; predictive inference; quantile estimation; hierarchical models

Co-authors: None

1.29. Catherine Forbes

Institution: Monash University

Department: Department of Econometrics and Business Statistics

Email: catherine.forbes@monash.edu

Website: <https://research.monash.edu/en/persons/catherine-forbes>

Talk title:

Moment-relevant sensitivities for Bayesian moment condition models

Abstract:

Moment condition models are popular as they offer flexibility in handling potentially complex models and data structures. Based on these model, semi-parametric estimation techniques with desirable properties have been proposed for use even when distributional assumptions are violated, or when facing endogeneity or non-linearities. Despite this, in finite samples, statistical analyses may be sensitive to small perturbations in the data which may suggest very different conclusions to that of the original data. In this paper we propose a general framework for constructing moment-relevant perturbations for the purpose of evaluating local influence in a Bayesian moment condition model context. The approach enables the analyst to explore potential moment-related frailties of a Bayesian empirical likelihood posterior given moment conditions and in light of available data. The information that results may help the analyst to better understand both the strengths and weaknesses of the moment condition model. The approach will be illustrated in a selection of empirically relevant examples.

Keywords: Empirical likelihood, local influence, perturbations, Kullback-Leibler divergence

Co-authors: Shelly Xie and David Frazier

1.30. Beatrice Franzolini

Institution: Bocconi University

Department: Bocconi Institute for Data Science and Analytics

Email: beatrice.franzolini@unibocconi.it

Website: <https://beatricefranzolini.github.io/>

Talk title:

Random Partitions for Repeated Measures Design in Multi-View and Network Data

Abstract:

Recent advances in Bayesian nonparametrics have extended traditional exchangeable partition models to settings with complex dependencies. This talk first introduces conditional partial exchangeability (CPE), a dependence condition that, through its conditional nature and requirement for marginal invariance, captures local dependence in multivariate partition models. It then explores how CPE unifies models for multi-view and dynamic clustering in repeated measurement designs and can be used to extend stochastic block models for analyzing multiplex networks, where it allows inferring both layer-specific and global block structures. By aligning symmetry assumptions in the model with the experimental design of the data, our approach provides a coherent framework for modeling diverse and highly structured data.

Keywords: Clustering; Exchangeability; Network data; Random Partitions;

Co-authors: Maria De Iorio, Daniele Durante; Valentina Ghidini

1.31. David Frazier

Institution: Monash University

Department: Business Statistics

Email: david.frazier@monash.edu

Talk title:

The Statistical Accuracy of Neural Posterior and Likelihood Estimation

Abstract:

Neural posterior estimation (NPE) and neural likelihood estimation (NLE) are machine learning approaches that provide accurate posterior, and likelihood, approximations in complex modeling scenarios, and in situations where conducting amortized inference is a necessity. While such methods have shown significant promise across a range of diverse scientific applications, the statistical accuracy of these methods is so far unexplored. In this manuscript, we give, for the first time, an in-depth exploration on the statistical behavior of NPE and NLE. We prove that these methods have similar theoretical guarantees to common statistical methods like approximate Bayesian computation (ABC) and Bayesian synthetic likelihood (BSL). While NPE and NLE methods are just as accurate as ABC and BSL, we prove that this accuracy can often be achieved at a vastly reduced computational cost, and will therefore deliver more attractive approximations than ABC and BSL in certain problems. We verify our results theoretically and in several examples from the literature.

Keywords: Posterior Estimation; Intractable Models; Normalizing Flows; Posterior Concentration Rate

Co-authors: Ryan Kelly Christopher Drovandi David J. Warne

1.32. Francesco Gaffi

Institution: University of Maryland

Email: fgaffi@umd.edu

Talk title:

Exchangeable random permutations for graph matching

Abstract:

The graph matching problem finds applications in diverse fields such as computer vision, pattern recognition, bioinformatics, and social network analysis. Observing the realizations of two sets of edges insisting on two copies of the same set of nodes, one wants to reconstruct the unknown permutation that matches the nodes across graphs by minimizing edge discrepancies. The cardinality of the space of permutations introduces significant computational challenges, making classical maximum likelihood methods computationally intractable and uncertainty quantification difficult to explore. A fully integrated Bayesian framework for graph matching is largely absent. To fill this gap, we develop a general theory of exchangeable random permutations, which can serve as priors for statistical graph matching and other permutation-based inferential problems. Building on the literature on random partitions, we define and characterize exchangeable random

permutations, derive their predictive structure, assess their projectivity, and introduce the position-aware Chinese restaurant process. The cornerstone of our proposal is to combine the matching with a latent partition structure that captures nodes' connectivity behavior: since the optimal matching is commonly inferred maximizing similarities in connectivity patterns, we assume the cycle structure of the matching permutation to coincide with a latent stochastic block structure. This approach enables simultaneous inference of a partition encoding connectivity behavior and a permutation minimizing edge discrepancies. Posterior inference is conducted via a node-wise blocked Gibbs sampler tailored to the structure of the novel class of priors. Moreover, we propose a two-step procedure for summarizing the posterior sample: first a variation-of-information averaging fixes the cycle structure, then a cycle-wise minimization of the Cayley distance selects a point estimate. Computational complexity is reduced through a pruning procedure and a tree-structured search, that can be also employed for uncertainty quantification. Finally, the predictive properties of our priors are leveraged to perform online matching tasks.

Keywords: Bayesian nonparametrics, unaligned multiplex network, position-aware Chinese restaurant process, correlated stochastic block model

Co-authors: Nathaniel Josephs and Lizhen Lin

1.33. Subhashis Ghoshal

Institution: North Carolina State University

Department: Statistics

Email: sghosal@ncsu.edu

Website: <https://www4.stat.ncsu.edu/~sghosal/>

Talk title:

Bayesian Learning of Relational Graph in Semiparametric High-dimensional Time Series

Abstract:

Time series data arising in many applications nowadays are high-dimensional. A large number of parameters describe features of these time series. Sensible inferences on these parameters with limited data are possible if some underlying lower-dimensional structure is present. We propose a novel approach to modeling a high-dimensional time series through several independent univariate time series, which are then orthogonally rotated and sparsely linearly transformed. With this approach, any specified intrinsic relations among component time series given by a graphical structure can be maintained at all time snapshots. We call the resulting process an Orthogonally-rotated Univariate Time series (OUT). The OUT model can easily accommodate key structural properties of time series, such as stationarity and causality. For Bayesian inference, we put suitable prior distributions on the spectral densities of the independent latent times series, the orthogonal rotation matrix, and the common precision matrix of the component times series at every time point. A likelihood is constructed using the Whittle approximation for univariate latent time series. An efficient Markov Chain Monte Carlo (MCMC) algorithm is developed for posterior computation. We study the convergence of the pseudo-posterior distribution based on the Whittle likelihood for the model's parameters upon developing a new general

posterior convergence theorem for pseudo-posteriors. We find that the posterior contraction rate for independent observations essentially prevails in the OUT model under very mild conditions on the temporal dependence described in terms of the smoothness of the corresponding spectral densities. While establishing the result, we develop a new general theorem on the contraction rate of a pseudo-posterior distribution that is potentially applicable in other situations.

Keywords: Contraction rate, Graphical model, Pseudo-posterior, Time series, Whittle likelihood

Co-authors: Arkaprava Roy, University of Florida, and Anindya Roy, University of Maryland-Baltimore County.

1.34. Maria F. Gil Leyva Villa

Institution: IIMAS-UNAM

Department: Department of Probability and Statistics

Email: marifer@sigma.iimas.unam.mx

Website: <https://marifergil.github.io>

Talk title:

On a new Stick-breaking characterization of Pitman-Yor processes

Abstract:

Stick-breaking is one of the most popular and intuitive constructions of Bayesian nonparametric (BNP) priors. It consists in expressing the weights of a random discrete distribution in terms of a sequence of $[0,1]$ -valued random variables, which we term length variables. This construction yields particularly tractable BNP models when the stick-breaking weights are invariant under size-biased permutations. In this case, key quantities such as the exchangeable partition probability function (EPPF), the predictive distribution and some posterior distributions have closed form expressions. Pitman (1996) showed that, under mild conditions, the Pitman-Yor process is the only stick-breaking process with independent length variables that remains invariant under size-biased permutations of the weights. Here we generalize this result by relaxing the independence assumption of the length variables and requiring a Markov symmetry instead. Thus providing a novel characterization of the Pitman-Yor process.

Keywords: Stick-breaking; Size-biased permutation; Pitman-Yor process; Markov process; Distributional invariance

Co-authors: Antonio Lijoi, Ramses Mena and Igor Prunster.

1.35. Matteo Giordano

Institution: University of Turin

Department: ESOMAS Deptment

Email: matteo.giordano@unito.it

Website: <https://matteogiordano.weebly.com>

Talk title:

Statistical algorithms for low-frequency diffusion data: A PDE approach

Abstract:

The talk will consider the problem of making nonparametric inference in a class of multi-dimensional diffusions in divergence form, from low-frequency data. Statistical analysis in this setting is notoriously challenging due to the intractability of the likelihood and its gradient. We propose a new computational approach which is motivated by PDE theory and is built around the characterisation of the transition densities as solutions of the associated heat equation. Employing optimal regularity results from the theory of parabolic PDEs, we prove a novel characterisation for the gradient of the likelihood. Using these developments, for the nonlinear inverse problem of recovering the diffusivity, we then show that the numerical evaluation of the likelihood and its gradient can be reduced to standard elliptic eigenvalue problems, solvable by powerful finite element methods. This enables the efficient implementation of a large class of popular statistical algorithms, including (i) preconditioned Crank-Nicolson and Langevin-type methods for posterior sampling, and (ii) gradient-based descent optimisation schemes to compute maximum likelihood and maximum-a-posteriori estimates. Joint work with Sven Wang.

Keywords: Diffusion process; Gaussian priors; MAP estimator; Markov chain Monte Carlo; Spectral methods

Co-authors: Sven Wang (Humboldt University Berlin)

1.36. Clara Grazian

Institution: University of Sydney

Department: School of Mathematics and Statistics Email: clara.grazian@sydney.edu.au

Website: <https://www.sydney.edu.au/science/about/our-people/academic-staff/clara-grazian.html>

Talk title:

Consistency Results on Estimating the Number of Clusters for the Pitman-Yor Process

Abstract:

Recently, there has been interest in the literature to provide consistency results for estimating the number of components in mixture models, focusing on approaches based on Dirichlet and Pitman-Yor process priors. While Dirichlet process models are widely used in nonparametric Bayesian clustering, their ability to recover the true number of mixture components remains a subject of ongoing research. We explore theoretical and empirical findings on the asymptotic behavior of these models and discuss key conditions under which they yield consistent estimates. Building on this foundation, we demonstrate that Pitman-Yor process models, when equipped with a uniform prior on the concentration parameter, can serve as a consistent estimator for the number of components in a mixture model under certain conditions. Our results provide insights into how the choice of prior affects model flexibility and consistency, contributing to a deeper understanding of Bayesian nonparametric mixture modeling.

Keywords: Pitman-Yor process; clustering; consistency

Co-authors: Thomas Zheng (University of Sydney)

1.37. Yuqi Gu

Institution: Columbia University

Department: Department of Statistics

Email: yuqi.gu@columbia.edu

Talk title:

Deep Discrete Encoders: Identifiable Deep Generative Models for Rich Data with Discrete Latent Layers

Abstract:

In the era of generative AI, deep generative models (DGMs) with latent representations have gained tremendous popularity. Despite their impressive empirical performance, the statistical properties of these models remain underexplored. DGMs are often overparametrized, non-identifiable, and uninterpretable black boxes, raising serious concerns when deploying them in high-stakes applications. Motivated by this, we propose an interpretable deep generative modeling framework for rich data types with discrete latent layers, called Deep Discrete Encoders (DDEs). A DDE is a directed graphical model with multiple binary latent layers. Theoretically, we propose transparent identifiability conditions for DDEs, which imply progressively smaller sizes of the latent layers as they go deeper. Identifiability ensures consistent parameter estimation and inspires an interpretable design of the deep architecture. Computationally, we propose a scalable estimation pipeline of a layerwise nonlinear spectral initialization followed by a penalized stochastic approximation EM algorithm. This procedure can efficiently estimate models with exponentially many latent components. Extensive simulation studies validate our theoretical results and demonstrate the proposed algorithms' excellent performance. We apply DDEs to three diverse real datasets for hierarchical topic modeling, image representation learning, response time modeling in educational testing, and obtain interpretable findings.

Keywords: Identifiability; Interpretable Artificial Intelligence; Representation Learning; Deep Belief Network; Directed Graphical Model

Co-authors: Seunghyun Lee

1.38. Rajarshi Guhaniyogi

Institution: Texas A&M University

Email: rajguhaniyogi@tamu.edu

Talk title:

Amortized Inference for Large Structured Data

Abstract:

The rapid expansion of large, structured datasets has brought forth both valuable opportunities and formidable challenges for statistical inference. This talk focuses on two pressing scientific problems emerging in this context: (1) building functional surrogates for

high-fidelity computer simulations, and (2) predicting expensive or less reliable images using low-cost, high-quality alternatives. Although Bayesian methods offer a rigorous framework for uncertainty quantification, they often fall short in terms of scalability when applied to high-dimensional and large-scale datasets. To overcome these limitations, we propose deep neural network (DNN)-based generative models designed for (a) functional outputs with vector-valued inputs, and (b) functional outputs with functional- and network-valued inputs. These models are capable of capturing intricate dependencies between heterogeneous input and output structures, while enabling interpretability through well-calibrated uncertainty estimates and maintaining computational efficiency in large-sample, high-resolution settings. This work contributes to the broader domain of interpretable machine learning. We demonstrate the effectiveness of the proposed methods in two motivating applications: large-scale computer simulations, and a neuroimaging study aimed at predicting task-based brain activation from resting-state fMRI networks and structural MRI data. This research is conducted in collaboration with colleagues from Los Alamos National Laboratories and the University of California, San Francisco.

Keywords: Deep neural networks; functional data; uncertainty quantification; large structured data; efficient computation

1.39. Luis Gutierrez

Institution: Pontificia Universidad Catolica de Chile

Department: Department of Statistics

Email: lgutieino@gmail.com

Website: <https://statistics-luisgutierrez.com/>

Talk title:

Bayesian hypothesis testing using flexible models

Abstract:

Using statistical tests to test hypotheses is still the basis for answering research questions in medicine, public health, biology, agronomy, and other fields. The statistical literature on hypothesis testing is focused on models with somewhat restrictive assumptions that do not fit well with databases derived from modern problems. In this talk, we will discuss a general strategy for constructing hypothesis tests under a Bayesian approach of inference using flexible models. Such a strategy is based on identifying each possible hypothesis with a binary vector that allows the definition of a prior distribution over the hypotheses space in a straightforward manner. Given a particular hypothesis, the data are assumed to be generated by a nonparametric Bayesian model that provides flexibility and allows the identification of differences between treatment levels beyond location. Strategies for posterior inference that enable exploration of the model space are discussed. Finally, applications in the context of ANOVA, MANOVA, and Granger causality are presented.

Keywords: Dependent Dirichlet process; Geometric Stick-breaking process; Vector autoregressive process; Partial Exchangeability; Multiple testing

Co-authors: Ivan Gutierrez, Danilo Alvares

1.40. Matthew Heiner

Institution: Brigham Young University

Department: Statistics

Email: heiner@stat.byu.edu

Website: <https://heiner.byu.edu/>

Talk title:

Partially Geometric Stick-Breaking Processes

Abstract:

Empirical distributions for count data often exhibit idiosyncrasies among low values, such as zero inflation, and stable tail behavior. We propose a flexible model for counts that combines nonparametric estimation of early probabilities with fixed decay after a single, unknown change point. We use a stick-breaking construction with variables that are beta-distributed before the change point and share a single value thereafter, inducing a geometric tail. The resulting process model parsimoniously balances needed flexibility where data are abundant, with a parametric representation where data are naturally sparse. The construction admits a collapsed posterior distribution for the change point, avoiding transdimensional MCMC. We illustrate by modeling rally lengths in men's professional tennis, where the change point may indirectly measure an effect of server advantage. Finally, we use the new process to construct a countably infinite mixture model, extending the geometric stick-breaking process of Fuentes-García, Mena, and Walker, and demonstrate its effectiveness for density estimation. The random change point accommodates a variety of behaviors, ranging from a parsimonious approximation of Poisson-Dirichlet process mixtures to mixtures of finite mixtures.

Co-authors: None

1.41. Amy Herring

Institution: Duke University

Department: Statistical Science

Email: amy.herring@duke.edu

Website: <https://scholars.duke.edu/person/Amy.Herring>

Talk title:

From the Back Yard to the Parlor: Collaboration-Inspired Methodology

Abstract:

As thousands of graduate school applicants will attest, many choose statistical science, at least in part, because of our curiosity and the desire to "play in everyone's back yard." Numerous successful senior researchers note their methods research is driven by their collaborations -- while the "why" may be obvious, the "how" is often mysterious. Using a collaborative project, the Sepsis Characterization in Kilimanjaro (SICK) study, we describe our original analysis goals and plans, unforeseen challenges, resulting methodological

innovations, and opportunities still being explored. The goal is to demystify the process by which questions beget questions that beget questions....

Keywords: clustering; prior distributions for clustering; dimension reduction; collaboration

Co-authors: Alex Dombowsky Maoran Xu Federica Stolf Lorenzo Mauri David Dunson

1.42. Nhat Ho

Institution: The University of Texas, Austin

Department: Statistics and Data Sciences

Email: minhnhhat@utexas.edu

Website: <https://nhatptnk8912.github.io/>

Talk title:

Data-Driven DRO and Economic Decision Theory: An Analytical Synthesis With Bayesian Nonparametric Advancements

Abstract:

We develop an analytical synthesis that bridges data-driven Distributionally Robust Optimization (DRO) and Economic Decision Theory under Ambiguity (DTA). By reinterpreting standard regularization and DRO techniques as data-driven counterparts of ambiguity-averse decision models, we provide a unified framework that clarifies their intrinsic connections. Building on this synthesis, we propose a novel DRO approach that leverages a popular DTA model of smooth ambiguity-averse preferences together with tools from Bayesian nonparametric statistics. Our baseline framework employs Dirichlet Process (DP) posteriors, which naturally extend to heterogeneous data sources via Hierarchical Dirichlet Processes (HDPs), and can be further refined to induce outlier robustness through a procedure that selectively filters poorly-fitting observations during training. Theoretical performance guarantees and convergence results, together with extensive simulations and real-data experiments, illustrate the method, and its favorable performance in terms of prediction accuracy and stability.

Keywords: Data-Driven Optimization, Distributionally Robust Optimization, Economic Decision Theory, Ambiguity Aversion, Bayesian Nonparametrics.

Co-authors: Nicola Barileto, Khai Nguyen

1.43. Maria Kalli

Institution: King's College London

Department: Mathematics

Email: maria.kalli@kcl.ac.uk

Website: <https://www.kcl.ac.uk/people/maria-kalli>

Talk title:

Bayesian nonparametric models with BART components

Abstract:

BART has emerged as an important method for Bayesian nonlinear regression. However, the need for a fully nonparametric version of BART has been investigated in previous work (see e.g. George et al, 2019). In this talk, I will describe Bayesian nonparametric models built using BART elements and which can be easily implemented using existing BART R packages. The most general density regression model uses an NCoRM mixture (Griffin and Leisen, 2018) where the components are heteroscedastic BART models. The performance of the method will be evaluated on simulated and real data. References E. I. George, P. Laud, B. Logan, R. E. McCulloch and R. Sparapani (2019) "Fully Nonparametric Bayesian Additive Regression Trees." In Topics in Identification, Limited Dependent Variables, Partial Observability, Experimentation, and Flexible Modeling: Part B, Volume 40B (eds. I. Jeliaskov and J. L. Tobias), Emerald: Leeds, U.K. J. E. Griffin and F. Leisen (2018) "Modelling and Computation Using NCoRM Mixtures for Density Regression." Bayesian Analysis, 13, 897-916. M. T. Pratola, H. A. Chipman, E. I. George and R. E. McCulloch (2020) "Heteroscedastic BART via Multiplicative Regression Trees." Journal of Computational and Graphical Statistics, 29, 405-417.

Keywords: Bayesian nonparametrics, Density regression, infinite mixtures, BART

Co-authors: Jim Griffin

1.44. Paul Kirk

Institution: University of Cambridge

Department: MRC Biostatistics Unit

Email: paul.kirk@mrc-bsu.cam.ac.uk

Website: <https://www.mrc-bsu.cam.ac.uk/staff/paul-kirk>

Talk title:

Federated Variational Inference for Bayesian Mixture Models

Abstract:

I will present a one-shot, unsupervised federated learning approach for Bayesian clustering of large-scale binary and categorical datasets, motivated by the need to identify patient clusters in privacy-sensitive electronic health record (EHR) data. We introduce a principled 'divide-and-conquer' inference procedure using variational inference with local merge and delete moves within batches of the data in parallel, followed by 'global' merge moves across batches to find global clustering structures. These merge moves require only summaries of the data in each batch, enabling federated learning across local nodes without requiring the full dataset to be shared. Empirical results on simulated and benchmark datasets demonstrate that our method performs well relative to comparator clustering algorithms. We validate the practical utility of the method by applying it to a large-scale British primary care EHR dataset to identify clusters of individuals with common patterns of co-occurring conditions (multimorbidity).

Keywords:

Federated learning; Divide and conquer; Distributed computing; Clustering

Co-authors: Jackie Rao, Francesca L Crowe, Tom Marshall, Sylvia Richardson

1.45. Athanasios Kottas

Institution: University of California, Santa Cruz

Department: Statistics

Email: thanos@soe.ucsc.edu

Talk title:

Nonparametric Markov renewal processes for earthquake modeling

Abstract:

We will present a nonparametric Bayesian modeling approach for temporal point patterns recorded along with discrete marks. The stochastic model builds from Markov renewal processes, which describe characteristics of recurring events observed along with discrete state labels. The motivation stems from flexible modeling for transition-specific distributions of inter-arrival times between earthquake occurrences, where the transitions are among earthquake magnitude states. A key modeling objective is to allow dependence among the inter-arrival time distributions, thus extending the stochastic model structure of Markov renewal processes. This property of the prior probability model is particularly relevant to seismology applications, and, in general, it can bolster estimation efficiency for scenarios where some transition cases consist of a small number of events. We develop a mixture model for the transition-specific inter-arrival time densities, placing emphasis on the choice of both the mixture kernel densities and mixture weights. The kernel density options ensure connections with renewal process properties, as well as domain-relevant decreasing inter-arrival densities. We construct transition-specific mixture weights through a hierarchical structure which enables efficient posterior simulation and general dependence between inter-arrival time distributions. The inferential scope of the modeling approach will be illustrated using a point pattern of earthquakes from the Southern California Earthquake Data Center.

Keywords: Bayesian nonparametrics; Logit stick-breaking process; Markov chain Monte Carlo; Point processes.

Co-authors: W. Zachary Horton

1.46. Jaeyong Lee

Institution: Seoul National University

Email: jlee@snu.ac.kr

Talk title:

Conditional Dirichlet Processes and Functional Condition Models

Abstract:

n this talk, we consider the conditional Dirichlet process (cDP) when a functional of

1.47. Juho Lee

Institution: KAIST

Department: Kim Jaechul Graduate School of AI

Email: juholee@kaist.ac.kr

Website: <https://juho-lee.github.io/>

Talk title:

Bayesian nonparametric models in the era of foundation models

Abstract:

a random distribution is given. We specifically apply the cDP to the functional condition

Keywords: Foundation models; fine-tuning; data-driven priors; uncertainty estimation

Co-authors: Hyungi Lee, Giung Nam, Chaeyun Jang, Dong Bok Lee, Edwin Fong

1.48. Cheng Li

Institution: National University of Singapore

Department: Department of Statistics and Data Science

Email: stalic@nus.edu.sg

Website: <https://blog.nus.edu.sg/stalic/>

Talk title:

Radial Neighbors for Provably Accurate Scalable Approximations of Gaussian Processes

Abstract:

model. This model is a nonparametric model in which the finite dimensional parameter of

Keywords: Approximations; Directed acyclic graph; Gaussian process; Spatial statistics; Wasserstein distance

Co-authors: Yichen Zhu, Michele Peruzzi, David B. Dunson

1.49. Didong Li

Institution: University of North Carolina at Chapel Hill

Department: Biostatistics

Email: didongli@unc.edu

Website: <https://sites.google.com/view/didongli/>

Talk title:

Parameter Inference for Gaussian Processes

Abstract:

interest is defined as a solution to a functional equation of the distribution. We derive both

Keywords: Identifiably, interpretability, truncation, manifold

Co-authors: Ameer Qaqish

1.50. Meng Li

Institution: Rice University

Department: Statistics

Email: meng@rice.edu

Website: <https://meng.rice.edu>

Talk title:

Just plug in: optimal Gaussian process modeling for rate of change

Abstract:

the posterior distribution of the parameter of interest and that of the underlying distribution

Keywords: derivatives; Gaussian process; posterior construction; climate change; plug-in

Co-authors: Zejian Liu

1.51. Antonio Lijoi

Institution: Bocconi University

Email: antonio.lijoi@unibocconi.it

Talk title:

A Bayesian nonparametric approach to estimation and prediction with competing risks.

Abstract:

itself. We investigate the properties of the functional condition model with cDP and propose

Keywords: Completely random measures, Hierarchical processes, Prediction curve, Random Partition

Co-authors: Claudio Del Sole, Igor Pruenster

1.52. Antonio Linero

Institution: University of Texas at Austin

Department: Statistics and Data Sciences

Email: antonio.linero@austin.utexas.edu

Website: <https://theodds.github.io>

Talk title:

Challenges and Opportunities for Bayesian Nonparametrics in Causal Inference

Abstract:

In this talk we will explore the intersection of Bayesian nonparametric methods and causal inference, highlighting the unique ways in which the Bayesian paradigm is suited to estimating causal effects. In particular, appropriately-designed Bayesian nonparametric priors have empirically performed very well in settings relevant to the real world, where typically effect sizes are small, noise levels are large, and treatment effects are relatively homogeneous. It is very easy to go wrong, however, and we illustrate this in several examples. First, simply plugging in existing Bayesian nonparametric priors in a naive "causally agnostic" fashion can easily encode absurd prior beliefs about selection bias parameters. Second, when inferring heterogeneity, there is some inherent tension between Bayesian methods and Frequentist operating characteristics due to how Bayesian methods address post-selection inference. Ultimately, we will argue that the strengths of the Bayesian nonparametric methods arise from genuine engagement with the underlying philosophy: when our priors are modeled to match genuine prior beliefs, Bayesian nonparametric methods can offer very large practical advantages over competing causal frameworks like debiased machine learning or targeted maximum likelihood estimation.

Keywords: Nuisance parameters, regularization, BART, post-selection inference, subgroup identification

Co-authors: Entejar Alam, Poorbita Kundu, Angela Ting

1.53. Li Ma

Institution: Duke University

Department: Statistical Science

Email: li.ma@duke.edu

Website: <https://lm186.github.io>

Talk title:

Generative models, two-sample problem, and additive tree ensembles

Abstract:

Generative models have become a popular tool for simulating observations from complex, high-dimensional distributions, and they share a close connection with Bayesian nonparametrics (BNP). Many popular BNP models correspond to specific generative processes, although the traditional focus of BNP has been on inferring the underlying distribution or function and quantifying the associated uncertainty. However, effectively assessing the quality of generative models remains a challenge. Standard approaches, such as reporting empirical distance estimates between real and generated samples, generally do not reveal the nature of the discrepancy. Nor do they allow for reliable uncertainty quantification. At its core, this is a two-sample problem involving high-dimensional, nonparametric objects—a setting that has been studied in the BNP literature.

We show that a powerful approach to understanding such discrepancies is to learn the density ratio between the generative model and the true data-generating model. We demonstrate several strategies for characterizing the discrepancy once reliable density ratio estimates and Bayesian uncertainty quantification become available. To this end, we construct a reliable estimator for the density ratio. Our starting point is a loss function for learning density ratios under a two-sample design, based on the variational form of the Hellinger distance between the two sample densities. Drawing a connection between this loss and the exponential loss commonly used in training additive tree ensembles for supervised learning, we show that the new loss can be optimized by fitting a sequence of weak tree learners that iteratively maximize the Hellinger distance between two sets of pseudo-observations, constructed by reweighting the original observations at each iteration.

This leads to a class of tree boosting algorithms that yields approximate M-estimators of the underlying density ratio. We then exploit the resemblance between the Hellinger-based loss and an exponential family kernel to design conjugate priors on the coefficients of each weak tree learner, under a loss-based generalized Bayesian framework. This enables us to construct a BART-like Bayesian backfitting sampler targeting the Gibbs posterior of the resulting tree ensemble. In this way, we achieve generalized Bayesian uncertainty quantification for the density ratio between the two samples. We illustrate applications of this approach in the context of generative models for microbiome compositional data.

Keywords: Generalized Bayes, trees, ensemble models, additive models, generative models

Co-authors: Naoki Awaya, Waseda University, nawaya@waseda.jp Yuliang Xu, Duke University, yuliang.xu@duke.edu

1.54. John Maheu

Institution: McMaster University

Department: DeGroote School of Business

Email: maheujm@mcmaster.ca

Talk title:

Equity vs Bonds, What Happened During COVID-19? A Bayesian Nonparametric Analysis

Abstract:

an algorithm for the posterior inference.

Keywords: Asset Allocation; Dirichlet Process Mixture; COVID-19 Market Impact; Multivariate Density Estimation; Multivariate GARCH

Co-authors: Hamidreza Khoshro

1.55. Eleni Matechou

Institution: Queen Mary University of London

Department: School of Mathematical Sciences

Email: e.matechou@qmul.ac.uk

Talk title:

Applications of Bayesian nonparametric models to ecological data

Abstract:

Ecological surveys often track individuals from wildlife populations using methods such as traps or nets, aiming to estimate population size and to monitor processes such as migration patterns, behavioural change, or life state transitions. These processes are typically complex, latent, and vary over time and between individuals. Over the last decade, we have developed a suite of flexible Bayesian nonparametric models tailored to such data, building on the Dirichlet process (DP), the Hierarchical DP, the Dependent DP, and Pólya trees. These models account for complex observation mechanisms and enable inference on the latent ecological processes of interest. This talk will present models for several common survey types—including capture-recapture, count, and batch-mark surveys—and demonstrate how the models describe complex latent patterns and corresponding observation processes, solving open inference problems in the field.

Keywords: ecological surveys, capture-recapture, count data, batch-mark, population size

Co-authors: François Caron, Jim Griffin, Alex Diana, Raffaele Argiento, Ioannis Rotous

1.56. Steve McEachern

Institution: The Ohio State University

Email: snm@stat.osu.edu

Talk title:

Bayesian nonparametrics – from theoretical foundations to modelling

Abstract:

Bayesian nonparametrics provides a coherent platform for statistical inference and decision-making. The literature, from Ferguson and Blackwell and McQueen's early work on, develops a rich collection of models for probability measures (distributions). These models have progressed to those for single distributions to collections of distributions. The progression has been accompanied by the development of a remarkable collection of theoretical results and a variety of computational implementations. The models have found widespread application across many disciplines. This talk, with a focus on the canonical problem in statistics (regression), emphasizes the importance of a framework for modelling efforts. The framework suggests directions for further development of models and provides novel perspectives on existing work.

1.57. Ramses Mena

Institution: Universidad Nacional Autonoma de Mexico

Department: IIMAS

Email: ramses.mena@gmail.com

Website: <https://sigma.iimas.unam.mx/ramses/>

Talk title:

Fast computation of means of Gibbs-type priors

Abstract:

Leveraging the symmetry of exchangeable sequences driven by Gibbs-type priors, we derive and solve a system of partial difference equations that characterize the moments of random means. These results enable the efficient computation of these quantities and provide rapid estimations, facilitating the elucidation and inference of Bayesian non-parametric models. An application and illustrative examples using log-returns of financial assets will be presented.

Keywords: Exact computation; Posterior inference; Random means;

Co-authors: Ilia Naumkin

1.58. Jeff Miller

Institution: Harvard University

Department: Biostatistics

Email: jwmiller@hsph.harvard.edu

Website: <https://jwmi.github.io>

Talk title:

Bayesian model criticism using uniform parametrization checks

Abstract:

Models are often misspecified in practice, making model criticism a key part of Bayesian analysis. It is important to detect not only when a model is wrong, but which aspects are wrong, and to do so in a computationally convenient and statistically rigorous way. We introduce a novel method for model criticism based on the fact that if the parameters are drawn from the prior, and the dataset is generated according to the assumed likelihood, then a sample from the posterior will be distributed according to the prior. Thus, departures from the assumed likelihood or prior can be detected by testing whether a posterior sample could plausibly have been generated by the prior. Building upon this idea, we propose to reparametrize all random elements of the likelihood and prior in terms of independent uniform random variables. This makes it possible to aggregate across arbitrary subsets of the u-values for data points and parameters to test for model departures using classical hypothesis tests for dependence or non-uniformity. We demonstrate empirically how this method of uniform parameterization checks (UPCs) facilitates model criticism in several examples, and we develop supporting theoretical results.

Keywords: Goodness-of-fit testing; Model misspecification; Model criticism; Hierarchical models; Robustness

Co-authors: Christian Covington

1.59. Peter Mueller

Institution: UT Austin

Department: Economics Department

Email: pmueller085@gmail.com

Website: <https://web.ma.utexas.edu/users/pmueller/>

Talk title:

Common atoms mixture models in some biostatistical inference problems

Keywords: real world data; clinical trials; spatial transcriptomics; dependent Dirichlet process

Co-authors: Yunshan Duan, Noirit Chandra, Wenyi Wang, Ying Yuan, Shuai Guo

1.60. XuanLong Nguyen

Institution: University of Michigan

Email: xuanlong@umich.edu

Talk title:

Learning topic hierarchies by tree-directed latent variable models

Abstract:

We study a parametric family of latent variable models, namely topic models, equipped with a hierarchical structure among the topic variables. Such models may be viewed as a finite mixture of the latent Dirichlet allocation (LDA) induced distributions, but the LDA components are constrained by a latent hierarchy, specifically a rooted and directed tree structure, which enables the learning of interpretable and latent topic hierarchies of interest. A mathematical framework is developed in order to establish identifiability of the latent topic hierarchy under suitable regularity conditions, and to derive bounds for posterior contraction rates of the model and its parameters. We demonstrate the usefulness of such models and validate its theoretical properties through a careful simulation study and a real data example using the New York Times articles.

Keywords: Hierarchical models, tree-based topic hierarchy, identifiability, posterior contraction

Co-authors: Sunrit Chakraborty, Rayleigh Lei

1.61. Yang Ni

Institution: Texas A&M University

Department: Statistics

Email: yni@stat.tamu.edu

Talk title:

Covariate-Assisted Bayesian Graph Learning for Heterogeneous Data

Abstract:

In a traditional Gaussian graphical model, data homogeneity is routinely assumed with no extra variables affecting the conditional independence. In modern genomic datasets, there is an abundance of auxiliary information, which often gets underutilized in determining the joint dependency structure. In this article, we consider a Bayesian approach to model undirected graphs underlying heterogeneous multivariate observations with additional assistance from covariates. Building on product partition models, we propose a novel covariate-dependent Gaussian graphical model that allows graphs to vary with covariates so that observations whose covariates are similar share a similar undirected graph. To efficiently embed Gaussian graphical models into our proposed framework, we explore both Gaussian likelihood and pseudo-likelihood functions. For Gaussian likelihood, a G-Wishart distribution is used as a natural conjugate prior, and for the pseudo-likelihood, a product of Gaussian-conditionals is used. Moreover, the proposed model has large prior support and is flexible to approximate any $\mathcal{C}\Omega$ -Hölder conditional variance-covariance matrices with $\mathcal{C}\Omega, \alpha \in (0,1]$. We further show that based on the theory of fractional likelihood, the rate of posterior contraction is minimax optimal assuming the true density to be a Gaussian mixture with a known number of components. The efficacy of the approach is demonstrated via simulation studies and an analysis of a protein network for a breast cancer dataset assisted by mRNA gene expression as covariates.

Keywords: Graphical models, Dirichlet processes, protein-protein interaction networks, product partition models, cancer

Co-authors: Yabo Niu, Debdeep Pati, Bani K. Mallick

1.62. Bernardo Nipoti

Institution: University of Milano-Bicocca

Email: bernardo.nipoti@unimib.it

Talk title:

Bayesian nonparametric modeling of heterogeneous populations of networks

Abstract:

We discuss a flexible framework for modeling multiple network data using similarity metrics to compare networks. Within this setting, we introduce a novel Bayesian nonparametric model that identifies clusters of networks with similar connectivity patterns. Our approach is based on a location-scale Dirichlet process mixture of centered Erdos-Renyi kernels, where each component is defined by a representative network, or mode, and a univariate measure of dispersion around it. This model offers desirable properties, including full support in the Kullback–Leibler sense and strong consistency. For posterior inference and network clustering, we develop an efficient Markov chain Monte Carlo algorithm. The model's performance is evaluated through extensive simulations and applications to human brain network data.

Keywords: Francesco Barile and Simon Lunagomez

1.63. Andriy Norets

Institution: Brown University

Department: Economics Department

Email: andriy_norets@brown.edu

Website: <https://anorets.github.io/>

Talk title:

Bayesian Nonparametric Model for Nonseparable Instrumental Variable Regression

Abstract:

We propose a flexible Bayesian model for estimation of nonseparable instrumental variable regressions with a univariate endogenous covariate and a univariate unobservable responsible for endogeneity. Our model uses recently developed nonparametric priors for conditional distributions based on covariate dependent mixtures that deliver optimal adaptive posterior contraction rates in settings with discrete and continuous variables under possible smoothness and sparsity. We develop a Markov chain Monte Carlo algorithm for estimation of this model. The proposed methodology provides a simple parametric baseline model that can be gradually extended to more flexible parametric models and ultimately a nonparametric one by increasing the number of mixture components or setting a prior on it. Thus, robustness to nonlinearities, nonseparability, heteroskedasticity and other deviations from the baseline can be gradually introduced in a computationally tractable way. We illustrate the model performance in simulations and an application.

Keywords: Nonseparable IV model, Bayesian nonparametrics, mixtures of normal distributions, covariate dependent mixtures, mixtures of experts

Co-authors: Simone Martinelli

1.64. Ilsang Ohn

Institution: Inha University

Email: ilsang.ohn@inha.ac.kr

Talk title:

Early stopped aggregation for nonparametric variational Bayes adaptation

Abstract:

When considering a model selection or, more generally, an aggregation approach for adaptive inference, it is often necessary to compute estimators over large model classes, even when the true data-generating process is relatively simple, due to the lack of prior knowledge. This requirement leads to substantial computational inefficiency. In this work, we propose a novel variational Bayes method for efficient model aggregation. Our approach leverages the idea of early stopping: instead of computing variational posteriors for all candidate models, we compute them only for a few simpler models and then aggregate the resulting posteriors to form a final variational posterior. In other words, we 'stop' computation once a certain level of model complexity is reached. Importantly, the number of computed variational posteriors is determined by a fully data-driven rule. We prove that the proposed method achieves optimal adaptive contraction rates. We further

demonstrate several applications of our method and present numerical studies that highlight the efficiency and strong performance of the proposed approach.

Keywords: early stopping; variational Bayes; posterior construction rates; adaptive inference

Co-authors: Lizhen Lin

1.65. Hernando Ombao

Institution: KAUST

Department: Statistics

Email: hernando.ombao@kaust.edu.sa

Talk title:

Neural Granger Causality: A Novel View

Abstract:

Granger Causality (GC) offers an elegant statistical framework to study the association between multivariate time series data. The family of classical vector autoregressive models (VAR) is easy to compute but has a major limitation, i.e., it captures only linear relationships between components. There have been recent approaches that harness the functional approximation power of Deep Neural Networks (DNNs) for the task of GC estimation. These methods however treat GC as a variable selection problem employing sparse regression frameworks which require tuning multiple hyper-parameters. We present a novel paradigm which treats GC as a prediction problem where we model the association between any two components of a multivariate time series, and provide uncertainty quantification using Bayesian Deep Learning. The proposed approach is a two step self-supervised estimation procedure. In the first step, the multivariate time series is jointly modeled to learn the dependence structure between multivariate time series data collectively. Self supervision is used to create a labeled dataset. In the second step, the supervised dataset created in the first step is used by a classification model is used to predict the entries of the adjacency matrix (A), by modeling it as a classification problem using a Bayesian Neural Network, where each i,j entry of A characterizes $P(X_i(t-h) \rightarrow X_j(t))$. Monte Carlo Dropout is used for uncertainty quantification in the prediction. We show that the Self Supervised Probabilistic GC (SSGC) performs at par with respect to accuracy, precision and recall in detecting the true GC graph, in comparison with the state-of-the-art models using several simulation baselines. Moreover, another advantage of the proposed SSGC is its capability to be used in real time or in time varying settings in a zero-shot manner without specifying or tuning many hyper-parameters that standard methods require. We also analyze real life EEG datasets along with simulation baselines. This is joint work with Ph.D. student Malik Shahid Sultan.

Keywords: Brain signals, Granger causality, multivariate time series, neural network

Co-authors: Malik Shahid Sultan

1.66. Peter Orbanz

Institution: University College London

Department: Gatsby Unit

Email: p.orbanz@ucl.ac.uk

Talk title:

Efficiently vectorized MCMC on modern accelerators

Abstract:

Markov chain sampling algorithms are now routinely used with the vectorization features of modern hardware. It has been observed that, for various MCMC algorithms, such vectorized implementations are inefficient due to a synchronization problem -- loosely speaking, at each iteration all chains running in parallel must wait until the last chain has finished drawing its sample. We show how to address this problem by augmenting the state space of the Markov chain using what theoretical computer scientists call a finite state machine. This results in substantial speedups, and simple theoretical results explain how these speedups come about and help to optimize algorithm design.

Keywords: MCMC algorithms

Co-authors: Hugh Dance, Pierre Glaser, Ryan P Adams

1.67. Julia Palacios

Institution: Stanford University

Department: Statistics Department

Email: juliapr@stanford.edu

Website: <https://juliapalacios.github.io/>

Talk title:

Novel coalescent point process approaches for modeling and for Bayesian inference of disease and molecular dynamics.

Abstract:

Coalescent point processes are used to infer population dynamics at different time scales from molecular sequence data. These methods have been applied to study cell population dynamics during developmental processes, to infer disease transmission dynamics of rapidly evolving pathogens, and to infer ancestral human population size trajectories. In this talk I will describe two challenging situations for coalescent inference of single cell lineage tracing and low phylogenetic signal. In the former, we derive an inference approach for the bounded coalescent extending the exact Gaussian Cox process to a Gaussian death Cox process. In the latter, we propose a coalescent model of multifurcating genealogies beyond the Lambda-coalescent.

Keywords: Exact Gaussian Cox Process, Multifurcating Genealogies, Coalescent, Phase-Type Distribution

Co-authors: Bingjing Tang, Julie Zhang

1.68. Riccardo Passeggeri

Institution: Imperial College London

Department: Mathematics

Email: riccardo.passeggeri@imperial.ac.uk

Website: <https://sites.google.com/site/riccardopasseggeri>

Talk title:

Random signed measures

Abstract:

Point processes and, more generally, random measures are ubiquitous in Bayesian nonparametric statistics. However, they can only take positive values, which is a severe limitation in many situations. In this presentation, we introduce random signed measures, also known as real-valued random measures, and apply them to construct various Bayesian non-parametric models. In particular, we provide an existence result for random signed measures, allowing us to obtain a canonical definition for random signed measures and solve a long-standing open problem. Further, we provide a representation of completely random signed measures (CRSMs), which extends the celebrated Kingman's representation for completely random measures (CRMs) to the real-valued case. We then introduce specific classes of random signed measures, including the Skellam point process, which plays the role of the Poisson point process in the real-valued case, and the Gaussian random measure. We use the theoretical results to develop two Bayesian nonparametric models -- one for sentiment topic modeling and the other for random graphs -- and to investigate mean function estimation in Bayesian nonparametric regression.

Keywords: random signed measure; completely random signed measures; Gaussian random measure; Sentiment topic modelling; Sparse random signed graphs.

Co-authors: None

1.69. Debdeep Pati

Institution: University of Wisconsin-Madison

Department: Statistics

Email: dpati2@wisc.edu

Talk title:

Scientifically guided robust inference

Abstract:

Motivated by the need for interpretable yet flexible models for modern scientific applications, we shall propose a likelihood-based inference constrained to lie close to a parametric family. We shall draw connections with nonparametric Bayesian inference constrained to lie around a base distribution and show favorable asymptotic properties of the proposed estimator. Finally, we shall explore a few applications from neuroscience, trustworthy machine learning and inference with survey data.

Keywords: empirical likelihood; robust; trustworthy; Wasserstein

Co-authors: Anirban Bhattacharya; Abhisek Chakraborty; Congwei Wang; Nicolas Garcia Trillos

1.70. Sonia Petrone

Institution: Bocconi University

Department: Department of Decision Sciences

Email: sonia.petrone@unibocconi.it

Website: <https://faculty.unibocconi.eu/soniapetrone/>

Talk title:

Bayesian predictive efficiency

Abstract:

In "predictive inference", the inferential scheme (model and prior) is deduced from the predictive rule, which is the rule that specifies how one learns from past observations to predict future events. In the basic context of random sampling, the natural request is that the predictive rule characterizes an exchangeable law; then the inferential scheme is given by the celebrated de Finetti representation theorem. Predictive characterizations of priors have a long tradition in Bayesian nonparametrics. However, beyond classic predictive constructions of discrete random measures, it is hard, if not impossible, to have an exchangeable-and-tractable predictive rule. Computations become challenging, especially with streaming data. There is thus interest in tractable Bayesian predictive rules that are not exchangeable, but still imply an inferential representation asymptotically, and allow for (parametric or non-parametric) inference. A question is whether "quasi-exchangeable" Bayesian predictive rules entail a loss of information in the learning process. This calls for a notion of 'predictive efficiency', as we will discuss in the talk and detail in some examples. This is joint work with Sandra Fortini.

Keywords: Predictive inference; asymptotic exchangeability; efficiency; Doob theorem

Co-authors: Sandra Fortini

1.71. Igor Pruenster

Institution: Bocconi University

Department: Bocconi Institute for Data Science and Analytics

Email: igor@unibocconi.it

Website: <http://mypage.unibocconi.eu/igorpruenster>

Talk title:

Multivariate species sampling models

Abstract:

Species sampling processes have long served as the fundamental framework for modeling random discrete distributions and exchangeable sequences. However, when analyzing data from distinct, though related, sources, a broader notion of probabilistic invariance is required, and partial exchangeability represents the natural choice. Over the past two decades, numerous models for partially exchangeable data, collectively known as dependent nonparametric priors, have been proposed. These include the hierarchical, nested and additive processes, widely used in Statistics and Machine Learning. However, a unifying framework remains elusive, leaving key questions about their underlying learning

mechanisms unanswered. We fill this gap by introducing multivariate species sampling models, a new general class of nonparametric priors that encompasses most existing dependent nonparametric processes. These models are characterized by their partially exchangeable partition probability function, which encodes the induced multivariate clustering structure. We establish their core distributional properties and analyze their dependence structure, demonstrating that borrowing of information across groups is entirely determined by shared ties. This provides new insights into the underlying learning mechanisms, offering, for instance, a principled rationale for the previously unexplained correlation structure observed in existing models. Beyond providing a cohesive theoretical foundation, our approach serves as a constructive tool for developing new models and opens new research directions aimed at capturing even richer dependence structures beyond the framework of multivariate species sampling processes.

Keywords: Borrowing of information; Dependent nonparametric prior; Discrete random structure; Partial exchangeability; Species sampling model

Co-authors: Beatrice Franzolini (Bocconi University), Antonio Lijoi (Bocconi University) and Giovanni Rebaudo (University of Torino)

1.72. Fernando Quintana

Institution: Pontificia Universidad Catolica de Chile

Department: Department of Statistics

Email: quintana@uc.cl

Website: <https://www.mat.uc.cl/~quintana/>

Talk title:

Spatio-Temporal Models via Sequences of Random Spanning Trees

Abstract:

We develop a model for spatio-temporal clustering, featuring a prior distribution based on product partition models that correlate spatial partitions. In addition, we use random spanning trees as a tool to facilitate exploration of the partition search space and to induce regionalization. The approach also handles overdispersion through a Poisson mixture model, with mean and dispersion parameters varying with spatio-temporal covariates. The proposal is applied to model the number of dengue cases reported weekly from 2018 to 2023 in the Southeast region of Brazil.

Keywords: Spatio-temporal clustering; Overdispersion; Correlated partitions; Product partition models

Co-authors: Jessica Pavani, Rosangela Loschi

1.73. Vinayak Rao

Institution: Purdue University

Email: varao@purdue.edu

Talk title:

MCMC for Bayesian nonparametric mixture modeling under differential privacy

Abstract:

Estimating the probability density of a population while preserving the privacy of individuals in that population is an important and challenging problem that has received considerable attention in recent years. We consider a Bayesian nonparametric mixture model under differential privacy (DP) and present two Markov chain Monte Carlo (MCMC) algorithms for posterior inference. One is a marginal approach, resembling Neal's algorithm 5 with a pseudo-marginal Metropolis-Hastings move, and the other is a conditional approach. Our approach is general, applicable to any mixture model and privacy mechanism and allows us to leverage existing MCMC methodology for non-privatized data. We give results on the computational complexity, acceptance rate, and mixing properties of our MCMC. We discuss the performance of our algorithms on synthetic and real datasets.

Keywords: BNP, MCMC, Differential Privacy

Co-authors: Mario Beraha; Stefano Favaro

1.74. Tommaso Rigon

Institution: University of Milano-Bicocca

Department: Department of Economics, Management, and Statistics

Email: tommaso.rigon@unimib.it

Website: <https://tommasorigon.github.io>

Talk title:

Bayesian inference for biodiversity estimation

Abstract:

Community ecology aims to understand species interactions and the factors that drive them. Joint species distribution models (JSDMs) have emerged as powerful tools for quantifying interspecific associations and assessing the effects of covariates on species occurrence patterns. However, many existing approaches inadequately address the presence of rare or previously unobserved species in new samples. We propose a Bayesian mixture framework that simultaneously estimates the co-occurrence structure of both common and rare species, while allowing for the discovery of unseen species under a controlled growth regime. We develop an efficient Gibbs sampler for posterior inference and evaluate our method using both synthetic data and the Guelph Arthropod dataset.

Keywords:

Bayesian Nonparametrics; Indian buffet processes; Mixture model; Species co-occurrence

Co-authors: Ching-Lung Hsu, David B. Dunson

1.75. Alan Riva-Palacio

Institution: National Autonomous University of Mexico

Department: Applied Mathematics and Systems Research Institute, Department of Probability and Statistics

Email: alan@sigma.iimas.unam.mx

Website: <https://alan7riva.github.io/>

Talk title:

Flexible BNP multiple-sample modeling with asymptotic guarantees

Abstract:

We present a framework for Bayesian non parametric comparison of multiple samples that is capable of flexible calibration for the prior information on each sample and that allows for modulation of the borrowing of information between the samples in order to ensure consistency at optimal rates. The proposed framework is general in the sense that it can be applied to the classes of neutral to the right (NTR) and normalized random measures with independent increment (NRMI) priors. We show in simulation studies how the proposed approach outperforms alternative models where the borrowing of information is not regulated.

Keywords: borrowing of information; consistency; neutral to the right priors; normalized random measures; prior calibration

Co-authors: None

1.76. Veronika Rockova

Institution: University of Chicago

Email: veronika.rockova@chicagobooth.edu

Talk title:

AI-Powered Bayesian Inference

Abstract:

The advent of Generative Artificial Intelligence (GAI) has heralded an inflection point that changed how society thinks about knowledge acquisition. While GAI cannot be fully trusted for decision-making, it may still provide valuable information that can be integrated into a decision pipeline. Rather than seeing the lack of certitude and inherent randomness of GAI as a problem, we view it as an opportunity. Indeed, variable answers to given prompts can be leveraged to construct a prior distribution which reflects assuredness of AI predictions. This prior distribution may be combined with tailored datasets for a fully Bayesian analysis with an AI-driven prior. In this paper, we explore such a possibility within a non-parametric Bayesian framework. The basic idea consists of assigning a Dirichlet process prior distribution on the data-generating distribution with AI generative model as its baseline. Hyper-parameters of the prior can be tuned out-of-sample to assess the informativeness of the AI prior. Posterior simulation is achieved by computing a suitably randomized functional on an augmented data that consists of observed (labeled) data as well as fake data whose labels have been imputed using AI. This strategy can be parallelized and rapidly produces iid samples from the posterior by optimization as opposed to sampling

from conditionals. Our method enables (predictive) inference and uncertainty quantification leveraging AI predictions in a coherent probabilistic manner.

Keywords: AI, Data Augmentation, Dirichlet Process

Co-authors:
Sean O'Hagan

1.77. Matteo Ruggiero

Institution: New York University Abu Dhabi

Email: mr7690@nyu.edu

Talk title:

Quasi-conjugacy in two-parameter Poisson–Dirichlet hidden Markov models

Abstract:

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.

Keywords: integer partitions; Young diagrams; two-parameter Poisson-Dirichlet; filtering; intractable likelihood

Co-authors: Marco Dalla Pria (University of Torino) Dario Spanò (University of Warwick)

1.78. Huiyan Sang

Institution: Texas A&M University

Department: Statistics

Email: huiyan@stat.tamu.edu

Talk title:

Bayesian spatial local variable selection models

Abstract:

Large spatial datasets with many spatial covariates have become ubiquitous in many fields in recent years. A question of interest is to identify which covariates are likely to influence a spatial response, and whether and how the effects of these covariates vary across space, including potential abrupt changes from region to region. To solve this question, a new Bayesian sparse spatially clustered regression approach is proposed, which could identify spatially contiguous subregions and locally important variables simultaneously. We develop an efficient informed MCMC algorithm for Bayesian model inference. The performance of the method is illustrated through simulation and real data examples.

Keywords: Variable Selection, Spatial Clustering, Informed MCMC, High dimensional statistics

Co-authors: Kun Huang, Texas A&M University

1.79. Babak Shahbaba

Institution: UC Irvine

Email: babaks@uci.edu

Talk title:

Scaling Up Bayesian Neural Networks with Neural Networks

Abstract:

Bayesian Neural Networks (BNNs) offer a principled and natural framework for proper uncertainty quantification in the context of deep learning. They address the typical challenges associated with conventional deep learning methods, such as data insatiability, ad-hoc nature, and susceptibility to overfitting. However, their implementation typically either relies on Markov chain Monte Carlo (MCMC) methods, which are characterized by their computational intensity and inefficiency in a high-dimensional space, or variational inference methods, which tend to underestimate uncertainty. To address this issue, we propose a novel Calibration-Emulation-Sampling (CES) strategy to significantly enhance the computational efficiency of BNN. In this framework, during the initial calibration stage, we collect a small set of samples from the parameter space. These samples serve as training data for the emulator, which approximates the map between parameters and posterior probability. The trained emulator is then used for sampling from the posterior distribution at substantially higher speed compared to the standard BNN. Using simulated and real data, we demonstrate that our proposed method improves computational efficiency of BNN, while maintaining similar performance in terms of prediction accuracy and uncertainty quantification.

Co-authors: Zahra Moslemi (UCI), Yang Meng (UCI), Shiwei Lan (ASU), Babak Shahbaba (UCI)

1.80. Anna Simoni

Institution: CREST-CNRS

Email: anna.simoni@ensae.fr

Talk title:

Panel data models with randomly generated groups: Bayesian inference and density forecasts

Abstract:

We consider a dynamic panel data model that accounts for a latent group structure across individuals which is constant over time. Differently from the previous panel data literature, we adopt a structural modeling that assumes that the individual effects are generated from a finite mixture with an unknown number of components and unknown parameters for each components. We first establish identification of this model. Then, we specify a prior for the number of components, the parameters of the mixture as well as for the coefficients of the dynamic and exogenous covariates. This extends the mixture of finite mixtures model to panel data settings. We establish asymptotic frequentist properties for the posterior of the parameters of interest as well as for the number of components. A Monte Carlo exercise illustrates finite sample properties.

Keywords: Multivariate mixture model, latent group structure, dynamic model, posterior consistency, Wasserstein distance

Co-authors: Jean-Pierre Florens, Toulouse School of Economics, France

1.81. Sanvesh Srivastava

Institution: The University of Iowa

Department: Statistics and Actuarial Science

Email: sanvesh-srivastava@uiowa.edu

Talk title:

Federated Causal Inference with Gaussian Process Priors

Abstract:

Recent advances in Bayesian nonparametric regression techniques have led to the development of specialized methods for estimating causal effects. Tree-based methods have become prominent due to their flexibility in capturing complex and heterogeneous treatment effects. In contrast, Gaussian process-based methods have received relatively less attention, with limited sampling-based approaches available for inferring the Conditional Average Treatment Effect (CATE). To bridge this methodological gap, we propose a regression model for potential outcomes leveraging a two-dimensional Gaussian process prior and develop an efficient Markov chain Monte Carlo algorithm tailored for the estimation of CATE. Furthermore, recognizing the importance of scalability and data privacy, we extend our Gaussian process-based approach into a federated framework, enabling distributed inference on CATE while preserving data confidentiality. Through extensive numerical studies, we demonstrate that our proposed method achieves accurate estimation of the CATE in settings with homogeneous and heterogeneous treatment effects.

Keywords: Causal inference; distributed inference; federated learning; Gaussian process prior

Co-authors: Sumedha Dhar (sumedha-dhar@uiowa.edu) and Nathan Wikle (nathan-wikle@uiowa.edu).

1.82. Surya Tokdar

Institution: Duke University

Department: Statistical Science

Email: surya.tokdar@duke.edu

Website: <https://www.stat.duke.edu/~st118/>

Talk title:

Density estimation and extreme value analysis

Abstract:

Statistical analyses of heavy-tailed distributions are susceptible to the problem of sparse information in the tail of the distribution getting washed away by unrelated features of a hefty bulk. Bayesian methods can avoid this problem by incorporating smoothness and tail regularization through carefully specified semiparametric prior distribution. We study such a formulation by combining classical extreme value theory with Bayesian density estimation with logistic Gaussian process priors, and show that the method is able to consistently estimate both the density function and its tail index at near minimax optimal rates of contraction. A joint, likelihood driven estimation of the bulk and the tail is shown to help improve uncertainty assessment in estimating the tail index parameter and offer more accurate and reliable estimates of the high tail quantiles compared to thresholding methods.

Keywords: Regular variation, extreme value analysis, density estimation, Gaussian processes, posterior contraction

Co-authors: Erika Cunningham, Sheng Jiang

1.83. Aad van der Vaart

Institution: TU Delft

Department: DIAM

Email: a.w.van.der.vaart@tudelft.nl

Website: <https://diamhomes.ewi.tudelft.nl/~avandervaart/>

Talk title:

Bayesian nonparametrics: from consistency to uncertainty quantification, a review

Abstract:

R.V. Ramamoorthi contributed significantly to the study of consistency of posterior distributions, in particular in nonparametric and semiparametric statistics. In this talk we review some of the follow up on contraction rates, which is considerable, and present the question of uncertainty quantification, on which we still know relatively little. Although for consistency the prior often washes out as the data become more and more informative, this is not true for contraction and uncertainty quantification (through the spread of the posterior distribution). As an example we present Dirichlet mixtures in density estimation, for which R.V. Ramamoorthi proved consistency in 1999, which were subsequently shown to

give fully adaptive, near-optimal contraction rates in a natural fully Bayes setup, but for which the uncertainty quantification still has not received theoretical justification. As another example we may include some inverse problems related to data assimilation.

Keywords: Consistency, contraction, uncertainty quantification

Co-authors: -

1.84. Mark van der Wilk

Institution: University of Oxford

Department: Department of Computer Science

Email: mark.vanderwilk@cs.ox.ac.uk

Website: <https://mvdw.uk>

Talk title:

Lessons from BNP on how should we grow/shrink neural network layers

Abstract:

Determining the size ("number of parameters") of a model is a question asked in both small statistical models, as well as the largest of current neural network models, because of its impact on both data efficiency and computational cost. Bayesian Nonparametrics argues for making models infinitely large, and relying on the prior to specify inductive biases for data efficiency. This strategy is loosely followed in deep learning, where models must be significantly overparameterised to obtain optimal performance. However, in deep learning, models are often compressed after training to reduce computational cost in deployment. Given the conflicting prescription of (infinitely) large models, with the desire for small models for computations sake, is there a single principle that can unify the two? We investigate this question in single-layer neural networks and Gaussian processes, and show that approximate inference combined with Bayesian nonparametrics provides a principle for both training and compressing models, even at the same time. We give a procedure for adjusting model size during training, and show that the required model size depends on the inductive bias, leading to behaviours where models grow and shrink, or start by memorising the data and then shrink as structure in the data is learned. We finish with some thoughts on how these ideas can scale in deep learning, with the hope of improving deep learning using insights from Bayesian nonparametrics.

Keywords: Occam's razor; Gaussian processes; approximate inference; deep learning; neural networks

Co-authors: Guiomar Pescador-Barrios, Tycho van der Ouderaa, Sarah Lucie Filippi

1.85. Sara Wade

Institution: University of Edinburgh

Department: School of Mathematics

Email: sara.wade@ed.ac.uk

Website: <https://sarawade.owlstown.net>

Talk title:

Understanding Uncertainty in Bayesian Cluster Analysis

Abstract:

The Bayesian approach to clustering is often appreciated for its ability to provide uncertainty in the partition structure. However, summarizing the posterior distribution over the clustering structure can be challenging, due the discrete, unordered nature and massive dimension of the space. While recent advancements provide a single clustering estimate to represent the posterior, this ignores uncertainty and may even be unrepresentative in instances where the posterior is multimodal. To enhance our understanding of uncertainty, we propose to summarize the posterior samples with not one, but multiple clustering estimates, each corresponding to a different part of the space of partitions that receives substantial posterior mass. In this work, we propose to find such clustering estimates by approximating the posterior distribution in a Wasserstein distance sense, equipped with a suitable metric on the partition space. An interesting byproduct is that a locally optimal solution to this problem can be found using the k-medoids algorithm on the partition space to divide the posterior samples into different groups, each represented by one of the clustering estimates. Using both synthetic and real datasets, we show that our proposal helps to improve the understanding of uncertainty, particularly when the data clusters are not well separated, or when the employed model is misspecified.

Keywords: clustering, uncertainty, variation of information, Wasserstein distance

Co-authors: Cecilia Balocchi

1.86. Richard Warr

Institution: Brigham Young University

Department: Statistics

Email: warr@stat.byu.edu

Talk title:

Clustering Craters on the Moon with Dysfunctional Families

Abstract:

Summaries of craters on terrestrial bodies, such as the number and size distribution, are essential for understanding the history of the Solar System. Identifying craters, however, has not been automated and thus relies on expert crater-counters marking static images. Robbins et al. [2014] showed that, contrary to previously held assumptions, there exists large variability across expert crater-counters' identified crater lists. How best to combine identified crater lists across multiple experts for the purposes of learning about the Solar System is an open and consequential question. Robbins et al. [2014] combined identified crater lists via clustering through a modification of the popular DBSCAN clustering

method. Their approach did not, however, make use of all the constraining information available nor did it provide an estimate of clustering uncertainty. To address the shortcomings of the DBSCAN method, we present a novel clustering approach that can combine multiple lists of identified objects of interest from the same image. The key innovation is incorporating a dysfunctional family constraint into the Bayesian nonparametric clustering approach, the Chinese restaurant process (CRP), which naturally takes into account information about the crater identifier. The dysfunctional family Chinese restaurant process (DFCRP) provides an estimate of clustering uncertainty. In this work, we provide guidance on hyperparameter specification, present a collapsed Gibbs sampler, and perform a simulation study to compare the performance of the DFCRP to the CRP. Finally, we apply the DFCRP to the crater identification problem of Robbins et al. [2014], comparing results, and also demonstrate the types of analyses that can be performed with posterior draws of cluster assignments.

Keywords: Bayesian nonparametrics; constrained clustering; model based clustering; exchangeability; moon craters

Co-authors: Emily Casleton, Dave Osthus, Brian Weaver

2. Contributed Talks

2.1. Davide Agnoletto

Institution: Duke University

Department: Department of Statistical Science

Email: davide.agnoletto@duke.edu

Website: <https://davideagno.github.io/>

Talk title:

Nonparametric predictive inference for discrete data via Metropolis-adjusted Dirichlet sequences

Abstract:

Bayesian inferences on unknown discrete data distributions, and in particular on count data, is typically challenging. Motivated by disadvantages of traditional mixture models in terms of model flexibility and/or efficiency of posterior inference, we develop a novel Bayesian predictive approach. In particular, our Metropolis-adjusted Dirichlet (MAD) sequence model characterizes the predictive measure as a mixture of a base measure and Metropolis-Hastings kernels centered on previous data points. The resulting MAD sequence is asymptotically exchangeable and the posterior on the data generator takes the form of a martingale posterior. This structure leads to straightforward algorithms for inference on count distributions, with easy extensions to multivariate, regression and binary data cases. We obtain a useful asymptotic Gaussian approximation and illustrate the methodology on a variety of applications.

Keywords: Count data; Martingale posterior; Nonparametric Bayes; Predictive inference; Sequential update.

Co-authors: Tommaso Rigon, David Dunson

2.2. Louise Alamichel

Institution: Bocconi University

Email: louise.alamichel@unibocconi.it

Talk title:

Enriched zero-inflated stochastic block models

Abstract:

Stochastic block models learn group structures among nodes sharing similar connectivity patterns. When sparse weighted networks are analyzed, these constructions have been recently extended to characterize zero inflation mechanisms in the observed edges by assuming zero-inflated Poisson likelihoods within each block with rates and zero-inflation probabilities varying across such blocks. However, state-of-the-art formulations assume that there is a unique nodes' partition characterizing stochastic equivalences in the

zero-inflation mechanism and in the counts of the observed edges. In practice, the node partition describing block structures in the zero-inflation probabilities might differ from the one regulating the Poisson rates, with the former often consisting of a coarsened version of the latter. For example, in criminal networks constrained by security-efficiency tradeoffs and complex organizational structures, there is evidence that the block patterns in the rates of interaction are obtained as fragmentation of those underlying the zero-inflation mechanism. To infer the above architectures, we develop a novel Bayesian zero-inflated stochastic block model relying on two different partition structures for the zero-inflation probabilities and rates, respectively, with the second resulting from a fragmentation of the first in different nested subgroups. This is accomplished through a novel treatment of enriched Gibbs-type constructions that leverages these formulations to induce a joint prior on the two partitions of the nodes under the desired structural constraint that one is a coarsened version of the other. Inference under the resulting enriched zero-inflated stochastic block model proceeds via a carefully-designed data-augmentation collapsed Gibbs-sampler. Preliminary results in simulated data and in a study of summit co-attendances within a complex Mafia organization showcase the strengths of the proposed formulation along with its ability to incorporate and learn relevant structures in networks with evidence of zero-inflation mechanisms.

Keywords: Criminal networks; Enriched Gibbs-type process; Stochastic block model.

Co-authors: Daniele Durante Tommaso Rigon

2.3. Martina Amongero

Institution: Università di Torino

Department: Department of Economics, Social Studies, Applied Mathematics, and Statistics (ESOMAS)

Email: martina.amongero@unito.it

Website: <https://sites.google.com/view/amongeromartina/home>

Talk title:

A Gibbs sampler for community detection in assortative Stochastic Block Model with unknown number of communities

Abstract:

Available data in the form of networks are gaining an increasing attention in modern research, examples include social networks, biological networks, information networks and many others. A statistical problem of key interest is community detection, that is to divide the nodes into strongly connected clusters with relatively weak cross-cluster connections. The Stochastic Block Model (SBM) provides a well suited generative process for explaining the formation of communities. By leveraging the SBM, researchers gain insights into the underlying structure of a network, uncovering interaction patterns that may not be apparent from raw data, and exploring network properties such as assortativity. In particular, a SBM is assortative when the probability of a connection is higher when nodes belong to the same rather than to different clusters. A recent line of works uses Bayesian non parametric methods for the recovery of communities in classical SBM by placing a prior distribution on the number of clusters and estimating cluster assignments with collapsed Gibbs samplers. However, the development of an efficient Gibbs sampler for

assortative SBM is still an open problem. In this work we aim at enforcing the assortative property through the prior and at studying its effect on community detection by implementing a Gibbs sampler for posterior inference. A simulation study based on a class of benchmark datasets is conducted to demonstrate the benefits of the assortative case over the standard one.

Keywords: network; community detection; Gibbs sampler; Stochastic Block Model;

Co-authors: Pierpaolo De Blasi (Università di Torino)

2.4. Filippo Ascolani

Institution: Duke University

Department: Department of Statistical Science

Email: filippo.ascolani@duke.edu

Website: <https://filippoascolani.github.io/>

Talk title:

Posterior results on the stick-breaking representation for Dirichlet process mixtures of Gaussians

Abstract:

The stick-breaking representation is a popular way of defining the Dirichlet process by the associated sequence of probability weights. It is particularly appealing when the discrete random measure is convolved with a suitable kernel: in this context, the stick-breaking construction is often truncated and posterior inference can be performed using a finite number of parameters. Despite its relevance, little is known about the posterior distribution of the weights in a mixture framework. Considering a Gaussian kernel and conducting a frequentist investigation, we combine prior properties and tools from the Bayesian asymptotic literature to deduce some results on the posterior behavior of the mixing measure: in particular we prove that if the “true” density is supersmooth then the set of relevant weights grows at most logarithmically with the number of observations. We use such results to shed some light on the clustering properties of Dirichlet process mixtures (e.g. number of clusters) and to provide posterior guarantees for computational methods based on truncation. This is a joint work with Surya Tokdar (Duke University).

Keywords: Clustering; asymptotic theory; computational scalability

Co-authors: Surya Tokdar (Duke University)

2.5. Ray Bai

Institution: University of South Carolina

Department: Statistics

Email: rbai@mailbox.sc.edu

Website: <https://raybai.net/>

Talk title:

Distributed Semiparametric Bayesian Inference for Massive Survival Data

Abstract:

Bayesian methods for survival analysis are attractive because of their flexibility and their ability to provide natural inference for both the covariate effects and nonparametric quantities of interest (e.g. the cumulative hazard or the survival function) through the posterior distribution. However, fitting Bayesian methods with Markov chain Monte Carlo (MCMC) is challenging when the sample size is large. To alleviate these computational challenges, we propose distributed Bayesian inference algorithms for the semiparametric Cox proportional hazards model based on the divide-and-conquer paradigm. Using random histogram priors for the unknown baseline hazard, we consider both the Cox model with independent observations and the mixed effects Cox model for clustered survival data. For the independent Cox model, we establish theoretically that the aggregated pseudo-posterior from our divide-and-conquer method provides a good approximation to the true posterior conditioned on the full data and quantifies uncertainty in the same order as the true posterior. For clustered survival data, we propose a novel data partitioning scheme to handle imbalanced cluster sizes. Our distributed MCMC algorithms greatly reduce the cost of running MCMC on large survival datasets, thus facilitating more efficient posterior inference. Our algorithms are illustrated on synthetic data and a kidney transplant dataset with 194,246 patients.

Keywords: survival analysis; random histogram prior; scalable Bayes; distributed inference

Co-authors: Zile Zhao, Sanvesh Srivastava, Dipankar Bandyopadhyay, Ray Bai

2.6. Elisabeth Bersson

Institution: MIT

Department: LIDS

Email: ebersson@mit.edu

Website: <https://betsybersson.github.io/>

Talk title:

Covariance Meta Regression

Abstract:

This article considers the problem of inferences on moderate to high-dimensional covariance matrices. A popular approach induces a low rank plus diagonal factorization of the covariance through a latent factor model. Current Bayesian and penalization approaches mostly focus on shrinking factor loadings towards zero, effectively shrinking towards a diagonal covariance. We propose a much more flexible class that adaptively

shrinks towards a variety of non-diagonal structures leveraging on meta covariates, which provide features of the outcomes instead of the samples. Even if no meta covariates are available, the proposed model more accurately estimates non-diagonal population covariance matrices than standard alternatives. We illustrate the approach through application to inferring covariance among chemical exposures, with the meta covariates encoding chemical molecular structure features.

Keywords: covariance estimation; shrinkage; factor analysis; high dimensional; meta features

Co-authors: David Dunson

2.7. Benedetta Bruni

Institution: Duke University

Department: Department of Statistical Sciences

Email: bb387@duke.edu

Talk title:

A partial likelihood approach to tree-based density modeling and its application in Bayesian inference

Abstract:

Tree-based models for probability distributions are usually specified using a predetermined, data-independent collection of candidate recursive partitions of the sample space. To characterize an unknown target density in detail over the entire sample space, candidate partitions must have the capacity to expand deeply into all areas of the sample space with potential non-zero sampling probability. Such an expansive system of partitions often incurs prohibitive computational costs and makes inference prone to overfitting, especially in regions with little probability mass. Existing models typically make a compromise and rely on relatively shallow trees. This hampers one of the most desirable features of trees, their ability to characterize local features, and results in reduced statistical efficiency. Traditional wisdom suggests that this compromise is inevitable to ensure coherent likelihood-based reasoning, as a data-dependent partition system that allows deeper expansion only in regions with more observations would induce double dipping of the data and thus lead to inconsistent inference. We propose a simple strategy to restore coherency while allowing the candidate partitions to be data-dependent, using Cox's partial likelihood. This strategy parametrizes the tree-based sampling model according to the allocation of probability mass based on the observed data, and yet under appropriate specification, the resulting inference remains valid. Our partial likelihood approach is broadly applicable to existing likelihood-based methods and in particular to Bayesian inference on tree-based models. We give examples in density estimation in which the partial likelihood is endowed with existing priors on tree-based models and compare with the standard, full-likelihood approach. The results show substantial gains in estimation accuracy and computational efficiency from using the partial likelihood.

Keywords: Bayesian Tree Models; Partial likelihood; Density Estimation

Co-authors: Li Ma

2.8. Alessandro Carminati

Institution: Politecnico di Milano

Department: Department of Mathematics

Email: alessandro.carminati@polimi.it

Talk title:

Hierarchical Shot Noise Cox Process Mixture Models

Abstract:

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.

Keywords: Partial exchangeability, hierarchical mixture model; shot-noise Cox process; Gibbs sampler

Co-authors: Mario Beraha, Federico Camerlenghi, Alessandra Guglielmi

2.9. Sunrit Chakraborty

Institution: University of Michigan - Ann Arbor, MI

Department: Department of Statistics

Email: sunritc@umich.edu

Website: <https://sites.google.com/umich.edu/sunrit-homepage/home>

Talk title:

Learning Topic Hierarchies by Tree-Directed Latent Variable Models

Abstract:

We study a parametric family of latent variable models, namely topic models, equipped with a hierarchical structure among the topic variables. Such models may be viewed as a finite mixture of the latent Dirichlet allocation (LDA) induced distributions, but the LDA components are constrained by a latent hierarchy, specifically a rooted and directed tree structure, which enables the learning of interpretable and latent topic hierarchies of interest. A mathematical framework is developed in order to establish the identifiability of the latent topic hierarchy under suitable regularity conditions and to derive bounds for posterior contraction rates of the model and its parameters. We demonstrate the

usefulness of such models and validate their theoretical properties through careful simulation studies and a real data example using the New York Times articles.

Keywords: topic model; topic hierarchy; directed tree; identifiability; inverse bound

Co-authors: Rayleigh Lei, XuanLong Nguyen

2.10. Claudia Contardi

Institution: University of Pavia

Department: Mathematics

Email: claudia.contardi01@universitadipavia.it

Talk title:

Gaussian credible intervals in Bayesian nonparametric estimation of the unseen

Abstract:

We investigate the topic of uncertainty quantification in the Bayesian nonparametric (BNP) approach to the problem of the estimation of the number of unseen species. This is a classical problem in statistics, with various applications from biological sciences to machine learning and linguistics. n observations taking values in a countable space of "species" are modelled as a sample from a random discrete distribution distributed as a Pitman-Yor (PY) prior. We are interested in providing a point estimator and confidence intervals for the number $K_{m,n}$ of hitherto unseen (distinct) species that would be observed in an additional sample of size m . As the posterior distribution of $K_{m,n}$ is available in closed form, it is possible to use the posterior mean as the natural BNP estimator. However, the direct construction of credible intervals is often infeasible, due to the computational complexity of evaluating the posterior for large m . Existing literature addresses this problem either via Monte Carlo approximation, or sampling from the limiting (for large m) posterior distribution, known as Pitman's posterior alpha-diversity. However, it is known that when the sample sizes n and m are comparable - a case which is interesting in many applications - the limiting distribution produces intervals that are shorter than the exact ones. We propose a novel, completely analytical approach to the problem, based on the study of the asymptotical behaviour of the distribution of $K_{m,n}$ when m goes to infinity and n is allowed to depend linearly on m . Exploiting recent probabilistic representations of $K_{m,n}$, we prove a weak law of large numbers with deterministic limit and a Gaussian central limit theorem, a behaviour which differs significantly from that observed in the "usual" asymptotical regime. Moreover, the CLT yields a Gaussian approximation of the posterior distribution of $K_{m,n}$ which can be used to construct credible intervals analytically. We validate our method on both synthetic and real-world data, showing that the Gaussian intervals outperform the state-of-the-art ones when n and m are comparable.

Keywords: Bayesian nonparametrics, unseen-species problem, Gaussian Central Limit Theorem, Uncertainty quantification, Pitman-Yor process

Co-authors: Emanuele Dolera (university of Pavia) Stefano Favaro (university of Turin and Collegio Carlo Alberto)

2.11. Laura D'Angelo

Institution: University of Milano-Bicocca

Department: Department of Economics, Management and Statistics

Email: laura.dangelo@unimib.it

Website: <https://laura-dangelo.github.io/>

Talk title:

Dependent Dirichlet processes via thinning

Abstract:

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.

Keywords: Mixture model; Model-based clustering; Partially exchangeable data; Stick-breaking prior

Co-authors: Bernardo Nipoti and Andrea Ongaro

2.12. Claudio Del Sole

Institution: Università of Milan - Bicocca

Email: claudio.delsol@unimib.it

Website: <https://claudiodelsole.github.io/>

Talk title:

A principled approach to covariate-dependent random partition models

Abstract:

In the Bayesian nonparametric framework, the inclusion of continuous covariates within a fully nonparametric regression model represents a challenging task, and existing models in the literature face a trade-off between flexibility in modelling the latent partition structure, its analytical tractability, and its consistency for new observations. We introduce and thoroughly analyze a novel class of covariate-dependent random probability measures, arising from the normalization of suitable random measures, which account for covariates through a kernel structure. Specifically, the jumps of a common discrete random measure are rescaled via multiplication by a similarity kernel: the resulting random probability measure is termed kernel-weighted normalized random measure. Noteworthy examples arise when the common random measure is a gamma process, or its distribution is a specific transformation of the distribution of a stable process, leading to kernel-weighted Dirichlet and Pitman-Yor processes, respectively. This construction induces a random

partition model with dependence on covariates, which is characterized by great flexibility and inherent consistency for new observations, while retaining some analytical tractability, thanks to the introduction of suitable latent variables. In fact, conditionally on such latent variables, the resulting partition model is closely related to the popular PPMx models. The induced partition probability function and the posterior distribution of the common random measure are derived in closed form, given the latent variables, and the predictive scheme for both the new observations and corresponding latent variables is characterized. This proposal can be effectively exploited as the building block for the construction of nonparametric regression models, as well as a clustering or species sampling model which incorporates information available through different types of covariates. Finally, the structure of the posterior distribution implied by the kernel-weighted Pitman-Yor process specification suggests the introduction of a novel nonhomogeneous process, which extends the standard Pitman-Yor process and acts as quasi-conjugate prior for our proposal.

Keywords: bayesian nonparametric regression; partition model; completely random measure; Pitman-Yor process; kernel weighting

Co-authors: Antonio Lijoi Igor Pruenster

2.13. Miheer Dewaskar

Institution: University of New Mexico

Department: Mathematics and Statistics

Email: mdewaskar@unm.edu

Website: <https://miheerdewaskar.com/>

Talk title:

Bayesian Level Set Clustering

Abstract:

Classically, Bayesian clustering interprets each component of a mixture model as a cluster. The inferred clustering posterior is highly sensitive to any inaccuracies in the kernel within each component. As this kernel is made more flexible, problems arise in identifying the underlying clusters in the data. To address this pitfall, in this talk we propose a fundamentally different approach to Bayesian clustering that decouples the problems of clustering and flexible modeling of the data density f . Starting with an arbitrary Bayesian (typically non-parametric) model for f and a loss function for defining clusters based on f , we develop a Bayesian decision-theoretic framework for density-based clustering. Within this framework, we develop a Bayesian level set clustering method to cluster data into connected components of a level set of f . We provide theoretical support, including clustering consistency, and highlight performance in a variety of simulated examples. An application to astronomical data illustrates improvements over the popular DBSCAN algorithm in terms of accuracy, insensitivity to tuning parameters, and providing uncertainty quantification.

Keywords: Bayesian nonparametrics, DBSCAN, Decision theory, Density-based clustering, Loss function, Nonparametric density estimation

Co-authors: David Buch and David B. Dunson

2.14. Alexander Dombowsky

Institution: Duke University

Department: Department of Statistical Science

Email: alex@dombowsky.net

Website: <https://adombowsky.github.io>

Talk title:

Hierarchical Directed Dirichlet Networks for Discrete Graphical Modeling

Abstract:

There are ample methods for modeling continuous features with a graphical dependence structure. Until recently, the setting of discrete type features has been largely overlooked relative to the massive literature on Gaussian graphical models. In this paper, we propose the Hierarchical Directed Dirichlet Network (HiDDeN), a fundamentally new Bayesian framework for the discrete belief networks, in which dependence between features is described by an underlying directed acyclic graph (DAG). For any node in the DAG, the conditional probabilities associated with its parents are modeled from a hierarchical Dirichlet distribution, i.e. a Dirichlet distribution centered on an unknown Dirichlet distributed probability measure. Deviation from this latent probability measure is controlled with a real-valued concentration parameter. This construction allows the sharing of information between the edges in the DAG, results in a reasonable shrinkage estimator for the conditional probabilities, and leads to an efficient marginal Markov chain Monte Carlo (MCMC) algorithm. In addition, we derive a closed form expression of the marginal likelihood for the HiDDeN model by integrating out the latent Dirichlet vectors associated with each node. From the marginal likelihood, we develop a Bayesian hypothesis testing procedure for graphical inference and an algorithm for empirical Bayes selection of the concentration parameters. HiDDeN is evaluated in several simulation studies, where we vary the sample size, dimension, number of categories, and the underlying DAG; in all cases, HiDDeN excels at learning the underlying conditional probability tensor. We apply our methodology to an application in patient monitoring, where we use HiDDeN to infer the conditional probabilities of diagnosis and to test edges in the DAG.

Keywords: Graphical models; contingency tables; belief networks; Bayesian hypothesis testing; hierarchical Dirichlet processes

Co-authors: David B. Dunson

2.15. Joseph Feldman

Institution: Duke University

Department: Statistical Science

Email: joseph.feldman@duke.edu

Website: <https://joseph-feldman-phd.com>

Talk title:

Data integration for multiple imputation of nonignorable missing data: A copula-based approach

Abstract:

We present an approach for modeling and imputation of nonignorable missing data. Our approach uses Bayesian data integration to combine (1) a Gaussian copula model for all study variables and missingness indicators, which allows arbitrary marginal distributions, nonignorable missingness, and other dependencies, and (2) auxiliary information in the form of marginal quantiles for some study variables. We prove that, remarkably, one only needs a small set of accurately-specified quantiles to estimate the copula correlation consistently. The remaining marginal distribution functions are inferred nonparametrically and jointly with the copula parameters using an efficient MCMC algorithm. We also characterize the (additive) nonignorable missingness mechanism implied by the copula model. Simulations confirm the effectiveness of this approach for multivariate imputation with nonignorable missing data. We apply the model to analyze associations between lead exposure and end-of-grade test scores for 170,000 North Carolina students. Lead exposure has nonignorable missingness: children with higher exposure are more likely to be measured. We elicit marginal quantiles for lead exposure using statistics provided by the Centers for Disease Control and Prevention. Multiple imputation inferences under our model support stronger, more adverse associations between lead exposure and educational outcomes relative to complete case and missing-at-random analyses.

Keywords: Bayesian semi-parametrics; missing-not-at-random; data integration

Co-authors: Jerome P. Reiter and Daniel Kowal

2.16. Jared Fisher

Institution: Brigham Young University

Department: Statistics

Email: fisher@stat.byu.edu

Website: <https://jareddf.github.io/>

Talk title:

A Bayesian Classification Trees Approach to Treatment Effect Variation with Noncompliance.

Abstract:

Estimating varying treatment effects in randomized trials with noncompliance is inherently challenging since variation comes from two separate sources: variation in the impact itself and variation in the compliance rate. In this setting, existing flexible machine learning methods are highly sensitive to the weak instruments problem, in which the compliance rate is (locally) close to zero. Our main methodological contribution is to present a Bayesian Causal Forest model for binary response variables in scenarios with noncompliance. By repeatedly imputing individuals' compliance types, we can flexibly estimate heterogeneous treatment effects among compliers. Simulation studies demonstrate the usefulness of our approach when compliance and treatment effects are heterogeneous. We apply the method to detect and analyze heterogeneity in the treatment effects in the Illinois Workplace Wellness Study, which not only features heterogeneous and one-sided compliance but also several binary outcomes of interest. We demonstrate the methodology on three outcomes one year after intervention. We confirm a null effect on the presence of a chronic condition, discover meaningful heterogeneity impact of the intervention on metabolic parameters though the average effect is null in classical partial

effect estimates, and find substantial heterogeneity in individuals' perception of management prioritization of health and safety.

Keywords: Causal inference; BART; principal stratification; heterogeneous treatment effects

Co-authors: David W. Puelz, Sameer K. Deshpande

2.17. Sandra Fortini

Institution: Bocconi University

Department: Decision Sciences

Email: sandra.fortini@unibocconi.it

Website: <https://dec.unibocconi.eu/people/sandra-fortini>

Talk title:

A sequential approach to the Poisson compound decision problem

Abstract:

The Poisson compound decision problem is a classical problem in statistics, for which parametric and nonparametric empirical Bayes methodologies are available to estimate the Poisson's means in static or batch domains. We consider the Poisson compound decision problem in a streaming or online domain. By relying on a quasi-Bayesian approach, often referred to as Newton's algorithm, we obtain sequential Poisson's mean estimates that are of easy evaluation, computationally efficient and with a constant computational cost as data increase, which is desirable for streaming data. Large sample asymptotic properties of the proposed estimates are investigated, also providing frequentist guarantees in terms of a regret analysis. We validate empirically our methodology, both on synthetic and real data, comparing against the most popular alternatives.

Keywords: Empirical Bayes; Newton's algorithm; Poisson compound decision; quasi-Bayes; streaming data

Co-authors: Stefano Favaro

2.18. Wenyu Gao

Institution: UNC Charlotte

Department: Department of Mathematics and Statistics

Email: wgao4@charlotte.edu

Website: <https://tinagwy.netlify.app/>

Talk title:

A Weighted Dirichlet Process Mixture Model for Spatial Clustering of Child Trafficking Vulnerabilities in Sierra Leone

Abstract:

Child trafficking remains a critical social challenge characterized by complex spatial and contextual variations. This study employs a weighted Dirichlet process mixture (WDPM) model to investigate trafficking vulnerability patterns across 138 Sierra Leonean

enumeration areas (EAs). Incorporating areal information from the EAs as random effects, a generalized linear mixed model (GLMM) is developed to examine the relationship between trafficking prevalence and potential risk factors. By applying the nonparametric Bayesian clustering approach to the random effects, this research innovatively identifies latent risk factors and spatial dependencies. The WDPM model is able to integrate socioeconomic and cultural factors into its weight function, enhancing clustering precision and model estimation accuracy. Unlike traditional clustering techniques, the WDPM approach dynamically determines cluster configurations without requiring predefined cluster numbers. Clustering results are rigorously validated through community-level hypothesis testing, ensuring robustness and reliability. This novel framework provides a sophisticated analytical method for understanding trafficking risks in resource-constrained settings. By uncovering latent vulnerability patterns, the study offers actionable insights for designing targeted, context-specific intervention strategies. The findings highlight a data-driven approach with significant potential to inform policy-making and optimize resource allocation to combat child trafficking effectively.

Keywords: Child Trafficking, Weighted Dirichlet Process Mixture (WDPM), Spatial Patterns, Sierra Leone

Co-authors: Hui Yi, Assistant Research Scientist / Data Scientist, Center on Human Trafficking Research & Outreach (CenHTRO), University of Georgia

2.19. Lorenzo Ghilotti

Institution: University of Milano-Bicocca

Department: Department of Economics, Management and Statistics

Email: l.ghilotti@campus.unimib.it

Website: <https://sites.google.com/campus.unimib.it/lorenzoghilotti/home-page>

Talk title:

Bayesian nonparametric processes for clustering count data with potentially unobserved traits

Abstract:

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.

Keywords: Nonparametric clustering; Completely random measures; Unseen traits; Mixture models

Co-authors: Federico Camerlenghi Michele Guindani Tommaso Rigon

2.20. Alice Giampino

Institution: University of Milano-Bicocca

Department: Department of Economics, Management and Statistics

Email: giampinoalice@gmail.com

Website: <https://alicegiampino.github.io/>

Talk title:

Local Level Dynamic Random Partition Models for Changepoint Detection

Abstract:

Motivated by an increasing demand for models that can effectively describe features of complex multivariate time series, e.g. from sensor data in biomechanics, motion analysis, and sports science, we introduce a novel state-space modeling framework where the state equation encodes the evolution of latent partitions of the data over time. Building on the principles of dynamic linear models, our approach develops a random partition model capable of linking data partitions to previous ones over time, using a straightforward Markov structure that accounts for temporal persistence and facilitates changepoint detection. The selection of changepoints involves multiple dependent decisions, and we address this time-dependence by adopting a non-marginal false discovery rate control. This leads to a simple decision rule that ensures more stringent control of the false discovery rate compared to approaches that do not consider dependence. The method is efficiently implemented using a Gibbs sampling algorithm, leading to a straightforward approach compared to existing methods for dependent random partition models. Additionally, we show how the proposed method can be adapted to handle multi-view clustering scenarios. Simulation studies and the analysis of a human gesture phase dataset collected through various sensing technologies show the effectiveness of the method in clustering multivariate time series and detecting changepoints.

Keywords: Bayesian nonparametrics; Bayesian clustering; Chinese restaurant process; Gibbs-type prior; spike-and-slab prior

Co-authors: Michele Guindani, Bernardo Nipoti, Marina Vannucci

2.21. Matteo Gianella

Institution: Politecnico di Milano

Department: Department of Mathematics

Email: matteo.gianella@polimi.it

Website: <https://sites.google.com/view/matteogianella/home>

Talk title:

“Slice” the tables in the Hierarchical Dirichlet Process

Abstract:

The hierarchical Dirichlet process (HDP) is a well-known and well-established nonparametric Bayesian model that extends the Dirichlet process to handle partially exchangeable data. It is particularly useful in scenarios where data are naturally organised into groups, and the statistical target is to identify shared clusters across these groups. Since its introduction back in 2004, the HDP has been consistently used as a mixing measure in parametric mixtures for clustering in various contexts, e.g., in topic modelling, collaborative filtering and gene expression analysis. Posterior inference in the HDP is typically based on the so-called Chinese restaurant franchise metaphor, whereby customers sit at different tables as in the original Chinese restaurant process, but different tables might share the same dish. This makes posterior inference cumbersome as one must introduce two sets of latent indicator variables that interact. The aim is to demonstrate how, by simply assuming a prior distribution on the concentration parameter of the HDP, another representation of the HDP can be obtained. In this new representation, all tables serve distinct dishes, paving the way for a slice sampling algorithm that reduces the amount of bookkeeping required and significantly improves posterior computation in terms of efficiency.

Keywords: computational statistics; hierarchical processes; completely random measures; partial exchangeability

Co-authors: Mario Beraha, Federico Camerlenghi and Alessandra Guglielmi

2.22. Isaac Goldstein

Institution: Stanford University

Department: Statistics

Email: igold@stanford.edu

Website: <https://isaachgoldstein.netlify.app>

Talk title:

Semiparametric inference of effective reproduction number dynamics from pathogen genealogies

Abstract:

The effective reproduction number is an important descriptor of an infectious disease epidemic. Coalescent models can be used to infer the effective reproduction number from genealogies constructed from viral genomes sampled from infected individuals during an epidemic. Most methods based on the coalescent model either assume individuals to be infectious as soon as they are infected, or that the transmission rate is constant over time,

both are strong assumptions violated by many real world diseases such as SARS-CoV-2. We propose a computationally tractable coalescent model which models only latent (infected but not yet infectious) and infectious individuals. We combine this model with a non-parametric Gaussian Markov Random Field prior on the time-varying effective reproduction number to conduct semiparametric inference of the effective reproduction number. Using simulated genealogies, we demonstrate the advantages of our approach over previous methodologies. We apply our new model to estimating the effective reproduction number of SARS-CoV-2 in Los Angeles County.

Keywords: Nowcasting; Kingman's Coalescent; Phylodynamics; Compartmental Models; GMRF;

Co-authors: Julia Palacios

2.23. Akira Horiguchi

Institution: University of California, Davis

Department: Statistics

Email: ahoriguchi@ucdavis.edu

Website: <https://ahoriguchi.github.io>

Talk title:

Multidimensional Scalable Generative Wavelet Tree Ensembles

Abstract:

We consider the problem of learning data representations that preserve existing geometric structures in an image. An existing approach is to incorporate adaptivity into discrete wavelet transforms by placing a Bayesian hierarchical model on the space of representations of an image. Such an approach provides both image denoising and generation of new images via sampling from the posterior distribution. Rather than using (multidimensional) balanced Haar wavelets, we propose using their unbalanced counterparts, which allows asymmetric partitions and hence enlarges the prior support on the space of representations. This Bayesian nonparametric approach improves upon existing approaches by better detecting edges and homogeneous regions in an image. We also propose using this more flexible wavelet basis to produce an ensemble of weak learners, since each weak learner can more flexibly capture a high-signal region of a residual image, which is not possible with balanced basis. Furthermore, we incorporate sequential Monte Carlo for computational scalability and to mitigate overfitting. We evaluate the performance of our method on synthetic and real images and find that (1) it detects edges while avoiding splits in homogeneous regions, (2) it produces smaller MSE and MAE than the balanced Haar counterpart, and (3) boosting produces effective representations that can be used for subsequent tasks.

Keywords: image reconstruction; adaptive wavelet basis; generative model;

Co-authors: Hengrui Luo (Assistant Professor at the Department of Statistics at Rice University; Li Ma (Professor at the Department of Statistical Science at Duke University)

2.24. Michael Jauch

Institution: Florida State University

Department: Department of Statistics

Email: mjauch@fsu.edu

Website: <https://michaeljauch.github.io/>

Talk title:

More on Bayesian nonparametric inference for likelihood ratio ordered distributions

Abstract:

Inference under shape or stochastic order constraints is an important topic in Bayesian nonparametrics. In recent work, we introduced a mixture representation for likelihood ratio ordered distributions, establishing that the ratio of two probability densities is monotone if and only if one can be expressed as a mixture of one-sided truncations of the other. We leveraged the mixture representation to develop a consistent Bayesian nonparametric density estimation procedure for likelihood ratio ordered distributions. In this talk, we will discuss two extensions of that work: 1) a Bayesian nonparametric procedure for hidden Markov models with likelihood ratio ordered emission densities and 2) a mixture representation for a pair of multivariate densities whose ratio is component-wise monotone.

Keywords: stochastic order; shape constraint; density estimation; mixture model

Co-authors: Andrés F. Barrientos, Víctor Peña, David S. Matteson

2.25. Alex Jiang

Institution: University of Washington - Seattle

Department: Department of Statistics

Email: jiang14@uw.edu

Website: <https://alexjiang1125.github.io/>

Talk title:

Posterior Consistency Under Model Misspecification - A Two-level Mixture Model Approach

Abstract:

Consistency in the number of mixture components is a desired property for Bayesian mixture models, which helps us better understand the underlying subpopulations and latent structures within a population. While the Dirichlet process mixture (DPM) model is asymptotically optimal for density estimation, it fails to provide consistent estimates for the number of mixture components in the model. The mixture of finite mixtures (MFM) model guarantees such consistency under mild identifiability assumptions. However, the guarantee breaks down under the regime where the mixture kernel is misspecified. In this work, we propose a two-level MFM-DPM model, where each mixture component in the MFM model is again modeled as a Dirichlet process mixture. The hierarchical structure of MFM-DPM allows the bottom-level DPM mixtures to approximate the top-level MFM mixture, even under kernel misspecification. Our study aims to address the following two

key questions: (1) Within the misspecified regime, is it possible to establish theoretical guarantees of mixture component consistency through a posteriori processing of the samples? (2) Can we make informed decisions on prior elicitation for better finite-sample results? Additionally, we develop a Markov chain Monte Carlo (MCMC) algorithm to implement our model, and we demonstrate its performance via simulation studies and an application to cancer genomics data.

Keywords: Mixture models; Mixture of finite mixtures; Dirichlet process mixtures; Posterior consistency; Hierarchical models;

Co-authors: Abel Rodriguez (Department of Statistics, University of Washington) Wei Sun (Fred Hutchinson Cancer Center)

2.26. George Karabatsos

Institution: University of Illinois-Chicago

Department: Mathematics, Statistics, and Computer Science, and Educational Statistics

Email: gkarabatsos1@gmail.com

Website: <https://georgek.people.uic.edu/>

Talk title:

Bayesian Nonparametric Sensitivity Analysis of Multiple Comparisons Under Dependence

Abstract:

This presentation will introduce a sensitivity analysis method for Multiple Testing Procedures (MTPs), based on marginal p-values and the Dirichlet process prior distribution. The method measures each p-value's insensitivity towards a significance decision, with respect to the entire space of MTPs controlling either the family-wise error rate (FWER) or the false discovery rate (FDR) under arbitrary dependence between p-values, supported by this nonparametric prior. The sensitivity analysis method is illustrated through 1,081 hypothesis tests of the effects of the COVID-19 pandemic on educational processes for 15-year-old students, performed on a 2022 public dataset. Extensions of the MTP sensitivity analysis method are discussed. Software code for the method is provided. Manuscript: <https://arxiv.org/abs/2410.08080>

Keywords: Sensitivity analysis; Multiple testing; False discovery rate; Familywise error rate; Dirichlet process.

Co-authors: None.

2.27. Matthew Koslovsky

Institution: Colorado State University

Department: Statistics

Email: matt.koslovsky@colostate.edu

Website: <https://mkoslovsky.quarto.pub/matt-koslovsky/>

Talk title:

Dynamic Functional Variable Selection for Multimodal mHealth Data

Abstract:

Mobile health (mHealth) methods allow researchers to monitor study participants in their natural environments in order to improve health-related outcomes through behavior change. MHealth investigators are interested in understanding the feasibility of supplementing or even replacing actively collected data with passively collected data to reduce participant burden, while also increasing the temporal resolution of the data. In this talk, we present a novel Bayesian dynamic functional variable selection method to explore the relations between multimodal mHealth data collected on different time scales. Specifically, our approach leverages spiked hierarchical species sampling priors to identify critical moments when a participant experiences momentary spikes in a low-frequency outcome, characterize high-frequency outcome trajectories which are related to these critical moments, and cluster the relational trends to explore potential subpopulations. We introduce continuous-time multistate Markov model priors to inform selection based on information learned at previous assessments. We demonstrate the variable selection and clustering performance of our model in various simulation settings motivated by the data structures found in mHealth studies. We then apply our model to multimodal intensive longitudinal data collected in the Pathways between Socioeconomic Status and Behavioral Cancer Risk Factors Study to explore relations between physical activity passively collected with accelerometers and mood actively collected with ecological momentary assessment methods.

Keywords: accelerometry; ecological momentary assessment; joint modeling, species sampling priors.

Co-authors: Kelley Pettee Gabriel, Michael Businelle, David W. Wetter, and Darla Kendzor

2.28. Dan Kowal

Institution: Cornell University

Department: Statistics and Data Science

Email: dan.kowal@cornell.edu

Website: <https://www.danielrkowal.com/>

Talk title:

Monte Carlo inference for semiparametric Bayesian regression

Abstract:

Data transformations are essential for broad applicability of parametric regression models. However, for Bayesian analysis, joint inference of the transformation and model parameters typically involves restrictive parametric transformations or nonparametric representations that are computationally inefficient and cumbersome for implementation and theoretical analysis, which limits their usability in practice. We introduce a simple, general, and efficient strategy for joint posterior inference of an unknown transformation and all regression model parameters. The proposed approach directly targets the posterior distribution of the transformation by linking it with the marginal distributions of the independent and dependent variables, and then deploys a Bayesian nonparametric model via the Bayesian bootstrap. Crucially, this approach delivers (1) joint posterior consistency under general conditions, including multiple model misspecifications, and (2) efficient

Monte Carlo (not Markov chain Monte Carlo) inference for the transformation and all parameters for important special cases. These tools apply across a variety of data domains, including real-valued, positive, and compactly-supported data. Simulation studies and an empirical application demonstrate the effectiveness and efficiency of this strategy for semiparametric Bayesian analysis with linear models, quantile regression, and Gaussian processes. The R package SeBR is available on CRAN. Paper:

<https://doi.org/10.100/01621459.2024.2395586> Software documentation:

<https://drkowal.github.io/SeBR/>

Keywords: Bayesian nonparametrics; Gaussian processes; Nonlinear regression; Quantile regression; Transformations

Co-authors: Bohan Wu

2.29. Suprateek Kundu

Institution: University of Texas MD Anderson Cancer Center

Department: Biostatistics

Email: skundu2@mdanderson.org

Talk title:

Flexible Bayesian Nonparametric Product Mixtures for Multi-scale Functional Clustering

Abstract:

There is a rich literature on clustering functional data with applications to time-series modeling, trajectory data, and even spatio-temporal applications. However, existing methods routinely perform global clustering that enforces identical atom values within the same cluster. Such grouping may be inadequate for high-dimensional functions, where the clustering patterns may change between the more dominant high-level features and the finer resolution local features. While there is some limited literature on local clustering approaches to deal with the above problems, these methods are typically not scalable to high-dimensional functions, and their theoretical properties are not well-investigated. Focusing on basis expansions for high-dimensional functions, we propose a flexible non-parametric Bayesian approach for multi-resolution clustering. The proposed method imposes independent Dirichlet process (DP) priors on different subsets of basis coefficients that ultimately results in a product of DP mixture priors inducing local clustering. We generalize the approach to incorporate spatially correlated error terms when modeling random spatial functions to provide improved model fitting. An efficient Markov chain Monte Carlo (MCMC) algorithm is developed for implementation. We show posterior consistency properties under the local clustering approach that asymptotically recovers the true density of random functions. Extensive simulations illustrate the improved clustering and function estimation under the proposed method compared to classical approaches. We apply the proposed approach to a spatial transcriptomics application where the goal is to infer clusters of genes with distinct spatial patterns of expressions. Our method makes an important contribution by expanding the limited literature on local clustering methods for high-dimensional functions with theoretical guarantees.

Keywords: Global and local clustering; product of Dirichlet process mixtures; posterior consistency; spatial functions; spatial transcriptomics.

Co-authors: Tsung-Hung Yao

2.30. Kyeongwon Lee

Institution: University of Maryland

Department: Department of Mathematics

Email: kwlee1718@gmail.com

Talk title:

Bayesian Sparse Neural Networks in General Besov Spaces with Intrinsic Dimensions

Abstract:

Neural networks have shown notable performance in machine learning for their flexibility and ability to adapt model complexity to data. This talk focuses on Bayesian sparse neural networks for nonparametric regression problems. We explore how these models achieve posterior consistency and nearly optimal convergence rates when the true regression belongs to the Besov spaces with intrinsic dimensions. These function spaces encompass a diverse array of mathematical objects making them particularly suitable for addressing problems with anisotropic smoothness and compositional structures. A key aspect of our discussion is the introduction of priors and their role in achieving optimal convergence rates. We demonstrate that these priors not only attain the desired rate of convergence but also maintain this performance even when the smoothness of the regression function is unknown, showcasing their adaptive behavior. This research contributes to the growing body of work in Bayesian nonparametrics, offering insights into the design and analysis of models for high-dimensional data and varying functional smoothness.

Keywords: Bayesian sparse neural networks, Posterior consistency, Besov space, Intrinsic dimension

Co-authors: Lizhen Lin, Jaewoo Park, Seonghyun Jeong

2.31. Changwoo Lee

Institution: Duke University

Department: Department of Statistical Science

Email: changwoo.lee@duke.edu

Website: <https://changwoo-lee.github.io/>

Talk title:

Logistic-beta processes for dependent random probabilities with beta marginals

Abstract:

The beta distribution serves as a canonical tool for modeling probabilities in statistics and machine learning. However, there is limited work on flexible and computationally convenient stochastic process extensions for modeling dependent random probabilities. We propose a novel stochastic process called the logistic-beta process, whose logistic

transformation yields a stochastic process with common beta marginals. Logistic-beta processes can model dependence on both discrete and continuous domains, such as space or time, and have a flexible dependence structure through correlation kernels. Moreover, its normal variance-mean mixture representation leads to effective posterior inference algorithms. We illustrate the benefits through nonparametric binary regression and conditional density estimation examples, both in simulation studies and in a pregnancy outcome application.

Keywords: Bayesian nonparametrics; Dependent Dirichlet process; Multivariate beta distribution; Nonparametric binary regression; Polya distribution

Co-authors: Alessandro Zito, Huiyan Sang, David Dunson

2.32. Yisheng Li

Institution: University of Texas MD Anderson Cancer Center

Department: Department of Biostatistics

Email: ysli@mdanderson.org

Website: https://faculty.mdanderson.org/profiles/yisheng_li.html

Talk title:

Bayesian nonparametric model for heterogeneous treatment effects with zero-inflated data

Abstract:

One goal of precision medicine is to develop effective treatments for patients by tailoring to their individual demographic, clinical, and/or genetic characteristics. To achieve this goal, statistical models must be developed that can identify and evaluate potentially heterogeneous treatment effects in a robust manner. The oft-cited existing methods for assessing treatment effect heterogeneity are based upon parametric models with interactions or conditioning on covariate values, the performance of which is sensitive to the omission of important covariates and/or the choice of their values. We propose a new Bayesian nonparametric (BNP) method for estimating heterogeneous causal effects in studies with zero-inflated outcome data, which arise commonly in health-related studies. We employ the enriched Dirichlet process (EDP) mixture in our BNP approach, establishing a connection between an outcome DP mixture and a covariate DP mixture. This enables us to estimate posterior distributions concurrently, facilitating flexible inference regarding individual causal effects. We show in a set of simulation studies that the proposed method outperforms two other BNP methods in terms of bias and mean squared error (MSE) of the conditional average treatment effect estimates. In particular, the proposed model has the advantage of appropriately reflecting uncertainty in regions where the overlap condition is violated compared to other competing models. We apply the proposed method to a study of the relationship between heart radiation dose parameters and the blood level of high-sensitivity cardiac troponin T (hs-cTnT) to examine if the effect of a high mean heart radiation dose on hs-cTnT varies by baseline characteristics.

Keywords: enriched Dirichlet process, heterogeneous effects, high-sensitivity cardiac troponin T, missing at random

Co-authors: Chanmin Kim, Ting Xu, Zhongxing Liao

2.33. Linxi Liu

Institution: University of Pittsburgh

Department: Department of Statistics

Email: li155@pitt.edu

Website: <https://scholar.google.com/citations?user=tCLE5aUAAAAJ&hl=en>

Talk title:

A generalized Bayesian tree ensemble for nonparametric density estimation

Abstract:

Tree-based methods cover a class of effective and computationally efficient nonparametric learning algorithms for a wide range of inference objectives. Ensembles of trees---such as forests, boosting, and Bayesian additive regression trees---are among the most successful and widely adopted algorithms, even when facing the competition of deep learning methods, given their ability to capture nonlinear effects and superior empirical performance in supervised problems. In unsupervised learning, tree-based methods have been proven effective in a range of statistical inference tasks, such as density estimation, clustering, multi-resolution sketching of distributional variations, localization of high-density regions, and data compression. In this work, we address the density estimation problem, a fundamental task in unsupervised learning, and introduce a novel generalized Bayesian tree ensemble approach for density estimation, termed "Gibbs density forests". Unlike classical Bayesian approaches that rely on likelihood-based inference, our method employs risk-based generalized Bayesian inference, achieving an optimal bias-variance tradeoff through information-risk minimization (Zhang, 2006). An appropriately chosen risk function plays a central role under this framework. We propose a new risk function for density forests and show that the risk function can effectively measure goodness of each forest. To reduce correlation between the trees in the forest, we fit each tree under a small perturbation. Under mild conditions, we demonstrate that the Gibbs posterior on forests under the generalized Bayesian framework contracts fast around the true density. In specific, for density functions lying in certain H^∞ spaces, we show that the Gibbs density forest can achieve faster convergence compared to individual trees. Additionally, the generalized Bayesian framework naturally facilitates a stochastic search over the forest space. We provide several numerical results to illustrate the performance of Gibbs density forests in moderately high-dimensional settings.

Keywords: Tree-based methods; generalized Bayesian inference; multivariate density estimation

Co-authors: Li Ma, Duke University

2.34. Zihang Lu

Institution: Queen's University

Department: Department of Public Health Sciences and Department of Mathematics and Statistics

Email: zihang.lu@queensu.ca

Talk title:

A Bayesian non-parametric mixture sparse factor model for clustering high-dimensional longitudinal data

Abstract:

Recent advances in engineering technologies have enabled the collection of a large number of longitudinal features. However, analyzing such kind of data can be difficult due to its high dimensionality, heterogeneity and computational challenges. In this talk, we will discuss a novel Bayesian nonparametric mixture model for clustering high-dimensional mixed-type (eg, continuous, discrete and categorical) longitudinal features. We employ a sparse factor model on the joint distribution of random effects and the key idea is to induce clustering at the latent factor level instead of the original data to escape the curse of dimensionality. The number of clusters is estimated through a Dirichlet process prior. We will also discuss an efficient Gibbs sampler to estimate the posterior distribution of the model parameters. Analysis of real and simulated data will be presented and discussed.

Keywords: Bayesian nonparametric model; High dimensional data; Longitudinal data; Mixture model; Sparse factor model;

Co-authors: Noirrit Kiran Chandra

2.35. Pulong Ma

Institution: Iowa State University

Department: Department of Statistics

Email: plma@iastate.edu

Talk title:

Residual Tree Gaussian Processes

Abstract:

Gaussian processes (GP) enjoy wide popularity in spatial statistics, uncertainty quantification, and machine learning. With the advance of measurement technologies and increasing computing power, large numbers of measurements and large-scale numerical simulations make traditional GP models and computational strategies inadequate in dealing with spatially heterogeneous and big data, especially in multi-dimensional domains. In recent years, several multi-scale or tree-based extensions of the GP have been introduced to model spatial nonstationarity and/or achieve scalable computation. In this talk, I introduce a new Bayesian tree-based GP inference framework called residual tree GP (ResGP). ResTGP combines key features of the treed GP and the multi-resolution GP, thereby enjoying the computational efficiency of the former and the flexibility of the latter. Our main idea is to decompose a Gaussian process as well as the data at a cascade of resolutions across locations through iteratively computing predictive and residual

processes, thereby characterizing the underlying covariance structure and achieving divide-and-conquer on the data points simultaneously. We also introduce a new computational strategy for Bayesian inference for ResTGP that does not rely on Metropolis-Hastings based stochastic tree search algorithms but is based on recursive message passing. We demonstrate the predictive performance of the proposed model with extensive numerical examples and also provide posterior consistency results for the proposed model in the nonparametric regression setting.

Keywords: Gaussian process; multi-scale modeling; random partitioning; stochastic tree

Co-authors: Li Ma, Duke University

2.36. Giuseppe Mignemi

Institution: Bocconi University

Department: Institute for Data Science and Analytics (BIDSA)

Email: giuseppe.mignemi@unibocconi.it

Website: <https://gmignemi.github.io/>

Talk title:

Bayesian nonparametric estimation of species richness: flexible and analytically tractable models

Abstract:

Data originating from populations/communities of different types/species arise in various applied settings and pose several challenging inferential issues. For instance, in Ecology, the assessment of species richness provides a measure of diversity, which is important for assessing environmental sustainability and for supporting policymakers' decision-making. Based on the sampling data from a population, the estimation of species richness (observed plus undetected) is statistically challenging, especially for highly heterogeneous populations with many rare species. Most of the contributions to this topic have been developed from a frequentist perspective, and despite their popularity, they have some limitations. The parametric methods rely either on conditional or unconditional maximum likelihood estimators which may not be obtainable in some cases due to convergence issues. The existing nonparametric estimators overcome these drawbacks, but an arbitrary cut-off value is needed to separate species frequencies into "rare" and "abundant". As a result, the choice of the estimator is not always straightforward and is led by heuristic reasons. On the other side, early contributions from a Bayesian perspective assume a symmetric Dirichlet distribution as a model for the species proportions and a zero-truncated negative binomial distribution on the number of species. Alternative more recent approaches rely either on the so-called broken stick sequential model or on the asymmetric multinomial Dirichlet distribution as a model for the species proportions. These methods make use of a parametric specification for modeling the data. We show how a Bayesian nonparametric model, based on a discrete random probability measure, naturally leads to the use of flexible and analytically tractable models for random partitions. Specifically, we overcome most of the abovementioned methods' drawbacks by relying on Gibbs-type priors, implying a mixture of symmetric Dirichlet distributions as a model for the species proportions. We derive Exchangeable Probability Partition Functions for different mixing distributions and use them as empirical Bayes estimators for the hyperparameters, i.e. the mixing

distribution and the symmetric Dirichlet parameters. Results from simulation studies suggest our proposal outperforms the other estimators as it is the most flexible across different data generative processes. Applications to real ecological data are presented.

Keywords: Gibbs-type priors, Species ramplng, Species richness, Ecological data, Empirical Bayes

Co-authors: Antonio Lijoi, Igor Prünster

2.37. Swapnil Mishra

Institution: National University of Singapore

Department: Saw Swee Hock School of Public Health

Email: s.mishra@nus.edu.sg

Website: <https://mlgh-sg.com/>

Talk title:

Deep Generative Spatial Priors: Efficient Inference with aggVAE

Abstract:

Obtaining accurate high-resolution estimates from lower-resolution data is a common challenge across various fields. Often, measurements or data are available only at broader administrative or aggregate levels, such as regional or national scales, while finer-grained details at sub-regional or local levels are scarce or nonexistent. This limitation can hinder effective decision-making, resource allocation, and the implementation of targeted strategies. For example, having comprehensive data at the state level but lacking detailed information at the city or neighborhood level can complicate the development of tailored policies or interventions. Addressing the gap between available lower-resolution estimates and the need for higher-resolution insights is essential for enhancing understanding, optimizing planning processes, and ensuring that strategies are appropriately informed by detailed, localized information. Bayesian Hierarchical models are widely used for areal data, such as aggregated disease prevalence at specific administrative levels, due to their flexibility and ability to incorporate parameter uncertainty. However, these models rely on spatial priors based on neighboring geographic boundaries, limiting their ability to capture continuous spatial variation and population heterogeneity within regions. They also struggle to estimate high-resolution prevalence from low-resolution data, such as extrapolating provincial estimates to district levels. To address these limitations, this work introduces extension of aggregated Variational Autoencoder (aggVAE) method. aggVAE models latent Gaussian Processes (GP) on an artificial grid, enabling unit-level estimates that can be aggregated to any desired region. By utilizing a Variational Autoencoder, aggVAE encodes spatial GPs into a low-dimensional latent space, significantly reducing computational costs compared to traditional Markov Chain Monte Carlo (MCMC) methods—from approximately four hours to around fifteen minutes for posterior evaluations. Applied to influenza prevalence estimation, aggVAE effectively extrapolated low-resolution data from nine regions to predict prevalence across 49 subregions. While initial results showed some underestimation, refinements such as tuning the latent space size and GP kernel functions promise enhanced accuracy. This methodological innovation offers a scalable and computationally efficient solution for high-resolution mapping, particularly valuable in resource-limited settings. By maintaining spatial coherence and

improving scalability, our work advances geostatistical modeling and supports better public policy planning.

Keywords: Gaussian Processes; MCMC; spatial modeling; deep generative models

Co-authors: Imantha Gunasekera Jessica Widyawati

2.38. Kai Nguyen

Institution: University of Texas at Austin

Department: Statistics and Data Sciences

Email: khainb@utexas.edu

Website: <https://khainb.com>

Talk title:

Summarizing Bayesian Nonparametric Mixture Posterior - Sliced Optimal Transport Metrics for Gaussian Mixtures

Abstract:

Existing summarization methods for mixture models focus on identifying a point estimate of the implied random partition for clustering, with density estimation as a secondary goal~\citep{wade2018bayesian,dahl2022search}. We propose a new approach for summarizing the posterior on mixing measures in Bayesian nonparametric mixture models, prioritizing density estimation as an inference target. One of the key features is the model-agnostic nature of the approach, which remains valid under arbitrarily complex dependence structures in the underlying sampling model. Using a decision-theoretic framework, our method identifies a point estimate by minimizing the posterior expected loss, employing a discrepancy between mixing measures as the loss function. Estimating the mixing measure implies inference on the mixture density. Exploiting the discrete nature of the mixing measure, we use a version of sliced Wasserstein distance. Additionally, we introduce two specific variants for Gaussian mixtures. The first, mixed sliced Wasserstein, applies generalized geodesic projections on the product of the Euclidean space and the manifold of symmetric positive definite matrices. The second, sliced mixture Wasserstein, leverages the linearity of Gaussian mixture measures for efficient projection. Empirical analyses on simulated and real data demonstrate that the proposed approach achieves comparable clustering performance while yielding more accurate density estimation.

Co-authors: Peter Mueller

2.39. Dennis Nieman

Institution: Humboldt University Berlin

Department: department of mathematics

Email: dennis.nieman@hu-berlin.de

Website: <https://dennisnieman.github.io>

Talk title:

Adaptive variational inference for Gaussian processes

Abstract:

We study model selection using variational Bayes in a general setting. We establish a general variational procedure to approximate a hierarchical Bayes posterior. Under regularity assumptions on the prior, hyper-prior and the variational class, this is shown to lead variational posteriors that adapt to smoothness. Theoretical results are complemented with simulation and real data examples.

Keywords: Adaptation, model selection, variational Bayes, hyperparameters, Gaussian processes

Co-authors: Botond Szabo

2.40. Yu-Chien Ning

Institution: Harvard T.H. Chan School of Public Health

Department: Department of Epidemiology

Email: bycning@hsph.harvard.edu

Website: <https://bo-ning.github.io/>

Talk title:

Empirical Bayes large-scale multiple testing for high-dimensional binary outcome data

Abstract:

In this talk, I will introduce the multiple testing problem for high-dimensional sparse binary sequences. I adopt an empirical Bayes approach to estimate potentially sparse sequences with Bernoulli noise and find that, when choosing the conjugate spike-and-slab uniform prior, the multiple testing procedure (also known as the ℓ -value procedure) constructed directly from the posterior distribution tends to be overly conservative in estimating the false discovery rate (FDR). To address this issue, I propose two new procedures, for which I obtained the first series of sharp frequentist theoretical results. My results demonstrate that both procedures can effectively control the FDR uniformly for signals under the sparsity assumption. Finally, numerical experiments are provided to validate the theory in finite samples.

Keywords: Multiple testing, empirical Bayes, frequentist analysis of posterior distribution, high-dimensional inference, Binomial distribution

Co-authors: I am the single author

2.41. Widemberg Nobre

Institution: Universidade Federal do Rio de Janeiro

Department: Department of Statistical Methods

Email: widemberg@im.ufrj.br

Website: <https://sites.google.com/im.ufrj.br/widemberg-nobre>

Talk title:

A Bayesian multilevel approach for causal inference under mis-specification

Abstract:

We investigate Bayesian methods for dealing causal effects estimation in light of mis-specified models. Specifically, we focus on propensity score regression while relaxing unmeasured confounding assumption. Motivated by an application to Tuberculosis treatment in Brazil, the proposal involves the use of multilevel models and non-parametric techniques in order to achieve good Bayesian and Frequentist properties. Simulation studies have shown good performance of the proposed method when the number of units per cluster increases uniformly.

Keywords: Bayesian Analysis, Propensity score regression, Latent ignorability, Statistical Method

Co-authors: Alexandra Schmidt, David Stephens, Erica Moodie

2.42. Sean O'Hagan

Institution: University of Chicago

Department: Statistics

Email: seanohagan@uchicago.edu

Website: <https://seanohagan.com>

Talk title:

Tree Bandits for Generative Bayes

Abstract:

In generative models with obscured likelihood, Approximate Bayesian Computation (ABC) is often the tool of last resort for inference. However, ABC demands many prior parameter trials to keep only a small fraction that passes an acceptance test. To accelerate ABC rejection sampling, this paper develops a self-aware framework that learns from past trials and errors. We apply recursive partitioning classifiers on the ABC lookup table as nonparametric tree-based surrogates to the intractable likelihood function (and consequently, the posterior). These trees sequentially refine high-likelihood regions into boxes. Each box is regarded as an arm in a binary bandit problem treating ABC acceptance as a reward. Each arm has a proclivity for being chosen for the next ABC evaluation, depending on the prior distribution and past rejections. The method places more splits in those areas where the likelihood resides, shying away from low-probability regions destined for ABC rejections. We demonstrate accurate ABC approximability at much lower simulation cost. We justify the use of our tree-based bandit algorithms with nearly optimal regret bounds. Finally, we successfully apply our approach to the problem of masked image classification using deep generative models.

Keywords: Approximate Bayesian Computation; Tree-based models; Thompson sampling; likelihood-free inference

Co-authors: Jungeum Kim, Veronika Ročková

2.43. Francesca Panero

Institution: Sapienza University of Rome

Email: francesca.panero@uniroma1.it

Website: <https://francescapanero.github.io>

Talk title:

Fast additive Gaussian Processes to model food security

Abstract:

Measuring food security is a complex task, involving taking a decision on which quantitative definition to pick and running continuous surveys to estimate it. Surveys, though, are expensive and sometimes difficult or dangerous to run. In the last years, organisations and governments have been investigating to which extent it is possible to use statistical models to predict food insecurity measures using auxiliary covariates which are more easily accessible. In this talk, we will explore the possibility of doing so introducing a family of additive Gaussian Processes that avoid the well-known curse of dimensionality thanks to a clever Gram matrix decomposition which exploits the Kronecker product on the cartesian space of the covariates. Our models are identifiable and allows to incorporate various degrees of main and interaction effects, without resorting to the main effects only or the saturated model.

Keywords: Gaussian processes; additive models; spatio-temporal; computational; food security

Co-authors: Sahoko Ishida (University of Oxford)

2.44. Michele Peruzzi

Institution: University of Michigan

Department: Biostatistics

Email: peruzzi@umich.edu

Talk title:

Inside-out cross-covariance for spatial multivariate data

Abstract:

The increasing prevalence of multivariate spatial data in fields like ecology, climate sciences, genomics, and cancer research has emphasized the need for innovative spatially-aware models that balance flexibility, interpretability, and scalability. We develop inside-out cross-covariance (IOX) models for multivariate Gaussian Processes. IOX is a novel approach to multivariate spatial inference that addresses limitations in traditional models such as the ubiquitous Linear Model of Coregionalization (LMC) and the multivariate Matérn model. IOX leads to valid cross-covariance matrix functions which we interpret as inducing spatial dependence on independent replicates of a correlated random vector. The resulting sample cross-covariance matrices are “inside-out” relative to the LMC. However, unlike LMCs, our methods offer direct marginal inference, easy prior elicitation of covariance parameters, the ability to model outcomes with unequal smoothness, and flexible dimension reduction. As a covariance model for a q -variate Gaussian process, IOX leads to scalable models for noisy vector data as well as flexible latent models. For large n cases, IOX complements Vecchia approximations and related process-based methods based on sparse graphical models. This work provides the theoretical development for IOX and demonstrates the superior performance of IOX on synthetic datasets as well as on

colorectal cancer proteomics data. An R package implementing the proposed methods is available at github.com/mkln/spiox.

Keywords: Gaussian process; co-kriging; dimension reduction; coregionalization; Bayesian hierarchical models

Co-authors: None

2.45. Giovanni Rebaudo

Institution: University of Torino and Collegio Carlo Alberto

Email: giovanni.rebaudo@unito.it

Website: <https://giovannirebaudo.github.io>

Talk title:

Probabilistic discovery of new species and homogeneous subpopulations

Abstract:

Bayesian hierarchical models are particularly effective when observations come from different groups or studies since they naturally allow borrowing information across groups while allowing the subjects in the same group to share the same unknown distribution. In the so-called partially exchangeable framework, we consider discrete observations arising from species-sampling models. In particular, we specify a discrete prior probability on the unknown distributions governing the observations in the different groups, allowing groups to share the same distribution for the observations, i.e., to arise from the same population. Moreover, in each group, the observations are distributed according to a proper species-sampling model with common atoms across groups allowing them to share species within and across groups. Our goal is two-fold: predict future observations in different groups and test for homogeneity among populations. The former problem is usually the main focus in species-sampling models, while the latter task is not feasible with the state-of-the-art methods available in the literature, since they generally consider all the populations distributions different almost surely. To do so, we extend what is arguably the most popular species sampling model in Bayesian nonparametrics in this partially exchangeable framework, that is the Hierarchical Pitman-Yor process. Adding a latent structure to the distributions allows us to have ties across the population's distributions, performing both the above-mentioned tasks at the same time. We show that the distribution of the induced random partition admits analytically tractable expression, which allows us to gain a deeper insight into the theoretical properties of the model. Posterior inference is feasible thanks to a Markov Chain Monte Carlo routine which allows us to evaluate the functionals of interest. Finally, a Bayesian nonparametric test of homogeneity between different populations is displayed and is complemented by real-data applications.

Keywords: Multivariate Species Sampling Models; Species Sampling Problems; Partial Exchangeability; Hierarchical Pitman-Yor Process; Nested Nonparametric Process

Co-authors: Augusto Fasano, Antonio Lijoi, Igor Prünster

2.46. Federica Stolf

Institution: Duke University

Email: federica.stolf@duke.edu

Website: <https://federicastolf.netlify.app/>

Talk title:

Infinite Joint Species Distribution Models

Abstract:

Joint species distribution models are popular in ecology for modeling covariate effects on species occurrence, while characterizing cross-species dependence. Data consist of multivariate binary indicators of the occurrences of different species in each sample, along with sample-specific covariates. A key problem is that current models implicitly assume that the list of species under consideration is predefined and finite, while for highly diverse groups of organisms, it is impossible to anticipate which species will be observed in a study and discovery of unknown species is common. To address this challenge, we propose a novel class of models for inference on multivariate binary outcomes having a growing dimension. We develop a general new infinite joint species distribution modeling framework, borrowing key ideas from the literature on infinite latent feature models to appropriately extend traditional multivariate probit models to accommodate large numbers of rare species and new species discovery. We discuss theoretical properties of the proposed modeling paradigm and implement efficient algorithms for posterior computation. Our approach maintains many of the desirable properties of the popular Indian buffet process, while addressing important statistical complexities that arise in applications, such as dependence in species occurrence, covariate effects and the need to borrow information in conducting inferences for rare species. Simulation studies and applications to fungal biodiversity data provide compelling support for the new modeling class.

Keywords: Bayesian; Ecology; Indian buffet process; Multivariate binary response; Multivariate probit model

Co-authors: David Dunson (Duke University)

2.47. Johan Van der Molen Moris

Institution: Pontifical Catholic University of Chile

Email: johan.vandermolen@uc.cl

Talk title:

Efficient computation of the Posterior Similarity Matrix for Bayesian nonparametric clustering

Abstract:

Mixture models, especially Dirichlet Process mixtures, are widely used in Bayesian clustering. The Posterior Similarity Matrix (PSM), typically estimated using Markov Chain Monte Carlo (MCMC) methods, is crucial for understanding the data's clustering structure and summarising clustering uncertainty. However, MCMC in this context can be computationally intensive, requires careful initialization, and often suffers from slow convergence, frequently exploring only a small fraction of the possible data partitions —

sometimes even becoming trapped at a single partition. This work introduces a novel and faster alternative to MCMC for PSM computation. It is based on an algorithm that directly approximates PSM entries by leveraging an analytical expression for a Dirichlet Process mixture model. This approach not only reduces computational cost but also improves the accuracy of PSM estimation, thus providing a more precise representation of the corresponding posterior distribution. The impact of this approach on inferred clusters is also assessed for different estimates based on the PSM. Results from various datasets demonstrate its advantages over MCMC, with discussions on potential extensions and challenges, particularly for large-scale datasets.

Keywords: Posterior Similarity Matrix; Dirichlet Process Mixtures; Bayesian Clustering; Computational Efficiency

Co-authors: None

2.48. Yunyi Zhang

Institution: Bocconi University

Department: BIDSa

Email: junyi.zhang@unibocconi.it

Talk title:

Dependent beta process in Bayesian Factor Analysis

Abstract:

Bayesian factor models are a popular tool for factor analysis. Current state-of-the-art Bayesian factor analysis approaches leverage the beta-Bernoulli process prior to characterise the factors. This prior, however, ignores the diversity across multiple populations. In this talk, we propose a new framework based on the dependent beta processes, which allows the development of an innovative hierarchical modelling methodology for Bayesian factor analysis. We discuss the distributional properties of the dependent beta process. Then, we fit the beta-NRMI process into the celebrated beta process factor analysis model and devise the MCMC posterior inference schemes. We illustrate the usefulness of our hierarchical modelling methodology using a psychological assessment dataset.

Keywords: Bayesian modelling; beta process; completely random measures; factor analysis; Indian buffet process

Co-authors: Prof. Antonio Lijoi (Bocconi University) Prof. Igor Prünster (Bocconi University)

2.49. Yichen Zhu

Institution: Bocconi University

Department: Institute for Data Science and Analytics

Email: yichen.zhu@unibocconi.it

Website: <https://sites.google.com/view/yichenzhu/home>

Talk title:

Vecchia Gaussian Processes: Probabilistic Properties, Minimax Rates and Methodological Developments

Abstract:

Gaussian Processes (GPs) are widely used to model dependency in spatial statistics and machine learning, yet the exact computation suffers an intractable time complexity of $O(n^3)$. Vecchia approximation allows scalable Bayesian inference of GPs in $O(n)$ time by introducing sparsity in the spatial dependency structure that is characterized by a directed acyclic graph (DAG). Despite the popularity in practice, it is still unclear how to choose the DAG structure and there are still no theoretical guarantees in nonparametric settings. In this paper, we systematically study the Vecchia GPs as standalone stochastic processes and uncover important probabilistic properties and statistical results in methodology and theory. For probabilistic properties, we prove that the conditional distributions of the Mat $\backslash\{e\}$ GPs, as well as the Vecchia approximations of the Mat $\backslash\{e\}$ GPs, can be characterized by polynomials. This allows us to prove a series of results regarding the small ball probabilities and RKHSs of Vecchia GPs. For statistical methodology, we provide a principled guideline to choose parent sets as norming sets with fixed cardinality and provide detailed algorithms following such guidelines. For statistical theory, we prove posterior contraction rates for applying Vecchia GPs to regression problems, where minimax optimality is achieved by optimally tuned GPs via either oracle rescaling or hierarchical Bayesian methods. Our theory and methodology are demonstrated with numerical studies, where we also provide efficient implementation of our methods in C++ with R interfaces.

Keywords: Gaussian Process, Minimax Rate, Norming Set, Vecchia Approximation

Co-authors: Botond Szabo

2.50. Alessandro Zito

Institution: Harvard T.H. Chan School of Public Health

Department: Department of Biostatistics

Email: azito@hsph.harvard.edu

Website: <https://alessandrozito.github.io/web/>

Talk title:

Localized mutational signature analysis via inhomogeneous Poisson process factorizations

Abstract:

In cancer genomics, it has become standard practice to summarize mutational patterns in the genomes of tumor cells using non-negative matrix factorization. However, the current status of mutational signature analysis considers only an aggregate scale by summing all mutations detected along the genome for a given mutation type, ignoring potential local patterns that target certain genomic regions. We introduce a novel class of stochastic processes to infer mutational signatures and their activity as they vary along the genome. Using a location-dependent intensity function, our inhomogeneous Poisson process factorization generalizes the standard non-negative matrix factorization typically used for count data. We illustrate how our approach is highly flexible and can account for both

continuous and discrete covariates characterizing patients and genomic variables such as gene density and chromatin accessibility. By selecting suitable hierarchical priors, our method also allows for a parsimonious representation of the signature activities, effectively zeroing out the contribution of unneeded signatures in the posterior. Inference can be easily performed via Markov Chain Monte Carlo algorithms and variational inference. We apply our method to the pan-cancer analysis of whole genome data, which comprises 30 tumor types and 3700 patients.

Keywords: Poisson process, non-negative matrix factorization, mutational signature analysis, cancer genomics

Co-authors: Catherine Xue Giovanni Parmigiani Jeffrey W. Miller

2.51. Dafne Zorzetto

Institution: Brown University

Department: Data Science Institute

Email: dafne_zorzetto@brown.edu

Website: <https://www.dafnezorzetto.com/>

Talk title:

Bayesian Nonparametrics for Principal Stratification with Continuous Post-Treatment Variables

Abstract:

Principal stratification provides a causal inference framework that allows adjustment for confounded post-treatment variables when comparing treatments. Although the literature has focused mainly on binary post-treatment variables, there is a growing interest in principal stratification involving continuous post-treatment variables. However, characterizing the latent principal strata with a continuous post-treatment presents a significant challenge, which is further complicated in observational studies where the treatment is not randomized. In our work, we introduce the Confounders-Aware SHared atoms BAYesian mixture (CASBAH), a novel approach for principal stratification with continuous post-treatment variables that can be directly applied to observational studies. CASBAH leverages a dependent Dirichlet process--utilizing shared atoms across treatment levels--to effectively control for measured confounders and facilitate information sharing between treatment groups in identifying principal strata membership. CASBAH also offers a comprehensive quantification of uncertainty surrounding the membership of the principal strata. Through Monte Carlo simulations, we show that the proposed methodology has excellent performance in characterizing the latent principal strata and estimating the effects of treatment on post-treatment variables and outcomes. Finally, CASBAH is applied to a case study in which we estimate the causal effects of US national air quality regulations on pollution levels and health outcomes.

Keywords: Causal Inference; Dependent Dirichlet Process; Environmental Health; Principal Strata; Shared Atoms Mixture Model

Co-authors: Antonio Canale, Fabrizia Mealli, Francesca Dominici, and Falco J. Bargagli-Stoffi

3. Posters

3.1. Entejar Alam

Institution: University of Texas at Austin

Department: Department of Statistics and Data Sciences

Email: entejar@utexas.edu

Talk title:

DPGLM: A Semiparametric Bayesian GLM with Inhomogeneous Normalized Random Measures

Keywords: Dependent Dirichlet process; Normalized random measures; Density regression; Semiparametric generalized linear model

Co-authors: Peter Müller Paul J. Rathouz

3.2. Sarwan Ali

Institution: Georgia State University

Department: Computer Science

Email: sali85@student.gsu.edu

Website: <https://sarwanpasha.github.io/>

Talk title:

Bayesian Nonparametric Models for Integrative Genomic Data Analysis in Healthcare Applications

Abstract:

Moreover, the framework can be extended to link subject-level partitions across datasets via an underlying shared latent structure, promoting information sharing and facilitating robust, interpretable inference. By jointly modelling within-subject dependencies and marginal relationships across views, the approach enhances both the flexibility and interpretability of the resulting clustering structures.

Keywords: Genomic Data Integration; Healthcare Applications; Precision Medicine; Multi-Omics Analysis

Co-authors: Sarwan Ali

3.3. Youngsoo Baek

Institution: Duke University

Department: Biostatistics

Email: youngsoo.baek@duke.edu

Talk title:

Frequentist Properties of Bayes Posteriors in Nonlinear Inverse Problems

Keywords: Inverse problems; Gaussian process; Frequentist coverage

Co-authors: Katerina Papagiannouli, Sayan Mukherjee

3.4. Falco J. Bargagli Stoffi

Institution: UCLA

Department: Biostatistics

Email: falco@ucla.edu

Website: <https://www.falcobargaglistoffi.com/>

Talk title:

Confounder-dependent Bayesian mixture model: Characterizing heterogeneity of causal effects in air pollution epidemiology

Abstract:

We also describe scalable Bayesian solutions, such as Bayesian Distance Clustering, which strike a balance between computational efficiency and probabilistic rigour. These methods are particularly effective in handling large, high-dimensional datasets, offering accurate predictive performance while maintaining scalability.

Keywords: Bayesian nonparametrics, causal inference, dependent Dirichlet process, group average treatment effect, heterogeneous causal effects

Co-authors: Dafne Zorzetto, Antonio Canale, Francesca Dominici

3.5. Nicola Barileto

Institution: UNIVERSITY OF TEXAS AT AUSTIN

Department: Statistics and Data Sciences

Email: nicola.bariletto@utexas.edu

Website: <https://nbariletto.github.io/>

Talk title:

A new look at strong posterior consistency through the lenses of identifiability

Keywords: posterior consistency; posterior asymptotics; identifiability; infinite mixture models

Co-authors: Bernardo Flores Lopez, Stephen G. Walker

3.6. Felipe Bedoya

Institution: Rice University

Department: Statistics

Email: fb31@rice.edu

Website: <https://meng.rice.edu/students/>

Talk title:

Large Bayesian prior: an adaptative and optimal method

Abstract:

We validate these methods through extensive simulations and applications to real-world datasets, demonstrating their effectiveness across a variety of contexts. These advancements provide a robust and versatile toolkit for addressing modern data analysis challenges, particularly in the context of multiview and longitudinal data.

Keywords: Nonparametric bayesian estimation, random series, basis functions.

Co-authors: Meng Li, Associate Professor with tenure at Rice University.

3.7. Davis Berlind

Institution: UCLA

Department: Statistics

Email: dberlind@g.ucla.edu

Website: <https://davis-berlind.github.io/>

Talk title:

Bayesian change-point detection and uncertainty quantification via variational inference

Abstract:

Change-point detection is a well-studied problem in statistical inference dating back to the use of control charts during World War II. Despite this long history, there are only a few methods that provide measures of uncertainty for their estimated change-point locations. We introduce a novel Bayesian method that can detect multiple structural breaks in the mean and variance of a sequence and return α -level credible sets around the locations of the breaks. In the case of a single change in the mean and/or variance of the data, our method consistently localizes the change at a rate that is minimax optimal up to either i) a $\log T$ factor if we observe an independent sub-Gaussian sequence, or ii) a $\log^2 T$ factor for an α -mixing sequence. For multivariate data, our method continues to localize mean changes at the optimal rate, and in high-dimensions exact recovery of the change-point is possible when the mean signal is dense or spatially clustered and $d \lesssim \log T$. For multiple change-points, we modularly combine many single change-point models, which allows for efficient approximation of the posterior distribution using a variational algorithm. Extensive simulation studies demonstrate that our method is competitive with the state-of-the-art and returns credible sets that are an order of magnitude smaller than those returned by competitors, all without sacrificing nominal coverage guarantees. We test our method on real data by detecting i) changes in the lithological structure of an oil well, and ii) gating of the ion channels in the outer membrane of a bacterial cell.

Keywords: change-point detection; variational inference; localization rate; high-dimensional

Co-authors: Oscar Madrid Padilla Lorenzo Capello

3.8. Christian Bernal Zelaya

Institution: University of California, Irvine

Department: Statistics

Email: cbernalz@uci.edu

Talk title:

Forecasting healthcare utilization using wastewater surveillance data.

Abstract:

Forecasting healthcare utilization is a critical task for now and future public crises caused by infectious diseases. Several researchers have used Bayesian nonparametrics to forecast healthcare utilization. However, many of these methods have utilized cases or tests along with the healthcare utility data. This can lead to problems due to factors such as underreporting, misreporting, and changes in testing protocols. Those who have used wastewater surveillance data have not effectively used it to forecast healthcare utilization. This paper attempts to address this gap by using wastewater surveillance data to forecast healthcare utilization. The removal of problems seen in other forms of surveillance data, such as cases or tests, makes wastewater surveillance data an ideal candidate for forecasting healthcare utilization such as in forecasting hospitalizations caused by some infectious disease. This paper will explore this use, in combination with Bayesian nonparametric methods, to forecast hospitalizations. The results that come from fitting Orange County, California hospitalization data will be compared to a Bayesian nonparametric model that does not use wastewater surveillance data. This paper will also explore two methods for picking the hyperparameters of the Bayesian nonparametric models. A simulation study will be conducted to validate our model. Ultimately, this paper will compare our model against a wastewater only model when estimating the effective reproduction number curve.

Keywords: Forecasting; infectious diseases; wastewater; Bayesian nonparametrics

Co-authors: Volodymyr Minin

3.9. Dila Bhandari

Institution: Tribhuvan University, Kathmandu, Nepal

Department: Statistics

Email: drbhandari2012@gmail.com

Talk title:

Forecasting of the Nepalese Tourism Industry Using a Modeling Approach

Abstract:

Nepal has become one of the largest economies in the world to which tourism and hospitality have significantly contributed; however, the growth rate of the tourism industry has been greatly affected during the COVID-19 pandemic. In this paper, we have used the modeling approach to analyze and understand the growth pattern of the Nepalese tourism industry. To achieve this, we consider the data on international tourist arrivals before and after the lockdown. The Dickey-Fuller test, AIC, and BIC methods are used to obtain the best-fitted model and further, the accuracy of the obtained model is also analyzed. Data

and forecasting indicate that the weather and public holidays significantly affect the tourism industry. Employing time-series analysis, machine learning algorithms, and econometric modeling, the research aims to accurately forecast tourist arrivals, revenue generation, and sector growth. The results highlight key trends and potential disruptions, offering valuable insights for policymakers, stakeholders, and industry practitioners. By leveraging these forecasts, the study provides strategic recommendations to enhance decision-making and foster sustainable growth within the Nepali tourism industry.

Keywords: AIC, BIC, Tourism, forecasting, AI

Co-authors: None

3.10. Srijato Bhattacharyya

Institution: Texas A&M University

Department: Statistics

Email: srijato@tamu.edu

Talk title:

A Random Spanning Tree based Bayesian Partitioning model for Spatial Clustering of Multivariate Spatial data

Abstract:

In disease-mapping applications, incorporating spatial structure by estimating latent spatial clustering is crucial for accurately identifying regions based on disease homogeneity. Additionally, estimating the underlying cluster-wise mean trend of diseases helps capture spatial patterns effectively. This study proposes a novel Bayesian hierarchical model built on a set-partitioning prior over random spanning tree cuts in order to model Multivariate Spatial data. The latent spatial clustering structure is captured through a well-designed random spanning tree prior while the mean trend and dependence of the disease are modeled through a Multivariate Normal-Inverse Wishart framework. We develop an efficient Markov chain Monte Carlo (MCMC) algorithm tailored to this model setup for posterior inference. We illustrate our method with appropriate simulation studies and real data analysis of cancer mortality data.

Keywords: Inverse Wishart, Random Spanning Trees, Reversible-Jump MCMC, Spatial Clustering, Stirling Number of Second Kind

Co-authors: Huiyan Sang, Bani Mallick

3.11. Soumyabrata Bose

Institution: University of Texas at Austin

Department: Department of Statistics and Data Sciences

Email: sbose@utexas.edu

Talk title:

Nonparametric Within-Between Model: Extending BART for Multilevel Modeling

Abstract:

The within-between (WB) model is a robust approach that addresses the constraints inherent in both fixed effects (FE) and random effects (RE) models by distinctly modeling within-group and between-group effects. This paper introduces a nonparametric extension of the Within-Between model for the analysis of hierarchical data using Bayesian Additive Regression Trees (BART). Our extension permits flexible nonlinear relationships while preserving the interpretability benefits of the linear Within-Between framework. We establish theoretical guarantees on posterior concentration rates under appropriate conditions and present a framework for deriving interpretable summaries of the intricate nonparametric effects using surrogate models. Through simulation studies, we demonstrate the superior performance of our approach compared to existing methods, including linear fixed effects, random effects, and standard BART extensions (BART-RE), particularly when the true relationships are nonlinear. We illustrate the practical applicability of our method through its application to the National Education Longitudinal Study, wherein we analyze student dropout status while accounting for both student-level and school-level effects.

Keywords: BART; Multilevel Modelling; Within-Between Model; Nonparametric Regression

Co-authors: Dr. Antonio R. Linero; Dr. Jared S. Murry

3.12. Abhijit Brahme

Institution: University of California, Santa Barbara

Department: Statistics

Email: abhijitbrahme@ucsb.edu

Talk title:

A Nonparametric Bayesian Model for Concave Functions with Applications to Modeling Aging in Sports

Abstract:

In this work, we propose a Bayesian non-parametric model for jointly inferring multiple metrics which co-evolve in time. We introduce a novel Concave Latent Variable Gaussian Process (Conc. LVGP) which allows us to simultaneously model the dependence between observational units through a latent variable, as well as capture the time dependence through a strictly concave function. We conduct inference on the posterior predictive trajectories by developing an MCMC algorithm which learns the concave functions, conditional on a latent embedding. We make spectral method based approximations to the respective covariance functions in order to make this inference method tractable. Finally, we illustrate the applications of our proposed approach by modeling mixed-output performance data as a function of the age of professional basketball players.

Keywords: gaussian process; convex functions; non-parametric; aging; spectral approximation; MCMC;

Co-authors: None

3.13. Martin Burda

Institution: University of Toronto

Department: Economics

Email: martin.burda@utoronto.ca

Website: <https://www.economics.utoronto.ca/index.php/index/person/person/faculty/1055>

Talk title:

Bayesian Adaptive Sparse Copula

Abstract:

Bayesian nonparametric density estimation procedures are typically based on single-scale priors, such as Dirichlet process mixtures. Alternative multiscale density priors built on deep decision trees have a number of well-known advantages, including the ability to characterize abrupt local changes and to provide an estimate with a desired level of resolution. Despite their theoretical appeal, multiscale methods have typically been developed in the literature as univariate. Their multivariate versions are generally very costly to implement in practical applications, rendering such methods infeasible in many empirical cases of interest. One of the key reasons is the rapidly increasing number of multiscale mixture components required to represent dependence structures in higher dimensions. In this paper, we propose a random Bernstein polynomial prior on the unit hypercube of arbitrary dimension with a spike-and-slab shrinkage structure. The polynomial components with near-zero posterior weights are shrunk towards zero and thus omitted from posterior sampling. This results in posterior sparsity of the multiscale decision tree, alleviating the curse of dimensionality. We embed the proposed model in the form of a copula link function along with nonparametric marginals in a composite prior over general spaces of densities. We provide conditions for posterior consistency under the weak topology. We further illustrate the practical use of the model in an application to forecasting the Value at Risk and Expected Shortfall of a financial portfolio in a scenario where sampling from the non-sparse posterior would be infeasible.

Keywords: nonparametric copula, multiscale estimation, Bernstein polynomial

Co-authors: Artem Prokhorov (University of Sydney, Australia)

3.14. Jiachen Cai

Institution: Gladstone Institutes

Department: Data science and biotechnology

Email: jiachen.cai@gladstone.ucsf.edu

Website: <https://www.linkedin.com/in/jiachen-cai-872ab9191/>

Talk title:

Study tumor-immune microenvironment by connecting single-cell RNA-seq data with spatial transcriptomics data using hierarchical Dirichlet processes

Abstract:

Tumor cells interact with immune cells during cancer progression, forming multicellular hubs known as tumor-immune microenvironment (TIMEs) that can affect clinical

outcomes. Technology advancement has enabled the collection of single-cell RNA sequencing (scRNA-seq) and spatial transcriptomics (ST) data to facilitate studying TMEs. It is common in practice that the design of ST panel is informed by the scRNA-seq analysis, which selects marker genes that can distinguish between cell types. Under this type of design, scRNA-seq and ST data have their own unique advantages: scRNA-seq data usually has larger number of genes yet smaller number of cells, whereas ST data has smaller number of genes yet larger number of cells with spatial locations at single-cell resolution. Integrating both types of data is promising to promote the understanding of TMEs, yet existing research rely on only scRNA-seq or only ST data. In this work, we propose to connect scRNA-seq data with ST data in a statistically principled way, by using hierarchical Dirichlet processes. Thanks to the vast number of genes measured in scRNA-seq experiment, this framework will build a comprehensive dictionary of cell functions based on gene expression patterns. Thanks to the preservation of single-cell spatial location in ST technology, we will then apply the built dictionary to ST data for visualizing the spatial organization of cell functions. Furthermore, we will develop a scalable nonnegative matrix factorization method to characterize the potential interactions between different cell functions, which is critical to understanding activities happening in multicellular hubs.

Keywords: hierarchical Dirichlet processes; spatial transcriptomics; scRNA-seq; tumor-immune microenvironment; nonnegative matrix factorization.

Co-authors: Mozghan Saeidi, Jack Demaray, Han Chen, Ragnhild Laursen, Karin Pelka, Barbara E Engelhardt

3.15. Arhit Chakrabarti

Institution: Texas A&M University

Department: Statistics

Email: arhit.chakrabarti@stat.tamu.edu

Talk title:

Global-Local Dirichlet Processes for Clustering Grouped Data in the Presence of Group-Specific Idiosyncratic Variables

Abstract:

We consider the problem of clustering grouped data for which the observations may include group-specific variables in addition to the variables that are shared across groups. This type of data is quite common. For example, in cancer genomic studies, molecular information is available for all cancers whereas cancer-specific clinical information may only be available for certain cancers. Existing grouped clustering methods only consider the shared variables but ignore valuable information from the group-specific variables. To allow for these group-specific variables to aid in the clustering, we propose a novel Bayesian nonparametric approach, termed global-local (GLocal) Dirichlet process, that models the "global-local" structure of the observations across groups. We characterize the GLocal Dirichlet process using the stick-breaking representation and the representation as a limit of a finite mixture model. We theoretically quantify the approximation errors of the truncated prior, the corresponding finite mixture model, and the associated posterior distribution. We develop a novel variational Bayes algorithm for scalable posterior

inference, which we illustrate with extensive simulations and a TCGA pan-gastrointestinal cancer dataset.

Keywords: Bayesian nonparametrics; clustering; group-specific variables; approximation error; variational Bayes

Co-authors: Yang Ni; Debdeep Pati; Bani K. Mallick

3.16. Jongwoo Choi

Institution: University of Connecticut

Department: Statistics

Email: jongwoo.choi@uconn.edu

Website: <https://scholar.google.com/citations?user=3Ltr0akAAAAJ&hl=en>

Talk title:

Robust Bayesian Predictive Model Selection Using Bregman Divergence

Abstract:

In Bayesian model selection, traditional methods often rely on the Kullback-Leibler (KL) divergence, which is highly sensitive to regions where the model underestimates the true probability. For heavy-tailed distributions, this sensitivity often results in selecting models that fit extreme values well but perform poorly in capturing central data patterns. We propose a decision-theoretic predictive model selection framework based on Bregman divergences. Under model misspecification, our method asymptotically selects the model closest to the data-generating process in Bregman divergence. Through simulations and applications with heavy-tailed and skewed data, we demonstrate that our method provides a robust alternative to traditional approaches, accurately capturing central data patterns without overemphasizing tail behavior.

Keywords: Bayesian Model Selection; Robustness; Model Misspecification; Bregman Divergence; Decision Theory

Co-authors: Dipak K. Dey, Neil A. Spencer

3.17. Simone Colombara

Institution: Politecnico di Milano

Department: Department of Mathematics

Email: simone.colombara@polimi.it

Talk title:

Bayesian Non-Parametric Spatio-Temporal Model for Analyzing Crop Yield under Time-Varying Climatic Stress

Abstract:

Crop yield is influenced by complex interactions between climatic variables such as soil temperature and moisture. Traditional models often fail to account for these non-linear and time-varying relationships, limiting the ability to make accurate inferences and predictions. This research develops a Bayesian non-parametric spatio-temporal model to

analyze the joint effects of climatic factors on crop yield. Specifically, the model captures non-linear exposure effects, identifies time periods of heightened plant vulnerability to climate, and performs clustering to reveal regions with similar vulnerability patterns. We specify a spatial BNP model that can accommodate to capture time varying covariates, allowing us to address spatio-temporal heterogeneity in plant response and expand beyond standard scalar-on-function regression approaches. Additionally, our modelling framework allows the clustering of spatial regions based on shared yield responses to climate factors. This clustering could reveal spatial heterogeneity in climate resilience, facilitating the identification of priority areas for agricultural interventions. By identifying periods and regions of vulnerability, this research contributes to the development of targeted strategies to improve climate resilience in agricultural systems.

Keywords: Spatio-Temporal Models; Crop yield prediction; Time-varying climatic effects; Spatial Clustering; Plant vulnerability

Co-authors: Veronica Berrocal, Alessandra Guglielmi

3.18. Alessandro Colombi

Institution: University of Milan-Bicocca

Department: Department of Economics, Management and Statistics

Email: a.colombi10@campus.unimib.it

Talk title:

Predicting distinct and shared species in multiple areas

Abstract:

In ecology, the description of species composition and biodiversity is known as the “species sampling problem”. This is a well-studied area in statistics that involves estimating features of interest in unobserved samples based on an observed one. In the last decade, the Bayesian nonparametrics literature has thoroughly investigated the case where data arise from a homogeneous population, providing tools to address the challenges of estimating coverage probabilities and predicting the number of unseen species in arbitrarily large future samples. In this work, we propose a novel framework to address heterogeneous populations, specifically dealing with scenarios where data arise from two groups (or areas). This setting significantly increases the mathematical complexity of the problem and, as a consequence, it received limited attention in the literature. While early approaches leverage on computational methods, we introduce a new model relying on vectors of random probabilities with a finite, yet random, number of atoms shared across the two populations and a Dirichlet distribution for the group-specific weights. Under this framework, we provide a distributional theory for the in-sample analysis of any observed sample and we enable out-of-sample prediction for the number of unseen distinct and shared species in additional samples of arbitrary sizes. Furthermore, our results can be applied to address sample size determination in sampling problems aimed at detecting shared species.

Keywords: species sampling problem; partial exchangeability; ecology; vector of finite Dirichlet processes

Co-authors: Raffaele Argiento; Federico Camerlenghi; Lucia Paci

3.19. Fernando Corrêa

Institution: University of São Paulo

Email: fptcorrea@gmail.com

Talk title:

Kullback-Leibler Consistency of p -dimensional Pólya Tree Posteriors and Differential Entropy Estimation

Abstract:

We exploit the multiplicative structure of Pólya Tree priors for density and differential entropy estimation in p -dimensions. We establish: (i) a representation theorem of entropy functionals and (ii) conditions on the parameters of Pólya Trees to obtain Kullback-Leibler and Total Variation consistency for vectors with compact support. Those results motivate a novel differential entropy estimator that is consistent in probability for compact supported vectors under mild conditions. In order to enable applications of both results, we also provide a theoretical motivation for the truncation of Univariate Pólya Trees at level $3\log 2n$.

Keywords: Bayesian nonparametrics; Pólya Trees; Density estimation; Differential Entropy Estimation

Co-authors: Rafael Bassi Stern, Julio Michael Stern

3.20. Fuheng Cui

Institution: The University of Texas at Austin

Department: Department of Statistics and Data Sciences

Email: cuifuheng@utexas.edu

Website: <https://cuifh.github.io>

Talk title:

Martingale Posterior Distributions for Log-concave Density Functions.

Abstract:

The family of log-concave density functions contains various kinds of common probability distributions. Due to the shape restriction, it is possible to find the nonparametric estimate of the density, for example, the nonparametric maximum likelihood estimate (NPMLE). However, the associated uncertainty quantification of the NPMLE is less well developed. The current techniques for uncertainty quantification are Bayesian, using a Dirichlet process prior combined with the use of Markov chain Monte Carlo (MCMC) to sample from the posterior. In this paper, we start with the NPMLE and use a version of the martingale posterior distribution to establish uncertainty about the NPMLE. The algorithm can be implemented in parallel and hence is fast. We prove the convergence of the algorithm by constructing suitable submartingales. We also illustrate results with different models and settings and some real data, and compare our method with that within the literature.

Keywords: Bayesian nonparametrics; Asymptotic exchangeability; Uncertainty quantification; Nonparametric maximum likelihood estimation.

Co-authors: Stephen G. Walker

3.21. Elvis Han Cui

Institution: School of Medicine, UCI

Department: Department of Neurology

Email: cuieh@hs.uci.edu

Talk title:

A Semiparametric Bayesian Method for Instrumental Variable Analysis with Partly Interval-Censored Time-to-Event Outcome

Abstract:

This paper develops a semiparametric Bayesian instrumental variable analysis method for estimating the causal effect of an endogenous variable when dealing with unobserved confounders and measurement errors with partly interval-censored time-to-event data, where event times are observed exactly for some subjects but left-censored, right-censored, or interval-censored for others. Our method is based on a two-stage Dirichlet process mixture instrumental variable (DPMIV) model which simultaneously models the first-stage random error term for the exposure variable and the second-stage random error term for the time-to-event outcome using a bivariate Gaussian mixture of the Dirichlet process (DPM) model. The DPM model can be broadly understood as a mixture model with an unspecified number of Gaussian components, which relaxes the normal error assumptions and allows the number of mixture components to be determined by the data. We develop an MCMC algorithm for the DPMIV model tailored for partly interval-censored data and conduct extensive simulations to assess the performance of our DPMIV method in comparison with some competing methods. Our simulations revealed that our proposed method is robust under different error distributions and can have superior performance over its parametric counterpart under various scenarios. We further demonstrate the effectiveness of our approach on an UK Biobank data to investigate the causal effect of systolic blood pressure on time-to-development of cardiovascular disease from the onset of diabetes mellitus.

Keywords: Instrumental variable analysis, Partly interval-censored time-to-event data, Dirichlet process mixture, MCMC algorithm.

Co-authors: Xuyang Lu, Jin Zhou, Hua Zhou, and Gang Li

3.22. Marco Dalla Pria

Institution: University of Turin

Department: Department of Economics, Social Studies, Applied Mathematics and Statistics (ESOMAS)

Email: marco.dallapria@unito.it

Website: <https://marcodallapria.carrrd.co/>

Talk title:

Filtering 2-parameter Poisson-Dirichlet diffusions

Abstract:

Temporally evolving systems whose components are unlabeled and only associated to a normalized value arise in a variety of fields. In many such cases, the system components can be obtained by grouping single elements that share a characteristic in an underlying structure, and the dynamics of the system are due to single microscopic changes, so that they are well approximated by continuous trajectories. In such scenarios, data take the form of unlabeled partitions or Young diagrams, whose likelihood is a symmetric polynomial which integrates out the unknown categories. In this paper we show how to obtain exact distributions for all quantities of interest on an evolving unlabeled system with unknown number of categories, conditional on observing Young diagrams collected at discrete times, when the system's evolution is described by a 2-parameter Poisson-Dirichlet diffusion on the infinite-dimensional ordered simplex. After developing some necessary results related to coagulating unlabelled partitions and to drawing from conditional two-parameter Chinese restaurant processes, we derive all conditional distributions of the system which allow to perform online or offline inference, when respectively the data are made available over time or when they come in a batch. We also find the predictive distributions for extrapolating further Young diagrams at arbitrary times, which are shown to be certain finite mixtures of Ewens-Pitman sampling formulas. After discussing practical implementation issues and illustrating the inferential procedures with numerical experiments, we apply the introduced methodology to the dynamic estimation of live social interaction patterns given graph-valued data.

Keywords: 2-parameter Poisson-Dirichlet diffusions; Chinese restaurant process; stochastic duality; stochastic filtering; integer partitions

Co-authors: Matteo Ruggiero Dario Spanò

3.23. Luca Danese

Institution: University of Milan - Bicocca

Department: DEMS

Email: danese.luca1@gmail.com

Website: <https://lucadanese.github.io/>

Talk title:

Model-based clustering of time-dependent observations with common structural changes

Abstract:

In this work, we propose a model-based clustering method for time series, where observations within the same cluster share the same change points, and no other commonalities are assumed. To induce ties among observations we resort to a representation based on latent orders, where orders characterize each cluster since they represent the change points structure shared among the observations. We propose a sampling algorithm based on a split-and-merge procedure to obtain a point estimate of the final partition and show its properties through a simulation study. Our model is highly general, as it can be applied to any kind of time-dependent data, such as survival functions in the epidemiological context. We present an application to COVID-19 in the member states of the European Union, where countries that belong to the same cluster share the same structural changes in the spreading process of the virus.

Keywords: Change points; Model-based clustering; Time series; Bayesian statistics;

Co-authors: Riccardo Corradin; Wasiur R. KhudaBukhsh; Andrea Ongaro

3.24. Subhasish Das

Institution: University of Texas at San Antonio

Department: Department of Computer Science

Email: subhasish.das@utsa.edu

Website: <https://subhasish-das.mystrikingly.com>

Talk title:

Bradycardia Prediction in Preterm Infants Using Nonparametric Kernel Density Estimation

Abstract:

Bradycardia, characterized by an abnormally slow heart rate, poses a significant risk to preterm infants due to their increased metabolic demands. Existing bradycardia prediction systems often suffer from high false positive rates, hindering clinical decision-making. This research work proposes a novel statistical method for predicting the onset of bradycardia in preterm infants without requiring any prior knowledge. By leveraging nonparametric kernel density estimation, the proposed method models the distribution of the QRS complex R-wave on the electrocardiogram (ECG) signals. This approach provides a statistical guarantee of 95% accuracy and has been validated through simulations on ECG data from ten preterm infants. Furthermore, a theoretical analysis reveals the inherent limitations of this method, demonstrating that the lowest achievable false positive rate is 5%, which is the clinically and statistically acceptable threshold.

Keywords: Nonparametric Kernel Density Estimation; Bradycardia Prediction

Co-authors: Bahman Moraffah, Ayan Banerjee, Sandeep K.S. Gupta, Antonia Papandreou-Suppappola

3.25. Snigdha Das

Institution: Texas A&M University

Department: Department of Statistics

Email: snigdha@stat.tamu.edu

Website: <https://www.linkedin.com/in/snigdha-das1812/>

Talk title:

Scalably Efficient Inference in Complex Surveys through Targeted Resampling of Weights

Abstract:

Survey data routinely arises from complex sampling designs, like stratified or multistage sampling, where individuals in the finite population have unequal probabilities of inclusion into the sample. Sampling designs can often be informative, where inclusion probabilities are correlated with the variable of interest. In such cases, the sample lacks representativeness and traditional inference methods manifest in biased estimators with poor coverage. While various adjustments exist to incorporate survey weights, few are suitable for model-based inference. We investigate two popular approaches for Bayesian inference under informative sampling, namely, the Bayesian pseudo posterior estimator (BPPE) which leverages the powered likelihood, and the estimator based on weighted Bayesian bootstrap (WBB). While BPPE achieves consistency, it falters at accurate uncertainty quantification. WBB provides good empirical coverage but is sensitive to the number of bootstrap samples and is computationally demanding, posing challenges for large datasets. To address these limitations, we propose a Survey-adjusted Weighted Likelihood Bootstrap (S-WLB) algorithm that resamples random weights from a carefully chosen distribution centered around the underlying sampling weights. The S-WLB simultaneously mitigates the sampling bias and performs valid inference without requiring complex analytic approximations or extensive post-processing. It is computationally efficient and offers theoretical guarantees for consistency and accurate asymptotic coverage. We demonstrate our method through simulations and applications, including the construction of risk indices using complex, nationally representative survey datasets such as NHANES and NSDUH.

Keywords: Survey designs; informative sampling; sampling weights; weighted likelihood bootstrap; uncertainty quantification.

Co-authors: Dipankar Bandyopadhyay (Virginia Commonwealth University) and Debdeep Pati (University of Wisconsin-Madison)

3.26. Dat Do

Institution: University of Michigan

Department: Department of Statistics

Email: dodat@umich.edu

Website: <https://dodat619.github.io>

Talk title:

Robust and Interpretable Bayesian Mixture Modeling by using Dendrogram of mixing measures

Abstract:

Bayesian mixture models are powerful devices for fitting heterogeneous data. The output of these models, named mixing measure, captures the parameters of subpopulations in data, and its posterior can be efficiently obtained from approximation algorithms such as MCMC or Variational Inference. However, when the data is too complex or the model is misspecified, the sampled mixing measures tend to have many similar atoms or small weights, leading to difficulties in interpretation. In this work, we propose to summarize those posterior samples using hierarchical trees (dendrogram) of mixing measures for better visualization and interpretation of Bayesian mixture models. We illustrate the usefulness of the dendrogram in (1) misspecified models, (2) mixture models with an increasing number of components, and (3) real data. We also show that when the model is well-specified, the parameters on the dendrogram initially constructed from an over-fitted mixture model can converge to the truth at the fast root-n rate. Some connections to Split-Merge sampling will also be discussed.

Keywords: Over-fitted Mixtures, Hierarchical Clustering, Convergence rate, Dirichlet Process Mixtures, Mixture of Finite Mixtures

Co-authors: Linh Do (Tulane University); Scott McKinley (Tulane University); Jonathan Terhorst (University of Michigan); XuanLong Nguyen (University of Michigan)

3.27. Jianghu (James) Dong

Institution: University of Nebraska Medical Center

Department: Biostatistics

Email: jianghu.dong@unmc.edu

Website: <https://www.unmc.edu/publichealth/departments/biostatistics/facultyandstaff/jianghu-dong.html>

Talk title:

Dynamic Joint Model.

Abstract:

Accurate prediction of overall survival (OS) in patients, such as those with end-stage renal disease (ESRD) or cancer, requires advanced biostatistical methods that utilize serially measured biomarkers, such as blood or creatinine levels, both pre- and post-treatment. Longitudinal biomarkers often exhibit complex trends over time, with both the rate of change and absolute expression levels providing critical insights into disease progression and survival. However, existing methods rarely incorporate joint models capable of effectively capturing these temporal dynamics. Dynamic prediction approaches, particularly Bayesian joint modeling of longitudinal and survival data, have gained significant traction for their potential in personalized medicine and tailored clinical decision-making. We propose a Bayesian model that integrates longitudinal biomarkers to improve OS prediction compared to traditional approaches, such as the extended Cox and frailty models. This methodology has the potential to enhance individualized treatment strategies and advance precision oncology.

Keywords: Longitudinal, Survival, Joint model, Bayesian

Co-authors: None

3.28. Yunshan Duan

Institution: University of Texas at Austin

Department: Department of Statistics and Data Sciences

Email: yunshan@utexas.edu

Talk title:

Spatially aligned random partition models on spatially resolved transcriptomics data

Abstract:

We propose spatially aligned random partition (SARP) models for two distinct types of experimental units, incorporating dependence in a sub-vector of the cluster specific parameters related to spatial information. This model is motivated by investigating the co-location of the immune and non-immune cell sub-populations, in order to understand the recruitment of immune cell subpopulations by tumor cells as spatial dependence. This is achieved by constructing Bayesian nonparametrics models on immune and non-immune cells with a corresponding structure in the priors. The structure of the model naturally leads to a convenient MCMC algorithm for posterior inference. The model shows superior performance in identifying co-location between immune and non-immune cells, as is shown in the simulation studies. We apply this method on young-onset colorectal cancer (YOCRC) data. Our long-term goal is to fill the gaps in understanding of the distinct nature of TME in YOCRC, to gain biologic insight that may help YOCRC treatment and enable reducing the burden of this disease.

Keywords: Random partitions; Spatial dependence; Spatial transcriptomics data

Co-authors: Wenyi Wang, Peter Mueller

3.29. Mael Duverger

Institution: Paris Dauphine University

Department: Ceremade (mathematics department)

Email: duverger@ceremade.dauphine.fr

Talk title:

Bernstein-von Mises property for functionals of multivariate linear Hawkes process

Abstract:

Multivariate Hawkes processes are point processes which have been extensively applied to model event times dependent on both past events and interaction phenomena, such as neuronal spike trains, online messages, financial transactions and seismic activity. In the linear multivariate case, where the number of components K is fixed, the Hawkes process is defined by a set of parameters composed of K positive real numbers (the background rates) and K^2 non-negative functions (called excitation functions). There have been some recent results on Bayesian estimation of these parameters, in particular Donnet et al. (2020) have obtained posterior concentration rates. In this work, based on these recent results, we

propose general prior conditions to get Bernstein-von Mises property for smooth functionals of the parameters. Broadly speaking, Bernstein-von Mises property states that the posterior distribution behaves asymptotically as a Gaussian distribution, centered at an efficient estimator. More precisely, we first derive asymptotic (LAN) expansions for such processes, and use these to develop the theory of statistical efficiency and Bernstein von Mises properties of the posterior distribution for smooth functionals. We illustrate our result by applying it to common prior distributions.

Keywords: Hawkes processes, semi-parametric inference, Bernstein-von Mises property, LAN expansion

Co-authors: Judith Rousseau (Paris Dauphine University)

3.30. Shitao Fan

Institution: University of Maryland

Department: Mathematics

Email: sfan211@umd.edu

Talk title:

Robust and Scalable Variational Bayes

Abstract:

We present a robust and scalable approach to Variational Bayes (VB) for handling outliers in large datasets. Our method partitions the data into disjoint subsets, applies VB posterior sampling to each, and aggregates the resulting posteriors using the median of probability measures, computed via the Wasserstein distance. This results in the variational median posterior (VM-Posterior) distribution, a novel method within VB. We prove that the VM-Posterior retains contraction properties similar to those of the true posterior, while accounting for errors introduced by the variational approximation. Additionally, we establish a variational Bernstein-von Mises theorem for both the multivariate Gaussian distribution with a general covariance structure and the mean-field variational family. We adapt existing algorithms for the computation of the VM-Posterior and validate its performance through numerical experiments, demonstrating its robustness and scalability for Bayesian inference in complex datasets.

Keywords: Variational Bayes; Robust and Scalable; outlier; Median

Co-authors: Carlos Misael Madrid Padilla, Lizhen Lin

3.31. Laura Ferrini

Institution: Università degli Studi di Milano-Bicocca

Email: laura.ferrini@unicatt.it

Talk title:

BNP Mixtures of graphs for clustering categorical data

Abstract:

Clustering categorical data is a pervasive issue in many applied domains, including social and biomedical science. In a multivariate setting, model-based clustering should account for differences in the joint distribution of variables across clusters, thus allowing for both shifts in the marginal distributions and changes in their dependence structure. This is crucial in cases where clusters are not well-separated, and differences are mostly captured by parameters indexing the joint distribution rather than the marginals. In such contexts, incorporating dependence relations between variables can significantly enhance the clustering performance. To address this issue, we propose a novel Bayesian methodology based on graphical models which explicitly accounts for dependencies in the joint distribution of the variables. Specifically, we consider an infinite mixture model based on a Dirichlet Process prior, where each mixture component corresponds to a decomposable graphical model. We provide full Bayesian inference for the model and develop a Markov chain Monte Carlo scheme for posterior inference on the clustering and graphs. Our methodology is evaluated through simulation studies and applied to real datasets. Results show that the proposed method outperforms benchmark approaches for clustering multivariate categorical data that do not explicitly account for dependencies in the joint distribution of variables.

Keywords: Dirichlet Process; Categorical undirected graphs;

Co-authors: Federico Castelletti

3.32. Bernardo Flores

Institution: University of Texas at Austin

Department: Department of Statistics & Data Sciences

Email: bernardofl@utexas.edu

Website: <https://www.bernardofl.github.io>

Talk title:

Predictive Coresets

Abstract:

Modern data analysis often involves massive datasets with hundreds of thousands of observations, making traditional inference algorithms computationally prohibitive. Coresets are methods designed to select a smaller subset of observations while maintaining similar learning performance. Traditional approaches involve finding sparse weights that minimize the Kullback-Leibler (KL) divergence between the likelihoods of the original and weighted datasets; however, they are ill-posed for nonparametric models, where the likelihood is often intractable. We propose an alternative variational method which leverages randomized posteriors and finds weights to match the unknown posterior predictive distributions conditioned on the full and reduced datasets. Our approach provides a general algorithm based on predictive recursions suitable for nonparametric priors. We evaluate our method's performance on diverse problems, including random partitions and density estimation.

Keywords: Dimensionality reduction; coresets; random partitions; species sampling model

Co-authors: None

3.33. Michela Frigeri

Institution: Politecnico di Milano

Department: Department of Mathematics

Email: michela.frigeri@polimi.it

Talk title:

A Bayesian hierarchical model for air pollution source apportionment

Abstract:

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.

Keywords: Bayesian source apportionment; air pollution; functional warping; latent factors model; spatio-temporal data

Co-authors: Veronica Berrocal, Alessandra Guglielmi

3.34. Jairo Alberto Fuquene Patino

Institution: UC Davis

Department: Statistics

Email: jafuquenepatino@ucdavis.edu

Talk title:

A Semiparametric Bayesian Extreme Value Model Using a Dirichlet Process Mixture of Gamma Densities

Abstract:

In this work we present a model with a Dirichlet process mixture of gamma densities in the bulk part below threshold and a generalized Pareto density in the tail for extreme value estimation. The proposed model is simple and flexible allowing us posterior density estimation and posterior inference for high quantiles. The model works well even for small sample sizes and in the absence of prior information. We evaluate the performance of the proposed model through a simulation study. Finally, the proposed model is applied to a real environmental data.

Keywords: Generalized Pareto Distribution, Threshold Estimation, Dirichlet Process Mixture.

Co-authors: None

3.35. Isaac Fwemba

Institution: University of Zambia

Department: Biostatistics

Email: isaac.fwemba@unza.zm

Talk title:

Comparing Bayesian nonparametric latent-class modeling (BLCM) and Bayesian Anchor multiplier as tools for HIV epidemic control through estimation of Key population size in Zambia.

Abstract:

Background: Zambia has one of the highest HIV incidences in Sub Saharan Africa which has been driven largely by key population groups. Key populations, including female sex workers, men who have sex with men, and people who inject drugs, are more vulnerable to HIV than the general population due to stigmatized and criminalized behaviors.

Methodologies needed to produce reliable estimates on the size of the key population at sub national level are needed to strengthen HIV epidemic control effort. Objective: The objective of our study was to use respondent driven sampling method to reach out to members of the key population in places that are hard to reach. Use of robust statistical method to demonstrate improved statistical estimation efforts that will result in obtaining robust and reliable estimates. Methods: The study was based on the cross sectional three wave concurrent data collection designs covering five districts from five provinces of Zambia using 3-source capture-recapture methods. Hotspots were mapped before 3-source capture-recapture started, and was guided by the key informants working in the area of HIV prevention and Key population in Zambia. We sampled female sex workers, men who

have sex with men, and people who inject drugs during 3 independent captures about one week apart. During hotspot mapping, key population members were offered inexpensive, memorable objects unique to each capture round. In second and third rounds, key population members were offered an object and asked to identify objects received during previous rounds (if any). Correct responses were tallied and recorded on tablets. Data were aggregated by key population and province for analysis. Three statistical approaches were used to analyse the collected data. These were the Bayesian anchor multiplier, Wisdom of the crowd and Bayesian nonparametric latent-class models. The Bayesian approach to latent class analysis allowed for the integration of prior information about the latent class membership probabilities and the conditional response probabilities. The prior probabilities were constructed using historical data from previous literature and expert opinion. Posterior Median population size estimates were derived using both Bayesian nonparametric latent-class models and Anchor multiplier with 95% confidence intervals provided. As expected the

Keywords: Bayesian Latent Class Modelling , Non parametric, Key population

Co-authors: Patrick Kaonga, Peter Fusani, Andrew Banda, Melvin Mwansa

3.36. Soham Ghosh

Institution: Texas A & M University

Department: Statistics

Email: sohamghosh@tamu.edu

Talk title:

Optimal Posterior Contraction for Principal Subspace in High Dimension

Abstract:

Principal Component Analysis (PCA) is a heavily used statistical tool for low-rank estimation and dimension reduction. Traditionally, some low-rank structure, like sparsity, is assumed for estimating PCs in high dimensions. However, very recently, consistent estimation of PCs has been established under the low effective rank assumption ($r(\Sigma) = \text{tr}(\Sigma)/\|\Sigma\|$, where tr denotes trace, $\|\cdot\|$ denotes operator norm). In this work, we propose a Bayesian method for estimating the principal subspace that achieves the frequentist optimal rate. Instead of assuming sparsity, we assume a relatively large signal-to-noise ratio, ensuring that the effective rank remains small. We use a very simple prior that facilitates the formulation of an efficient computational strategy.

Keywords: Bayesian estimation; principal component analysis; posterior contraction; rate of convergence.

Co-authors: Anirban Bhattacharya

3.37. Soham Ghosh

Institution: University of Wisconsin, Madison

Department: Department of Statistics

Email: sghosh39@wisc.edu

Website: <https://scholar.google.com/citations?user=VwTydBwAAAAI&hl=en>

Talk title:

High-Dimensional Mixed-Outcome Regression with Sparse Structure Estimation via Multivariate Spike-and-Slab LASSO

Abstract:

We consider a high-dimensional multi-outcome regression setting in which q , possibly dependent outcomes of mixed-types, are regressed on p covariates. We model these outcomes, which contain a mix of continuous and discrete variables, as potentially truncated and transformed versions of a latent multi-outcome linear model. Our goal is to estimate a sparse matrix (B) of latent regression coefficients and a sparse latent residual precision matrix (Ω). We specify continuous spike and slab priors to all entries of B and to the off-diagonal elements of Ω . Under a set of mild yet realistic assumptions, we establish the posterior contraction rate for our model in the asymptotic regime where both p and q diverge with the sample size n . Our findings are corroborated through extensive simulations and a real data analysis involving Chronic Kidney Disease (CKD), showcasing the effectiveness of our method in accurately recovering the true support and structure of B and Ω across various challenging mixed-type scenarios.

Keywords: Mixed-type responses, Spike-and-slab priors, Posterior contraction, Precision matrix estimation, Latent variable modelling

Co-authors: Sameer K. Deshpande

3.38. Francesco Gili

Institution: TU Delft

Department: Mathematics - Statistics

Email: f.gili@tudelft.nl

Website: <https://www.francescogili.nl/>

Talk title:

Semiparametric Bernstein-von Mises Phenomenon via Projected Posterior in Wicksell's problem.

Abstract:

Consider these two problems. The first one: astronomers are interested in the study of the way stars are distributed in the universe. In this setting, globular clusters - spherical aggregations of stars held together by gravity - are a topic of particular focus. But how to determine the distribution of stars within these three-dimensional structures when given only a 2D photo of the clusters? The second one: having material that presents a globular microstructure and that cannot be scanned (e.g. steel) but only sectioned, how can we determine the distribution of particles' size inside the material when given only a limited number of 2D cross sections? Despite their seemingly distant nature, both of these problems share a common thread: the inherent structure of an inverse problem, originally

conceptualized by Wicksell. In the astronomical setting, the focus is on estimating the distribution of the squared distance of a star to the center of the globular cluster or galaxy $X = Z^2_{p_1} + Z^2_{p_2} + Z^2_{p_3} \sim F$ using $Z = Z^2_{p_1} + Z^2_{p_2} \sim G$. In this paper, we explore the properties of nonparametric Bayesian estimators based on the Dirichlet process prior in this problem, which is the first time the Dirichlet process is studied in an inverse setting. We deviate from the classical Bayesian nonparametric approach, which would place a Dirichlet Process prior on the distribution function of the unobservables F , by directly placing a DP prior on the distribution function of the observables G . This method offers computational simplicity due to the conjugacy of the posterior and allows for asymptotically efficient estimation by projecting the posterior onto the $(\mathbb{L}_2)$ subspace of increasing, right-continuous functions. In particular, we demonstrate how our methodology enables uncertainty quantification, achieving Bernstein–von Mises-type results and we compare it to the state-of-the-art nonparametric frequentist shape-constrained estimator, recently derived in. These results have theoretical relevance because they provide an instance of a semiparametric setting, embedded in an inverse problem, where the nonparametric Bayesian procedure can be asymptotically compared to the state-of-the-art nonparametric shape-constrained estimator.

Keywords: Wicksell's problem; Isotonic estimation; Efficiency theory; Bayesian Nonparametrics; Dirichlet Process Prior

Co-authors: Aad van der Vaart, Geurt Jongbloed

3.39. Nicola Gorini

Institution: Bocconi University

Department: Statistics and Computer Science

Email: nicola.gorini@phd.unibocconi.it

Website: <https://nicolagorini.github.io/>

Talk title:

Linear Methods for Non-linear Inverse Problems

Abstract:

I will introduce different Bayesian methodologies to recover an unknown function from a noisy observation of the solution of a certain class of partial differential equations. Explicit applications include the time-independent Schroedinger equation, the Heat equation with an absorption term, the Darcy problem and more.

Keywords: Bayesian non-parametric, Gaussian processes, PDE, Inverse problems

Co-authors: Botond Szabo, Aad van der Vaart

3.40. Hwangwan Gwon

Institution: University of California-Irvine

Department: Statistics

Email: hwanggw@uci.edu

Talk title:

Improving In-Situ Soil Moisture Monitoring with the CUSP-MRA Prior: A Local Stationarity Approach

Abstract:

Soil moisture refers to the amount of water stored in soil pores. It is a critical factor in terrestrial hydrology, influencing land surface processes such as crop fertility, wildfire dynamics, and hydrologic extremes. Although advanced technologies like remote sensing provide information on soil moisture levels, calibration issues along with the sparse temporal frequency of satellite data, highlight the essential role of in-situ measurements from ground probes to obtain an accurate quantification of soil dryness over space and time. However, mapping soil moisture over the United States is not a simple and straightforward task given the limited number of locations with in-situ probes and the well-known non-stationarity of soil moisture. To assist in identifying the locations where additional soil moisture data should be collected, we introduce the CUSP-MRA prior, a new prior capable of accommodating both stationary and nonstationary spatial processes. We build this prior by combining the idea of the Cumulative Shrinkage Prior (CUSP) — a Bayesian nonparametric prior useful for determining the shrinkage of basis-function coefficients— with that of the Multi-Resolution Approximation (MRA). The CUSP-MRA prior represents a spatial process through an infinite basis-function expansion, where the basis functions are constructed via the MRA framework and their coefficients are increasingly shrunk toward zero according to the CUSP, with the extent of shrinkage learned from the data. Additionally, since the shrinkage of the basis-function coefficient is largely driven by the underlying spatial variability and range of the process, inferring upon these basis function coefficients enables us to identify “regions of local stationarity”, that is, areas where the spatial process exhibits homogeneous levels of spatial variability and strength of spatial dependence.

Keywords: Spatial process; local stationarity; multiresolution; shrinkage process; stick breaking process

Co-authors: Veronica Berrocal, Department of Statistics, University of California-Irvine
Francesco Denti, Department of Statistics, University of Padua

3.41. Torey Hilbert

Institution: Ohio State University

Department: Statistics

Email: hilbert.54@osu.edu

Talk title:

Symmetrization of Martingale Posterior Distributions

Abstract:

The martingale posterior framework, recently proposed by Fong et al., is based on a sequence of one step ahead predictive distributions. It leads to computationally efficient inference in parametric and nonparametric settings. The predictive distributions implicitly provide a joint model for an infinite sequence of data. The observed data, typically arbitrarily considered to be Y_1 through Y_n in this infinite sequence, form the beginning of the sequence and the tail of the sequence is missing. Filling in the remainder of the sequence allows one to summarize Y_1 through Y_{∞} (or through Y_N , N large in practice). A typical summary, such as the mean, is regarded as a parameter. Generation of a large number of sequences gives access to the posterior distribution of the parameter. In cases where the martingale posterior model does not match a de Finetti model, the joint distribution over the Y 's is not exchangeable, and so the indices $\{1, \dots, n\}$ of the data affect the analysis. We investigate methods of symmetrizing inference in these models. In some settings, re-indexing the observed data to $\{a_1 < \dots < a_n\}$ and sending a_1 to infinity is analytically tractable, and we recover classical Bayesian models with known priors. Additionally, we investigate the effect of nesting the nonexchangeable models within exchangeable models.

Keywords: martingale posterior; de Finetti model; bayesian nonparametrics; nonexchangeable models

Co-authors: Steven MacEachern

3.42. Yujin Jeong

Institution: Yonsei university

Department: Statistics and Data science

Email: yujin3257@yonsei.ac.kr

Website: <https://github.com/YujinJeong0202>

Talk title:

Bayesian Nonparametric Modeling for Multivariate Conditional Copula Regression with Varying Coefficients

Abstract:

Modeling multi-response data with both discrete and continuous measurements is a challenging task due to the lack of natural multivariate distributions for mixed-type observations. Copula regression models provide a flexible yet accessible framework to address this challenge, specifying the joint distribution of responses through a copula function and their marginals. While conventional copula models typically rely on parametric linear predictors and are limited to fixed dependence structures, real-world data often demonstrate dynamic relationships influenced by additional covariates. Motivated by datasets such as the BRFSS (Behavioral Risk Factor Surveillance System) where the associations among variables evolve with age or other covariates, this study develops a Bayesian nonparametric framework for conditional copula regression with multivariate responses. Our model captures the varying effects of explanatory variables in the marginal distributions using varying-coefficient models, with regression coefficients modeled as smooth functions of an additional covariate. To enhance local adaptability, we employ Bayesian knot selection with suitably designed natural cubic spline basis functions. For the copula component, we construct a mixture of Gaussian copulas with

covariate-dependent mixture weights, using the probit stick-breaking process for greater flexibility compared to single parametric copulas. This framework eliminates the need for pair copulas or vine copulas and accommodates high-dimensional responses without imposing constraints on copula parameters. The resulting joint model integrates dynamic marginals and copula structures, which allows both the marginal distributions and dependence structures to flexibly adapt as functions of an additional covariate, such as time. This approach provides clear and meaningful insights into complex multivariate settings.

Keywords: Copula regression; Bayesian nonparametrics; Varying-coefficient models; Time-varying copula models; Probit stick-breaking process;

Co-authors: Seonghyun Jeong

3.43. Niladri Kal

Institution: Texas A&M University

Department: Statistics

Email: niladrik@tamu.edu

Talk title:

Adaptive Divide and Conquer with Two Rounds of Communication

Abstract:

Existing adaptive, distributed computing methods restrict either the number of machines or the smoothness range over which adaptation is possible, limiting their potential applicability. We introduce a novel two-round communication strategy that enables adaptive, rate-optimal estimation without such stringent restrictions. In the first round, local machines send summary statistics to the central node to estimate the hyperparameters that depend on the underlying smoothness. In the second round, another set of statistics are transmitted, enabling the central machine to aggregate and produce a final estimator that adapts to the true smoothness level. This approach achieves optimal convergence rates across a wider range of regularities, offering a potential improvement in the adaptability and efficiency of distributed estimation.

Keywords: Divide-Conquer; Bayesian; Optimal rate of convergence; Distributed learning

Co-authors: Botond Szabo, Rajarshi Guhaniyogi, Natesh Pillai, Debdeep Pati

3.44. Jennifer Kampe

Institution: Duke University

Department: Statistical Science

Email: jennifer.kampe@duke.edu

Talk title:

Nested exemplar latent space models for dimension reduction in dynamic networks

Abstract:

Dynamic latent space models are widely used for characterizing changes in networks and relational data over time. These models assign to each node latent attributes that characterize connectivity with other nodes, with these latent attributes dynamically changing over time. Node 25 attributes can be organized as a three-way tensor with modes corresponding to nodes, latent space dimension, and time. Unfortunately, as the number of nodes and time points increases, the number of elements of this tensor becomes enormous, leading to computational and statistical challenges, particularly when data are sparse. We propose a new approach for massively reducing dimensionality by expressing the latent node attribute tensor as low rank. This leads to an interesting new nested exemplar latent space model, which characterizes the node attribute tensor as dependent on low-dimensional exemplar traits for each node, weights for each latent space dimension, and exemplar curves characterizing time variation. We study properties of this framework, including expressivity, and develop efficient Bayesian inference algorithms. The approach leads to substantial advantages in simulations and applications to ecological networks.

Keywords: dynamic network modeling, Bayesian nonparametrics, ecology application, latent factor, tensor factorization

Co-authors: Luca Silva, Tomas Roslin, David Dunson

3.45. George Kanchaveli

Institution: Bocconi University

Email: george.kanchaveli@phd.unibocconi.it

Talk title:

A two-sample testing for exchangeable sequences via optimal transport

Abstract:

Two-sample testing is crucial for determining whether two populations differ significantly, a fundamental problem in data-driven decisions. Standard methods assume independent and identically distributed observations, which can be restrictive. In this work, we develop a two-sample test for distinguishing laws of exchangeable sequences. The key ingredient is de Finetti's theorem, which establishes a one-to-one correspondence between such laws and laws of random probability measures; therefore, we focus on the latter and develop a test to determine whether two such laws coincide. Our goal is to obtain the testing scheme with a low False Positive Rate (probability of rejecting the null hypothesis when it is True) and a high True Positive Rate (probability of rejecting the null hypothesis when it is false), and also provide the bounds for the number of samples needed to have the testing scheme with the desired level of TPR. Our approach extends the distance-based framework used in i.i.d. two-sample testing, where the null hypothesis of homogeneity is rejected if the distance between the sample laws exceeds a threshold. To adapt this framework, we first establish a metric structure on the space of probability measures over probability measures. Using optimal transport tools, we introduce two natural distance functions: the Wasserstein-over-Wasserstein and the Hierarchical Integral Probability Metric. A key challenge, compared to the i.i.d. setting, is that we do not observe samples of probability measures, but rather samples of random variables sampled from them, leading to a two-level sampling scheme. We use two methods for each distance function to obtain the threshold. The first one leverages Talagrand's concentration inequality. The second one is

referred to as the permutation approach whose properties are well studied in the i.i.d case using the tools for empirical process theory. We demonstrated that the testing scheme based on the concentration inequality approach is consistent and FP can be controlled. However, in practice, we need many samples to get the high TPR for both of the distances. We are working on adapting and providing theoretical justifications for the permutation approach for exchangeable case.

Keywords: Optimal transport, Talagrand's inequality, Donsker

Co-authors: Marta Catalano, Hugo Lavenant

3.46. Valentin Kilian

Institution: University of Oxford

Department: Department of Statistics

Email: valentin.kilian@mansfield.ox.ac.uk

Website: <https://valentinkil.github.io>

Talk title:

Modelling extremely sparse networks

Abstract:

In recent years, Completely Random Measures (CRMs) have emerged as a fundamental building block of modern Bayesian methods, with applications ranging from clustering to network analysis. A key feature to study a CRM is its Laplace exponent, whose behaviour at infinity tunes important asymptotic properties of the CRM and the hierarchical models built upon it. Remarkably, there is currently no tractable model for rapidly varying CRMs characterised by an almost linear Laplace exponent at infinity. This case is however important. For instance, many real-world networks exhibit extreme sparsity, where the number of edges scales linearly (up to a slowly varying function) with the number of nodes. In the Caron-Fox sparse network models based on CRMs, rapidly varying CRMs correspond to these extremely sparse networks. Developing a tractable model for rapidly varying CRMs would make it possible to generate such extremely sparse networks while capturing other essential real-world properties, such as heavy-tailed degree distributions, while maintaining computational tractability for Bayesian inference. Existing models struggle to provide these critical properties simultaneously. In response to this challenge, we present a novel class of CRM models that include rapidly varying CRMs. The CRM is constructed as a mixture of generalised gamma CRMs, has a simple expression, interpretable parameters, and includes the generalised gamma and stable processes as special limiting cases. We also present a simple size-biased construction for the CRM. To facilitate posterior inference on extremely sparse graphs, we adapt the Caron-Fox Hamiltonian Monte Carlo algorithm, which we use to analyse network properties of real data sets.

Keywords: completely random measure; graphex model; random graphs; sparsity; Caron-Fox model

Co-authors: François Caron (University of Oxford); Benjamin Guedj (INRIA and University College of London)

3.47. Hyotae Kim

Institution: Duke University

Department: Department of Biostatistics & Bioinformatics

Email: hyotae.kim@duke.edu

Talk title:

Modeling Cross-Site Dependency in Microbiome Compositional Data: A Latent Factor Approach for Improved Analysis and Prediction

Abstract:

Microbiome compositional data are often collected from several body sites and exhibit dependency among them. Analyzing microbial compositions from different sites jointly allows for effective borrowing of information by exploiting the underlying cross-site correlation, which can lead to more effective statistical analysis, especially when the sample size at one or both sites is limited. To this end, we introduce a joint model for microbiome compositions at two (or more) sites within the same subjects. Our model incorporates (i) latent factors shared across two body sites to explain the common subject effects and to serve as the source of correlation between the two sites; and (ii) mixtures of latent factors to allow heterogeneity among the subjects in their level of cross-site association. Simulation studies show how common subject effects influence regression analysis results; a stronger association between two sites in the data causes a greater degree of bias in the analysis. The model with latent factors mitigates the bias present in the model without latent factors, whereas the two models perform comparably for the data set without paired associations. In a case study involving samples of the urinary and vaginal microbiome collected from women, our model leads to the detection of covariate associations of the vaginal and urinary microbiome composition that are otherwise not statistically significant under a similar regression model applied to the two sites separately. Our model also enables prediction of the microbial abundance at one site based on observations from another, improving predictive power. Finally, we note that the latent variable modeling approach is useful as a foundation for developing more sophisticated models, such as multinomial models for all taxa of microbial communities with two types of presumptive associations: among taxa and between communities (body sites). Current research is exploring logistic-tree normal models for microbiome data collected from multiple surfaces of a built-environment. Our aim is to capture both covariances among taxa and covariances between surfaces by incorporating latent factors with Gaussian Process regression approaches. In this talk, I also introduce the ongoing development of this model, along with some examples.

Keywords: Bayesian modeling; latent factor models; paired microbiome data; P'olya-Gamma augmentation; compositional data

Co-authors: Nazema Siddiqui, Lisa Karstens, and Li Ma

3.48. Seongmin Kim

Institution: Seoul National University

Department: Department of Statistics
Email: zlatjdals@snu.ac.kr
Website: <https://smkim55.wordpress.com>

Talk title:

Bayesian analysis of the eigenstructure of the high-dimensional covariance with shrinkage inverse Wishart prior

Abstract:

The eigenstructure of covariances is essential in understanding the covariance of multivariate observations. The inverse Wishart prior is commonly used for the covariance estimation in Bayesian inference. However, in the high-dimensional setting where the number of covariates increases with the sample size, it is well-known that the estimates of eigenvalues tend to spread out. Surprisingly, there has been little prior study in Bayesian statistics on eigenstructures of unconstrained covariance. Recently, the shrinkage inverse-Wishart (SIW) prior had been proposed to mitigate the limitation of the inverse Wishart. In this study, we investigate the asymptotic properties of the posterior of both the eigenvalues and eigenvectors. We discuss the pros and cons of the SIW in the high-dimensional setting. Specifically, we examine the behavior of the eigenvalues estimated using the shrinkage inverse Wishart prior in comparison to the inverse Wishart prior.

Keywords: Covariance; eigenstructure; high-dimensional; shrinkage inverse-Wishart prior; Asymptotic property

Co-authors: Jaeyong Lee (Department of Statistics, Seoul National University) Kwangmin Lee (Department of Big Data Convergence, Chonnam National University) Sewon Park (Security Algorithm Lab, Samsung SDS)

3.49. Emma Landry

Institution: University of California, Los Angeles
Department: Biostatistics
Email: emmalandry@g.ucla.edu
Website: <https://emma-landry.github.io/>

Talk title:

Bayesian regression in the Wasserstein space for density-valued data

Abstract:

We propose a non-parametric regression model to explain the effect of vector-valued covariates on density-valued data. Motivated by the Fréchet regression methodology, we assume our data to lie in a complete separable metric space equipped with the 2-Wasserstein distance. It has been demonstrated in the optimal transport literature that the W_2 distance is a canonical metric for the estimation of deformations, our approach hence bypasses the need for recovery of warping functions to account for phase variation. Focusing on the case of one-dimensional distributions, we transform the density functions to quantile functions, allowing for the representation of the Wasserstein Fréchet mean under the L_2 metric. The quantile functions are modeled using B-splines, where the B-spline

coefficients are represented hierarchically with covariate dependence. An appropriate linear transformation is applied to ensure monotonicity of the estimated quantile functions. Estimation is tackled from a loss minimization perspective -- under the mean squared error loss, the problem is equivalent to analysis with the Gaussian likelihood, but extensions to more robust alternatives may be considered using empirical likelihood methods. Inference for the model parameters is then conducted using Hamiltonian Monte Carlo. We illustrate the proposed method on a case study of autism spectrum disorder (ASD) with data consisting of spectral densities obtained from EEG measurements from a cohort of children. We also provide a simulation study and theoretical guarantees.

Keywords: functional data analysis; distributional regression; optimal transport; loss minimization; monotonic regression

Co-authors: Donatello Telesca

3.50. Riccardo Lazzarini

Institution: Bocconi University

Department: Department of Decision Sciences

Email: riccardo.lazzarini@phd.unibocconi.it

Website: <https://riccardo-lazzarini.github.io/>

Talk title:

Statistical Guarantees for Robust Gaussian Process Regression

Abstract:

We study the frequentist properties of a robust Gaussian Process Regression model. We focus on the approach introduced by Altamirano et al. (2024). The authors derived a generalized Bayes method which is robust for outliers while still obtaining conjugate computations. We analyze the asymptotic behaviour of the corresponding posterior distribution as the number of data points increases, assuming that the observed data are subject to contamination. In contrast to Altamirano et al. (2024), we let the number of contaminated data points increase with the sample size and analyze the theoretical properties of this approach.

Keywords: Gaussian Process regression; Bayesian Nonparametric; Data Contamination;

Co-authors: Botond Szabo, Ismael Castillo, Paul Egels

3.51. Andrew Le

Institution: University of California, Santa Cruz

Department: Statistics

Email: ale988@ucsc.edu

Talk title:

Analysis of Intensities for Replicated Point Patterns

Abstract:

We develop a Bayesian nonparametric model for replicated point patterns under a Poisson process assumption, where each replicate arises from a shared underlying baseline process density. Our approach builds on a foundation of parsimonious mixture representations through the Bernstein-Dirichlet process prior model, enabling efficient estimation of complex functional shapes while supporting full Bayesian inference without the need for likelihood approximations. A key feature of this prior is that it provides a unified framework linking Poisson process density and intensity formulations. By hierarchically extending the Dirichlet process prior model, we capture shared structural characteristics across replicates, with the concentration parameter controlling how closely each replicate follows the baseline process. The model is fit tractably using standard Pólya-urn based Markov chain Monte Carlo techniques performed in parallel across replicates, pooling the distinct components to estimate the baseline process. We demonstrate the inferential capabilities of our method on simulated data and also apply it to data from the Divvy bike-share system in Chicago.

Keywords: Non-homogeneous Poisson process; Bernstein-Dirichlet process prior; Bayesian nonparametrics; analysis of variance

Co-authors: Athanasios Kottas

3.52. Tâm Le Minh

Institution: Inria, Univ. Grenoble Alpes

Email: tam.le-minh@inria.fr

Talk title:

Nonparametric models for bipartite edge-exchangeable networks with node dependence: applications to ecological interactions

Abstract:

Ecological interactions between species define the functioning of an ecosystem. These interactions can be represented as a network, where nodes correspond to species and edges represent interactions. Analyzing the structure and variability of these networks provides insights into how ecosystems respond to external factors. Probabilistic model-based approaches are both versatile and well-suited for describing the generative process of these networks while accounting for sources of random variability, such as sampling effects. Many random network models rely on node-exchangeability assumptions, such as the stochastic block model or the expected degree distribution model. However, ecological networks are typically constructed by directly sampling interactions (edges), rather than sampling species (nodes) and their associated interactions. In this context, edge-exchangeable models are more appropriate for ecological data, as they directly model the sampled interactions between pairs of species. These models treat each interaction as a statistical unit, linking two species. This can be reframed as a species sampling problem, where observations are pairs of sampled species. Previous studies have explored edge-exchangeable network models based on species sampling models (SSMs) (Williamson, 2016). However, most of these models assume that the two nodes associated with each edge are sampled conditionally independently, an assumption that is often unrealistic in many network analysis settings, including ecological interactions. This work proposes a class of edge-exchangeable models for networks based on dependent SSMs, such as hierarchical,

enriched, and product (Dombowsky and Dunson, 2024) Gibbs-type processes, incorporating dependency between nodes at each edge end. This class of models provides a new framework for nonparametric model-based network analysis, allowing hypothesis testing, link prediction, and inference of the number of unobserved nodes. This framework is generalized to bipartite networks and applied to the study of ecological interaction networks, such as pollination, seed dispersal, herbivory, plant-ant, and host-parasite networks. References Dombowsky, A. and Dunson, D. B. (2024). Product centered Dirichlet processes for Bayesian multiview clustering. arXiv preprint. Williamson, S. A. (2016). Nonparametric network models for link prediction. Journal of Machine Learning Research.

Keywords: edge-exchangeable networks; dependent Gibbs-type processes; nonparametric inference; ecological interactions

Co-authors: Louise Alamichel, Julyan Arbel

3.53. Kai Lehman

Institution: LMU Munich

Department: Physics

Email: kai.lehman@physik.lmu.de

Talk title:

Bayesian Nonparametric Inference for Misspecified Cosmological Models

Abstract:

Bayesian nonparametric inference provides for a framework which is robust to misspecified models (e.g. Fong et al. 2019, Dellaporta et al. 2022). In the field of cosmology we are presented with a number of high fidelity cosmological simulations which could be used for simulation-based inference (sbi) pipelines. A major concern for this approach, however, is that synthetic observables produced in these simulations are highly dependent on simulation specifics, in particular astrophysical effects (e.g. Chisari et al. 2019). The large disagreement between different simulations in this regard is a major concern for possible systematics in the resulting inference process. In the analysis of state-of-the-art observation campaigns such as EUCLID or DESI, which equip us with enormous statistical power, large parts of the data are discarded due to these modeling uncertainties. The community's inability to accurately describe these physical processes is consequently one of the dominating systematics in current and future observation campaigns, hindering our ability to take full advantage of the statistical power of our data. In our recent paper Lehman et al. 2024, we provided a framework to optimally extract information from cosmological simulations to perform parameter inference. In the spirit of that work, we present an application of nonparametric methods to handle possible misspecifications in this setting. We will take special care in the robustness of our methods and aim to introduce sbi as a reliable inference tool for cosmology.

Keywords: simulation-based inference; model misspecification; applications

Co-authors: Sven Krippendorff, Jochen Weller, Klaus Dolag

3.54. Rayleigh Lei

Institution: University of Michigan

Department: Institute for Social Research

Email: rayleigh@umich.edu

Website: <https://rayleigh.github.io>

Talk title:

Calibrating hierarchical Bayesian subdomain inference with a binary outcome

Abstract:

Researchers are often interested in subdomain estimation and fit hierarchical Bayesian models to accommodate the data structure, introduce prior information, and stabilize small area estimation with small sample sizes. Further, as demonstrated by Calibrated Bayes techniques, Bayesian procedures are desired to have good frequentist properties. However, hierarchical Bayesian models shrink subdomain estimates toward an overall mean and may result in credible intervals that fail to achieve nominal coverage. For binary outcome data, we examine the subdomain inference and extend the Frequentist, but Assisted by Bayes (FAB) interval introduced by Hoff and et al. to improve the nominal coverage and estimation efficiency. We model sparse and heterogeneous group data with hierarchical Bayesian models and flexible prior specifications, including Gaussian Processes and global-local shrinkage prior. We describe how to numerically compute the FAB intervals in the binary case and demonstrate the improvement of FAB through repeated simulation studies. We apply the proposed methods to estimate the COVID infection rates in subgroups based on geography and demographic characteristics.

Keywords: Credible intervals; subdomain estimation; Frequentist, but Assisted by Bayes (FAB) intervals;

Co-authors: Yajuan Si

3.55. Mengbing Li

Institution: University of Michigan

Department: Biostatistics

Email: mengbing@umich.edu

Talk title:

Bridging Exposure and Outcome Clustering: A Bayesian Multilayered Simplex Factor Regression Model for Complex Epidemiological Data

Abstract:

Dietary acculturation can greatly impact chronic disease progression in migrant populations. Yet modeling the interrelationship between a large set of multivariate acculturation exposures and multivariate dietary consumption outcomes is challenging. Mixed membership models are useful for identifying the underlying dietary patterns by positing individuals to a few outcome subgroups of varying degrees, but difficulties arise when these outcomes relate to latent exposure subgroups. We propose a new multilayered simplex factor regression model that simultaneously derives latent acculturation exposure and dietary outcome patterns while capturing the exposure-outcome relationship. The

relationship is modeled via a stick-breaking multivariate logistic regression on the individual membership scores, allowing for easy introduction of auxiliary information and flexible choices of the number of subgroups. We discuss model identifiability conditions and demonstrate this method on a migrant population in the United States.

Keywords: Bayesian methods; Latent class models; Grade of membership models; Multivariate categorical data; Epidemiology

Co-authors: Briana Stephenson, Zhenke Wu

3.56. Jiguang Li

Institution: UChicago Booth

Department: Econometrics & Statistics

Email: jiguang@chicagobooth.edu

Talk title:

Bayesian Computerized Adaptive Testing

Abstract:

Multidimensional computerized adaptive testing (MCAT) has been widely adopted in psychological and educational measurement for its ability to efficiently select a small number of items from large item pools. Despite their theoretical appeal in quantifying measurement uncertainties as distributions, Bayesian approaches to MCAT do not scale well with the number of items and the number of factors due to the computational intensity of Markov Chain Monte Carlo (MCMC) sampling. In this work, we show that many Bayesian item selection rules can be significantly accelerated by directly sampling from latent factor posterior distributions, thereby eliminating the need for expensive MCMC sampling. We propose a novel item selection rule based on Bayesian predictive distributions, which not only connects to the widely used mutual information approach but also achieves much greater computational efficiency. Recognizing the limitations of existing Bayesian item selection approaches which are potentially suboptimal, we further develop a general double Q-learning algorithm to learn an optimal item selection policy. Through simulation and real-data studies, we demonstrate the scalability of our direct sampling methods and the superior performance of the proposed item selection rules. Notably, the Q-learning algorithm achieves faster posterior variance reduction, enabling earlier test termination with significantly reduced online computation.

Keywords: Q-learning, unified skew normal distribution, psychometrics

Co-authors: Veronika Rockova & Robert Gibbons

3.57. Jiaming Liu

Institution: Rice University

Department: Department of Statistics

Email: jl259@rice.edu

Talk title:

Trustworthy variational Bayes with guarantees

Abstract:

Variational Bayes (VB) methods have become increasingly popular for approximating the Bayesian posterior due to their computational efficiency. However, credible intervals derived from VB posterior distributions often exhibit undercoverage, failing to achieve nominal frequentist coverage probabilities. We propose a calibration method that combines sample splitting strategy with the general posterior calibration (GPC) method under the framework of fractional posteriors to address this limitation. Our approach selects an optimal smoothing parameter via sample splitting Bootstrap to ensure proper coverage of credible intervals while maintaining the computational advantages of VB. We also establish theoretical guarantees for our method, showing that the calibrated credible intervals achieve asymptotically correct frequentist coverage. Through extensive simulation studies on Gaussian mixture models and Bayesian mixture linear regression models, we demonstrate that our method outperforms both standard VB and existing calibration approaches and achieves normal frequentist coverage in finite samples.

Keywords: Variational Bayes, fractional posterior, sample splitting Bootstrap, credible interval calibration

Co-authors: Dr. Meng Li

3.58. Joshua Loyal

Institution: Florida State University

Department: Statistics

Email: jd122h@fsu.edu

Website: <https://joshloyal.github.io/>

Talk title:

Generalized Bayesian Inference for Dynamic Random Dot Product Graphs

Abstract:

The random dot product graph (RDPG) is a popular model for network data with extensions that accommodate dynamic (time-varying) networks. However, two significant deficiencies exist in the dynamic RDPG literature: (1) no coherent Bayesian way to update one's prior beliefs about the model parameters due to their complicated constraints, and (2) no approach to forecast future networks with meaningful uncertainty quantification. This work proposes a generalized Bayesian framework that addresses these needs using a Gibbs posterior that represents a coherent updating of Bayesian beliefs based on a least-squares loss function. Furthermore, we establish the consistency and contraction rate of this Gibbs posterior under commonly adopted Gaussian random walk priors. For estimation, we develop a fast Gibbs sampler with a time complexity that is linear in both the number of time points and observed edges in the dynamic network. Simulations and

real data analyses show that the proposed method's in-sample and forecasting performance outperforms that of competitors.

Keywords: gibbs posterior; dynamic network data; network latent space models; statistical network analysis

Co-authors: None

3.59. Luigi Maria Malgieri

Institution: Duke University

Department: Department of Statistical Science

Email: lm145@duke.edu

Talk title:

Consistency for Fleming-Viot Dependent Dirichlet Process

Abstract:

A key feature of the Dirichlet Process is its weak consistency with respect to the true data-generating distribution under iid sampling. We investigate similar results in a partially exchangeable scenario related to a family of Dependent Dirichlet Processes driven by Fleming–Viot diffusions (FV-DDP). These are widely used in population genetics to model large evolving populations undergoing genetic drift and mutation, and were recently shown to induce analytically tractable priors on the space of temporal trajectories whose marginal values are atomic probability measures. We consider vectors of observations collected at different time points, with each vector composed by iid draws from the value of the process at that time. Under this assumption, we show that the smoothing distributions induced by a FV-DDP prior, which are conditional on past, concurrent and subsequent data, converge weakly to the true emission distribution as the vector of sample sizes goes uniformly to infinity. In particular, infinite vectors of observations collected at adjacent times have asymptotically negligible influence on the time of interest. The proof blends classical techniques for Bayesian Non-Parametric consistency with novel bounds on the death process of Kingman's coalescent.

Keywords: Consistency; Partial exchangeability; Dependent Dirichlet Process; Fleming–Viot diffusion; Population genetics

Co-authors: Filippo Ascolani, Matteo Ruggiero

3.60. Shuhei Mano

Institution: The Institute of Statistical Mathematics

Department: Department of Fundamental Statistical Mathematics

Email: smano@ism.ac.jp

Talk title:

A measure-on-graph-valued diffusion: a particle system with collisions and its application

Abstract:

*A diffusion taking value in probability measures on a graph with a vertex set is studied. The masses on each vertices satisfy the stochastic differential equation of the form $dx_i = \sum_{j \in N(i)} \sqrt{x_i x_j} dB_{ij}$ on the simplex, where $\{B_{ij}\}$ are independent standard Brownian motions with skew symmetry and $N(i)$ is the neighbour of the vertex i . A dual Markov chain on integer partitions to the Markov semigroup associated with the diffusion is used to show that the support of an extremal stationary state of the adjoint semigroup is an independent set of the graph. We also investigate the diffusion with a linear drift, which gives a killing of the dual Markov chain on a finite integer lattice. The Markov chain is used to study the unique stationary state of the diffusion, which generalizes the Dirichlet distribution. Although no explicit expression of the density of the stationary state is available, a coupling from the past algorithm enables use of the density as a prior distribution. As an application of the latter diffusion, we discuss a Bayesian graph selection based on computation of marginal likelihoods. This work is published in *Mathematics: Special Issue Random Combinatorial Structures*, 10: 4081 (2022).*

Keywords: Measure valued diffusion; Integer partition; Duality; Dirichlet distribution; Graph selection

Co-authors: None

3.61. Francesco Mascari

Institution: Bocconi University

Department: Department of Decision Sciences

Email: francesco.mascari@phd.unibocconi.it

Website: <https://francescomascari.github.io>

Talk title:

Measuring Partial Exchangeability with Reproducing Kernel Hilbert Spaces

Abstract:

In Bayesian modeling, partial exchangeability is crucial for multi-group settings where the data is structured in distinct but potentially related groups or blocks. The blocks' laws are characterized by as many random probability measures, jointly distributed according to their de Finetti measure. Given two groups, the dependence structure can show a continuous spectrum of behaviors between two extremes: total independence or almost sure equality of the associated random probability measures, resulting in full exchangeability of the observables. This work introduces a novel approach to measure partial exchangeability using Reproducing Kernel Hilbert Spaces (RKHS). We propose a correlation index that detects full exchangeability for a vast class of nonparametric models, such as the hierarchical Dirichlet Process (hDP), preserving analytical tractability under mixing. Our index overcomes the limitations of existing methods as it can be applied both a priori and a posteriori while maintaining stability and tractability. Notably, our RKHS-based index generalizes the set-wise correlation, a popular measure in the literature, which can be retrieved as a special case lacking desirable properties for a kernel, such as continuity and universality. We use our index to investigate the posterior asymptotic behavior of the dependence structure under an hDP prior for two groups, proving that it goes to zero as $O((n_1 n_2)^{-1/2})$ when $\max(n_1, n_2)$ grows to infinity, where n_1 and n_2 are the numbers of observations for the first and second group, respectively. Numerical

simulations confirm our theoretical findings, highlighting the effectiveness of Gaussian and Laplace kernels in this context. This work opens new perspectives on ways to measure and analyze the posterior dependence of Bayesian nonparametric models.

Keywords: Partial Exchangeability; Dependence Measure; Hierarchical Dirichlet Process; Reproducing Kernel Hilbert Spaces;

Co-authors: Hugo Lavenant; Marta Catalano

3.62. Carson McKee

Institution: King's College London

Department: Mathematics

Email: carson.mckee@kcl.ac.uk

Website: <https://carsonmckee.github.io/carsonmckee.io/>

Talk title:

Mutually Dependent Bernoulli Processes for Multivariate Change-Point Detection

Abstract:

Financial and hydrological time series data often exhibit sharp structural breaks, or change-points, driven by events such as market crashes or sudden heavy rainfall. In the multivariate setting, these changes may occur asynchronously across series. It is often the case that the occurrence of a change-point in one series affects the probability of a subsequent change occurring in another. For example, financial data often exhibits “contagion”, where turmoil spreads from one country to another due to latent intermarket dependencies. Similarly, the flow rates of rivers which are part of the same network often exhibit change-points at similar times. Here, we develop a multivariate change-point prior, which explicitly models this dependence using discrete time and mutually dependent point processes. Under this prior, the probability of a change-point occurring in a series at a given time is dependent on recent changes in other series. This dependence is driven by a latent network structure linking the change-points in different series. We show that the conditional distribution of the change-points in each series follows a Product Partition Model with a cohesion function that depends on the change-points in the other series. Obtaining posterior estimates under this model is nontrivial and various MCMC strategies are discussed. The model is demonstrated in simulated and real datasets (financial and hydrological) and is shown to uncover latent dependencies linking change-points in different series.

Keywords: Change-point; time series; multivariate

Co-authors: Maria Kalli

3.63. Daniele Micheletti

Institution: University of California, Irvine

Department: Statistics
Email: dmichele@uci.edu

Talk title:

Towards Deep Generative Models for Exchangeable Graphs: Exploring Connections with Bayesian Nonparametrics

Abstract:

The statistical analysis of random graphs, where nodes represent entities and edges represent their relationships, is fundamental in many fields including the biological and social sciences. While various approaches exist for modeling network data, this work focuses on developing flexible models that respect the theoretical properties of exchangeable random graphs, as characterized by the Aldous-Hoover representation theorem. This representation establishes that any exchangeable random graph can be generated through a continuous function, known as a graphon, that determines the probability of edge formation between any pair of nodes. This framework is adopted by several Bayesian nonparametric (BNP) models, including the broad family of Extended Stochastic Block Models (ESBM), which assign flexible Gibbs-type priors to node-specific variables. However, restrictive assumptions are often made on the form that node-specific variables take, and on the way that edge probabilities are parameterized. We propose a deep generative model that aims to learn the latent graphon function directly from observed graph data. Our approach leverages recent developments in graph neural networks as computational tools to flexibly approximate infinite-dimensional graphons. Specifically, we develop a variational inference framework which learns both a latent representation for each node, and the continuous function governing their interactions. We compare these deep generative models to established BNP models, such as the ESBM and Latent Space Models, to provide insights into how different modeling approaches capture the underlying structure of exchangeable random graphs.

Keywords: graph neural networks; network data; deep generative models; graph exchangeability

Co-authors: Federica Zoe Ricci; Erik Sudderth

3.64. Giosue Migliorini

Institution: University of California, Irvine
Department: Statistics
Email: gmiglior@uci.edu
Website: <https://giosueio.github.io/>

Talk title:

Neural Guidance for Variational Interacting Particle Systems with Discrete Latent States

Abstract:

Many phenomena such as wildfires and pandemics can be cast as systems of interacting particles, taking values in a discrete state space, evolving in continuous time. Inferring complex latent dynamics from partial, noisy observations remains a challenge, especially in

high dimensions where traditional methods struggle with scalability. We propose a novel amortized variational approach to smoothing and forecasting that goes beyond restrictive mean-field assumptions. Our method parameterizes a guidance function with a neural model to approximate Doob's h-transform, trained using a novel gradient estimator. An EM scheme enables learning the parameters via maximum likelihood alongside the variational posterior. The removal of factorization constraints improves performance on problems involving high-dimensional, highly coupled stochastic processes. We demonstrate our method's effectiveness on synthetic epidemic data with graph-structured interactions and real wildfire spread trajectories.

Keywords: markov jump processes, interacting particle systems, neural networks, variational inference, latent dynamics

Co-authors: Padhraic Smyth

3.65. Sepideh Mosaferi

Institution: University of Massachusetts Amherst

Department: Mathematic and Statistics Department

Email: smosaferi@umass.edu

Website: <https://www.sepidehmosaferi.com>

Talk title:

Unbiased Predictor for Skewed Response Variable with Measurement Error in Covariate

Abstract:

We introduce a new predictor when the Fay-Herriot normal error model is fitted to a logarithmically transformed response variable, and the covariate is measured with error. This framework has been previously studied in the literature, but the empirical Bayesian predictor cannot perform uniformly better than the direct estimator. Our proposed predictor in this talk is unbiased and can perform uniformly better than the previous one in the literature. We derive an approximation of the mean squared error (MSE) for the predictor. The prediction intervals based on the MSE suffer from coverage problems. Thus, we propose a non-parametric bootstrap prediction interval which is more accurate. This problem is of great interest in small area applications since statistical agencies and agricultural surveys are often asked to produce estimates of right skewed variables with covariates measured with errors. With Monte Carlo simulation studies and two Census Bureau's data sets, we demonstrate the superiority of our proposed methodology.

Keywords: Bayesian estimator; prediction interval; transformation.

Co-authors: Malay Ghosh: Distinguished Professor, Department of Statistics, University of Florida, USA Shonosuke Sugawara: Associate Professor, Faculty of Economics, Keio University, Japan

3.66. Andre Nunes

Institution: Instituto Superior Técnico da Universidade de Lisboa - IST da Universidade de Lisboa Portugal

Department: Mathematics Department
Email: andrenunes.pro.vocal.vox@gmail.com

Talk title:

Joint modeling of longitudinal binary responses: A nonparametric Bayesian approach using acute and chronic malnutrition data

Abstract:

In the analysis of longitudinal data, Gaussian random effects are usually adopted to deal with individual dependence over time. We question the Gaussian assumption, especially for analyzing longitudinal binary response variables that are naturally dependent and jointly distributed according to Bayesian Nonparametric (BNP) random effects. That is, conditionally on common random effects, the two longitudinal binary response variables are considered independent, facilitating the construction of the corresponding likelihood function and therefore the inference on the model parameters. Nonparametric and semiparametric approaches for statistical data analysis have the advantage of not imposing a given distribution for either responses or common random effects. For a longitudinal binary two outcomes, we assume Bernoulli sampling model and BNP random effects for linking the two binary responses. Regarding BNP models, Dirichlet process, Pitman-Yor process, and some of their variations are performed. These models were implemented in R package NIMBLE, evaluating costs and benefits on computational time and fitting BNP models with pre-clustering of the patients based on similarities matrices. Markov Chain Monte Carlo (MCMC) methods have been also employed for inferential purposes of some parameters such as regression coefficients. The motivation is due to a binary data study from a longitudinal four-arm randomized parallel trial conducted in Bengo Angolan province to explore the effects of four interventions on wasting and stunting outcomes. A total of 121 children with intestinal parasitic infections were allocated to the four arms. Finally, a simulation study is run to evaluate the performance of the proposed BNP joint model.

Keywords: Bayesian Nonparametric; Mixed Model; Longitudinal Data; Dirichlet Process; Pitman-Yor Process

Co-authors: Giovani L. Silva Luzia Gonçalves

3.67. Jeunghun Oh

Institution: Seoul National University
Department: Department of Statistics
Email: nomad1994@snu.ac.kr
Website: <https://jh0h.github.io>

Talk title:

Asymptotic Properties of Lévy Adaptive B-spline Regression

Abstract:

One of the interesting fields of nonparametric regression analysis is to examine how well a proposed model approximates a particular function space. In this paper, we investigate the asymptotic properties of the LABS (Lévy Adaptive B-Spline regression) model. Also, we

show LABS can approximate a true function in a Besov space under reasonable regularity conditions. The LABS model is an implementation of a LARK using B-spline bases as its generating kernel. The mean function of the LABS model comprises linear combinations of B-spline bases with varying polynomial degrees. This enables the LABS model to capture the varying local smoothness of a true function, i.e. jump discontinuities, sharp peaks, etc. As a result, the model can express functions in general function spaces such as Besov spaces. In this regard, we verify that the asymptotic performance of LABS attains near optimal posterior concentration rate.

Keywords: Nonparametric Regression; LARK; B-splines; Besov Space; Posterior Concentration

Co-authors: Sewon Park (Samsung SDS), Jaeyong Lee (Seoul National University)

3.68. Oyebayo Ridwan Olaniran

Institution: University of Ilorin, Nigeria

Department: Statistics

Email: olaniran.or@unilorin.edu.ng

Talk title:

Mixed-Effects Bayesian Additive Regression Trees for High-Dimensional Longitudinal Genomic Data

Abstract:

Random Forest (RF) and its Bayesian counterpart, Bayesian Additive Regression Trees (BART), have emerged as powerful tools for analyzing high-dimensional classification and regression problems. However, their application has predominantly focused on cross-sectional data or single-response endpoints, limiting their utility for longitudinal studies. High-dimensional genomic data, particularly in cancer research, often features repeated measurements with the number of predictors (genes) vastly exceeding the number of observations (patients), posing unique analytical challenges. This study extends the BART nonparametric framework to accommodate high-dimensional longitudinal data by integrating mixed-effects modelling. The proposed Mixed-Effects Bayesian Additive Regression Trees (ME-BART) methodology introduces a tree-based method that effectively addresses sparsity and variability inherent in such data. Extensive simulation studies and real-life applications demonstrate the superiority of ME-BART over conventional methods in terms of predictive accuracy and interpretability.

Keywords: Mixed-Effects Model; Bayesian Additive Regression Trees (BART); High-Dimensional Data; Longitudinal Data; Genomic Data Analysis

Co-authors: Saidat Fehintola Olaniran

3.69. Saidat Fehintola Olaniran

Institution: Kwara State University, Nigeria

Department: Statistics and Mathematical Sciences

Email: saidat.olaniran@kwasu.edu.ng

Talk title:

Bayesian Tapered Narrowband Least Squares for Fractional Cointegration Testing in Panel Data

Abstract:

Fractional cointegration has been extensively examined in time series analysis, but its extension to panel data with unobserved heterogeneity remains underdeveloped. This paper introduces the Bayesian Tapered Narrowband Least Squares (BTNBLs) methodology, a robust framework for testing fractional cointegration in panel data with fixed effects. The proposed test statistic, derived under the Bayesian nonparametric framework, incorporates variability in long-term cointegrating vectors and fractional cointegration parameters across cross-sectional units. The methodology leverages tapered periodograms to minimize boundary effects and employ narrowband least squares to enhance precision in estimating fractional differencing parameters. The performance of the BTNBLs test is evaluated using extensive Monte Carlo experiments across diverse data-generating processes, accounting for heteroskedasticity and structural differences between units. Empirical validation is provided through an application to agricultural market data from 36 countries, demonstrating the test's capability to detect fractional cointegrating relationships. Results reveal a statistically significant long-run equilibrium between exchange rates and labour wages, even in the presence of idiosyncratic cross-sectional disturbances.

Keywords: Fractional Cointegration; Tapered Narrowband Least Squares (TNBLs); Panel Data; Fixed Effects; Bayesian Kernel Estimation

Co-authors: Oyebayo Ridwan Olaniran

3.70. AKINLOLU OLOSUNDE

Institution: OBAFEMI AWOLOWO UNIVERSITY, UILE-IFE

Department: MATHEMATICS

Email: aolosunde@oauife.edu.ng

Talk title:

Advancing Inventory Management with Bayesian Probabilistic Demand: A Novel Framework for Decision-Making Under Uncertainty

Abstract:

Inventory management is a crucial element in supply chain operations, where efficient stock control significantly impacts cost reduction and service quality. Traditional inventory models often assume deterministic or stationary probabilistic demand patterns, which may not accurately reflect real-world scenarios with dynamic and uncertain demand. non-parametric Bayesian inference provides a powerful statistical approach for updating

demand distributions as new information becomes available. This project aims to develop innovative inventory models incorporating Bayesian probabilistic demand forecasting to enhance decision-making under uncertainty with applications to industrial stocks

Keywords: inventory model, nonparametric Bayesian, stocks, decision theory

Co-authors: None

3.71. Garritt Page

Talk title:

Uncertainty Quantification in Bayesian Clustering

Abstract:

We carefully establish an approach to assessing uncertainty in partition estimation that is interpretable, easily computed, and is available for any Bayesian clustering method that employs MCMC techniques. Our approach is based on constructing a subset of the space of partitions that is anchored to a sub-partition that accumulates a pre-specified amount of posterior probability. The sub-partition can then be used to construct metrics that are able to express uncertainty on subsets of the partition and also on the entire partition in a holistic way. Additionally, we show how the sub-partition can be used to estimate cluster-specific parameters for clusters that accumulate a pre-specified amount of posterior mass.

Keywords: Bayesian clustering, uncertainty quantification, partition modeling, cluster-specific parameter estimation.

Co-authors: Andrés Felipe Barrientos, David B Dahl, David B Dunson

3.72. Ansh Pal

Institution: University of Bordeaux

Department: ED SP2

Email: anshpal@gmail.com

Website: <https://ansh07.github.io/anshpal/>

Talk title:

Collapsed variational inference for Dirichlet process mixture model inference with unknown covariance and adaptive concentration parameter

Abstract:

Application of Dirichlet process has gained popularity in the field of mixture modeling, rendering promising results for density estimation and data clustering. Although MCMC methods have bridged the gap between theory and application for such non-parametric Bayesian models, issues like slow and difficult convergence still persists. An alternative is to look for a deterministic approximation approach, namely variational inference methods. We propose a collapsed variational inference method that performs Dirichlet Process mixture modeling while varying the concentration parameter α . As long as this concentration parameter α is estimated a posteriori, consistency for the posterior number

of clusters is achieved. Our model adopts a stick-breaking representation of the Dirichlet process, collapses the mixture weights by integration and uses a weakly-informative gamma prior for α , in order to optimize the evidence lower bound (ELBO) leveraging a coordinate ascent algorithm. The covariance matrix of the observed data is generally unknown. To mitigate the impact of its prior distribution on the posterior number of non-empty clusters, we propose to incorporate weakly informative scalar priors on elements of the lower triangular matrix obtained by the Cholesky factorization of the covariance instead of a more common multivariate prior on the full symmetric matrix itself in our Bayesian model. We evaluate the performance of our approach in controlled numerical simulation, and use it to study cancer subtypes in a gene expression data set.

Keywords: Dirichlet mixture model; Variational inference; Clustering; Unstructured covariance; Concentration parameter

Co-authors: Rodolphe Thiébaud and Boris Hejblum

3.73. Ji Won Park

Institution: Genentech

Department: Prescient Design

Email: park.ji_won@gene.com

Website: <https://jiwoncpark.github.io>

Talk title:

Targeted neural posterior estimation

Abstract:

We present Targeted Neural Posterior Estimation (TNPE), a framework that integrates Targeted Minimum Loss Estimation (TMLE) with Monte Carlo Efficient Influence Functions (MC-EIF) into the sequential simulation-based updates of the variational posterior distribution. TNPE focuses on specific parameters of interest by leveraging TMLE to correct for high-dimensional nuisance parameters and MC-EIF to rapidly compute influence functions on top of neural network posterior models. As a motivating application, we demonstrate TNPE in the context of inferring binding kinetics from surface plasmon resonance (SPR) sensorgrams --- a key technique in drug discovery for quantifying interactions between drug molecules and their targets. While SPR analysis is often complicated by instrumental artifacts and secondary biophysical processes, these nuisance effects can be simulated at high fidelity using commercially available software. We jointly model nuisance effects and primary binding parameters with simulation-based neural posterior estimation, while prioritizing the binding parameters of interest. Empirically, TNPE achieves accurate and precise recovery of binding parameters in simulated settings and generalizes well to real SPR data. By enabling robust inference of finite-dimensional functionals in the presence of high-dimensional nuisance parameters, TNPE promises to extend simulation-based inference to complex biological settings.

Keywords: semiparametric statistics, simulation-based inference, efficient influence functions, targeted minimum loss estimation

Co-authors: John Quinn, Richard Bonneau, Kyunghyun Cho

3.74. Miguel Pebes-Trujillo

Institution: Nanyang Technological University, Singapore

Department: School of Materials Science & Engineering

Email: mpebes@ntu.edu.sg

Talk title:

Bayesian Methods for Crop Trait Optimization in Precision Agriculture Research

Abstract:

We present a model for identifying the optimal treatment to achieve one or multiple desired traits in plants, such as maximum yield and mineral content for improved nutrition. When applied over time, our model enables continuous optimization of these traits. Experiment-based research involves applying treatments to plants and assessing treatment performance by comparing how closely the traits exhibited after the intervention align with the ideal. Often, conflicting traits need to be achieved, requiring trade-offs that complicate the goal of determining the overall optimal treatment. Our framework defines overall optimality through treatment rankings and infers these rankings to order treatments from most to least effective. The reliability of ranking estimates is demonstrated by analyzing synthetic and real data from precision agriculture research. We showcase intervention through a crop yield-maximization experiment, comparing plants irrigated under a reinforcement learning (RL) regime steered by our model with those under standard treatments. Plants optimized by our model showed significant improvement over time compared to the rest. Our model adopts a Bayesian approach and uses a Markov Chain Monte Carlo algorithm for parameter estimation. We also discuss its potential for adoption as a standard tool for experiment evaluation and as a principled counterpart to classic RL in dynamic trait-optimization contexts. Additionally, we explore other Bayesian applications that inform this ranking model, including change-point detection and model selection in proximal spectral sensing. Finally, we examine the potential of applying Bayesian nonparametric methods to the different components of our framework to improve controllability in urban agriculture systems.

Keywords: Probabilistic ranking; paired comparisons; multi-trait optimization; diachronic Bayesian.

Co-authors: Itamar Shenhar, Aravind Harikumar, Ittai Herrmann, Menachem Moshelion, Kee Woei Ng, Matan Gavish

3.75. Anna Petranzan

Institution: University of Milano - Bicocca

Email: a.petranzan@campus.unimib.it

Talk title:

The Gnedin model for biodiversity estimation

Abstract:

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.

Keywords: Species Sampling problems; Gibbs-type priors; Gnedin model

Co-authors: Federico Camerlenghi, Tommaso Rigon

3.76. Sihyeon Pyeon

Institution: Yonsei University

Department: Department of Statistics and Data Science

Email: shpyeon@yonsei.ac.kr

Talk title:

Bayesian Adaptive Triangulation

Abstract:

Conventional nonparametric regression methods for bivariate function estimation struggle with irregular domain shapes and varying levels of function smoothness. This study proposes Bayesian adaptive triangulation, a novel framework that addresses these challenges by developing locally adaptive splines over polygonal domains within the Bayesian framework. Our approach determines the optimal constrained Delaunay triangulation of a domain based on the data, allocating finer resolution to regions with high variability while simplifying smoother areas. This ensures both flexibility and precision in capturing localized patterns while preserving the geometric characteristics of the spatial domain. A carefully designed prior incorporates a penalty term for spline smoothness and a model selection term, effectively balancing estimation accuracy and model complexity. We demonstrate that Bayesian adaptive triangulation achieves the optimal posterior contraction rate, adapting to the unknown smoothness of the target function. Simulation studies show that our method outperforms existing models, achieving improved estimation accuracy while maintaining lower model dimensionality.

Keywords: Bayesian nonparametrics; bivariate splines; constrained Delaunay triangulation; spatial adaptation; posterior contraction

Co-authors: Sunwoo Lim (Marshall School of Business, University of Southern California, Los Angeles, CA, USA) ; Jaewoo Park (Department of Statistics and Data Science, Department of Applied Statistics, Yonsei University, Seoul, Republic of Korea); Seonghyun Jeong (Department of Statistics and Data Science, Department of Applied Statistics, Yonsei University, Seoul, Republic of Korea)

3.77. SYED EHSAN AR RAFI

Institution: Texas Tech University

Department: Mathematics and Statistics

Email: srafi@ttu.edu

Website: <https://www.linkedin.com/in/syed-ehsan-ar-rafi/>

Talk title:

Application of Quantile Transform Method while Simulating Random Variates from Densities with Non-Closed Form of Cumulative Distribution Function

Abstract:

Computer-generated random variates from different densities have been used in various simulation applications as stochastic inputs. The quantile transform method is one of the most common and efficient variates generation techniques, which requires the inverse of the cumulative distribution function (cdf) of a desired density. However, the application of the quantile transform method is limited to variates for which the cdf or inverse of the cdf is in simple closed form. This paper discusses the applicability of quantile transform method to generate variates for non-closed form of the cdf or inverse of the cdf and numerically solves the non-closed form of the cdf or inverse of the cdf. The application of the quantile transform method is then investigated through a simulation study where normal, gamma and non-central densities are considered. From the simulation study, it is observed that generating variates from such densities via quantile transform method is not precluded because of the non-closed form

Keywords: Statistics, Simulation, Quantile Transformation Method

Co-authors: Anamul Haque Sajib, Sabina Sharmin

3.78. Federica Zoe Ricci

Institution: University of California Irvine

Department: Statistics

Email: fzricci@uci.edu

Website: <https://federicazoe.github.io/>

Talk title:

Approximately sparse mixed-membership stochastic blockmodels for the analysis of large networks

Abstract:

Mixed-membership stochastic blockmodels and their extensions have proven successful at modeling network data based on inferring nodes' memberships in a set of latent communities. But despite empirical evidence suggesting that many real-world networks are sparse, models that regard networks as exchangeable arrays (including stochastic blockmodels) can only generate dense networks, where the number of edges grows quadratically with the number of nodes. Some Bayesian nonparametric approaches to community detection, based on completely random measures, overcome this limitation. But how important is it to model sparsity? To address this question, we design novel modeling

and computational approaches to construct (approximately) sparse mixed-membership stochastic blockmodels that can be fit to networks with hundreds of thousands of nodes. We achieve this by leveraging finite approximations of infinite-activity, completely random measures and stochastic optimization techniques. Our work aims to unlock the full benefits of sparse models, which can theoretically both yield a better fit to sparse network data and achieve much better asymptotic complexity than their dense counterparts. In this poster, we illustrate our methods and present experiments on real-world networks that compare our framework to sparse and dense alternatives on a set of model-fit and prediction metrics.

Keywords: network analysis; community detection; clustering; MCMC; dirichlet process

Co-authors: Daniele Micheletti; Michele Guindani; Erik Sudderth

3.79. Carlos E. Rodriguez

Institution: IIMAS-UNAM

Department: Department of Probability and Statistics

Email: carloserwin@sigma.iimas.unam.mx

Website: <https://sigma.iimas.unam.mx/carloserwin/index.html>

Talk title:

Martingale Posterior Inference for Finite Mixture Models and Clustering

Abstract:

Martingale posterior inference is employed to quantify uncertainty for a finite mixture model with an unknown number of components. New ideas for Bayesian analysis, particularly focusing on martingale posterior distributions, are applied to focus on clustering. The fundamental concept involves constructing appropriate martingales for the unknown parameters. A key outcome of this approach is the ability to conduct posterior analysis of clusters while circumventing the label-switching problem. This methodology is further extended to finite populations, where the mixture is utilized to impute missing part of the population and perform inference for the parameter of interest. The proposed methodology is demonstrated through the analysis of four real datasets.

Keywords: Akaike weights; EM algorithm; Finite population; Model averaging; Predictive inference.

Co-authors: Stephen G. Walker and Ramsés H. Mena

3.80. Giovanni Romanò

Institution: Bocconi University

Department: Decision Sciences

Email: giovanni.romano4@phd.unibocconi.it

Website: <https://giovanni-romano.github.io/>

Talk title:

Dependent stochastic block models for sequences of bipartite networks

Abstract:

We develop a Bayesian nonparametric model for inferring changes in the stochastic block structures within a sequence of networks indexed by a predictor. This model is originally motivated by the attempt to learn age-specific group structures in bipartite causes-of-death networks encoding co-occurrences among underlying and multiple death causes for different age groups in the US population. To address the above goal we substantially generalize state-of-the-art stochastic block models to account for (i) categorical edges, (ii) bipartite settings and (iii) sequences of networks indexed by a predictor. This is accomplished by assuming separate partition structures for the rows and columns of the adjacency matrices characterizing the observed sequence of networks, while allowing these partition structures to change smoothly as a function of the predictor indexing such a sequence. To this end, we extend the dependent random partition process from Page et al. (2022) to a general Gibbs-type prior framework, further supervised by external node attributes via a PPMx reinforcement. Given these partitions, the edges within each block are modeled jointly under a Dirichlet-categorical likelihood. Bayesian inference under the above formulation proceeds through a collapsed Gibbs sampler combined with a parallel tempering scheme to further improve mixing and convergence. Such a latter strategy facilitates the exploration of the domain of the posterior that is hindered by the intricate nature of the space of partitions. As illustrated in simulation studies and in the original motivating application, the novel formulation we propose provides a powerful model to infer changes in grouping structures underlying sequences of networks indexed by a predictor. Preliminary results in the causes-of-death networks application unveil interesting and yet-unexplored patterns in the composition, evolution and interactions of causes-of-death clusters, with relevant policy implications.

Keywords: age-specific causes-of-death networks; dependent random partition models; sequences of bipartite networks; stochastic block model

Co-authors: Daniele Durante Cristian Castiglione

3.81. Artiom Rumiancev

Institution: King's College London

Department: Department of Mathematics

Email: artiom.rumiancev@kcl.ac.uk

Website: <https://github.com/artrumiancev>

Talk title:

Adaptively slice sampling mixture models

Abstract:

The slice-efficient sampler of Kalli et al. (2011) is a popular method for fitting infinite mixture models with a wide range of stick-breaking priors. The sampler relies on the automatic truncation of the infinite mixture through the use of latent uniform random variables, which determine the minimum number of mixture components required to proceed with the chain. The general representation of the slice-efficient scheme introduces an auxiliary positive sequence, which can be used to control the balance between efficiency and computational time. The form of the sequence and parameter choice is a non-trivial task since there are no optimality results. In order to tune the sequence's parameters, we

propose using heteroscedastic Bayesian optimization as a method to maximize the relative expected squared jumping distance, an objective function that captures the relationship between the mixing time of the Markov chain and the running cost of the algorithm. This general optimization procedure guarantees diminishing adaptation is parallelizable and can be applied to a variety of stick-breaking processes.

Keywords: Stick breaking priors; Slice sampling mixture models; Bayesian optimization; Gaussian process; Markov chain Monte Carlo

Co-authors: Maria Kalli

3.82. S. Yaser Samadi

Institution: Southern Illinois University Carbondale IL

Department: School of Mathematical and Statistical Sciences

Email: ysamadi@siu.edu

Talk title:

Modeling Dynamic Interactions in Time Series Mixed Data: Bayesian CopulaFactor Autoregressive Approach

Abstract:

In this talk, we present two novel Bayesian copula factor autoregressive models designed specifically for analyzing time series mixed data, with a particular focus on accommodating main effects and capturing dynamic interactions. Our research is motivated by the challenging task of inferring and understanding the intricate interdependencies among macroeconomic variables and stock market indices. The proposed models incorporate latent factors in both the response time series and the high-dimensional mixed-type covariates, leveraging the quadratic autoregression (QAR) model while assuming conditional independence. By including latent variables in the response time series, our models not only enable efficient dimension reduction but also effectively characterize the main effects and interactions of the covariates. We employ a semiparametric time series extended rank likelihood approach to model the margins of explanatory variables, resulting in a reduction of estimated parameters and facilitating the development of a fast computational algorithm. To compute the posterior distribution of the latent factors and model parameters, we propose flexible Bayesian algorithms, including the Metropolis-Hastings (MH) and Forward Filtering Backward Sampling (FFBS) algorithms within the Gibbs sampling framework. The theoretical results and the proposed algorithms are thoroughly evaluated through comprehensive simulation studies. Furthermore, we demonstrate the practical applicability of the proposed model by applying it to a macroeconomic dataset, thereby showcasing its effectiveness in real-world scenarios. Overall, our research contributes to advancing the understanding and modeling of dynamic interactions within time series mixed data, offering valuable insights and tools for various domains.

Keywords: Multivariate time series, Quadratic autoregression, Bayesian inference, Factor analysis, Copula models, Parameter-expansion method

Co-authors: Hadi Safari-Katesari and Samira Zaroudi

3.83. Hristo Sariev

Institution: Sofia University "St. Kliment Ohridski", Bulgaria

Department: Faculty of Mathematics and Informatics

Email: h.sariev@math.bas.bg

Talk title:

Characterization of exchangeable Hoeffding decomposable sequences

Abstract:

The Hoeffding decomposition is a classical method for studying U-statistics based on sequences of i.i.d. random variables. The concept was extended by G. Peccati [Ann. Probab. 32 (2004) 1796-1829] to exchangeable processes, but apart from a few examples, it was not known which exchangeable processes were explicitly Hoeffding decomposable. In this study, we provide a complete characterization of the class of exchangeable Hoeffding decomposable processes in terms of their predictive distributions. In particular, we show that they are mixtures of exchangeable measure-valued Polya urn sequences, from which we derive the general form of their prior distributions.

Keywords: Hoeffding decomposition; Exchangeability; Polya urns; Predictive distributions

Co-authors: Mladen Savov; Stefan Gerdjikov

3.84. Mahlon Scott

Institution: University of Central Florida

Department: Dept. of Statistics and Data Science

Email: mahlon.scott@ucf.edu

Talk title:

Spatial Storm Surge Risk Modeling

Abstract:

Storm surges present a severe risk to coastal communities and infrastructure, underscoring the critical importance of accurately estimating extreme events such as the 100-year return surge. These estimates are essential not only for effective hazard assessment but also for informing resilient coastal design. Inspired by principles of extreme value modeling, this project introduces a Bayesian hierarchical model integrated with Gaussian processes to account for spatial random effects. Furthermore, we integrate the fixed effects of atmospheric covariates (such as pressure), enhancing extensibility to nearby unmonitored locations. This approach enhances the precision of long-return period storm surge estimates and enables the seamless generalization of predictions to unmonitored nearby coastal regions, much like the way advanced Bayesian frameworks are applied to high-dimensional neuroimaging or spatiotemporal data, bridging gaps between observations and uncharted territories. Storm surges present a severe risk to coastal communities and infrastructure, underscoring the critical importance of accurately estimating extreme events such as the 100-year return surge. These estimates are essential not only for effective hazard assessment but also for informing resilient coastal design. Inspired by principles of extreme value modeling, this project introduces a Bayesian

hierarchical model integrated with Gaussian processes to account for spatial random effects. Furthermore, we integrate the fixed effects of atmospheric covariates (such as pressure), enhancing extensibility to nearby unmonitored locations. This approach enhances the precision of long-return period storm surge estimates and enables the seamless generalization of predictions to unmonitored nearby coastal regions, much like the way advanced Bayesian frameworks are applied to high-dimensional neuroimaging or spatiotemporal data, bridging gaps between observations and uncharted territories.

Keywords: coastal hazard assessment; gaussian process; storm surge; 100-year returns; bayesian hierarchical modeling

Co-authors: Hsin-Hsiung Huang (UCF)

3.85. Jessalyn Sebastian

Institution: University of California, Irvine

Department: Statistics

Email: jessalys@uci.edu

Talk title:

Gaussian Process Priors with Markov Properties for Effective Reproduction Number Estimation

Abstract:

Many quantities characterizing infectious disease outbreaks — like the effective reproduction number R_t , which describes the average number of secondary infections a newly infected individual will cause over the course of their infection — need to be modeled as time varying parameters. It is common practice to use Gaussian random walks as priors for estimating such functions in Bayesian analyses. In this setting, however, the random walk prior may be too flexible, as it can result in high posterior variance and fails to capture prior scientific knowledge about the estimand. For example, the effective reproduction number often oscillates smoothly around the critical value 1 over time. This suggests that other Gaussian process priors such as the Ornstein-Uhlenbeck process, a mean-reverting modified random walk, or Integrated Brownian Motion, the time integral of the Brownian Motion, may be more appropriate. We use simulated outbreak data to compare the performance of these priors with the Gaussian random walk prior in estimating the effective reproduction number and contrast their performance when applied to real California SARS-CoV-2 data collected at the county level.

Keywords: nowcasting; Bayesian inference; COVID-19; infectious diseases, stochastic processes

Co-authors: Volodymyr Minin

3.86. Gözde Sert

Institution: Texas A&M University

Department: Statistics

Email: gozdesert@tamu.edu

Talk title:

Bayesian Semi-supervised Inference via a Debiased Modeling Approach

Abstract:

Inference in semi-supervised (SS) settings has received a great amount of attention in recent years due to increased relevance in modern big-data problems. In a typical SS setting, there is a much larger sized unlabeled data containing only observations for predictors, in addition to a moderately sized labeled data involving observations for both an outcome and a set of predictors. Such data arises naturally from settings where the outcome, unlike the predictors, is costly to obtain. One of the primary statistical objectives in SS settings is to explore whether parameter estimation can be improved by exploiting the unlabeled data. This article proposes a novel Bayesian approach to SS inference for the population mean estimation problem. The proposed approach provides improved and optimal estimators both in terms of estimation efficiency as well as inference. The method itself has several interesting artifacts. The central idea behind our method is to model certain summary statistics of the data rather than specifying a probability model for the entire raw data itself. Specifying appropriate summary statistics crucially relies on a debiased representation of the population mean in terms of nuisance parameters. Combined with careful usage of sample splitting, our debiasing approach mitigates the effect of bias due to slow rates or misspecification of the nuisance parameter from the posterior of the final parameter of interest. We establish concrete theoretical results, via Bernstein--von Mises theorems, validating all our claims and further supporting them through extensive numerical studies. To the best of our knowledge, this is the first work in Bayesian inference for SS settings. We also believe that the central idea of this article will be more broadly applicable to Bayesian semi-parametric inference.

Keywords: Robustness and efficiency; Bayesian semi-parametric inference; Debiasing; Sample splitting; Bernstein--von Mises theorem.

Co-authors: Abhishek Chakraborty and Anirban Bhattacharya

3.87. Mostafa Shams Esfand Abadi

Institution: Wake Forest University

Department: Department of Statistical Sciences

Email: shamsm@wfu.edu

Website: <https://github.com/m0stafa-shams>

Talk title:

Bayesian nonparametric regression for insurance claims data

Abstract:

We often want to model insurance claims as a function of covariates, which leads to a regression problem. Parametric regression models assume the same functional form tying the response to the covariates for each data point. These models are not flexible enough and can fail to accurately capture at the individual level, the relationship between the covariates and the insurance claims, which are often multimodal, highly skewed, and heavy-tailed. In this poster presentation, I explore the use of Dirichlet process mixture models (DPMM) to model claims frequency and severity based on covariates. In particular, I model claims frequency as a mixture of Poisson regression and the logarithm of claims severity as a mixture of normal regression and use the Dirichlet process (DP) as a prior for the mixing distribution over the regression parameters. Unlike parametric regression, such models allow each data point to have its individual regression parameter vector and due to the clustering property of Dirichlet process prior, the data points are then clustered by their shared regression parameters. I describe model fitting using Markov chain Monte Carlo (MCMC) methods and illustrate their applicability using real motor insurance claims data.

Keywords: Bayesian nonparametric regression; Dirichlet process mixture model; Insurance claims modeling

Co-authors: Kaushik Ghosh

3.88. Ziyi Song

Institution: University of California, Irvine

Department: Department of Statistics

Email: ziyi9@uci.edu

Website: <https://ziyi-song-stats.github.io/>

Talk title:

Clustering computer mouse tracking data with informed hierarchical shrinkage partition priors

Abstract:

Mouse-tracking data, which record computer mouse trajectories while participants perform an experimental task, provide valuable insights into subjects' underlying cognitive processes. Neuroscientists are interested in clustering the subjects' responses during computer mouse-tracking tasks to reveal patterns of individual decision-making behaviors and identify population subgroups with similar neurobehavioral responses. These data can be combined with neuro-imaging data to provide additional information for personalized interventions. In this article, we develop a novel hierarchical shrinkage partition (HSP) prior for clustering summary statistics derived from the trajectories of mouse-tracking data. The HSP model defines a subjects' cluster as a set of subjects that gives rise to more similar (rather than identical) nested partitions of the conditions. The proposed model can incorporate prior information about the partitioning of either subjects or conditions to facilitate clustering, and it allows for deviations of the nested partitions within each subject group. These features distinguish the HSP model from other bi-clustering methods that typically create identical nested partitions of conditions within a subject group. Furthermore, it differs from existing nested clustering methods, which define clusters based on common parameters in the sampling model and identify subject groups by different distributions. We illustrate the unique features of the HSP model on a mouse tracking

dataset from a pilot study and in simulation studies. Our results show the ability and effectiveness of the proposed exploratory framework in clustering and revealing possible different behavioral patterns across subject groups.

Keywords: Bayesian nonparametrics; clustering; Dirichlet process; nonexchangeable random partitions.

Co-authors: Weining Shen; Marina Vannucci; Alexandria Baldizon; Paul M. Cinciripini; Francesco Versace; Michele Guindani

3.89. Navin Souda

Institution: UCLA

Department: Statistics and Data Science

Email: nvsouda@ucla.edu

Talk title:

Nonparametric Bayesian Biclustering of Count Data Using the Zero-Inflated Poisson SBM

Abstract:

Biclustering methods aim to group together both the observations and features of a dataset. Such methods have become of increasing interest in recent years, particularly in application genomic data such as RNA-seq, where one may try to find combinations of genes that are expressed similarly over a particular set of samples. We propose a zero-inflated Poisson stochastic block model that takes advantage of the fact that the data are sparse counts. Our method uses a Dirichlet process prior on the number of blocks in both the rows and columns to allow for these values to be determined as part of the model-fitting process, rather than having to be predetermined as in many other common procedures. Furthermore, this method allows for straightforward estimation of the posterior through a Gibbs sampler, which we demonstrate on both simulated and real data.

Keywords: biclustering; Dirichlet Process; stochastic blockmodel; community detection;

Co-authors: Arash Ali Amini, UCLA

3.90. Adam Stone

Institution: Durham University

Department: Mathematical Sciences

Email: adam.stone2@durham.ac.uk

Talk title:

AddiVortes: (Bayesian) Additive Voronoi Tessellations.

Abstract:

Many statistical models, including Random Forests, Gradient Boosted Machines, and Bayesian Additive Regression Trees (BART), rely on decision trees to partition the covariate space. While effective, these methods are often constrained by rigid, axis-aligned splits, which can limit their ability to capture nuanced relationships in the data. We introduce

Additive Voronoi Tessellations (AddiVortes), a novel Bayesian nonparametric model that uses Voronoi tessellations instead of decision trees to create more flexible and adaptive partitions. AddiVortes builds on the BART framework, incorporating regularisation priors and a Bayesian backfitting MCMC algorithm to ensure interpretability and prevent overfitting. Its performance is demonstrated through extensive comparisons on real-world datasets and simulation studies, showing that AddiVortes consistently outperforms traditional tree-based methods. By leveraging the unique properties of Voronoi tessellations, AddiVortes offers a robust and interpretable alternative for regression problems, bridging the gap between flexibility and practical usability. In this talk, I will outline the theoretical foundations of AddiVortes and demonstrate its practical advantages through simulations and applications to real-world datasets.

Keywords: Bayesian methods; Black Box algorithm; Multivariate analysis; Nonparametric regression; Voronoi tessellations.

Co-authors: John Paul Gosling, Department of Mathematical Sciences, Durham University, Durham, UK

3.91. Edric Tam

Institution: Stanford University

Email: edrictam@stanford.edu

Website: <https://www.edrictam.com>

Talk title:

On the Statistical Capacity of Deep Generative Models

Abstract:

Deep generative models are routinely used in generating samples from complex, high-dimensional distributions. Despite their apparent successes, their statistical properties are not well understood. A common assumption is that with enough training data and sufficiently large neural networks, deep generative model samples will have arbitrarily small errors in sampling from any continuous target distribution. We set up a unifying framework that debunks this belief. We demonstrate that broad classes of deep generative models, including variational autoencoders and generative adversarial networks, are not universal generators. Under the predominant case of Gaussian latent variables, these models can only generate concentrated samples that exhibit light tails. Using tools from concentration of measure and convex geometry, we give analogous results for more general log-concave and strongly log-concave latent variable distributions. We extend our results to diffusion models via a reduction argument. We use the Gromov-Levy inequality to give similar guarantees when the latent variables lie on manifolds with positive Ricci curvature. These results shed light on the limited capacity of common deep generative models to handle heavy tails. We illustrate the empirical relevance of our work with simulations and financial data.

Keywords: Diffusion Models; Isoperimetry; Concentration of measure; Generative Adversarial Networks; Variational Autoencoders

Co-authors: David Dunson

3.92. Giovanni Toto

Institution: University of Padova

Department: Department of Statistical Sciences

Email: giovanni.toto@phd.unipd.it

Talk title:

Bayesian Multivariate Density Regression with Coordinate-Wise Predictor Selection

Abstract:

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 constrained 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.

Keywords: Common atoms models; Copula; Density regression; Variable selection; Tensor factorization

Co-authors: Peter Müller, Abhra Sarkar

3.93. Vince Velkey

Institution: University College London

Department: Gatsby Computational Neuroscience Unit

Email: vince.velkey.23@ucl.ac.uk

Talk title:

Finite Exchangeability, de Finetti's Theorem and Exponential Families

Abstract:

We establish a connection between exchangeability and maximum entropy distributions by studying the mixing measure in de Finetti's theorem. The main technical tool is the convex core of a measure, originally invented to study kl-projections in information theory. Along the way, we recover a representation result for finite exchangeable sequences from a functional analytic perspective. Our method can be applied for more general symmetric structures, we discuss an abstract proof recipe. We apply the method for finite

exchangeable graphs, using results from the theory of graph limits. Using the convex core, we draw the analogous connection to exponential family random graph distributions.

Keywords: finite exchangeability; maximum entropy distributions; kl-projections; de Finetti's theorem

Co-authors: Peter Orbanz

3.94. Spencer Wadsworth

Institution: Carleton College

Department: Mathematics and Statistics

Email: swadsworth@carleton.edu

Talk title:

Bayesian Stacking via Proper Scoring Rule Optimization using a Gibbs Posterior

Abstract:

Combining probabilistic forecasts into an ensemble forecast is standard practice in collaborative forecast projects where linear pooling is one of the most common methods. A common idea is that weight selection methods should be tailored to the specific research question and often selecting weights via optimization of a proper scoring rule. Bayesian predictive synthesis has also emerged as a model probability updating scheme which provides a Bayesian solution to selecting model weights more flexible than Bayesian model averaging. Existing methods may or may not improve ensembles for given datasets, and there is room for additional methodology. In this manuscript, we introduce a Gibbs posterior on stacked model weights based on minimizing a proper scoring rule. The Gibbs posterior extends stacking into a Bayesian framework by allowing for optimal weight solutions to be influenced by a prior distribution and include uncertainty quantification. We compare ensemble forecast performance with model averaging methods and equal weighted models in simulation studies and in a real data example from the 2023-24 US Centers for Disease Control FluSight competition. In both the simulation studies and the FluSight analysis, the stacked Gibbs posterior produces ensemble forecasts which perform better than the ensembles from other methods considered.

Keywords: Optimal linear pooling · Gibbs posterior · Probabilistic forecasting · Proper scoring rules

Co-authors: Jarad Niemi

3.95. Ziyou Wang

Institution: King's College London

Department: Department of Mathematics

Email: k21179324@kcl.ac.uk

Talk title:

Spatial heterogeneity in the effects of covariates on income distributions using Bayesian nonparametric methods

Abstract:

Understanding the spatial heterogeneity of income and the effect of covariates such as gender, race and last educational attainment is a key problem for social scientists. The American Community Survey (ACS) is an annual nationwide survey designed to provide reliable economic, social, housing and demographic data to provide answers to this problem. We use individual-level personal income data which are organized in Public Use Microdata Areas (PUMAs). These are non-overlapping geographical areas containing at least 100,000 residents. Beraha and Griffin (2023) introduced Normalized Latent Measure Factor Models (NLMFM) which model a collection of distributions using linear combinations of a small number of interpretable latent factor measures. We extend their prior to allow for covariates to define a NLMFM with Covariates and introduce a novel adaptive Gibbs sampling scheme to achieve automatic inference on the number of factors. This scheme makes factors more interpretable and avoids redundant latent factors. We apply the NLMFM with Covariates to personal income and introduce gender and race in a spatial prior over the PUMAs. Rather than summaries such as the mean, our approach provides estimates of the effects of race and gender on the distribution of personal income within each PUMA. This allows us to understand the spatial heterogeneity in the effects and provides overall measures of the covariate effect through averaging over PUMAs. It disentangles their impact on the personal income distribution and provides inference on the parts of the distributions which are affected. We illustrate this on PUMA for several states and two years.

Keywords: Latent factor models; Comparing probability distributions; Normalized random measures; Adaptive MCMC

Co-authors: Maria Kalli; Jim Griffin

3.96. Hanyu Wang

Institution: University of St Andrews

Department: School of Mathematics and Statistics

Email: hw209@st-andrews.ac.uk

Talk title:

Divide-and-Conquer Dirichlet Process mixture model for Uncertainty Evaluation in Large-Scale Clustering: Data Allocation and Cluster Numbers

Abstract:

Analyzing large-scale datasets with Bayesian Nonparametric (BNP) models presents significant computational challenges. This work introduces a scalable, divide-and-conquer Dirichlet Process mixture model (DPMM) for efficient partitioning of massive datasets. Our approach employs a hierarchical, two-layer clustering strategy: first, local DPMM clustering is performed independently on data subsets (workers), followed by a global DPMM clustering that aggregates local results and performs label alignment. To minimize communication overhead and preserve cluster characteristics, we transfer local cluster centroids (as coordinate information) and principal eigenvectors (as directional information) from the local clusters (workers) to the global cluster (masters), and this process enables high-quality final clustering results. To provide a statistical property of data allocation in the presence of overlapping or complex cluster structures, we introduce a novel post-processing Markov Chain Monte Carlo (MCMC) method. This method explicitly

quantifies the uncertainty associated with data point assignments at both local and global levels and also provides an overall uncertainty for each observation. Furthermore, by leveraging Bayesian Credible Balls, we provide a probabilistic measure of the uncertainty in the number of clusters identified, enabling the detection of minor clusters that may be overlooked by traditional methods. The efficacy of the proposed method in addressing data allocation uncertainty and minor cluster detection is evaluated through a series of simulation studies.

Keywords: Dirichlet Process Mixture Model; Divide-and-Conquer; Uncertainty Qualification; Bayesian Credible

Co-authors: Dr Michail Papathomas, Dr Nicolò Margaritella

3.97. Ganchao Wei

Institution: Duke University

Department: Statistical Science

Email: ganchao.wei@duke.edu

Talk title:

Stream-level flow matching with Gaussian processes

Abstract:

Flow matching (FM) is a family of training algorithms for fitting continuous normalizing flows (CNFs). Conditional flow matching (CFM) exploits the fact that the marginal vector field of a CNF can be learned by fitting least-squares regression to the conditional vector field specified given one or both ends of the flow path. In this paper, we extend the CFM algorithm by defining conditional probability paths along "streams", instances of latent stochastic paths that connect data pairs of source and target, which are modeled with Gaussian process (GP) distributions. The unique distributional properties of GPs help preserve the "simulation-free" nature of CFM training. We show that this generalization of the CFM can effectively reduce the variance in the estimated marginal vector field at a moderate computational cost, thereby improving the quality of the generated samples under common metrics. Additionally, adopting the GP on the streams allows for flexibly linking multiple correlated training data points (e.g., time series). We empirically validate our claim through both simulations and applications to image and neural time series data.

Keywords: Generative Model, Normalizing Flows, Flow Matching, Gaussian process

Co-authors: Li Ma

3.98. Bohan Wu

Institution: Columbia University

Department: Statistics

Email: bw2766@columbia.edu

Talk title:

Adaptive Nonparametric Perturbations of Parametric Bayesian Models

Abstract:

Parametric Bayesian modeling offers a powerful and flexible toolbox for scientific data analysis. Yet the model, however detailed, may still be wrong, and this can make inferences untrustworthy. In this paper we study nonparametrically perturbed parametric (NPP) Bayesian models, in which a parametric Bayesian model is relaxed via a distortion of its likelihood. We analyze the properties of NPP models when the target of inference is the true data distribution or some functional of it, such as in causal inference. We show that NPP models can offer the robustness of nonparametric models while retaining the data efficiency of parametric models, achieving fast convergence when the parametric model is close to true. To efficiently analyze data with an NPP model, we develop a generalized Bayes procedure to approximate its posterior. We demonstrate our method by estimating causal effects of gene expression from single cell RNA sequencing data. NPP modeling offers an efficient approach to robust Bayesian inference and can be used to robustify any parametric Bayesian model.

Keywords: Generalized Bayes, Robust Bayes, Model criticism, Causal inference, Semiparametrics

Co-authors: Eli N. Weinstein, Sohrab Salehi, Yixin Wang, David M. Blei

3.99. Daniel Wurzler Barreto

Institution: The University of Texas at Austin

Department: Department of Statistics and Data Sciences

Email: daniwbarreto@utexas.edu

Talk title:

A Relaxation of Conditionally Identically Distributed Sequences for Predictive Bayesian Inference

Abstract:

Bayesian methods commonly rely on the notion of exchangeability and the de Finetti's representation theorem to justify attributing a prior distribution to a latent parameter. This way of conducting inference naturally leads to the use of Bayes' theorem as a way of updating prior to posterior beliefs, which requires computing an often high dimensional and analytically intractable integral over the parameter space. Alternatively, one can focus on directly specifying a joint distribution for the observations, bypassing the need to make subjective assessments over unobservable quantities, removing the necessity of a posterior update via Bayes' theorem to obtain predictions, and leading to less computationally costly algorithms. A further development (Fong et al, 2023) was only to require a sequence of predictives for future observations conditional on the observed. Such a strategy, for the existence of limits and hence posterior type distributions, can use conditionally identically distributed (c.i.d.) sequences (see for example Berti et al, 2021), which generalize the notion of exchangeability. This predictive Bayesian approach often yields methods that are well-suited for online learning and are highly parallelizable, making it attractive when considering modern statistical challenges and computing resources. In this context, we propose the notion of second-order c.i.d. sequences, as a mirror to the idea of second-order exchangeable sequences (see for example Goldstein & Woof (2007)), by focusing only on predictions for the first and second moments. We show that c.i.d. and second-order c.i.d.

sequences have similar properties, but the latter provides a more flexible framework, leading to simpler and more accessible predictive Bayesian methods.

Keywords: predictive Bayesian inference; Bayes linear inference; c.i.d. sequences

Co-authors: Stephen G. Walker

3.100. Yuliang Xu

Institution: Duke University

Department: Department of Statistical Science

Email: yuliang.xu@duke.edu

Website: <https://sites.google.com/view/yuliang-xu/>

Talk title:

A Hierarchical Tree Model for Multi-sample Density Estimation and Clustering

Abstract:

Motivated by the clustering problem in the DNase-seq data, we propose a novel hierarchical tree model that accounts for the within cluster variation, as well as potential long-range dependencies on different gene locations. The traditional nonparametric density learners such as Polya tree usually assign independent prior to the splitting probabilities and ignore long-range dependence where different locations in the support of the density might correlate. Our proposed correlated tree model incorporates such long-range dependence by modeling the splitting probabilities using multivariate normal through Polya-gamma representation and assigning sparse covariance prior to account for such dependence. The proposed method is demonstrated to have better clustering accuracy compared to the default method with independent variance assumption. When applied to the ENCODE project with REST transcription factor data, our method is able to identify more stable cluster patterns compared to k-means and independent variance tree methods such as hierarchical Polya tree models.

Keywords: Hierarchical Tree; Sparse Covariance; Density Estimation; Clustering;

Co-authors: Li Ma

3.101. Ryan Yee

Institution: University of Wisconsin--Madison

Department: Statistics

Email: ryee2@wisc.edu

Website: <https://ryanyee3.github.io>

Talk title:

Scalable piecewise smoothing with BART

Abstract:

BART is a powerful Bayesian non-parametric regression model that inherently produces discontinuous outputs. Initial attempts to overcome this limitation proposed ensembles of regression trees that output Gaussian Processes instead of constants. These early attempts

are unfortunately limited in their ability to smooth over many dimensions and to scale to large datasets because of their considerable computational burden. In this work, we propose ridgeBART, an extension of BART built with trees that output linear combinations of ridge functions (i.e., a composition of an affine transformation of the inputs and non-linearity); that is, we build a Bayesian ensemble of localized neural networks with a single hidden layer. We develop a new MCMC sampler that updates trees in linear time and establish nearly minimax-optimal posterior contraction rates for estimating Sobolev and piecewise anisotropic Sobolev functions. We demonstrate ridgeBART's effectiveness on synthetic data and use it to estimate the probability that a professional basketball player makes a shot from any location on the court in a spatially smooth fashion.

Keywords: Regression tree ensembles; Perceptron; Spatial smoothing; Ensemble learning; Posterior contraction

Co-authors: Soham Ghosh; Sameer K. Deshpande

3.102. MAN CHUN YEUNG

Institution: California State University, Fullerton

Department: Mathematics Department

Email: jasonyeungmanchun99@gmail.com

Talk title:

Bayesian nonparametric inference for unordered bivariate time to event data

Abstract:

Bivariate survival data arises when two time to event measurements are recorded on one individual or across a pair of individuals. Whether directly or indirectly of interest, the correlation between paired event times is essential from a modeling perspective. Often, the two event times naturally arise or can be labeled as sequentially ordered event times. The correlation between ordered event times is necessarily positive. Many modeling techniques exist for this setting in the survival and reliability literature. However, there are scenarios in which the events do not occur in a sequential fashion nor can they be labeled as such. For example, consider the time until a patient experiences therapeutic benefits along with the time to develop side effects from a medication. In this research, we propose a Bayesian nonparametric model structure for unordered bivariate event times. Specifically, we consider the popular Dirichlet Process (DP) mixture model and specify a (bivariate) folded normal kernel distribution. We offer insights into the correlation structure under this model as well as the equivalent model forms for the joint survival and mean residual life functions. We demonstrate the model's inferential potential and utility via simulated and real data examples with right censoring.

Keywords: Bivariate mean residual life; Bivariate survival analysis; Bayesian nonparametrics; Dirichlet process; unordered paired time-to-event

Co-authors: Valerie Poynor; Sam Behseta

3.103. Shijie Yuan

Institution: The University of Texas at Austin

Department: Statistics

Email: sj.yuan@austin.utexas.edu

Website: <https://scholar.google.com/citations?user=Ce-G948AAAAJ&hl=en>

Talk title:

Monitoring Adverse Events Through Bayesian Nonparametric Clustering Across Studies

Abstract:

Ensuring patient safety in clinical trials is essential due to ethical and regulatory considerations. Adverse event (AE) monitoring plays a crucial role in detecting potential safety signals and mitigating risks associated with investigational treatments. One common approach to AE evaluation involves comparing the incidence rate in an ongoing trial to a predefined background rate to determine whether a specific AE occurs more frequently than expected. However, setting an appropriate background rate is challenging due to differences in study populations (e.g., age, sex, race) between external studies and the current trial, as well as inconsistencies in data collection, where external data may be available at the patient level, covariate-specific summaries, or only as aggregate study-level statistics. To address these challenges, we introduce a Bayesian nonparametric clustering framework designed to enhance AE monitoring across multiple studies and establish a more context-specific background rate. This framework effectively incorporates external data while accommodating varying levels of covariate granularity, ranging from individual patient-level details to aggregated study-level summaries. For these different levels of covariate granularity, we define experimental units based on available covariate characteristics, which serve as the fundamental element of analysis. Our approach employs a covariate-dependent product partition model (PPM_x) to facilitate data integration across diverse external studies, increasing the probability that experimental units with similar covariates are assigned to the same cluster. Additionally, we introduce a pairwise similarity measure to quantify the relatedness between experimental units based on observed covariates. Last, based on the similarity of other units to the unit in the current trial, we construct the corresponding background rate. By leveraging Bayesian nonparametric inference, our framework is not only capable of handling missing data but also fully incorporating clinical experts' domain knowledge by allowing them to assign appropriate weights to different covariates within the pairwise similarity measure. Through application examples, we demonstrate the effectiveness of our framework in detecting safety signals under various trial scenarios. The flexibility of the model in handling different levels of data resolution makes it particularly suitable for integrating external data sources, especially when patient-level data from external studies are unavailable.

Keywords: Product partition models; Historical data integration; Covariate-dependent clustering

Co-authors: Peter Müller; Yuan Ji; Noirrit Kiran Chandra; Kevin Roberts

3.104. Percy Zhai

Institution: The University of Chicago
Department: Booth School of Business
Email: percy.zhai@chicagobooth.edu

Talk title:

Minimax Bayesian Predictive Inference with the Horseshoe Prior

Abstract:

This work is focused on providing a distributional prediction of a high-dimensional Gaussian vector with a sparse mean. The accuracy of the predictive density is measured by the Kullback-Leibler loss. Several priors have been considered in the current literature, including discrete priors and Laplace priors deployed inside the spike-and-slab framework. This work complements the toolbox by considering the Horseshoe prior, which is known to attain adaptive rate-minimax performance for the estimation problems. We show that it is also the case for the predictive inference problem. We start with the oracle case where the sparsity level is known, and demonstrate that the Horseshoe prior provides a predictive risk that attains the minimax rate with the properly calibrated parameters. We also discuss such calibration in the predictive inference settings may be different to that in the point estimation problem. In the more general case where the sparsity level is unknown, we consider the full Bayes method that imposes an uninformative prior on the Horseshoe prior's parameter. Such hierarchical priors are able to achieve adaptive rate-minimaxity. We also discuss the difference between this method and the empirical Bayes approach that plugs in an estimate of the parameter. These hierarchical priors are continuous and fully automatic (i.e. without the need to specify hyper-parameters), and are therefore easy to implement. Since the Horseshoe is a continuous mixture of Gaussian priors, under which the predictive density can be written as a continuous mixture of normal densities, the predictive inference is computationally inexpensive, a property desired by the practitioners.

Keywords: Horseshoe Prior; Predictive Inference; Sparse Normal Means; Kullback-Leibler Loss; Asymptotic Minimaxity.

Co-authors: Veronika Rockova

3.105. Zhaoxi Zhang

Institution: University of Edinburgh
Department: School of Mathematics
Email: z.zhang-156@sms.ed.ac.uk
Website: <https://zhaoxizhang.owlstown.net/>

Talk title:

Measuring dependence of partition on covariates in cluster analysis

Abstract:

Clustered structures are very common in data, particularly in areas where latent groupings underlie observed patterns. Examples include biological data, where genetic or phenotypic similarities often form natural clusters, and social science data, where behavior or preferences may group individuals into distinct categories. To uncover these structures, a

range of clustering methods have been developed. In recent years, nonparametric mixture models have become popular for their flexibility, allowing data-driven clustering without strong parametric assumptions. A major challenge in clustering is determining the role of covariates in explaining the partition observed in a response variable. This is particularly relevant when a covariate may not directly affect the response within the cluster but instead influences the likelihood of belonging to a particular cluster. Quantifying the degree to which covariates impact the clustering structure can aid in understanding underlying associations and improve the interpretability of models. To investigate the cluster structure by nonparametric mixture model, there are generally three approaches: the marginal approach, the conditional approach, and the joint approach. Initially designed to quantify separation between distributions, we propose the use of underlap coefficient to quantify the degree of dependence of the partition structure of the response, as inferred by the mixture model, on the covariates in the marginal approach and the conditional approach. We posit that the underlap coefficient can serve as an effective complement to posterior predictive checks when using mixture models for clustering purposes. While the posterior predictive check can identify model inadequacies, the underlap offers insights into where to make model adjustments, particularly whether or not to allow weights to depend on covariates. We further propose Bayesian estimators to accurately estimate the underlap coefficient for this task.

Keywords: Mixture models; Cluster analysis; Covariate dependence; Underlap coefficient

Co-authors: Vanda Inácio, University of Edinburgh; Sara Wade, University of Edinburgh

3.106. Yilei Zhang

Institution: University of Michigan--Ann Arbor

Department: Statistics

Email: yileizh@umich.edu

Talk title: A Bayesian Method for Learning Mixture Models of Non-parametric Components

Abstract:

Mixture models are widely used in modeling heterogeneous subpopulations in data. Mixture models of parametric components (e.g., Gaussian mixture models) have been thoroughly studied on both statistical and algorithmic fronts. However, in the face of increasing complexity of large scale data, parametric assumptions such as Gaussianity are often unrealistic, and very little literature has been found on learning the mixture models of non-parametric components. In an effort to fill this gap, we first address the identifiability issue in mixture models of non-parametric components. Building on this, we establish a framework using mixture of Dirichlet processes to learn such models, and develop an efficient MCMC algorithm to implement our method. Our method can learn each component density without resorting to solving the mixing measure, thus provides a sample-efficient framework for learning subpopulation properties from data. We also show the posterior contraction rate of our component density estimator is of an almost polynomial order, which is a significant improvement from the logarithm convergence rate

of solving mixing measures. This substantiates the sample-efficiency and applicability of our method in learning non-parametric component densities.

Keywords: non-parametric mixture models, Dirichlet process, component density estimation, heterogeneous data, subpopulations

Co-authors: Yun Wei, Department of Mathematical Science, University of Texas at Dallas
Aritra Guha, AT&T Data Science and AI Research XuanLong Nguyen, University of Michigan -- Ann Arbor

3.107. Xinyu Zhang

Institution: The Ohio State University

Department: Statistics

Email: zhang.11591@osu.edu

Talk title:

Bayesian Restricted Likelihood, Generalized Bayes and Model Misspecification

Abstract:

Model misspecification is problematic for Bayesians. Various methods have been proposed to modify the update from prior distribution to posterior distribution when one acknowledges that one's model is imperfect. Two current proposals are Bayesian restricted likelihood (BRL) methods and generalized Bayesian (GB) methods. The first focuses on aspects of the model that are believed to be modeled well and derives the posterior distribution by conditioning on an insufficient statistic (e.g., Huber's M-estimate) to capture those aspects of the model. The second focuses on a particular inference, making use of a loss function to define the target of inference. The usual Bayesian update is altered: the likelihood function is replaced with the exponentiated negative loss. We compare these two approaches by investigating both finite sample and asymptotic behavior when the data come from a location family. The results provide guidance on the choice of method in different cases.

Keywords: Bayesian Restricted Likelihood; Generalized Bayes; Model Misspecification; M-estimator

Co-authors: Steven MacEachern

3.108. Shuangjie Zhang

Institution: TEXAS A&M UNIVERSITY

Department: Statistics

Email: szhan209@tamu.edu

Website: <https://shuang-jie.github.io>

Talk title:

Generalized Bayes Compositional Directed Acyclic Graph

Abstract:

Bayesian network, represented as Directed Acyclic Graph (DAG), provides a structured way to model probabilistic relationships. However, classical Bayesian inference often relies on

likelihood-based inference, which can be sensitive to model misspecification and restrictive assumptions about data distributions. Motivated by complex count data from multi-community microbiome studies, we introduce the Generalized Bayes Compositional Directed Acyclic Graph (GBC-DAG), a novel framework that extends Bayesian updates by incorporating generalized Bayes inference within DAG-based probabilistic models. Our approach replaces the traditional likelihood with a Kullback–Leibler divergence based likelihood for compositional data, enabling greater flexibility and robustness in posterior inference. We prove the properness of loss-based likelihood and the identifiability of the DAG. GBC-DAG captures the interaction between microorganisms across various communities and our findings highlight the advantages of generalized Bayesian methods in microbiome compositional data analysis.

Keywords: Generalized Bayes; Directed Acyclic Graph; Compositional Microbiome Data; Kullback–Leibler Divergence

Co-authors: Bani K. Mallick, Yang Ni

3.109. Wei Zhang

Institution: Bocconi University

Email: zhngw129@gmail.com

Talk title:

A Bayesian Nonparametric Approach to Multinomial Logistic Tensor Regression with Numerous Response Categories

Abstract:

We consider the task of making statistical inferences and predictions when the data consist of a multinomial response of numerous categories and multiple categorical predictors. To perform such tasks, multinomial logistic regression has been widely applied, however, including only linear combinations of categorical predictors in the regression may fail to account for possible interactions. Therefore, we adapt a scalar-on-tensor regression framework and further decompose the coefficient tensor associated with each response category into a core tensor and a set of factor matrices, utilizing the data structure of multiple categorical predictors. Since the multinomial response has many categories, only the core tensor is assumed to be category-specific to reduce the number of parameters in the model, while the set of factor matrices is identical across categories. A Dirichlet process prior is imposed on the core tensors with a base distribution that allows atoms of varying dimensions. The Dirichlet process prior also facilitates clustering of categories so that these categories can be merged together. We illustrate the model using a Swiss COVID-19 mortality data set where we regress the main cause of death on underlying diseases, consequential diseases and companion diseases.

Keywords: multinomial logistic regression, tensor regression, Dirichlet process

Co-authors: Antonietta Mira and Ernst C. Wit

3.110. Yuan Zhong

Institution: University of Michigan

Department: Biostatistics

Email: ylzhong@umich.edu

Talk title:

Deep kernel learning based Gaussian processes for Bayesian image regression analysis

Abstract:

Regression models are widely applied in neuroimaging studies to learn associations between clinical and image variables. Gaussian process (GP) priors are common Bayesian nonparametric approaches to model unknown regression functions in high-dimensional data. However, existing GP methods depend on pre-specified parametric covariance kernels, which often lack the flexibility to capture data complexity, have limited generalizability across study populations, and face computational challenges in large-scale datasets. We propose a scalable fully Bayesian deep kernel learning framework for GP priors with applications in various image regression models. Our method leverages the estimation power of deep neural networks (DNNs) to adaptively learn the kernel basis functions from the data to capture complex spatial correlations. We establish theoretical properties by deriving posterior concentration rates of regression and kernel function estimation. Simulation studies demonstrate improved accuracy in estimation and signal detection across different scenarios. We further validate the proposed method through two neuroimaging applications: fMRI activation detection for general cognitive ability and EEG-based brain-computer interfaces.

Keywords: Deep Neural Networks; Neuroimaging; Nonparametric Regression; Spatial Correlation

Co-authors: Weining Shen, Jian Kang

3.111. Jungang Zou

Institution: Columbia University

Department: Department of Biostatistics

Email: jz3183@cumc.columbia.edu

Talk title:

Nonparametric Bayesian Additive Regression Trees for Prediction and Missing Data Imputation in Longitudinal Studies

Abstract:

Missing data is a common challenge in healthcare longitudinal studies, particularly when both covariates and outcomes are incomplete. Traditional multiple imputation methods, such as MICE and joint models, are widely used under the missing at random assumption but often suffer from uncongeniality, increased complexity with numerous covariates, and bias due to parametric model misspecification. To address these limitations, we propose a flexible Bayesian nonparametric sequential imputation framework tailored for longitudinal data. At the core of this framework is the Bayesian ensemble-tree

mixed-effects model BMTrees, and its variants, which leverage nonparametric priors to capture complex, non-linear relationships and handle non-normal random effects and errors. Our approach adapts the sequential imputation framework to the longitudinal setting, effectively modeling relationships between observed and missing variables while incorporating a fitting-with-imputing strategy to enhance computational efficiency. We establish the convergence of the MCMC algorithm within this framework. Simulation studies demonstrate that BMTrees outperforms established ensemble-tree methods, including mixedBART and mixedRF, in both prediction and imputation tasks, particularly in challenging scenarios with non-normal data structures. Through a case study with MESA data, we demonstrate the use of our sequential imputation method combined with longitudinal causal inference techniques to evaluate the comparative effectiveness of antihypertensive treatment initiation thresholds for reducing long-term systolic blood pressure. The R package SBMTrees, implementing the proposed methods, is publicly available on CRAN and GitHub.

Keywords: Bayesian non-parametric methods, longitudinal missing data, sequential imputation, mixed-effects models, multiple imputation

Co-authors: Liangyuan Hu