Abstracts

Contents

Updated: July 5, 2024

Raffaele Argiento
University of Bergamo, Italy

Covariate-Informed Model-Based Clustering

Bayesian nonparametrics is the primary probabilistic tool for defining clustering models. In particular, product partition models assign prior probability weights to each cluster using cohesion functions. Product partition models with covariates modify the probabilistic model by including a new factor in the cohesion function, i.e., the similarity function. Under this prior, the probability of experimental units with similar covariates being included in the same cluster is increased by the similarity values. Moreover, including covariate information in prior specifications enhances posterior predictive performance and facilitates the interpretation of estimated clusters.

The benefits of this modelling approach will be exemplified by a brief discussion of three real-world applications. Firstly, we will deal with the prediction of gap times between successive blood donations of the primary Italian blood supplier. Secondly, motivated by precision medicine, we cluster patients based on their genetic characteristics and treatment responses. Lastly, we address the Bergamo (Italy) province public transportation system, aiming to cluster municipalities according to their transportation network.

Back to Contents Page →

Cecilia Balocchi
University of Edinburgh, UK

Improving uncertainty quantification in Bayesian cluster analysis

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. Wade and Ghahramani (2018) proposed to summarize the posterior samples using a single optimal clustering estimate, which minimizes the expected posterior Variation of Information (VI). In instances where the posterior distribution is multimodal, it can be beneficial to summarize the posterior samples using 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 VI-based Wasserstein distance sense. An interesting byproduct is that this problem can be seen as using the k-means algorithm 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.

Back to Contents Page →

Marta Catalano
LUISS University, Italy

Merging rate of opinions via optimal transport on random measures

The Bayesian approach to inference is based on a coherent probabilistic framework that naturally leads to principled uncertainty quantification and prediction. Via posterior distributions, Bayesian nonparametric models make inference on parameters belonging to infinite-dimensional spaces, such as the space of probability distributions. The development of Bayesian nonparametrics has been triggered by the Dirichlet process, a nonparametric prior that allows one to learn the law of the observations through closed-form expressions. Still, its learning mechanism is often too simplistic and many generalizations have been proposed to increase its flexibility, a popular one being the class of normalized completely random measures. Here we investigate a simple yet fundamental matter: will a different prior actually guarantee a different learning outcome? To this end, we develop a new distance between completely random measures based on optimal transport, which provides an original framework for quantifying the similarity between posterior distributions (merging of opinions). Our findings provide neat and interpretable insights on the impact of popular Bayesian nonparametric priors, avoiding the usual restrictive assumptions on the data-generating process.

This is joint work with Hugo Lavenant.

Back to Contents Page →

Noirrit Chandra
UT Dallas, USA

Functional connectivity across the human subcortical auditory system using an autoregressive matrix-Gaussian copula graphical model approach with partial correlations

The auditory system comprises multiple subcortical brain structures that process and refine incoming acoustic signals along the primary auditory pathway. Due to technical limitations of imaging small structures deep inside the brain, most of our knowledge of the subcortical auditory system is based on research in animal models using invasive methodologies. Advances in ultra-high field functional magnetic resonance imaging (fMRI) acquisition have enabled novel non-invasive investigations of the human auditory subcortex, including fundamental features of auditory representation such as tonotopy and periodotopy. However, functional connectivity across subcortical networks is still underexplored in humans, with ongoing development of related methods. Traditionally, functional connectivity is estimated from fMRI data with full correlation matrices. However, partial correlations reveal the relationship between two regions after removing the effects of all other regions, reflecting more direct connectivity. While most existing methods for learning conditional dependency structures based on partial correlations assume independently and identically Gaussian distributed data, fMRI data exhibit significant deviations from Gaussianity as well as high temporal autocorrelation. In this paper, we developed an autoregressive matrix-Gaussian copula graphical model approach to estimate the partial correlations and thereby infer the functional connectivity patterns within the auditory system while appropriately accounting for autocorrelations between successive fMRI scans. Our results are highly stable when splitting the data in halves according to the acquisition schemes and computing partial correlations separately for each half of the data, as well as across cross-validation folds. In contrast, full correlation-based analysis identified a rich network of interconnectivity that was not specific to adjacent nodes along the pathway. Overall, our results demonstrate that unique functional connectivity patterns along the auditory pathway are recoverable using novel connectivity approaches and that our connectivity methods are reliable across multiple acquisitions.

Back to Contents Page →

Yunshan Duan
UT Austin, USA

Immune Profiling among Colorectal Cancer Subtypes using Dependent Mixture Models

Comparison of transcriptomic data across different conditions is of interest in many biomedical studies. In this paper, we consider comparative immune cell profiling for early-onset (EO) versus late-onset (LO) colorectal cancer (CRC). EOCRC, diagnosed between ages 18-45, is a rising public health concern that needs to be urgently addressed. However, its etiology remains poorly understood. We work towards filling this gap by identifying homogeneous T cell subpopulations that show significantly distinct characteristics across the two tumor types, and identifying others that are shared between EOCRC and LOCRC. We develop dependent finite mixture models where immune subtypes enriched under a specific condition are characterized by terms in the mixture model with common atoms but distinct weights across conditions, whereas common subtypes are characterized by sharing both atoms and relative weights. The proposed model facilitates the desired comparison across conditions by introducing highly structured multi-layer Dirichlet priors. We illustrate inference with simulation studies and data examples. Results identify EO- and LO-enriched T cells subtypes whose biomarkers are found to be linked to mechanisms of tumor progression. The findings reveal distinct characteristics of the immune profiles in EOCRC and LOCRC, and potentially motivate insights into treatment of CRC.

Back to Contents Page →

Subhashsis Ghosal
North Carolina State University, USA

Bayesian Inference for High-dimensional Time Series by Latent Process Modeling

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). Key structural properties of time series such as stationarity and causality can be easily accommodated in the OUT model. 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. In the course of establishing the result, we develop a new general theorem on contraction rate of a pseudo-posterior distribution that is potentially applicable in other situations. Through a simulation study, we compare the accuracy of estimating the parameters and identifying the graphical structure with other approaches. We apply the proposed methodology to analyze a dataset on different industrial components of the US gross domestic product between 2010 and 2019 and predict future observations.

Based on a collaboration with Arkaprava Roy (University of Florida) and Anindya Roy (University of Maryland-Baltimore County).

Back to Contents Page →

Alessandra Guglielmi
Politecnico di Milano, Italy

Bayesian clustering of high-dimensional data via latent repulsive mixtures

Model-based clustering of moderate or large dimensional data is notoriously difficult. We propose a model for simultaneous dimensionality reduction and clustering by assuming a mixture model for a set of latent scores, which are then linked to the observations via a Gaussian latent factor model. This approach was recently investigated by Chandra and coauthors. They use a factor-analytic representation and assume a mixture model for the latent factors. However, performance can deteriorate in the presence of model misspecification. Assuming a repulsive point process prior for the component-specific means of the mixture for the latent scores is shown to yield a more robust model that outperforms the standard mixture model for the latent factors in several simulated scenarios. To favor well-separated clusters of data, the repulsive point process must be anisotropic, and its density should be tractable for efficient posterior inference. We address these issues by proposing a general construction for anisotropic determinantal point processes. We apply the new model to simulated examples and to an application to the occurrence of plant species at different sites of the Bauges Natural Regional Park, France.

Joint work with Lorenzo Ghilotti (University of Milano-Bicocca, Italy) and Mario Beraha (Politecnico di Milano, Italy).

Back to Contents Page →

Nhat Ho
UT Austin, USA

Bayesian Nonparametrics Meets Data-Driven Robust Optimization

Training machine learning and statistical models often involves optimizing a data-driven risk criterion. The risk is usually computed with respect to the empirical data distribution, but this may result in poor and unstable out-of-sample performance due to distributional uncertainty. In the spirit of distributionally robust optimization, we propose a novel robust criterion by combining insights from Bayesian nonparametric (i.e., Dirichlet Process) theory and recent decision-theoretic models of smooth ambiguity-averse preferences. First, we highlight novel connections with standard regularized empirical risk minimization techniques, among which Ridge and LASSO regressions. Then, we theoretically demonstrate the existence of favorable finite-sample and asymptotic statistical guarantees on the performance of the robust optimization procedure. For practical implementation, we propose and study tractable approximations of the criterion based on well-known Dirichlet Process representations. We also show that the smoothness of the criterion naturally leads to standard gradient-based numerical optimization. Finally, we provide insights into the workings of our method by applying it to high-dimensional sparse linear regression and robust location parameter estimation tasks. This is based on the joint work with Nicola Bariletto.

Back to Contents Page →

Akira Horiguchi
Duke University, USA

A tree perspective on stick-breaking models in covariate-dependent mixtures

Stick-breaking (SB) processes are often adopted in Bayesian mixture models for generating mixing weights. When covariates influence the sizes of clusters, SB mixtures are particularly convenient as they can leverage their connection to binary regression to ease both the specification of covariate effects and posterior computation. Existing SB models are typically constructed based on continually breaking a single remaining piece of the unit stick. We view this from a dyadic tree perspective in terms of a lopsided bifurcating tree that extends only in one side. We show that two unsavory characteristics of SB models are in fact largely due to this lopsided tree structure. We consider a generalized class of SB models with alternative bifurcating tree structures and examine the influence of the underlying tree topology on the resulting Bayesian analysis in terms of prior assumptions, posterior uncertainty, and computational effectiveness. In particular, we provide evidence that a balanced tree topology, which corresponds to continually breaking all remaining pieces of the unit stick, can resolve or mitigate these undesirable properties of SB models that rely on a lopsided tree.

Back to Contents Page →

Alejandro Jara

Pontificia Universidad Catolica de Chile, Chile

Bayesian Copula Density Estimation Using Bernstein Yett-Uniform Prior

Probability density estimation is a central task in statistics. Copula-based models provide a great deal of flexibility in modelling multivariate distributions, allowing for the specifications of models for the marginal distributions separately from the dependence structure (copula) that links them to form a joint distribution. Choosing a class of copula models is not a trivial task and its misspecification can lead to wrong conclusions. We introduce a novel class of random Bernstein copula functions, and studied its support and the behavior of its posterior distribution. The proposal is based on a particular class of random grid-uniform copulas, referred to as Yett-uniform copulas. Alternative Markov chain Monte Carlo algorithms for exploring the posterior distribution under the proposed model are also studied. The methodology is illustrated by means of simulated and real data.

Back to Contents Page →

Yuan Ji
University of Chicago, USA

Plaid Atoms Model with Application to Clinical Trials

We propose the Plaid Atoms Model (PAM), a Bayesian nonparametric model for grouped data based on `atom skipping'. Atom skipping refers to stochastically assigning 0 weights to atoms in an infinite mixture. Deploying atom skipping across groups, PAM produces a dependent clustering pattern with overlapping and non-overlapping clusters across groups. As a result, interpretable posterior inference is possible such as reporting the posterior probability of a cluster being exclusive to a single group or shared among a subset of groups. We discuss the theoretical properties of the proposed and related models. In clinical trials, it is highly desirable to borrow information from external data to augment a control arm in a randomized clinical trial, especially in settings where the sample size for the control arm is limited. We apply PAM to identify overlapping and unique subpopulations across datasets, with which we restrict the information borrowing to the common subpopulations. This forms a hybrid control (HC) that leads to more precise estimation of treatment effects. Simulation studies demonstrate the robustness of the new method, and an application to an Atopic Dermatitis dataset shows improved treatment effect estimation.

Joint work with Dehua Bi.

Back to Contents Page →

Jaeyong Lee
Seoul National University, Korea

Constrained Dirichlet Processes and Moment Condition Models

In this talk, we consider the constrained Dirichlet process (cDP), the conditional distribution of the Dirichlet process when a functional of a random distribution is given. We specifically apply the cDP to the moment condition model. This model is a nonparametric model in which the finite dimensional parameter of interest is defined as a solution to a functional equation of the distribution. We derive both the posterior distribution of the parameter of interest and that of the underlying distribution itself. We investigate the properties of the moment condition model with cDP and propose an algorithm for the posterior inference.

Back to Contents Page →

Antonio Lijoi
Bocconi University, USA

New insights on hierarchical random measures

Models for heterogeneous, or grouped, data rely on latent random structures that account for dependence across different groups. This is common in hierarchical nonparametric models, where the prior is specified as a composition of discrete random probability measures. Here we illustrate a new strategy that leads to closed-form expressions for the marginal and posterior distributions. It is based on two main tools: (i) a set of latent variables that can be seen as marks associated to the atoms of the random measure; (ii) an identity for the moment measure associated to a suitable exponentially tilted random measure. Illustrations with hierarchical priors for hazard rate functions and covariate-dependent distributions are discussed.

Back to Contents Page →

Steven McEachern
Ohio State University, USA

The posterior distribution as a computational object

Typical parametric models are relatively simple, being governed by a handful of parameters. Summaries in terms of moments and probabilities are often available in closed form. In contrast, high dimensional parametric models are inherently more complex. Substantial computation may be needed to extract meaningful summaries, many of which may not be available in closed form. Nonparametric models involve an infinite number of parameters, further racheting up the complexity.

For Bayesians, Markov chain Monte Carlo methods opened the door to models that move well beyond conjugacy, including BNP. The methods produce a discrete approximation to the posterior distribution that is then used for a variety of purposes. We find value in viewing the posterior distribution as a computational object that can be probed to yield various summaries. In turn, this suggests that we focus on creating structures that make these probes as useful as possible. Speed, accuracy, and the choice of effective probes are paramount.

Back to Contents Page →

Ramses Mena
IIMAS-UNAM, Mexico

From multivariate Bernoulli observations to Bayesian nonparametrics: A clinical footprint application

The COVID-19 pandemic has prompted the development of various statistical models aimed at understanding the intricate relationships among variables such as comorbidities, symptoms, hospitalizations, and deaths among millions of patients. In response, we employed a straightforward yet robust Bayesian methodology to uncover significant insights within such datasets.

Our approach involved encoding these variables into multivariate binary variables, termed as Clinical Footprints, which effectively captured all potential relationships among tagged comorbidities, symptoms, and other factors. In this presentation, we will outline this methodology and delve into its Bayesian nonparametric extension, which expands beyond binary or tagged variables, allowing for a more comprehensive analysis.

Back to Contents Page →

Peter Müller
UT Austin, USA

DP-SPGLM: semiparametric Bayesian inference for a GLM using inhomogeneous normalized random measures

We introduce an instance of a varying weight dependent Dirichlet process (DDP) model to implement a semi-parametric GLM. The model extends the recently developed semi-parametric generalized linear model (SPGLM) by adding a nonparametric Bayesian prior on the centering distribution of the GLM. We show that the resulting model takes the form of an inhomogeneous normalized random measure that arises from exponential tilting of a normalized random measure. Building on familiar posterior simulation methods for mixtures with respect to normalized random measures we introduce modification to implement posterior simulation in the resulting semi-parametric GLM model.

Back to Contents Page →

Garritt Page
Brigham Young University, USA

Informed Random Partition Models with Temporal Dependence

Model-based clustering is a powerful tool that is often used to discover hidden structure in data by grouping observational units that exhibit similar response values. Recently, clustering methods have been developed that permit incorporating an ``initial'' partition informed by expert opinion. Then, using some similarity criteria, partitions different from the initial one are down weighted, i.e. they are assigned reduced probabilities. These methods represent an exciting new direction of method development in clustering techniques. We add to this literature a method that very flexibly permits assigning varying levels of uncertainty to any subset of the partition. This is particularly useful in practice as there is rarely clear prior information with regards to the entire partition. Our approach is not based on partition penalties but considers individual allocation probabilities for each unit (e.g., locally weighted prior information). We illustrate the gains in prior specification flexibility via simulation studies and an application to a dataset concerning spatio-temporal evolution of PM10 measurements in Germany.

Back to Contents Page →

Fernando Quintana
Pontificia Universidad Catolica de Chile, Chile

A change-point random partition model for large spatio-temporal datasets

Spatio-temporal areal data can be seen as a collection of time series which are spatially correlated, according to a specific neighboring structure. Motivated by a dataset on mobile phone usage in the Metropolitan area of Milan, Italy, we propose a semi-parametric hierarchical Bayesian model allowing for time-varying as well as spatial model-based clustering. To accommodate for changing patterns over work hours and weekdays/weekends, we incorporate a temporal change-point component that allows the specification of different hierarchical structures across time points. The model features a random partition prior that incorporates the desired spatial features and encourages co-clustering based on areal proximity. We explore properties of the model by way of extensive simulation studies from which we collect valuable information. Finally, we discuss the application to the motivating data, where the main goal is to spatially cluster population patterns of mobile phone usage.

Back to Contents Page →

Lorenzo Trippa
Harvard University, USA

Bayesian Combinatorial Multi-Study Factor Analysis

Analyzing multiple studies allows leveraging data from a range of sources and populations, but until recently, there have been limited methodologies to approach the joint unsupervised analysis of multiple highdimensional studies. A recent method, Bayesian Multi-Study Factor Analysis (BMSFA), identifies latent factors common to all studies, as well as latent factors specific to individual studies. However, BMSFA does not allow for partially shared factors, i.e. latent factors shared by more than one but less than all studies. We extend BMSFA by introducing a new method, Tetris, for Bayesian combinatorial multi-study factor analysis, which identifies latent factors that can be shared by any combination of studies. We model the subsets of studies that share latent factors with an Indian Buffet Process. We test our method with an extensive range of simulations, and showcase its utility not only in dimension reduction but also in covariance estimation. Finally, we apply Tetris to highdimensional gene expression datasets to identify patterns in breast cancer gene expression, both within and across known classes defined by germline mutations.

Joint work with Isabella N. Grabski, Roberta De Vito and Giovanni Parmigiani.

Back to Contents Page →

Sara Wade
University of Edinburgh, UK

Investigating neuronal connectivity patterns from MAPseq data: a hierarchical Bayesian mixture approach

We develop a semiparametric Bayesian approach to missing outcome data in longitudinal studies in the presence of auxiliary covariates. We consider a joint model for the full data response, missingness, and auxiliary covariates. We include auxiliary covariates to “move” the missingness “closer” to missing at random. In particular, we specify a semiparametric Bayesian model for the observed data via Gaussian process priors and Bayesian additive regression trees. These model specifications allow us to capture nonlinear and nonadditive effects, in contrast to existing parametric methods. We then separately specify the conditional distribution of the missing data response given the observed data response, missingness, and auxiliary covariates (i.e., the extrapolation distribution) using identifying restrictions. We introduce meaningful sensitivity parameters that allow for a simple sensitivity analysis. Informative priors on those sensitivity parameters can be elicited from subject-matter experts. We use Monte Carlo integration to compute the full data estimands. Performance of our approach is assessed using simulated datasets. Our methodology is motivated by, and applied to, data from a clinical trial on treatments for schizophrenia.

Back to Contents Page →

Junyi Zhang
Hong Kong Polytechnic University

Posterior Sampling from Truncated Inverse-Lévy Measure Representation of NRMI Mixtures

In this talk, we discuss the finite approximation of the normalised random measure with independent increments (NRMI) by truncating its inverse-Lévy measure representation. The approximation is obtained by keeping the N largest atom weights of the underlying completely random measure (CRM) unchanged and combining the smaller atom weights into a single term. We develop the simulation algorithms for the approximation and characterise its posterior distribution, for which a blocked Gibbs sampler is devised. We demonstrate the usage of the approximation in the Bayesian nonparametric mixture model and the Caron-Fox network model. Numerical implementations are given based on the gamma, stable and generalised gamma processes.

Back to Contents Page →

Tianjian Zhou
Colorado State University, USA

A Semiparametric Bayesian Approach to Dropout in Longitudinal Studies with Auxiliary Covariates

We develop a semiparametric Bayesian approach to missing outcome data in longitudinal studies in the presence of auxiliary covariates. We consider a joint model for the full data response, missingness, and auxiliary covariates. We include auxiliary covariates to "move" the missingness "closer" to missing at random. In particular, we specify a semiparametric Bayesian model for the observed data via Gaussian process priors and Bayesian additive regression trees. These model specifications allow us to capture nonlinear and nonadditive effects, in contrast to existing parametric methods. We then separately specify the conditional distribution of the missing data response given the observed data response, missingness, and auxiliary covariates (i.e., the extrapolation distribution) using identifying restrictions. We introduce meaningful sensitivity parameters that allow for a simple sensitivity analysis. Informative priors on those sensitivity parameters can be elicited from subject-matter experts. We use Monte Carlo integration to compute the full data estimands. Performance of our approach is assessed using simulated datasets. Our methodology is motivated by, and applied to, data from a clinical trial on treatments for schizophrenia.

Back to Contents Page →