Abstracts

Contents

Tutorial Speakers

Abstracts

Contents

Speakers

Kerrie Mengersen
Queensland University of Technology, Australia

Tutorial on Bayesian nonparametrics in practice

This tutorial will focus on a number of applications of Bayesian nonparametric modelling to substantive issues in environment, sport, health and society. These projects have all been undertaken in collaboration with organisations that want to use the answers for decision-making. The Australian Antarctic Division asked: how large a boundary should we put around an oil spill based on soil contamination? The Queensland Academy of Sport asked: can we predict injuries so our players can win well? The Australian Government called for support and strategies to manage covid. The Red Cross cares about understanding and predicting natiomal and global terrorism events. In answering these questions, we will discuss a variety of models and computational algorithms. We will also have hands-on activities and invite discussion about improvements and alternativies, implementation and translation.

The research discussed in the tutorial was led by Julyan Arbel, Raiha Browning, John Worrall and Judith Rousseau.

Back to Contents Page →

Li Ma
Duke University, USA

Tutorial on Generative modeling and nonparametric inference with trees and recursive partitions

---

Trees and recursive partitions are most well-known in supervised learning for predictive tasks, such as regression and classification. Famous examples include CART, random forest, and boosting and their Bayesian cousins such as Bayesian CART and BART. A natural question is whether such successes can be replicated in the context of unsupervised problems and modeling unlabeled data. In this short course, I will first survey some classical Bayesian generative models and nonparametric priors based on trees and partitions, followed by several more recent examples of tree-based approaches for unsupervised learning and generative modeling, where the two primary objectives are to (i) learn the underlying nature of complex multivariate, possibly high-dimensional distributions based on unlabeled i.i.d. training data, and (ii) generate new data samples from the trained model. In these examples, the employment of trees and partitions leads to highly efficient, statistically rigorous inference algorithms that scale approximately linearly in the sample size and accommodate moderately high (e.g., hundreds) dimensions. Some examples from biomedical applications such as microbiome compositional analysis will be provided.

Back to Contents Page →

Peter Orbanz
University College London, UK

Tutorial on Exchangeability and symmetry in statistics and machine learning

---

The fundamental theorem of Bayesian statistics is a symmetry result: de Finetti's theorem characterizes distributions that are invariant under permutations. Invariance under a class of transformations is just how physicists and mathematicians define symmetry. In the last few years, symmetry is suddenly being discovered and rediscovered everywhere: In generative modeling, in data augmentation, in conformal prediction, in machine learning for science, and so forth. I will try to explain in the simplest possible terms that these ideas are all related mathematically; that exchangeability implies not just the de Finetti representation, but also the law of large numbers and the central limit theorem; that this is not only true for exchangeability, but for a much larger class of symmetries; and how all of this relates to work in machine learning that studies symmetries of functions rather than distributions. I will also sketch some open problems. There is a lot of work to be done.

Back to Contents Page →

Pierre Alquier
ESSEC Business School, UK

PAC-Bayesian Bounds for Offline Contextual Bandits

The PAC-Bayesian theory provides tools to understand the accuracy of Bayes-inspired algorithms learning that learn probability distributions on parameters. In this talk, we will apply these tools to offline policy learning in contextual bandits. After a short introduction to PAC-Bayes bounds, I will explain why they are convenient in this context: a policy is a probability distribution on decisions. We can thus learn a policy by minimizing a PAC-Bayes bound. We prove that the derived bounds are tighter than previous guarantees for bandits, and can be optimized directly to confidently improve upon the logging policy offline. Our approach learns policies with guarantees, uses all available data and does not require tuning additional hyperparameters on held-out sets. We demonstrate through extensive experiments the effectiveness of our approach in providing performance guarantees in practical scenarios.

Joint work with Otmane Sakhi and Nicolas Chopin.

Back to Contents Page →

Minwoo Chae
Pohang University of Science and Technology, Korea

Online Bernstein-von Mises Theorem

Suppose that mini-batches from the entire dataset become available in sequential order. The Bayesian paradigm, which updates the degree of belief after each observation of a mini-batch, is ideally suited for this context. At each step, we update the posterior distribution based on the observations from the mini- batch and then approximate it with a normal distribution. This approximation is justified by the celebrated Bernstein-von Mises theorem. The approximated normal distribution is subsequently used as the new prior for the next step. The accuracy of the normal approximation at each step is crucial, as it is expected that the updated normal distribution at the final step will closely approximate the full posterior distribution. Additionally, the size of each mini-batch and the number of mini-batches play important roles in controlling the approximation errors. This talk will present some rigorous arguments to support this online updating approach.

Back to Contents Page →

Noirrit Chandra
University of Texas at Dallas, USA

Inferring Covariance Structure from Multiple Data Sources via Subspace Factor Analysis

---

Factor analysis provides a canonical framework for imposing lower-dimensional structure such as sparse covariance in high-dimensional data. High-dimensional data on the same set of variables are often collected under different conditions, for instance in reproducing studies across research groups. In such cases, it is natural to seek to learn the shared versus condition-specific structure. Existing hierarchical extensions of factor analysis have been proposed, but face practical issues including identifiability problems. To address these shortcomings, we propose a class of SUbspace Factor Analysis (SUFA) models, which characterize variation across groups at the level of a lower-dimensional subspace. We prove that the proposed class of SUFA models lead to identifiability of the shared versus group-specific components of the covariance, and study their posterior contraction properties. Taking a Bayesian approach, these contributions are developed alongside efficient posterior computation algorithms. Our sampler fully integrates out latent variables, is easily parallelizable and has complexity that does not depend on sample size. We illustrate the methods through application to integration of multiple gene expression datasets relevant to immunology.

Back to Contents Page →

Alessandro Colombi
University of Milano-Bicocca, Italy

Discovering species richness in multiple populations

In ecology, a classic problem is comparing two populations sampled from different sites or from the same area at different times, e.g., before and after pollution. Such a comparison relies on measures of similarity or dissimilarity, which are always functions of the number of shared species between the two populations. However, if unobserved species are present, relying solely on the observed number of shared species introduces a substantial bias which must be corrected by estimating the number of unobserved shared species. To address this issue, we assume partially exchangeable observations and we formulate a statistical model based on a vector of random probability, with atoms shared across the two populations. Incorporating a prior on the number of species and adopting a Dirichlet distribution for group-specific weights, we are able to predict both the number of new species in additional samples and the number of new shared species. Notably, our work's mathematical contribution lies in providing closed-form expressions for these distributions.

Back to Contents Page →

Claudia Contardi
University of Pavia, Italy

Poisson credible intervals in Bayesian nonparametric estimation of the unseen

We investigate the topic of uncertainty quantification in the Bayesian nonparametric (BNP) approach to the estimation of the number of unseen species, K. This is a classical problem in statistics, with applications in various fields.

Under a Pitman-Yor (PY) prior, the posterior distribution of K, and hence the posterior mean as the natural BNP estimator, are available in closed form. However, the direct construction of credible intervals is infeasible, due to the computational complexity of evaluating the posterior. Existing literature addresses this problem via sampling from the limiting posterior distribution in the non-degenerate PY case, and relies on a Gaussian central limit theorem in the Dirichlet case.

We propose a novel analytical approach to the problem, which allows for a unified treatment over all the range of PY priors and under different asymptotical regimes. Our method relies on a compound Poisson approximation of the distribution of K for a large number of new observations. We show that our distribution asymptotically approximates the variance of the true distribution up to a higher order when compared to the limiting posterior. This motivates the use of the Poisson approximation for the construction of credible intervals with given coverage in asymptotical regimes.

Back to Contents Page →

Claudio Del Sole
Università Bocconi, Italy

Hierarchically dependent mixture hazard rates for modelling competing risks

A popular approach in Bayesian modelling of non-exchangeable data relies on the specification of hierarchical nonparametric priors, which induce dependence across groups of observations. In the analysis of grouped survival data, subject to a single disease or failure, hierarchies of completely random measures have been used as mixing measures to model multivariate dependent mixture hazard rates. This modelling approach can be recast to tackle a competing risks scenario, in which groups correspond to different diseases or causes of failures affecting each subject: in this setting, the multivariate construction acts at a latent level, as only the minimum time-to-event and the corresponding cause of death or failure are actually observed. The posterior hierarchy of random measures, as well as the posterior estimates of both survival function and cause-specific incidence functions are explicitly determined, conditionally on a suitable latent partition that admits a characterization in terms of a novel variant of the Chinese restaurant franchise process. These results are pivotal for devising marginal and conditional sampling algorithms. The performances of our proposal are compared with those of its non-hierarchical counterpart, which models hazard rates independently for each disease, in order to assess the benefits of information borrowing across groups. Some applications to clinical datasets are also discussed.

Back to Contents Page →

Yunshan Duan
University of Texas at Austin, USA

Spatial transcriptomics profiling of the tumor microenvironment with dependent random partitions

---

Young-onset colorectal cancer (YOCRC), diagnosed at ages 18-50 years, is an increasingly urgent global health problem that remains poorly understood. Colorectal cancer has high intra-tumor heterogeneity, especially in the tissue architecture. We propose to develop state-of-the-art single-cell and spatial molecular profiling strategies to investigate the tumor microenvironment (TME) of YOCRC, filling a gap in current literature. We will develop model-based inference for spatial partitioning of samples using spatial transcriptomics and unmatched single-cell data for YOCRC and late-onset colorectal cancer (LOCRC) for comparison. Related existing methods for the analysis of spatial transcriptomics data require matched single-cell data from the same tissue, which are costly to obtain. Also, spot-level inference under existing methods cannot provide inference on cell-level behavior and immune profiles. The innovative features of the proposed work are (i) simulation-based approximate Bayesian computation to implement a deconvolution of spatial transcriptomics data into cell-level data; (ii) spatial random partition of imputed single cells with preference for spatially connected clusters; and (iii) dependent random partition of immune and non-immune cells. The proposed analysis aims to investigate spatial segmentation and cell interactions. Downstream tasks, including differential gene expression and immune lineage analysis, will be conducted. 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.

Back to Contents Page →

Edwin Fong
University of Hong Kong, Hong Kong

Bayesian Nonparametric Inference with the Martingale Posterior

---

While the prior distribution is the usual starting point for Bayesian uncertainty, recent work has reframed Bayesian inference as the predictive imputation of missing observations. In particular, the martingale posterior distribution arises when the Bayesian model is a chosen sequence of predictive distributions on future observables, which then induces a posterior distribution on the parameter of interest without the need for a likelihood and prior. This generalization greatly broadens the range of nonparametric models one can use for Bayesian inference, and offers substantial advantages in computation and flexibility, while posing new challenges. In this talk, we introduce the framework and present some recent advances.

Back to Contents Page →

Shufei Ge
ShanghaiTech University, China

Poisson Hyperplane Processes with Rectified Linear Units

Neural networks have shown state-of-the-art performances in various classification and regression tasks. Rectified linear units (ReLU) are often used as activation functions for the hidden layers in a neural network model. In this work, we establish the connection between the Poisson hyperplane processes (PHP) and two-layer ReLU neural networks. We show that the PHP with a Gaussian prior is an alternative probabilistic representation to a two-layer ReLU neural network. In addition, we show that a two-layer neural network constructed by PHP is scalable to large-scale problems via the decomposition propositions. Finally, we propose an annealed sequential Monte Carlo algorithm for Bayesian inference. Our numerical experiments demonstrate that our proposed method outperforms the classic two-layer ReLU neural network.

Back to Contents Page →

Lorenzo Ghilotti
Università degli Studi di Milano, Italy

Bayesian analysis of product-form feature allocation models

Species sampling models represent a large class of Bayesian nonparametric priors tailored for a population of animals, where each animal belongs to a single species. The random partition induced by a sample of animals is characterized by the Exchangeable Partition Probability Function. Feature allocation models constitute a primary extension of the species framework, where subjects can display multiple features recorded by binary variables. Feature allocations, analogous to clustering, are described by the Exchangeable Feature Probability Function (EFPF). The aim of this work is to provide distribution results for a fundamental class of feature allocation models with EFPFs in product-form, which have been recently investigated from a probabilistic perspective. These models serve as prominent priors in the feature setting, akin to Gibbs-type priors in the species framework, offering a balance between tractability and flexibility. We develop a general theory, analysing the predictive structure, marginal distribution, and posterior distribution of the underlying statistical process. Noteworthy examples, such as mixtures of the Indian Buffet Process and Beta-Bernoulli models, are examined. The methodology has significant applications in ecology, addressing species richness estimation using the accumulation curve, and in genomics, dealing with extrapolation problems for estimating the number of unseen genetic variants.

Back to Contents Page →

María Fernanda Gil Leyva Villa
Universidad Nacional Autónoma de México, Mexico

Markov stick-breaking processes

Stick-breaking has a long history and represents the most popular procedure for constructing random discrete distributions in Statistics and Machine Learning. In particular, due to their intuitive construction and computational tractability they are ubiquitous in modern Bayesian nonparametric inference. Most widely used models, such as the Dirichlet and the Pitman-Yor processes, rely on iid or independent length variables. Here we pursue a completely unexplored research direction by considering Markov length variables and investigate the corresponding general class of stick breaking processes, which we term Markov stick-breaking processes. We derive conditions ensuring that the associated species sampling process is proper and that the distribution of a Markov stick-breaking process has full topological support, two fundamental desiderata for Bayesian nonparametric models. We also analyze the stochastic ordering of the weights and provide a new characterization of the Pitman-Yor process as the only stick-breaking process invariant under size-biased permutations, under mild conditions. Moreover, we identify two notable subclasses of Markov stick-breaking processes, that enjoy appealing properties and include Dirichlet, Pitman-Yor and Geometric priors as special cases. Our findings further include distributional results on the latent allocation variables and a posterior characterization of the length variables: besides their theoretical interest, these allow us to develop computational algorithms for posterior inference. Interesting methodological implications are drawn from numerical implementations of Markov stick-breaking mixture models.

Back to Contents Page →

Clara Grazian
University of Sydney, Australia

A Bayesian Semiparametric Approach for Long Memory Analysis

Dirichlet processes and their extensions have gained significant popularity in Bayesian nonparametric statistics. They have also been applied to spatial and spatio-temporal data analysis, serving as a valuable tool for surface analysis and prediction. A prevalent approach in modeling Dirichlet processes in a spatial context involves a stick-breaking representation of the process, where the spatial dependence is defined within the stick-breaking probabilities. However, extensions to include temporal dependence remain somewhat limited. This is especially crucial for phenomena that exhibit rapid changes over time and space, with numerous localized fluctuations.

In this talk, I will delve into a Dirichlet process where the stick-breaking probabilities are designed to incorporate both spatial and temporal dependencies. I will demonstrate that this approach is not merely a straightforward extension of existing methodologies but can, in fact, surpass available approaches in terms of prediction accuracy. The result is a predictive model that doesn't rely on the assumptions of Gaussianity or separability of time and space.

Back to Contents Page →

Kapil Gupta
Indian Institute of Management Bangalore, India

A divide-and-conquer approach for spatio-temporal analysis of large house price data from Greater London

Statistical research in real estate markets, particularly understanding spatio-temporal dynamics of house prices, has garnered attention in recent times. Albeit Bayesian methods are common in spatio-temporal modeling, standard Markov chain Monte Carlo (MCMC) techniques are usually slow for large datasets such as the house price data. We propose a divide-and-conquer spatio-temporal modeling approach to tackle this problem. The method involves partitioning the data into multiple subsets and utilizing an appropriate Gaussian process model for each subset in parallel. The results from each subset are then combined via the Wasserstein barycenter technique to obtain the global parameters for the original problem. The divide-and-conquer approach allows multiple observations per spatial and time unit, thereby offering added benefit for practitioners. As a real life application, we analyze house price data of 0.65 million properties from 983 middle layer super output areas in London for a period of eight years. The methodology renders insightful findings about the effects of various amenities, trend pattern, and relationship of price to carbon emission. Further, as demonstrated from a cross-validation study, it records good predictive accuracy while balancing the computational need, and is proved to be more effective than a conventional Bayesian approach.

Back to Contents Page →

Akira Horiguchi
Duke University, USA

Sampling depth trade-off in function estimation under a two-level design

Many modern statistical applications involve a two-level sampling scheme that first samples subjects from a population and then samples observations on each subject. These schemes often are designed to learn both the population-level functional structures shared by the subjects and the functional characteristics specific to individual subjects. Common wisdom suggests that learning population-level structures benefits from sampling more subjects whereas learning subject-specific structures benefits from deeper sampling within each subject. Oftentimes these two objectives compete for limited sampling resources, which raises the question of how to optimally sample at the two levels. We quantify such sampling-depth trade-offs by establishing the L2 minimax risk rates for learning the population-level and subject-specific structures under a hierarchical Gaussian process model framework where we consider a Bayesian and a frequentist perspective on the unknown population-level structure. These rates provide general lessons for designing two-level sampling schemes given a fixed sampling budget. Interestingly, they show that subject-specific learning occasionally benefits more by sampling more subjects than by deeper within-subject sampling. We show that the corresponding minimax rates can be readily achieved in practice through simple adaptive estimators without assuming prior knowledge on the underlying variability at the two sampling levels. We validate our theory and illustrate the sampling trade-off in practice through both simulation experiments and two real datasets. While we carry out all the theoretical analysis in the context of Gaussian process models for analytical tractability, the results provide insights on effective two-level sampling designs more broadly.

Back to Contents Page →

Gregor Kastner
University of Klagenfurt, Austria

Bayesian Machine Learning meets Formal Methods: An application to spatio-temporal data

---

We propose an interdisciplinary framework that combines Bayesian predictive inference, a well-established tool in Machine Learning, with Formal Methods rooted in the computer science community. Bayesian predictive inference allows for coherently incorporating uncertainty about unknown quantities by making use of methods or models that produce predictive distributions, which in turn inform decision problems. By formalizing these decision problems into properties with the help of spatio-temporal logic, we can formulate and predict how likely such properties are to be satisfied in the future at a certain location. Moreover, we can leverage our methodology to evaluate and compare models directly on their ability to predict the satisfaction of application-driven properties. The approach is illustrated in an urban mobility application, where the crowdedness in the center of Milan is proxied by aggregated mobile phone traffic data. We specify several desirable spatio-temporal properties related to city crowdedness such as a fault-tolerant network or the reachability of hospitals. After verifying these properties on draws from the posterior predictive distributions, we compare several spatio-temporal Bayesian models (with and without nonparametric aspects) based on their overall and property-based predictive performance. (joint with Laura Vana, Ennio Visconti, Laura Nenzi and Annalisa Cadonna)

Back to Contents Page →

Santiago Marin Ardila
The Australian National University, Australia

BOB: Bayesian Optimized Bootstrap for Uncertainty Quantification in Gaussian Mixture Models

A natural way to quantify uncertainties in Gaussian mixture models (GMMs) is through Bayesian methods. That said, sampling from the joint posterior distribution of GMMs via standard Markov chain Monte Carlo (MCMC) imposes several computational challenges, which have prevented a broader full Bayesian implementation of these models. A growing body of literature has introduced the Weighted Likelihood Bootstrap and the Weighted Bayesian Bootstrap as alternatives to MCMC sampling. The core idea of these methods is to repeatedly compute maximum a posteriori (MAP) estimates on many randomly weighted posterior densities. These MAP estimates then can be treated as approximate posterior draws. Nonetheless, a central question remains unanswered: How to select the random weights under arbitrary sample sizes. We, therefore, introduce the Bayesian Optimized Bootstrap (BOB), a computational method to automatically select these random weights by minimizing, through Bayesian Optimization, a black-box and noisy version of the reverse KL divergence between the Bayesian posterior and an approximate posterior obtained via random weighting. Our proposed method outperforms competing approaches in recovering the Bayesian posterior, it provides a better uncertainty quantification, and it retains key asymptotic properties from existing methods. BOB's performance is demonstrated through extensive simulations, along with real-world data analyses.

Back to Contents Page →

Rafael Mediros Cabral
National University of Singapore, Singapore

Scalable Bayesian Clustering for Multi-View Data: Unveiling Dependencies in High Dimensions

We introduce a novel Bayesian clustering model for multi-view data, allowing for flexible information sharing between

domains. A Bayesian non-parametric model is specified for each domain, and the different domains are linked through appropriate prior specification. Our approach is based on recent results on distance-based clustering and allows capturing multi-view dependencies, such as those in partitions that evolve across time, enhancing its applicability to complex datasets. We introduce two strategies, one allowing for nested partitions and the other for overlapping partitions across the domains. The model outperforms traditional Bayesian non-parametric algorithms regarding speed, scalability, and robustness, enabling effective clustering in high dimensions. Our primary application involves a numismatic study aimed at clustering 523 coin images—both obverse and reverse—according to the distinct dies used in their minting.

Back to Contents Page →

Deyu Ming
University College London, UK

Deep Gaussian Process Emulation for Model Networks using Stochastic Imputation

Modern scientific problems often span multiple disciplines, necessitating the integration of distinct computer models, each with unique functional complexities, computation times, and programming environments. In this talk, we introduce deep Gaussian process (DGP) emulations that offer powerful tools to address emulation challenges related to computer model networks and computer models with non-stationary behaviors, by conceptualizing a computer model network as a DGP with partial exposure of its hidden layers. We develop a method called Stochastic Imputation, which integrates the Expectation-Maximization algorithm with elliptical slice sampling, for inferring the partially exposed non-parametric deep networks in a Bayesian fashion. Through a series of synthetic and empirical examples, including the Joint UK Land-use simulator and a pulmonary simulator, we demonstrate that our DGP emulators, augmented by sequential designs and automatic structural pruning, deliver significantly better performance than conventional Gaussian process emulators in terms of predictive accuracy, uncertainty quantification, and computation. The implementations of the examples are facilitated by our package, 'dgpsi', which is publicly available for both R and Python users on CRAN and CONDA.

Back to Contents Page →

Michael Minyi Zhang
University of Hong Kong, Hong Kong

A Semi-Bayesian Nonparametric Estimator of the Maximum Mean Discrepancy Measure

A classic inferential statistical problem is the goodness-of-fit (GOF) test. Such a test can be challenging when the hypothesized parametric model has an intractable likelihood and its distributional form is not available. Bayesian methods for GOF can be appealing due to their ability to incorporate expert knowledge through prior distributions. However, standard Bayesian methods for this test often require strong distributional assumptions on the data and their relevant parameters. To address this issue, we propose a semi-Bayesian nonparametric (semi-BNP) procedure in the context of the maximum mean discrepancy (MMD) measure that can be applied to the GOF test. Our method introduces a novel Bayesian estimator for the MMD, enabling the development of a measure-based hypothesis test for intractable models. Through extensive experiments, we demonstrate that our proposed test outperforms frequentist MMD-based methods by achieving a lower false rejection and acceptance rate of the null hypothesis. Furthermore, we showcase the versatility of our approach by embedding the proposed estimator within a generative adversarial network (GAN) framework. It facilitates a robust BNP learning approach as another significant application of our method. With our BNP procedure, this new GAN approach can enhance sample diversity and improve inferential accuracy compared to traditional techniques.

Back to Contents Page →

Alexander Mozdzen
University of Klagenfurt, Austria

Non-local mixture models with repulsive weights

Mixture models are a popular tool for studying data which is assumed to arise from multiple clusters. Common problems when employing standard mixture models to uncover clusters within a sample are the estimation of overlapping clusters and interpretability. Repulsive mixtures aim to ameliorate this drawback by favouring the estimation of well separated component locations. In a Bayesian framework, this is achieved by incorporating the distance between the locations into their prior distribution. Such a distribution can be constructed by multiplying a standard, potentially conjugate prior with a repulsive term. Numerous ways of incorporating a repulsive term into the locations of mixture models as well as computational methods for their estimation have been proposed in the literature. We extend this development and propose a mixture model with a repulsive {\it potential} acting on both locations and weights. This strategy leads not only to well separated components but also to non-negligible weights, greatly improving interpretability and reducing redundancy in the estimates. Posterior inference is performed through a tailored MCMC scheme, including a birth-and-death step to update the random number of components. The performance of the model is demonstrated in a simulation study as well as on real data applications.

Joint with Andrea Cremaschi, Maria de Iorio, and Gregor Kastner

Back to Contents Page →

Giovanni Rebaudo
University of Turin, Italy

Understanding partially exchangeable nonparametric priors for discrete structures

---

Species sampling models provide a general framework for random discrete distributions that are tailored for exchangeable data. However, they fall short when used for modeling heterogeneous data collected from related sources or distinct experimental conditions. To address this, partial exchangeability serves as the ideal probabilistic framework. While numerous models exist for partially exchangeable observations, a unifying framework, like species sampling models, is currently missing for this framework. Thus, we introduce multivariate species sampling models, a general class of models characterized by their partially exchangeable partition probability function. They encompass existing nonparametric models for partial exchangeable data, highlighting their core distributional properties. Our results allow the study of the induced dependence structure and facilitate the development of new models.

This is a joint work with Beatrice Franzolini, Antonio Lijoi, and Igor Pruenster.

Back to Contents Page →

Tommaso Rigon
University of Milano-Bicocca, Italy

A Bayesian theory for the estimation of biodiversity

The assessment of biodiversity is a major challenge in ecology. An influential ecological theory suggests the existence of a fundamental biodiversity number, say alpha, which coincides with the precision parameter of a Dirichlet process. In this paper, we review the statistical implications of this theory and highlight generalizations developed in Bayesian nonparametrics, such as Gibbs-type priors. We argue that the sigma-diversity is the most natural extension of the fundamental biodiversity number and discuss strategies for its estimation. Additionally, we contribute to Bayesian nonparametric theory by carefully examining a model termed Aldous-Pitman process, which serves as the building block for any Gibbs-type prior with a square-root growth rate. Finally, we propose a modeling framework that accommodates the hierarchical structure of Linnean taxonomy, offering a more refined approach to quantifying biodiversity. The analysis of a large and comprehensive dataset on Amazonian tree flora integrates our theoretical discussion.

Back to Contents Page →

Ziyi Song
University of California, Irvine, USA

Clustering Computer Mouse Tracking Data with Informed Hierarchical Shrinkage Partition Priors

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. 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 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.

Back to Contents Page →

Victoria Volodina
The University of Exeter, UK

Propagating moments in probabilistic graphical models for decision support systems

---

Probabilistic graphical models are widely used to model complex systems with uncertainty. Traditionally, Gaussian directed graphical models are applied for analysis of large networks with continuous variables since they can provide conditional and marginal distributions in closed form simplifying the inferential task. The Gaussianity and linearity assumptions are often adequate, yet can lead to poor performance when dealing with practical applications. In this research, we propose to model each variable in graph G as a polynomial regression of its parents to capture complex relationships between individual variables and with utility function of polynomial form. Since the marginal posterior distributions of individual variables can become analytically intractable, we develop a nonparametric message-passing algorithm to propagate information throughout the network solely using moments which enables the expected utility scores to be calculated exactly. We illustrate how the proposed methodology works with examples and in an application to decision problems in energy planning and healthcare.

Back to Contents Page →

Shijia Wang
ShanghaiTech University, China

An adaptive ABC-MCMC with Global-Local proposals

In this paper, we address the challenge of Markov Chain Monte Carlo (MCMC) algorithms within the approximate Bayesian Computation (ABC) framework, which often get trapped in local optima due to their inherent local exploration mechanism, particularly for multi-modal distributions and small ABC threshold. To remedy this, we propose a novel global-local MCMC algorithm that combines the "exploration" capabilities of global proposals with the "exploitation" finesse of local proposals. We integrate iterative importance resampling into the likelihood-free framework to establish an effective global proposal distribution, and adapt a normalizing flow-based probabilistic distribution learning model to iteratively improve the algorithmic performance. Furthermore, in order to optimize the efficiency of the local sampler and overcome the limitations caused by random walk behavior in high-dimensional space, we utilize Langevin dynamics to propose candidate parameters and utilize completely random numbers (CRN) to enhance the stability of the gradient estimation. We numerically demonstrate that our method is able to improve sampling efficiency and achieve more reliable convergence for complex posteriors.

Back to Contents Page →

Junyi Zhang
Hong Kong Polytechnic University, Hong Kong

Posterior Inference for Truncated Inverse-Lévy Measure Representation of CRMs

---

In this talk, we discuss the finite approximation of the completely random measure (CRMs) by truncating its inverse-Lévy measure representation. The approximation is obtained by keeping the N largest atom weights of the 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 latent feature model. Numerical implementations are given based on the generalised gamma and beta processes.

Back to Contents Page →

Zhaoxi Zhang
University of Edinburgh, UK

The underlap coefficient as a measure of covariate dependence in cluster analysis

---

Mixture models are invaluable tools for density estimation and clustering tasks. After obtaining a partition of responses by the mixture model, assessing the dependency of these partitions on covariates is of great importance. This is particularly relevant in applications where understanding the influence of covariates on clusters or subpopulations is crucial, such as in precision medicine for targeted interventions. In this context, we propose the use of the underlap coefficient as a metric for measuring covariate dependence in cluster analysis. Initially designed to quantify separation between distributions, we posit that the underlap coefficient can also 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 UNL 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.

Back to Contents Page →

Xiaotian Zheng
University of Wollongong, Australia

Model-based inference for spatial biological-dissimilarity indices

A pairwise dissimilarity index is a measure used to quantify site-pairwise dissimilarity in species composition within a spatial region. The index is often calculated using species occurrence data, and typically, statistical models are fitted to indices in order to identify key ecological drivers of biodiversity change. However, such a modeling approach does not make a distinction between the noisy, incomplete species data and the underlying ecological process. Consequently, dependence in the process's pairwise dissimilarities is not properly accounted for. Here, we propose a hierarchical statistical framework that takes into consideration data-sampling errors and enables inference on the latent pairwise dissimilarities directly from species data. Moreover, we propose a generalized chi-squared process that can describe the spatial dependence in pairs of dissimilarities. We study the model analytically and through simulation studies, and we apply our methodology to an ecological data set.

This research is joint with Noel Cressie and Andrew Zammit-Mangion of the Centre for Environmental Informatics at the University of Wollongong.

Back to Contents Page →