List of Posters

Tuesday Session

Adaptive MCMC for hierarchical Bayesian models applied to biomass data of fishes

Antonio Abbruzzo Tuesday, June 28

Department of Economics, Business and Statistics

International bottom trawl survey in the Mediterranean Programm (MEDITS) collects data concerning biomass of several fishes' species in the area extending from the southern coasts of Sicily. The availability of such spatially referenced data, given by the constant technological refinement, leads to a strong interest in developing statistical models that deal with spatial processes. We adopt a hierarchical Bayesian model that considers the spatial domain.

The model consists of three parts which identify the variability due to the explanatory variables, the variability due to the spatial processes (seen as a Gaussian Process) and the noise component. The hierarchical Bayesian model suffers identifiability problems of the parameters, reduced by prior information. Finally, we explore adaptive MCMC algorithms, which allow convergence to high-dimensional target distributions much more quickly than straightforward methods such as random walk Metropolis or Gibbs sampling.

INLA+ : A Computationally Efficient Method for Fitting Intrinsic Spatiotemporal Models

Esmail Abdul Fattah Tuesday, June 28

PhD Candidate

The massive growth of spatiotemporal datasets partly due to advanced technologies, emphasizes on the increased demand for developing new and computationally efficient statistical models for analyzing big data. When we account for the space-time interactions in these statistical models, inference becomes more useful in revealing unknown patterns in the data. However, when the number of locations and/or number of time points are large, the inference gets computationally challenging due to the high number of required constraints necessary for precise inference, and this holds for various inference architectures including Markov chain Monte Carlo (MCMC) Gilks et al. (1997) and Integrated Nested Laplace Approximations (INLA) Rue et al. (2009). INLA is still a promising approach in many cases, although it is restricted to a certain dependence structure between the canonical parameter and the linear additive predictor. The computational complexity increases when the ratio of the number of constraints and the number of paramaters exceeds a certain threshold. Based on INLA methodology, we propose a new efficient approach for models with many constraints on the parameter space, time and their interactions using hybrid parallel method.

Intuitive Joint Priors for Bayesian Multilevel Models: The R2D2M2 prior

Javier Enrique Aguilar Romero Tuesday, June 28

SimTech University of Stuttgart

Regression models are ubiquitous in the quantitative sciences making up a big part of all statistical analysis performed on data. In the quantitative sciences, data often contains multilevel structure, for example, because of natural groupings of individuals or repeated measurement of the same individuals. Multilevel models (MLMs) are designed specifically to account for the nested structure in multilevel data and are a widely applied class of regression models. From a Bayesian perspective, the widespread success of MLMs can be explained by the fact that they impose joint priors over a set of parameters with shared hyper-parameters, rather than separate independent priors for each parameter. However, in almost all state-of-the-art approaches, different additive regression terms in MLMs, corresponding to different parameter sets, still receive mutually independent priors. As more and more terms are being added to the model while the number of observations remains constant, such models will overfit the data. This is highly problematic as it leads to unreliable or uninterpretable estimates, bad out-of-sample predictions, and inflated Type I error rates. The primary objective of our project is thus to develop, evaluate, implement, and apply intuitive joint priors for Bayesian MLMs. We hypothesize that our developed priors will enable the reliable and interpretable estimation of much more complex Bayesian MLMs than was previously possible.

Bayesian Functional Emulation of CO2 Emissions on Future Scenarios

Luca Aiello Tuesday, June 28

University of Milano Bicocca - DEMS

We propose a statistical emulator for a climate-economy deterministic model ensemble, based on a functional regression framework. Inference on the unknown parameters is carried out through a Mixed Effects Hierarchical Model using a fully Bayesian framework with a prior distribution on the vector of all parameters. We also suggest an autoregressive parameterization of the covariance matrix of the error, with matching marginal prior. In this way, we allow for a functional framework for the discretized output of the simulators that allows their time continuous evaluation.

Modeling Renewable Energy Consumption in Africa Using Bayesian Model Averaging

Olalekan Akintande Tuesday, June 28

University of Ibadan

The increasing concern over global warming and energy security has rejuvenated the renewable energy option as the most vibrant option to sustain future energy needs. This paper developed a renewable energy consumption model using annual data spanning between 1996 and 2016 in the five most populous countries (Ethiopia, South Africa, Nigeria, DR Congo, and Egypt) in Africa. Following the existing literature on the subject, the driving factors investigated were categorized into three broad areas. These include macroeconomic, socioeconomic, and institutional variables. Altogether, thirty-four predictor variables are analyzed. The study employed Bayesian Model Averaging (BMA) procedures to account for the uncertainty associated with model choice and variable selection. The results of the analysis indicate that population growth, urban population, energy use, electric power consumption, human capital are the main determinants of renewable energy consumption in the selected countries. Also, an increase in any of these determinants (population growth, urban population, energy demand/use, electricity power demand/consumption) causes an increase in renewable energy consumption.

Trade-off between farm production and flood alleviation using tillage as natural flood management (NFM) strategy.

Qaisar Ali Tuesday, June 28

University of Reading, United Kingdom

Tillage is practiced for management of crop production and additionally adept as NFM strategy to create soil surface roughness across local gradient, and resistance in generating runoff through efficient water absorption, and infiltration. However, mechanical working with soils can hardly limit soil structural changes which may cause unavoidable consequences beyond impacting farm production. Soil compaction, erosion, depletion of soil organic carbon, and soil structural loss are few of them which reduce soil water permeability, water retention and storage in soil profile in longer run hampering sustainable production. Later, these bring adverse soil conditions with lack of evidence with the changing climate in due course. Bayesian approach provides an exceptional tool to highlight interacting variables with causation, and their strength of effects in involved phenomena. Hence, a Bayesian Belief Network (BBN) model was developed to eminence the impacts of tillage on flood alleviation and farm production to quantify their influence among various interacting variables potentially in a diverse fashion. This study discovered a trade-off between farm production and flood alleviation using tillage as NFM strategy. Model can help users in decision support against certain choices. Model contains the potential for the improvement towards customized adjustment from farm level to catchment scale.

Bayesian Inference in Point Process Models Based on Thinning Procedures

Renaud Alie Tuesday, June 28

McGill University

In this poster, we discuss point process models based on thinning procedures, where each point of a base configuration is either kept or discarded. We present a principled way of carrying Bayesian inference in such models via data augmentation. Basically, imaginary thinned locations are instantiated at each step of the MCMC algorithm according to their conditional distribution. This distribution is not always straightforward to derive; we detail some of the potential issues.

We then provide a general colouring theorem (named after the homonymous result for Poisson processes) that streamlines most of the measure theoretic details involved in obtaining the joint distribution of thinned and observed locations in any model with a tractable density. We demonstrate its usefulness on common models based on thinning procedures and introduce interesting variations. Finally, we showcase applications on datasets provided in the spatstat package in R.

Estimating capital asset price model with many instruments

Cássio Alves Tuesday, June 28

University of Sao Paulo

We propose a Bayesian approach to estimate the capital asset price model (CAPM) using a large set of instruments and shrinkage priors over the parameters associated with the instrumental variables. The CAPM model suffers from the error-in-variables problem since market return is unobservable, and a proxy, such as SP500, is used instead, introducing measurement errors in the regressor. When using the instrumental variable approach to estimate the CAPM, the challenge is to find "strong" instruments to the market return. The data-rich environment available in finance allows us to use many-instrument settings. We use regularization priors to deal with a large number of instrumental variables. Using simulated data with 20, 80, and 140 instruments and 300 observations, we find that our approach may reduce the average bias caused by error-in-variables up to 90% concerning the traditional two-stage least square. This reduction, however, is attenuated as the number of instruments increases.

A Zero-Inflated Poisson model with spatio-temporal varying coefficients

Alaa Amri Tuesday, June 28

The University of Edinburgh

We propose, in a Bayesian context, a zero-inflated Dynamic Generalized Linear model where the coefficients vary both spatially and temporally to account for the excess of zeros in discrete spatio-temporal data while reasonably forecasting non-zero counts. We particularly allow the zero-inflation probability to vary temporally in order to make the model more flexible. An inference algorithm, that is based on Particle Gibbs, is described. Our approach is compared with other existing models, and it is illustrated with both simulated and real datasets where observations are observed at different points in space and time. The real dataset is about cycling counts in the City of Glasgow collected from Strava and automated bike counters.

Bayesian hierarchical modeling to account for complex patterns of measurement error in cohort studies. Application in radiation epidemiology.

Sophie Ancelet Tuesday, June 28

IRSN

Despite its deleterious consequences for statistical inference and its ubiquity in observational research, exposure measurement error is rarely accounted for in epidemiological studies. Basically, standard correction methods, like regression calibration or SIMEX, often lack the flexibility to account for complex patterns of exposure uncertainty. However, in occupational cohort studies, for instance, changes in the methods of exposure assessment can lead to complex error structures. Moreover, a strategy of group-exposure assessment and individual worker characteristics may lead to error components that are shared between or within individuals. In this work, we thus propose and fit several Bayesian hierarchical models, combining survival models with time-dependent covariates, measurement and exposure models, to obtain a corrected estimate of the potential association between exposure to radon and mortality for several cancer types in the French cohort of uranium miners. A simulation study is under progress to assess the impact of model misspecification on risk estimates.

Optimal lower bounds for logistics likelihoods

Niccoló Anceschi Tuesday, June 28

Bocconi University

The use of logit mapping in binary regression models notoriously hinders tractable analytical inference. Within the Bayesian framework, data-augmentation strategies addressing this problem have received considerable attention. Conversely, unconstrained and penalized maximum likelihood estimation typically exploits quadratic approximations of the logistic log-likelihood, either arising from Newton's method or from tangent lower bounds. As the former does not guarantee stable convergence, we focus on the latter showing that the lower bound corresponding to the Polya-Gamma data-augmentation scheme is optimal among quadratic bounds. Furthermore, we derive a novel tangent minorizer dominating the Polya-Gamma one, by adding a piece-wise linear contribution depending on the L1-norm of the linear predictors. Such novel lower bound still allows for a simple coordinate-wise minimization algorithm, as routinely implemented in the literature for lasso and elastic-net penalized logistic regression. Empirical results confirm that an optimization procedure exploiting the novel minorization scheme benefits from an improved convergence rate.

The flexible Dirichlet-multinomial regression model

Roberto Ascari Tuesday, June 28

University of Milano-Bicocca

Count compositions are vectors of non-negative integers summing to a fixed constant. They are widespread in biomedical research, especially in microbiome data analysis. A widespread distribution for human microbiome data is obtained by compounding the multinomial distribution with the Dirichlet; this approach leads to the Dirichlet-multinomial (DM). The DM distribution fits real data better than the multinomial one, but its covariance structure may still be too rigid to deal with real data.

This work aims to propose a new distribution for count compositions and to develop a regression model based on it. The new distribution can be expressed as a structured finite mixture with particular DM components. Thanks to this mixture structure and the additional parameters introduced, the new distribution can provide a better fit and an interesting interpretation in terms of latent groups. Inferential issues are dealt with by a Bayesian approach through the Hamiltonian Monte Carlo algorithm.

Trees of random probability measures and Bayesian nonparametric modelling

Filippo Ascolani

Bocconi University

Tuesday, June 28

We introduce a way to generate trees of random probability measures, where the link between two nodes is given by a hierarchical procedure: starting from a common root, each node of the tree is endowed with a random probability measure, whose baseline distribution is again random and given by the associated node in the previous layer. The data can be observed at any node of the tree and different branches may have different length: the split mechanism can be also considered random or based on covariates of interest. When the branches have the same length and the observations are linked only to the leaves, we recover the well known family of discrete hierarchical processes.

We prove that, if the distribution at each node is given by the normalization of a completely random measure (NRMI), the model is analytically tractable: conditional on a suitable latent structure, the posterior is still given by a tree of NRMIs. Furthermore, the asymptotic behaviour of the number of clusters is derived, when either the sample size at a particular layer diverges or the number of levels grows. Finally, applications to real data are discussed.

Bayesian Multilevel Analysis of Determinants of Acute Respiratory Infection in Children under the Age of Five Years in Ethiopia

Tilahun Asena Tuesday, June 28

Arba Minch University

Background: Acute respiratory tract infection (ARI) is one of the causes for morbidity and mortality in children under the age of 5 years in the world. Pneumonia, which is caused by respiratory tract infection, accounts for approximately 1.9 million deaths globally in children under the age of five years. Among these deaths majority occurs in the developing world. In Ethiopia, the prevalence rate of ARI was 7% according to 2016 Ethiopia Demographic and Health Survey (EDHS) estimates. Method: Bayesian multilevel approach was employed to assess factors associated with the prevalence of ARI among children under five in Ethiopia. The data was collected from 10,641 children under the age of five years out of which 9,918 children were considered in this study. Result: The ARI prevalence rate for children under five years was estimated as 8.4%, which was slightly higher than the estimated prevalence level of the country. The highest proportion of the prevalence of ARI was observed for children whose mothers had no education. The major health, environmental and nutritional related background characteristics of the proportion of children who had ARI varied from one region to another. The highest prevalence of ARI was observed in Tigray (15.31%) followed by Oromia (14.40%) as opposed to the low prevalence which was recorded in Benishangul Gumuz (2.58%). The utilization of vitamin A was analyzed and the results shows that about 43.10% who received vitamin A had the lowest proportion on the prevalence of ARI (7.75%) compared to not having vitamin A. About 11.13% of children under five had Diarrhea with the highest prevalence of ARI (24.64%) and the highest prevalence of ARI was observed for the child whose source of drinking water were unprotected/unimproved (9.39%). Conclusion: The age of the child, household wealth index, mother educational level, and vitamin A supplement, history of diarrhea, maternal work, stunting and source of drinking water was found to be significantly affecting the prevalence of ARI among children under five years. Furthermore, the study revealed that there is a significant variation of incidence of ARI between and within the regions of Ethiopia. Attention should be given to those predictor variables while planning to increase the health status of children in Ethiopia.

Hierarchical Bayesian modeling under covariate shift in supernova cosmology

Maximilian Autenrieth Tuesday, June 28

Imperial College London

Supernova type Ia (SNIa) are an essential tool to constrain cosmological parameters, including dark energy, thought to make up 70% of the energy density of the universe. In order to identify transients as SNIa, upcoming large surveys will rely on supervised classification from low-resolution photometric data given a non-representative set of confirmed SNIa from high-quality spectra. To account for uncertainties and contamination from the probabilistic classification of SNIa, we discuss a fully hierarchical and pragmatic Bayesian framework to improve accuracy and precision of cosmological parameters estimates. Missing a scientifically justified model for the contaminated data, the pragmatic approach circumvents the need of a contamination model, while the fully Bayesian approach relies on correct model specification. The pragmatic approach further prevents potential complications of using data twice by not updating class probabilities. To obtain class probabilities, our framework includes a new general-purpose method to improve supervised learning from non-representative training sets (covariate shift). Exploiting methodology from causal inference, we show that the effects of covariate shift can be reduced or eliminated by conditioning on propensity scores through stratification, leading to balanced covariates and much-improved target prediction.

Leveraging External Data in Bayesian Adaptive Platform Designs

Alejandra Avalos Pacheco Tuesday, June 28

University of Florence

There is growing interest in trial designs that incorporate non-concurrent trial data with the goal of increase power. However, if the outcome distributions of the external and internal data differ, the integration of external data may lead to biased treatment effects estimates, reduced/increased power/type I error rates. We introduce novel designs that leverage external data via a Bayesian model averaging approach. Our main contributions are two-fold, we provide: (i) ethical designs that reduce the amount of patients assigned to control, thus maximizing the possible benefit to patients, (ii) flexible procedures to test the effectiveness of novel treatments, which satisfy a set of constraints on the operating characteristics required by regulators. We discuss the asymptotic characteristics of the test statistic, and study the finite-sample operating characteristics. We illustrate the performance of our proposed designs in several simulation studies based on data from real phase II/III trials of cancer therapies for glioblastoma.

Bayesian additive tree random measures

Naoki Awaya Tuesday, June 28

Duke University

Bayesian additive regression trees (BART) is a powerful predictive model widely used in supervised learning. Given the recent development of Polya tree (PT) models equipped with flexible partitioning priors, which is analogous to Bayesian CART in the supervised setting, a natural question that arises is whether the Bayesian additive tree-based approach can be adopted as a nonparametric prior for unsupervised learning. We introduce such a random measure based on an additive ensemble of PT base learners. This new random measure is constructed through adding up a sequence of PT measures under a recently introduced notion of addition on probability measures involving compositions of the so-called “tree-CDFs". To avoid overfitting this random measure, we propose a new regularizing prior for the base PT learners so that each PT works as a weak learner. We also introduce a stochastic back-fitting algorithm for posterior inference. The performance of the new model is evaluated in extensive numerical experiments, and it shows the drastic improvement in the fit to i.i.d. observations than the state-of-the-art single-tree based models.

Posterior Concentration Rates for Bayesian Penalized Splines

Paul Bach Tuesday, June 28

Humboldt-Universität zu Berlin

Despite their widespread use in practice, the asymptotic properties of Bayesian penalized splines have not been investigated so far. We close this gap and study posterior concentration rates for Bayesian penalized splines in a Gaussian nonparametric regression model. A key feature of the approach is the hyperprior on the smoothing variance, which allows for adaptive smoothing in practice but complicates the theoretical analysis considerably as it destroys conjugacy. Our main tool for the derivation of posterior concentration rates is a novel spline estimator that projects the observations onto the first basis functions of a Demmler-Reinsch basis. Our results show that posterior concentration at near optimal rate can be achieved if the hyperprior on the smoothing variance strikes a fine balance between oversmoothing and undersmoothing. Overall, our results are the first posterior concentration results for Bayesian penalized splines and can be generalized in many directions.

Causal Mediation in Weight Management Trials using Bayesian Nonparametrics

Woojung Bae Tuesday, June 28

University of Florida

In weight management trials, it is of interest to understand how interventions are effective. In a recent trial, we examine how the number of days calorie goals are met mediates the effect of a behavioral intervention on maintaining 18-month weight loss (after a 4-month weight loss program). To do this, we use a flexible Bayesian nonparametric (BNP) method and assume all confounders of the mediator-outcome relationship are measured. We present results from the analysis of this trial and compare them to simpler parametric and BNP approaches. Simulation studies are also conducted to examine the frequentist operating characteristics of this approach.

Calibrating generalized Bayesian posteriors for inverse problems

Youngsoo Baek Tuesday, June 28

Duke University

We propose a general framework for calibrating generalized Bayesian posteriors in inverse problems. The complex non-linear operators involved in many forward model simulations lead to intractable likelihoods, hence diverse proposals for approximate Bayesian computations in the literature. We instead adopt a non-generative, loss-based framework as in Bissiri, Holmes, and Walker (2016) to develop a principled approach to calibrating generalized posteriors, as follows. First, we propose a sequential Monte Carlo-based gradient descent for choosing a tuning parameter that weights the prior against the loss by cross-validation. Second, we formulate a joint variational principle on the space of posteriors and predictive distributions that avails model comparison, for which the appropriate cross-validation statistic is computable. Finally, we suggest guidelines for designing varied loss functions in presence of significant non-linearity in the forward model. We support the merits of our method through simulated experiments and ultrasound vibrometry application.

Informed Bayesian survival analysis

František Bartoš Tuesday, June 28

Department of Psychological Methods, University of Amsterdam

Parametric survival analysis is a powerful method for parameter estimation, hypothesis testing, and survival extrapolation of censored event history outcomes. We outline and implement a Bayesian model-averaging framework for parametric survival analysis that allows us to incorporate historical data, test informed hypotheses, continuously monitor evidence, and incorporate uncertainty about the true data generating process. We illustrate the Bayesian framework by re-analyzing data from a colon cancer trial. In a simulation study, we compare the Bayesian framework to maximum likelihood estimation of survival models with AIC/BIC model selection in fixed-n and sequential designs. We find that the Bayesian framework (1) produces faster decisions in sequential designs, (2) has slightly higher statistical power and false-positive rates in fixed-n designs, and (3) yields more precise estimates of the treatment effect in small and medium sample sizes. We did not find a clear advantage in predicting survival.

Bayesian optimal designs for choice experiments involving mixtures of ingredients and process variables

Mario Becerra Tuesday, June 28

KU Leuven

Discrete choice experiments are frequently used to quantify consumer preferences. Choice experiments involving mixtures of ingredients and process variables have been largely overlooked in the literature, even though several products can be described as mixtures of ingredients. For example, the ingredients to make up a fruit cocktail and the temperature at which it is served. As experiments in general are expensive and time-consuming, efficient experimental designs are required to provide reliable statistical modeling. Two optimality metrics have usually been studied: D-optimality and I-optimality. The Bayesian version of these metrics is obtained by assigning a prior distribution to the parameters of the model and averaging over the prior. We will show and compare the properties of Bayesian D- and I-optimal designs.

Scalable Bayesian generalized linear mixed models using stochastic gradient MCMC

Samuel Berchuch Tuesday, June 28

Duke University

Generalized linear mixed models (GLMM) are used to analyze correlated and longitudinal data and are used extensively in many application areas, including biomedical settings, where datasets are becoming larger and larger. Standard posterior sampling algorithms, such as Markov chain Monte Carlo (MCMC) procedures, are not inherently scalable and have limited utility in large datasets. To overcome this limitation, we introduce a stochastic gradient MCMC (SGMCMC) algorithm that uses mini-batch samples to approximate the true gradient over the whole dataset. Our algorithm uses a gradient computed using Fisher's Identity, which bypasses the intractable marginal likelihood associated with GLMM. A sampling convergence criterion is introduced that is based on the empirical relationship between the SGMCMC algorithm and an underlying stochastic differential equation. Through simulation, we show that our method scales to large data settings while maintaining parameter estimation performance. Finally, we apply the method to a large electronic health records database.

Gaussian processes at the Helm(holtz): A better way to model ocean currents

Renato Berlinghieri Tuesday, June 28

Massachusetts Institute of Technology

Understanding the behavior of ocean currents has the potential to improve ecosystem management, forecasting of oil spill dispersion, and the comprehension of ocean transportation. Since we expect current dynamics to be smooth but highly non-linear, Gaussian processes (GPs) offer an attractive model. However, existing naive approaches fail to capture real-life current structure, such as continuity of currents and the shape of vortices. By contrast, these physical properties are captured by divergence and curl-free components of a vector field obtained through a Helmholtz decomposition. In this paper, we thus propose instead to model these components with a GP directly. We show that, because this decomposition relates to the original vector field just via mixed partial derivatives, we can still perform inference given the original data with only a small constant multiple of additional computational expense. We illustrate our method on simulated and real oceans data.

Optimal Conformal Prediction for Small Areas

Elizabeth Bersson Tuesday, June 28

Duke University

Existing inferential methods for small area data involve a trade-off between maintaining area-level frequentist coverage rates and improving inferential precision via the incorporation of indirect information. In this article, we propose a method to obtain an area-level prediction region for a future observation which mitigates this trade-off. The proposed method takes a conformal prediction approach in which the conformity measure is the posterior predictive density of a working model that incorporates indirect information. The resulting prediction region has guaranteed frequentist coverage regardless of the working model, and, if the working model assumptions are accurate, the region has minimum expected volume compared to other regions with the same coverage rate. When constructed under a Normal working model, we prove such a prediction region is an interval and construct an efficient algorithm to obtain the exact interval. We illustrate the performance of our method through simulation studies and an application to EPA radon survey data.

Approximations of piecewise deterministic Markov processes and their convergence properties

Andrea Bertazzi Tuesday, June 28

TU Delft

Piecewise deterministic Markov processes (PDMP) received substantial interest in recent years as an alternative to classical Markov chain Monte Carlo algorithms. The applicability of PDMP to real world scenarios is currently limited by the fact that these processes can be simulated only when specific properties of the target distribution are known beforehand. In order to overcome this problem, we introduce an Euler-type discretisation scheme for PDMP which does not need such pre-requisite knowledge. For the resulting scheme we study both pathwise convergence to the continuous process as the step size converges to zero and convergence in law to the target measure in the long time limit. (This is a joint work with Paul Dobson and Joris Bierkens.)

Forecasting Models for Predictive Crime Mapping in South Asian Countries

Dila Ram Bhandari Tuesday, June 28

Nepal Commerce Campus, Tribhuvan University

South Asian countries facing the big challenges of various cases of crime. While the use of mapping in criminal justice has increased over the last 30 years, most applications are retrospective that is, they examine criminal phenomena and related factors that have already occurred. While such retrospective mapping efforts are useful, the true promise of crime mapping lies in its ability to identify early warning signs across time and space, and inform a proactive approach to police problem solving and crime prevention. Recently, attempts to develop predictive models of crime have increased, and while many of these efforts are still in the early stages, enough new knowledge has been built to merit a review of the range of methods employed to date. This chapter identifies the various methods, describes what is required to use them, and assesses how accurate they are in predicting future crime concentrations, or hot spots. Factors such as data requirements and applicability for law enforcement use will also be explored, and the chapter will close with recommendations for further research and a discussion of what the future might hold for crime forecasting.

fiBAG: Functional Integrative Bayesian Analysis of High-dimensional Multiplatform Genomic Data

Rupam Bhattacharyya Tuesday, June 28

University of Michigan

Large-scale multi-omics datasets offer complementary, partly independent, high-resolution views of the human genome. Modeling and inference using such data poses challenges like high-dimensionality and structured dependencies but offers potential for understanding the complex biological processes characterizing a disease. We propose fiBAG, an integrative hierarchical Bayesian framework for modeling the fundamental biological relationships underlying such cross-platform molecular features. Using Gaussian processes, fiBAG identifies mechanistic evidence for covariates from corresponding upstream information. Such evidence, mapped to prior inclusion probabilities, informs a calibrated Bayesian variable selection (cBVS) model identifying genes/proteins associated with the outcome. Simulation studies illustrate that cBVS has higher power to detect disease-related markers than non-integrative approaches. A pan-cancer analysis of 14 TCGA cancer datasets is performed to identify markers associated with cancer stemness and patient survival. Our findings include both known associations like the role of RPS6KA1/p90RSK in gynecological cancers and interesting novelties like EGFR in gastrointestinal cancers.

Semiparametric Variational Inference for sparse time-varying parameter model

Nicolas Bianco Tuesday, June 28

University of Padova

Time-varying parameter models are widely used in statistics for the analysis of dynamical systems. They link observed variables and state variables to draw statistical inference about the unobserved states. Here we consider time-varying regressions within a Bayesian context. Since the risk of over-parametrization is high, sparsity is a desired property in such a models. The latter is defined over two directions: over the parameter vector at a fixed time and across the timeline for a given variable. However, the formulation of the parameter as a latent process makes estimation more complicated. Approximate inference can be helpful in this scenario to reduce the computational effort. We propose a variational Bayes approach for parameter estimation and signal extraction that relies on a global flexible approximation of the latent states through a non-stationary Gaussian Markov random field.

Modelling wildlife movement with Piecewise Deterministic Markov Processes

Paul Blackwell Tuesday, June 28

University of Sheffield, UK

Piecewise Deterministic Markov Processes have a natural application as models of wildlife movement, capturing aspects of the movement more realistically than the diffusion models or discrete-time Hidden Markov Models that are often used. Inference for such models from discrete observations can be challenging; an existing approach using Approximate Bayesian Computation allows some estimation of parameters, but not the detailed reconstruction that is necessary for many ecological problems. I will describe an inference approach using a form of Reversible Jump Markov Chain Monte Carlo that does permit the detailed reconstruction of movement paths, opening up the possibility of linking the parameters to behaviour and environmental covariates.

Estimating the potential to prevent locally acquired HIV infections in a UNAIDS Fast-Track City, Amsterdam

Alexandra Blenkinsop Tuesday, June 28

Imperial College London/AIGHD

Amsterdam and other UNAIDS Fast-Track cities aim for zero new HIV infections. A milestone is to characterise the number of HIV infections acquired from sources within cities. We located diagnosed HIV infections in Amsterdam postcodes. Infection dates were estimated from biomarker data (Pantazis, 2019), and used to estimate the proportion of undiagnosed individuals at population level. Transmission chains were phylogenetically reconstructed from routinely available pol sequences. To account for missing data, Bayesian branching process models were used to model growth of transmission chains. An estimated 516 of diagnosed infections in Amsterdam were acquired between 2014-2018, and over 20% of infections remained undiagnosed by May 2019. 68% [61-74%] of infections in Amsterdam MSM, and 57% [41-71%] in heterosexuals, were estimated to be locally acquired, with 43% [37-49%] of local infections in foreign-born MSM. Our analyses indicate potential to further curb local transmission, with foreign-born MSM benefitting most from intensified interventions.

Modelling Populations of Path-observed Networks via Distance Metrics

George Bolt Tuesday, June 28

Lancaster University

Network data arises through observation of relational information between a collection of entities. In this talk, a new Bayesian modelling framework will be introduced to analyse a particular (and currently unconsidered) type of network data where (i) we observe not a single but multiple networks, and (ii) within each network paths are the units of observation, e.g. page visits to a website for a sample of users. Through use of distance metrics between observations, we construct models based on location and scale parameters, akin to a Gaussian distribution. Parameter inference subsequently provides analogues of the mean and variance respectively, permitting output of statistical summaries respecting the data structure. This is supplemented with MCMC algorithms to sample from the models and their associated posterior distributions; a task made non-trivial thanks to our target space including discrete objects with variable dimensions and the presence of intractable normalising constants in the likelihood.

Approximate Bayesian Computation (ABC) for the natural history of breast cancer, with application to data from a Milan cohort study

Laura Bondi Tuesday, June 28

MRC Biostatistics Unit, University of Cambridge

We propose a new class of multi-state models for the natural history of breast cancer, where the main events of interest are the start of asymptomatic detectability of the disease and the start of symptomatic detectability. We develop a cure rate parametric specification that allows for dependence between the times from birth to the two events, and present the results of the analysis of longitudinal data from the Milan breast cancer screening program. Due to the intractability of the observed likelihood function arising from the complex missing data structure, we rely on Approximate Bayesian Computation (ABC) for inference. We discuss issues that arise from the use of ABC for model choice and parameter estimation, with a focus on the problem of choosing appropriate summary statistics. The estimated latent disease process allows for the study of the effect of different examination schedules and adherence patterns on a population of asymptomatic subjects.

Alternatives for approximation in likelihood-free inference: Box-ABC

Elena Bortolato Tuesday, June 28

Department of Statistical Sciences, University of Padova

For some statistical models, the likelihood function is intractable, or a considerable amount of time could be necessary for it to be evaluated. Likelihood-free inference and in particular Approximate Bayesian Computation methods avoid the problem of evaluating the likelihood by comparing actual to some simulated data according to distances. In this context, we propose algorithms that avoid the choice of distances and tuning parameters, aiming at reducing the arbitrariness of the approximation. The acceptance rule implicitly makes use of a pseudo-likelihood that inherits some desirable properties from confidence distributions, such as median unbiasedness, and consistency guarantees of the related pseudo-posterior. A corrected version of the pseudo-likelihood that makes use of Efron's implied likelihood (1993) can also be retrieved.

Bayesian extrapolation from multi-source and multi-species pre-clinical data to human

Sandrine Boulet Tuesday, June 28

Inserm, Centre de Recherche des Cordeliers, Sorbonne Université, Université de Paris

The development of a new drug requires a preclinical trial phase that includes in vitro, in vivo and in silico experiments. These experiments allow to determine a safe and effective starting dose range for the first-in-human study. Nevertheless, usually each experiment is analyzed independently, without using previous analysis results, and the final doses are often chosen based on one experience only. We propose a Bayesian framework for the extrapolation (from preclinical to clinical) of multi-source and multi-species data to predict the doses of interest (e.g. the minimum effective dose, the maximum tolerated dose, etc.) in human. A full Bayesian approach, divided in four main steps, is built to deal with the sequential estimation nature, the extrapolation, the commensurability checking and the information merging. The new framework is evaluated via an extensive simulation study, inspired by a real-life example in oncology, the inhibition of TGF-beta signaling to block tumor growth.

Trans-dimensional histogram kernel for the discrete-time self-exciting process

Raiha Browning Tuesday, June 28

Queensland University of Technology

Hawkes processes are a self-exciting stochastic process first introduced as a continuous-time point process to describe phenomena whereby past events increase the probability of the occurrence of future events. This work presents a flexible approach for modelling a variant of these, namely discrete-time Hawkes processes (DTHP). These are less well studied than their continuous-time counterparts, but occur frequently in practice. Most standard models of Hawkes processes rely on a parametric form for the function describing the influence of past events, referred to as the triggering kernel. This is likely to be insufficient to capture the true excitation pattern, particularly for complex data. By utilising trans-dimensional MCMC inference techniques, our proposed model for the triggering kernel can take the form of any step function, affording much more flexibility than a parametric form.

Through a comprehensive simulation study, we illustrate the situations in which the proposed model performs well or otherwise. Initially, a univariate DTHP is considered followed by multivariate scenarios where there are multiple interacting Hawkes processes. In particular, the bivariate process that we consider demonstrates the multivariate case and is an exemplar for extending to higher dimensions where sufficient data and computational resources are available. Moreover, a key finding is how the balance between prior informativeness and the informativeness of the data affects inference in a high-dimensional parameter space. At least one of these is required to produce reliable inference. For higher-dimensional problems, this is of particular importance.

We also apply the proposed model to real-world data for several case studies. First, we model the interaction between two countries during the COVID-19 pandemic, taking France and Italy as an exemplar of two countries that are likely to have an interaction due to spatial proximity. We recover excitation patterns between these countries that align with what occurred historically. We then capture the interaction of terrorist activity in two countries of close spatial proximity, Indonesia and The Philippines. We find that excitation mainly occurs within each region rather than cross-excitation between areas.

Inferring Transmission Structure from HIV Sequence Data via Latent Spatial Poisson Processes

Fan Bu Tuesday, June 28

University of California, Los Angeles

Viral deep-sequencing technologies are finding increasing use toward understanding disease transmission networks and population-level transmission flows, but the phylogenetic analysis outcomes are accompanied with uncertainties. To utilize such rich data to uncover HIV transmission flow patterns between different age groups, we develop a spatial Poisson process model to characterize point patterns of multiple types where the type labels are latent variables that represent the unobserved transmission statuses between potential sources and recipients. In contrast to existing methods, our framework does not require pre-classification of the transmission statuses of data points, instead learning them probabilistically through a fully Bayesian inference scheme. Such an approach jointly leverages the deep-sequencing information as well as spatial structures in disease transmission. Moreover, our model makes direct use of continuous processes, and thus enjoys improved computational efficiency compared to previous methods that rely on discretization. Through simulations and a real data case study, we demonstrate that our framework can capture age structures in HIV transmission and bring new insights into valuable epidemiological questions.

Bayesian Design On River Network Systems

Katie Buchhorn Tuesday, June 28

QUT

Monitoring of the environment, more specifically river networks, provide us with important ecological information to maintain healthy ecosystems and inform subsequent decision-making and policy. Increasingly, 'in-situ' sensors are being deployed to collect data across river networks. Optimal design offers a framework for the placement of such sensors to provide maximum information about the state of the river network (i.e., maximising prediction certainty, and/or minimising uncertainty of model parameters, and/or increasing the reliability of data...). However, river networks pose statistical challenges due to the branching network topology and the directionality and volume of flow. Recently, a suite of spatial statistical models has been proposed for such purposes. For this, we propose to consider Bayesian design methods and apply these to design a real river network monitoring program in the north-western USA. To address practical constraints of collecting data in natural environments, we propose an algorithm to find Bayesian sampling windows (rather than sampling points) to identify regions of high sampling efficiency. We also propose a non-parametric version of a known coordinate exchange algorithm to locate our Bayesian designs.

Spatially-Varying Bayesian Predictive Synthesis for Flexible and Interpretable Spatial Prediction

Danielle Cabel Tuesday, June 28

Temple University

We consider model uncertainty and predictive synthesis for spatial data. Spatial data are characterized by their spatial dependence, which are often complex, non-linear, and difficult to capture with a single model. This leads to significant model uncertainty that cannot be resolved by simple ensemble methods. We propose a novel method to deal with model uncertainty in spatial data by developing a spatial version of the Bayesian predictive synthesis framework by defining a latent factor spatially varying coefficient model as a synthesis function. This specification, using a Gaussian process, captures spatially dependent biases and dependencies amongst models and effectively learns spatially varying synthesis weights. We implement an MCMC strategy for full uncertainty quantification, as well as a variational inference strategy for fast point inference. Several simulation examples and two real data applications are presented. Our proposed spatial Bayesian predictive synthesis outperforms standard and state-of-the-art approaches in point and distributional predictions.

Transformed Elliptical Slice Sampler

Alberto Cabezas Tuesday, June 28

Lancaster University

A Bayesian model can be viewed as a collection of connected components, where the complexity of the overall posterior model, and the challenge of interfering the model's parameters, depends on which components are chosen. Replacing simple linear components with nonparametric terms, nonlinearities, or hierarchies, leads to more complex posterior geometries. This is usually alleviated using reparameterization techniques, marginalization, increasing prior information and/or getting more data, depending on the specific model at hand. These techniques can, however, impact the quality of the model, constrain the researcher to using a simpler model or force him to include arbitrary prior information. In this research we generalize the Elliptical slice sampler of Murray, et. al. (2010) to allow for general priors in the modelling as a method to sample with minimal tuning parameters from a general posterior distribution. Furthermore, we use reparameterizations derived from notions of optimal transport to alleviate the problem of complex geometries on these posterior distributions and allow the generalized elliptical slice sampler to efficiently sample from complex geometries. We study specific situations arising from funnel geometries in the case of hierarchies or Gaussian processes with scarce data and/or correlated length scales, arithmetically unstable tail geometries, geometries arising from highly correlated groups of parameters and multimodal geometries.

Controlling the flexibility of non-Gaussian models

Rafael Cabral Tuesday, June 28

KAUST

The normal inverse Gaussian (NIG) and generalized asymmetric Laplace (GAL) distributions can be seen as skewed and heavy-tailed extensions of the Gaussian distribution. Models driven by these more flexible noise distributions are then regarded as generalizations of simpler Gaussian models. Inferential procedures tend to overestimate the degree of non-Gaussianity in the data and we propose controlling the flexibility of these non-Gaussian models by adding sensible priors in the inferential framework that contract the model towards Gaussianity. The methods are derived for a generic class of non-Gaussian models that include non-Gaussian spatial Matérn fields, autoregressive models for time series, and simultaneous autoregressive models for aerial data. The results are illustrated with a simulation study, where priors that penalize model complexity were shown to lead to more robust estimation and give preference to the Gaussian model, while at the same time allowing for non-Gaussianity if there is sufficient evidence of asymmetry and leptokurtosis in the data. We also propose a new method for assessing if there is enough support in the data for choosing the non-Gaussian model over the simpler Gaussian model, and show how to determine which time points or regions in space the Gaussian model lacks flexibility.

Developing dynamic latent process methodology for high-dimensional biomarker data with correlated latent biological processes

Jiachen Cai Tuesday, June 28

MRC Biostatistics Unit, University of Cambridge

The increasing availability of high-dimensional biomarker measurements taken longitudinally can facilitate analysis of the biological mechanisms underlying disease and clustering of patients, as required for precision medicine. Existing approaches can only deal with part of the data structure but fail to jointly model all of them: specifically, Bayesian Latent Factor Analysis (BLFA) [Carvalho et al, 2008] can be used to uncover the latent structure, drastically reducing the dimension; whereas, treating the longitudinal factors as functional data, Dependent Gaussian Processes (DGP) constructed through kernel convolutions [Shi & Choi, 2011] may be appropriate for modeling time-dependent factor trajectories and correlation between factors simultaneously. We proposed an integrative model combining BLFA and DGP to address this gap, and developed an Empirical Bayes/Gibbs Sampler [Casella, 2001] for estimation and inference. We assessed the model performance in simulations, and applied it to longitudinal high-dimensional gene expression data as described in [Chen et al, 2011].

An Efficient Approach for Computation and Interpretation of Bayesian Credibility Models for Experience Rating

Sebastian Calcetero Tuesday, June 28

University of Toronta

Credibility models for insurance are mathematically intractable due to their complex structure, and therefore the calculation of credibility premiums must be obtained via simulations from the predictive distribution using MCMC. However, such simulations are computationally expensive and prohibitive for large portfolios. In addition, the computations end up being "black-box" procedures, as there is no clear expression to know how the observed experience is used to upgrade premiums. Here, we address these two challenges. At first, we propose an efficient simulation setup in which simulations are drawn from the prior distribution, instead of the posterior one. Secondly, we propose a methodology to estimate a closed-form credibility formula from which approximated Bayesian credibility premiums can be computed for any model, therefore allowing for practical interpretations of how the previous claim experience of a policyholder can be used to derive credibility premiums.

Bayesian Mediation Analysis Methods to Explore Racial Disparities in the diagnostic age of breast cancer

Wentao Cao Tuesday, June 28

LSU Health - New Orleans

A mediation effect refers to the effect transmitted by a mediator intervening the relationship between an exposure variable and a response variable. Mediation analysis is widely used to identify significant mediators and to make inference on mediation. Bayesian method allows researchers to incorporate prior information from previous knowledge into the analysis, deal with the hierarchical structure of variables, and estimate the quantities of interest from the posterior distributions. In this research, we propose three Bayesian mediation analysis methods to make inferences on the mediation effects. Through a series of simulations, we compare the accuracy and efficiency of the estimates of mediation effects using three Bayesian mediation methods. We apply the three Bayesian mediation methods to explore the racial disparity in the diagnostic age of breast cancer patients in Louisiana. The purpose of this study is to identify contributing factors in explaining the racial disparity in the diagnostic age of breast cancer from a large pool of potential risk factors.

BayesDLMfMRI: Bayesian Matrix-Variate Dynamic Linear Models for Task-based fMRI Modeling in R

Johnatan Cardona Jiménez Tuesday, June 28

Institución Universitaria Pascual Bravo

This work introduces a new R package for task-based fMRI data analysis to perform statistical analysis at individual and group levels. The analysis to detect brain activation at the individual level is based on modeling the fMRI signal using Matrix-Variate Dynamic Linear Models (MDLM). Therefore, the analysis for the group stage is based on posterior distributions of the state parameter obtained from the modeling at the individual level. In this way, this package offers several R functions with different algorithms to perform inference on the state parameter to assess brain activation for individual and group stages. Those functions allow for parallel computation when the analysis is performed for the entire brain and analysis at specific voxels when it is required.

Institución Universitaria Pascual Bravo

Mariana Carmona Baez Tuesday, June 28

McGill University

Quebec is one of the provinces in Canada that has been greatly affected by the COVID-19 pandemic. We have daily data by age group on the number of hospitalizations and ICU admissions due to COVID-19 in Quebec. We model the number of hospitalizations and ICU admissions, from March 2020 until October 2021, as following a bivariate dynamic generalized linear model. In particular, we assume that hospitalizations and ICU admissions, conditioned on a set of parameters, follow independent distributions. We focus on the evolution of both the risk of being admitted to the hospital given a positive test for COVID-19, and the probability of being admitted to the ICU. We use covariates such as the proportion of hospitalized males and the proportion of fully vaccinated people in Quebec. Our model provides estimates of the probability of hospitalizations and of the correlation between hospitalizations and ICU admissions.

Approximate General Bayesian Inference via Semiparametric Variational Bayes

Cristian Castiglione Tuesday, June 28

University of Padova

We present a new variational algorithm to approximate the general posterior distribution for Bayesian models that combine subjective prior beliefs with an empirical risk function. In particular, we consider the class of loss functions with a piecewise polynomial behavior, which includes support vector machines, quantile regression and expectile regression. Our iterative procedure lies in the class of semiparametric variational Bayes and enjoys closed-form updating formulas along with an analytic integration of the evidence lower bound. We require neither conjugacy nor elaborate data augmentation strategies. Structured prior distributions, e.g., cross-random effects, spatial-temporal processes, or inducing shrinkage priors, can be easily accommodated into such a framework without additional effort since the modularity of mean field variational Bayes is preserved. The properties of our algorithm are assessed through a simulation study, where we compare the proposed method with MCMC and MFVB both in terms of posterior approximation accuracy and prediction error.

Divide-and-Conquer Bayesian Fusion

Ryan Chan Tuesday, June 28

The Alan Turing Institute

Combining several (sample approximations of) distributions, which we term sub-posteriors, into a single distribution proportional to their product, is a common challenge. For instance, in distributed ‘big data’ problems, or when working under multi-party privacy constraints. Many existing approaches resort to approximating the individual sub-posteriors for practical necessity, then representing the resulting approximate posterior. The quality of the posterior approximation for these approaches is poor when the sub-posteriors fall out-with a narrow range of distributional form. Recently, a Fusion approach has been proposed which finds a direct and exact Monte Carlo approximation of the posterior (as opposed to the sub-posteriors), circumventing the drawbacks of approximate approaches.

Synergistic Interaction Modeling

Shounak Chattopadhyay Tuesday, June 28

Duke University

There is abundant interest in assessing the joint effects of multiple exposures on human health. This is often referred to as the mixtures problem in environmental epidemiology and toxicology. Classically, studies have examined the adverse health effects of different chemicals one at a time, but there is concern that certain chemicals may act together to amplify adverse health effects. Such amplification is referred to as synergistic interaction, while chemicals that inhibit each other's effects have antagonistic interactions. Current methods for assessing the health effects of chemical mixtures do not explicitly consider synergy or antagonism in the modeling, instead focusing on either parametric or unconstrained nonparametric dose response surface modeling. The parametric case can be too inflexible, while nonparametric methods face a curse of dimensionality that leads to overly wiggly and uninterpretable surface estimates. We propose a Bayesian method that decomposes the response surface into additive main effects and pairwise synergistic or antagonistic interactions, while providing a methodology for variable selection for each component. This Synergistic Interaction Modeling (SIM) framework is evaluated relative to existing approaches using simulation experiments and an application to data from NHANES.

Bayesian Inference for a Common Coefficient of Variation from Inverse Gaussian Distributions

Yogendra Chaubey Tuesday, June 28

Concordia University

In agricultural research, the coefficient of variation (CV) is used to measure the field heterogeneity and stability in genotype x environment interaction, where it could be common across fields or genotypes. The datasets in the associated experiments contain valuable information on parameters such as CV that can serve the purpose of obtaining prior information on these parameters (see Singh et al., 2015, Crop Science, 55(2), 501-513). This paper discusses the posterior distribution of the common CV from inverse Gaussian (IG) distributions when the prior distribution of the CV has been derived from some empirical distributions; the priors for means are assumed to be generalized inverse Gaussian (GIG) as studied in Chaubey et al. (2021, Applied Statistics and Data Science: Proceedings of Statistics 2021 Canada, New York: Springer, 97–114). We present the methods of evaluation and discuss the results of the application on multi-environment international yield trials in small-seeded lentils.

Modeling Neural Population Coordination via a Block Correlation Matrix

Yunran Chen Tuesday, June 28

Duke University

How neural population preserves multiple simultaneously presented stimuli is a fundamental question, but has not yet been fully understood by neuroscientists. Related studies suggest a single neuron may present a trial-wise multiplexing and a time division 'Mixture' dynamic under dual stimuli, where a single neuron may encode information by fluctuating between two stimuli from trials to trials. Here we are interested in whether or how such turn-taking activities may present a coordination pattern in neural population. We define a fluctuate weight to measure selection preference for a single neuron at a trial by a Poisson mixture model. To capture the neural population coordination, we introduce a block correlation matrix of fluctuating weights by a Gaussian copula model. We estimate the block correlation matrix with unknown cluster assignment in a Bayesian framework. Specifically, we consider a canonical representation of a block matrix to facilitate the prior specification and design a MCMC sampling scheme.

Fast Approximate Inference for Spatial Extreme Value Models

Meixi Chen Tuesday, June 28

University of Waterloo

The generalized extreme value (GEV) distribution is a popular model for analyzing and forecasting extreme weather data. To increase prediction accuracy, spatial information is often pooled via a latent Gaussian process on the GEV parameters. Inference for such hierarchical GEV models is typically carried out using Markov chain Monte Carlo (MCMC) methods. However, MCMC can be prohibitively slow and computationally intensive when the number of latent variables is moderate to large. In this paper, we develop a fast Bayesian inference method for spatial GEV models based on the Laplace approximation. Through simulation studies, we compare the speed and accuracy of our method to both MCMC and a more sophisticated but less flexible Bayesian approximation. A case study in forecasting extreme wind speeds is presented.

PDMP Monte Carlo methods for piecewise-smooth densities

Augustin Chevallier

Lancaster university

Tuesday, June 28

There has been substantial interest in developing Markov chain Monte Carlo algorithms based on piecewise-deterministic Markov processes. However existing algorithms can only be used if the target distribution of interest is differentiable everywhere. The key to adapting these algorithms so that they can sample from to densities with discontinuities is defining appropriate dynamics for the process when it hits a discontinuity. We present a simple condition for the transition of the process at a discontinuity which can be used to extend any existing sampler for smooth densities, and give specific choices for this transition which work with popular algorithms such as the Bouncy Particle Sampler, the Coordinate Sampler and the Zig-Zag Process. Our theoretical results extend and make rigorous arguments that have been presented previously, for instance constructing samplers for continuous densities restricted to a bounded domain, and we present a version of the Zig-Zag Process that can work in such a scenario. Our novel approach to deriving the invariant distribution of a piecewise-deterministic Markov process with boundaries may be of independent interest.

Approximate inference for stochastic kinetic models from multiple data sources for joint estimation of infection dynamics from aggregate reports and virological data

Oksana Chkrebtii

The Ohio State University

Tuesday, June 28

Acute respiratory infections (ARI) are infections of the upper and lower respiratory tract caused by multiple etiological agents. Prior to the current pandemic, influenza and respiratory syncytial virus (RSV) were the leading etiological agents of seasonal ARI around the world, and were largely diagnosed based on symptoms alone, without the use of virological tests necessary to identify individual viruses, limiting the ability to study the interaction between multiple pathogens and make public health recommendations. We develop an approximate marginal sampling approach based on the Linear Noise Approximation to fit a stochastic kinetic model (SKM) for a system of interacting ARI pathogens circulating in a large population to data on aggregate infection reports from six epidemic seasons collected by the state health department, and a subset of virological tests from a sentinel program at a general hospital in San Luis Potosi, Mexico.

Phylogeny from matricial datasets: application to sign languages history

Grégoire Clarté Tuesday, June 28

University of Helsinki

We propose a phylogenetic bayesian inference based on matricial datasets where lines and columns are correlated. We apply this model to sign language phylogenies, with a generative model describing jointly the phonological and lexical evolution of the language. The computational techniques include SMC algorithm with exotic tempering. The results on simulated data are satisfying, and on real data confirm some of the linguists' hypotheses while confronting others.

Hierarchical neutral to the right priors

Riccardo Cogo Tuesday, June 28

University of Milano-Bicocca

The Beta-Stacy process introduced by Walker and Muliere (1997) is a well-known Bayesian nonparametric prior for survival functions, which is typically used in presence of censored survival times. Such a process belongs to the more general class of neutral to the right priors, which are suitable for a single group of homogeneous, i.e. exchangeable, survival times. We aim at introducing a hierarchical version of neutral to the right priors, tailored for heterogeneous, though related, groups of survival times. In particular, we introduce a hierarchical version of the Beta-Stacy process, we investigate its properties and characterize its posterior distribution. We finally provide Bayesian estimators of the random survival functions highlighting the connections with well-known frequentist estimators. The model is applied to a real data example.

A semi-supervised Bayesian mixture modelling approach for joint batch correction and classification

Stephen Coleman Tuesday, June 28

University of Cambridge

Systematic differences between batches of samples present significant challenges when analysing biological data. Such batch effects are well-studied and are liable to occur in any setting where multiple batches are assayed. Many existing methods for accounting for these have focused on high-dimensional data such as RNA-seq and have assumptions that reflect this. Here we focus on batch-correction in low-dimensional classification problems. We propose a semi-supervised Bayesian classifier based on mixture models that jointly predicts class labels and models batch effects. Our model allows observations to be probabilistically assigned to classes in a way that incorporates uncertainty arising from batch effects. A simulation study demonstrates that our method performs well compared to popular off-the-shelf machine learning methods. We apply our model to two datasets collected in 2020 to measure seropositivity for SARS-CoV-2. We use our model to estimate seroprevalence in the populations studied.

Impacts of random effects misspecification on statistical Inference for spatial processes

Kiswendsida Julien Compaore Tuesday, June 28

Université Laval

Studies have highlighted the consequences of correlation structure misspecification in spatial processes modelling, nevertheless there are hardly any studies on the impacts of misspecification of spatial latent processes distribution. Our work is directed towards the latter aspect. Indeed, it is common in spatial statistics to make assumption of a Gaussian distribution of random effects. However, this hypothesis is not always verified, random effects can have a heavy tailed distribution. It is therefore necessary to diagnose how a misspecification of random effects distribution affects estimation and inference. This work is part of Fisheries Oceans Canada's fish stock assessment project. It is carried out in collaboration with Louis-Paul Rivest from Laval University (Université Laval) and Hugues Benoît from Fisheries Canada.

Posterior Representation of Compound Random Measures

Riccardo Corradin Tuesday, June 28

University of Nottingham

Dependent random measures have been studied extensively over the past decades. Among the possible way to define a vector of dependent random measures, a remarkable strategy is given by the family of compound random measures. We derived a posterior representation for compound random measures. Such a characterization is fundamental to define conditional sampling strategies, e.g., a Ferguson and Klass algorithm. We further embedded a vector of compound random measures in a nested structure. By normalizing these random measures, we can easily obtain a class of prior distributions that can be used to perform clustering among distributions and observations simultaneously, in a mixture framework. We further applied the proposed model to cluster the main towns in Lombardy, with respect to their distributions of the daily concentration of particulate matter.

Inducing high spatial correlation with randomly edge-weighted neighborhood graphs.

Danna Lesley Cruz Reyes Tuesday, June 28

Grupo de Investigación Clínica, Escuela de Medicina y Ciencias de la Salud, Universidad del Rosario, Bogotá, Colombia.

Traditional models for areal data assume a hierarchical structure where one of the components is the random effects that spatially correlate the areas. The conditional autoregressive (CAR) model is the most popular distribution to jointly model the prior uncertainty about these spatial random effects. One limitation of the CAR distribution is the inability of producing high correlations between neighboring areas. We propose a robust model for areal data that alleviates this problem. We represent the map by an undirected graph where the nodes are the areas and randomly-weighted edges connect nodes that are neighbors. The model is based on a multivariate Student-t distribution, spatially structured, in which the precision matrix is indirectly built assuming a multivariate distribution for the random edges effects. The edges effects' joint distribution is a spatial multivariate Student-t that induces another t distribution for the areas' spatial effects which inherit its capacity to accommodate outliers and heavy-tail behavior. Most important, it can produce a higher marginal correlation between the spatial effects than the CAR model overcoming one of the main limitations to this model. We fit the proposed model to analyze real cancer maps and compared its performance with several state-of-art competitors. Our proposed model provides better fitting in almost all cases.

Clustering activation patterns of spatially-referenced neurons

Laura D'Angelo Tuesday, June 28

University of Milano-Bicocca

Estimation of groups of co-activating neurons in calcium imaging studies is a challenging problem due to the need to deconvolve the activations and, then, to cluster the latent binary time series of activity. These time series describe, at each time point, the presence or absence of a spike, which corresponds to the active or resting state of neurons, respectively. We describe a nonparametric mixture model that allows for simultaneous deconvolution and clustering of binary time series based on common patterns of activity. The model makes use of a latent continuous process for the spike probabilities to identify groups of co-activating cells. Neurons' spatial dependence is introduced by informing the mixture weights with their location, following the common neuroscience assumption that neighboring neurons often activate together. The model's performance is illustrated on simulated data and a real dataset of hippocampal neurons.

Bayesian nonparametric change point detection for multivariate time series with missing observations

Luca Danese Tuesday, June 28

University of Milano-Bicocca

Time series are a commonly observed type of data. Among the possible analysis which can be performed on such a data, change points detection aims to describe patterns in the underlying structure of time-indexed observations. Within a Bayesian nonparametric framework, we propose a model to detect multiple change point in multivariate time series, by extending one of the main state-of-the-art approaches. The approach mainly relies on a model-based clustering approach to detect the change points, combined with a multivariate kernel for time-dependent realizations. Our studies are motivated by detecting possible changes among the compositions of COVID-19 daily new cases among different areas in Italy. Our findings are consistent with the breakdowns of the pandemic and the government's policies.

Fitting Structural Equation Models via Variational Approximations

Khue-Dung Dang Tuesday, June 28

The University of Melbourne

Structural equation models are commonly used to capture the relationship between sets of observed and unobservable variables. Traditionally these models are fitted using frequentist approaches but recently researchers and practitioners have developed increasing interest in Bayesian inference. In Bayesian settings, inference for these models is typically performed via Markov chain Monte Carlo methods, which may be computationally intensive for models with a large number of manifest variables or complex structures. Variational approximations can be a fast alternative; however, they have not been adequately explored for this class of models. We develop a mean field variational Bayes approach for fitting elemental structural equation models and demonstrate how bootstrap can considerably improve the variational approximation quality. We show that this variational approximation method can provide reliable inference while being significantly faster than Markov chain Monte Carlo.

Analysis of bias-variance trade-off in estimators for variational inferencing

Niladri Das

Sandia National Laboratories

Variational inference (VI) is a method that approximates probability densities through optimization. VI has been used in many applications and tends to be faster than classical methods, such as Markov chain Monte Carlo sampling. Traditionally, instead of KL divergence, the evidence lower bound is used as a cost function. We then optimize the parameters of a variational distribution. We explore the technicalities in using KL divergence in such a scenario, synthesizing estimators, and leveraging upon the delta method to study the bias-variance trade-off. Our objective is to study the gradient space of the KL divergence with respect to the variational parameter and to understand how we can tune the bias and variance by optimizing importance sampling. With lower variance in the gradient information, we can achieve larger optimization steps, lower the number of samples needed, which we compare against the ELBO based optimization.

A new way to model responses on (0,1): the Generalized Flexible Beta regression model

Ludovica De Carolis Tuesday, June 28

Università di Milano-Bicocca

This contribution addresses the issue of modeling responses on (0,1), such as scores or proportions, typical of many fields. To this end, a regression model based on the Generalized Flexible Beta (GFB) distribution is proposed. The Flexible Beta (FB) distribution is a special mixture of two betas that extends the shapes of the beta distribution while ensuring strong identifiability and likelihood boundedness. The GFB distribution generalizes the latter, allowing the maximum distance between the means of the two components to arbitrarily vary. Simulation studies and applications to real datasets show that the GFB regression model has a better fitting capacity than other models usually applied in the literature, under various data patterns (unimodal, bimodal, with heavy tails or outliers). Furthermore, the novel regression model can potentially fit data with better accuracy than the FB regression model, since it models the group regression means more flexibly while preserving easiness of interpretation.

Generalizing Impacts of Voluntary Interventions Using Bayesian Nonparametric Regressions

Irina Degtiar Tuesday, June 28

Mathematica Policy Research

Impact evaluation participants may not be representative of the population to whom the intervention may eventually be scaled, leading to uncertainty regarding the relevance of evaluation findings to future decisions. Generalizability is further challenged by voluntary policy scale-ups, as future volunteers are not enumerable. We present a novel approach for estimating target population average treatment effects among the treated (PATT) by generalizing results from an observational study to target population volunteers. Our approach accommodates flexible outcome regressions such as Bayesian Causal Forests (BCF) and Bayesian Additive Regression Trees (BART), and propagates uncertainty regarding who will volunteer into the posterior credible intervals to reflect the uncertainty of scaling up a voluntary intervention. In a simulation based on real data, we estimate the PATT of a national scale-up of a voluntary health policy model and demonstrate that nonparametric approaches (BCF and BART) improve performance over estimators that rely on parametric regressions.

Multivariate Hawkes Processes with Inhibition

Isabella Deutsch Tuesday, June 28

University of Edinburgh

Hawkes processes are point processes that model data where events occur in clusters through the self-exciting property of the intensity function. This model can be extended two ways. First we consider a multivariate setting where multiple processes can influence each other. Second, we modify the intensity function to allow for excitation and inhibition, both within and across processes. We present a summary of this multivariate Hawkes model where we focus on the specification of the influences in the intensity function that also allows for sparsity. We highlight difficulties in the estimation procedure caused by a potentially negative intensity function and provide approximate and exact solutions. Moreover, we leverage the notion of direct offspring events to examine the number of general offsprings and demonstrate its usefulness for prior predictive checks in the Bayesian workflow. These concepts are then showcased in a novel Bayesian application of Hawkes processes for fashion wholesale data.

Robust Bayesian inference using coarsened posteriors computed via data reweightings

Miheer Dewaskar Tuesday, June 28

Duke University

Bayesian inference typically assumes that the data generating density belongs to a chosen model class. However, particularly as sample size increases, even a small violation of this assumption can have a large impact on the outcome of the Bayesian procedure. To address this problem, Miller and Dunson (2018) introduce the coarsened posterior, where rather than condition on the data exactly, one conditions on a neighborhood of the empirical distribution of the observed data. Under suitable relative-entropy neighborhoods, the previous authors show that the coarsened posterior can be approximated by simply raising the likelihood to a fractional power. In this work, for neighborhoods of a fixed size based on integral-probability-metrics (IPM), we use Large Deviation theory to show that the coarsened posterior can be asymptotically approximated by solving a variational optimization problem at each point in the model. For an IPM like Maximum Mean Discrepancy, this allows us to construct a coarsened-likelihood that is robust to outliers in the data, and which is computed by solving convex-optimization to find optimal weights for the data points.

Joint Modeling of Sepsis Outcomes and Subtypes Using a Mixed-Data Grade of Membership Model

Alex Dombowsky Tuesday, June 28

Duke University

Identifying subtypes of life-threatening conditions such as sepsis is critical for designing effective therapies. Clinical data–consisting of readily measurable patient signs upon presentation to hospital–have been modeled in past studies for determining sepsis subtypes. While hard clustering is appealing in this setting, assigning each patient a cluster label may be too simplistic, as the pathophysiology and clinical features of sepsis can be heterogeneous and dynamic throughout the course of illness. This work proposes instead to use a joint model on clinical data, which simultaneously infers illness phenotypes using a Grade of Membership model and validates them by predicting patient outcomes. The model incorporates three common clinical data types: continuous, binary, and ordinal. We develop an MCMC algorithm for computation and provide code to implement it. The approach is applied to the Sepsis Characterization in Kilimanjaro study, which aims to define sepsis subtypes present in sub-Saharan Africa.

Coupled Markov Switching Count Models for the Detection and Forecasting of COVID-19 Outbreaks in Quebec Hospitals

Dirk Douwes-Schultz Tuesday, June 28

McGill University

The accurate detection and forecasting of COVID-19 outbreaks in the 30 largest hospitals in Quebec would greatly aid in COVID-19 planning within the province. We develop a novel Bayesian Markov switching model in which each hospital switches between outbreak and non-outbreak periods through a series of coupled nonhomogeneous hidden Markov chains. Unlike previous Markov switching approaches to outbreak detection/forecasting, the probability of entering the outbreak period is allowed to depend on space-time covariates, such as lagged COVID-19 cases, and the outbreak status of other hospitals previously, which allows the outbreaks to spread between hospitals. The effects of outbreak spread can change over space and time to account for differing levels of connectivity in the hospital network. Hospitalizations in each period are assumed to follow flexible autoregressive count processes and the model can also be used to forecast demand in the outbreak period to distinguish between small and large outbreaks.

A Bayesian Competition-Density Model of Forest Structure

Mark Ducey Tuesday, June 28

University of New Hampshire

Quantifying the upper frontier of tree size and density in forests is critical for understanding potential future carbon sequestration in forests, and for benchmarking current conditions and practices. Nearly all work on this problem has used frequentist approaches that require a large amount of species-specific data, including numerous samples close to the upper frontier. Such data are unavailable for many species in regions with complex forests. An alternative model, Sterba's competition-density framework, allows use of samples away from the frontier, but it has four free parameters that lack obvious interpretation. A Bayesian approach seems promising. Reparameterization of the model allows more intuitive specification of priors, including informative priors, and it also suggests possible simplifications to the model. Application is illustrated with historic data on hemlock (Tsuga canadensis), an ecologically-important species currently threatened by an invasive insect.

Using Variational Bayes to correct Laplace Approximations

Shourya Dutta Tuesday, June 28

King Abdullah University of Science and Technology (KAUST)

Variational Bayes (VB) and Laplace Approximations (LA) are two well known remedies to provide approximate Bayesian inference. In the poster, we will discuss how to combine these two approaches so that VB can act as a correction to the initial estimates provided by the LA. The justification for our two-step approach follows from the Zellner (1988) variational formulation of Bayes theorem as the optimal information processing rule. I will present some examples within the class of Latent Gaussian models for which the LA provides reasonably good estimates but the VB corrections are shown to enhance the estimates without adding essentially any computational burden.

Approximate Bayesian Computation with Path Signatures

Joel Dyer Tuesday, June 28

University of Oxford

Simulation models of scientific interest often lack a tractable likelihood function, precluding standard likelihood-based statistical inference. A popular likelihood-free method for inferring simulator parameters is approximate Bayesian computation, where an approximate posterior is sampled by comparing simulator output and observed data. However, effective measures of closeness between simulated and observed data are generally difficult to construct, particularly for time series data which are often high-dimensional and structurally complex. Existing approaches typically involve manually constructing summary statistics, requiring substantial domain expertise and experimentation, or rely on unrealistic assumptions such as iid data. Others are inappropriate in more complex settings like multivariate or irregularly sampled time series data. In this paper, we introduce the use of path signatures as a natural candidate feature set for constructing distances between time series data for use in approximate Bayesian computation algorithms. Our experiments show that such an approach can generate more accurate approximate Bayesian posteriors than existing techniques for time series models.

Spatial Power Prior Elicitation: a new approach for Brain TMS

Osafu Augustine Egbon Tuesday, June 28

University of São Paulo, Brazil

This study proposed a spatial power prior distribution for spatial mapping and hotspot findings in spatial data. The proposed prior distribution mines information from finitely available historical spatial data to elicit informative prior knowledge in the current spatial mapping task. It takes the relationship between the current and historical data into account, technically preventing the historical data from overshadowing the current inference. Spatial projection over the current study site was achieved using defined basis functions constructed with a finite element method given a triangular mesh representing the spatial domain. The proposed framework was applied to transcranial magnetic stimulation data to find the hotspot on a patient's scalp that produces the highest motor evoked potential.

Joint cohort and predictive modelling

Samuel Emerson Tuesday, June 28

Durham University

Bayesian logistic regression models are common for binary response data, although in many circumstances the predictive performance and interpretability can be improved with multiple logistic regression models on some partition of the data. These partitions represent natural clusters in the covariate space where the model may systematically differ. For example, in health modelling settings there may be natural patient cohorts, where interest would often lie in any differences each model reveals between cohorts. We propose a method to jointly find these cohorts and fit the Bayesian model by constructing a graph in covariate space, which is explored by a scheme proposing cuts that form cohorts. A sequential Monte Carlo sampler for the model marginal enables efficient growth and shrinkage of cohorts, making alternatives to logistic regression an easy extension of this work. There are links to related methods such as mixture of experts models and model based clustering.

A Bayesian approach for clustering and exact finite-sample model selection in longitudinal data mixtures

Elena Erosheva Tuesday, June 28

University of Washington

We consider mixtures of longitudinal trajectories, where one trajectory contains individual-specific measurements over time of the variable of interest and each individual belongs to one cluster. The number of clusters as well as individual cluster memberships are unknown and must be inferred. We propose an original Bayesian clustering framework that allows us to obtain an exact finite-sample model selection criterion. Our approach is more flexible and parsimonious than asymptotic alternatives such as Bayesian Information Criterion (BIC) or Integrated Classification Likelihood (ICL) criterion in the choice of the number of clusters. Moreover, our approach has other desirable qualities: i) it keeps the computational effort of the clustering algorithm under control and ii) it generalizes to several families of regression mixture models, from linear to purely non-parametric.

Machine Learning Techniques of Analyzing Spectroscopy Data for Age Grading of Mosquito

Md Manzur Farazi Tuesday, June 28

Marquette University

Older mosquitoes have higher chances of carrying infectious parasites that cause malaria. Hence, mosquitoes' age is very important to understand the capability of the mosquito to spread diseases, for assessing the risk of breaking out a mosquito-borne disease near future, and for evaluating the effectiveness of mosquito control interventions. We develop a statistical model based Near-Infrared (NIR) spectroscopy that measures the amount of light absorbed by the head or thorax of mosquitoes at different wavelengths is expected to be easier and quick. Machine learning tools such as partial least squares regression (PLSR) and Neural Network has been used widely in literature to forecast age and age categories. However, these methods do not appropriately take into account the physiological changes mosquito go though as they age. This change point has been well established in medical literature. In this study, we proposed a change-point statistical model. We, first, use some pre-processing techniques on NIRS data and develop a new algorithm to predict age. Change-point technique is applied to estimate age from spectra using PLSR model. We compare the result with standard machine learning methods. The change-point PLSR model shows better performance in terms of age estimation of the mosquitoes.

A Bayesian Semiparametric Approach to Treatment Effect Variation with Noncompliance

Jared Fisher Tuesday, June 28

Brigham Young University

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 Frequentist and ML-based methods are quite flexible but are highly sensitive to the so-called weak instruments problem, in which the compliance rate is (locally) close to zero, and require pre-specifying subgroups of interest. Parametric Bayesian approaches, which account for noncompliance via imputation, are more robust in this case, but are much more sensitive to model specification. In this paper, we propose a Bayesian semiparametric approach that combines the best features of both approaches. Our main contribution is to embed Bayesian Additive Regression Trees (BART) in a broader Bayesian noncompliance framework in which we repeatedly impute individuals' compliance types. This allows us to flexibly estimate varying treatment effects among compliers while mitigating the weak instruments problem. We then apply our method to the Oregon health insurance experiment and show that analyses that only focus on a single source of variation can miss important heterogeneity.

clusterBMA: Bayesian Model Averaging for Clustering

Owen Forbes Tuesday, June 28

Queensland University of Technology

To combine inference across multiple sets of results for unsupervised clustering, Bayesian Model Averaging (BMA) offers some attractive benefits over other existing approaches. Benefits include intuitive probabilistic interpretation of an overall cluster structure integrated across multiple sets of clustering results, with quantification of model-based uncertainty.

BMA has been applied in the clustering context for averaging multiple finite mixture models, using the Bayesian Information Criterion (BIC) to approximate posterior model probability for weighted averaging across selected models. In this work we propose an extension to BMA methodology to enable weighted model averaging across results from multiple different clustering algorithms, using a combination of clustering internal validation criteria in place of the BIC to weight results from each model. From a combined posterior similarity matrix representing a weighted average of the clustering solutions across models, we apply symmetric simplex matrix factorisation to calculate final probabilistic cluster allocations. We present results from a case study and simulation study to explore the utility of this technique for identifying robust integrated clusters. This method is implemented in an accompanying R package, "clusterBMA".

The Bernstein-von Mises theorem for semiparametric mixtures

Stefan Franssen Tuesday, June 28

TU Delft

Many semiparametric models can be reformulated as a mixture model. Notable examples are the symmetric location mixtures, exponential frailty and errors-in-variables. In a mixture model, the mixture density is defined, given a kernel $p_\theta(x|z)$ and a mixing distribution F, by:

$$p_{\theta,F}(x) := \int p_\theta(x|z)dF(z).$$

We focus on the semiparametric inference of the parameter of interest θ. Following Bayesian methodology, we equip θ and F with priors and then study its frequentist properties: we prove a semiparametric Bernstein-von Mises theorem which shows that, under conditions, the posterior distribution is asymptotically normal and efficient for the parameter θ. When the mixing distribution F is equipped with a Dirichlet process prior, we can verify these conditions in the exponential frailty and the symmetric location mixture models. Extending the result to errors-in-variables is a work in progress. Joint work with Minh-Lien Jeanne Nguyen and Aad van der Vaart.

Scalable automated Bayesian inference for databases with InferenceQL

Cameron Freer Tuesday, June 28

Massachusetts Institute of Technology

InferenceQL is a probabilistic programming framework for scalable automated Bayesian inference from database tables. InferenceQL is designed to help make Bayesian approaches to data analysis and predictive modeling accessible to a much broader audience, and to assist experts in auditing and improving quality of data, models, and inferences.

Unlike Stan and Gen, InferenceQL provides automation for learning models online, via non-parametric Bayesian structure learning of probabilistic programs. Experts can override these models with custom learning algorithms and custom probabilistic programs for specific subsets of variables and conditional distributions. For a broad class of models, InferenceQL also automates calculation of exact conditional probabilities, marginal probabilities, and conditional mutual informations, as well as generation of posterior samples.

Finally, InferenceQL broadens access to rigorous Bayesian approaches to model criticism (via posterior predictive checking), data quality screening (via conditional probability calculation), fairness auditing (via conditional probability ratios), and synthetic proxy data generation (as a privacy enhancing technology). This is accomplished via constructs that interleave ordinary database queries with Bayesian inference.

Exploring the determinants of aortic stenosis treatments across geographical regions in Quebec

Jingyan Fu Tuesday, June 28

McGill University

First introduced in 2002 and standardized two years later, the trans-catheter aortic valve replacement (TAVR) is a new treatment to cure aortic stenosis (AS) without opening the chest, resulting in lower mortality rates. However, only a limited number of cardio centers in hospitals are capable of this costly operation, and the number of TAVR operations undergone is not enough compared to the AS patients diagnosed each year. Therefore, it is essential to ensure that the treatment decision of aortic stenosis is made transparently. In this research, we intended to find if there existed any influential regional or social-economic factor that made a potential AS patient more likely to receive TAVR operation. After studying the baseline characteristics, we used a latent spatial effect to model the regional difference with variance depending on the contiguous neighbourhood structure. We compared it to another model that incorporated the spatial information through the shortest distance between patients and TAVR operation centers. Due to the multi-collinearity among predictors, we applied the regularized horseshoe prior to the logistic regression, including personal, social-economic, and clinical variables, and drew the posterior samples through Hamilton Monte Carlo. Based on the estimates, female patients and patients with severe diseases had a higher chance of receiving TAVR than patients in metropolitan areas. Though not significant, signs of inequality existed among patients with social-economic burdens, making them the new treatment's least favourite.

Bayesian adaptive lasso for the multidimensional four-parameter logistic item response model

Zhihui Fu Tuesday, June 28

Minnan Normal University

The multidimensional four-parameter logistic (M4PL) item response model, which includes an upper asymptote for the correct response probability, has drawn increasing interest due to its suitability for many practical scenarios. Given that it involves latent variables and dichotomous responses, which induce an intractable likelihood function, the likelihood-based methods are not directly applicable. On the contrary, the sampling-based Bayesian approach is feasible to provide an efficient and reliable analysis for the proposed joint model. In this presentation, based on a data augmentation scheme using the Gibbs sampler, we develop a Bayesian adaptive lasso procedure to conduct simultaneous estimation and variable selection. Nice features including empirical performance of the proposed methodology are demonstrated by simulation studies.

Joint Bayesian Inference of Phylogeny and Mutation Order in Cancer

Yuan Gao Tuesday, June 28

Allen Institute

Although the role of evolutionary process in cancer progression is widely accepted, increasing attention is being given to the evolutionary mechanisms that can lead to different clinical outcomes. Recent studies suggest that the temporal order in which somatic mutations accumulate during cancer progression is important. Single-cell sequencing provides a unique opportunity to examine the effect of mutation order during cancer progression. However, the errors associated with single-cell sequencing complicate this task. We propose a novel Bayesian method for jointly inferring the cancer phylogeny and the mutation order using noisy single-cell sequencing data that incorporates the errors from the data collection process. Through analyses of simulations and of real data from cancer patients, we demonstrate that our method outperforms existing models.

Detecting Renewal States in Chains of Variable Length via Intrinsic Bayes Factors

Nancy Gaqrcia Tuesday, June 28

University of Campinas

Markov chains with variable length are useful parsimonious stochastic models able to generate most stationary sequence of discrete symbols. The idea is to identify the suffixes of the past, called contexts, that are relevant to predict the future symbol. Sometimes a single state is a context, and looking at the past and finding this specific state makes the further past irrelevant. These states are called renewal states and they split the chain into independent blocks. In order to identify renewal states for chains with variable length, we propose the use of Intrinsic Bayes Factor to evaluate the plausibility of each set of renewal states. In this case, the difficulty lies in finding the marginal posterior distribution for the random context trees for general prior distribution on the space of context trees and Dirichlet prior for the transition probabilities. To show the strength of our method, we analyzed artificial datasets generated from two binary models models and one example coming from the field of Linguistics.

A Bayesian joint modelling approach to analyzing irregular longitudinal data from electronic health records: leveraging recommended visit intervals

Rose Garrett Tuesday, June 28

University of Toronto; the Hospital for Sick Children

Using routinely collected data from electronic health records offers a low-cost approach to investigating disease progression over multiple years of follow-up. However, this study design can lead to a biased sample since patients interact with the healthcare system more often when they are unwell and thus there is an overrepresentation of measurements on sicker patients. Although several analytic methods have been established to account for this, Bayesian approaches have been limited. We formulate a Bayesian joint model of the disease outcome and irregular visit processes that relaxes some of the restrictive and often unrealistic simplifying assumptions about the visit process made by existing methods. To achieve this, we leverage information that has never been used before but is often already recorded in patient charts: physician recommendations on when the next visit should occur. We illustrate our approach using data from a clinic-based cohort of patients with juvenile dermatomyositis.

Linear models with assumptions-free residuals: a Bayesian Nonparametric approach.

Valentina Ghidini Tuesday, June 28

Bocconi University

Linear models are ubiquitously employed, thanks to their simplicity and all their possible generalized versions; however, assumptions on the residuals may be limiting, especially if the model is somehow misspecified. For this reason, this work proposes a Bayesian nonparametric approach to relax some common assumptions on errors (such as homoskedasticity and symmetry), exploiting a nonparametric prior on the space of distributions of residuals. In particular, we design a methodology aimed at creating a much more flexible model, yet retaining some desirable parametric properties (such as interpretability). This modus operandi is general, and can be applied to any model which is hypothesizing some residual distribution (for example linear regression, AutoRegressive models, Gaussian processes). Our proposal also naturally presents some useful clustering properties which allow us to unveil hidden patterns in the data and to automatically recognize outliers.

Bayesian clustering of high-dimensional data via latent repulsive mixtures

Lorenzo Ghilotti Tuesday, June 28

University of Milan-Bicocca

We address the problem of clustering high-dimensional data. To this end, we propose a factor-analytic model whereby the clustering is performed at the latent factors level, similarly to the Lamb model in Chandra et al. (2020). A matrix of loadings links the latent factors and the data. Our proposal differs from the Lamb model in the prior for the cluster centers, that we assume to be an anisotropic repulsive point process. Anisotropy ensures that separation is induced between the high-dimensional centers of different clusters. In particular, we propose an anisotropic determinantal point process, since it guarantees analytical availability of its density, crucial for sampling purposes. We also design an efficient Metropolis-within-Gibbs sampling algorithm. The update of the matrix of loadings is problematic. Therefore, we suggest a derivative-based Metropolis-Hastings step, obtaining an analytical expression of such derivative. Finally, the methodology is compared to alternative models in the literature.

A Bayesian Method for Biclustering Multivariate Ordinal Data with Informative Censoring

Alice Giampino Tuesday, June 28

Università di Milano-Bicocca

Multivariate ordinal data are extensively analyzed with several clustering approaches. These algorithms typically consider missing values as absent information, rather than a valuable element itself for profiling consumers' preferences. In this work, we propose a Bayesian nonparametric model for biclustering multivariate ordinal data. As a distinctive contribution, our approach considers censored observations as informative. The ordinal nature of the data is handled by introducing latent variables. Our model exploits the flexibility of two independent Dirichlet processes, allowing us to make inference on the number of clusters characterizing the latent structure. The large dimensionality is tackled by specifying a matrix factorization model. The conjugate specification of the model allows for an explicit derivation of the full conditional distributions of all the random elements in the model. Posterior inference is then carried out by means of Gibbs sampling algorithm. The performance of the method is illustrated by analyzing movie rating data.

Flexible spatial modeling of areal data: Introducing the Hausdorff-Gaussian Spatial Process

Lucas Godoy Tuesday, June 28

University of Connecticut

Accounting for spatial dependence when analyzing areal data is extremely important for efficient and valid statistical inferences. Most existing models for areal data employ adjacency matrices to quantify the spatial dependence structure, where spatial polygons of different shapes and sizes are treated the same way. Such methodologies impose some limitations. Remarkably, computing predictions in different maps may become impractical. We propose a flexible model for area data from a geostatistical perspective with a Gaussian process defined based on the Hausdorff distance instead of the Euclidean distance to circumvent these limitations. We present the benefits of the proposed method in Bayesian spatial modeling comparing its performance to popular spatial models via a simulation study. The model is used to fit the average income at the census tracts in a Brazilian city called Barbacena. Our methodology has performed better (WAIC = 409) than a CAR model (WAIC = 433.1).

Bayesian change point models for cancer incidence trends

Lovedeep Gondara Tuesday, June 28

Simon Fraser University

Change point models are often used in cancer surveillance to show recent trends in cancer incidence/mortality (such as by using joinpoint software from National Cancer Institute). However, the focus has been on the Frequentist paradigm, in this poster we show that similar can be accomplished in the Bayesian paradigm, which offers advantage over the Frequentist analysis of change point modelling.

Structure induced by a multiple membership transformation on the conditional autoregressive model

Marco Gramatica Tuesday, June 28

Queen Mary University of London

We investigate the theoretical properties of a new method for spatially misaligned areal data. The usual context for disease mapping is to model data aggregated at the areal level. When data is available at different (misaligned) resolution and where no spatial framework is available, it is possible to specify relative risks by using the multiple membership principle (MM). Using a weighted average of conditional autoregressive (CAR) spatial random effects is possible to embed spatial information for a misaligned outcome and estimate relative risks for both frameworks. We investigate the theoretical properties of the application of the MM principle to the CAR prior in terms of its parameterisation, properness and identifiability. Simulation results show that overall posterior samples are well calibrated for both frameworks across all simulation scenarios. The spatial MM modelling strategy is illustrated by an application to diabetes prevalence data for GP practices in South London.

Zero-inflated hierarchical generalized Dirichlet multinomial Bayesian regression model with cyclic splines for analysis of TDP-43 on the ALS-FTD spectrum

Patrick Gravelle Tuesday, June 28

Brown University

Amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD) are neurodegenerative diseases that are characterized by motor function impairment and by executive, cognitive, and behavioral symptoms, respectively. The ALS-FTD disease spectrum can be characterized by the RNA-binding protein, TDP-43. Defining the role of TDP-43 in the neurodegeneration process may help lead to treatment development of ALS-FTD.

Our study considers mouse behavior continuously recorded over 5 days. Mice were categorized into the control (wild-type, WT) or TDP-43 mutation groups. Our goal was to gain understanding for the time spent performing each behavior of the two groups. We investigated the appropriateness of Bayesian ANOVA methods and hierarchical normal Bayesian models. However, due to poor model fits as a result of inappropriate assumptions of the data, we extended our analysis to hierarchical non-normal Bayesian models.

We developed a sophisticated model to appropriately and accurately model the time mice spend performing a set of 9 behaviors. The zero-inflated hierarchical generalized Dirichlet multinomial (ZIHGDM) regression model with cyclic splines captures the relationships and characteristics of the data that our preliminary models were unable to do.

The ZIHGDM model performed better than all previous models through the posterior predictive checks. These checks indicated that the ZIHGDM model was able to accurately predict observed values at nearly every hour of the day, for every behavior, across all 5 test statistics, for both TDP-43 and WT mice, with the greatest precision of all models considered. This lead to the identification of several phenotypes expressions unique to TDP-43 mice which will aid future ALS-FTD research.

Multivariate Nearest-Neighbors Gaussian Processes with Random Covariance Matrices

Isabelle Grenier Tuesday, June 28

University of California, Santa Cruz

Computational efficiency is at the forefront of many cutting edge spatial modeling techniques. Non-stationarity, on the other hand, has often been an unintentional feature of the approximations used in these spatial models. Deriving from the well known multivariate linear regression model, we propose a non-stationary and non-isotropic spatial model. In order to remain relevant with today's massive datasets challenges, we apply the concept of nearest-neighbors to our normal-inverse-Wishart framework. The model, called Nearest-Neighbor Gaussian Process with Random Covariance matrices (NN-RCM) is developed for both univariate and multivariate spatial settings and allow for specific characteristics such as duplicate observations and missing data. The model is illustrated in a case study of albedo assessments over CONUS from the Geostationary Operational Environmental Satellites (GOES) East and West. We apply the bivariate NN-RCM model using each satellite as a source of information. The objective is to merge the albedo assessments while also quantifying the discrepancy between the sources.

Spying on the prior of the number of data clusters and the partition distribution in Bayesian cluster analysis

Jan Greve Tuesday, June 28

WU Vienna University of Economics and Business

In applied Bayesian Model-Based clustering, it is common to deal with problems where the number of clusters is unknown. Bayesian mixture models employed in such applications usually specify a flexible prior that takes into account the uncertainty with respect to the number of clusters. However, a major empirical challenge involving the use of these models is in the characterisation of the induced prior on the partitions. This work introduces an approach to compute descriptive statistics of the prior on the partitions for three selected Bayesian mixture models developed in the areas of Bayesian finite mixtures and Bayesian nonparametrics. The proposed methodology involves computationally efficient enumeration of the prior on the number of clusters in-sample (termed as 'data clusters') and determining the first two prior moments of symmetric additive statistics characterising the partitions. The accompanying reference implementation is made available in the R package fipp.

Forecasting macroeconomic data with Bayesian VARs: Sparse or dense? It depends!

Luis Gruber Tuesday, June 28

University of Klagenfurt

Vectorautogressions (VARs) are widely applied when it comes to modeling and forecasting macroeconomic variables. In high dimensions they are prone to overfitting. Bayesian methods, more concretely shrinking priors, have shown to be successful in curing the curse of dimensionality. In the present paper we introduce the recently developed R^2-induced Dirichlet-decomposition prior to the VAR framework and compare it to refinements of well-known priors in the VAR literature: Hierarchical Minnesota prior, Stochastic Search Variable Selection prior and Dirichlet-Laplace prior. We demonstrate the virtues of the proposed prior in an extensive simulation study and in an empirical application forecasting data of the US economy. Further we shed more light on the ongoing Illusion of Sparsity debate. We find that forecasting performances under sparse/dense priors vary across evaluated economic variables and across time frames; dynamic model averaging, however, can combine the merits of both worlds. All priors are implemented using the reduced-form VAR and all models feature stochastic volatility in the variance-covariance matrix.

Mixture of Factor Analysers and Adaptive Telescoping Sampler for Automatic Inference on Number of Clusters and Factors

Margarita Grushanina Tuesday, June 28

Vienna University of Economics and Business

Factor analyzers are a common tool for finding structure in high-dimensional data. While in most applications the number of clusters and latent factors within clusters is fixed in advance, in recent literature models with automatic reference on both the number of clusters and latent factors have been introduced.

The current work contributes to the literature on mixtures of infinite factor analyzers (MIFA) with automatic inference of the number of clusters and the number of latent cluster-specific factors. The model employs a dynamic mixture of finite mixtures (MFM) and a telescoping sampler for automatic inference on the cluster structure of the data. Automatic inference on the cluster-specific factors within each cluster is performed via the cumulative shrinkage process (CUSP) prior and an adaptive Gibbs sampler. The performance of the model is demonstrated on several benchmark data sets as well as the Eurozone inflation rates data.

A scalable algorithm for incorporating external judgement in Bayesian forecasts via entropic tilting

Rishab Guha Tuesday, June 28

Amazon

We present a procedure which builds on the technique of entropic tilting in order to incorporate external judgement into forecasts produced by Bayesian state-space models. Our method avoids the problems of scalability and lack of support associated with classical implementations of entropic tilting, and provides improved interpretability relative to hard / soft conditioning.

Half Logistic-Truncated Exponential Distribution: Classical and Bayesian Estimation Using Different Approaches

Ahtasham Gul Tuesday, June 28

Pakistan Bureau of Statistics

Gul and Mohsin (2020) developed a new modified form of renowned ""Half logistic"" distribution introduced by Balakrishnan (1991) and named it Half Logistic-Truncated Exponential Distribution (HL-TEXPD). In this paper different approximation methods with squared error loss function (SELF) have been used to develop the bayes estimators for complete sample of the HL-TEXPD for unknown parameter. In classical setup, Monte Carlo simulations are performed using the maximum likelihood estimates. Similarly, Metropolis–Hastings algorithm has been used to simulate sample from the posterior densities using informative priors as well as non-informative priors of the parameter under different situations. To illustrate the usefulness and goodness of fit of HL-TEXPD distribution, we considered four real data sets. Additionally, TTT (total time on test) plot is drawn to study the failure rate of the three data sets.

Wednesday Session

Bayesian flexible models for the analysis of factorial experimental designs

Luis Gutierrez Wednesday, June 29

Pontificia Universidad Catolica de Chile

This work presents flexible models for the analysis of factorial experimental designs. The literature for analyzing factorial experiments is antique and based on restrictive assumptions. Hence, available models are inappropriate to analyze modern and diverse datasets correctly. We propose to use Bayesian nonparametric inference and elements from the model selection literature to develop a flexible model including the associated hypotheses tests. Specifically, our strategy considers the definition of latent binary variables, which identify each experiment with different random distributions. The random distributions are modeled via infinite mixture models. In our specification, the hypothesis space directly maps the binary variables. Then, the prior distribution on the hypothesis space is defined with a distribution over the binary vector. We relaxed the classical analysis of variance assumptions and produced models for data in different supports. The posterior consistency of the associated Bayes factor is studied. All the models are implemented in the R-package factorialDDP.

Bayesian Sparse Vector Autoregressive Approximate Hidden Semi-Markov Models

Beniamino Hadj-Amar Wednesday, June 29

Rice University

In neural data applications, it is of great interest to estimate hidden states of the brain as well as temporal and spatial dependencies between locations within these hidden regimes. We propose a sparse vector autoregressive (VAR) hidden semi-Markov model (HSMM) for modelling temporal and contemporaneous (e.g. spatial) dependencies in high-dimensional nonstationary time series. The HSMM's generic state distribution is embedded in a special transition matrix structure, facilitating efficient likelihood evaluations and arbitrary approximation accuracy. We deploy an l1-ball projection prior, which combines differentiability with a positive probability of obtaining exact 0's to conduct variable selection within each different switching regime and facilitate posterior estimation via HMC. Our proposed approach is able to identify both the dynamic properties and stationary structure of multiple signals by simultaneously identifying the active components of the switching VAR matrices as well as the non-zero elements of the regime-specific covariances, while flexibly modelling the general hidden state dwell distribution. We illustrate the performances of our proposed approach via several simulation studies and a data application in neuroscience.

Curvature Process: Directional Concavity in Gaussian Random Fields

Aritra Halder Wednesday, June 29

University of Virginia

Spatial process models offer a rich framework for modeling dependence in response variables arising from diverse scientific domains. Analyzing properties of the resulting surface generated through such modeling provides deeper insights into the nature of latent dependence within the studied response. This manuscript contributes toward a Bayesian inferential framework for a novel stochastic process, termed as the directional curvature process, that quantifies variability in local surface geometry for random surfaces. The resulting distribution theory produces inference on trajectories or curves over the domain that track rapid changes on the response surface with the aim of assessing directional curvature. Such trajectories or curves of rapid change, often referred to as *wombling* boundaries, occur in geographic space in the form of rivers in a flood plain, roads, mountains or plateaus or other topographic features considered instrumental in effecting high gradients. We demonstrate fully model based Bayesian inference on directional curvature processes to analyze differential behavior in responses along wombling boundaries. We illustrate our methodology with a number of simulated experiments followed by multiple applications featuring the Boston Housing data; Meuse river data; and temperature data from the Northeastern United States.

Advancements in Characterizing Warhead Fragmentation

John Haman Wednesday, June 29

Institute for Defense Analyses

Fragmentation analysis is a critical piece of the evaluation of warheads. But the traditional methods for data collection are expensive and laborious. Optical tracking technology is promising to increase the fidelity of fragmentation data, and decrease the time and costs associated with data collection. However, the new data will be complex, three dimensional 'fragmentation clouds'. This raises questions about how testers can effectively summarize spatial data to draw conclusions for sponsors. In this poster, we will present the Bayesian spatial models that are fast and effective for characterizing the patterns in fragmentation data, along with several exploratory data analysis techniques that help us make sense of the data. Our analytic goals are to

– Produce simple statistics and visuals that help the live fire analyst compare and contrast warhead fragmentations;

– Characterize important performance attributes or confirm design/spec compliance; and

– Provide data methods that ensure higher fidelity data collection translates to higher fidelity modeling and simulation down the line.

Lagrangian manifold Monte Carlo on Monge patch

Marcelo Hartmann Wednesday, June 29

University of Helsinki

MCMC methods for Bayesian analysis are the gold-standard stochastic approximations to recover the posterior distribution necessary for statistical inference. Among many methods proposed in the literature, the Hamiltonian Monte Carlo (HMC) has been the method of choice for practicioners but it may fail to explore posterior probability measures with sharp corners and strong curvatures. The Riemann manifold Hamiltonian Monte Carlo (RMHMC) and the Lagrangian Monte Carlo (LMC) circumvent this issues by accounting the curvature of the posterior using a metric-tensor G based on the Fisher-information matrix. Is this work we proposed a Riemann embedding which renders a model-independent metric-tensor that has closed-form inverse and determinant which scales down the computational cost of the numerical integrator while still preserving the quality of the samples proposed by the geometric MCMC algorithm.

Use of Approximate Bayesian Computation for Partial Discharge Analysis

Kai Hencken Wednesday, June 29

ABB Schweiz AG

Partial Discharge are short breakdown inside electrical equipment. As they indicate weaknesses of the insulation strength, they are seen as important precursors to a failure of the system. Therefore the measurement and analysis of the patterns of instances in time and strength of the discharge is an important tool, that has been addressed already using different methods in the past. In this work we explore how a combination of a physics-based stochastic process can be combined within Approximate Bayesian Computation as a new way to analyze them. Especially the ABC-SMC method was found to be useful. Different summary statistics were explored and real Partial Discharge measurement data was used not only for parameter estimation, but also to do model comparison in order to compare different physical models proposed in the literature.

Balanced tree stick-breaking priors for covariate-dependent mixture models

Akira Horiguchi Wednesday, June 29

Duke University

Stick-breaking priors are often adopted in Bayesian nonparametric mixture models for generating the mixture weights. When covariates influence the sizes of the clusters, stick-breaking mixtures can leverage Pólya-gamma data augmentation to ease the demands of posterior computation. However, despite this computational convenience, we show that in the presence of covariates, existing stick-breaking processes in fact often induce artificial correlation among the class of random measures they model, and introduce excessive posterior uncertainty into inference on the covariates effects on all but the largest of the clusters, which propagates into posterior uncertainty on the cluster sizes and the number of clusters. We believe that these phenomena arise because existing stick-breaking models are constructed based on continually breaking a single remaining piece of a unit stick. This design results in an underlying one-sided unbalanced tree structure, which in turn leads to an inefficient representation of covariate effects that confounds with the stochastic ordering of cluster sizes. Instead, we propose to generalize the stick-breaking priors to allow the continual breaking of all remaining pieces of a unit stick at each stage of the stick breaking. This corresponds to a balanced tree along which random weights are generated. When covariates are incorporated, this model decouples the stochastic ordering of the cluster sizes from that of the effect of covariates, thereby circumventing the challenges faced by existing stick breaking models. The computation under this new covariate-dependent model, however, remains essentially identical to that of existing stick-breaking mixtures. We demonstrate through both simulated and a data set from flow cytometry that the new model induces more robust posterior inference for both the number and size of clusters as well as the covariate effects on the cluster sizes.

On Statistical Inference in Factor Analysis: Polynomial-Time Verification of the Anderson-Rubin Condition

Darjus Hosszijni Wednesday, June 29

WU Vienna

In sparse factor models, identifiability of the variance decomposition has received little attention perhaps due to its difficulty. Anderson and Rubin (1956) famously establish identifiability under a rank assumption on a large number of submatrices of the factor loading matrix. This number is potentially exponential in the input dimensions and identifiability is therefore infeasible to verify for computers. In our paper, the computational complexity is reduced to the speedy inspection of just one special rotation, the generalized lower triangular (GLT) form. As part of exploratory factor analysis, our method is deployed to avoid nonsensical models independently of observation ordering in both Bayesian and frequentist contexts. Furthermore, a fully Bayesian sampling procedure is developed, which leverages on the GLT rotation while estimating the unknown number of latent factors. The procedure is applied to financial and economic data. Joint work with Sylvia Frühwirth-Schnatter.

Modeling Populations

Leanna House Wednesday, June 29

Virginia Tech

Applied areas such as, epidemiology, social policy, transportation, etc., often rely on complex simulation models (e.g., agent-based) to assess the viability of potential mitigation and/or policy strategies. Among other inputs, these models tend to require specific, individual-level details for entire populations of interest; e.g., the number, age, and income for every home in a municipality. Yet, rarely is such detail available, or even possible to collect. Some success has resulted from pairing simulation models with synthetic population generators (e.g., iterated conditional modes and other imputation methods), but challenges remain in such cases. In particular, describing and accounting for uncertainty in analyses imposed by the use of generated (not true) populations remains a difficult task. When the uncertainty is ignored or estimated poorly, inferences derived from complex simulation models are likely unduly dependent on random, unknown features of supplied input populations. In this talk, we develop approaches for generating realizations of populations a posteriori, which can be incorporated directly into simulation-based analyses. This poster is co-presented by Leanna House and Dave Higdon.

Goal-Oriented Optimal Experimental Design for Nonlinear Systems

Xun Huan Wednesday, June 29

University of Michigan

Optimal experimental design (OED) provides a systematic approach to quantify and maximize the value of experimental data. Conventional Bayesian OED maximizes the expected information gain (EIG) on model parameters, but often we are interested in predictive quantities of interest (QoIs) that depend on the parameters in a nonlinear manner. Goal-oriented OED (GOOED) thus seeks the design that provides the greatest EIG on the predictive QoIs. GOOED avoids unnecessary shrinkage of posterior in regions unimportant to the QoIs, and computes EIG on the posterior pushforward that is often lower dimensional than the parameter space. Under a double-loop Monte Carlo structure, observations are simulated in the outer loop and posterior pushforward samples are generated with MCMC in the inner loop. Kernel density estimation is then used to estimate the Kullback-Leibler divergence. We demonstrate GOOED on a number of nonlinear applications, and illustrate the difference between optimal designs from conventional OED versus GOOED.

Change Point Detection of Transmission Rate in a Stochastic Epidemic Model

Jenny Huang Wednesday, June 29

Duke University

Throughout the course of an epidemic, the transmission rate of a disease is expected to fluctuate. These changes, however, are not expected to occur very often. We can thus model the change in transmission rate using a sparse binary vector. A Bayesian approach to change point detection has been to use sparsity inducing priors to tease out signal from irrelevant noise. These priors, including discrete mixture distributions like the spike-and-slab and single continuous distributions like the horseshoe and double-exponential, have been well-studied within linear models. Considerably less work, however, has been done to apply these priors within more complicated settings like the stochastic SIR. This study investigates the use of sparsity-inducing priors to locate and estimate changes in transmission rate within a stochastic SIR model.

Improved estimation of the mean of a Bernoulli with Monte Carlo

Mark Huber Wednesday, June 29

Claremont McKenna College

Consider the problem of estimating the mean p of a 0-1 (Bernoulli) random variable through Monte Carlo. The Gamma Bernoulli Approximation Scheme (GBAS) give a method with a user specified relative error distribution, that is, the distribution of the relative error is independent of p. This allows for very precise targeting of the run time to achieve any desired relative loss. GBAS operates by smoothing geometric random variables of mean $1/p$ into exponential random variables of mean $1/p$. While the benefits of using scalable exponentials versus geometric random variables are clear, this does result in the elimination of a factor of $1-p$ in the variance. This can become significant for problems where p is large. This work considers several different methods for addressing this problem. First, modifications to GBAS are considered, including the use of multiple intervals created by using antithetic and lattice smoothing of the generated exponential random variables. Second, more precise bounds on sums of geometric random variables are developed than were used in earlier work.

A subsampling approach for Bayesian model selection

Aliaksandr Hubin Wednesday, June 29

University of Oslo

It is common practice to use Laplace approximations to compute marginal likelihoods in Bayesian versions of generalized linear models (GLM). Marginal likelihoods combined with model priors are then used in different search algorithms to compute the posterior marginal probabilities of models and individual covariates. This allows performing Bayesian model selection and model averaging. For large sample sizes, even the Laplace approximation becomes computationally challenging because the optimization routine involved needs to evaluate the likelihood on the full set of data in multiple iterations. As a consequence, the algorithm is not scalable for large datasets. To address this problem, we suggest using a version of a popular batch stochastic gradient descent (BSGD) algorithm for estimating the marginal likelihood of a GLM by subsampling from the data. We further combine the algorithm with Markov chain Monte Carlo (MCMC) based methods for Bayesian model selection and provide some theoretical results on the convergence of the estimates. Finally, we report results from experiments illustrating the performance of the proposed algorithm.

Mixture Representations for Likelihood Ratio Ordered Distributions

Michael Jasuch Wednesday, June 29

Cornell University

In many statistical applications, subject matter knowledge or theoretical considerations suggest that two distributions should satisfy a stochastic order, with samples from one distribution tending to be larger than those from another. In these situations, incorporating stochastic order constraints can lead to improved inferences. This poster will introduce mixture representations for distributions satisfying a likelihood ratio order. To illustrate the value of the mixture representations, I'll address the problem of density estimation for likelihood ratio ordered distributions. In particular, I propose a nonparametric Bayesian solution which takes advantage of the mixture representations. The prior distribution is constructed from Dirichlet process mixtures and has large support on the space of pairs of densities satisfying the monotone ratio constraint. With a simple modification to the prior distribution, we can also test the equality of two distributions against the alternative of likelihood ratio ordering. I'll demonstrate the approach in two biomedical applications.

Network-based Trajectory Topic Interaction Map for Text Mining of COVID-19 Biomedical Literature

Ye Seul Jeon Wednesday, June 29

Yonsei University

Since the emergence of the worldwide pandemic of COVID-19, relevant research has been published at a dazzling pace, which makes it hard to follow the research in this area without dedicated efforts. It is practically impossible to implement this task manually due to the high volume of the relevant literature. Text mining has been considered to be a powerful approach to address this challenge, especially the topic modeling, a well-known unsupervised method that aims to reveal latent topics from the literature. However, in spite of its potential utility, the results generated from this approach are often investigated manually. Hence, its application to the COVID-19 literature is not straightforward and expert knowledge is needed to make meaningful interpretations. In order to address these challenges, we propose a novel analytical framework for estimating topic interactions and effective visualization for topic interpretation. Here we assumed that topics constituting a paper can be positioned on an interaction map, which belongs to a high-dimensional Euclidean space. Based on this assumption, after summarizing topics with their topic-word distributions using the biterm topic model, we mapped these latent topics on networks to visualize relationships among the topics. Moreover, in the proposed approach, we developed a score that is helpful to select meaningful words that characterize the topic. We interpret the relationships among topics by tracking the change of relationships among topics using a trajectory plot generated with different levels of word richness. These results together provide deeply mined and intuitive representation of relationships among topics related to a specific research area. The application of this proposed framework to the PubMed literature shows that our approach facilitates understanding of the topics constituting the COVID-19 knowledge.

Bayesian Updating Rules for Entry and Exit of Forecasts

Matt Johnson Wednesday, June 29

Amazon

We develop coherent Bayesian rules for ensemble methods when forecasts, whether coming from people and/or models, enter and exit. While the literature of ensemble methods, including forecast combination, model averaging, and ensemble learning, is vast, virtually none address the issue of unbalanced ensembles. The issue is particularly pertinent for time series – especially economic – where the introduction and/or discontinuation of indices, models, forecasts, and forecasters are commonplace. This fact motivates our macroeconomic application, where forecasts are removed and added based on macroeconomic conditions, mirroring real situations in economic decision making. The updating rules build on the recently developed framework of Bayesian predictive synthesis, and relies solely on the implicit prior/posterior updating of Bayes' theorem, making it applicable to any Bayesian ensemble procedure. Though the proposed rules are general, we specifically discuss the cases for linear pooling (e.g. Bayesian model averaging) and dynamic Bayesian predictive synthesis.

Using decoupled shrinkage and selection to produce sparse factor models for non-Gaussian data

Beatrix Jones Wednesday, June 29

University of Auckland

Decoupled shrinkage and selection provides a framework for extending existing techniques to contexts requiring sparsity. Here we take advantage of that strength to estimate sparse factor models for two non-Gaussian settings. The first uses the copula framework in the existing package bfa to extract "dietary patterns" (underlying factors) from a set of responses to a food frequency questionnaire. Conventionally, foods associated with a particular pattern are identified via an arbitrary cut-off on the factor loadings. Our analysis both accommodates the non-normality of this data and provides automatic selection by producing a sparse loading matrix. In the second example, we construct factors to explain the presence/absence of particular fish species at survey sites in the waters around New Zealand.

On a class of prior distributions accounting for uncertainty in the data

Chaitanya Joshi Wednesday, June 29

University of Waikato

A new class of prior distributions that can be used to assess the sensitivity of the Bayesian posterior inference to uncertainty in the data is proposed. This class is derived starting from an initial distribution and the likelihood function. We establish the mathematical properties of this class and the conditions under which ordering can be established within this class. We show how the sensitivity analysis can be performed using a standard MCMC procedure for any model whose likelihood, or an approximation, is available in a closed form and illustrate using examples.

Approximate Bayesian Computation with Guiding Principles for Stochastic Differential Equations

Petar Jovanovski Wednesday, June 29

Chalmers University of Technology

Stochastic differential equations (SDE) are employed in many areas of science as a powerful tool for modeling processes that are subject to random fluctuations. Bayesian inference for SDEs is problematic because in the majority of cases the likelihood function is analytically intractable. Approximate Bayesian Computation (ABC) methods can be employed because forward simulation is made possible with numerical methods. By leveraging probabilistic symmetries in discretely observed Markov processes, we employ partially exchangeable neural networks (PEN) to learn the summary statistics needed in ABC. The summary statistics are sequentially learned by exploiting an ABC-SMC sampler, which provides new training data. Furthermore, we propose a new simulator that depends on the observation, as opposed to the standard forward simulator. Our ambition is therefore to provide a learning tool for the SDEs model parameters while simultaneously learning the summary statistics needed in ABC.

Bayesian Cox Regression for Population-scale Inference in Electronic Health Records

Alexander Wolfgang Jung Wednesday, June 29

EMBL-EBI

The Cox model is an indispensable tool for time-to-event analysis, particularly in biomedical research. However, medicine is undergoing a profound transformation, generating data at an unprecedented scale, which opens new frontiers to study and understand diseases. Here we propose a Bayesian version for the counting process representation of Cox's partial likelihood for efficient inference on large-scale datasets with millions of data points and thousands of time-dependent covariates. Through the combination of stochastic variational inference and a reweighting of the log-likelihood, we obtain an approximation for the posterior distribution that factorizes over subsamples of the data, enabling the analysis in big data settings. Crucially, the method produces viable uncertainty estimates for large-scale and high-dimensional datasets. We show the utility of our method through a simulation study and an application to myocardial infarction in the UK Biobank. Our framework extends the Cox model to new data sources like biobanks and EHR, the combination of which can provide new insights into our understanding of diseases."

Individualized Inference using Bayesian Quantile Directed Acyclic Graphical Models

Ksheera Sagar K N Wednesday, June 29

Purdue University

We propose an approach termed "qDAGx" for Bayesian quantile regression under directed acyclic graphs (DAGs) where these DAGs are individualized, in the sense that they depend on individual-specific covariates. A key distinguishing feature of the proposed approach is that the DAG structure is learned simultaneously with the model parameters, instead of being assumed fixed, as in most existing works. To scale up the proposed method to a large number of variables and covariates, we use for the model parameters the popular global-local horseshoe prior that affords a number of attractive theoretical and computational benefits to our approach. By modelling the conditional quantiles, qDAGx overcomes the common limitations of mean regression for DAGs, which can be sensitive to the choice of likelihood (e.g., an assumption of multivariate normality), as well as to the choice of priors. We demonstrate the performance of qDAGx via thorough numerical simulations and via an application in precision medicine by inferring person-specific protein–protein interaction networks in patients with lung cancer.

Nonparametric Bayesian modeling of spatially distributed networks

Jennifer Kampe Wednesday, June 29

Duke University

Networks represented as binary adjacency matrices provide simple, interpretable summaries of highly complex ecological systems. The expansion of ecological monitoring technologies has dramatically increased the availability of species co-occurrence network data, and the spatial locations represented. We develop a network-valued stochastic process with dependencies between networks are described via a latent space model in which the factor coordinates emerge from a spatial Gaussian process. We implement this model via an efficient Gibbs sampler, illustrate its performance on simulated and real ecological data sets, and present theoretical results on the properties of the network-valued stochastic process.

Advancing Non-reversibility in Trans-dimensional Samplers

Oktay Karakus Wednesday, June 29

Cardiff University

In this study, we investigate leveraging superior mixing and convergence properties of non-reversible samplers in exploring nonlinear time series models via the reversible jump Markov chain Monte Carlo (RJMCMC), which is one of the most competent ways to tackle the uncertainty caused by variable dimensionality. Despite its potential to make inferences for different sizes of parameter spaces and classes, since performing random walk, RJMCMC suffers from limited mixing capability and slow convergence for advanced applications. Non-reversible transitions have attracted significant attention in recent years and offer promising results thanks to their favourable convergence and mixing properties. Despite this, their performance is still unexplored specifically for time-series model selection, estimation, and prediction. This work utilises various non-reversible samplers for benchmarking and a novel non-reversible sampler based on the Hamiltonian dynamics is proposed via embedding it in the RJMCMC algorithm to explore both continuous (within class) and discrete (trans-class) parameter spaces efficiently.

Statistical Disclosure Risk with Differential Privacy, with Application to the 2020 Decennial Census

Zekican Kazan Wednesday, June 29

Duke University

We propose Bayesian methods to assess the statistical disclosure risk of data released under differential privacy, focusing on settings with a strong hierarchical structure. The risk assessment is performed by hypothesizing Bayesian intruders with various amounts of prior information and examining the distance between their posteriors and priors. We discuss applications of these risk assessment methods to differentially private data releases from the 2020 decennial census and perform simulation studies using public individual-level data from the 1940 decennial census. Among these studies, we examine how the data holder's choice of privacy parameter affects the disclosure risk and quantify the increase in risk when a hypothetical intruder incorporates substantial amounts of hierarchical information.

Lagged couplings for Markov chain Monte Carlo phylogenetic inference

Luke Kelly Wednesday, June 29

Université Paris-Dauphine

Phylogenetic inference is an intractable statistical problem on a complex sample space. Markov chain Monte Carlo (MCMC) methods are the primary tool for Bayesian phylogenetic inference, but it is challenging to construct efficient schemes to explore the associated posterior distribution or assess their performance. Existing approaches are unable to diagnose mixing or convergence of Markov schemes jointly across all components of a phylogenetic model. Building on recent work developing couplings of MCMC algorithms to diagnose convergence and construct unbiased MCMC estimators, we describe a procedure to couple Markov chains targeting a posterior distribution over a space of phylogenetic trees with branch lengths, scalar parameters and latent variables. We use these couplings to check mixing and convergence of Markov chains jointly across all components of the phylogenetic model; samples from our coupled chains may also be used to construct unbiased estimators.

Sequential Bayesian Registration for Functional Data

Yoonji Kim Wednesday, June 29

Department of Statistics, The Ohio State University

In many modern applications, discretely-observed data may be naturally understood as a set of functions. Functional data often exhibit two confounded sources of variability: amplitude (y-axis) and phase (x-axis). The extraction of amplitude and phase, a process referred to as registration, is essential in exploring the underlying structure of functional data. While such data are often gathered sequentially with new functional observations arriving over time, most available registration procedures are only applicable to batch learning, leading to inefficient computation. To address these challenges, we introduce a Bayesian framework for sequential registration of functional data, which updates statistical inference as new sets of functions are assimilated. This Bayesian model-based sequential learning approach utilizes sequential Monte Carlo sampling to recursively update the alignment of functions. Consequently, distributed computing significantly reduces computational cost. Simulations and applications to real data reveal that the proposed approach is flexible in a variety of challenging scenarios.

A Bayesian Survival Model for Time-Varying Coefficients and Unobserved Heterogeneity

Peter Knaus Wednesday, June 29

Vienna University of Economics and Business

Two sources of heterogeneity are often overlooked in the applied survival literature. On the one hand, time-varying hazard contributions of explanatory variables cannot be captured in the widely used Cox proportional hazard model. To this end this paper investigates a dynamic survival model in the spirit of Hemming and Shaw (2005) within a Bayesian framework. Such a specification allows parameters to gradually evolve over time, thus accounting for time-varying effects. On the other hand, unobserved heterogeneity across (a potentially large number of) groups is often ignored, leading to invalid estimators. This paper makes accounting for such effects feasible for even large numbers of groups through a shared factor model, which picks up unexplained covariance in the error term. Building on the Markov Chain Monte Carlo scheme of Wagner (2011) allows the usage of shrinkage priors to avoid overfitting in such a highly parameterized model. This paper uses global-local shrinkage priors to this effect, in the spirit of Bitto and Frühwirth-Schnatter (2019) and Cadonna et al. (2020), among others, to detect which regressors to include and which are allowed to vary over time. Finally, an R package which makes the routine easily available is introduced.

Adaptive Bayesian methods for non-linear inverse problems

Geerten Koers Wednesday, June 29

Delft University of Technology

The non-linear problem of adaptive recovery of a potential $f_0 \in H^\beta$ for the Schrödinger Equation $\frac{1}{2}\Delta u_{f_0} = f_0 u_{f_0}$ on $[0,1]^d$, with $u = g$ on $\partial[0,1]^d$ is considered. The observation is a noisy version of u_{f_0}, with noise that asymptotically vanishes. The Laplacian of u_{f_0} is endowed with an SVD prior, indexed by a hyperparamater specifying its smoothness. No knowledge on the smoothness parameter β of f_0 is assumed, as the hyperparamater is selected by an empirical procedure. The proposed method first considers the linear problem of recovering the Laplacian, and afterwards constructs a posterior on f_0 using this Laplacian and boundary conditions. It is shown that the contraction rate is nearly $n^{-\beta/(d+2\beta+4)}$, corresponding to the minimax-rate of estimating β-smooth functions. A numerical analysis illustrates the results for various choices of f_0.

Dynamic Functional Variable Selection for Multimodal mHealth Data

Matthew Koslovsky Wednesday, June 29

Colorado State University

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. Motivated by data collected in the Pathways between Socioeconomic Status and Behavioral Cancer Risk Factors Study, we design a novel Bayesian joint modeling framework for dynamic functional variable selection to identify and cluster relations between functional activity level trajectories passively measured with accelerometers and momentary negative affect collected with ecological momentary assessment methods. Our approach leverages spiked hierarchical species sampling priors to identify critical moments when a participant experiences momentary spikes in negative affect, characterize activity level trajectories which are related to these critical moments, and cluster the relational trends within and between participants to explore potential subpopulations. The results of our analysis may have relevance for personalized just-in-time adaptive interventions focused on negative affect management, where moments of increased negative affect may be identified and targeted with activity prompts and/or other motivational or educational messaging via mHealth technologies to improve mood.

A graphical model for multivariate discrete data using spatial and expert information

Christopher Krapu Wednesday, June 29

Oak Ridge National Lab

When modeling sparsely observed multivariate spatial data, strong prior information can be used to bolster predictive accuracy. We propose a novel graphical model for a spatial Bayesian network and which combines a dimension-reduced latent Gaussian spatial field with parameters enforcing a DAG-derived cross-variable covariance structure. This modeling form is especially suitable for the usage of prior information in the form of rulesets derived from expert information. To perform inference with missing data, we implement a Markov chain Monte Carlo scheme composed of alternating steps of discrete Gibbs sampling and Hamiltonian Monte Carlo. A case study on inferring the properties of buildings in Knoxville, Tennessee, USA is presented to highlight the advantages and limitations of this approach.

Robust Variational Inference Methods for Model Misspecification

Ines Krissaane Wednesday, June 29

Nottingham University

In many complex scientific problems, we often find ourselves working with a model that is misspecified relative to the data generating process (DGP), in the sense that there is no parameter setting that allows the model to perfectly replicate the data. In this case, Bayesian methods can give misleading inferences about the quantities of interest (QoI). Although misspecified, the model may be the only link between the data and QoI, and may still contain useful structure that we can learn from. Should we proceed with a standard Bayesian analysis as if nothing was wrong, or model the discrepancy, or change the inferential approach? Here, we will consider alternatives to standard Bayesian inference, and instead of using Kullback-Leibler divergence to measure distances between the DGP and the model, we will focus on alternatives that may give inferences that are more robust to model misspecification. We use variational inference as a flexible approach for approximating the 'posterior' distribution and investigate whether this can lead to better approximate distributions under model misspecification.

Efficient and Scalable Bayesian Bipartite Matching through Fast Beta Linkage

Brian Kundinger Wednesday, June 29

Duke University

Recently, researchers have developed Bayesian versions of the Fellegi Sunter model for record linkage. These have the crucial advantage of quantifying uncertainty from imperfect linkages. However, current implementations of Bayesian Fellegi Sunter models are computationally intensive, making them challenging to use on larger-scale record linkage tasks. We propose a variation on Bayesian Fellegi Sunter models that we call fast beta linkage, or fabl. Specifically, in fabl we use independent prior distributions over the matching space, allowing us to use hashing techniques that reduce computational overhead. This also allows us to complete pairwise record comparisons over large datasets through parallel computing and reduce memory costs through a new technique called storage efficient indexing. Through simulations and two case studies, we show that fabl has increased speed with minimal loss of accuracy.

Shrinkage in Space — Priors for Spatial Models

Nikolas Kuschnig Wednesday, June 29

Vienna University of Economics and Business

In an ever more connected world, spillover effects lie at the centre of much applied research. Spatial econometric models are commonly used to analyse such spillovers empirically. However, these models suffer from rigid specifications and strong assumptions regarding connectivity between units. These drawbacks can be addressed by adopting a fully Bayesian approach. I propose spatial shrinkage priors for flexible modelling that impose regularisation where appropriate and limit the need for assumptions otherwise. Assumptions such as a known connectivity structure can be loosened, with it being learned from the data instead. The result is a more credible and extensible empirical framework that natively accounts for uncertainty. In this paper, I dismantle the spatial econometric framework, discuss prior information in the spatial context, and sketch out a Bayesian approach to it. I demonstrate the merits of my approach by means of simulation and an empirical exercise.

Simultaneous Graphical Dynamic Linear Models for Stock Price Forecasting

Nelson Kyakutwika Wednesday, June 29

Stellenbosch University

Forecasting in a multivariate system calls for use of approaches that can take into consideration the dependence effects among variables. Simultaneous Graphical Dynamic Linear Models are a recently introduced class of models that capture these multivariate dependencies. Attention is put on constructing a distinct dynamic linear model for each univariate time series; each day, the series are recoupled using importance sampling to capture the contemporaneous cross-series dependencies, the series are then decoupled by variational Bayes into individual univariate series to move to the next day. A challenge in computation is encountered while scaling up to high dimensions. We do a Bayesian analysis with the help of Graphics Processing Unit (GPU) computation to forecast returns of 100 stocks listed on the Johannesburg Stock Exchange. The results suggest that recoupling improves forecast accuracy. Also, GPU computation speeds up the analysis.

Compartmental models in epidemiology: Application on Smoking Habits

Alessio Lachi Wednesday, June 29

DiSIA (UNIFI)

Over 85% of Lung Cancer deaths are attributable to smoking. We develop a Bayesian approach to estimate the parameters that governing a Compartmental model. The goal is to analyze smoke attributable deaths and describe the dynamics of transition from a given compartment to another one through a system of Stochastic Ordinary Differential equations in continuous time but discretized through a Markov chain model with fixed or calibrated parameter rates. The model divides the population in three compartments, never - current and former smokers, each other interconnected. Fertility and mortality rates are assumed fixed over time, but this latter distinguished by age, gender and smoking status. Initiation and cessation rates are estimate through piecewise Spline special function and the relapse rate follows a Negative Exponential distribution. We derive the posterior distributions of the parameter of piecewise Spline special functions and the parameters governing the relapse rate through approximate Bayesian computation algorithms.

Why the Rich Get Richer? On the Balancedness of Random Partition Models

Changwoo Lee Wednesday, June 29

Texas A&M University

Random partition models are widely used in Bayesian methods for various clustering tasks, such as mixture models, topic models, and community detection problems. While the number of clusters induced by random partition models has been studied extensively, another important model property regarding the balancedness of cluster sizes has been largely neglected. We formulate a framework to define and theoretically study the balancedness of exchangeable random partition models, by analyzing how a model assigns probabilities to partitions with different levels of balancedness. We demonstrate that the "rich-get-richer" characteristic of many existing popular random partition models is an inevitable consequence of two common assumptions: product-form exchangeability and projectivity. We propose a principled way to compare the balancedness of random partition models, which gives a better understanding of what model works better and what doesn't for different applications. We also introduce the "rich-get-poorer" random partition models and illustrate their application to entity resolution tasks.

Fisher meets BART: Integrating causal machine learning and randomization tests

JungHo Lee Wednesday, June 29

University of Chicago

This paper presents a method for assessing the significance of causal estimates from machine learning models using a randomization-based testing framework. The starting point will be testing weak null hypotheses that contain nuisance parameters that are treatment effect estimates. While randomization tests of these weak nulls have historically been underpowered, we show how using estimates from modern causal machine learning models can alleviate these issues. Therefore, a contribution of this paper will be integrating randomization testing and machine learning methods for causal inference. We demonstrate our methodology on weak null hypotheses involving the sample average treatment effect and treatment effect heterogeneity.

Asymptotic Properties for Bayesian Neural Network in Besov Space

Kyeongwon Lee Wednesday, June 29

Seoul National University

Neural networks have shown great predictive power when dealing with various unstructured data such as images and natural languages. The Bayesian neural network captures the uncertainty of prediction by assuming a prior distribution for the parameter of the model and computing the posterior distribution. We show that the Bayesian neural network using spike-and-slab prior has consistency with nearly minimax convergence rate when the true regression function is in the Besov space. Furthermore, we extend the result to (1) when the smoothness parameters are unknown and (2) using shrinkage prior. In other words, we propose a practical Bayesian neural network with guaranteed asymptotic properties.

Scalable Spatiotemporally Varying Coefficient Modelling with Bayesian Kernelized Tensor Regression

Mengying Lei Wednesday, June 29

McGill University

As a regression technique in spatial statistics, the spatiotemporally varying coefficient model (STVC) is an important tool for discovering nonstationary and interpretable response-covariate associations over both space and time. However, it is difficult to apply STVC for large-scale spatiotemporal analyses due to the high computational cost. To address this challenge, we summarize the spatiotemporally varying coefficients using a third-order tensor structure and propose to reformulate the spatiotemporally varying coefficient model as a special low-rank tensor regression problem. The low-rank decomposition can effectively model the global patterns of the large data sets with a substantially reduced number of parameters. To further incorporate the local spatiotemporal dependencies, we use Gaussian process (GP) priors on the spatial and temporal factor matrices. We refer to the overall framework as Bayesian Kernelized Tensor Regression (BKTR). For model inference, we develop an efficient Markov chain Monte Carlo (MCMC) algorithm, which uses Gibbs sampling to update factor matrices and slice sampling to update kernel hyperparameters. We conduct extensive experiments on both synthetic and real-world data sets, and our results confirm the superior performance and efficiency of BKTR for model estimation and parameter inference.

Recursive Monte Carlo and Variational Inference with Auxiliary Variables

Alexander Lew Wednesday, June 29

Massachusetts Institute of Technology

A key challenge in applying Monte Carlo and variational inference (VI) is the design of proposals and variational families that are flexible enough to closely approximate the posterior, but simple enough to admit tractable densities and variational bounds. We present recursive auxiliary-variable inference (RAVI), a new framework for exploiting flexible proposals, for example based on involved simulations or stochastic optimization, within Monte Carlo and VI algorithms. The key idea is to estimate intractable proposal densities via meta-inference: additional Monte Carlo or variational inference targeting the proposal, rather than the model. RAVI generalizes and unifies several existing methods for inference with expressive approximating families, which we show correspond to specific choices of meta-inference algorithm, and provides new theory for analyzing their bias and variance. We illustrate RAVI's design framework and theorems by using them to analyze and improve upon Salimans et al. (2015)'s Markov Chain Variational Inference, and to design a novel sampler for Dirichlet process mixtures, achieving state-of-the-art results on a standard benchmark dataset from astronomy and on a challenging data-cleaning task with Medicare hospital data.

Bayesian Group LASSO Regression for Genome-wide Association Studies

Lanxin Li Wednesday, June 29

School of Mathematics and Statistics, University of Glasgow

Genome-wide association studies (GWAS) are designed to search across a genome-wide set of genetic variations (SNPs) from different individuals to find SNPs that are associated with a trait of interest. Many statistical methods for GWAS have limitations in accurately identifying SNPs underlying complex diseases (like heart disease), due to weak association signals, local correlations between SNPs, and numbers of candidate SNPs vastly exceeding available sample sizes. We propose a Bayesian model framework, adapting ideas from Bayesian group Lasso regression, that clusters correlated SNPs into groups, and a population-based MCMC method to conduct powerful group selection in GWAS, to improve the accuracy and efficiency of detecting trait-associated regions. In this model, signals from causative SNPs and SNPs correlated with causative ones can be accumulated together, vastly reducing the total number of variables that need to be tested, and both statistically and biologically informed priors are used to further improve its precision.

Automatic Marginalization of Discrete Variables

Yucen L Wednesday, June 29

Meta

Markov Chain Monte Carlo (MCMC) inference methods are an effective way to estimate posterior distributions of probabilistic programs or models given a set of observations. However, some of its more effective forms, such as Hamiltonian Monte Carlo and No U-Turn Sampler, are only applicable to models where all variables are continuous. In the case of hybrid models (that is, those involving both continuous and discrete variables), these methods can be applied only after marginalizing the discrete variables and sampling from the remaining all-continuous marginalized distribution. In fact, even for MCMC methods that can be applied to hybrid models (such as Metropolis-Hastings), generating samples from the marginalized distribution may offer enormous gains due to the variance reduction afforded by Rao-Blackwellization.

Marginalizing the discrete variables from a probabilistic program is typically an ad hoc and time-consuming manual process requiring expertise in probabilistic inference. In this paper, we present an automatic method for marginalizing the discrete variables from probabilistic programs that have static structure (in other words, those equivalent to a Bayesian network). We describe the details of the method, prove its correctness, and show how it can improve results for an experimentation Bayesian analysis application.

Adaptive Random Neighbourhood Informed MCMC on Bayesian Variable Selection

Xitong Liang Wednesday, June 29

UCL

Bayesian variable selection methods (BVS) are growing in popularity. BVS both enables quantification of model uncertainty and also helps investigate underlying low-dimensional structure in predicting the response of interest. However, the increasing adoption of BVS is hindered by a lack of efficient computational algorithms for discrete posterior distributions and large-scale datasets. Our main contribution is developing a new MCMC scheme named PARNI. It first randomly constructs a neighbourhood using ideas from [1], then proposes a new model within this neighbourhood according to a locally balanced proposal [2]. Additional challenges arise for generalised linear models where the marginal posterior is not analytically available. In such cases, we employ methods such as data augmentation to produce informed proposals. We conduct numerical studies on both simulated and real datasets. The results show the PARNI sampler is competitive with state-of-the-art schemes.

[1] Griffin, J., Łatuszyński, K. and Steel, M., 2020. In search of lost mixing time: adaptive Markov chain Monte Carlo schemes for Bayesian variable selection with very large p. Biometrika, 108(1), pp.53-69.

[2] Zanella, G., 2019. Informed Proposals for Local MCMC in Discrete Spaces. Journal of the American Statistical Association, 115(530), pp.852-865.

Shared Differential Clustering across Single-cell RNA Sequencing Datasets with the Hierarchical Dirichlet Process

Jinlu Liu Wednesday, June 29

University of Edinburgh

Single-cell RNA sequencing (scRNA-seq) is powerful technology that allows researchers to understand gene expression patterns at the single-cell level. However, analysing scRNA-seq data is challenging due to issues and biases in data collection. In this work, we construct an integrated Bayesian model that simultaneously addresses normalization, imputation and batch effects and also nonparametrically clusters cells into groups across multiple datasets. Specifically, the Hierarchal Dirichlet process (HDP) is used to discover clusters of cells with similar mean-expression and dispersion patterns that may be unique or shared across datasets. In addition, the mean-variance relationship is directly accounted for through an informative regression model, which provides robust estimates, particularly for sparse data and/or small clusters. A Gibbs sampler based on a finite-dimensional approximation of the HDP is developed for posterior inference. On simulated datasets, we show that our model is robust in terms of the ability to capture the clustering structure and the true relationship between the mean-expression and dispersion parameters. Our work is motivated by experimental data collected to study prenatal development of cells under conditions when the transcription factor, Pax6, is knocked out in mutant mice. In this case, our model is used to identify clusters of cells which behave differently under the experimental conditions.

A Bayesian Phase I/II Design for Immunotherapy and Chemotherapy Combination

Suyu Liu Wednesday, June 29

MD Anderson Cancer Center

Immunotherapy is an innovative treatment approach that harnesses a patient's immune system to treat cancer. It has provided an alternative and complementary treatment modality to conventional chemotherapy. Combining immunotherapy with cytotoxic chemotherapy agent has become the leading trend and the most active research field in oncology. To accommodate this growing trend, we propose a Bayesian phase I/II dose-finding design to identify the optimal biological dose combination (OBDC), defined as the dose combination with the highest desirability in the risk-benefit tradeoff. We propose new statistical models to describe the relationship between the doses and treatment outcomes, including immune response, toxicity, and progression-free survival (PFS). During the trial, based on accrued data, we continuously update model estimates and adaptively assign patients to dose combinations with high desirability. The simulation study shows that our design has desirable operating characteristics.

Leave-Group-Out Cross-Validation in INLA

Zhedong Liu Wednesday, June 29

King Abdullah University of Science and Technology

After a model is fitted, the predictive performance of the model is usually interesting. Leave-one-out cross-validation (LOO-CV) is viewed as a good estimator to estimate the predictive performance. We argue that LOO-CV is a good estimator of the predictive performance of a model for non-structured data, but it may not be a good estimator of the predictive performance of a model for structured data given a prediction task. When the prediction task is given in a structured data model, the cross-validation strategy should be modified to adapt to the task. When the prediction task is not given in a structured data model, we proposed a cross-validation strategy according to the data structure implied by the model, which makes the testing data relatively independent of the training data. The mentioned cross-validation strategies can be represented by a leave-group-out cross-validation(LGO-CV) format. The LGO-CV and the data structure extraction are implemented in R-INLA.

A spatio-temporal random-censoring Poisson model for handling underreported data

Guilherme Lopes de Oliveira Wednesday, June 29

CEFET-MG

Estimates of mortality rates, mainly in underdeveloped regions around the world, are biased due to the large amount of deaths that go unreported. Oliveira et al. (2017) proposed the called random-censoring Poisson model (RCPM) in which censorship is used as a mechanism to correct underestimation of the mortality rates. The RCPM does not consider the effect of time in the event occurrence rates neither in the censoring probabilities. We extended the RCPM to allow a temporal trend analysis in both the data occurrence and data reporting processes. Time-indexed covariates and some time-dependent functions are considered for such a purpose. Our model is applied to early neonatal mortality data from Minas Gerais State, Brazil, in period 2000 to 2017. Joint work with Rosangela Loschi, Breno Valente and Vinícius Fonseca.

Bayesian modeling in an adaptive platform trial in COVID-19

Elizabeth Lorenzi Wednesday, June 29

Berry Consultants

Traditional fixed, parallel group randomized controlled trials (RCTs) have long been considered the “gold standard” in drug development. However, there are many questions that are too costly, time-consuming, or challenging to answer with traditional designs. At the start of the COVID-19 pandemic, there was an urgent need to learn about how best to treat this new disease and to do so efficiently. Many trial groups around the world began setting up RCTs to fill this void, many of which were adaptive platform trials. Adaptive platform trials are RCT designs that through a pre-specified decision algorithm, can perpetually study multiple interventions, add or remove arms over time, and study multiple patient populations. The adaptive nature of platform trials – modifying key design parameters as information is gained – as well as the dynamic sharing of information across populations, treatments and time, makes them particularly well-suited for the pandemic setting. In this poster, I will present on the Bayesian analysis model, trial design and results of a platform trial in the COVID-19 pandemic, REMAP-CAP. The REMAP-CAP trial is an international, multi-factorial, adaptive platform trial designed to learn about effective treatment strategies for patients with severe pneumonia or COVID-19 in both pandemic and non-pandemic settings. The platform has resulted in conclusions on the effectiveness of interventions including corticosteroids, immune modulators, convalescent plasma, anticoagulation therapy, antiviral interventions, and antiplatelet interventions.

Power laws distributions in objective priors

Francisco Louzada Wednesday, June 29

Universidade de São Paulo, Brazil

Using objective priors in Bayesian applications has become a common practice to analyze data without subjective information. Formal rules usually obtain these prior distributions, and the data provide the dominant information in the posterior distribution. However, these priors are typically improper and may lead to improper posterior. Here, for a general family of distributions, we show that the obtained objective priors for the parameters either follow a power-law distribution or have an asymptotic power-law behavior. As a result, we observed that the exponents of the model are between 0.5 and 1. Understanding these behaviors allows us to easily verify if such priors lead to proper or improper posteriors directly from the exponent of the power-law. The general family considered in our study includes essential models such as Exponential, Gamma, Weibull, Nakagami-m, Half-Normal, Rayleigh, Erlang, and Maxwell Boltzmann distributions, to list a few. In summary, we show that comprehending the mechanisms describing the shapes of the priors provides essential information that can be used to understand the properties of the posterior distributions.

Latent Space Modelling of Hypergraph Data

Simon Lunagomez Wednesday, June 29

ITAM

The increasing prevalence of relational data describing interactions among a target population has motivated a wide literature on statistical network analysis. In many applications, interactions may involve more than two members of the population and this data is more appropriately represented by a hypergraph. In this poster, we present a model for hypergraph data which extends the well established latent space approach for graphs and, by drawing a connection to constructs from computational topology, we develop a model whose likelihood is inexpensive to compute. A delayed-acceptance MCMC scheme is proposed to obtain posterior samples and we rely on Bookstein coordinates to remove the identifiability issues associated with the latent representation. We theoretically examine the degree distribution of hypergraphs generated under our framework and, through simulation, we investigate the flexibility of our model and consider estimation of predictive distributions. Finally, we explore the application of our model to real data.

Bayesian doubly robust causal inference via loss functions

Yu Luo Wednesday, June 29

Imperial College London

Doubly robust causal inference has a well-established basis in frequentist semi-parametric theory, with estimation of causal parameters typically conducted via outcome regression and propensity score adjustment. A Bayesian counterpart, however, is not obvious as doubly robust estimation involves a semi-parametric formulation in the absence of a fully specified likelihood function. In this paper, we propose a Bayesian approach for doubly robust causal inference via two general Bayesian updating approaches based on loss functions. First, we specify a loss function for a doubly robust propensity score augmented outcome regression model and apply the traditional Bayesian updating mechanism which uses a prior belief distribution to calculate the posterior. Secondly, we draw inference for the posterior from a Bayesian predictive distribution via a Dirichlet process model, extending the Bayesian bootstrap. We show that these updating procedures yield valid posterior distributions of parameters which exhibit double robustness. Simulation studies show that the proposed methods can recover the true causal effect efficiently and achieve frequentist coverage even when the sample size is small or if the propensity score distribution is highly skewed. Finally, we apply our methods to evaluate the causal impact of speed cameras on traffic collisions in England.

Down the Rabbit Hole of Sexually Violent Person Laws – Bayes' Theorem Shows the Way Out By Tackling Controversies, Cognitive Biases and Educational Deficits in Analysis of Evidence in Legal and Mental Health Fields

Diane Lytton Wednesday, June 29

Independent forensic psychologist

"What's Bayes have to do with science and this case?" a prosecutor asked during a Sexually Violent Person (SVP) commitment trial that requires experts' diagnostic and recidivism risk opinions. Opinions become trial evidence whether sex offenders about to be released from prison should be released, or indefinitely committed to maximum security treatment facilities until found no longer dangerous typically through treatment, aging or terminal medical conditions. Bayes is inherent in common SVP risk tools, and useful for diagnostic opinions similar to other medical fields; however, on-going Bayesian controversies impede unbiased, evidence-based cost-effective decisions. Models from early 1900s mathematics to 1950s medicine, learning and cognitive sciences illustrate that the intuitive odds-ratio Bayes formula and pictorial representations are understood by most persons with limited math or statistics knowledge, can advance science out of rabbit holes in the legal and mental health fields, and enhance critical thinking skills from an early age.

GP-BART: a novel Bayesian additive regression trees approach using Gaussian processes

Mateus Maia Marques Wednesday, June 29

Maynooth University

The Bayesian additive regression trees (BART) model is an ensemble method extensively and successfully used in regression tasks due to its consistently strong predictive performance and its ability to quantify uncertainty. BART combines "weak" tree models through a set of shrinkage priors, whereby each tree explains a small portion of the variability in the data. However, the lack of smoothness and the absence of a covariance structure over the observations in standard BART can yield poor performance in cases where such assumptions would be necessary. We propose Gaussian processes Bayesian additive regression trees (GP-BART) as an extension of BART which assumes that the prediction of each terminal node among all trees is given by a Gaussian process (GP) model. We illustrate our model on simulated and real data and compare its performance to traditional modelling approaches, outperforming them in many scenarios.

Improving multiple-try Metropolis with local balancing

Florian Maire Wednesday, June 29

Department of Mathematics and Statistics, Université de Montréal

Multiple-try Metropolis (MTM) is a popular Markov chain Monte Carlo algorithm amenable to parallel computing with appealing potential. At each iteration, it samples several candidates for the next state of the Markov chain and selects one of them based on a weight function. We show that the preferred choice of weight function in the literature, which is proportional to the target density, induces pathological behaviours in high dimensions, leveraging a connection to the work of Zanella (2020). We propose different weight functions for which those pathological behaviours do not arise. Also, we provide a scaling-limit analysis that allows to characterize the high-dimensional behaviour of MTM when using the preferred weight function and when using a proposed weight function. In each case, MTM is seen as an approximation to a limiting sampling scheme which is approached under conditions on the rate at which the number of candidates increases with the dimension.

Bayesian model-based clustering for multiple network data

Anastasia Mantziou Wednesday, June 29

Imperial College London

There is increasing appetite for analysing multiple network data. This is different to analysing traditional data sets, where now each observation in the data comprises a network. Recent technological advancements have allowed the collection of this type of data in a range of different applications. This has inspired researchers to develop statistical models that most accurately describe the probabilistic mechanism that generates a network population. Only a few studies developed to date consider the heterogeneity that can exist in a network population. We propose a Mixture of Measurement Error Models for identifying clusters of networks in a network population, with respect to similarities detected in the connectivity patterns among the networks' nodes. Simulation studies show our model performs well in both clustering multiple network data and inferring the model parameters. We apply our model on two real world multiple network data sets resulting from the fields of Computing and Neuroscience.

Bayesian Functional Partial Membership Models

Nicholas Marco Wednesday, June 29

University of California, Los Angeles

Partial membership models, or mixed membership models, are a flexible un-supervised clustering method that allows observations to belong to multiple clusters at the same time. In this paper, we propose a Bayesian partial membership model for functional data. By using the multivariate Karhunen-Loʻeve theorem, we are able to derive a scalable model that does not make many assumptions on the covariance structure of the data. Compared to previous work on partial membership models, our proposed model is more flexible and allows for direct interpretation of the mean and covariance structure. To illustrate the usefulness of our model, we fit our partial membership model on EEG signals from a cohort containing children with Autism Spectrum Disorder (ASD). We found that the results from our model tend to agree with the results previously found in the scientific literature, however our model allows clinicians to analyze the data in a novel way.

Bayesian mixture models for the extremal dependence

Giulia Marcon Wednesday, June 29

Università degli studi di Palermo

Multivariate extremes can be modelled through the estimation of the angular measure. Our proposal focuses on a non-parametric Bayesian model defined as a mixture of Beta densities which allows the estimation over the (d-1)-dimensional simplex. The constraints required to provide a valid extremal dependence are addressed in a straightforward manner and in a matrix form. This extends current models based on Bernstein polynomials, retaining desirable properties while improving flexibility. Joint work with Isadora Antoniano Villalobos.

Money laundering control in Mexico, A risk management approach through regression trees (data mining)

José Francisco Martínez Wednesday, June 29

Universidad Autónoma del Estado de Hidalgo, México

Purpose – This paper is aimed at developing a regression tree model useful to quantify the Money Laundering (ML) risk associated to a customer profile and his contracted products (customer's inherent risk). ML is a risk to which different entities are exposed, but mainly the financial ones because of the nature of their activity, so that they are legally obliged to have an appropriate methodology to analyze and assess such a risk.
Design/methodology/approach – This paper uses the technique of regression trees to identify, measure and quantify the ML customer's inherent risk.
Findings – After classifying customers as high- or low-risk based on a probability threshold of 0.5, this study finds that customers with 56 months or more of seniority are more risky than those with less seniority; the variables "contracted product" and "customer seniority" are statistically significant; the variables origin, legal entity and economic activity are not statistically significant for classifying customers; institution collection, business products and individual product are the most risky; and the percentage of effectiveness, suggested by the decision tree technique, is around 89.5 per cent.
Practical implications – In the daily practice of ML risk management, the two main issues to be considered are: 1) the knowledge of the customer, and 2) the detection of his inherent risk elements.
Originality/value – Information from the customer portfolio and his transaction profile is analyzed through BigData and data mining.

Nonparametric Mixture of Envelope Models

Andrea Mascaretti Wednesday, June 29

Università degli Studi di Padova

Envelope models (Cook et al. 2010, Stat. Sinica) address the problem of regression for multivariate responses by leveraging sufficient dimension reduction techniques. In particular, predictors are categorised in ""material"" and ""immaterial"" to the response variables as two different stochastic structures are assumed to co-exist in the model. The main advantage of the approach is its greater efficiency in comparison to standard techniques. Bayesian envelopes, however, have not been explored in detail in the literature. One of the main reasons is the complex nature of the parameter space over which the prior distribution needs to be defined. In this work, we build upon existing contributions on the topic and extend it to allow for nonparametric mixtures of envelopes that allows us to relax the hypothesis that the set of material and immaterial variables is the same across all data points. The proposed approach is illustrated via the analysis of synthetic and real datasets.

Bayesian learning of causal structures from general interventions

Alessandro Mascaro Wednesday, June 29

University of Milano-Bicocca

Graphical models based on Directed Acyclic Graphs (DAGs) are often used to represent causal relationships between variables. In this setting, the process of identifying the DAG structure from the data is referred to as causal discovery. However, if only observational data are available, the DAG is identifiable only up to its Markov equivalence class, collecting all the causal structures that are consistent with the same conditional independence relations. Experimental data, i.e. produced after external interventions on variables, allow the identification of smaller sub-classes of DAGs, known as I-Markov equivalence classes. Moreover, different types of interventions define different characterization of I-Markov classes. Current causal discovery algorithms from experimental data assume that interventions do not modify the parents of the intervened node in the DAG. We relax this assumption by proposing a novel Bayesian methodology for causal discovery from Gaussian experimental data arising from general interventions.

Imputation of Missing Data Using Gaussian Linear Cluster-Weighted Modeling

Luis Alejandro Masmela Caita Wednesday, June 29

Universidad Distrital F.J.C. Bogotá D.C. - Colombia

Missing data theory deals with the statistical methods in the occurrence of missing data. Missing data occurs when some values are not stored or observed for variables of interest. However, most of the statistical theory assumes that data is fully observed. An alternative to deal with incomplete databases is to fill in the spaces corresponding to the missing information based on some criteria, this technique is called imputation. We introduce a new imputation methodology for databases with univariate missing patterns based on additional information from fully-observed auxiliary variables. We assume that the non-observed variable is continuous, and that auxiliary variables assist to improve the imputation capacity of the model. In a fully Bayesian framework, our method uses a flexible mixture of multivariate normal distributions to model the response and the auxiliary variables jointly. Under this framework, we use the properties of Gaussian Cluster-Weighted modeling to construct a predictive model to impute the missing values using the information from the covariates. Simulations studies and a real data illustration are presented to show the method imputation capacity under a variety of scenarios and in comparison to other literature methods.

Bayesian Functional Principal Component Analysis using Orthogonal Gaussian Processes

James Matuk Wednesday, June 29

Duke University

Functional Principal Component Analysis (FPCA) is a prominent tool to characterize variability and reduce dimension of longitudinal and functional datasets. Bayesian implementations of FPCA are advantageous because of their ability to propagate uncertainty in subsequent modeling. To ease computation, many modeling approaches rely on the restrictive assumption that functional principal components can be represented through a pre-fixed basis. Alternatively, we propose a flexible Bayesian FPCA model using orthogonal Gaussian processes (GPs). The covariance functions used to define the GPs are carefully designed to enforce mutual orthogonality between principal components to ensure identifiability of model parameters. The priors that we construct, using Bayesian constraint relaxation, balance the degree to which the orthogonality constraint is met and ease of posterior computation. We demonstrate our proposed model using extensive simulation experiments and in an environmental health application, through modeling toxicological dose response curves to study the relationship between chemical exposure and human health.

Bayesian Bootstrap Uncertainty Quantification for Spatial Lesion Regression Modelling

Anna Menacher Wednesday, June 29

University of Oxford

The analysis of white matter lesions based on MRI scans provides insights into the aging brain and neurodegenerative diseases. Current mass-univariate approaches for modelling binary lesion images are unable to provide probabilistic statements about spatial features, like cluster size. Bayesian spatial models can reliably quantify parameter uncertainty via computationally costly MCMC methods but are infeasible for large-scale studies. Alternative approximate methods, like variational inference, are faster however severely underestimate posterior variance. We propose a scalable approximate posterior sampling technique for parameter estimation and inference of structured spike-and-slab regression models by performing independent variational MAP optimizations of pseudo-posteriors. This approach builds on concepts of weighted likelihood Bootstrap and jittered spike-and-slab priors with random shrinkage targets. We hereby provide accurate uncertainty quantification of spatially-varying coefficients and introduce novel cluster-size based imaging statistics. Lastly, we validate our results via simulation studies and an application to the UK Biobank, a large-scale lesion mapping study.

Bayesian estimation of cosmological time delays with Continuous Auto-Regressive Moving Average (CARMA) processes

Antoine Meyer Wednesday, June 29

Imperial College London

Strong gravitational lensing occurs when the gravitational field of a galaxy bends the light emitted by a distant source, causing multiple images of the same source to appear in the sky when viewed from Earth. Fluctuations in the source brightness are observed in the images at different times, due to the different paths the lensed images take to travel to the observer. The time delay between brightness fluctuations can be used to constrain cosmological parameters such as the expansion rate of the Universe. We develop an objective Bayesian method to estimate cosmological time delays, using Continuous Auto-Regressive Moving Average (CARMA) processes to model the irregularly sampled time series of brightness data from the several observed images of the source. Our model accounts for heteroskedastic measurement errors and an additional source of independent extrinsic long-term variability in the source brightness, known as microlensing. We employ the Kalman Filter algorithm for efficient likelihood computation and perform posterior sampling using the MultiNest implementation of nested sampling to deal with posterior multimodality.

A Bayesian hierarchical model for disease mapping that accounts for scaling and heavy-tailed latent effects

Victoire Michal Wednesday, June 29

McGill University

In disease mapping, a disease's relative risk is estimated across areas of a region of interest. The number of cases in an area is assumed to follow a Poisson distribution whose mean is decomposed as the product between an offset and the disease's relative risk logarithm. The log risk may be written as the sum of fixed effects and latent random effects. The BYM2 model decomposes each latent effect into a weighted sum of independent and spatial effects. We build on this model to allow for heavy-tailed effects and accommodate potentially outlying risks, after accounting for the fixed effects. We assume a scale mixture structure wherein the variance of the latent process changes across areas and allows for outlier identification. We explore two prior specifications of this scale mixture structure in simulation studies and in the analysis of Zika cases from the 2015-2016 epidemic in Rio de Janeiro.

A Bayesian multilevel hidden Markov model for uncovering mood states in bipolar patients

Sebastián Mildiner Moraga Wednesday, June 29

Utrecht University

Identifying manic and depressive episodes is central to treating bipolar disorder, yet challenging to perform. To this end, most traditional methods aggregate items of validated questionnaires, such that the individual item dynamics are lost. Here, we present a novel Bayesian multilevel hidden Markov model that explains the variability in the item-based response patterns of patients as switches in their mood states. The proposed model allows for probabilistic estimation of the sequence of mood states in each patient and offers a framework to measure between-patient variability formally. We illustrate our proposed model on empirical data consisting of 20 type-I/II bipolar patients completing five questionnaires a day for four months. We uncover four clinically meaningful mood states and describe their dynamics. Our results are one of the first to comprise patient-specific parameters and improve over the previous time-resolution, thus expanding our understanding of the individual differences in the course of the disorder.

Tumor radiogenomics in gliomas with Bayesian layered variable selection

Shariq Mohammed Wednesday, June 29

Boston University

We propose a statistical framework to integrate radiological magnetic resonance imaging (MRI) and genomic data to identify the underlying radiogenomic associations in lower grade gliomas (LGG). We devise a novel imaging phenotype by dividing the tumor region into concentric spherical layers that mimics the tumor evolution process. MRI data within each layer is represented by voxel–intensity-based probability density functions which capture the complete information about tumor heterogeneity. Subsequently, we build Bayesian variable selection models, using structured spike-and-slab priors, for each layer with the density function-based imaging phenotypes as the response and the genomic markers as predictors. Our novel hierarchical prior formulation incorporates the interior-to-exterior structure of the layers, and the correlation between the genomic markers. Genes implicated with survival and oncogenesis are identified as being associated with the spherical layers, which could potentially serve as early-stage diagnostic markers for disease monitoring, prior to routine invasive approaches.

Regularised B-splines projected Gaussian Process priors to estimate time-trends in age-specific COVID-19 deaths

Mélodie Monod Wednesday, June 29

Imperial College London

The COVID-19 pandemic has caused severe public health consequences in the United States. In this study, we use a hierarchical Bayesian model to estimate the age-specific COVID-19 attributable deaths over time in the United States. The model is specified by a novel non-parametric spatial approach over time and age, a low-rank Gaussian Process (GP) projected by regularised B-splines. We show that this projection defines a new GP with attractive smoothness and computational efficiency properties, derive its kernel function, and discuss the penalty terms induced by the projected GP. Simulation analyses and benchmark results show that the B-splines projected GP may perform better than standard B-splines and Bayesian P-splines, and equivalently well as a standard GP, at considerably lower runtimes. We apply the model to weekly, age-stratified COVID-19 attributable deaths reported by the US Centers for Disease Control, which are subject to censoring and reporting biases. Using the B-splines projected GP, we can estimate longitudinal trends in COVID-19 associated deaths across the US by 1-year age bands. These estimates are instrumental to calculate age-specific mortality rates, describe variation in age-specific deaths across the US, and for fitting epidemic models. The B-splines projected GP priors that we develop are likely an appealing addition to the arsenal of Bayesian regularising priors.

The ALPS algorithm for multi-modal distributions

Matt Moores Wednesday, June 29

University of Wollongong

Multi-modal distributions pose major challenges for the usual algorithms that are employed in statistical inference. These problems are exacerbated in high-dimensional settings, where techniques such as Markov Chain Monte Carlo (MCMC) and Expectation Maximisation (EM) typically rely upon localised update mechanisms: such localised algorithms can effectively become trapped in one of the local modes, leading to biased inference and underestimation of uncertainty.

This poster presents the Annealed Leap-Point Sampler (ALPS), an MCMC algorithm that augments the state space of the target distribution with a sequence of modified, annealed (cooled) distributions. The temperature of the coldest state is chosen such that the corresponding annealed target density of each individual mode can be closely fitted by a Laplace approximation. As a result, independent MCMC proposals based on a mixture of Gaussians can jump between modes even in high-dimensional problems. The ability of this method to "mode hop" at the super-cold state is then filtered through to the target state by swapping information between neighbours, in a similar manner to parallel tempering. ALPS also incorporates the best aspects of current gold-standard approaches to multi-modal sampling in high-dimensional contexts.

We have implemented ALPS as an open-source R package. Our method is demonstrated using examples of multi-modal distributions that arise in econometrics and chemistry.

Multivariate Functional Model With Spatially Varying Coefficient

Hossein Moradir Rekabdarkolaee Wednesday, June 29

South Dakota State University

Hurricanes are massive storm systems with enormous destructive capabilities. Understanding the trends across space and time of a hurricane track and intensity leads to improved forecasts and minimizes their damage. Viewing the hurricane's latitude, longitude, and wind speed as functions of time, we propose a novel spatiotemporal multivariate functional model to simultaneously allow for multivariate, longitudinal, and spatially observed data with noisy functional covariates. The proposed procedure is fully Bayesian and inference is performed using MCMC. This new approach is illustrated through simulation studies and analyzing the hurricane track data from 2004 to 2013 in the Atlantic basin. Simulation results indicate that our proposed model offers a significant reduction in the mean square error and averaged interval and increases the coverage probability. In addition, our method offers a 10% reduction in location and wind speed prediction error.

The Posterior Predictive Null

Gemma Moran Wednesday, June 29

Columbia University

Bayesian model criticism is an important part of the practice of Bayesian statistics. Traditionally, model criticism methods have been based on the predictive check, an adaptation of goodness-of-fit testing to Bayesian modeling and an effective method to understand how well a model captures the distribution of the data. In modern practice, however, researchers iteratively build and develop many models, exploring a space of models to help solve the problem at hand. While classical predictive checks can help assess each one, they cannot help the researcher understand how the models relate to each other. This paper introduces the posterior predictive null check (PPN), a method for Bayesian model criticism that helps characterize the relationships between models. The idea behind the PPN is to check whether data from one model's predictive distribution can pass a predictive check designed for another model. This form of criticism complements the classical predictive check by providing a comparative tool. A collection of PPNs, which we call a PPN study, can help us understand which models are equivalent and which models provide different perspectives on the data. With mixture models, we demonstrate how a PPN study, along with traditional predictive checks, can help select the number of components by the principle of parsimony. With probabilistic factor models, we demonstrate how a PPN study can help understand relationships between different classes of models, such as linear models and models based on neural networks. Finally, we analyze data from the literature on predictive checks to show how a PPN study can improve the practice of Bayesian model criticism.

A Flexible Class of Time-Varying Stochastic Epidemic Models

Raphael Morsomme Wednesday, June 29

Duke University

Over the course of an epidemic, the infection rate may naturally vary over time as a result of changes in behavior, new policies, or other factors. However, existing methods for incorporating time-varying rates within stochastic epidemic models do not scale well to data from large outbreaks; as a result, many practitioners resort to simplifying assumptions with rigid forms on the infection rate. In this paper, we present a new class of time-varying stochastic epidemic models amenable to scalable Bayesian inference under the exact model posterior. Drawing on ideas from volatility modeling, we capture the time-varying infection rate with a discrete-time gamma process model. The model can track changes in the infection rate and, through the use of time-specific discount factors, allows for abrupt variations. We show the practicality of this model by developing an efficient Gibbs sampler for the infection rate under a discrete-time stochastic SEIR model, enabling Bayesian inference for large outbreaks with missing data. We apply the method to assess the impact of interventions on the effective reproduction number in the 1995 Ebola pandemic in Congo.

Bayesian modeling and clustering for spatio-temporal areal data

Alexander Mozdzen Wednesday, June 29

University of Klagenfurt

Spatio-temporal areal data can be seen as a collection of time series which are spatially correlated according to a specific neighboring structure. We propose a Bayesian hierarchical model that allows for spatial model based clustering of the areal units using a nonparametric prior. Exploiting the sparse structures of the precision matrix of a spatio-temporal Gaussian Random Markov Field we develop an efficient MCMC algorithm. The performance of the model is assessed via an application to study regional unemployment patterns in Italy. When compared to other spatial and spatio-temporal competitors, our model shows more precise estimates and the additional information obtained from the clustering allows for an extended economic interpretation.

Tucker decomposition for DNA methylation data

Anjali Nagulpally Wednesday, June 29

UMass Amherst

Recent advances in microarray and sequencing technology have provided large-scale DNA methylation data sets that can be used for understanding the molecular pathways involved in cancer development. However, there is a lack of principled and computationally tractable statistical models for drawing inferences from DNA methylation data. We introduce a novel Poisson–Beta Bayesian statistical model whose generative process is inspired by the true data generation mechanism. We show that though the model is non-conjugate, the complete conditional distribution of the latent Poisson counts is a form of the Bessel distribution, and the marginal likelihood is the doubly non-central Beta (DNCB) distribution. We derive an elegant and efficient auxiliary variable Gibbs sampler for posterior inference. We improve on existing matrix factorization methods by constraining entries to [0, 1] and allowing the number of sample and feature clusters to differ, enabling a flexible representation of the data. We show that this model is competitive with the state-of-the-art in terms of stability and out-of-sample prediction accuracy. This model is more broadly useful for representation learning in real-valued data that is bounded by the unit interval, which includes many types of genomic data sets.

Bayesian adaptive variable selection in linear models: A generalization of Zellner's informative g-prior

Djibril Ndiaye Wednesday, June 29

Université Laval

Bayesian inference is about recovering the full conditional posterior distribution of the parameters of a statistical model. This exercise, however, can be challenging to undertake if the model specification is not available a priori, as is typically the case in practice. This thesis proposes a new framework to select the subset of regressors that are the relevant features that explain a target variable in linear regression models. We generalize Zellner's g-prior with a random matrix, and we present a likelihood-based search algorithm, which uses Bayesian tools to compute the posterior distribution of the model parameters over all possible models generated, based on the maximum a posteriori (MAP). We use Markov chain Monte Carlo (MCMC) methods to gather samples of the model parameters and specify all distributions underlying these model parameters. We then use these simulations to derive a posterior distribution for the model parameters by introducing a new parameter that allows us to control how the selection of variables is done. Using simulated datasets, we show that our algorithm yields a higher frequency of choosing the correct variables and has a higher predictive power relative to other widely used variable selection models such as adaptive Lasso and Bayesian adaptive Lasso, and relative to well-known machine learning algorithms. Taken together, this framework and its promising performance under various model environments highlight that simulation tools and Bayesian inference methods can be efficiently combined to deal with well-known problems that have long loomed the variable selection literature.

Dynamic Stochastic MIDAS Copula Models

Hoang Nguyen Wednesday, June 29

Örebro University

Stock and oil relation depends on the current economic conditions that their dependence structure is usually time-varying. In this study, we propose a dynamic stochastic mixed frequency data sampling (DSM) copula model to utilize the information of the macroeconomic variables into a dynamic stochastic copula model of stock returns and oil returns. A DSM copula decomposes the relationship into a short-term dynamic stochastic dependence and a long-term dependence driven by related macro-finance factors. We find that inflation and interest rate, uncertainty, liquidity are the main drivers of long-term dependence. The proposed DSM copula model also helps to improve the accuracy in risk management and optimal portfolio allocation compared to other alternatives.

Many Processors, Little Time: MCMC for Partitions via Optimal Transport Couplings

Tin Nguyen Wednesday, June 29

Massachusetts Institute of Technology

Markov chain Monte Carlo (MCMC) methods are often used in clustering since they guarantee asymptotically exact expectations in the infinite-time limit. In finite time, though, slow mixing often leads to poor performance. Modern computing environments offer massive parallelism, but naive implementations of parallel MCMC can exhibit substantial bias. In MCMC samplers of continuous random variables, Markov chain couplings can overcome bias. But these approaches depend crucially on paired chains meetings after a small number of transitions. We show that straightforward applications of existing coupling ideas to discrete clustering variables fail to meet quickly. This failure arises from the "labelswitching problem": semantically equivalent cluster relabelings impede fast meeting of coupled chains. We instead consider chains as exploring the space of partitions rather than partitions' (arbitrary) labelings. Using a metric on the partition space, we formulate a practical algorithm using optimal transport couplings. Our theory confirms our method is accurate and efficient. In experiments ranging from clustering of genes or seeds to graph colorings, we show the benefits of our coupling in the highly parallel, time-limited regime.

Multiscale Analysis of Bayesian Cox Piecewise Constant Hazards Model

Bo Y.-C. Ning Wednesday, June 29

University of California, Davis

The piecewise constant prior is routinely used in the Bayesian Cox proportional hazards model (hereafter, Cox model) for survival analysis. Despite its popularity, the large sample properties of this Bayesian method have not yet been fully understood. In this talk, several new results will be given, including a Bernstein-von Mises theorem for the joint posterior distribution of the bivariate linear functional of the regression coefficients and the baseline hazard function, a Bayesian Donsker theorem for the conditional cumulative hazard and survival functions, and a sharp supremum-norm convergence rate for the conditional baseline hazard function. The novelties of our study include 1) we study the joint posterior distribution (of the regression coefficient and the baseline hazard function) instead of their marginal distributions; 2) we do not reply on conjugacy analysis, and hence the results are applicable to a wider class of priors, including those that are non-conjugate; 3) by adopting the multiscale analysis technique, the regularity of the true function only needs to be ?$\beta gt; 1/2$, which enabling us to deal with piecewise constant functions. The techniques we developed for studying the joint posterior distributions can serve as a useful tool for future studies of other semiparametric and nonparametric models. Simulation studies for ?finite sample datasets are conducted to verify our theory. Their results also reveal that the Bayesian method can be a competitive alternative to the frequentist approach.

Bayesian inference on impact of COVID-19 in G7 countries: A State space model.

Oluwadare Ojo Wednesday, June 29

Federal University of Technology Akure, Nigeria

This work examines the impact of Corona virus disease (COVID-19) on the stock market of Group of Seven (G7) countries using daily data from March, 1st of 2020 to December, 31st of 2020. A Bayesian Structural Time Series Model (BSTSM) was used to capture the effects of COVID-19 on the stock market performance of these G7 countries through a Markov Chain Monte Carlo (MCMC) method. We considered an AR(p) model with time-varying parameters and local linear trend models to know if the stock price of these countries during the period of the first wave of COVID-19 is changing overtime. There was a stochastic trend in stock prices of G7 countries during the period of the first wave of COVID-19 while the autoregressive process itself was also changing overtime. The stock market of the USA followed by Japan performed well than other G7 countries during the first phase of the COVID-19 pandemic while the stock market of France was affected during the COVID-19 pandemic.

Bayesian Analysis of Breast Feeding Pattern in Nigeria

Oladapo Oladoja Wednesday, June 29

First Technical University, Ibadan, Nigeria.

Adequate nutrition is essential to children's growth and development. To this end, breastfeeding is universally endorsed by the World's health and scientific organizations as the best way of feeding infants, therefore early initiation of breastfeeding is important for both mother and child because inadequate nutrition can lead to infant mortality. This study is aimed to give an update on the average number of last-born children that were breastfed across the geopolitical zones in Nigeria in 2008 and 2013. Bayesian approach was used to analyse this problem. The lastborn children breastfed in Nigeria follows a Normal density and using Normal prior yields a Normal Posterior. The average number of lastborn children breastfed (2008 and 2013 respectively) in North-Central (2473.676 and 1663.98), North-East (2678.324 and 2040.36), North-West (4903.023 and 4032.60), South-East (1697.717 and 1207.6), South-South (2295.461 and 1235.38) and South-West (2952.894 and 1685.03) were determined. It was recommended that there is a need for more enlightenment on the part of Government and health agencies to nursing mothers on the relevance of breastfeeding to the growth of a child and the health of mothers.

Individual model selection

Prince Peprah Osei Wednesday, June 29

Caleton University

In the study of decision-making patterns by individuals, such as the Iowa Gambling Task, there are competing models appropriate for a participant or group of participants. The subjects in the data must be allocated appropriately to these competing models. The calibration of the models requires re-running these models several times whenever an individual or set of individuals are moved or reshuffled among these models. Usually, we have a fixed computational budget and can only afford to run each model once. In this paper, we propose a model selection strategy that uses the Bayesian information criterion that is simple and easy to implement to allocate the subjects to these competing models. We illustrate our proposed model selection strategy by application to the Iowa Gambling Task data set.

Bayesian nonparametric panel Markov-switching GARCH models

Ayokunle Anthony Osuntuyi Wednesday, June 29

University Ca' Foscari Venice, Italy

This paper introduces a new model for panel data with Markov-switching GARCH effects to deal with regime changes and temporal clustering of conditional volatility and expected return in a large group of US financial assets. The model incorporates a series-specific hidden Markov chain process that drives the GARCH parameters. To cope with the high-dimensionality of the parameter space, the paper exploits the cross-sectional clustering of the series by first assuming a soft parameter pooling through a hierarchical prior distribution with a two-step procedure, and then introducing clustering effects in the parameter space through a Bayesian nonparametric prior distribution. A numerical analysis is carried out to evaluate the performance of the new model. More specifically, some simulation experiments are run along with an empirical application to financial returns data in the US.

Bayes linear emulation and history matching of the June model for the transmission of Covid-19

Jonathan Owen Wednesday, June 29

University of Leeds

The Covid-19 pandemic has resulted in global health, social and economic damage. Governments and NGOs have implemented increasingly complex intervention strategies to mitigate the effects of the pandemic warranting more detailed spatially resolved computer models for the transmission of Covid-19. The June model, developed at Durham University, UK, is a highly resolved, agent-based model which has been used to inform the UK Government and National Health Service strategy, as well as the UN for refugee camps. However, major limitations include: its complex structure; large number of inputs and outputs; the presence of many sources of uncertainty; further compounded by a substantial evaluation time. We apply robust Bayes linear emulation and history matching methodology to the June model for the UK Omicron wave, and in conjunction with the UN to Cox's Bazar, the world's largest refugee camp; in each case, to identify parameter settings which are consistent with the observed data.

Clustering Functional Data via Variational Inference

Camila P. E. de Souza Wednesday, June 29

University of Western Ontario

Functional data analysis (FDA) deals with data recorded densely over time (or any other continuum) with one or more observed curves per subject. Conceptually, functional data are continuously defined, but in practice, they are usually observed at discrete points. Among different kinds of functional data analyses, clustering analysis aims to determine underlying groups of curves in the data when there is no information on the group membership of each individual curve. In this work, we propose a new model-based approach for clustering and smoothing functional data simultaneously via variational inference (VI). Therefore, we derive a coordinate ascent variational inference (CAVI) algorithm to approximate the posterior distribution of our model parameters by finding the variational distribution with the smallest Kullback-Leibler divergence to the posterior. To our best knowledge, there are no studies in the literature on clustering functional data through VI. Our CAVI algorithm is implemented as an R package, and its performance is evaluated using simulated data and publicly available datasets.

Likelihood-Free Inference with Generative Neural Networks via Scoring Rule Minimization

Lorenzo Pacchiardi Wednesday, June 29

University of Oxford

Bayesian Likelihood-Free Inference methods yield approximate posteriors for simulator models with intractable likelihood. Some recent methods approximate the posterior using normalizing flows, namely neural networks implementing invertible transformations. When transforming, say, Gaussian random variables, normalizing flows parametrize a distribution whose density is accessible thanks to the invertibility property. The transformation weights can thus be optimized via maximum likelihood (equivalently, Kullback-Leibler divergence minimization) from simulated parameter-observation pairs. Here, we generalize this approach to the framework of Scoring Rules (SRs) minimization. SRs evaluate probability distributions with respect to an outcome; we employ them to assess how the approximate posterior for a simulated observation matches the corresponding parameter value. The training objective for “strictly proper" SRs is minimized when approximate and true posteriors coincide; additionally, some SRs can be minimized without evaluating the approximate posterior, but instead sampling from it. This allows relaxing the invertibility requirement and using more general neural networks.

NuZZ: numerical Zig-Zag sampling for general models

Filippo Pagani Wednesday, June 29

University of Cambridge

Markov chain Monte Carlo (MCMC) is a key algorithm in computational statistics, and as datasets grow larger and models grow more complex, many popular MCMC algorithms become too computationally expensive to be practical. Recent progress has been made on this problem through development of MCMC algorithms based on Piecewise Deterministic Markov Processes (PDMPs), irreversible processes that can be engineered to converge at a rate which is independent of the size of data. While there has understandably been a surge of theoretical studies following these results, PDMPs have so far only been implemented for models where certain gradients can be bounded, which is not possible in many statistical contexts. Focusing on the Zig-Zag process, we present the Numerical Zig-Zag (NuZZ) algorithm, which is applicable to general statistical models without the need for bounds on the gradient of the log posterior. This allows us to perform numerical experiments on: (i) how the Zig-Zag dynamics behaves on some test problems with common challenging features; and (ii) how the error between the target and sampled distributions evolves as a function of computational effort for different MCMC algorithms including NuZZ. Moreover, due to the specifics of the NuZZ algorithms, we are able to give an explicit bound on the Wasserstein distance between the exact posterior and its numerically perturbed counterpart in terms of the user-specified numerical tolerances of NuZZ.

Spatial Aggregation with Respect to a Population Distribution

John Paige Wednesday, June 29

NTNU

The common workflow for spatial aggregation based on point-referenced observations is to specify a spatial model, estimate it, and extract areal predictions. We focus on binary responses and small area estimation where the targets are areal prevalences. When the observations arise from a population of individuals, we argue that spatial aggregation requires a 'sampling frame model' that incorporates both a response model for the observations and uncertainty about the population. We demonstrate the importance of incorporating different sources of population uncertainty, and highlight that a spatial aggregation of risk is an expected prevalence and lacks some of the uncertainty of the prevalence.

We describe how different assumptions in the spatial aggregation model lead to other common approaches and compare them with the new approach. In a simulation study designed based on neonatal mortality rate (NMR) data from the 2014 Kenya demographic and health survey (KDHS2014), we show that incorporating population uncertainty can increase predictive variance at small aggregation scales. We demonstrate the practical implications for the analysis of NMR from KDHS2014 at different spatial scales.

Bayesian nonparametric disclosure risk assessment

Francesca Panero Wednesday, June 29

University of Oxford

Any decision about the release of microdata for public use is supported by the estimation of the number of sample uniques that are also population uniques, the most popular measure of disclosure risk. In such a context, parametric and nonparametric partition-based models have been shown to have the strength of leading to estimators with desirable features (ease of implementation, computational efficiency and scalability), and the weakness of producing underestimates in realistic scenarios, with the underestimation getting worse as the tail behaviour of the empirical distribution of microdata gets heavier. To fix this underestimation phenomenon, we propose a Bayesian nonparametric partition-based model that can be tuned to the tail behaviour of the empirical distribution of microdata. Our model relies on the Pitman–Yor process prior, and it leads to a novel estimator with all the desirable features of partition-based estimators and that, in addition, allows to reduce underestimation by tuning a “discount" parameter. We show the effectiveness of our estimator through its application to synthetic data and real data.

Bayesian methodology to perform statistical analysis of paired samples for count data

Alejandra Estefanía Patiño Hoyos Wednesday, June 29

IU Pascual Bravo

Counting results arise quite frequently in both clinical and engineering research. In the frequentist inferential context, significance statistical analysis for count data have focused mainly on the independent samples scenario, which does not cover the case in which, for example, pairs of measurements are taken from individual patients before and after a treatment. This experimental setup requires tests for paired samples. For this type of count results, neither does there a wide discussion in the literature associated with Bayesian statistical analysis. In this work we propose a statistical methodology under the Bayesian approach to compare paired samples for count data. For this purpose, we develop the Bayesian probabilistic model for the inferential analysis of the paired samples, later this model is adjusted and validated via simulation, and finally the proposed approach is applied in a set of data identified from the specialized literature.

Bayesian learning and forecasting of age-specific period mortality via locally adaptive spline processes

Federico Pavone Wednesday, June 29

Bocconi University

While the analysis of mortality has a long and well-established history, the attempt to accurately learn and forecast changes in mortality across ages and periods still attracts active research. We propose a novel locally adaptive spline process which is carefully constructed to incorporate the main structures of period mortality curves while allowing interpretable inference, accurate prediction and efficient computation within a single model. This is obtained by modeling age-specific mortality curves through spline functions with coefficients evolving over time via a novel dependent nested Gaussian process having locally varying smoothness. Inspired by Zhu and Dunson (2013), we formulate a multivariate version of the nested Gaussian process prior which allows to incorporate dependence across components. Spline bases combined with dependent nGP prior allow direct modeling of mortality trends and their derivatives with interpretable inference for age classes. The proposed model results appealing both for inference and prediction compared to competing alternatives in the literature.

Fast Bayesian estimation for ranking models

Michael Pearce Wednesday, June 29

University of Washington, Department of Statistics

The Mallows' model is a common location-scale family of ranking distributions. In the frequentist setting, estimation is challenging as the parameter space is high-dimensional and discrete, leading to slow, if not intractable, estimation. We propose a Bayesian procedure that allows for fast estimation of the Mallows' model via a posterior that can be expressed as a finite mixture of continuous parameter models. We demonstrate how prior distributions may be chosen to align Bayesian posteriors with those from the frequentist setting, allowing for substantial increases in estimation speed. In addition, we show how our procedure may be applied to various extensions of the Mallows' model in the literature. The model is applied to real grant panel review data to demonstrate fast preference aggregation and how to estimate its uncertainty.

Bayesian Nonparametric Predictive Modeling for Personalized Treatment Selection

Matteo Pedone Wednesday, June 29

University of Florence

Integrating genomics into clinical oncology has gained significant attention in the field of personalized medicine (PM). PM's mission is to tailor treatment to individual patient characteristics, leveraging various sources of heterogeneity. We propose a Bayesian nonparametric predictive model for personalized treatment selection that leverages prognostic and predictive biomarkers. We built an integrative predictive framework adopting a product partition model with covariates to induce clusters of observations that are more homogeneous with respect to predictive biomarkers, building partitions that are only partially exchangeable. Inference is conducted through a Markov Chain Monte Carlo algorithm. We demonstrate the capability of the method to select the optimal treatment through simulation studies. Finally we illustrate the results obtained with our approach on publicly available data of lower grade glioma.

Impact Assessment Methods Based on Multivariate and Spatial Dynamic Bayesian Models: An Application in the Assessment of the Impact of Strict Confinement by Covid-19 on Air Quality in the City of Medellín

Carlos Andrés Pérez Aguirre Wednesday, June 29

Master's Student

The preventive isolation or strict quarantine measures caused by the SARS-CoV-2 (Covid-19) pandemic during the month of April 2020 caused a drastic reduction in vehicle flow and industrial activities. Thus, in this project we propose a statistical method based on dynamic spatio-temporal models with a Bayesian approach to estimate the impact of the reduction of vehicular flow and industrial activities in the presence of pollutants (PM 10, PM 2.5, NO, NO2, NOx), that degrade air quality in the city of Medellín and some municipalities of the Metropolitan Area. The impact is computed by building a counterfactual model and calculating the difference between it and the values observed after the strict quarantine period, the results are evaluated under different scenarios and specifications of the spatio-temporal models using statistical simulation tools.

Media Bias and Polarization via a Markov-Switching Latent Space Network Model

Antonio Peruzzi Wednesday, June 29

Ca' Foscari Venice

The news consumption landscape has drastically changed in the last decades and several old issues return to the fore. One of these is whether and to which extent news outlets bias information. We propose a new dynamic latent-space model (LS) for news outlets in which we exploit both time-varying online audience duplication-network data as well as textual contents from published articles to measure media bias and social-media polarization. Our model, estimated within the Bayesian framework, recovers the latent coordinates of news outlets in a 2-dimensional euclidean space, while providing a proper interpretation respectively in terms of media slant and online engagement. The aim is twofold: making advancements both concerning the analysis of the timely evolution of audience duplication networks and concerning the determination of media slant and polarization. The developed model is applied to a Facebook dataset regarding European news outlets in the years 2015 and 2016.

A Bayesian factor-based calibration model: An application to measuring the impact of sustainability on option-implied distributions

Giovanni Pianon Wednesday, June 29

Ca' Foscari University of Venice

Option-Implied data represent an essential source of information for modeling the distribution of stock returns. However, as the scientific literature has shown, agents' risk-preferences and irrationality can lead to a severe misalignment from the objective distribution. We propose a factor-based calibration model to correct for such a bias. By linking the calibration function parameters to a set of common latent factors, the model captures the time-varying nature of the forces driving the distortion of Option-Implied distributions and allows for cross-sectional information pooling in implementing univariate calibration. A Bayesian inference approach and an efficient Monte Carlo posterior approximation allow us to deal with the model's high-dimensionality and intractability. We employed the model to perform density forecasting and to shed light on the underlying factors determining the bias of risk-neutral densities, investigating, in particular, the role of ESG ratings and other sustainability indicators

Sequentially guided MCMC proposals for synthetic likelihoods and correlated synthetic likelihoods

Umberto Picchini Wednesday, June 29

Chalmers University of Technology and University of Gothenburg

Synthetic likelihood (SL) is a simulation-based strategy for parameter inference when the likelihood function is intractable. In SL, the likelihood function of the data is replaced by a multivariate Gaussian density over summary statistics of the data. SL requires the simulation of many datasets at each parameter of a sampling algorithm, such as MCMC, making the method computationally-intensive. We propose two strategies alleviating the computational burden [1]. First, we introduce an algorithm producing a proposal distribution that is sequentially tuned and made conditional to data, thus rapidly “guiding" the parameters towards high posterior density regions. This could be potentially used also with MCMC samplers for ABC. Second, we exploit strategies borrowed from the correlated pseudo-marginal MCMC literature, to increase chains mixing. Our methods enable inference via SL for challenging case studies, when the posterior is multimodal or when the chain is initialized in low posterior probability regions, where standard samplers failed.

[1] U. Picchini, U. Simola, J. Corander (2022). Sequentially guided MCMC proposals for synthetic likelihoods and correlated synthetic likelihoods. Bayesian Analysis, doi:10.1214/22-BA1305.

Gaussian processes on Hypergraphs

Thomas Pinder Wednesday, June 29

Lancaster University

We derive a Matern Gaussian process (GP) on the vertices of a hypergraph. This enables estimation of regression models of observed or latent values associated with the vertices, in which the correlation and uncertainty estimates are informed by the hypergraph structure. We further present a framework for embedding the vertices of a hypergraph into a latent space using the hypergraph GP. Finally, we provide a scheme for identifying a small number of representative inducing vertices that enables scalable inference through sparse GPs. We demonstrate the utility of our framework on three challenging real-world problems that concern multi-class classification for the political party affiliation of legislators on the basis of voting behaviour, probabilistic matrix factorisation of movie reviews, and embedding a hypergraph of animals into a low-dimensional latent space.

Accurate skewed asymptotic approximations of posterior distributions

Francesco Pozza Wednesday, June 29

University of Padova

Common Bayesian asymptotic theory results rely on convergence of posterior distributions to a Gaussian limiting one. However, such a limiting behavior may require an excessively large sample size before becoming visibile in practice. In fact, in situations with small-to-moderate sample size, even simple parametric models often yield posterior distributions which are far from being Gaussian, typically due to skewness and kurtosis. By adopting as limiting law a skewed generalization of the Gaussian distribution, we show that it is possible to obtain substantially more accurate results, both theoretical and empirical, in finite-sample and asymptotic regimes.

Measurement Error Models for Spatial Network Lattice Data: Analysis of Car Crashes in Leeds

Luca Presicce Wednesday, June 29

University of Milan-Bicocca

Road casualties represent an alarming concern for modern societies demanding evidence-based interventions. Statistical models for road safety analysis typically include socio-economic variables and traffic volumes. However, the latter variables usually suffer from measurement error (ME), which can severely bias the statistical inference. This paper presents a Bayesian hierarchical model to analyse car crashes occurrences taking into account ME in the spatial covariates and the lattice structure of the road segments. Using a CAR prior, this work introduces a spatial dependence structure within the classical ME model. The suggested methodology is exemplified considering road collisions in the road network of Leeds (UK). Traffic volumes are approximated at the street segment level using an extensive dataset obtained from mobile devices. Estimation was carried out with the INLA methodology, which allows for computational advantages. Our results show that omitting ME adjustment considerably worsens the model's fit and attenuates the effects of imprecise covariates.

Parametric modeling of strong ground motions

Isaias Ramirez Wednesday, June 29

CIMAT

A parametric characterization of earthquakes accelerograms in Mexico it is proposed. To obtain this characterization it was necessary to fit a Bayesian non linear regression, in which informative prior distributions were used. The idea essentially is to exploit that the evolutionary power spectrum (ePS) is a non negative function. Using that, we can see it as a joint density in time and frequency letting us estimate it. The ePS can be easily parameterized in its marginals, it is necessary to fit this parameters to each record, and then made a regression using as explicative variables characteristics of the earthquake as its magnitud or distance to epicenter.

Thursday Session

Optional Pólya trees: posterior rates and uncertainty quantification

Thibault Randrianariosa Thursday, June 30

Sorbonne Université

We consider statistical inference in the density estimation model using a tree-based Bayesian approach, with Optional Pólya trees as prior distribution. We derive near-optimal convergence rates for corresponding posterior distributions with respect to the supremum norm. For broad classes of Hölder-smooth densities, we show that the method automatically adapts to the unknown Hölder regularity parameter. We consider the question of uncertainty quantification by providing mathematical guarantees for credible sets from the obtained posterior distributions, leading to near-optimal uncertainty quantification for the density function, as well as related functionals such as the cumulative distribution function. The results are illustrated through numerical simulations.

Modelling of aortic dissection as Beta random fields and uncertainty propagation with a Bayesian variational auto-encoder

Sascha Ranftl Thursday, June 30

Graz University of Technology

Aortic dissection is a dangerous disease that is linked to stochastic, spatially heterogeneous degradation of aorta tissue. Here, this degradation is modelled as a Beta random field. Based on this, the mechanical stresses in the tissue, in response to a mechanical load, are computed with Finite Elements. The computational effort of this latter step makes direct sampling infeasible, requiring a surrogate model instead. The structure of the input data, random field realizations, and the output data, mechanical stress fields, define a problem similar to image-to-image (I2I) regression. This I2I problem structure suggests a Bayesian Variational Auto-Encoder (B-VAE) as a surrogate. After training with Stein Variational Gradient Descent, the B-VAE can subsequently faster predict approximate stress field samples from random field samples, and with non-parametric Variational Inference estimate the PDF of the predicted mechanical stress fields. A rigorous Bayesian formulation is presented, and the biological implications of the results are discussed.

Modeling Infectious Disease Dynamics: Integrating Contact Tracing-based Stochastic Compartment and Spatio-temporal Risk Models

André Victor Ribeiro Amaral Thursday, June 30

King Abdullah University of Science and Technology

Major infectious diseases such as COVID-19 significantly impact population lives and put enormous pressure on healthcare systems globally. Strong interventions, imposed to prevent these diseases from spreading, may also negatively impact society, leading to jobs losses and mental health problems—making crucial the prioritization of riskier areas when applying these protocols. The modeling of mobility data derived from contact-tracing data can be used to forecast infectious trajectories and help design strategies for prevention and control. In this work, we propose a new spatio-temporal stochastic model that allows us to characterize the temporally varying spatial risk better than existing methods. We demonstrate the use of the proposed model by simulating an epidemic in Valencia, Spain, and comparing it with a contact tracing-based stochastic compartment reference model. The results show that by accounting for the spatial risk values in the model, the peak of infected individuals and the overall number of infected cases are reduced. Thus, adding a spatial risk component into compartment models may give finer control over the epidemic dynamics, helping policymakers to make better decisions.

Modeling sparse graphs with overlapping communities via thinned random measures

Federica Zoe Ricci Thursday, June 30

University of California, Irvine

Classic exchangeable graph models, including most stochastic block models, generate dense graphs with a limited ability to capture many characteristics of real-world social and biological networks. A class of models based on completely random measures like the generalized gamma process (GGP) have recently addressed some of these limitations. By thinning edges from realizations of GGP random graphs, here we define a novel sparse graph model in which nodes have mixed memberships in a potentially large set of latent communities. Our representation of sparse graphs as thinned random measures enables efficient Monte Carlo methods that scale linearly with the number of observed edges, and thus (unlike dense block models) sub-quadratically with the number of entities or nodes. We demonstrate effective recovery of latent community structure on real-world networks with thousands of nodes.

Ensemble Learning with Generalized Predictive Synthesis

Joseph Rilling Thursday, June 30

Temple University

We discuss the ensemble of multiple forecasts under bias, misspecification, and dependence. While linear combination strategies are ubiquitous in the literature, we show that this strategy is fundamentally flawed, being underparameterized when all forecasts are misspecified. To develop a method to overcome this, which we call generalized predictive synthesis, we propose a novel theoretical strategy based on stochastic processes that identifies a more general class of ensemble methods. Examining the predictive properties of this new class, we identify the conditions and mechanism for which generalized predictive synthesis improves over linear combinations, in terms of expected squared forecasts error. We present and extensive simulation study as well as two real applications and show that this method improves over existing strategies.

Metropolis Adjusted Langevin Trajectories: a robust alternative to Hamiltonian Monte Carlo

Lionel Riou-Durand Thursday, June 30

University of Warwick

Hamiltonian Monte Carlo (HMC) is a widely used sampler, known for its efficiency on high dimensional distributions. Yet HMC remains quite sensitive to the choice of integration time. Randomizing the length of Hamiltonian trajectories (RHMC) has been suggested to smooth the Auto-Correlation Functions (ACF), ensuring robustness of tuning. We present the Langevin diffusion as an alternative to control these ACFs by inducing randomness in Hamiltonian trajectories through a continuous refreshment of the velocities. We connect and compare the two processes in terms of quantitative mixing rates for the 2-Wasserstein and L2 distances. The Langevin diffusion is presented as a limit of RHMC achieving the fastest mixing rate for strongly log-concave targets. We introduce a robust alternative to HMC built upon these dynamics, named Metropolis Adjusted Langevin Trajectories (MALT). Studying the scaling limit of MALT, we obtain optimal tuning guidelines similar to HMC, and recover the same scaling with respect to the dimension without additional assumptions. We illustrate numerically the efficiency of MALT compared to HMC and RHMC.

Calibration of Bayesian Predictive Distributions using Generalised Bayesian Inference

Oliver Robinson Thursday, June 30

University of Warwick

When constructing prediction models, we frequently encounter the problem of miscalibration of our posited model with the true data generating mechanism. To deal with this, we take this simple, miscalibrated model and then use Generalised Bayesian Inference to improve robustness by introducing a learning rate parameter, which needs to be tuned. Approaches to do this include Wu and Martin's GPrC algorithm, which tunes this parameter by ensuring it achieves a specific coverage probability consistent with the empirical data. Our approach extends this by computing the corresponding tuning parameter for a range of different coverage probabilities, thereby allowing computation of importance sampling weights for a predicted value from our misspecified model. This achieves correction of the full model misspecification, rather than just for a specific quantile, hence enabling accurate, calibrated prediction in settings such as when we have spatio-temporal data with extreme values.

Estimation of Optimal Dynamic Treatment Regimes via Gaussian Processes

Daniel Rodriguez Duque Thursday, June 30

McGill University

In precision medicine, identifying optimal sequences of decision rules, termed dynamic treatment regimes (DTRs), is an important undertaking. One approach investigators may take to infer about optimal DTRs is via Bayesian Dynamic Marginal Structural Models (MSMs). These model an outcome under adherence to a DTR; unfortunately, straightforward models may lead to bias in the optimum. If computationally tractable, a grid search for the optimal DTR may obviate this difficulty. We seek to alleviate these inferential challenges by pairing Gaussian Process (GP) optimization methods with estimators for the causal effect of adherence to a DTR. We examine how to identify optimal DTRs in multi-modal settings and conclude that a GP modeling approach that acknowledges noise in the estimated response surface leads to improved results. Additionally, we find that a grid search may not always be a robust solution. We exemplify the proposed methods by analyzing a clinical dataset on HIV therapy.

Bayesian density estimation on an unknown submanifold with Dirichlet process mixtures

Paul Rosa Thursday, June 30

University of Oxford

In order to overcome the curse of dimensionality in multivariate analysis, it is common to assume that the data points are supported on a low-dimensional submanifold of the ambient space. Here, we consider the framework where the data live on a small tubular neighbourhood around an unknown submanifold, encompassing in particular the case of noisy manifold-valued data. We propose a Bayesian nonparametric approach to estimate the density based on location-scale Dirichlet process mixtures of Gaussians. We study the posterior concentration rates in such models, investigating in particular the combined effects of the low dimensionality of the sub-manifold, the width of the tubular neighbourhood and the (anisotropic) regularity of the density. An interesting aspect of our results is that it sheds light on the advantage of certain types of location-scale mixtures over location mixtures for such setups.

Flood hazard model calibration using multiresolution model output

Samantha Roth Thursday, June 30

Pennsylvania State University Department of Statistics

Riverine floods occur when water levels exceed the capacity of their channels, posing challenges to life and property. Hydraulic models can project riverine flood heights, and these projections can inform risk management policies. The input parameters for hydraulic models can be highly uncertain. Calibration methods use observations to quantify parametric uncertainty. With limited computational resources, calibration often proceeds with either a small number of expensive high resolution model runs or many cheaper low resolution runs. We propose a Gaussian process-based Bayesian emulation-calibration approach that assimilates model outputs and observations at multiple resolutions. We demonstrate our approach on the LISFLOOD flood hazard model for a flood-prone region of the Susquehanna River basin. Compared to existing single-resolution approaches, our method yields more accurate parameter estimates and flood predictions. The problem of utilizing model runs at different resolutions occurs in many scientific disciplines. Hence, our methodology has the potential to be broadly applicable.

A Bayesian approach for high dimensional spatial skewed processes

Paritosh Kumar Roy Thursday, June 30

McGill University

Environmental processes are commonly observed across different locations and result in skewed distributions. Recent proposals in the literature that handle the data in their original scale involve two independent Gaussian processes (GP) that capture the data's skewness and spatial structure. We aim at investigating how recent proposals that approximate high dimensional GPs perform in the estimation of skewed processes built through marginal skew-normal or log-Gaussian processes. In particular, we focus on the Nearest Neighbor GP (NNGP), the lattice kriging approach, and reduced-rank GP. We investigate the performance of these approximations with artificial data and temperature data observed across 1000 locations over the grid of latitude and longitude values of [34.29519, 37.06811] and [-95.91153, -91.28381]. The results suggest that the NNGP approach performs consistently better than the lattice kriging and reduced-rank approaches in terms of the mean absolute error, root-mean-squared error, prediction interval coverage, and Watanabe-Akaike information criterion.

An UnCover analysis of bed pathways and hospital Length of Stay (LoS): A retrospective cohort study of Colombian COVID-19 patients

Lina Marcela Ruiz Galvis Thursday, June 30

Universidad de Antioquia

Understanding the impact of COVID-19 on hospital capacity requires the estimation of the total Length of Stay (LoS) in a particular bed type (i.e. Hospital or Intensive Care Unity bed). Here we used an Accelerated Failure Time Model (AFT) to estimate the LoS for COVID-19 Colombian patients. We used the R software to write this Linear regression model using Bayesian analysis in JAGS. The database includes fifty thousand patients between March 2020 to August 2021. The model allows to identify whether age, gender, vaccination period are covariable affecting the LoS. It also allows estimating whether LoS are different according to the patient outcome (i.e. recovered, deceased, transferred). With this first approach, past data can be appropriately used to better prepare for the next phases of the COVID-19 pandemic.

Efficient estimation of joint models for multivariate longitudinal and survival data using Integrated Nested Laplace Approximations (INLA)

Denis Rustand

King Abdullah University of Science and Technology

Joint models for longitudinal and survival outcomes have recently gained a lot of interest in clinical research. These complex models involve multiple likelihoods (i.e. for each longitudinal and survival outcome), usually linked through correlated or shared random-effects. In this context, exact inference methods reach limitations due to long computation times and convergence issues. We introduce a Bayesian approximation for these joint models based on the INLA algorithm implemented in the R package INLA to alleviate the computational burden and allow the estimation of multivariate joint models with less restrictions. Our simulation studies show that INLA reduces the computation time substantially as well as the variability of the parameter estimates compared to alternative strategies. We further apply the methodology to analyze 5 longitudinal markers (3 continuous, 1 count, 1 binary, and 16 random effects) and competing risks of death and transplantation in a clinical trial on primary biliary cholangitis. INLA provides a fast and reliable inference technique for applying joint models to the complex multivariate data encountered in health research.

Analysis of ARGO float data via multivariate type G Matérn stochastic partial differential equation random fields

Damilya Saduakhas Thursday, June 30

King Abdullah University of Science and Technology

Gaussian random fields with multivariate cross-covariance functions have been widely used to model spatial data sets with multiple variables, like climate data. However, many datasets have features that cannot be captured by Gaussian models, like non-Gaussian dependence or exponential tails. In this work, we wish to further develop our knowledge of multivariate Non-Gaussian random fields. In particular, we will look at a class of multivariate non-Gaussian models based on systems of stochastic partial differential equations as proposed by Bolin and Wallin (2020). We will extend the work by Bolin and Wallin (2020) by investigating suitable parametrizations for p-variate random fields, there p>3. Further, the random fields will be included in geostatistical models with correlated measurement errors, which will be fitted to data through stochastic gradient descent methods. An application to multivariate modeling of Argo float data (ocean temperature, pressure) will be investigated.

Development of a spatially regularized detector for emergent/re-emergent disease outbreaks

Cosmin Safta Thursday, June 30

Sandia National Laboratories

We propose a Bayesian framework to detect outbreaks of emerging diseases. Our method triggers when the observed case-count data disagree significantly from forecasted levels. Existing detectors compare data to baseline levels, which are difficult to set for emerging diseases because of noisy/missing data. In contrast, forecasting allows us to use socioeconomic parameters and spatiotemporal data on disease prevalence to compensate for low-quality epidemiological information. The proposed framework incorporates Poisson statistics to accommodate low-counts and Markov random fields to account for spatio-temporal correlations in the disease spread rate. Posterior distributions would be approximated with both via Markov-Chain Monte Carlo (at regional level) and variational inference (at the national level). Results computed with COVID-19 data from Nee Mexico, US, will be used to demonstrate the method. A figure-of-merit will be the time/date at which the new method detects the start of the outbreak.

Bayesian Joint Modeling and Selection among Many Biomarkers Measured Longitudinally

Soumya Sahu Thursday, June 30

University of Illinois Chicago

The literature on joint modeling is diverse; however, current approaches can typically jointly analyze one or a few longitudinal processes and a time-to-event outcome. This work is motivated by imaging features of the eye, measured longitudinally at multiple visits of patients with early-stage age-related macular degeneration (AMD). A primary scientific question in this context is selection of a panel of features that can prognosticate conversion to neovascular AMD. We develop a Bayesian nonparametric joint model that (1) flexibly models the longitudinal trajectories, (2) provides flexibility in modeling the association between the longitudinal processes and time-to-event outcome, and (3) addresses selection among the multiple longitudinally measured features. We compare performance of the proposed approach with other existing models and machine-learning models used in scientific literature. Our analysis of all imaging features simultaneously in the proposed model highlights unique features associated in multiple ways with prognostication of conversion to neovascular AMD that are distinct from findings based on marginal joint modeling involving one longitudinal feature at a time.

Leveraging Bayesian Methods to Quantify Uncertainty in Inverse Optimization

Nathan Sandholtz Thursday, June 30

Brigham Young University

In applied optimization tasks, human decisions often exhibit "noise" around theoretical optimums. Noise may arise from measurement error and/or human error in decision-making. The presence of noise adds significant complexity when considering the inverse optimization problem, which is the problem of inferring unknown features of an optimization model such that a collection of observed decisions are rendered as optimal with respect the inferred model. The Bayesian paradigm provides a natural framework to make inference on the unknown optimization model parameters while simultaneously enforcing optimality constraints. In this paper, we assume a likelihood over observed human decisions, but we constrain a pre-specified functional of the likelihood (e.g. the mean) to be inversely optimal by careful construction of the prior distributions. We then approximate the posterior distributions via MCMC methods, enabling us to generate credible regions for the unknown model parameters, all of which adhere to the required optimality constraints. This is joint work with Timothy Chan and Nasrin Yousefi.

Unsupervised attack pattern detection in cyber-security using Bayesian topic modelling

Francesco Sanna Passino Thursday, June 30

Imperial College London

Cyber-systems are constantly under threat of intrusion attempts. Attacks are usually carried out with one underlying specific intent, or from groups of actors with similar objectives. Therefore, discovering such patterns is extremely valuable to threat experts. From a statistical point of view, this objective translates into a clustering task. This work explores Bayesian topic models for clustering session data collected on honeypots, particular hosts designed to entice malicious intruders. These session commands provide a rare insight into the operational modes of cyber attackers, such as their automated or interactive nature, the individual scripting styles and their overall objectives. The main practical implications of clustering the sessions are two-fold: finding similar groups and identifying outliers. An array of Bayesian models are considered, suitably adapted to the challenges encountered with computer network data. In particular, the concepts of primary topics, session-level and command-level topics are introduced, along with a secondary topic for instructions representing common high frequency commands. Furthermore, the proposed method is extended to allow for an unbounded vocabulary size and number of latent intents. The methodologies are used to discover an unusual MIRAI variant which attempts to take over existing coin miner infrastructure.

The Bayesian “Time Machine" for Multi-arm Randomized Clinical Trials

Ben Saville Thursday, June 30

Berry Consultants

Multi-arm platform trials investigate multiple agents simultaneously, typically with staggered entry and exit of experimental treatment arms versus a shared control arm. In such settings, there is considerable debate whether to limit analyses for a treatment arm to concurrently randomized control subjects, or to allow comparisons to both concurrent and non-concurrent (pooled) control subjects. The potential bias from temporal drift over time is at the core of this debate. We propose time-adjusted analyses, including a Bayesian “Time Machine", to model potential temporal drift in the entire study population, such that primary analyses can incorporate all randomized control subjects from the platform trial. We conduct a simulation study to assess performance relative to utilizing concurrent or pooled controls. In multi-arm platform trials with staggered entry, analyses adjusting for temporal drift have superior estimation of treatment effects and favorable testing properties compared to analyses using either concurrent or pooled controls.

Real-time clinical Trial Oversight via Online Non-parametric Bayesian Structure Learning of Probabilistic Programs

Ulrich Schaechtle Thursday, June 30

Massachusetts Institute of Technology

Bayesian statisticians typically analyze data in later phases of clinical trials. If analyses were available from the earliest stages, we could catch and correct avoidable issues early, preventing trial failure and unnecessarily high costs.

We have developed a new Bayesian approach to early and continuous analysis of ongoing clinical trials based on probabilistic programming. Our approach builds on online Bayesian structure learning for probabilistic programs via sequential Monte Carlo inference given a non-parametric prior. This automates time-consuming model-building that would otherwise have to be done by statisticians with clinical trial expertise. We then use fast symbolic inference to condition the learned probabilistic programs on incoming trial data. These capabilities are integrated into InferenceQL, a probabilistic programming platform designed to make automated Bayesian inference sufficiently automated so that domain experts can apply it to solve routine inference problems.

By computing conditional probabilities in high dimensions, automatically learned models help identify unlikely data points (e.g. data entry errors, or violations of inclusion/exclusion criteria), discover confounders, and reveal dataset drift. We will provide quantitative evidence that this approach is not just appealing in principle but also effective in practice, using a combination of simulation studies with known ground truth and applications to real-world trials in multiple disease areas. Our poster focuses on results from one phase III oncology trial, but the approach we propose has so far been successfully applied to three clinical trials at Takeda, including two in oncology and one in gastrointestinal disease.

Scalable and flexible UQ framework for an interconnected system of codes and data in nuclear science

Georg Schnabel Thursday, June 30

IAEA

Nuclear science covers many subdomains, such as nuclear astrophysics, reactor physics and nuclear medicine. They rely on nuclear data, i.e., probabilities of various physical interactions of the atomic nuclei, and predictions of computer codes, e.g., to simulate particle collisions and the propagation of particles in nuclear reactors. These codes may take minutes to hours and there is a complex interplay between them. Collision codes are incorporated into reactor codes and simulations rely on nuclear data which are the result of a calibration process. Many sources of uncertainty at various levels exist, and also physical constraints must be considered. A flexible framework to interlink codes and do rigorous uncertainty quantification in such systems involving codes, data and IT infrastructure can be an accelerator to improve all components involved. We discuss and demonstrate that Gaussian processes in combination with Bayesian networks are a promising and scalable framework for this purpose.

Bayesian group sequential designs for cluster-randomized trials

Junwei Shen Thursday, June 30

McGill University

Adaptive approaches, allowing for more flexible trial design, have been proposed for individually randomized trials to save time or reduce sample size. However, adaptive designs for cluster-randomized trials in which groups of participants are randomized to treatment arms are less common. Motivated by a potential real-world cluster-randomized trial, two Bayesian group sequential designs for cluster-randomized trials are proposed to allow for early stopping for efficacy at pre-planned interim analyses. The difference between the two designs lies in the way that participants are sequentially recruited. The design operating characteristics are explored via simulations for a variety of scenarios and two outcome types for the two designs. The simulation results show that for different outcomes the design choice may be different. We make recommendations for Bayesian group sequential designs for cluster-randomized trial based on the simulation results.

Distribution Compression in Near-linear Time

Abhishek Shetty Thursday, June 30

University of California, Berkeley

In distribution compression, one aims to accurately summarize a probability distribution P using a small number of representative points. Near-optimal thinning procedures achieve this goal by sampling n points from a Markov chain and identifying $\sqrt{n}$ points with $\widetilde{O}(1/\sqrt{n})$ discrepancy to P. Unfortunately, these algorithms suffer from quadratic or super-quadratic runtime in the sample size n. To address this deficiency, we introduce Compress++, a simple meta-procedure for speeding up any thinning algorithm while suffering at most a factor of 4 in error. When combined with the quadratic-time kernel halving and kernel thinning algorithms of Dwivedi and Mackey (2021), Compress++ delivers $\sqrt{n}$ points with $O(\sqrt{\log n/n})$ integration error and better-than-Monte-Carlo maximum mean discrepancy in $O(n \log^3 n)$ time and $O(\sqrt{n} \log^2 n)$ space. Moreover, Compress++ enjoys the same near-linear runtime given any quadratic-time input and reduces the runtime of super-quadratic algorithms by a square-root factor. In our benchmarks with high-dimensional Monte Carlo samples and Markov chains targeting challenging differential equation posteriors, Compress++ matches or nearly matches the accuracy of its input algorithm in orders of magnitude less time.

Randomized Stein Points (RaSPs): efficient and confident sampling of expensive posterior distributions

Chen Shi Thursday, June 30

Duke University

Stein points is a deterministic sampling method that generates samples for an un-normalized Bayesian posterior distribution by minimizing the kernel stein discrepancy between the point set and the target distribution. However, current optimization of the discrepancy measure depends on random search over the parameter space and requires numerous evaluations of the posterior distribution. Moreover, since the algorithm is deterministic, stein points do not provide inference on estimated quantities, which are usually integrals over posterior samples. My study focuses on developing a randomized version of stein points that could provide probabilistic confidence intervals on desirable quantities for error estimation. In addition, surrogate models are used to improve optimization and reduce the required evaluations of posterior distribution. With these improvements, randomized stein points can be utilized to sample from expensive posterior distribution. Given a constrained computational budget, randomized stein points can generate samples that have a higher effective size and provide narrower confidence intervals on estimated quantities than Markov Chain Monte Carlo.

Joint modelling of spatio-temporal forest fire ignition, count and burned proportion outcomes

Giovani Silva Thursday, June 30

University of Lisbon

In this work, we present Bayesian hierarchical models to jointly analyze distinct data formats involving discrete, categorical and continuous outcomes. In the proposed modelling, latent processes are adopted to characterize the dependency among different type outcomes. Our modeling motivation involves forest fire data in Portugal, where the outcomes are usually occurrence of fire (binary), number of fires (count), and burned area (proportion). Since these responses are observed by region over time, this work aims to analyze spatiotemporal forest fire data when the fire ignition, number of fires, and proportion of burned area are jointly modeled. We look for space and time effects on these three outcomes among municipalities over last years. For getting estimates of the model parameters, we have used Integrated Nested Laplace Approximation (INLA) methods, as well as for some short term prediction. This is joint work with Elias Krainski, Denis Rustand, and Haavard Rue.

A Joint Bayesian Framework for Measurement Error and Missing Data in INLA

Emma Sofie Skarstein Thursday, June 30

Norwegian University of Science and Technology

Missing data and measurement error are common problems in most applied data sets. However, while the former is receiving considerable attention, many researchers are still not routinely accounting for varying types of measurement error in their variables. By viewing missing data as a limiting case of measurement error, we propose Bayesian hierarchical models to account for continuous covariate measurement error and missingness simultaneously. The investigated models encompass both the well-known classical measurement error, but also the less considered Berkson measurement error, which often occurs, among other places, in experimental setups and exposure studies. The Bayesian framework is very flexible, and allows us to incorporate any prior knowledge we may have about the nature of the measurement error. We illustrate how the respective methods can then be efficiently implemented via integrated nested Laplace approximations (INLA).

Estimating COVID-19 incidence and infection fatality rates

Justin Slater Thursday, June 30

University of Toronto

Naive estimates of incidence and infection fatality rates (IFR) of COVID-19 suffer from a variety of biases, many of which relate to preferential testing. This has motivated epidemiologists from around the globe to conduct serosurveys that measure the immunity of individuals by testing for the presence of SARS-CoV-2 antibodies in the blood. These quantitative measures (titre values) are then used as a proxy for previous or current infection. However, statistical methods that use this data to its full potential have yet to be developed. Previous researchers have discretized these continuous values, discarding potentially useful information. In this work, we demonstrate how multivariate mixture models can be used in combination with poststratification to estimate cumulative incidence and IFR in a(n) (approximate) Bayesian framework without discretization. In doing so, we propagate error from both the estimated number of infections and deaths to provide estimates of IFR. This method is demonstrated using data from the Action to Beat Coronavirus (Ab-C) serosurvey in Canada.

Bayesian modelling of weekly mortality incorporating annual trends

Peter Smith Thursday, June 30

University of Southampton

The COVID-19 pandemic has focused attention on methods for calculating excess deaths. Statistics based entirely on cause-of-death data underestimate the true toll of the pandemic, for example due to the difficulty in determining underlying cause of death and the potential for indirect effects on other causes. This work contributes to the growing excess mortality literature by incorporating information about long-term mortality trends from an annual forecasting model. This provides a more realistic baseline for excess mortality estimation, whilst enabling decomposition of weekly deaths by single year of age according to established mortality age patterns. Within a coherent Bayesian framework that fully integrates all sources of uncertainty, we investigate the utility and suitability of Gaussian Process and dynamic harmonic regression models for capturing weekly mortality variation. Obtaining estimates of the weekly deaths expected in the absence of the pandemic, we estimate age-specific and overall excess deaths and compare with alternative estimates.

Co-clustering of Spatially Resolved Transcriptomic Data

Andrea Sottosanti Thursday, June 30

University of Padova

Spatial transcriptomics is a modern sequencing technology that allows measuring the activity of thousands of genes in a tissue sample and mapping where the activity is occurring. This technology has enabled the study of those genes which exhibit spatial variation across the tissue. Comprehending their functions and their interactions is fundamental to understanding several key biological mechanisms. However, adequate statistical tools that exploit the newly spatial mapping information are still lacking. We introduce SpaRTaCo, a new model that clusters the spatial expression profiles of the genes according to the tissue areas. This is accomplished by performing a co-clustering, i.e., inferring the latent block structure of the data and inducing a simultaneous clustering of the genes and of the tissue areas. Our proposed methodology is validated with a series of simulations and its usefulness in responding to specific biological questions is illustrated with an application to a human brain sample.

How to publish with Chapman & Hall/CRC Press

Lara Spieker Thursday, June 30

Chapman & Hall/CRC Press

The poster will discuss and display how the publishing process from the initial idea to the proposal, and proposal review process, and the writing process aimed at prospective authors to provide best practices for shaping your ideas and submitting a book proposal.

Bayesian calibration of imperfect computer models using Physics-informed priors

Michail Spitieris Thursday, June 30

NTNU

We introduce a computational efficient data-driven framework suitable for quantifying the uncertainty in physical parameters of computer models represented by differential equations. We construct physics-informed priors for differential equations, which are multi-output Gaussian process (GP) priors that encode the model's structure in the covariance function. We extend this into a fully Bayesian framework which allows quantifying the uncertainty of physical parameters and model predictions. Since physical models are usually imperfect descriptions of the real process, we allow the model to deviate from the observed data by considering a discrepancy function. We use Hamiltonian Monte Carlo (HMC) sampling to obtain the posterior distributions. To demonstrate our approach, we use the arterial Windkessel model, which is a time-dependent differential equation with interpretable physical parameters that are considered important to hypertension. We also apply our approach to the heat equation, a space-time dependent partial differential equation.

Spatial non-stationary models for pavement deterioration

Ingelin Steinsland Thursday, June 30

NTNU

Maintaining a sufficient quality of the road pavement surface is of high importance for retaining safe driving conditions. Rut depth is depression in the surface formed by wheels and is commonly used to describe the condition of the road surface. In this work we modell rutting, i.e. annual change in rut depth, and aim to contribute to both predicting future rutting and find location of excelling rutting that might be due to drainage challenges. We set up and use Bayesian latent Gaussian models with relevant explanatory variables such as trafic intensity and road cover type . This includes models with non-stationary spatial dependency where the non-stationarity is driven by traffic intensity. The proposed models for rutting seems to work and provide insight into the physical nature of the spatial dependencies, with results indicating nonstationarity for rutting, with increasing standard deviation and decreasing spatial dependency as the traffic intensity increases.

Doubly online changepoint detection for monitoring health status during sport activities

Mattia Stival Thursday, June 30

Department of Statistical Sciences, University of Padova

We provide an online framework to analyze data recorded from smart watches during running activities. In particular, we focus on identifying variations in the behavior of one or more measurements caused by physical condition changes such as physical discomfort, periods of prolonged de-training, or even malfunction of measuring devices. Our framework considers data as a sequence of running activities represented by multivariate time series of physical and biometric data. We combine classical changepoint detection models with an unknown number of components with Gaussian state space models to detect distributional changes between a sequence of activities. The model considers multiple sources of dependence due to the sequential nature of subsequent activities, the autocorrelation structure within each activity, and contemporaneous dependence between different variables. We provide an online Expectation-Maximization (EM) algorithm involving a sequential Monte Carlo approximation of changepoint predicted probabilities. As a byproduct of our model assumptions, our proposed approach processes sequences of multivariate time series in a doubly online framework. While classical changepoint models detect changes be- tween subsequent activities, the state space framework coupled with the online EM algorithm provides the additional benefit of estimating real-time probabilities that a current activity is a changepoint.

A hierarchical Bayesian non-asymptotic extreme value model for spatial data

Federica Stolf Thursday, June 30

University of Padova

Spatial maps of extreme precipitation are crucial in flood protection. With the aim of producing maps of precipitation return levels, we propose a novel approach to model a collection of spatially distributed time series of extreme values where the asymptotic assumption, typical of the traditional extreme value theory, is relaxed.

We introduce a Bayesian hierarchical model that accounts for the possible underlying variability in the distribution of event magnitudes and occurrences, which are described through latent temporal and spatial processes. Spatial dependence is characterized by geographical covariates and effects not fully described by the covariates are captured by spatial structure in the hierarchies.

The performance of the approach is illustrated through simulation studies and an application to daily rainfall extremes across several sites in North Carolina (USA). The results show that we significantly reduce the estimation uncertainty with respect to state of the art methods.

A Bayesian development of probabilistic methods for Agent-based models of migration

Peter Strong Thursday, June 30

University of Warwick

Agent-Based Models (ABMs) are often used to model migration and are increasingly used to simulate individual migrant decision-making and unfolding events through a sequence of heuristic if-then rules. However, ABMs lack the methods to embed more principled strategies of performing inference to estimate and validate the models, both of which are significant for real-world case studies. Here, we give a worked example of how to build a probabilistic model in the subjective Bayesian paradigm that represents an ABM of migration based on the intrinsic Markov nature of the process and demonstrate its suitability to capture the uncertainties associated with such egocentric models. We embellish this model into a Chain Event Graph (CEG), a class of probabilistic graphical models able to provide a compact representation of complex independence statements.

Resampled Stochastic Approximation in Neural Networks

Stephen Styles Thursday, June 30

University of Alberta

Neural networks typically involve some sort of stochastic gradient descent (SGD) method to achieve their estimates. The current standard for these SGD methods are adaptive gradient techniques, like ADAM, which can have speed of convergence issues. Convergence for these algorithms requires a large quantity of observations to achieve a high level of accuracy and, with certain classes of functions, could take multiple epochs of data points to catch on. We explore the possibility of boosting the speed of convergence using a set of decoupled linear stochastic approximations for a wide neural network via a bootstrap technique. The convergence of decoupled systems is shown empirically using additional resampled observations to have remarkably quick convergence with a lower amount of data points. With this boost in speed, we are able to approximate classes of functions within a fraction of the observations that are needed with traditional neural network training methods. Some prior mathematical results are stated to reinforce our empirical study.

Bayesian estimation of nonlinear Hawkes processes

Deborah Sulem Thursday, June 30

University of Oxford

Temporal point processes (TPP) are widely applied to model the occurrences of events, e.g., natural disasters, online message exchanges, financial transactions or neuronal spike trains. The Hawkes model is a TPP model which allows the probability of occurrences of future events to depend on the past of the process and is particularly popular for modelling interactive phenomena such as disease expansion. In this work we consider the nonlinear multivariate Hawkes model, which can notably account for excitation and inhibition between interacting entities. We provide theoretical guarantees for applying nonparametric Bayesian estimation methods; in particular, we obtain concentration rates of the posterior distribution on the parameters, and convergence rates of Bayesian estimators. Another object of interest in event-data modelling is to infer the graph of interaction - or Granger causal graph. In this case, we provide consistency guarantees for Bayesian methods.

Fast Bayesian Inference for Item Response Theory models

Elias T Krainski Thursday, June 30

KAUST

One aim of an Item Response Theory (IRT) model is to estimate the students ability based on the item parameters. The joint estimation of the item parameters can be challenging for traditional algorithms, even when only a few hundred students are considered, and real data usually comprises of thousands of students. INLA is a well-known approximate inference framework for complex models and after some recent developments, IRT models can now be inferred in very low computing time using INLA. Additionally, we propose an extension of the two parameter IRT model to better account for the non-symmetric behavior, where we consider a penalized complexity prior to contract to the simpler model if the data necessitates it. An application of the proposed IRT model to a large data set from a high school national level exam in Brazil is presented. This is joint work with Janet Van Niekerk and Haavard Rue.

Dynamics of the competition between two languages

Achille Ecladore Tchahou Tchendjeu Thursday, June 30

University of Bamenda

This paper reports a new kind of mathematical model for language competition dynamics using compartmental epidemiological modeling approach. The model describes the competition between two languages, say the predominant one called language 1, and the less spoken language 2, both in the same community. We distinguish three groups of population building the concerned community: one majority speaking language 1, one minority speaking language 2 and a last minority speaking neither language 1 nor language 2. The study of the proposed model includes an analysis of the evolution of the number of speakers over time. This model predicts that language 2 can inevitably disappear if the threshold parameter R0 is less than one. We show that our model is well-posed mathematically and linguistically. We also show that the model has basically two linguistic equilibrium points. A monolingual equilibrium point that corresponds to the case where only one language is spoken: some individuals speak only the language 1 and some others do speak neither language 1, nor language 2. A bilingual equilibrium point which corresponds to the case where all the languages are spoken: some individuals speak both languages 1 and 2, some others speak only language 1, while a third group does speak only language 2, and the last group speaks none of the two languages. In this bilingual equilibrium point, both languages coexist. Depending on the threshold parameter, we demonstrate the stability of these equilibria. The model presents on one hand, a direct bifurcation phenomenon, in which we have a stable equilibrium point without bilingual speakers when the associated basic reproduction number is less than one. On the other hand, it presents a stable bilingual equilibrium point when the number of associated basic reproductions is greater than one. The analysis of the overall sensitivity of the model is performed and the impact of the system parameters on the basic reproduction number was performed using sensitivity analysis to determine the impact of each parameter of the system on bilingualism. The numerical simulations carried out are in agreement with the presented theory.

Bayesian nonparametric scalar-on-image regression via Potts-Gibbs random partition models

Mica Teo Thursday, June 30

University of Edinburgh

Scalar-on-image regression aims to investigate changes in a scalar response of interest based on high-dimensional imaging data. We propose a novel Bayesian nonparametric scalar-on-image regression model that utilises the spatial coordinates of the voxels to group voxels with similar effects on the response to have a common coefficient. We employ the Potts-Gibbs random partition model as the prior for the random partition in which the partition process is spatially dependent, thereby encouraging groups representing spatially contiguous regions. In addition, Bayesian shrinkage priors are utilised to identify the covariates and regions that are most relevant for the prediction. The proposed model is illustrated using the simulated data sets.

Author Clustering and Topic Estimation for Short Texts

Graham Tierney Thursday, June 30

Duke University

Analysis of short text, such as social media posts, is extremely difficult because of their inherent brevity. In addition to classifying topics of such posts, a common downstream task is grouping the authors of these documents for subsequent analyses. We propose a novel model that expands on the Latent Dirichlet Allocation by modeling strong dependence among the words in the same document, with user-level topic distributions. We also simultaneously cluster users, removing the need for post-hoc cluster estimation and improving topic estimation by shrinking noisy user-level topic distributions towards typical values. The model is estimated both with an exact collapsed Gibbs sampler and a fast variational approximation. We demonstrate that the accuracy of the variational approximation improves as the size of the corpus increases, the scenario when the approximation is most useful. Our method performs as well as — or better — than traditional approaches, and we demonstrate its usefulness on a dataset of tweets from United States Senators, recovering both meaningful topics and clusters that reflect partisan ideology. We also develop a measure of echo chambers among these politicians by characterizing insularity of topics discussed by groups of Senators and provide uncertainty quantification using posterior predictive simulations.

Principal stratification with measurement errors for post-infection outcome vaccine efficacy and effectiveness without monotonicity

Rob Trangucci Thursday, June 30

University of Michigan

Traditional measures of post-infection outcome vaccine efficacy using principal stratification rely on the assumption that infection and outcome can be measured without error and that an individual's vaccine efficacy for infection is nonnegative. In reality, diagnostic tests for infection are imperfect and outcomes like symptoms or severe disease can be misclassified. Nonnegativity of individual vaccine efficacy, also known as monotonicity, cannot always be assumed to hold in observational studies of vaccine effectiveness. We extend the principal stratification framework applied to vaccine efficacy in two ways: we allow for nondifferential measurement errors in both outcome and infection and we eliminate the monotonicity assumption. When monotonicity cannot be assumed but there are multiple treatment groups, error-free measurements, homogeneity of principal causal effects across treatment groups, and a discrete covariate we derive verifiable sufficient conditions for identifiability of the causal model. When monotonicity holds, but there are measurement errors in the outcome and infection, we derive conditions for the nonparametric identifiability of the principal strata proportions. With neither error-free measurements nor monotonicity, we combine the results to yield a causal estimand that is dependent on two partially identifiable parameters. We formulate a parametric model for post-infection outcome vaccine efficacy and apply the model to a simulation study to measure sensitivity to deviations from our identifiability conditions. Given our dependence on the assumptions of homogeneity, and conditional unconfoundedness we extend the parametric model to allow for nonhomogeneous principal causal effects between treatment groups, and unobserved confounders. Finally we apply our parametric model to an influenza vaccine clinical trial and an observational study of influenza vaccine effectiveness.

A Time-Dependent Poisson-Gamma Model for Recruitment Forecasting in Multi-center Clinical Trials

Armando Turchetta Thursday, June 30

McGill University

Estimating the recruitment time in multicenter clinical trials is a key component of the feasibility assessment. Yet, deterministic models mainly based on trial investigators' recruitment assumptions are still used. A Bayesian approach built on a doubly stochastic Poisson process, known as the Poisson-Gamma model, was introduced to address the lack of a strong and consistent statistical methodology in this field. This approach is based on the modeling of enrollments as a Poisson process where the recruitment rates are assumed to be constant over time and to follow a common Gamma prior distribution. However, the constant-rate assumption is a restrictive limitation that is rarely appropriate for applications in real clinical trials. In this presentation, we illustrate a flexible generalization of this methodology which allows the enrollment rates to vary over time by modeling them through B-splines, and we show the suitability of this approach for a wide range of recruitment behaviors.

Zig-Zag for Approximate Bayesian Computation

Giorgos Vasdekis Thursday, June 30

University of Warwick

Piecewise Deterministic Markov Processes (PDMPs) (see Fearnhead et.al. 2018) have recently caught the attention of the MCMC community for having a non-diffusive behaviour and being able to explore the state space more efficiently. This makes them good candidates to generate MCMC algorithms. One important problem in Bayesian computation the last ten years is inference for models where the likelihood is intractable. A popular method to deal with problems in this setting is the Approximate Bayesian Computation (ABC). In this poster we describe a PDMP algorithm, based on the Zig-Zag process (see Bierkens and Roberts 2019), that is designed to target ABC posteriors. This way we combine the areas of PDMPs and ABC. We show that the algorithm targets the distribution of interest and we provide numerical examples to show its effectiveness. This is joint work with Richard Everitt.

Sparse recovery of dynamical systems with inference

Sara Venkatraman Thursday, June 30

Cornell University

In many scientific disciplines, time-evolving phenomena are frequently modeled by nonlinear ordinary or partial differential equations. A common statistical approach to the recovery of such models from data relies on penalized regression, whereby equations of the form $dx/dt = f(x(t))$ are estimated by regressing time derivatives on a large set of candidate functions, such as polynomials. However, this technique is prone to the selection of incorrect terms and is lacking in rigorous uncertainty quantification for the estimated equations. We propose leveraging recent advances in Bayesian and frequentist penalized regression to estimate differential equations as sparse combinations of terms that are statistically significant or have high posterior probabilities. Using noisy data generated from several dynamical systems, we demonstrate that this method is able to correctly identify the polynomial terms comprising such systems and quantify their variability.

A Bayesian approach to assess efficacy in ALS platform trials

Matteo Vestrucci Thursday, June 30

Berry Consultants

The HEALEY ALS Platform Trial tests multiple novel treatment regimens in persons with ALS. Regulatory guidance recommends an integrated analysis of survival and function to assess efficacy of novel therapies. The traditional approach is a joint rank analysis (CAFS) that lacks 1- a clinically meaningful summary of treatment benefit and 2- the flexibility to accommodate the challenges arising from a platform trial, such as modeling potential differences in shared placebo participants over time and across regimens. The proposed primary analysis model for the trial is a novel Bayesian integrated analysis of survival and function that overcomes the two deficiencies. We compare the proposed analysis to CAFS by means of simulations to show that it also has better operating characteristics. The simulations correctly account for the association between survival and function by utilizing a novel hybrid approach built on simulating synthetic patients based on bootstrapped blueprints, generated from the PROACT database.

Modeling vulnerability through a hierarchical spatial Bayesian model for mixed dichotomous and continuous variables

Gabrielle Virgili-Gervais Thursday, June 30

McGill University

Rapid urbanization has put pressure on services and infrastructures provided by cities. This has led to social inequalities, leaving parts of the population vulnerable. This poster proposes a model enabling policy makers to identify vulnerable areas and decisive factors of local vulnerability, to provide meaningful aid to those in need. We present a model capturing the multifaceted concept of vulnerability through a Bayesian hierarchical model. This model extends Quinn's (2004) model on mixed dichotomous and continuous variables, using 10% of the 2010 Ghana census in the Greater Accra Metropolitan area. Three models were fit to provide a vulnerability index: one with independent parameters, one inducing a spatial structure on the index and one with hierarchical intercepts, wherein different areas are allowed to have different weights. The advantage of this approach is that different variables can contribute differently to the index across space. Unlike most vulnerability assessment models proposed by the literature, the hierarchical nature of our model enabled us to use the census observations at the household level without needing to aggregate any information. Thus, it better accommodates the variability of the vulnerability between census tracts. In all three models, some of the most discriminating variables were related to access to clean water and waste disposal, both of which could be addressed by local authorities. Amongst the proposed models, the best fit was achieved using a hierarchical prior on the intercepts. This is joint work with Alexandra M. Schmidt, Jill Baumgartner and Brian Robinson from McGill University.

A Bayesian approach to multiview learning

Lasse Vuursteen Thursday, June 30

TU Delft

In multi-view learning, one considers prediction based on different sources of information called views. Each view constitutes the observations of the same label Y_i but with a different set of covariates. When one has multiple views concerning the same prediction problem, how does one find out which view is most predictive? And how should predictions based on separate views be combined? For example, say that for each patient in a biomedical study, one can obtain separately different types of medical scans. Finding out which type of scan(s) are most valuable and how the predictions of these scans should be combined is valuable when obtaining these scans is costly. We study the performance of Bayesian methods in a variety of simulated multi-view settings such as linear- and logistic regression. We compare the performance to established frequentist methods. In particular, we investigate the potential of Bayesian methods in quantifying the degree to which a view has predictive value.

Fully Bayesian Structured Sparse Group Lasso

Dan Wang Thursday, June 30

TU Delft

In statistical modeling, group-selection problems arise naturally, especially in applications with high dimensional covariates. Several penalization and Bayesian methods have been proposed in the literature. The Sparse group lasso estimator can realize variable selection both on a group and within-group level. However, it does not provide meaningful variance estimates for the regression coefficients. In order to overcome this problem, we consider a novel fully Bayesian method: The Bayesian structured sparse group lasso. We have designed a fully Gibbs sampling procedure to sample from the posterior distribution of the parameters based on the exchange algorithm and the double Metropolis-Hastings algorithm. For the variable selection, we also consider an empirical posterior credible interval method. Simulations and real data analysis using our proposed method show promising performance in terms of parameter estimation, prediction, and variable selection, in comparison with a variety of existing methods.

Coverage of Credible Intervals in Bayesian Multivariate Isotonic Regression

Kang Wang Thursday, June 30

North Carolina State University

We consider the nonparametric multivariate isotonic regression problem, where the regression function is assumed to be nondecreasing with respect to each predictor. Our goal is to construct a Bayesian credible interval for the function value at a given interior point with assured limiting frequentist coverage. A natural prior on the regression function is given by a random step function with a suitable prior on increasing step-heights, but the resulting posterior distribution is hard to analyze theoretically due to the complicated order restriction on the coefficients. We instead put a prior on unrestricted step-functions, but make inference using the induced posterior measure by an "immersion map" from the space of unrestricted function to that of multivariate monotone functions. This allows maintaining the natural conjugacy for posterior sampling. A natural immersion map to use is a projection with respect to a distance function, but in the present context, a block isotonization map is found to be more useful. The approach of using the induced "immersion posterior" measure instead of the original posterior to make inference provides a useful extension of the Bayesian paradigm, particularly helpful when the model space is restricted by some complex relations. We establish a key weak convergence result for the posterior distribution of the function at a point in terms of some functional of a multi-indexed Gaussian process that leads to an expression for the limiting coverage of the Bayesian credible interval. Analogous to a recent result for univariate monotone functions, we find that the limiting coverage is slightly higher than the credibility, the opposite of a phenomenon observed in smoothing problems. Interestingly, the relation between the credibility and the limiting coverage does not involve any unknown parameter. Hence by a recalibration procedure, we can get a predetermined asymptotic coverage by choosing a suitable credibility level smaller than the targeted coverage, and thus also shorten the credible intervals.

Bayesian Genetic Mark-Recapture Methods for Estimating Seasonal River Run Size of Stock Populations

Yiran Wang Thursday, June 30

University of Waterloo

Genetic mark-recapture (GMR) is a statistical technique used in estimating population size in ecology. By combining genetic data on the relative abundance of species from a sample with population counts obtained for a subset of the species, GMR allows the estimation of the total population size and the contributions of each species. This is typically done with a type of Lincoln-Petersen estimator which provides an asymptotically unbiased estimate for the total population size. However, the accompanying variance estimator does not account for the uncertainty in the sampling process of the genetics data. As a result, this approach can suffer from a significantly underestimated variance, especially when the relative proportions in the genetic sample differ from those in the population. In this work, we propose a novel Bayesian GMR framework to address this issue. The Bayesian framework can explicitly incorporate the sampling error in the genetic sample and readily lends itself to combining additional sources of data into a single model, such as capture-recapture data or telemetry data, which are also frequently used to estimate population size. The effectiveness of the new method is investigated via simulation studies and used to estimate the abundance of Sockeye Salmon in the Taku River.

Data Augmentation for Bayesian Deep Learning

Yuexi Wang Thursday, June 30

University of Chicago

Deep Learning (DL) methods have emerged as one of the most powerful tools for functional approximation and prediction. While the representation properties of DL have been well studied, uncertainty quantification remains challenging and largely unexplored. Data augmentation techniques are a natural approach to provide uncertainty quantification and to integrate stochastic MCMC search with stochastic gradient descent (SGD) methods. The purpose of our paper is to show that training DL architectures with data augmentation leads to efficiency gains. To demonstrate our methodology, we develop data augmentation algorithms for a variety of commonly used activation functions: logit, ReLU and SVM. Our methodology is compared with traditional stochastic gradient descent with back-propagation. Our optimization procedure leads to a version of iteratively re-weighted least squares and can be implemented at scale with accelerated linear algebra methods providing substantial performance improvement. We illustrate our methodology on a number of standard datasets. Finally, we conclude with directions for future research.

Tree-based models for high-dimensional compositional data in microbiome studies

Zhuoqun Wang Thursday, June 30

Department of Statistical Science, Duke University

The human gut microbiome is associated with various diseases and health outcomes. A key characteristic of microbiome compositional data is its large and complex cross-sample heterogeneity. Appropriately accounting for these functional "variance components" is critical for several common inference tasks, including identifying latent structures, carrying out hypothesis testing on cross-group differences, and modeling dynamics, but is complicated by the key features of microbiome compositional data including high-dimensionality, sparsity, and compositionality. These characteristics incur the need for structural constraints on covariance modeling while maintaining the analytical and computational tractability of the resulting models and methods. We present several recently proposed methods that aim to utilize a tree structure—namely the phylogeny of the microbial species—to incorporate flexible covariance components while maintaining computational scalability. In particular, we present probabilistic models for microbiome compositional data based on the Dirichlet-tree (DT) distribution and the logistic-tree normal (LTN) distribution and demonstrate their wide applicability in a range of applications including cross-sample comparison, mixed-effects modeling, covariance estimation, and clustering analysis. This is based on joint work with Jialiang Mao and Li Ma.

Behavioural Change Individual-Level Models for Infectious Disease Transmission

Madeline Ward Thursday, June 30

University of Calgary

Individual-level models are a flexible class of statistical model which can incorporate information on individual risk factors, including spatial location, to account for the high degree of heterogeneity that is characteristic of population mixing, and, thus, disease transmission. However, these models have typically assumed stable population behaviour over time. Yet, as we have observed throughout the COVID-19 pandemic, behaviour often changes based on the current perceived risk of contracting the disease. In turn, this behaviour change can have a large impact on the transmission dynamics of the disease. This poster presents a new class of behavioural change individual-level models where prevalence affects susceptibility and/or population mixing and illustrate their use through simulated and real data on foot and mouth disease. We will detail the MCMC model-fitting methodology used, including the use of spike and slab priors to identify whether prevalence-dependency is truly present in an epidemic.

Using Dirichlet Processes Mixture Models and Machine Learning to Estimate Crash Risk on Roadways

Richard Warr Thursday, June 30

Brigham Young University

Historically, specifying models for point pattern data has had to balance flexibility with interpretability. On the one hand, mixture model specifications for Poisson process intensity surfaces can flexibly capture the nonlinear nature of the intensity surface but don't yield interpretable regression parameters. On the other hand, log-Gaussian Cox processes can give interpretable regression coefficients for the intensity surface but suffer from computational issues. In this project we provide a partial solution to this balancing act by using a Dirichlet process mixture model (DPMM) to flexibly model an intensity surface for a Poisson process. We then project the resulting DPMM, using all posterior draws, onto a set of basis functions using penalized regression to obtain an estimate (with the associated Bayesian uncertainty) of a corresponding log-Gaussian Cox process fit. We demonstrate this process by estimating the intensity surface and the associated effects of roadway characteristics on the frequency of crashes on I-15 in Utah over time. This is a joint work with Matthew Heaton, Philip White, Caleb Daley, and Benjamin Dahl.

Bayesian Modeling of Effective and Functional Brain Connectivity using Hierarchical Vector Autoregressions

Bertil Wagmann Thursday, June 30

Department of Computer and Information Science, Linköping University, SWEDEN

Analysis of brain connectivity is important for understanding how information is processed by the brain. We propose a novel Bayesian vector autoregression (VAR) hierarchical model for analyzing brain connectivity in a resting-state fMRI data set with autism spectrum disorder (ASD) patients and healthy controls. Our approach models functional and effective connectivity simultaneously, which is new in the VAR literature for brain connectivity, and allows for both group- and single-subject inference as well as group comparisons. We combine analytical marginalization with Hamiltonian Monte Carlo (HMC) to obtain highly efficient posterior sampling. The results from more simplified covariance settings are, in general, overly optimistic about functional connectivity between regions compared to our results. In addition, our modeling of heterogeneous subject-specific covariance matrices is shown to give smaller differences in effective connectivity (EC) compared to models with a common covariance matrix to all subjects.

Clustering Neural Populations with an Extended Poisson Dynamic Factor Model

Ganchao Wei Thursday, June 30

University of Connecticut, Department of Statistics

Modern neural recording techniques allow neuroscientists to observe the spiking activity of many neurons simultaneously. Although previous work has illustrated how activity within and between known populations of neurons can be summarized by low-dimensional latent vectors, in many cases what determines a discrete population may be unclear. Neurons differ in their anatomical location, but, also, in their cell types and response properties. When the neural activities are globally nonlinear, using single population analysis is inappropriate. However, defining the populations is usually difficult and wrong cluster assignments will lead to bias in latent structure inferences. To tackle this challenge, here we develop a clustering method based on a mixture of extended Poisson dynamic factor model including individual baselines, with the number of cluster is treated as a parameter in mixture of finite mixtures (MFM) model. To sample efficiently from the posteriors, we approximate the full conditional distribution of latent state by Gaussian and marginalize the loading out when clustering, by making use of the Poisson-Gamma conjugacy. We further apply our method to multi-region neuropixels recordings for illustration. The propose method provides a tool to cluster neurons based on functionality.

Bayesian Data Selection

Eli Weinstein Thursday, June 30

Columbia University

Insights into complex, high-dimensional data can be obtained by discovering features of the data that match or do not match a model of interest. To formalize this task, we introduce the "data selection" problem: finding a lower-dimensional statistic - such as a subset of variables - that is well fit by a given parametric model of interest. A fully Bayesian approach to data selection would be to parametrically model the value of the statistic, nonparametrically model the remaining “background" components of the data, and perform standard Bayesian model selection for the choice of statistic. However, fitting a nonparametric model to high-dimensional data tends to be highly inefficient, statistically and computationally. We propose a novel score for performing both data selection and model selection, the “Stein volume criterion", that takes the form of a generalized marginal likelihood with a kernelized Stein discrepancy in place of the Kullback-Leibler divergence. The Stein volume criterion does not require one to fit or even specify a nonparametric background model, making it straightforward to compute - in many cases it is as simple as fitting the parametric model of interest with an alternative objective function. We prove that the Stein volume criterion is consistent for both data selection and model selection, and we establish consistency and asymptotic normality (Bernstein-von Mises) of the corresponding generalized posterior on parameters. We validate our method in simulation and apply it to the analysis of single-cell RNA sequencing datasets using probabilistic principal components analysis and a spin glass model of gene regulation.

Improving Hidden Population Size Estimation

Justin Weltz Thursday, June 30

Duke University

Many subpopulations defined by illegal or stigmatized behavior are difficult to sample using conventional survey methodology. Response Driven Sampling (RDS) is a participant referral process frequently employed in this context to collect information without privacy-violating survey mechanisms. Previous methods have attempted to estimate missing edges in the partially observed network inherent to the RDS sample. Unfortunately, clear topological biases in these subgraph estimates cause problems for downstream estimation. Treating the RDS subgraph as a nuisance parameter, we propose a double pronged approach for correcting the bias of population size estimates. The first extends simulation-based iterative bias correction methods to the infinite dimensional graph domain. The second leverages prior information on the combinatorial space of potential graphs in a Bayesian framework to obtain more realistic network features. We demonstrate significant bias reduction in estimated population size using these methods in a variety of simulated paradigms.

Bayesian unanchored additive models for component network meta-analysis

Augustine Wigle Thursday, June 30

University of Waterloo

Component Network Meta-Analysis (CNMA) models are an extension of standard Network Meta-Analysis (NMA) models which account for the use of multicomponent treatments in the network. This poster contributes innovatively to several statistical aspects of CNMA. First, by introducing a unified notation, we establish that currently available methods differ in the way they assume additivity, an important distinction that has been overlooked so far in the literature. In particular, one model uses a more restrictive form of additivity than the other which we term an anchored and unanchored model, respectively. We show that an anchored model can provide a poor fit to the data if it is misspecified. Second, given that Bayesian models are often preferred by practitioners and that there are not currently Bayesian CNMA models that benefit from the more flexible additivity assumption, we develop two such unanchored Bayesian CNMA models. An extensive simulation study examining bias, coverage probabilities and treatment rankings confirms the favourable performance of the novel models. This is the first simulation study to compare the statistical properties of CNMA models in the literature. Finally, the use of our novel models is demonstrated on a real dataset, and the results of all existing CNMA models on the dataset are compared.

Arianna: A Domain-Specific Language for MCMC Algorithms

Daniel Winkler Thursday, June 30

Vienna University of Economics and Business

The development of MCMC algorithms involves an implementation in a mathematical language, in addition to one in a programming language. Often further versions are written in faster, lower-level languages (e.g., C++). This development cycle comes with obvious drawbacks (e.g., multiple manual implementations).

We contribute “Arianna" (honoring Rosenbluth), a system for MCMC algorithms based on domain-specific languages (DSLs; special-purpose languages like SQL). Our DSL allows the implementation of algorithms using mathematical notation, which can be translated to different programming languages.

Our DSL constitutes a concise, easily comprehensible, and extensible yet powerful system to streamline MCMC development. We highlight different non-trivial MCMC algorithms including global-local shrinkage priors for factor models.

While researchers retain complete control over the algorithm (no black box), highly optimized backends (e.g., GPU) can be provided by domain specialists. In our presentation we will highlight several backends (Rust for speed, Javascript for in-browser execution, among others).

Supervised Bayesian Nonparametric Clustering Techniques for Survey Data

Stephanie Wu Thursday, June 30

Harvard University

Dietary intake is a major modifiable risk factor for cardiovascular disease. The effect of dietary patterns on cardiovascular disease can be characterized using supervised Bayesian nonparametric clustering methods. However, when dietary data are sourced from surveys where unequal probabilities of selection are inherent in the design, the complex survey design must be accounted for to avoid biased estimation. Working from an overfitted finite mixture model framework, we explore two approaches that use sampling weights to adjust for survey design and apply them to a supervised cluster setting. The first replaces the likelihood with a weighted pseudo-likelihood in the posterior update. The second uses a bootstrap approach to generate a pseudo-population that is integrated into the MCMC algorithm. Using dietary consumption data from nationally representative surveys, we apply these two methods and discuss their performance via simulation studies in order to better understand the impact of diet on cardiovascular disease risk.

Inferring the sources of HIV infection in Africa from deep-sequence data with semi-parametric Bayesian Poisson flow models

Xiaoyue Xi Thursday, June 30

University of Cambridge and Imperial College London

Pathogen deep-sequencing is an increasingly routinely used technology in infectious disease surveillance. We present a semi-parametric Bayesian Poisson model to exploit these emerging data for inferring infectious disease transmission flows and the sources of infection at the population level. The framework is computationally scalable in high-dimensional flow spaces thanks to Hilbert Space Gaussian process approximations, allows for sampling bias adjustments, and estimation of gender- and age-specific transmission flows at finer resolution than previously possible. We apply the approach to densely sampled, population-based HIV deep-sequence data from Rakai, Uganda, and find substantive evidence that adolescent and young women are predominantly infected through age-disparate relationships.

Likelihood-based Inference for Stochastic Epidemic Models over Dynamic Networks

Jason Xu Thursday, June 30

Duke University

Stochastic epidemic models such as the Susceptible-Infectious-Removed (SIR) model are widely used to model the spread of disease at the population level, but fitting these models to data present significant challenges. In particular, the marginal likelihood is typically considered intractable in the presence of missing data, as practitioners resort to simulation methods or approximations. We discuss some recent contributions that enable direct inference using the likelihood of observed data, focusing on a perspective that makes use of latent variables to explore configurations of the missing data within a Bayesian framework. Motivated both by count data from large outbreaks and high-resolution contact data from mobile health studies, we show how a data-augmented MCMC approach successfully learns the interpretable epidemic parameters and scales to handle realistic data settings efficiently.

Bayes Factor with Conjugate Prior for Region-Based Rare Variant Analysis

Jingxiong Xu Thursday, June 30

Lunenfeld-Tanenbaum Research Institute

Next Generation Sequencing technology provides opportunities to discover rare variants (RVs) associated with complex human diseases. We previously introduced a region-based test using a Bayes Factor (BF) statistic (Xu et al., Biometrics, 2020) where the association between a set of RVs in the same region (e.g. a gene) and a disease was assessed. Here we extend this approach using the generalized linear model framework and its conjugate prior (Chen et al., Statistica Sinica, 2003), which can handle outcomes of different types (binary, continuous, count), informative functional annotation and unbalanced designs. We also implemented a variable selection of the RVs to improve the power of our approach using the birth-death algorithm. Simulation studies and application to UK Biobank cancer data are conducted to assess the finite sample properties of our method.

Objective Bayesian Model Selection for Generalized Linear Mixed Models

Shuangshuang Xu Thursday, June 30

Virginia Tech

We propose an objective Bayesian model selection approach for generalized linear mixed models. To deal with the issue of integration of random effects, we approximate the likelihood function using a Gaussian pseudo-likelihood. In addition, we assume approximate reference priors for the parameters of the model. In addition to the approximate reference prior, we also consider a non-local prior for the variance component of random effects. To deal with the impropriety of the prior, we develop a fractional Bayes factor approach with a minimum training fraction. We then perform model selection based on the resulting posterior probabilities of the several competing models. Simulation studies with Poisson generalized linear mixed models with spatial random effects and overdispersion random effects show that our approach performs favorably when compared to widely used competing Bayesian methods. We illustrate the usefulness and flexibility of our approach with three case studies on a Poisson longitudinal model, a Poisson spatial model, and a logistic mixed model.

Statistically robust discovery of mutational signatures using the power posterior

Catherine Xue Thursday, June 30

Harvard University

Mutational signatures are distinctive patterns of mutations resulting from carcinogenic molecular processes, such as UV radiation, molecular effects of chemical agents, and defective DNA repair mechanisms. Non-negative matrix factorization (NMF) models have been used to discover mutational signatures and deconvolve their respective contributions in individual tumors from sequencing data. However, any assumed model is only a rough approximation to reality, and as a consequence, the results are sometimes misleading and irreproducible. We propose an alternative approach to mutational signature inference that, by leveraging the power posterior, is robust to using an approximate model and, by using a novel sparsity-inducing prior, automatically infers the number of signatures. We demonstrate the robustness and accuracy of our approach on simulated data and real data from the ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium.

Multivariate Dynamic Modeling for Bayesian Forecasting of Business Revenue

Anna Yanchenko Thursday, June 30

Duke University

Forecasting enterprise-wide revenue is critical to many companies and presents several challenges and opportunities for significant business impact. This case study is based on model developments to address these challenges for forecasting in a large-scale retail company. Focused on multivariate revenue forecasting across collections of supermarkets and product Categories, hierarchical dynamic models are natural: these are able to couple revenue streams in an integrated forecasting model, while allowing conditional decoupling to enable relevant and sensitive analysis together with scalable computation. Structured models exploit multi-scale modeling to cascade information on price and promotion activities as predictors relevant across Categories and groups of stores. With a context-relevant focus on forecasting revenue 12 weeks ahead, this work highlights product Categories that benefit from multi-scale information, defines insights into when, how and why multivariate models improve forecast accuracy, and shows how cross-Category dependencies can relate to promotion decisions in one Category impacting others.

Stereographic Markov Chain Monte Carlo

Jun Yang Thursday, June 30

University of Oxford

High dimensional distributions, especially those with heavy tails, are notoriously difficult for off the shelf MCMC samplers: the combination of unbounded state spaces, diminishing gradient information, and local moves, results in empirically observed “stickiness" and poor theoretical mixing properties – lack of geometric ergodicity. In this paper, we introduce a new class of MCMC samplers that map the original high dimensional problem in Euclidean space onto a sphere and remedy these notorious mixing problems. In particular, we develop random-walk Metropolis type algorithms as well as versions of Bouncy Particle Sampler that are uniformly ergodic for a large class of light and heavy tailed distributions and also empirically exhibit rapid convergence in high dimensions.

Probabilistic Learning of Treatment Trees in Cancer

Tsung-Hung Yao Thursday, June 30

University of Michigan

Accurate identification of synergistic treatment combinations and their underlying biological mechanisms is critical across many disease domains, especially cancer. In translational oncology research, preclinical systems such as patient-derived xenografts (PDX) have emerged as a unique study design evaluating multiple treatments administered to samples from the same human tumor implanted into genetically identical mice. In this paper, we propose a novel Bayesian probabilistic tree-based framework for PDX data to investigate the hierarchical relationships between treatments by inferring treatment cluster trees, referred to as treatment trees (Rx-tree). The framework motivates a new metric of mechanistic similarity between two or more treatments accounting for inherent uncertainty in tree estimation; treatments with a high estimated similarity have potentially high mechanistic synergy. Building upon Dirichlet Diffusion Trees, we derive a closed-form marginal likelihood encoding the tree structure, which facilitates computationally efficient posterior inference via a new two-stage algorithm. Simulation studies demonstrate superior performance of the proposed method in recovering the tree structure and treatment similarities. Our analyses of a recently collated PDX dataset produce treatment similarity estimates that show a high degree of concordance with known biological mechanisms across treatments in five different cancers. More importantly, we uncover new and potentially effective combination therapies that confer synergistic regulation of specific downstream biological pathways for future clinical investigations. Our accompanying code, data, and shiny application for visualization of results are available at: https://github.com/bayesrx/RxTree

Operator-induced structural variable selection for identifying materials genes

Shengbin Ye Thursday, June 30

Rice University

We propose a new method for variable selection with operator-induced structure (OIS), in which the predictors are engineered from a limited number of primary variables and a set of elementary algebraic operators through compositions. Standard practice directly analyzes the high-dimensional candidate predictor space in a linear model; statistical analyses are then substantially hampered by the daunting challenge posed by millions of correlated predictors. The proposed method iterates nonparametric variable selection to achieve effective dimension reduction in linear models by utilizing the geometry embedded in OIS, leading to reduced computational costs and improved accuracy. An OIS screening property for variable selection methods in the presence of feature engineering is introduced. Finite sample assessment indicates that the employed Bayesian Additive Regression Trees (BART)-based variable selection method enjoys this property. We demonstrate the superior performance of the proposed method in simulation studies and a real data application to single-atom catalyst analysis.

Semiparametric posterior corrections

Andrew Yu Thursday, June 30

University of Oxford

In semiparametric Bayesian inference, the key condition for the marginal posterior of the target parameter to satisfy a Bernstein-von Mises theorem is that the prior predictive is approximately unchanged after shifting the likelihood in the "least favorable direction". This may exclude using certain types of priors e.g. adaptive, or force the user to tailor the analysis towards a specific parameter at the expense of other estimands of potential interest. We introduce an approach to bypass this condition using a Bayesian bootstrap correction. The corrected posterior satisfies the Bernstein-von Mises theorem under much less restrictive conditions, analogous to those for (frequentist) one-step estimation or targeted maximum likelihood estimation. Once samples have been drawn from the initial posterior, the procedure is practically instantaneous and can be applied to multiple parameters.

Group Structure Estimation for Panel Data - A General Approach

Lu Yu Thursday, June 30

University of Toronto

Consider a panel data setting where repeated observations on individuals are available. Often it is reasonable to assume that there exist groups of individuals that share similar effects of observed characteristics, but the grouping is typically unknown in advance. We propose a novel approach to estimate such unobserved groupings for general panel data models. Our method explicitly accounts for the uncertainty in individual parameter estimates and remains computationally feasible with a large number of individuals and/or repeated measurements on each individual. The developed ideas can be applied even when individual-level data are not available and only parameter estimates together with some quantification of uncertainty are given to the researcher.

Using AI and High-resolution fMRI Data to Learn the Backbone Functional Connectivity of the Human Brain

Zeyu Yuwen Thursday, June 30

University of Florida, Department of Statistics

Recent advances in brain imaging have produced large volume of neuroscience data. These data are valuable as they provide scans of the brain in a high resolution on the brain activities during important cognitive tasks. To understand how the brain works, a primary target of interest to extract is the functional connectivity, which changes dynamically over time. While some methods mainly view any non-zero correlation between two regions of brain as a functional connection and produce many false-positive findings. More importantly, in the face of high spatial resolution, subject-to-subject variation, and temporal change, these approaches aren't capable to produce useful interpretation. In this study, we will exploit the recent development of network models and AI deep learning to learn the backbone functional connectivity network. Compared to the crude correlation network, a fundamental difference is we focus on finding a backbone network that connects the important parts of the brain, and quantifying its dynamic change over time and variation among subjects. This project has a big impact on 1) biostatistical methodology to fill the gap in handling massive and temporally changing bioimaging data, and 2) neuroscience application to find the fundamental difference of brain connectivity between the task vs rest groups.

Bayesian nonparametric multiscale mixture models via Hilbert-curve partitioning

Daniele Zago Thursday, June 30

Università degli Studi di Padova

Bayesian nonparametric multivariate density estimation typically relies on mixture specifications, exceptions made for Pólya tree constructions. Herein, we develop a multivariate mixture model exploiting the multiscale stick-breaking prior recently proposed by Stefanucci and Canale (2021). The building block of the proposed approach is a base measure defined exploiting the Hilbert space-filling curve which allows to adapt a simple partitioning of a univariate parameter space to the multivariate case with minor adjustments. Alongside the theoretical discussion, we illustrate the model's performance by analyzing both synthetic and real datasets. The results suggest that the proposed multiscale model achieves competitive performance with respect to state-of-the-art Bayesian nonparametric methods both in scenarios presenting single- and multi-scale features.

Modelling temporally misaligned data across space: the case of total pollen concentration in Toronto

Sara Zapata-Marin Thursday, June 30

McGill University

Some spatio-temporal processes face the problem of temporal misalignment in the data when measurements are taken at different temporal scales. Rather than aggregating the data to the coarser scale, we can account for this temporal misalignment and take advantage of the finer temporal scale measurements. We propose a spatio-temporal model to account for temporal misalignment when one of the scales is the sum or average of the other. The inference is performed under the Bayesian framework and uncertainty about the unknown quantities of interest is naturally accounted for. We have available measurements of pollen concentration made between March and October of 2018 over 18 sites in Toronto, Canada. However, for some locations, observations were recorded daily, whereas, for others, observations were collected weekly. We show how the temporal aggregation of the pollen concentration measurements has an impact on the effect of the different covariates.

Multiplex Network Hawkes Model for the Assessment of Systemic Risk Propagation

Mante Zelvyte Thursday, June 30

UCL

Understanding the dynamics of connectedness within financial systems and the channels of risk propagation is of key importance for systemic risk management and to prevent the need of government intervention in future. Research in this area has been particularly active since the financial crisis of 2008-2009. Systemic risk is characterised by default contagion making Hawkes process a natural fit due to its ability to capture interactions between events. Network Hawkes model introduced by Linderman et al. (2014) offers an intuitive and interpretable way to infer the latent structure underlying systemic risk propagation. We extend the model by allowing for multiplex network structure with external covariates driving the network link formation. Our proposed model enables to explore the importance of different risk propagation channels by utilising different sources and types of data, incl. categorical company data, balance sheet information and market prices.

Automatic Reversible Jump MCMC for Bayesian Variable Selection in Poisson Regression Models

Gregor Zens Thursday, June 30

Bocconi University

Variable selection in Poisson regression models is a standard task for applied researchers in various fields. While frequentist penalized likelihood methods are well established, Bayesian frameworks have received considerably less attention. Existing Bayesian approaches are moreover often characterized by a high computational demand or extensive tuning requirements. We develop a novel, exact and computationally feasible hierarchical framework for variable selection in Poisson regression models. For estimation, automatic and efficient reversible jump Markov chain Monte Carlo techniques are utilized. A large-scale simulation study demonstrates the strengths of the framework relative to a number of competitor models and real data applications further illustrate the approach.

Comparing dependent undirected Gaussian networks

Hongmei Zhang Thursday, June 30

University of Memphis

A Bayesian approach is proposed that unifies Gaussian network constructions and comparisons between two longitudinal networks (identical or differential) for data collected at two-time points. Utilizing the concept of modeling repeated measures, we construct a joint likelihood of networks. The conditional posterior probability mass function for network differentiation determination is derived and its asymptotic proposition is theoretically assessed. An alternative approach built upon latent rather than manifest data is proposed to significantly reduce computing burden. Simulations are used to demonstrate and compare the two methods and compare them with existing approaches. Based on epigenetic data collected at different ages, the proposed methods are demonstrated on their ability to detect dependent network differentiations. Our theoretical assessment, simulations, and real data applications support the effectiveness of the proposed methods, although the approach relying on latent data is less efficient.

Dynamic Multivariate Generalized Double Pareto priors for modeling correlated spatio-temporal sparsity

Wei Zhang Thursday, June 30

Universita della Svizzera italiana

Dynamic regression models are natural extensions of static linear regression models that assume time-varying parameters. When both the number of parameters and the length of the time series are large, it is desirable to impose shrinkage sparsity priors that account for dependencies between neighboring time points. Here, we propose a multivariate extension of recently proposed normal-gamma auto-regressive priors that allows us to dynamically model the correlation between multiple time series, still retaining a generalized double Pareto distribution as stationary marginal distribution. We describe the properties of the proposed dynamic shrinkage process priors and devise an efficient MCMC scheme for posterior inference. We illustrate the performances of the model via a simulation study and an application to neuroimaging data.

Bayesian Inference for Cox Proportional Hazard Models with Partial Likelihoods, Semi-Parametric Covariate Effects and Correlated Observations

Ziang Zhang Thursday, June 30

University of Toronto

We propose a flexible and scalable approximate Bayesian inference methodology for the Cox Proportional Hazards model with partial likelihood. The model we consider accommodates semi-parametric covariate effects and correlated survival times. The proposed method is based on nested approximations and adaptive quadrature, and the computational burden of working with the log-partial likelihood is mitigated through automatic differentiation and Laplace approximation. We provide two simulation studies to show the accuracy of the proposed approach, compared with the existing methods. We demonstrate the practical utility of our method and its computational advantages over MCMC methods through the analysis of Kidney infection times, which are paired, and the analysis of Leukemia survival times with a semi-parametric covariate effect and spatial variation.

The Normal-Beta Prime Shrinkage Prior for the Estimation of Large Covariance and Precision Matrices

Peng Zheng Thursday, June 30

The University of Manchester

We propose a new fully Bayesian approach for estimating sparse covariance and precision matrices under the normal-beta prime shrinkage prior. Compared to the existing literature, the proposed estimation procedure has several advantages: first, no hyperparameters need to be prespecified and all of them can be obtained via maximum marginal likelihood (MML) method. Second, the data-driven choice of hyperparameters from the MML enables our method to accommodate the estimation of matrices with different sparsity levels whereas the existing frequentist and Bayesian approaches such as the graphical lasso and the graphical horseshoe, which we demonstrate in the simulation studies, can only handle either a very sparse setting or a relatively dense setting. Third, we establish the posterior convergence rate of the induced posterior under some sparsity assumptions on the matrices. Finally, we extend our approach to handle heavy-tailed data with a multivariate-t distribution where the sparsity pattern in the scale matrix and its inverse only indicates that the corresponding two random variables are marginally uncorrelated and conditionally uncorrelated respectively. Both simulations and real data analysis illustrate our estimator outperforms other approaches in terms of adaptation to various sparsity levels.

Proximal MCMC for Bayesian Inference of Constrained and Regularized Estimation

Xinkai Zhou Thursday, June 30

University of California, Los Angeles

This paper advocates proximal Markov Chain Monte Carlo (ProxMCMC) as a generic Bayesian inference framework for constrained or regularized estimation. Originally developed in the Bayesian imaging literature, ProxMCMC deploys the Moreau-Yosida envelop for a smooth approximation of the total variation regularization term, fixes nuisance and regularization parameters as constants, and relies on the Langevin algorithm for the sampling of the posterior. We extend the ProxMCMC to the full Bayesian framework with modeling and data adaptive estimation of all parameters including the regularization parameter. More efficient sampling algorithms such as the Hamiltonian Monte Carlo are employed to scale proximal MCMC to high-dimensional problems. Analogous to the proximal algorithms in optimization, ProxMCMC offers a versatile and modularized procedure for the inference for constrained and non-smooth problems. The power of ProxMCMC is illustrated on various statistical estimation and machine learning tasks. The inference in these problems is traditionally considered difficult from both frequentist and Bayesian perspectives.

Minimax Quasi-Bayesian estimation in sparse canonical correlation analysis via a Rayleigh quotient function

Qiuyun Zhu Thursday, June 30

Boston University

Canonical correlation analysis (CCA) is a popular statistical technique for exploring the relationship between datasets. The estimation of sparse canonical correlation vectors has emerged in recent years as an important but challenging variation of the CCA problem, with widespread applications. Currently available rate-optimal estimators for sparse canonical correlation vectors are expensive to compute. We propose a quasi-Bayesian estimation procedure that achieves the minimax estimation rate, and yet is easy to compute by Markov Chain Monte Carlo (MCMC). The method uses a re-scaled Rayleigh quotient function as a quasi-log-likelihood, and adopts a Bayesian framework that combines this quasi-log-likelihood with a spike-and-slab prior that serves to regularize the inference and promote sparsity. We investigated the empirical behavior of the proposed method on both continuous and truncated data, and we noted that it outperforms several state-of-the-art methods. Also, the convergence rate of the posterior mean estimator achieves the minimax rate of the CCA problem. As an application, we use the methodology to maximally correlate clinical variables and proteomic data for a better understanding of covid-19 disease.

Parallel Krylov Approximations for Latent Gaussian Models

Abylay Zhumekenov Thursday, June 30

KAUST

Gaussian Markov Random Fields (GMRFs) are widely used in Bayesian inference, and are central to Latent Gaussian Modeling (LGM). Dependencies between elements of a GMRF, e.g. spatial points, are modeled using sparse precision matrices. A key point is to find sparse Cholesky factors using a reordering, which grants drastic improvements in the complexity.

Nevertheless, computing a derivative of the log-determinant of the precision matrix with respect to a hyperparameter remains a somewhat challenging problem. Extracting correlation information is not easier either, since it necessitates partially inverting the matrix. Moreover, matrix factorization becomes prohibitively slow for huge problems. The scalability issue comes from the fact that direct methods are hard to parallelize, especially the sparse versions.

Instead, we propose recursive and embarrassingly parallel Krylov approximations to the derivative of the log-determinant and the inverse elements by solving only linear systems. Given the rise of multicore and specialized chip architectures, we believe that the method will be more suitable for large problems, where direct methods fail.

Inferring taxonomic placement from DNA barcoding allowing discovery of new taxa

Alessandro Zito Thursday, June 30

Duke University

Predicting the taxonomic affiliation of DNA sequences collected from biological samples is a fundamental step in biodiversity assessment. This task is performed by leveraging on existing databases containing reference DNA sequences whose taxa are known. However, environmental sequences can be from organisms that are either unknown to science or for which there are no reference sequences available. Thus, taxonomic novelty of a sequence needs to be accounted for when doing classification. We propose Bayesian nonparametric taxonomic classifiers, BayesANT, which use species sampling model priors to allow new taxa to be discovered at each taxonomic rank. Using a simple product multinomial likelihood with conjugate Dirichlet priors at the lowest rank, a highly efficient algorithm is developed to provide a probabilistic prediction of the taxa placement of each sequence at each rank. BayesANT is shown to have excellent performance when many sequences in the test set belong to unobserved taxa.

Bayesian nonparametrics for post-treatment variables in causal inference

Dafne Zorzetto Thursday, June 30

University of Padua

In causal inference, principal stratification requires to adjust the treatment comparison for confounded post-treatment variables. Post-treatment variables are potentially affected by the treatment while also affecting the response. A common strategy consists in classifying subjects into latent classes defined by potential posttreatment variables under each treatment. Continuous post-treatment variables pose some complications as they potentially induce an infinite number of possible principal strata. Starting from Schwartz et al. (2011), we propose a Bayesian nonparametric model that describes the distribution of the potential post-treatment outcomes, conditionally on the latent class variables. The model exploits the properties of Bayesian nonparametric mixtures which induce data clustering by encouraging borrowing of information and leading latent structures. To improve this model interpretability, we make use of inner spike and slab nonparametric mixtures. This construction allows to obtain a more informative and parsimonious model which could identify a different number of clusters under different treatments.

A Zero-Inflated Conway-Maxwell Poisson Regression Model with Spatially-Varying Dispersion for Spatiotemporal Data of US Vaccine Refusal

Bokgyeong Kang Thursday, June 30

The Pennsylvania State University

Vaccination is widely recognized as one of the most effective tools for preventing disease. However, parental refusal and delay of childhood vaccination has increased in recent years in the United States. This phenomenon challenges maintenance of herd immunity and increases the risk of outbreaks of vaccine-preventable diseases. Our aim is to identify demographic or socioeconomic characteristics associated with vaccine refusal, which could help public health professionals and medical providers develop interventions targeted to concerned parents. We examine US county-level vaccine refusal data for patients under five years of age collected on a monthly basis during the period 2012–2015. These data exhibit challenging features: zero inflation, spatially-varying dispersion, spatial dependence, and a large sample size (over 3,000 counties). We propose a spatial zero-inflated Conway–Maxwell–Poisson regression model that addresses these challenges. We use the asymptotically exact exchange algorithm and the sampling technique proposed by Chanialidis et al. (2018) to do Bayesian inference for our model. Our Bayesian framework permits efficient sampling and provides asymptotically exact estimates.

Bounds on Wasserstein distances between continuous distributions using independent samples

Tamas Papp Thursday, June 30

Lancaster University

The plug-in estimator of the Wasserstein distance is known to be conservative, however its usefulness is severely limited when the distributions are similar as its bias does not decay to zero with the true Wasserstein distance. We propose a linear combination of plug-in estimators for the squared 2-Wasserstein distance with a reduced bias that decays to zero with the true distance. The new estimator is provably conservative provided one distribution is appropriately overdispersed with respect the other, and is unbiased when the distributions are equal. We apply it to approximately bound from above the 2-Wasserstein distance between the target and current distribution in Markov chain Monte Carlo, running multiple identically distributed chains which start, and remain, overdispersed with respect to the target. Our bound consistently outperforms the current state-of-the-art bound, which uses coupling, improving mixing time bounds by up to an order of magnitude.

A Dynamic Stochastic Block Model for Multi-Layer Networks

Ovielt Baltodano Lopez Thursday, June 30

Ca' Foscari University of Venice

We propose a flexible stochastic block model for multi-layer networks, where layer-specific hidden Markov–chain processes drive the changes in the edge clustering. The changes in block membership in a given layer may be influenced by the lagged membership on the rest of the layers. This allows to identify clustering overlap, clustering decoupling or more complex relationships between layers including settings of unidirectional or bidirectional Granger–block causality. We cope with the overparameterization issue of a saturated specification by assuming a Multi–Laplacian prior distribution. Data augmentation and Gibbs sampling are used to make the inference problem more tractable. Through simulations we show that the standard BVAR models are not able to detect the Granger–block causality under the great majority of scenarios and we present an application to exemplify the use DSBMM finding new evidence of unidirectional causality from the block structure of the FTA network on the non-observable trade barriers network structure for 159 countries in the period 1995–2017.

Globally-centered autocovariances in MCMC

Medha Agarwal Thursday, June 30

University of Washington

Autocovariances are a fundamental quantity of interest in Markov chain Monte Carlo (MCMC) simulations with autocorrelation function (ACF) plots being an integral visualization tool for performance assessment. Unfortunately, for slow-mixing Markov chains, the empirical autocovariance can highly underestimate the truth. For multiple-chain MCMC sampling, we propose a globally-centered estimator of the autocovariance function (G-ACvF) that exhibits significant theoretical and empirical improvements. We show that the bias of the G-ACvF estimator is smaller than the bias of the current state-of-the-art. The impact of this improved estimator is evident in three critical output analysis applications: (1) ACF plots, (2) estimates of the Monte Carlo asymptotic covariance matrix, and (3) estimates of the effective sample size. Under weak conditions, we establish strong consistency of our improved asymptotic covariance estimator, and obtain its large-sample bias and variance. The performance of the new estimators is demonstrated through various examples.

A hierarchical prior for generalized linear models based on predictions for the mean response

Ethan Alt Thursday, June 30

Harvard Medical School

There has been increased interest in using prior information in statistical analyses. For example, in rare diseases, it can be difficult to establish treatment efficacy based solely on data from a prospective study due to low sample sizes. To overcome this issue, an informative prior for the treatment effect may be elicited. We develop a novel extension of the conjugate prior of Chen and Ibrahim (2003) that enables practitioners to elicit a prior prediction for the mean response for generalized linear models, treating the prediction as random. We refer to the hierarchical prior as the hierarchical prediction prior (HPP). For independent and identically distributed settings and the normal linear model, we derive cases for which the hyperprior is a conjugate prior. We also develop an extension of the HPP in situations where summary statistics from a previous study are available, drawing comparisons with the power prior. The HPP allows for discounting based on he quality of individual level predictions, and simulation results suggest that, compared to the conjugate prior and the power prior, the HPP efficiency gains (e.g., lower MSE) where predictions are incompatible with the data. An efficient MCMC algorithm is developed. Applications illustrate that inferences under the HPP are more robust to prior-data conflict compared to selected non-hierarchical priors.

Automatically tuning the amount of inference in probabilistic programs

McCoy Becker Thursday, June 30

Massachusetts Institute of Technology

Inference in probabilistic programs can be computationally costly. There is a widespread need for rigorous, automated methods that help probabilistic programmers control the cost of inference by tuning the number of SMC particles and the number of MCMC sweeps that they use. We introduce a new method for solving this problem for a broad class of probabilistic programs that use resample-move SMC for inference. Our approach is based on automating and integrating simulation-based calibration, for assessing average-case inference quality via synthetic data, with pseudo-marginal estimators of KL divergence, for estimating data dependence. We illustrate results using a prototype implementation for the Gen probabilistic programming platform.

Concrete Horseshoe for high-dimensional regression models

Anupreet Porwal Thursday, June 30

Department of Statistics, University of Washington

In recent years, a rich variety of continuous shrinkage priors have been proposed to enforce sparsity in statistical modelling and machine learning problems. Often represented as global-local scale mixtures of normals, these priors are designed to generate a horseshoe-shaped shrinkage profile mimicking the Bernoulli distribution. Maddison et al., 2016 proposed the Concrete distribution as a continuous relaxation of the Bernoulli random variable using the Gumbel-softmax reparameterization trick. In this work, we propose “Concrete Horseshoe” prior, obtained by modelling the shrinkage factor using the Concrete distribution. The implied prior on the regression coefficients can be expressed as a scale mixture of normals with generalised beta prime distribution as its mixing density. We show that the following regularly varying prior is tail-robust and automatically adapts its tail heaviness to account for varying sparsity levels in the observed data. The performance of the prior is tested through simulations and various applications.

Bayesian Clustering via Mixed Integer Nonlinear Programming

Ji Ah Lee Thursday, June 30

University of Massachusetts Amherst

We present a global optimization approach for solving a maximum a-posteriori clustering problem under a Gaussian mixture model. Since our focus is on the optimal partition of samples into clusters, we employ a Bayesian framework to analytically marginalize over the model parameters. Our approach can easily accommodate hard constraints on cluster membership and preserves the combinatorial structure of the MAP clustering problem by its formulation as a mixed-integer nonlinear optimization problem, in contrast to standard inferential methods such as expectation maximization or Markov chain Monte Carlo. Numerical experiments show that our method solves standard problems to optimality and consistently recovers high-quality clusterings as measured by the earthmover's distance and variation of information metric. Finally, we cluster a real breast cancer gene expression data set incorporating known intrinsic subtype information and find that the constraints yield more coherent and biologically meaningful clusters.

Enhanced Prediction Using Covariate Informed Product Partition Models

Nathan Hawkins Thursday, June 30

Brigham Young University

Covariate informed product partition models excel at modeling data that are non-linear in nature. These models exploit available covariates by increasing the probability of co-clustering for two individuals with similar covariate values. In practice, however, prediction using these models can be computationally expensive and constrained to specific model parameters. We develop a covariate-based prediction algorithm that runs at compiled speed and allows for changes in model parameters to improve fit. We show the utility of this algorithm using data from the 2018 men's world volleyball championship. First, we train a product partition model using team performance statistics from the round robin stage of the tournament to predict wins. We then use our algorithm to predict the win probability after every point of every match in the knockout stages.

Bayesian Spatiotemporal Modeling of County-Level Drug Overdose Death Increases in the United States During the COVID-19 Pandemic

Jay Xu Thursday, June 30

UCLA Biostatistics

Fatal drug overdoses, which have been a growing public health crisis, exceeded 91,000 during 2020 (an all time-record for a single calendar year) amid the social and economic fallout driven by the COVID-19 pandemic, an approximate 30% increase from the previous year. We characterize the uneven societal burden of drug overdose deaths in the United States by linking county-level panel data on drug overdose deaths from the CDC with county-level socioeconomic features from the US Census Bureau and the American Community Survey and performing a Bayesian spatiotemporal analysis to identify county-level socioeconomic attributes associated with an excess population-level burden of drug overdose deaths amid the COVID-19 pandemic. Using a hierarchical spatiotemporal Poisson model as a modeling strategy, we provide credible interval estimates of the parameters of interest and supply graphics illustrating important geographic “hotspots" corresponding to regions that were disproportionately vulnerable to drug overdose deaths during the COVID-19 pandemic.

Past convictions and the probability of guilt

Ian Hunt Thursday, June 30

University of Tasmania

This research is a critique of the literature that applies probability analysis to past convictions, in the context of determining guilt in criminal trials. Recent arguments for potentially relaxing rules that exclude past conviction evidence are sustained, but particular flaws and limitations in the likelihood-ratio-based theses from Hamer (2019) and Redmayne (2015) are exposed. The critique of Bayesian probabilistic analysis made by Robinson (2020) is largely dismissed. We should aim to foster a continued lively debate in the literature, gather more data, and narrow the distance between those arguing about theoretical probability analysis and those focused on actual courtroom usage of past conviction evidence.

A Bayesian non-parametric model for 2-multiple context free grammars and inference of grammar complexity by Bayes factors

Caroline Lawless Thursday, June 30

University of Oxford

The class of context-free grammars is believed to be too restrictive to fully describe all features of natural language. The class of context-sensitive grammars, on the other hand, is too complex: modelling with them would require an unrealistic amount of computational time. Various mildly context-free grammar formalisms, which may be placed between context-free grammars and context-sensitive grammars in terms of complexity, have thus been proposed in the last few decades. We will be interested in the class of 2-multiple context-free grammars (2-MCFGs), which properly include the class of context-free grammars.

We propose a Bayesian non-parametric model for 2-MCFGs within which a model for context-free grammars is naturally embedded. Our model is analagous to that of Ryder et. al (in progress) for context-free and regular grammars. We provide theoretical results on the probability of a grammar being context-free, and on the probability of a sentence being finite under our model. We apply our model to Muriquis monkey vocalisation data. As in Ryder et. al (in progress), we carry out model choice by Bayes factors using sequential Monte Carlo in Birch probabilistic programming language.

Power laws distributions in objective priors

Francisco Louzada Thursday, June 30

Universidade de Sao Paulo

The use of objective prior in Bayesian applications has become a common practice to analyze data without subjective information. Formal rules usually obtain these priors distributions, and the data provide the dominant information in the posterior distribution. However, they are typically improper and may lead to improper posterior. We show for a general family of distribution that the obtained objective priors for the parameters either follow a power-law distribution or has an asymptotic power-law behavior. As a result, we observed the exponents of the model are between 0.5 and 1. Understand these behaviors allow us to easily very if such priors lead to proper or improper posteriors directly from the exponent of the power-law. The general family considered in this study includes essential models such as Exponential, Gamma,Weibull, Nakagami-m, Haf-Normal, Rayleigh, Erlang, and Maxwell Boltzmann distributions to list a few. In summary, comprehending the mechanisms that describe the shapes of the priors provides essential information that can be used in situations where additional complexity is presented.

Introducing a Bayesian two-stage logistic-normal model for small area estimation of proportions

Jamie Hogg

Thursday, June 30

QUT

With the rise in popularity of digital Atlases to communicate spatial variation in health to the public, there is an increasing need for robust small area proportion estimates. Useful predictors of health outcomes are generally not available for all individuals and thus it is important to develop small area estimation methods that can incorporate survey-only covariates at the individual level. We propose a new Bayesian hierarchical model that accounts for survey design and leverages both individual-level survey-only and area-level census covariates to simultaneously reduce the variance and bias of estimates and allow for the prediction of proportions for non-sampled areas. Initial testing of our model's performance using simulated survey data, compared with existing Bayesian SAE methods, showed that our novel model provides optimal performance when important individual-level survey-only covariates are utilized. Finally, we conducted a case study to estimate how the prevalence of smoking varied between small areas in Australia.