Bayes4Health Bayesian Data Science for Health Research

Workshop 14th-16th September 2021, Cambridge

List of Abstracts

Sergio Bacallado

Bayesian CCA with structural constraints for the analysis of the microbial metabolome

Canonical Correlations Analysis is a classical technique for studying multiview data. In microbiome research, it has been applied successfully to analyse the relationship between the genetic material in a bacterial community (metagenomic or transcriptomic data), and the most prevalent metabolites in the sample. Starting from a probabilistic interpretation of CCA (Bach and Jordan, 2006), we propose a Bayesian procedure which (i) imposes causal constraints from metabolic networks through zeros in the precision matrix, and (ii) limits the effective rank of the cross-covariance matrix. This represents an alternative to sparse CCA for dealing with high-dimensional datasets, in situations where there is valuable prior information on the relationship between the variables.

Alice Corbella (University of Warwick)

Automatic Zig-Zag Sampling

Recent research showed that Piecewise Deterministic Markov Processes (PDMP) may be exploited to design efficient MCMC algorithms [1]. The Zig-Zag sampler is an example of this: it is based on the simulation of a PDMP whose switching rate $\lambda(t)$ is governed by the derivative of a (minus log) target density.

While many properties of this sampler have been derived [2, 3], less has been done to explore the applicability of the Zig-Zag sampler to solve Bayesian inference problems. One of the main obstacles towards a wider use of this method by the community is the computation of the derivative of the log-density in the rate $\lambda(t)$. This can be particularly challenging when dealing with likelihoods containing recursive relationships, for which manual differentiation is time-consuming.

To expand the applicability of the Zig-Zag sampler, we incorporate Automatic Differentiation (AD) tools in the Zig-Zag algorithm, to facilitate the computation of $\lambda(t)$ from the functional form of the log-target density. Moreover, to allow the simulation of a PDMP via Poisson thinning, we use univariate optimization routines to find a local upper bound for the bounding rate.

In this talk we present our Automatic Zig-Zag sampler; we discuss the challenges that arise with the simulation via thinning and the need of a new tuning parameter; and we comment on efficiencies and bottlenecks of AD for Zig-Zag.

Joint work with Simon Spencer and Gareth Roberts

[1] Fearnhead, P., Bierkens, J., Pollock, M., and Roberts, G.O., 2018. Piecewise deterministic Markov processes for continuous-time Monte Carlo. Statistical Science, 33(3), pp.386-412.

[2] Bierkens, J., Fearnhead, P., and Roberts, G., 2019. The zig-zag process and super-efficient sampling for Bayesian analysis of big data. The Annals of Statistics, 47(3), pp.1288-1320.

[3] Bierkens, J., Roberts, G.O., and Zitt, P.A., 2019. Ergodicity of the zigzag process. Annals of Applied Probability, 29(4), pp.2266-2301.

Rob Cornish

Variational Inference with Continuously-Indexed Normalizing Flows

Continuously-indexed flows (CIFs) have recently achieved improvements over baseline normalizing flows on a variety of density estimation tasks. CIFs do not possess a closed-form marginal density, and so, unlike standard flows, cannot be plugged in directly to a variational inference (VI) scheme in order to produce a more expressive family of approximate posteriors. However, we show here how CIFs can be used as part of an auxiliary VI scheme to formulate and train expressive posterior approximations in a natural way. We exploit the conditional independence structure of multi-layer CIFs to build the required auxiliary inference models, which we show empirically yield low-variance estimators of the model evidence. We then demonstrate the advantages of CIFs over baseline flows in VI problems when the posterior distribution of interest possesses a complicated topology, obtaining improved results in both the Bayesian inference and surrogate maximum likelihood settings.

Maria de Iorio (University College London)

Seemingly Unrelated Multi-State processes: a Bayesian semiparametric approach

Many applications in medical statistics as well as in other fields can be described by transitions between multiple states (e.g. from health to disease) experienced by individuals over time. In this context, multi-state models are a popular statistical technique, in particular when the exact transition times are not observed. The key quantities of interest are the transition rates, capturing the instantaneous risk of moving from one state to another. The main contribution of this work is to propose a joint semiparametric model for several possibly related multi-state processes (Seemingly Unrelated Multi-State, SUMS, processes), assuming a Markov structure for the transitions over time. The dependence between different processes is captured by specifying a joint random effect distribution on the transition rates of each process. We assume a flexible random effect distribution, which allows for clustering of the individuals, overdispersion and outliers. Moreover, we employ a graph structure to describe the dependence among processes, exploiting tools from the

Gaussian Graphical model literature. It is also possible to include covariate effects. We use our approach to model disease progression in mental health. Posterior inference is performed through a specially devised MCMC algorithm.

Ritabrata Dutta (University of Warwick)

Generalized Bayesian Likelihood-Free Inference Using Scoring Rules Estimators

We propose a framework for Bayesian Likelihood-Free Inference (LFI) based on Generalized Bayesian Inference using scoring rules (SRs). SRs are used to evaluate probabilistic models given an observation; a proper SR is minimised in expectation when the model corresponds to the data generating process for the observations. Using a strictly proper SR, for which the above minimum is unique, ensures posterior consistency of our method. Further, we prove finite sample posterior consistency and outlier robustness of our posterior for the Kernel and Energy Scores. As the likelihood function is intractable for LFI, we employ consistent estimators of SRs using model simulations in a pseudo-marginal MCMC; we show the target of such chain converges to the exact SR posterior by increasing the number of simulations. Furthermore, we note popular LFI techniques such as Bayesian Synthetic Likelihood (BSL) can be seen as special cases of our framework using only proper (but not strictly so) SR. We empirically validate our consistency and outlier robustness results and show how related approaches do not enjoy these properties. Practically, we use the Energy and Kernel Scores, but our general framework sets the stage for extensions with other scoring rules.

(Joint work with Lorenzo Pacchiardi)

Richard Everitt (University of Warwick)

Rare event ABC-SMC^2

We propose a new algorithm based on a combination of an SMC sampler for estimating the ABC likelihood in the case of high-dimensional data (Prangle et al, 2018) and ABC-SMC for exploring the parameter space. The new method has a similar structure to SMC2 (Chopin et al, 2012). To automate the approach, we make use of an adaptive scheme for both the sequence of ABC tolerances in the SMC, and also for the MCMC rejuvenation steps of the external parameter space SMC. This is joint work with Ivis Kerama and Tom Thorne.

Barbel Finkenstadt (Department of Statistics,University of Warwick)

Inference for Circadian Pacemaking

Organisms have evolved an internal biological clock which allows them to temporally regulate and organize their physiological and behavioral responses to cope in an optimal way with the fundamentally periodic nature of the environment. It is now well established that the molecular genetics of such rhythms within the cell consist of interwoven

transcriptional-translational feedback loops involving about 15 clock genes, which generate circa 24-h oscillations in many cellular functions at cell population or whole organism levels. We will present statistical methods and modelling approaches that address newly emerging large circadian data sets, namely spatio-temporal gene expression in SCN neurons and rest-activity actigraph data obtained from non-invasive e-monitoring, both of which provide unique opportunities for furthering progress in understanding the synchronicity of circadian pacemaking and address implications for monitoring patients in chronotherapeutic healthcare.

Sebastian Funk (London School of Hygiene and Tropical Medicine)

Bayesian nowcasting and forecasting of Covid-19

During an epidemic, forecasts are an essential tool for both policy makers and the general public to prepare or allocate present resources for future needs, while researchers can use the performance of different forecasting models to learn about strengths and weaknesses of different approaches. In this talk I will introduce some of the statistical and practical challenges in forecasting the dynamics of Covid-19, and present an approach we have developed and implemented in Stan, and made available to the broader community in the R package EpiNow2, that addresses some of these challenges. I will discuss the performance of this model in forecasting Covid-19 cases and deaths in the context of a broader class of models that have been used to make predictions of Covid-19 at different levels of spatial aggregation

Chris Jewell (University of Lancaster)

tbc

Brieuc Lehmann

A predictive approach to Bayesian nonparametric survival analysis

Bayesian nonparametric methods are a popular choice for analysing survival (i.e. right-censored) data due to their ability to flexibly model the distribution of survival times, enabling rich inference without relying on restrictive parametric assumptions. These methods typically employ a nonparametric, neutral-to-the-right process prior on the hazard or survival function, chosen for its conjugacy properties with respect to right-censored data. Here, we build on recent work that recasts Bayesian inference as assigning a predictive distribution on $y_{n+1:\infty}$ conditional on observing $y_{1:n}$, thus avoiding the need to specify such a prior. We describe a copula-based predictive update that properly accounts for right-censoring and admits a simple importance sampling algorithm to perform inference. We provide theory to justify the validity of our predictive approach and illustrate the method on a number of simulated and real data sets. Our approach adds to the suite of

tools for survival data, enabling analysts to perform inference through the specification of a predictive distribution instead of a prior distribution on the survival or hazard function. Joint work with Edwin Fong and Chris Holmes.

Ben Leimkuhler (University of Edinburgh)

Partitioned and Multirate Training of Neural Networks

I will describe some new algorithms for training neural networks within a Bayesian or 'cold posterior' setting. Our algorithms are based on partitioning the parameters of the network in various way and apply different algorithms, constraints, or different learning rates, to different blocks of components. By choosing appropriate partitionings we can obtain large computational speed-ups for various tasks including transfer learning in vision and natural language processing, where we can fine-tune deep neural networks in almost half the time without reducing the generalization performance. I will also briefly discuss several approaches to regularization that improve generalization outcomes. This talk is joint work with Tifffany Vlaar and Charles Matthews.

Renee Menezes (Netherlands Cancer Institute, Amsterdam, The Netherlands)

A Bayesian framework for efficient testing of differential exon usage

A gene can be defined as a genomic area coding for a specific protein. Genes involve one or more exons, each one encoding a part of the final mature mRNA produced. Splicing of a gene involves eliminating introns (parts of DNA not directly coding for protein) and joining exons together in the final mRNA; alternative splicing occurs when exons are included or excluded, yielding a protein with possibly a different function. This means that alternative splicing allows a gene to have isoforms, i.e. to produce different protein products. Evidence of alternative splicing can be found by studying changes in the distribution of counts across exons of a gene, using RNA-seq data. This is in general referred to as “differential exon usage” (DEU).

The few methods for DEU testing currently available require heavy multiple testing correction as testing is done per exon, and do not use the data in an efficient way. We propose ShrinkISO, a DEU test based upon a Bayesian regression model per gene, including data of all the gene’s exons. This yields a powerful test because it uses all exons, and requires multiple testing only across genes, which is less severe than across all exons of all genes. In addition, the test can be fine-tuned to ignore effects too small to be biologically relevant, helping researchers focus on findings more likely to be relevant. Furthermore, a covariance structure can be used to better model the association between exons. Finally, the model uses a zero-inflation negative binomial distribution for the counts, making it ideal for using with sequencing data. We will evaluate reproducibility of results obtained when studying different cancer types, for various sample sizes.

Filippo Pagani

Numerical ZigZag

In the field of MCMC (Markov chain Monte Carlo) there has recently been a surge of interest in algorithms based on stochastic processes with irreversible dynamics. Such algorithms possess appealing theoretical properties, such as potentially faster mixing than their reversible counterparts, and superefficient sampling on large datasets. In this talk we focus on the NuZZ (Numerical ZigZag) algorithm, an algorithm based on the ZigZag Sampler which computes the time to the next switch numerically, circumventing the need for analytical solutions. We use NuZZ to study the ZigZag dynamics on some test problems with important features, and we discuss theoretical bounds on the numerical error of approximated PDMPs.

Sam Power

Accelerated Sampling on Discrete Spaces with Non-Reversible Markov Jump Processes

In Bayesian inference problems and elsewhere, Markov Chain Monte Carlo (MCMC) algorithms are an indispensable tool for sampling from complex probability distributions. On continuous state-spaces, there has been a great deal of successful work on how to construct efficient MCMC dynamics which can converge quickly, under very general circumstances. Much of this success has stemmed from identifying continuous-time dynamical processes (ODEs, SDEs, PDMPs) which admit the desired invariant measure, and then discretising those processes to form tractable discrete-time chains. This approach has apparently seen less use in the setting of discrete spaces.

In this work, we aim to bridge this gap by identifying `canonical' Markov processes (both reversible and non-reversible) on structured *discrete* spaces which admit a given invariant measure, and then use them to derive new algorithms for efficient sampling on discrete spaces. The algorithms are parameter-free (no tuning is required) and can be simulated directly in continuous time, easily and without discretisation error. We provide theoretical results supporting the use of non-reversible dynamics, and a range of numerical experiments demonstrate the practical benefits of our algorithms.

This is joint work with Jacob Vorstrup Goldman (Cambridge).

Lorenzo Rimella

Exploiting locality in high-dimensional Factorial Hidden Markov models

We propose algorithms for approximate filtering and smoothing in high-dimensional Factorial Hidden Markov models. The approximation involves factorizing the target distributions and discarding likelihood factors according to a notion of locality in a factor graph associated with the emission distribution. This allows the exponential-in-dimension cost of exact filtering-smoothing to be avoided and results in low-cost algorithms which

depend on the properties of factor graph of the emission distribution (e.g. maximum number of neighbours, maximum number of common neighbours, etc.). We prove that the approximation accuracy, measured in a local total variation norm, is “dimension-free” in the sense that as the overall dimension of the model increases the error bounds we derive do not necessarily degrade. A key step in the analysis is to quantify the error introduced by localizing the likelihood function in a Bayes’ rule update. Computational cost and accuracy are both deeply analysed in a synthetic data scenario. The factorial structure of the likelihood function which we exploit arises naturally when data have known spatial or network structure. We demonstrate the new algorithms on synthetic examples and a London Underground passenger inflow-outflow problem, where the factor graph is effectively given by the train network. (https://arxiv.org/abs/1902.01639)

Stephanie van der Pas

Multiscale Bayesian Survival Analysis

Joint work with Ismaël Castillo. We consider Bayesian nonparametric inference in the right-censoring survival model, where modeling is made at the level of the hazard rate. We derive posterior limiting distributions for linear functionals of the hazard, and then for `many' functionals simultaneously in appropriate multiscale spaces. As an application, we derive Bernstein-von Mises theorems for the cumulative hazard and survival functions, which lead to asymptotically efficient confidence bands for these quantities. Further, we show optimal posterior contraction rates for the hazard in terms of the supremum norm. In medical studies, a popular approach is to model hazards a priori as random histograms with possibly dependent heights. This and more general classes of arbitrarily smooth prior distributions are considered as applications of our theory. A sampler is provided for possibly dependent histogram posteriors. Its finite sample properties are investigated on both simulated and real data experiments.

Lorenz Wernisch (Bios Health Ltd, Cambridge, https://www.bios.health/)

Bayesian optimization of protocols for neurostimulation

Neurostimulation, the application of small currents to specific peripheral nerves to trigger certain responses in organs or the brain, provides a promising alternative to traditional drug treatments. Multiple parameters determine the precise stimulation pattern making an exhaustive systematic search for optimal settings time consuming or even unfeasible. Moreover, response measurements are often noisy. Bayesian optimization provides an ideal tool for an efficient parameters search in this case.

Christopher Yau (University of Manchester)

Multimorbidity and the Wright-Fisher Indian Buffet Process

Multimorbidity refers to the acquisition of multiple diseases in the same individual. Individuals who have a complex array of multiple conditions, particularly later in life, poses severe challenges for the health system in which medicine is typically optimised only for the treatment of single conditions. Understanding how multimorbidity arises and its patterns could reveal novel insights into more effective treatment strategies.

One approach to doing this is to create probabilistic multimorbidity trajectory models from electronic health records. These trajectories would chart the time-dependent acquisition of disease conditions in an individual. We have developed a model based on a Bayesian nonparametric feature allocation model with a Wright--Fisher Indian Buffet Process prior. Our model, which we call the Multimorbidity Wright--Fisher Indian Buffet Process (m-WFIBP) defines a generative process in which a set of diseases of an individual is drawn from latent multimorbidity clusters whose dependency structure across time is governed by the Wright--Fisher diffusion.

We demonstrate the utility of our model in applications to simulated data and disease event data from patient electronic health records. In both settings, we show how the m-WFIBP can obtain intelligible representation of latent multimorbidity clusters and its time susceptibility and predict future disease acquisition.