Autumn school in Bayesian Statistics 2023

Research school

30th October - 3rd November 2023, CIRM, Marseille

https://bayesatcirm.github.io/2023/

Program with abstracts

Monday

Masterclass: Nicolas Chopin (Monday 9:00)

An introduction to state-space models, particle filters, and Sequential Monte Carlo samplers

This course will provide a general introduction to SMC algorithms, from basic particle filters and their uses in state-space (hidden Markov) modelling in various areas, to more advanced algorithms such as SMC samplers, which may be used to sample from one, or several target distributions. The course will cover ""a bit of everything"": theory (using Feynman-Kac models as a general framework), methodology (how to construct better algorithms in practice), implementation (examples in Python based on the library particles will be showcased), and applications.

Invited talk: François Caron (Monday 11:30)

Sparse graphs based on exchangeable random measures: properties, models and examples

Random simple and multigraph models based on exchangeable random measures, aka graphex processes or generalised graphon models, have recently been proposed as a versatile class of sparse random graph models. This class of models can be seen as a generalisation of the popular graphon models. I will present this class of models, discuss some of their asymptotic properties (degree distribution, clustering coefficients). I will also present some particular models with interpretable parameters within this class and their use for discovering latent communities in sparse real-world networks.

Lunch at 12:30

Tutorial Gabriel Victorino Cardoso and Yazid Janati El Idrissi (Monday 14:00)

Bayesian Statistics with Python: part I

Contributed talk: Liam Llamazares Elias (Monday 16:30)

Penalized complexity priors for stochastic partial differential equations

Penalized complexity priors are used in Bayesian inference to define priors that penalize the distance from a base model. We investigate how to use this framework to set priors for the coefficients of a spatial stochastic partial differential equation (SPDE). We first work with stationary solutions to the SPDE. In this case, the coefficients are finite-dimensional and not

functions of space, leading to a more straightforward analysis. We then extend this to the non-stationary setting. In this case, as the coefficients are themselves functions of space, the parameter space is infinite-dimensional. We show how to extend the concept of spectral density to non-stationary fields. We then use this non-stationary spectral density to define a distance between the parameters of the SPDE and calculate a penalized complexity prior.

Contributed talk: Joshua Bon (Monday 17:00)

Bayesian score calibration for approximate models

We propose a new method for adjusting samples from an approximate posterior to reduce bias and produce more accurate uncertainty quantification. We do this by optimising a transform of the approximate posterior that maximises a scoring rule. The procedure is Bayesian but has some interesting parallels with Frequentist calibration.

Posters and cocktail (Monday from 17:30)

Dinner at 19:30

Tuesday

Masterclass: Nicolas Chopin (Tuesday 9:00)

An introduction to state-space models, particle filters, and Sequential Monte Carlo samplers

Contributed talk: Davide Agnoletto (Tuesday 11:30)

Bayesian inference for generalized linear models via quasi-posteriors

Generalized linear models are a standard statistical tool for modeling the relation between a response variable and a set of covariates. Despite their popularity, they can incur in misspecification problems that could negatively impact inferential conclusions. A semi-parametric solution adopted in frequentist literature uses the quasi-likelihood function, which relies on the second-order assumptions, in place of the usual likelihood. This approach yields increased flexibility since only the first two moments of the data generator are specified rather than its entire distribution. We propose to integrate this solution in the Bayesian paradigm introducing the quasi-posterior distribution. This quantity represents a coherent Bayesian update according to the generalized Bayes notion. We show that quasi-posterior approximates the regression coarsened posterior in the case of exponential families, providing new insights on the choice of the coarsening parameter. Asymptotically, the quasi-posterior converges in total variation to a normal distribution, has important connections with the loss-likelihood bootstrap posterior, and is also well-calibrated in terms of frequentist coverage. Moreover, the loss-scale parameter has a clear interpretation in terms of the dispersion parameter, leading to the consolidated method of moments estimator for its quantification. We provide some applications to overdispersed counts and heteroscedastic continuous data.

Contributed talk: Sylvain Le Corff (Tuesday 12:00)

Monte Carlo guided Diffusion for Bayesian linear inverse problems

Ill-posed linear inverse problems that combine knowledge of the forward measurement model with prior models arise frequently in various applications, from computational photography to medical imaging. Recent research has focused on solving these problems with score-based generative models (SGMs) that produce perceptually plausible images, especially in inpainting problems. In this study, we exploit the particular structure of the prior defined in the SGM to formulate recovery in a Bayesian framework as a Feynman-Kac model adapted from the forward diffusion model used to construct score-based diffusion. To solve this Feynman-Kac problem, we propose the use of Sequential Monte Carlo methods. The proposed algorithm, MCGdiff, is shown to be theoretically grounded and we provide numerical simulations showing that it outperforms competing baselines when dealing with ill-posed inverse problems.

Lunch at 12:30

Invited talk: Marta Catalano (Tuesday 14:00)

Merging rate of opinions via optimal transport on random measures

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

Contributed talk: Francesca Romana Crucinio (Tuesday 15:00)

Optimal Scaling Results for a Wide Class of Proximal MALA Algorithms

We consider a recently proposed class of MCMC methods which uses proximity maps instead of gradients to build proposal mechanisms which can be employed for both differentiable and non-differentiable targets. These methods have been shown to be stable for a wide class of targets, making them a valuable alternative to Metropolis-adjusted Langevin algorithms (MALA); and have found wide application in imaging contexts. The wider stability properties are obtained by building the Moreau-Yoshida envelope for the target of interest, which depends on a parameter λ. In this work, we investigate the optimal scaling problem for this class of algorithms, which encompasses MALA, and provide practical guidelines for the implementation of these methods

Contributed talk: Mikołaj Kasprzak (Tuesday 15:30)

How good is your Laplace approximation? Finite-sample computable error bounds for a variety of useful divergences.

The Laplace approximation is a popular method for providing posterior mean and variance estimates. But can we trust these estimates for practical use? One might consider using rate-of-convergence bounds for the Bayesian Central Limit Theorem (BCLT) to provide quality guarantees for the Laplace approximation. But the bounds in existing versions of the BCLT

either: require knowing the true data-generating parameter, are asymptotic in the number of samples, do not control the Bayesian posterior mean, or apply only to narrow classes of models. Our work provides the first closed-form, finite-sample quality bounds for the Laplace approximation that simultaneously (1) do not require knowing the true parameter, (2) control posterior means and variances, and (3) apply generally to models that satisfy the conditions of the asymptotic BCLT. In fact, our bounds work even in the presence of misspecification. We compute exact constants in our bounds for a variety of standard models, including logistic regression, and numerically demonstrate their utility. We provide a framework for analysis of more complex models.

Invited talk: Jean-Michel Marin (Tuesday 16:30)

Goodness of Fit for Bayesian Generative Models

Goodness-of-fit methods (GOF) aim at evaluating the level of adequacy between the observed dataset and a given model of interest, typically using an hypothesis-testing approach. In an Approximate Bayesian Computation context, this question can be re-framed as a novelty detection problem, in which one seeks to evaluate to which extent the observed dataset is an outlier compared to the simulated datasets. Many scores have been used as metrics to construct GOF test statistics and have been extensively tested in the literature. Here we propose a score based on the Local Outlier Factor.

Social activities & sports (Tuesday from 18:00)

Dinner at 19:30

Wednesday

Masterclass: Silvia Chiappa (Wednesday 9:00)

Bayesian causal inference

Contributed talk: Jacopo Iollo (Wednesday 11:30)

Tempered sequential Monte Carlo for Bayesian experimental design via stochastic optimization

We propose a new procedure, for Bayesian experimental design, that performs sequential design optimization while simultaneously providing accurate estimates of successive posterior distributions for parameter inference. The sequential design process is carried out via a contrastive estimation principle, using stochastic optimization and tempered Sequential Monte Carlo (SMC) samplers to maximise the Expected Information Gain (EIG). As larger information gains are obtained for larger distances between successive posterior distributions, this EIG objective worsens classical SMC performance. To handle this issue, tempering is proposed to have both a large information gain and an accurate SMC sampling. %tempering is proposed to improve SMC sampling accuracy. This novel combination of stochastic optimization and tempered SMC allows to jointly handle design optimization and parameter posterior inference. We provide a proof that the obtained optimal design estimators benefit from some consistency property. Numerical experiments confirm the approach potential on various benchmarks where our procedure outperforms other recent existing approaches.

Contributed talk: Mengyan Zhang (Wednesday 12:00)

Bayesian optimisation with aggregated feedback

We consider the Bayesian optimisation problem, under a novel setting of aggregated feedback. This is motivated by applications where the precise rewards are impossible or expensive to obtain, while an aggregated reward or feedback, such as the average over a subset, is available. We adaptively construct a tree with nodes as subsets of the arm space and propose Gaussian Process Optimistic Optimisation (GPOO) algorithm.

Lunch at 12:30

Free time (Wednesday afternoon)

Dinner at 19:30

Thursday

Masterclass: Silvia Chiappa (Thursday 9:00)

Bayesian causal inference

Contributed talk: Elena Bortolato (Thursday 11:30)

Coupling MCMC algorithms on submanifolds

Manifolds arise in diverse problems in statistics, demanding effective methods for sampling probability distributions defined on them and specialized Markov Chain Monte Carlo (MCMC) methods since Andersen (1983) were developed. Such techniques can be also applied to a variety of sampling problems, even not naturally defined on manifolds. We propose to combine standard MCMC samplers with such manifold steps and develop implementable coupling schemes to diagnose convergence and to evaluate the efficiency.

Contributed talk: Thibaut Lemoine (Thursday 12:00)

Monte Carlo integration on complex manifolds

Lunch at 12:30

Tutorial Gabriel Victorino Cardoso and Yazid Janati El Idrissi (Thursday 14:00)

Bayesian Statistics with Python: part II

Invited talk: Fabrizia Mealli (Thursday 16:30)

TBA

Contributed talk: Filippo Ascolani (Thursday 17:30)

Complexity of Gibbs samplers through Bayesian asymptotics

Gibbs samplers are popular algorithms to approximate posterior distributions arising from Bayesian hierarchical models. Despite their popularity and good empirical performances, however, there are still relatively few quantitative theoretical results on their scalability or lack thereof, e.g. much less than for gradient-based sampling methods. We introduce a novel technique to analyse the asymptotic behaviour of mixing times of Gibbs Samplers, based on tools of Bayesian asymptotics. We apply our methodology to high-dimensional hierarchical models, obtaining dimension-free convergence results for Gibbs samplers under random

data-generating assumptions, for a broad class of two-level models with generic likelihood function. Specific examples with Gaussian, binomial and categorical likelihoods are discussed.

Quiz/games (Thursday 18:30)

Dinner at 19:30

Friday

Invited talk: Sophie Donnet (Friday 9:00)

Using a Sequential Monte Carlo algorithm to find the mesoscale structure of a network

This work is motivated by the analysis of ecological interaction networks. Stochastic block models are widely used in this field to decipher the structure that underlies a network or that is shared by a collection of networks. Efficient algorithms based on variational approximations exist for frequentist inference and sometimes for Bayesian inference, but without statistical guarantees as for the resulting estimates. We propose to combine the variational estimation with a sequential Monte-Carlo algorithm to efficiently sample the posterior distribution and to perform model selection.

Contributed talk: Anna Menacher (Friday 10:00)

Scalar-on-image regression with a relaxed Gaussian process prior

In this work, we provide a scalable hierarchical Bayesian spatial model for scalar-on-image regression problems with a relaxed thresholded Gaussian process prior on the spatially-varying parameters. We achieve the same properties as the soft thresholded Gaussian process prior, developed by Kang et al. (2018), by introducing an additional set of parameters that perform the thresholding; however, our algorithm does not rely on a Metropolis-within-Gibbs sampler to sample the parameters. We aim to utilize variational inference and thereby speed up the performance of the algorithm to enable its application to large-scale population health studies as well as to increase the number of imaging modalities that can be introduced to the model. We support our work with extensive simulation studies and provide a real data application to the UK Biobank.

Contributed talk: Claudio Del Sole (Friday 10:30)

Hierarchically dependent mixture hazard rates for modelling competing risks

A popular approach in Bayesian modelling of partially exchangeable data consists in imposing hierarchical nonparametric priors, which induce dependence across groups of observations. In survival analysis, hierarchies of completely random measures have been successfully exploited as mixing measures to model multivariate dependent mixture hazard rates, leading to a posterior characterization which may also accommodate censored observations. Such framework can be easily adapted to a competing risks scenario, in which groups correspond to different diseases affecting each individual: in this case, the multivariate construction acts at a latent level, as only the minimum time-to-event and the corresponding cause of death are actually observed. The posterior hierarchy of random measures, as well as the posterior estimates of both survival function and cause-specific incidence functions are explicitly

described, conditionally on a suitable latent partition structure which fits the Chinese restaurant franchise metaphor. Marginal and conditional sampling algorithms are also devised and tested on synthetic datasets. The performances of this proposal are finally compared with those of its non-hierarchical counterpart, which models the hazard rate of each disease independently: leveraging the information borrowed from other groups, the hierarchical construction is empirically shown to recover the shape of the incidence functions more efficiently, in presence of proportional hazards.

Contributed talk: Uribe Felipe (Friday 11:30)

Towards dimension reduction of Bayesian inverse problems with neural network priors

In many Bayesian inverse problems the change from prior to posterior is confined to a low-dimensional subspace of the parameter space. We explore gradient-based dimension reduction techniques to identify this crucial subspace, with the primary aim of enhancing the efficiency of MCMC methods employed in tackling the inverse problem.

Closing remarks at 12:00

Lunch at 12:30

Posters

Louise Alamichel

Bayesian mixture models (in)consistency for the number of clusters

Bayesian non-parametric mixture models are commonly used to model complex data. Although these models are well suited to density estimation, their application to clustering has certain limitations. Recent results proved posterior inconsistency of the number of clusters when the true number of clusters is finite for the Dirichlet and Pitman-Yor process mixture models. Some possible solutions have also been proposed recently to achieve consistency for the number of clusters, notably in the case of the Dirichlet process by using a post-processing algorithm or putting a hyperprior on the parameter. We discuss and extend these results to other non-parametric Bayesian priors such as Gibbs-type processes and their finite-dimensional representations such as the Dirichlet multinomial or Pitman-Yor multinomial processes. We prove that mixture models based on these processes are also inconsistent concerning the number of clusters. We also show that the post-processing algorithm can be extended to more general models and provides a consistent method for estimating the number of components. Finally we study, for the Pitman-Yor process, the role played in consistency by a hyperprior on the parameters.

Arka Banerjee

Fast MCMC: A new multivariate variance estimator

Monte Carlo error estimation is essential for measuring sample quality of samples generated from a population under study in the Bayesian paradigm. Error variance estimation is one of the standard practices for both univariate and multivariate Markov chains. Multivariate batch means estimator, being one of the fastest and less volatile among the widely used estimators, does not perform well for highly correlated Markov chains resulting in a biased confidence region for the concerned estimates of the population parameters. Here we modify the multivariate batch means estimators with a clever use of covariance-correlation transformation to get a fast and unbiased estimator even for highly correlated cases and compare its performance with other estimators.

André Felipe Berdusco Menezes

Bayesian automatic monitoring and intervention in Dynamic Linear Model: The pybats-detection package

This presentation focuses on the pybats-detection, an open source Python package designed to facilitate Bayesian automatic sequential monitoring and subjective intervention within the class of Dynamic Linear Model. In the pybats-detection the sequential automatic monitoring is

conducted by comparing the current model to alternative models designed to detect specific model departures such as outlier observations and abrupt structural changes. The Bayes factor, the relative predictive likelihood for two models, is used as a measure for model evaluation. Model adaptation actions are taken by increasing the parameter uncertainty for regime changes or ignoring potential outlier observations. To showcase the applicability of the pybats-detection package, three real-world time series analyses are utilized. Overall, the package was developed with the aim of encouraging users to include dynamic linear models in their projects.

Daria Bystrova

Approximating the clusters' prior distribution in Bayesian nonparametric models

In Bayesian nonparametrics, knowledge of the prior distribution induced on the number of clusters is key for prior specification and calibration. However, evaluating this prior is infamously difficult even for moderate sample size. We evaluate several statistical approximations to the prior distribution on the number of clusters for Gibbs-type processes, a class including the Pitman--Yor process and the normalized generalized gamma process. We introduce a new approximation based on the predictive distribution of Gibbs-type process, which compares favourably with the existing methods. We thoroughly discuss the limitations of these various approximations by comparing them against an exact implementation of the prior distribution of the number of clusters.

Stefano Cortinovis

Learning Dissipative Hamiltonian Dynamics with Gaussian Processes

Julie Fendler

Improving the inference of Bayesian profile regression mixture models to estimate the health effects of highly correlated co-exposures from censored survival outcomes.

We focus on the problem of estimating a disease risk from a highly censored survival outcome and a few highly correlated environmental exposures, for which simple multiple regressions may lead to unstable and unprecise risk estimates. We extend Bayesian profile regression mixture (BPRM) models to this context by assuming an instantaneous excess hazard ratio sub-model. This multilevel model incorporates a Dirichlet process mixture as an attribution sub-model. It allows clustering individuals with similar profiles, that is, with similar exposure characteristics, and estimating the associated excess hazard risk for each group, in a unique step. Inferring this model with a standard adaptive Metropolis-Within-Gibbs algorithm leads to convergence issues. The Markov chains are trapped in local maxima. Different sampling methods are compared to overcome this issue. Our BPRM model is applied to the estimation of the risk of death by lung cancer in the post-55 French cohort of uranium miners who were chronically and occupationally exposed to multiple and correlated sources of ionizing radiation: radon, gamma rays and

uranium dust. This case study shows that BPRM models are promising tools for exposome research and opens new avenues for methodological research in this class of models.

Samuele Garelli

A new class of predictive distributions for predictive resampling

Fong, Holmes, and Walker (2023) formalise a promising alternative approach to Bayesian inference. Rather than applying the usual “likelihood-prior to posterior” paradigm, they specify a predictive model for the data and reconstruct the whole population via a predictive resampling mechanism. Then, they estimate any parameter of interest by taking summary statistics of the imputed population. For this procedure to work, it is necessary that the predictive distributions converge to some random measure (to ensure that data are imputed from the same population) and that they have a good fit on the observed data (so that the imputed population is as close as possible to all that is known about the real one, i.e. the observed data). In this work, we focus on such framework and we study a new class of predictive distributions. The idea is that predictives should preserve the main features of the observed data, namely their modes/clusters and consequently clusters’ center and dispersion. Therefore, we decided to focus on the case of mixtures of Gaussian distributions. A clustering algorithm is run over the observed data to obtain the weight of each mixture component. Components’ center and dispersion are then set equal to cluster-specific mean and variance. We show a good fit of these models both on simulated and on real data. Moreover, we prove that these predictive distributions converge in total variation to some random measure. Since their parameters depend on data that are not IID, this is an interesting theoretical property of this class of predictive distributions. So far, we have shown convergence in the case of univariate data, multivariate data and regression. We are currently studying predictives for classification.
This work is in collaboration with Fabrizio Leisen, Luca Pratelli and Pietro Rigo.

References:

Fong, E., Holmes, C. and Walker, S.G. (2023). Martingale posterior distributions. To appear in Journal of the Royal Statistical Society, Series B (with discussion).

Francesco Gili

Efficient nonparametric estimation in Wicksell's problem

Valentin Kilian

Improving Gaussian Graphical Model inference by learning the graph structure

Gaussian Graphical Models (GGM) provide a suitable framework to infer direct links between variables. In a high-dimensional setting, when the number of variables exceeds the number of observations, the inference of a GGM is difficult and we propose a new method for GGM

inference in that context. This method is based on a multiple testing procedure that incorporates some learned latent graph structure. This allows to detect more significant links between variables while controlling the proportion of falsely detected links. Our procedure is based on the Noisy Stochastic Block Model (NSBM), an extension of the Stochastic Block Model (SBM) which models noisy interactions between entities. We propose to estimate the parameters of the NSBM with an extension of a greedy algorithm proposed by E. Côme in 2015 for the SBM and based on the exact integrated complete-data likelihood. Then we show how the NSBM can be used to infer GGM by taking as entries of the algorithm the values of some test statistics constructed to test if there is a direct link between two variables. Results on synthetic and real data are presented.

Emma Kopp

How far can we trust language phylogenies?

Antoine Luciano

Gibbs Sampling with Robust Statistics

In certain applied scenarios, the availability of complete data is restricted, often due to privacy concerns, and only robust statistics derived from the data are accessible. These robust statistics demonstrate reduced sensitivity to outliers and offer enhanced data protection due to their higher breakdown point. In this article, operating within a parametric framework, we propose a method to sample from the posterior distribution of parameters conditioned on different robust statistics: specifically, the pairs (median, MAD) or (median, IQR), or one or more quantiles. Leveraging a Gibbs sampler and the simulation of latent augmented data, our approach facilitates simulation according to the posterior distribution of parameters belonging to specific families of distributions, such as Gaussian, Cauchy, or translated Weibull.

Pedro Menezes de Araújo

Beta item response theory models applied to mortality rate data.

Beta item response theory models can be applied to (0,1) interval data to estimate latent traits and item-specific parameters. We employed this class of models to estimate the country's latent mortality effect using the mortality rate in different age groups as items. With that, we can summarise the country's effect on mortality over time with a probabilistic model, helping to objectively understand these mortality profile dynamics.
We used the two-parameter logistic model in different settings and also investigated the dimension of the latent country effect. We used Bayesian methods and Hamiltonian Monte Carlo to generate samples from the posterior distribution. We have evidence that two dimensions are enough to describe the data, one for the very older age group and one for younger age groups. We can also observe similar patterns for some countries, indicating the existence of clusters.

Exaucé Ngarti

Approximating Bayesian Posterior Distributions for Parameter Estimation in the Context of Shallow Water Equations

This work from my master's thesis focuses on inferring the posterior distribution of a specific parameter based on observations of its effects. This research is motivated by the need to estimate key parameters that govern natural phenomena, like fluid movement and wave propagation, which are described by physical equations. Some parameters, such as the roughness of the ocean floor, are challenging to directly observe due to their variability. Current approaches involve indirect observations of their effects on observable phenomena, introducing uncertainties into the estimation process, which is crucial for coastal regions.

Traditional parametric inference methods like maximum likelihood estimation sometimes struggle to incorporate prior knowledge about parameters, making a Bayesian approach more suitable. However, Bayesian point estimation techniques, such as maximum a posteriori estimation, may not fully represent parameter behavior. This research aims to derive the posterior distribution of the parameter, providing comprehensive information and enabling point estimates (e.g., mode or mean) while quantifying associated uncertainties.

The chosen approach is variational inference, which approximates the posterior distribution using a parameterized variational distribution. We show that this method works well in ideal cases but faces challenges with nuisance parameters. To address this, we are integrating neural networks to enhance the expressiveness of variational distributions and robustly estimate the parameter's posterior distribution.

Théo Silvestre

Assessing a dose-response relationship after brain radiotherapy via Mixture of Experts

Radiotherapy (RT) is one of the most important treatments for brain tumors. However, its potential toxicity on the central nervous system is a highly relevant clinical issue as cognitive dysfunction, mainly related to radiation-induced leukoencephalopathy (RIL), may alter the quality of life of patients. In this context, the aim of this work is to model and learn about the potential relationship between the dose of ionizing radiation absorbed in a voxel of the brain and the presence/absence of a white-matter hyperintensity (brain lesion) in this voxel, in patients treated with RT for glioblastoma. We propose a piecewise logistic regression that can be seen as an extension to the framework of a binary response variable of a locally linear Gaussian expert model, called GLIiM. Various Bayesian statistical learning methods (variational Bayes, MCMC) are implemented and compared from simulated data inspired by the EpiBrainRad prospective cohort, which includes patients treated with radiochemotherapy for glioblastoma. Many modeling perspectives and Bayesian computational challenges will also be discussed.

Federica Stolf

Dependent Infinite Latent Feature Models

The Indian Buffet Process (IBP) is a popular prior distribution in the rich literature on infinite latent feature models. The current literature focuses on the case where latent features are generated independently, which entails no within-sample dependence in feature selection. Motivated by ecology applications in which latent features correspond to discovered species in a sample, we propose a novel class of dependent infinite latent feature models. Our construction starts with a probit IBP and then incorporates a factor analytic hierarchical modeling structure. The proposed approach preserves many appealing properties of the IBP. We study basic theoretical properties of the construction, including motivation for truncation approximations. We additionally develop efficient Markov chain Monte Carlo algorithms for posterior computation. Simulation studies and applications to insect biodiversity data provide support for the new modeling class relative to competitors.

Man Ho Suen

Linearisation approach for aggregated data

In spatial statistics, it is not uncommon to have spatial misalignment in observed responses at point locations and covariates data at various resolutions and shapes. One of the common approaches is to aggregate the point observations into count data with respect to the area polygon. This takes away the point location information and introduces both bias and uncertainty. We start with a Poisson point process and discretise the domain into subspaces. The definition of these subspaces can be flexible based on various scenarios. Assuming the intensity of the process is log-linear, we use an implementation trick and the first order Taylor linearisation in the INLA and inlabru R packages. We compute the approximation bias with the help of the omitted second order terms. This turns out to provide insights into improving the modelling of aggregated data.

Justin Tan

Sampling from varieties in complex projective space

Calculating physically relevant quantities in the low-energy reduction of stringy models requires integration over manifolds embedded in projective space. We present an efficient approach to sampling and Monte Carlo integration on these manifolds for a popular class of low-energy string compactifications.

Konstantinos Tsampourakis

An augmented Gaussian sum filter through a mixture decomposition

Bayesian filtering is an approach to the inference problem for state-space models (SSM), arising in many disciplines of science and engineering. The well known Gaussian filters tackle the problem using Gaussian approximations of the distributions of the hidden state in the model. Gaussian filters however are unable to track multimodal distributions that commonly arise in complex dynamical systems. A variant known as the Gaussian sum filter (GSF) uses Gaussian mixture approximations of the filtering distribution. The GSF however has important limitations, requiring small width of the component covariances for good performance. Moreover, in many SSMs the estimates provided by the GSF blow up, due to covariance inflation. In this paper, we propose a way of controlling the covariances of the underlying Gaussian mixture. Our approach relies on a well known Gaussian identity, which helps us break down each component of the GSF, the parent, into several children components of smaller width. These smaller components are propagated through the nonlinearities by local linearization, which results in a smaller error than in the standard GSF. To reduce the resulting mixture we use resampling. We refer to our novel approach as augmented Gaussian sum filter (AGSF). We demonstrate the advantages of our approach using a toy example, for which the extended Kalman filter (EKF) and GSF perform poorly due to covariance inflation.

Rémi Vaucher

Sampling Signature induced causality chains in a set of time series using Metropolis hastings

Discovering the underlying topological structure describing the statistical dependencies inside a set of time series is a problem that has recently emerged as a n exciting topic in the recent years. In this poster contribution, we address the problem of sampling topological complexes that leverage the expressive power of Signatures, a new feature map invented by Terry Lyons and his collaborators in order to devise a relevant joint distribution. More precisely, following recent work of Giusti and Lee, we use the computed Signatures to understand the causality chains in the time series, that we aggregate into a random complex using sampling proportionally to the causality measure, combined with a noninformative prior enforcing the absence of cycles in the resulting causal directed graph. Signatures of order 2 build graphs and signatures of higher order combine into potentially higher order complexes. Samppling is performed using a simple Metropolis-Hastings algorithm.

Ilina Yozova

Bayesian Causal Forests Combining Randomised And Observational Data For Heterogeneous Treatment Effects Estimation

Bayesian Causal Forest (BCF) of Hahn et al. (2020) was developed to perform nonparametric causal regression model using Bayesian Additive Regression Trees (Chipman et al. (2010)). BCFs are specifically designed to estimate heterogeneous treatment effects using observational data, imposing regularisation directly on treatment effect rather than the response. BCF teases

apart the model into 3 distinct pieces: the prognostic effect, which is the influence of the covariates directly on the response; the treatment effect, and the traditional propensity score, which captures the treatment assignment mechanism. However, when using data from different sources, the treatment assignment mechanism, the prognostic and/or treatment effects, as well as the sets of covariates, may differ greatly between each dataset. A well-known example arises when combining randomised control trial (RCT) data with observational data – RCT have (ideally) balanced propensity and tend to have higher internal validity than observational studies. Therefore, we extend the BCF model to leverage the internal validity of randomised experiments to model heterogeneous treatment effects in observational studies by introducing an additional term in the prognostic effect, which can absorb differences between two two data sources, as well as capture some potential confounding. We implement our methods on a number of examples.