Statistical properties of the solutions to the stochastic Green function

Lucía Bautista Bárcena (University of Edinburgh), Inmaculada Torres Castro, Luis Landesa Porras

The analysis of electromagnetic fields in chaotic closures has been studied with great interest in the last few years. The deterministic results in free space are well-known, so an attempt has been made to develop the stochastic Green function from them, using the Random Matrix Theory, the Berry's hypothesis and the works from Wigner about Gaussian distributions. This allows us, for example, to improve the WiFi signal inside a ship.

Create, Criticise, Compare

Sohan Seth (University of Edinburgh)

Model criticism is usually carried out by assessing if replicated data generated under the fitted model looks similar to the observed data, see e.g. Gelman, Carlin, Stern, and Rubin (2004, p.165). We will present, aggregated posterior check, a method for model criticism for latent variable models that pulls back the data into the space of latent variables, and carries out model criticism in that space. Making use of a model's structure enables a more direct assessment of the assumptions made in the prior and likelihood. We will demonstrate the method with examples of model criticism in latent space applied to Gaussian processes.

Optimising Placement of Pollution Sensors in Windy Environments

Sigrid Passano Hellan (University of Edinburgh), Christopher G Lucas, Nigel H Goddard

Air pollution is one of the most important causes of mortality in the world. Monitoring air pollution is useful to learn more about the link between health and pollutants, and to identify areas for intervention. Such monitoring is expensive, so it is important to place sensors as efficiently as possible. Bayesian optimisation has proven useful in choosing sensor locations, but typically relies on kernel functions that neglect the statistical structure of air pollution, such as the tendency of pollution to propagate in the prevailing wind direction. We describe two new wind-informed kernels and investigate their advantage for the task of actively learning locations of maximum pollution using Bayesian optimisation.

Prediction of heart failure prognosis from gene expression data using CUR matrix decomposition and an ensemble of machine learning algorithms

Mintu Nath (University of Aberdeen), S. Romaine, A.A. Voors, J.A. Timmons, N.J. Samani and on behalf of the BIOSTAT-CHF Consortium

Identifying expression patterns of genes and their association with the survivability of heart failure patients could elucidate novel regulatory mechanisms and functional relevance of mRNAs as well as predict the disease risk. The data included the expression profile of 17904 mRNAs, generated with the Affymetrix Human Transcriptome Array 2.0 (HTA 2.0), from 944 age and sex-matched heart failure patients (318 died and 626 survived during two-year follow-up). The log2-transformed normalised expression data were partitioned into training (n=756, 80%) and validation (n=188, 20%) datasets. We considered 3867 most influential

mRNAs using the CUR matrix decomposition of the training data and employed 12 different machine learning models capturing both linear and non-linear class boundaries. We identified four different models based on performance and correlation between models (AdaBoost classification trees, high dimensional regularised discriminant analysis, elastic net penalised regression and neural networks) and applied the gradient boosting method to develop a prediction model as an ensemble of these four models. We conclude that the final prediction model - developed using CUR matrix decomposition and an ensemble of machine learning algorithms - enhanced computational efficiency, identified genes of biological interests and improved the risk prediction based on patients' gene expression profiles.

Bayesian semiparametric modeling of jointly heteroscedastic extremes

Karla Vianey Palacios Ramirez (University of Edinburgh)

We introduce a Bayesian semiparametric model for learning about the magnitude and frequency of joint extreme values. The joint scedasis function for joint extremes is here devised as a function that carries information on the frequency of joint extremes over time. We develop Bayesian estimators for the two parameters in the model—the joint scedasis function and the coefficient of tail dependence; to learn about the joint scedasis function we resort to finite mixtures of Polya trees. The simulation study shows that the proposed methods are able to recover the true magnitude and frequency of joint extremes in a variety of simulation scenarios. An application of the proposed methodology to the so-called FAANG (Facebook, Apple, Amazon, Netflix and Alphabet's Google) stocks reveals some interesting insights on the dynamics governing their joint extreme losses over time.

Inferential Data Modelling in a Query-Answering System

Thomas Fletcher (University of Edinburgh), Alan Bundy, Kwabena Nuamah

FRANK is a "Third wave of AI" query answering system which performs inferential and statistical reasoning on data from publicly available knowledge bases. Its input parsing, data processing and output capabilities are being extended in order to let it automatically select and apply appropriate statistical methods based on features of queries and retrieved data. To achieve this, FRANK's inference steps are interleaved with an expert system (SMART) whose components include an ontology-based reasoning engine, a varied catalogue of statistical methods and the infrastructure to produce appropriate types of outputs from them (numerical, graphical and descriptive).

Shared Clustering across single-cell RNA sequencing Datasets with the Hierachical Dirichlet Process

Jinlu Liu (University of Edinburgh), Sara Wade and Natalia Bochkina

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 (HDP) process 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. 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 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 Study of Automatic Metrics for the Evaluation of Natural Language Explanations

Miruna Clinciu (University of Edinburgh and Heriot-Watt University), Arash Eshghi and Helen Hastie

As transparency becomes key for Robotics and AI, it will be necessary to evaluate the methods through which transparency is provided, including automatically generated natural language explanations. Here, we explore parallels between the generation of such natural language explanations and the much-studied field of evaluation of Natural Language Generation (NLG), investigating which of the NLG metrics map well to explanations. We present a crowd-sourced corpus of natural language explanations derived from Bayesian Networks (ExBAN). We run correlations comparing human subjective ratings for perceived informativeness and clarity with automatic measures. We find that more recent state-of-the art automatic NLG methods, such as BERTScore and BLEURT, correlate well with these human ratings, compared to other more traditional metrics, such as BLEU and ROUGE.

Functional Covariate DDP with Application to European Debt Crisis

Emmanuel Bernieri (University of Edinburgh) and Miguel de Carvalho

We propose a nonparametric approach combined with functional data analysis to understand the relation between the yield curve of a country and its production index. We apply the method to various european countries with publicly available yield curve and production index on a monthly basis from January 2008 to June 2020. In addition we conduct a Monte Carlo simulations to demonstrate the effectiveness of our methodology. Our method allows us to test the theory that states that a inverted yield curve is the sign of a economical crisis for the country. Our model give us a rich mathematical object -- a conditional density that gives more information that a functional linear regression.

Couplings for Multinomial Hamiltonian Monte Carlo

Kai Xu (University of Edinburgh), Tor Erlend Fjelde, Charles Sutton, Hong Ge

Hamiltonian Monte Carlo (HMC) is a popular sampling method in Bayesian inference. Recently, Heng & Jacob (2019) studied Metropolis HMC with couplings for unbiased Monte Carlo estimation, establishing a generic parallelizable scheme for HMC. However, in practice a different HMC method, multinomial HMC, is considered as the go-to method, e.g. as part of the no-U-turn sampler. In multinomial HMC, proposed states are not limited to end-points as in Metropolis HMC; instead points along the entire trajectory can be proposed. In this paper, we establish couplings for multinomial HMC, based on optimal transport for multinomial sampling in its transition. We prove an upper bound for the meeting time – the

time it takes for the coupled chains to meet – based on the notion of local contractivity. We evaluate our methods using three targets: 1,000 dimensional Gaussians, logistic regression and log-Gaussian Cox point processes. Compared to Heng & Jacob (2019), coupled multinomial HMC generally attains a smaller meeting time, and is more robust to choices of step sizes and trajectory lengths, which allows re-use of existing adaptation methods for HMC. These improvements together paves the way for a wider and more practical use of coupled HMC methods.

Parametric Copula-GP Model for Information-Theoretic Analysis of Neuronal and Behavioral Data

Nina Kudryashova (University of Edinburgh), Theoklitos Amvrosiadis, Nathalie Dupuy, Nathalie Rochefort, Arno Onken

Information-theoretic analysis has been found to be useful in many disciplines. For instance, in systems neuroscience it helps to understand how neuronal populations convey and process information. However, estimating information from a large number of interacting variables is challenging. We propose a parametric copula model which separates the statistics of individual variables from their dependence structure. Such a decomposition of the joint distribution is well suited for mutual information estimation. To account for temporal or spatial changes in dependencies between variables, we model varying copula parameters by means of Gaussian Processes (GP). We improve the flexibility of this method by linearly mixing copula elements with qualitatively different dependencies. We validate the resulting Copula-GP framework on synthetic data and show that the focus on a parametric description of the dependence structure provides more accurate mutual information estimates, compared to other non-parametric methods (KSG, MINE), especially at higher dimensions. We also apply our framework to neuronal and behavioural recordings obtained in awake mice. We demonstrate that our Copula-GP framework allows modeling joint distributions of neuronal and behavioural variables characterized by different statistics and timescales, scales well to higher dimensions (~100) using vine copula constructions, and provides reliable mutual information estimates.

Tail Index Regression-Adjusted Functional Covariate

Anwar Alabdulathem (University of Edinburgh), Miguel de Carvalho

In this poster, I will propose a statistical model the aims to assess how the extreme values of a random variable can change with a functional covariate. The proposed model can be understood as a regression model for heave-tailed response and where the covariate is a random function. To learn about the proposed method a likelihood-based estimator is defined by solving an approximation to a certain variational calculus problem of integral. Simulation data are used to evaluate the performance of the proposed methods. Applications of the proposed methods are envisioned finance, in order to understand how the risk of an extreme loss in the stock market may change according to a certain functional covariate.

Parameter Estimation in Sparse Linear-Gaussian State-Space Models via Reversible Jump Markov Chain Monte Carlo

Benjamin Cox (University of Edinburgh), Victor Elvira

State-space models are ubiquitous for modelling complex systems that evolve over time. In such models, key parameters are usually unknown and must be estimated. In particular, linear systems can be parametrised by the transition matrix that encodes the dependencies among state dimensions. Due to physical and computational constraints, it is crucial to estimate this matrix by promoting sparsity in its components, in such a way that the interactions between dimensions are reduced. In this work, we propose a novel methodology to estimate model parameters and promote sparsity. The method is based on reversible jump Markov chain Monte Carlo and allows the exploration of the space of sparse matrices in an efficient manner by adapting the implicit model dimension. This novel methodology has strong theoretical guarantees and exhibits excellent performance in tricky numerical examples.

Refining Process Descriptions from Execution Data in Hybrid Planning Domain Models

Alan Lindsay (Heriot-Watt University), Santiago Franco Aixela, Rubiya Reba, Thomas L. McCluskey

The creation and maintenance of a domain model is a well recognised bottleneck in the use of automated planning; indeed, ensuring a planning engine is fed with an accurate model of an application is essential in order that generated plans are effective. Engineering domain models using a hybrid representation is particularly challenging as it requires accurately describing continuous processes, which can have complex numeric effects. In this work we consider the problem of the refinement of an engineered hybrid domain model, to capture the effect of the underlying processes more accurately. Our approach exploits the information content of the original model, utilising machine learning techniques to identify important situation and temporal features that indicate a variation in the original effect. We use the problem of modelling traffic flows in an Urban Traffic Management setting as a case study and demonstrate in our evaluation that the refined domain models provide more accurate simulation, which can lead to higher quality plans. The contribution of this work is a general approach to the automated refinement of hybrid planning domain models that reduces the knowledge engineering effort in producing a detailed process model. The approach can be used for refining the domain model during the initial stages of development, or for re-configuring the domain model when used in the same problem area but with a different scenario. We test out the approach within a real-world case study.