Centre for Statistics

Early Career Researchers Day 2024

17th June 2024

THE UNIVERSITY *of* EDINBURGH
Centre for Statistics

Contents

About

CfS Early Career Researches Day 2024

The Centre for Statistics (CfS) Early Career Researchers Day is an annual event which runs as a satellite event of the CfS Annual Conference. The aim is to foster knowledge exchange among young researchers working with data from across the institutions based in Edinburgh. (https://centreforstatistics.maths.ed.ac.uk/)

This conference is organised by PhD students at the School of Mathematics of the University of Edinburgh, led by the Chairperson, Johnny MyungWon Lee.

Organising Committee

Johnny MyungWon Lee	*johnny.myungwon.lee@ed.ac.uk*
Antoni Sieminski	*A.M.Sieminski@sms.ed.ac.uk*
Sergio Gomez Anaya	*S.A.Gomez-Anaya@sms.ed.ac.uk*
James Chok	*james.chok@ed.ac.uk*

Useful Information

Registration, Lunch & Networking will be held at the ground floor of the **Bayes Centre** in the foyer. The **talks** will be held in the seminar room next to the foyer in Bayes **G.03**. No key access will be required to enter this area.

Coffee breaks will similarly be offered at the Bayes Centre foyer.

The **poster session** will be held on level 5 of the **Bayes Centre in room 5.45 & 5.46**. **Key access will be provided** by the Bayes Centre reception.

The open-source LaTeX template, AMCOS_booklet, used to generate this booklet is available at
https://github.com/maximelucas/AMCOS_booklet | Redesigned by James Chok and Johnny MyungWon Lee

Timetable

Monday, 17th June 2024

12:00–13:00	**Registration & Lunch**		
13:00–13:10	**Welcome remarks**		
13:10–13:25	IS	**Dr. Carl Donovan** St Andrews, UK	Transitioning from Academia to Industry: Insights from Statistical Consulting
13:25–13:40	CT	**Alisa Sheinkman** Edinburgh, UK	Variational Bayesian Neural Networks with Shrinkage
13:40–13:55	CT	**Dr. Torben Sell** Edinburgh, UK	Using Sentinel-2 Data and Machine Learning to Map Land Cover in the Pentlands Regional Park, Scotland
13:55–14:25	**Coffee & Photo**		
14:25–14:40	CT	**Lambert De Monte** Edinburgh, UK	Multivariate Radial Pareto distributions: a Geometric Approach to the Statistical Modelling of Multivariate Extremes
14:40–14:55	CT	**Xiangruo Dai** Edinburgh, UK	Higher Education Typologies in Europe: A Mixed Methods Research Design
14:55–15:10	CT	**Huizi Zhang** Edinburgh, UK	Covariate-dependent Hierarchical Dirichlet Process for Cluster Analysis
15:10–15:25	CT	**Zhaoxi Zhang** Edinburgh, UK	The Underlap Coefficient as Measure of a Biomarker's Discriminatory Ability in a Multi-class Disease setting
15:25–17:00	**Poster session**		

CT: Contributed Talk, IS: Invited Speaker.

List of Talks

Monday, 17th June 2024

Transitioning from Academia to Industry: Insights from Statistical Consulting

C. Donovan[1,2]

[1] Honorary Lecturer in Statistics, University of St. Andrews, UK

[2] Director of DMP Statistical Solutions

We are pleased to welcome Dr. Carl Donovan, director of DMP Statistical Solutions, for a compelling talk on his transition from academia to industry. Carl will share his journey from a lecturer at the University of St Andrews to leading a boutique statistical consultancy. He will discuss the nature of the projects undertaken in the industry, the diverse client base, and the collaborative environment with a team of PhD statisticians.

Carl will also reflect on the contrasts between academia and industry, highlighting what he misses and what he doesn't. Additionally, Carl will provide practical advice for aspiring consultants, drawing from his extensive experience working with former PhD students. This talk offers valuable insights for young researchers considering a career in statistical consultancy.

Variational Bayesian Neural Networks with Shrinkage

A. Sheinkman, *University of Edinburgh, UK*

Despite the dominant role of deep models in machine learning, limitations persist, including overconfident predictions, susceptibility to adversarial attacks, and underestimation of variability in predictions. The Bayesian paradigm provides a natural framework to overcome such issues and has become the gold standard for uncertainty estimation with deep models, also providing improved accuracy and tuning of critical hyperparameters.

However, exact Bayesian inference is challenging, typically involving variational algorithms that impose strong independence and distributional assumptions. Moreover, existing methods are sensitive to the architectural choice of the network. We address these issues and construct a relaxed version of the standard feed-forward rectified neural network, employing Polya-Gamma data augmentation tricks to render a conditionally linear and Gaussian model.

Additionally, we use sparsity-promoting priors on the weights of the neural network for data-driven architectural design. To approximate the posterior, we derive a variational inference algorithm that avoids distributional assumptions and independence across layers and is a faster alternative to the usual Markov Chain Monte Carlo schemes.

Using Sentinel-2 data and machine learning to map land cover in the Pentlands Regional Park, Scotland

T. Sell, *Lecturer in Machine Learning, University of Edinburgh, UK*

Understanding land cover is key for landowners who are looking to understand e.g. the biodiversity or the carbon storage potential of their land. In my talk I will explain how the usually exorbitant costs can be reduced by using machine learning and publicly available satellite data. The Pentlands feature as an illustrative example close to home, for which land cover maps are derived for the last 6 years. I will end with a discussion of limitations and future research directions.

Multivariate Radial Pareto distributions: a Geometric Approach to the Statistical Modelling of Multivariate Extremes

L. De Monte, *University of Edinburgh, UK*

Multivariate extreme value theory (EVT) is a branch of probability and statistics concerned with the characterisation of the extremes of random vectors and the estimation of the probability of (joint) rare events. Due to the wide range of possible dependence structures exhibited by random vectors, many EVT frameworks relying on differing underlying assumptions have been proposed. However, most of them suffer from well-known drawbacks such as the impossibility to model positive and negative dependence between variables simultaneously. In this presentation, we develop a flexible framework arising from geometric considerations that addresses many challenges of previously established EVT frameworks. We demonstrate the benefits of our approach on two case studies in which we model 1) the risk of unusually low and high flows at rivers Pang and Windrush (England) and 2) the combinations of wave height, surge, and period leading to sea levels exceeding a dyke in Newlyn (England) at extreme rates.

A new class of multivariate distributions is identified, termed multivariate radial generalised Pareto distributions, and is shown to admit stability properties that permit extrapolation to extremal sets along any "extreme" direction. We show that these distributions arise as non-trivial limit distributions of radially re-normalised exceedances of a multivariate quantile. Using this novel class of multivariate distributions, our statistical models are fully Bayesian and hence allow us to quantify uncertainty in estimation using inference via the posterior distribution.

Higher Education Typologies in Europe: A Mixed Methods Research Design

X. Dai, *University of Edinburgh, UK*

Despite policies enacted across the EHEA to encourage widening participation amongst underrepresented groups, the socioeconomic gaps in higher education participation have remained high. My research design will investigate the impact of policy on outcomes and attitudes in higher education across selected nations using a mixed methods approach. In my presentation, I aim to explain the quantitative side of the approach, including my use longitudinal surveys to answer if educational typologies affect participation from lower class pupils.

Covariate-dependent Hierarchical Dirichlet Process for Cluster Analysis

H. Zhang, *University of Edinburgh, UK*

A recurring and important objective in handling unstructured data is to uncover its inherent structure through clustering observations into groups. We delve into problems related to identifying clusters across multiple datasets when additional covariate information is available. We formulate a novel Bayesian nonparametric approach based on mixture models, integrating ideas from the hierarchical Dirichlet process and single-atoms dependent Dirichlet process.

The proposed method accommodates covariates of various types through the utilization of appropriate kernel functions, exhibiting generality and flexibility. We construct a robust and efficient Markov chain Monte Carlo (MCMC) algorithm involving data augmentation to tackle the intractable normalized weights.

We demonstrate the application of the proposed method to two real-world datasets on single-cell RNA sequencing and calcium imaging, respectively. The versatility of the proposed model enhances our capability to discern the relationship between covariates and clusters.

The Underlap Coefficient as Measure of a Biomarker's Discriminatory Ability in a Multi-class Disease Setting

Z. Zhang, *University of Edinburgh, UK*

In the multi-class setting, the most commonly used measures of a diagnostic biomarker's discriminatory ability are the ROC-based measures, such as the volume under the receiver characteristic surface (VUS) and the Youden index (YI). However, these measures require a stochastic ordering assumption for the distributions of biomarker outcomes in different groups, which is always not plausible, particularly when covariates are involved.

To address this issue, we propose the underlap coefficient, a new summary index of a biomarker's diagnostic capacity, study its properties, as well as its relationship with the VUS and YI when a stochastic order is enforced in the three-class setting. We further propose Bayesian nonparametric estimators for both the unconditional underlap coefficient and for its covariate-specific counterpart. We illustrate the proposed approach through an application to an Alzheimer's disease (AD) dataset aimed to assess how four potential AD biomarkers, distinguish between individuals across different disease stages.

List of Posters

Monday, 17th June 2024

Probabilistic Scenarios for Wind Energy Generation - S. Gomez-Anaya, *University of Edinburgh, UK*

To effectively integrate renewable energy sources into the grid system, accurate power generation forecasts are necessary. Understanding the uncertainty around these estimates is crucial for measuring system reliability, planning for extreme scenarios, and optimising daily transmission and operation of the grid.

Limited access to detailed private datasets, aggregated data reports, Numerical Weather Predictions (NWP), and historical recreations of weather conditions result in a heterogeneous and sometimes conflicting array of information. This necessitates techniques optimised for handling large volumes of data.

In this work, modern statistical techniques are leveraged to provide probabilistic scenarios for wind power. By comparing the advantages and disadvantages of these techniques, I aim to offer insights into the benefits and limitations of current methods.

Mapping Weather to Electricity Demand for Forward-Looking Risk Calculations - A. Bhattacharya, *University of Edinburgh, UK*

This work presents a novel method for determining risk metrics like LOLE (Loss of Load Expectation), which measures the hours when electricity demand exceeds supply, particularly due to winter weather pattern shifts in the UK.

The approach involves mapping historical weather to daily peak demand using a new demand formula, scaling this historical demand to a target year by incorporating temperature sensitivity and year effects. This adjusted demand is then used to determine LOLE using future renewable energy scenarios. The method aims to assess the uncertainty in demand and renewable generation during peak winter hours.

Dependent Mixture Models for Extremes - V. Carcaiso, *University of Padova, Italy*

In the block maxima approach for extreme value analysis, it is commonly assumed that maxima are extracted from large samples of a stationary process. However, this assumption may not hold in many applications, such as analysing annual rainfall maxima influenced by different weather regimes.

To address this, we employ finite mixture models, specifically two-component mixtures of Gumbel distributions. Observations are labelled based on the generating physical process, but this information may be unavailable or unreliable. Our proposed model probabilistically allocates data points to mixture components using labels and additional variables, rather than deterministic allocation. We use a Bayesian hierarchical approach to facilitate borrowing information between groups and to allow for direct quantification of uncertainty in component allocation.

History Matching as a Calibration Method for Carbon Cycle Models - N. Fischer, *University of Edinburgh, UK*

Complex physical models implemented in computer code are fundamental for assessing ecosystem dynamics. Coupled with observations, these models can infer unobserved ecosystem properties. For example, the carbon cycle model DALEC Crop models the effect of Nitrogen fertilisation on wheat growth. However, its current calibration framework, CARDAMOM, has limitations. It is computationally expensive and does not allow for exploring the effects of structural discrepancy, thereby limiting our understanding of the crop field's physical processes.

This work introduces history matching combined with emulation as an alternative calibration method. History Matching identifies the set of parameter combinations that produce acceptable matches to observations. We use this input space in a DALEC Crop forward run to generate yield predictions, bypassing CARDAMOM. Finally, we compare the outputs of CARDAMOM and history matching calibrations.

The Structural Variation Landscape and its Role in Trait Architecture in the Genome of European Seabass (Dicentrarchus Labrax) - Z. Jiao, *University of Edinburgh, UK*

SVs are typically defined as genetic polymorphisms that affect $>$ 50bp of sequence. While SVs are an important source of genetic variation and an important cause of inter-individual differences, they have been neglected in genetics studies compared with SNPs.

Here, we defined the SV landscape including 21,428 high-confidence SVs in European seabass (Dicentrarchus labrax), a high value European aquaculture species with 90 animals. These SVs were annotated to estimate potential effects on genes. We imputed the SVs for 990 fish with phenotype data for viral nervous necrosis (VNN), one of the main infectious diseases in European seabass, allowing a GWAS analysis using the SVs, with 108 SVs in a single QTL region. The results will improve our understanding of the role of SVs in genetic architecture of traits relevant to aquaculture.

A Bayesian Lasso for Tail Index Regression - J. MW. Lee, *University of Edinburgh, UK*

Extreme events can be better comprehended through the lens of regression models tailored for extreme values. Our methodological contribution involves leveraging Bayesian regularization and generalized additive framework for tail index regression, thereby enabling a more flexible model for analyzing extreme values. This framework revolves around a conditional Pareto-type specification, enriched by the inclusion of Bayesian Lasso-type shrinkage priors and further refined through low-rank thin plate splines basis expansion.

The performance of the proposed method is then validated through a simulation study that recovers the true covariate-adjusted tail index, $\alpha(\boldsymbol{x})$ over a variety of scenarios along while regularizing the covariates. We illustrate our model to investigate extreme wildfire events in Portugal, delving into the key drivers behind these occurrences.

Taming the Interacting Particle Langevin Algorithm - the Superlinear Case - N. Makras, *University of Edinburgh, UK*

Recent advances in stochastic optimization have yielded the interacting particle Langevin algorithm (IPLA), which leverages the notion of interacting particle systems (IPS) to efficiently sample from approximate posterior densities. This becomes particularly crucial within the framework of Expectation-Maximization (EM), where the E-step is computationally challenging or even intractable.

Although prior research has focused on scenarios involving convex cases with gradients of log densities that grow at most linearly, our work extends this framework to include polynomial growth. Taming techniques are employed to produce an explicit discretization scheme that yields a new class of stable, under such non-linearities, algorithms which are called tamed interacting particle Langevin algorithms (tIPLA). We obtain non-asymptotic convergence error estimates in Wasserstein-2 distance for the new class under an optimal rate.

Score-Based Denoising Diffusion Models for Photon-Starved Image Restoration Problems - S. Melidonis, *Heriot-Watt University, UK*

Score-based denoising diffusion models have recently emerged as a powerful strategy to solve image restoration problems. Modern restoration approaches combine a data fidelity term and a pretrained diffusion model, which is used as an implicit prior in a Plug-and-Play (PnP) manner.

With extreme computer vision applications in mind, this paper presents the first PnP denoising diffusion method for photon-starved imaging problems. These problems involve highly challenging noise statistics, such as binomial, geometric, and low-intensity Poisson noise statistics, which are difficult because of high uncertainty about the solution and because the models exhibit poor regularity properties. The proposed method is demonstrated on a series of challenging photon-starved imaging experiments, where it delivers remarkably accurate solutions and outperforms alternative strategies from the state-of-the-art.

Partner Institutions

The Centre for Statistics (CfS) Early Career Researchers Day is a part of the CfS Annual Conference.

The Centre for Statistics (CfS) Early Career Researchers Day is affiliated with the University of Edinburgh Centre for Statistics, the University of Edinburgh School of Mathematics, and the Maxwell Institute for Mathematical Sciences.

We also acknowledge and thank Bayes Centre for offering the seminar room for our event.

THE UNIVERSITY *of* EDINBURGH
Centre for Statistics

THE UNIVERSITY *of* EDINBURGH
School of Mathematics

MAXWELL INSTITUTE FOR
MATHEMATICAL SCIENCES

BayesCentre