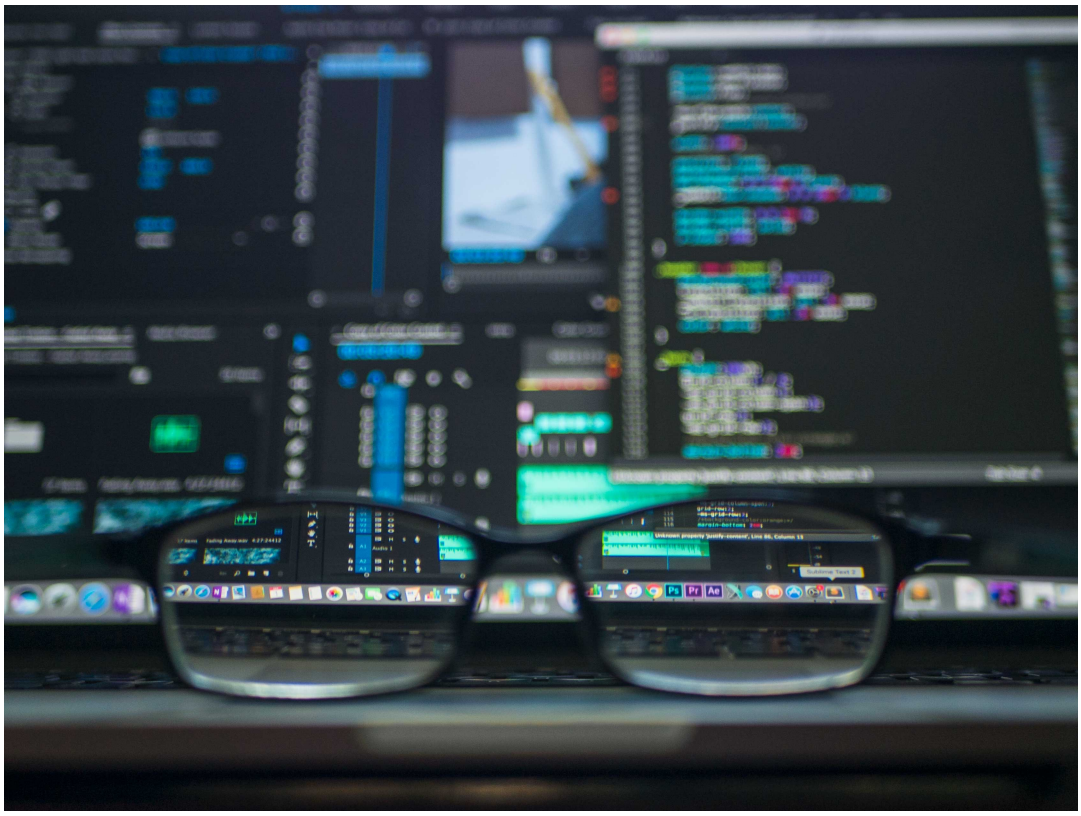
THE UNIVERSITY *of* EDINBURGH
Centre for Statistics

CFS Annual Conference 2024

June 18, 2024
Nucleus Building, Elm Lecture Theatre
University of Edinburgh, King's Buildings, EH9 3FD



Hosted by the **Centre for Statistics**
THE UNIVERSITY *of* EDINBURGH
Edinburgh, UK
https://centreforstatistics.maths.ed.ac.uk

Organized by Sara Wade (sara.wade@ed.ac.uk), Timothy Cannings (timothy.cannings@ed.ac.uk), Grégoire Clarte (gclarte@ed.ac.uk), and Torben Sell (torben.sell@ed.ac.uk).

CFS Annual Conference 2024

Our **Annual Conference** showcases the interdisciplinary data-driven research being carried out in and around Edinburgh. The one-day event features a number of **invited speakers** from academia and industry, a **poster session** and plenty of time for **networking**. The conference is preceded by our **Early Career Researchers Day**, which provides PhD students and other ECRs the opportunity to get together and discuss their research in a less formal setting.

Programme:

- 09:00-09:25: Registration with coffee/tea.
- 09:25-09:30: Housekeeping and Welcome (Timothy Cannings, Director of Centre for Statistics).
- Session 1: Chaired by Andrej Svetlošák (School of Mathematics)
  - 09:30-10:05: Roger Halliday (Research Data Scotland) *Unlocking data for research in the public good*
  - 10:05-10:40: Elliot Crowley (School of Engineering) *Finding the right neural network for the job*
- 10:40-11:10: Coffee.
- Session 2: Chaired by Glenna Nightingale (School of Health in Social Science)
  - 11:10-11:45: Cecilia Balocchi (School of Mathematics) *Improving uncertainty quantification in Bayesian cluster analysis*
  - 11:45-12:20: Edward Wallace (School of Biological Sciences) *Quantifying the messages that make cells tick: adventures in understanding regulation of messenger RNA in fungi*
- 12:20-14:20: Posters and Lunch.
- Session 3: Chaired by Takuo Matsubara (School of Mathematics)
  - 14:20-14:55: Xuebin Zhao (School of Geosciences) *Variational Prior Replacement in Bayesian Inference and Inversion*
  - 14:55-15:30: Roxanne Connelly (School of Social and Political Science) *Do You Like School? Social Class, Gender, Ethnicity and Pupils' Educational Enjoyment*
  - 15:30-16:05: Mirella Lapata (School of Informatics) *Prompting is *not* all you need! Or why Structure and Representations still matter in NLP*
- 16:05-16:10: Closing (Timothy Cannings, Director of Centre for Statistics).
- 16:10-17:00: Reception and Photo.

Invited talks are 25 min + 10 min Q&A and changeover. Further information can be found on the CfS 2024 website:

https://centreforstatistics.maths.ed.ac.uk/cfs/events/upcoming-events/cfs-annual-conference-2024

CfS Annual Conference 2024: Invited Talks

Invited talks feature speakers from six Schools and Institutes at the **University of Edinburgh**, including the School of Engineering, School of Mathematics, School of Biological Sciences, School of Geosciences, School of Social and Political Sciences, and School of Informatics, as well as an industry speaker from **Research Data Scotland**.

Unlocking data for research in the public good

Roger Halliday (roger.halliday@researchdata.scot)
Research Data Scotland

Abstract: Public policy challenges are complex and require many organisations to collaborate to solve them. To understand root causes and know whether we're taking the right actions, we need to know what is happening for a person, their household, a business or a place - not just getting a partial picture of what is happening by looking at data from a one public service at a time. Connecting data from across public services isn't straightforward, but is happening. This talk will outline the opportunities data linkage brings, what this means for the research community, and how we're making it happen in Scotland whilst keeping people's data safe.

†**Time slot**: 09:30-10:05

Finding the right neural network for the job

Elliot Crowley (elliot.j.crowley@ed.ac.uk)
School of Engineering, University of Edinburgh

Abstract: In machine learning, neural network architectures are important, with well-designed convolutional architectures powering advances in computer vision in the 2010s, and the transformer architecture forming the foundation of large language models like ChatGPT. However, given a new task, a manually designed architecture may be far from optimal. We can instead turn to neural architecture search (NAS) algorithms to automate this design. A NAS algorithm consists of a search space—the space of all possible architectures that can be chosen—and a search algorithm for navigating this space. In this talk, I will explain how NAS works at a high level, before stressing how important the search space is for success. I will then introduce recent work where we have designed an expressive NAS search space that contains many existing competitive architectures, and provides flexibility for discovering new ones on novel tasks.

†**Time slot**: 10:05-10:40

Improving uncertainty quantification in Bayesian cluster analysis

Cecilia Balocchi (cecilia.balocchi@ed.ac.uk)
School of Mathematics, University of Edinburgh

Abstract: The Bayesian approach to clustering is often appreciated for its ability to provide uncertainty in the partition structure. However, summarizing the posterior distribution over the clustering structure can be challenging. Wade and Ghahramani (2018) proposed to summarize the posterior samples using a single optimal clustering estimate, which minimizes the expected posterior Variation of Information (VI). In instances where the posterior distribution is multimodal, it can be beneficial to summarize the posterior samples using multiple clustering estimates, each corresponding to a different part of the space of partitions that receives substantial posterior mass. In this work, we propose to find such clustering estimates by approximating the posterior distribution in a VI-based Wasserstein distance sense. An interesting byproduct is that this problem can be seen as using the k-mediods algorithm to divide the posterior samples into different groups, each represented by one of the clustering estimates. Using both synthetic and real datasets, we show that our proposal helps to improve the understanding of uncertainty, particularly when the data clusters are not well separated, or when the employed model is misspecified.

†**Time slot**: 11:10-11:45

Quantifying the messages that make cells tick: adventures in understanding regulation of messenger RNA in fungi

Edward Wallace (edward.wallace@ed.ac.uk)
School of Mathematics, University of Edinburgh

Abstract: All living cells rely on messenger RNA, both to encode protein machines and structures, and also to control the precise time and location where those structures are made in the cells. My group studies how mRNA regulation allows fungal cells to grow and adapt to changing environments, with a focus on control of how the production of protein from mRNA templates is controlled. We've learned about how RNA and proteins stick together to detect heat stress, and how fungal cell walls are regulated by a specific RNA-binding protein factor. I'll give an overview of our work and of the statistical challenges involved in RNA-seq analysis, especially for non-standard experimental designs. RNA-seq involves chopping RNA molecules up into small fragments, enriching for fragments of biological interest, amplifying the fragments, sampling millions of these, and then re-assembling this into a quantitative picture of what happened in the cells.

†**Time slot**: 11:45-12:20

Variational Prior Replacement in Bayesian Inference and Inversion

Xuebin Zhao (Xuebin.Zhao@ed.ac.uk)
School of Geosciences, University of Edinburgh

Abstract: Many scientific investigations require that the values of a set of model parameters are estimated using recorded data – a process referred to as inference. In Bayesian inference, information from both observed data and prior knowledge that existed independently, is combined to update model parameters probabilistically. The result is represented by the so-called posterior probability distribution function. This describes the probability (density) of each possible set of parameter values according to their consistency with the data and prior information. Prior information is often described by a prior probability distribution. This represents our belief about the range of values that the variables can take, and their relative probabilities when considered independently of the current recorded data. Situations arise in which we wish to change prior information: (i) the subjective nature of prior information (your view about the results of previous studies may differ from ours), (ii) cases in which we wish to test different states of prior information as independent hypothesis tests, and (iii) information from new, independent studies may emerge during the course of an investigation so prior information may evolve over time. Estimating the solution to any single inference problem is usually computationally costly, as it typically requires thousands or millions of model samples to be drawn and simulation of the observed data set that would have been recorded if each sample was true. Therefore, recalculating the Bayesian inference solution every time prior information changes can be extremely expensive.

We develop a mathematical formulation that allows prior information to be changed in a solution using variational methods, without performing Bayesian inference on each occasion. In this method, existing prior information is removed from a previously obtained posterior distribution and is replaced by new prior information. We therefore call the methodologyvariational prior replacement (VPR). We demonstrate VPR using a 2D seismic full waveform inversion example, where VPR provides almost identical posterior solutions compared to those obtained by solving independent inference problems using different prior distributions. The former can be completed within minutes even on a laptop whereas the latter requires days of computations using high-performance computing resources. We demonstrate the value of the method by comparing the posterior solutions obtained using three different types of prior information: Uniform, smoothing and geological prior distributions.

†**Time slot**: 14:20-14:55

Do You Like School? Social Class, Gender, Ethnicity and Pupils' Educational Enjoyment

Roxanne Connelly (roxanne.connelly@ed.ac.uk)
School of Social and Political Science, University of Edinburgh

Abstract: Foundational theories in the sociology of education consistently suggest that the enjoyment of education is stratified

by social class, gender, and ethnicity. Analysing data from the UK Millennium Cohort Study we provide a detailed empirical analysis of the social stratification of educational enjoyment in a nationally representative sample. Our results challenge orthodox sociological views on the relationship between structural inequalities and educational enjoyment, and therefore question the existing theoretical understanding of the wider role of enjoyment in education.

†**Time slot**: 14:55-15:30

Prompting is *not* all you need! Or why Structure and Representations still matter in NLP

Mirella Lapata (mlap@inf.ed.ac.uk)

School of Informatics, University of Edinburgh

Abstract: Recent years have witnessed the rise of increasingly larger and more sophisticated language models (LMs) capable of performing every task imaginable, sometimes at (super)human level. In this talk, I will argue that there is still space for specialist models in today's NLP landscape. Such models can be dramatically more efficient, inclusive, and explainable. I will focus on examples from the domain of summarization and will illustrate the benefit of learning task-specific representations. I will show how such representations can be further structured to allow search and retrieval, and evidence-based generation. I will also discuss why we need to use LLMs for what they are good at and remove the need for them to do things that can be done much better by smaller models.

†**Time slot**: 15:30-16:05

Contributed posters feature presenters from the **University of Edinburgh** across Schools and Institutes, including e. g. the School of Mathematics, School of Geosciences, and the Roslin Institute. In total, **19** contributed posters will be presented across the different disciplines.

Towards a deep learning approach for data-driven short-term spatiotemporal earthquake forecasting

Foteini Dervisi (F.Dervisi@sms.ed.ac.uk)
School of GeoSciences, University of Edinburgh

Abstract: Recent developments in earthquake monitoring have led to an increase in the volume of available earthquake catalogue data, which indicates that machine learning techniques might be well-suited to uncover patterns within earthquake sequences. This work focuses on using neural networks to produce localised short-term seismicity forecasts after the occurrence of large magnitude events. We assemble a dataset of earthquake catalogues from diverse tectonic regions and build spatiotemporal sequences around large magnitude events. These sequences are then used as input to a neural network, whose aim is to produce next-day forecasts of seismicity. Our results suggest that machine learning is a promising alternative to disciplinary statistics and physics-based earthquake forecasting approaches.

Variational Bayesian Neural Networks with Shrinkage

Alisa Sheinkman (A.Sheinkman@sms.ed.ac.uk)
School of Mathematics, University of Edinburgh

Abstract: 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.

The underlap coefficient as measure of a biomarker's discriminatory ability in a multi-class disease setting

Zhaoxi Zhang (Z.Zhang-156@sms.ed.ac.uk)
School of Mathematics, University of Edinburgh

Abstract: The first step when evaluating a potential diagnostic biomarker is to determine the variation in its values across different disease groups. In the three-class disease setting, the volume under the receiver characteristic surface (VUS) and the three-class Youden index (YI) are the commonly used summary measures of a biomarker's discriminatory ability. However, these measures are only appropriate under a stochastic ordering assumption for the distributions of biomarker outcomes in the three groups. This assumption is stringent and not always plausible, particularly when covariates are involved. Violating this assumption may lead to incorrect conclusions about a biomarker's performance to distinguish between the three disease classes. To address this issue, we propose the underlap coefficient, a new summary index of a biomarker's capacity to distinguish between multiple disease groups, 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. A simulation study reveals a good performance of the proposed estimators across a range of conceivable scenarios. We illustrate the proposed approach through an application to an Alzheimer's disease (AD) dataset aimed to assess how four potential AD biomarkers, two of which not exhibiting a stochastic order, distinguish between individuals with normal cognition, mild impairment, and dementia, and how and if age and gender impact this discriminatory ability.

ASA DataFest 2024 at Edinburgh

Serveh Sharifi (Serveh.Sharifi@ed.ac.uk)

School of Mathematics, University of Edinburgh

Abstract: The American Statistical Association (ASA) DataFest is a celebration of data in which teams of undergraduate students work on a large, complex, and surprise dataset over a weekend to find and share meaning and insights. The dataset and the work are likely beyond the scope of what students see in their courses. Founded at UCLA in 2011, ASA DataFest has experienced rapid growth over the years. It is now hosted by many of the USA's most prestigious colleges and universities, as well as several renowned foreign institutions. This friendly competition provides students with invaluable real-world expression, an opportunity to showcase their skills and explore a data scientist's job, and a platform to network with professionals and peers. At the local run of this event in Edinburgh in March 2024, all UG students from the University of Edinburgh and Heriot-Watt University, with an interest in data, were invited to join. During the competition, academic staff, PhD students, and data scientists from industry guided students in their work. The event ended with brief presentations of the teams' work and assessing the students' ability to communicate results clearly and effectively. Teams were awarded in various categories, such as "Best Insights", "Best Visualisation", and "Best Use of Outside Data". This poster summarises our experience in running this event and how it can benefit students learning in data science.

Mapping Shifts in Winter Weather to Energy Demand for Forward-Looking Risk Assessments

Aninda Bhattacharya (A.Bhattacharya-9@sms.ed.ac.uk)

School of Mathematics, University of Edinburgh

Abstract:

This work introduces a new method to determine risk metrics such as LOLE (Loss of Load Expectation) to measure the number of hours when electricity demand exceeds the available supply in the system, arising from shifts in weather patterns during a winter season in the UK. The proposed method involves first mapping historic weather to daily peak demand using the newly derived demand formula. The historic peak demand series is then scaled to a future year by adding a temperature sensitivity and year effect to reflect the expected changes over time. Once the demand is mapped to a future year, it is then used to determine LOLE using future scenarios of renewables. The goal is to examine the uncertainty on demand and available renewable generation arising from different weather systems during the peak hours over UK. The reliability of the method is then validated against the previously established demand rescaling methodology in literature. The comparison ensures that the proposed method provides a reliable risk assessment under uncertainty due to changing weather, taking into account future renewable energy projections (offshore and onshore wind generation scenarios during winter).

Covariate-dependent hierarchical Dirichlet process

Huizi Zhang (H.Zhang-144@sms.ed.ac.uk)

School of Mathematics, University of Edinburgh

Abstract: 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.

Linearization approach for aggregated data

Man Ho Suen (M.H.Suen@sms.ed.ac.uk)

School of Mathematics, University of Edinburgh

Abstract: 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. One of the popular approaches in landslide literature is to aggregate based on slope units that cluster landslide observations beneath the surface. This takes away the point location information and introduces both bias and uncertainty. Assuming the intensity of the process is log-linear, an implementation trick is used and the first-order Taylor linearization in the INLA and inlabru R packages. The approximation bias is computed with the help of the omitted second-order terms. This turns out to provide insights into improving the modelling of aggregated data.

Shotgun sequencing read duplication parameters estimated from k-mer spectra

Hannes Becher (H.Becher@ed.ac.uk)

The Roslin Institute, University of Edinburgh

Abstract: Read duplication and mate overlap are common issues with modern shotgun DNA sequencing datasets. A form of pseudo replication, not accounting for duplication or overlap can lead to overconfident genetic variant calling. Mapping based assessments of read duplication is computationally expensive. Pseudo replication in sequencing data leads to a characteristic pattern of the k-mer spectrum, causing overdispersion compared to the expected Poisson distribution. Here, I report the relationship between the level of sequence (read) duplication and the shape of the associated k-mer spectrum, building on simulated and real world data. This work demonstrates how read duplication can be assessed efficiently in an assembly free fashion.

A Bayesian Lasso for Tail Index Regression

Johnny Lee (johnny.myungwon.lee@ed.ac.uk)

School of Mathematics, University of Edinburgh

Abstract: Extreme events can be better comprehended through the lens of regression models tailored for extreme values. Our methodological contribution involves leveraging Bayesian regularisation and generalised additive framework for tail index regression, thereby enabling a more flexible model for analysing 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, $\alpha(x)$ 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.

Unlocking administrative data linkage in Scotland for research

Cecilia MacIntyre (Cecilia.MacIntyre@gov.scot)

Data for Research Unit, Statistics and Data Access Division, Scottish Government

Extending the R number by applying hyperparameters of Log Gaussian Cox process models in an epidemiological context to provide insights into COVID-19 positivity in the City of Edinburgh and in students residing at Edinburgh University

Glenna Nightingale (Glenna.Nightingale@ed.ac.uk)

School of Health in Social Science, University of Edinburgh

Abstract: The impact of the COVID-19 pandemic on University students has been a topic of fiery debate and of public health research. This study demonstrates the use of a combination of spatiotemporal epidemiological models to describe the trends in COVID-19 positive cases on spatial, temporal and spatiotemporal scales. In addition, this study proposes new epidemiological metrics to describe the connectivity between observed positivity; an analogous metric to the R number in conventional epidemiology. The proposed indices, Rspatial, Rspatiotemporal and Rscaling will aim to improve the characterisation of the spread of infectious disease beyond that of the COVID-19 framework and as a result inform relevant public health policy. Apart from demonstrating the application of the novel epidemiological indices, the key findings in this study are: firstly, there were some Intermediate Zones in Edinburgh with noticeably high levels of COVID-19 positivity, and that the first outbreak during the study period was observed in Dalry and Fountainbridge. Secondly, the estimation of the distance over which the COVID-19 counts at the halls of residence are spatially correlated (or related to each other) was found to be 0.19km (0.13km to 0.27km) and is denoted by the index, Rspatial. This estimate is useful for public health policy in this setting, especially with contact tracing. Thirdly, the study indicates that the association between the surrounding community level of COVID-19 positivity (Intermediate Zones in Edinburgh) and that of the University of Edinburgh's halls of residence was not statistically significant. Fourthly, this study reveals that relatively high levels of COVID-19 positivity were observed for halls for which higher COVID-19 fines were issued (Spearman's correlation coefficient = 0.34), and separately, for halls which were non-ensuite relatively to those which were not (Spearman's correlation coefficient = 0.16). Finally, Intermediate Zones with the highest positivity were associated with student residences that experienced relatively high COVID-19 positivity (Spearman's correlation coefficient = 0.27).

Author list: Megan Ruth Laxton , Glenna Nightingale, Finn Lindgren , Arjuna Sivakumaran, Richard Othieno

Statistical inference for extremes of stochastic processes using a novel geometric approach

Xindi Song (X.Song-23@sms.ed.ac.uk)

School of Mathematics, University of Edinburgh

Abstract: Accurately estimating the frequency of future environmental events that surpass historical records is vital for risk assessment and prevention strategies. Extreme value theory provides mathematically justified models as the basis for extrapolations from observed extreme events out to more extreme ones. In the context of environmental processes, events are typically infinite-dimensional, meaning that the domain on which the process takes values is a continuum. Motivated by a novel geometrical framework, we employ a novel definition for characterizing extreme events of stochastic processes. The definition is based on converting a process to its magnitude, which is encoded by the supremum of the process, and its direction, which is encoded by the process scaled by its magnitude. Based on this decomposition, we develop functional extreme value models for the magnitude of the process, conditionally on the overall shape characteristics of the process. The statistical models that we build are generative and are implemented in a fully Bayesian manner, allowing extrapolation along any direction in function space, thereby permitting computation of posterior predictive probabilities of any rare event such the event that a future realization s stays abnormally large or low for some amount of time or over some region in space. In this poster, we showcase our results for extremes of temporal stochastic processes, but present modelling strategies that can be implemented to more general setups, such as to extremes of spatial and spatio-temporal processes.

Wind Energy Probabilistic Scenarios

Sergio Gomez Anaya (S.A.Gomez-Anaya@sms.ed.ac.uk)

School of Mathematics, University of Edinburgh

Abstract: To effectively integrate renewable sources into the energy grid system, accurate power generation forecasts are necessary. Additionally, understanding the uncertainty around these estimates is crucial for measuring system reliability, planning for extreme scenarios, and optimizing daily transmission and operation of the grid. Limited access to private datasets means that on-site wind speed information or wind power data is often unavailable. Typically, public information is reported at an aggregated level. Numerical Weather Prediction (NWP) scenarios are often used to understand weather conditions in specific regions. The high frequency of new spatio-temporal historical information, along with historical recreations of weather

conditions, necessitates techniques optimised for handling large volumes of data. In this work, modern statistical techniques are leveraged to consistently provide probabilistic scenarios for wind power. By comparing the advantages and disadvantages of these techniques, I aim to offer insights on which method works best given the vast amount of information and the required outputs.

Dependent mixture models for extremes

Viviana Carcaiso (V.Carcaiso@sms.ed.ac.uk)
School of Mathematics, University of Edinburgh
Department of Statistical Sciences, University of Padova

Abstract: 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.

A hierarchical Bayesian mixture approach for modelling neuronal connectivity patterns from MAPseq Data

Edward Agboraw (s1605280@sms.ed.ac.uk)
Centre for Discovery Brain Sciences, University of Edinburgh

Abstract: Structural motif analysis seeks to better understand brain connectivity by interrogating the long-range axonal projections that link distant brain regions at the single-cell resolution. This approach is enabled by MAPseq, a novel experimental methodology which labels cellular projections with viral barcodes. This recontextualizes single-neuron tracing as a problem of sequencing, overcoming the throughput limitations of prior optical techniques and capitalizing on the swift advancement of modern sequencing technology.

Current MAPseq-based long range projection motif analysis methods rely on the use of standard algorithmic clustering methods. These approaches require significant transformations of the data and are not based on formal models of neural projection, limiting their biological interpretability.

An alternative method overcomes these issues by modelling projection patterns via the Binomial Distribution. This method allows for proper hypothesis testing and biological interpretations but is based on a highly simplified model of neural projection which removes much of the information contained in a typical MAPseq dataset.

Here we introduce a new method for MAPseq motif analysis, utilizing a novel Hierarchical Bayesian Mixture Model of neural projection based on the Dirichlet-Multinomial. This approach models neural projection directly, accommodating features of MAPseq data ignored by the Binomial Model, such as projection strength. It also does not require any transformations of the data, simplifying the biological interpretation of the results and better enabling group comparisons.

The utility of this model is demonstrated here on a MAPseq dataset describing long-range projections from the Entorhinal Cortex to specific target regions in the neocortex.

Taming the Interacting Particle Langevin Algorithm – the superlinear case

Nikos Makras (N.Makras@sms.ed.ac.uk)
School of Mathematics, University of Edinburgh

Abstract: 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.

Multivariate radial Pareto distributions: a geometric approach to the statistical modelling of multivariate extremes

Lambert De Monte (l.demonte@ed.ac.uk)

School of Mathematics, University of Edinburgh

Abstract: 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. Joint work with Ioannis Papastathopoulos, Ryan Campbell, Håvard Rue.

Focusing on Gene Co-Expression Variance and Network Topology for Better Understanding of Cell Fate Decisions

Thanakorn Jaemthaworn (T.Jaemthaworn@sms.ed.ac.uk)

Centre for Regenerative Medicine, Institute for Regeneration and Repair, University of Edinburgh

Abstract: Limited understanding of cell fate decisions remains a major challenge in stem cell technology. Current single-cell RNA sequencing (scRNA-Seq) focuses on analysing individual gene expression to understand cell differentiation. To improve scRNA-Seq analysis, we propose applying the information encoded in the variance of gene co-expression, rather than relying solely on the variance of individual gene expression. We introduce a novel feature selection method called High Variance of Correlated Gene (HvCG), which demonstrates superior performance compared to the traditional highly variable gene (HVG) method. Furthermore, we present the topological analysis of Cell Type-Specific Gene Co-expression Networks (CTS-GCNs) constructed from zebrafish melanocyte differentiation scRNA-Seq data. As a result, the CTS-GCNs topology can reveal a deeper understanding of cell transition dynamics during cell differentiation.

A Flexible School and College Level Qualification in Data Science

Kate Farrell (kate.farrell@ed.ac.uk)

Moray House School of Education and Sport, Institute for Education, Community & Society, University of Edinburgh

Abstract: This poster abstract describes the design and development of a high school-level qualification in data science. The qualification has been available for 4 years; 1982 learners have completed the course to date across 30 educational institutions.

We describe the structure of the course and its pedagogical principles. We conclude with a set of recommendations for other educators who are designing similar qualifications in other school systems.

CFS Annual Conference 2024: Participant List

Name	Department	Institution	Email
Alan O'Callaghan	IGC	CGEM	alan.ocallaghan@ed.ac.u
Alisa Sheinkman	School of Math	UoE	a.sheinkman@sms.ed.ac.
Amanda Lenzi	School of Mathematics	University of Edinburgh	lenzi.amanda88@gmail.c
Amy Wilson	Mathematics	University of Edinburgh	Amy.L.Wilson@ed.ac.uk
Andrej Svetlosak	School of Mathematics	The University of Edinburgh	andrej.svetlosak@ed.ac.u
Andrew Ricketts	Mathematics	University of Edinburgh	andrew.g.ricketts@gmail
Aninda	School of Mathematics	University of Edinburgh	s2601509@ed.ac.uk
Antoni Sieminski	Centre for Statistics	University of Edinburgh	s2410784@ed.ac.uk
Benjamin Cox	School of Mathematics	University of Edinburgh	benjamin.cox@ed.ac.uk
Borja de Pedro Sarasola	School of Social and Political Sciences	Edinburgh University	s2240520@ed.ac.uk
Bruce Worton	School of Mathematics	University of Edinburgh	Bruce.Worton@ed.ac.uk
Cecilia Balocchi	School of Mathematics	University of Edinburgh	cecilia.balocchi@ed.ac.uk
cecilia macintyre	ADR Scotland	scottish government	cecilia.macintyre@gov.sc
Chris Dent	School of Mathematics	University of Edinburgh	chris.dent@ed.ac.uk
Christopher Wretman	School of Social and Political Science	The University of Edinburgh	christopher.wretman@ed
Clara Panchaud	School of Mathematics	University of Edinburgh	s2239964@ed.ac.uk
Colin Aitken	Mathematics	University of Edinburgh	Cgga@ed.ac.uk
Daisy Bao	Moray Housr	The University of Edinburgh	S1932050@ed.ac.uk
Damian Clancy	Actuarial Maths & Statistics	Heriot-Watt University	d.clancy@hw.ac.uk
Edward Agboraw	Centre for Discovery Brain Sciences	School of Mathematics	s1605280@ed.ac.uk
Elliot Crowley	School of Engineering	University of Edinburgh	elliot.j.crowley@ed.ac.uk
Finn Lindgren	School of Mathematics	The University of Edinburgh	Finn.Lindgren@ed.ac.uk
Foteini Dervisi	School of GeoSciences	University of Edinburgh/British Geological Survey	s2323222@ed.ac.uk
Galina Andreeva	Business School	University of Edinburgh	Galina.Andreeva@ed.ac.
gillian raab	geosciences/SCADR	univerdity og edinburgh	gillian.raab@ed.ac.uk
Glenna Nightingale	School of Health in Social Science	University of Edinburgh	Glenna.Nightingale@ed.a
Grégoire Clarté	School of Maths	University of Edinburgh	gclarte@ed.ac.uk
Hannes Becher	The Roslin Institute	University of Edinburgh	h.becher@ed.ac.uk
Harris Abdul Majid	School of Engineering	University of Edinburgh	H.abdulmajid@ed.ac.uk

Continued on next page

Name	Department	Institution	Email
Heather Yorston	Schools of Maths and Informatics	UoE	heather.yorston@ed.ac.u
Huizi Zhang	School of Mathematics	University of Edinburgh	H.Zhang-144@sms.ed.ac.
Inga Vermeulen	Childlight - Global Child Safety Institute	University of Edinburgh	Inga.Vermeulen@ed.ac.u
Ioannis Papastathopoulos	School of Mathematics	University of Edinburgh	i.papastathopoulos@ed.a
Jasmeen Kanwal	MHSES	University of Edinburgh	jasmeen.kanwal@ed.ac.u
Johnny Lee	School of Mathematics	University of Edinburgh	johnny.myungwon.lee@e
Jordan Richards	School of Mathematics	UoE	jordan.richards@ed.ac.ul
Kate Farrell	Data Education in Schools	University of Edinburgh	kate.farrell@ed.ac.uk
Katerina Karoni	School of Mathematics	University of Edinburgh	katerinakaron@gmail.co
Kimberly Lyons	Global Academy of Agriculture and Food Systems	University of Edinburgh	klyons@ed.ac.uk
Lambert De Monte	School of Mathematics	University of Edinburgh	l.demonte@ed.ac.uk
Lanxin Li	School of Mathematics	University of Edinburgh	lli11@ed.ac.uk
Laura Confalonieri	School of Mathematics	The University of Edinburgh	Laura.Confalonieri@ed.a
Lennart Hoheisel	Math	UoE	S1744523@ed.ac.uk
Liz Howell	Maxwell Institute	MACMIGS	s1523887@ed.ac.uk
Luz Pascual	School of Mathematics	University of Edinburgh	s2571924@ed.ac.uk
Man Ho Suen	Mathematics	University of Edinburgh	s1872841@ed.ac.uk
Margo Chase-Topping	Quantitative Biology	Roslin Institute	margo.chase@ed.ac.uk
Marifatul Amalia (Amalia)	Moray House School of Education and Sport	University of Edinburgh	M.Amalia@sms.ed.ac.uk
Mirella Lapata	School of Informatics	University of Edinburgh	mlap@inf.ed.ac.uk
Mr. John Wm. Dennis	Mathematics	University of St. Andrews	jwdennis22@gmail.com
Nicole Augustin	School of Mathematics	University of Edinburgh	nicole.augustin@ed.ac.uk
Nikolaos Makras	School of Mathematics	University of Edinburgh	s2451968@ed.ac.uk
Ozan Evkaya	School of Mathematics	University of Edinburgh	ozan.evkaya@ed.ac.uk
Professor Ailsa Henderson	PIR/SPS	University of Edinburgh	ailsa.henderson@ed.ac.ul
Rida Ayyaz	Statistics	University of Edinburgh	R.Ayyaz@sms.ed.ac.uk
Roger Halliday	also University of Glasgow	Research Data Scotland	roger.halliday@researchd
Rosie Wilkie	School of Maths	UoE	rosie.wilkie@ed.ac.uk
Ross Davidson	Animal and Veterinary Sciences	SRUC/Bioss	ross.davidson@sruc.ac.ul
Roxanne Connelly	School of Social and Political Science	University of Edinburgh	roxanne.connelly@ed.ac.
Ruth King	School of Mathematics	University of Edinburgh	Ruth.King@ed.ac.uk
Sara Wade	Mathematics	University of Edinburgh	sara.wade@ed.ac.uk

Continued on next page

Name	Department	Institution	Email
Sergio Gomez	School of Mathematics / Maxwell Institute Graduate School	University of Edinburgh	s2441782@ed.ac.uk
Serveh Sharifi	School of Mathematics	University of Edinburgh	serveh.sharifi@ed.ac.uk
Simon Taylor	School of Mathematics	University of Edinburgh	simon.taylor@ed.ac.uk
Simon Wood	Maths	Edinburgh	simon.wood@ed.ac.uk
Sjoerd Beentjes	School of Mathematics	University of Edinburgh	sjoerd.beentjes@ed.ac.uk
Takuo Matsubara	The School of Mathematics	The University of Edinburgh	takuo.matsubara@ed.ac.
Thanakorn Jaemthaworn	Centre of Regenerative Medicine	The University of Edinburgh	T.jaemthaworn@sms.ed.
Tiffany Vlaar	School of Mathematics & Statistics	University of Glasgow	Tiffany.Vlaar@glasgow.a
Tim Cannings	School of Mathematics	University of Edinburgh	timothy.cannings@ed.ac.
Torben Sell	School of Mathematics	University of Edinburgh	Torben.sell@ed.ac.uk
Ulrich Germann	School of Informatics	University of Edinburgh	ugermann@inf.ed.ac.uk
Usama Nadeem	Mathematics	Edinburgh	S1510444@ed.ac.uk
Vanda Inacio	School of Mathematics	University of Edinburgh	vanda.inacio@ed.ac.uk
Victor Elvira	School of Mathematics	University of Edinburgh	victor.elvira@ed.ac.uk
Viviana Carcaiso	Department of Statistical Sciences/School of Maths	University of Padova/University of Edinburgh	V.Carcaiso@sms.ed.ac.uk
Wenxing Zhou	School of Mathematics	University of Edinburgh	s2002002@ed.ac.uk
Xindi Song	School of Mathematics	University of Edinburgh	s2150450@ed.ac.uk
Xuebin Zhao	School of Geosciences	University of Edinburgh	xuebin.zhao@ed.ac.uk
Yubo Rasmussen	Actuarial Mathematics and Statistics	Heriot-Watt	yr2001@hw.ac.uk
Zhaoxi Zhang	School of Mathematics	University of Edinburgh	s2151978@ed.ac.uk
Zhuolin Pan	School of GeoSciences	College of Science & Engineering	s2601989@ed.ac.uk
Edward Wallace	School of Biological Sciences	University of Edinburgh	Edward.Wallace@ed.ac.u