4th Young-ISA meeting

Researchers' Toolkit: Publication, Progression, and People

November 17, 2023

LCI-G018, Institute for Lifecourse and Society (ILAS)

University of Galway

Ollscoil na Gaillimhe
University of Galway

09:30 – 10:00	Registration	
10:00 - 10:10	Welcome address	
Keynote talk 1 (Chair: John Newell)		
10:10 - 10:50	Kathleen O'Sullivan	People: You Are in the Middle Too
Session 1 (Chair: Amir Jalali)		
10:50 - 11:05	Eleanor Fallon	Developing statistical literacy and numeracy in journalists for health risk communication
11:05 - 11:20	Darshana Jayakumari	A new goodness-of-fit diagnostic for count data models based on half-normal plots
11:20 - 11:40	Coffee break	
Keynote talk 2 (Chair: Shirin Moghaddam)		
11:50 - 12:30	Carl Scarrott	Plan for Progression
Session 2 (Chair: Davood Roshan)		
12:30 - 12:45	Luiza Piancastelli	A Bayesian model for clustering spatially distributed compositions with application to species abundances in the Great Barrier Reef
12:45 - 13:00	Emily Gribbin	Developing statistical models to improve the efficiency of techniques in the field of single molecule imaging
13:00 - 13:15	Luke Kelly	Asymptotic guarantees for Bayesian phylogenetic inference
13:15 - 14:20	Lunch	
Keynote talk 3 (Chair: Kate Finucane)		
14:20 - 15:00	Brendan Murphy	Navigating the publication system: Directions from an author, reviewer, and editor
Round table (Chairs: James A Sweeney & Fatima Jaouimaa)		
15:00 - 15:40	Kathleen O'Sullivan, Carl Scarrott and Brendan Murphy	
15:40 - 16:00	Coffee break	
Joint Young-ISA and BIR Session (Chairs: Gabriel Rodrigues Palma and Catherine Higgins)		
16:00 - 16:20	Ana Silva Couto	Ecological modelling for offshore renewable energy impact assessment

16:20 - 16:40	**Lydia King**	**The Role of Genomic Data in Stratifying Patients within Predictive Models for Breast Cancer Survival Outcome**
16:40 - 17:00	**Estevao Batista Do Prado**	**Bayesian additive regression trees for genotype-by-environment interaction models**
17:00 - 17:20	**Joshua Tobin**	**Identifying Risk Factors for Mental Health Disorders in Older People using Multi-View Latent Block Models**
17:20 - 17:30	**Closing of the event and poster prizes**	

People: You Are In the Middle Too

Kathleen O'Sullivan

University College Cork

Academia is more than just producing scholarly work and advancing careers. It is about people. Nowadays, it is nearly impossible to succeed in academia without interacting with people. A repeatedly used maxim, "Publish or Perish", coined by Coolidge in 1932, pops into conversations in academia, to describe the pressure to publish scholarly work to succeed. People are a crucial component in all aspects of research. It is difficult in our profession to avoid people. John Tukey once said, "The best thing about being a statistician is that you get to play in everyone's backyard". To do this in a meaningful way and to add value, we must communicate, talk and actively listen, to experts in different disciplines. In other words, we need to build relationships to foster confidence so that colleagues in other fields are comfortable trusting us with their data to facilitate our need to "play". We must network, understanding that everyone communicates differently and, more importantly, receives communication differently.

In this event, focusing on the three P's, Publication, Progression and People, the last P may be the most important. It is easy to overlook the significance of people in one's career. To succeed in meeting targets such as successful grant applications, publications, graduating research students or career advancement, we look to others, sometimes to learn what to do and, more critically, what not to do. We seek out mentors and role models, which come in many forms, to provide templates for success. Mentors are instrumental in helping us navigate our professional lives. Solid, truthful advice can come from the most unusual sources and often when we least expect it. I am not an expert in human behaviour or communication, but I have, just like you, a lived experience.

This talk will focus on how people shape our careers and how we shape others' careers. Somehow, that always puts us in the middle. We must not forget that students are the beating pulse of academia, who engage with us daily to not only master competencies in their chosen field of study but also to learn how to do research. No academic can exist without professional, managerial and support staff to ensure efficient operations and maintain vital facilities.

In conclusion, the people we meet are instrumental in shaping our academic journey, notwithstanding its challenges at times. Colleagues in our own and disparate disciplines challenge us intellectually. Mentors guide us through the complexities of academia. Support staff keep us functioning and students bring fresh perspectives. We must endeavour to cultivate meaningful and impactful connections throughout our academic careers. Ultimately, as the Irish proverb says, "Ní neart go cur le chéile", there is no strength without unity, meaning we can accomplish more if we work together.

Developing statistical literacy and numeracy in journalists for health risk communication

Eleanor Fallon

University of Limerick

People's perception of risk can be influenced by the media, so journalists play a vital role in the communication of health risks to society. Numeracy is an important skill for health literacy. However, previous literature has found that some journalists and journalism students have low numeric confidence levels and a wide range of numeric ability. Some journalists also report having mathematical anxiety.

Our study aimed to develop an education intervention to improve journalism students' statistical literacy and numeracy levels to support effective health risk communication. First, we investigated journalism students' subjective and objective numeracy levels, their interpretation of qualitative descriptors of risks and their preference for health risk communication. The results informed the development of an education intervention with journalism students. The intervention was embedded within a journalism module, where journalism students at the University of Limerick attended a two-hour seminar (Titled: "Reporting risk: How to communicate the numbers in the news"). We tested the impact of this intervention using a before-and-after study. Results are presented through quantitative measures (self-assessment tests) and qualitative measures (student summary reports).

We also created an evaluation framework for a well-established education intervention for health journalists in practice in the United States to improve their ability to interpret and reliably report the results of medical research. We compare the statistical literacy and numeracy of the two groups and the impact of the interventions.

A new goodness-of-fit diagnostic for count data models based on half-normal plots

Darshana Jayakumari

Maynooth University

Statisticians use model selection methods to obtain a model that can effectively capture patterns in the data. Extended count data models, in the context of generalized linear models, are typically based on the Poisson distribution or two-parameter extensions involving a dispersion parameter. One of the approaches used in selecting the best performing model is by using half-normal plots with a simulated envelope. However, closely related models can result in very similar plots. We suggest a new numerical measure as an extension to the existing half-normal plots that can be jointly used to aid model selection. In this talk we intend to evaluate the role of the suggested diagnostic measure for the cases of overdispersion, under dispersion and zero inflation.

Plan for Progression

Carl Scarrott

University of Galway

The Young-ISA are the future of Statistical Science. This talk will provide some guidance on academic progression possibilities and pathways, in particular via promotion, the importance of having a plan, with a pipeline to build your portfolio, and the need to be proactive in getting involved in professional practice, seeking opportunities and promoting of your work to give it prominence. I will pass on some of my experiences from applying for posts, being on interview and promotion panels that may be useful in supporting the progression of our early career researchers.

A Bayesian model for clustering spatially distributed compositions with application to species abundances in the Great Barrier Reef

Luiza Piancastelli

University College Dublin

The relative abundance of species is often used in ecological surveys to estimate a habitat's composition, a metric that reflects patterns of commonness and rarity of biological assemblages. The focus of this research are measurements of four species abundances at several transect locations along Australia's Great Barrier Reef (GBR) gathered between 2012 and 2017. We undertake the task of finding clusters of locations with similar species composition and investigate changes in clustering dynamics over a time when an unprecedented sequence of extreme weather events impacted the area. To this aim, a model that incorporates the geographical location of compositions is proposed, accounting for the possibility that nearby reefs are of similar abundance. This is accomplished with a Dirichlet mixture model and a spatial process prior that takes a neighbourhood structure setup from a Voronoi partition of reefs latitude/longitude coordinates. Bayesian inference is carefully addressed for the posterior model, which is doubly-intractable. A decline in the number of clusters is evidenced, suggesting a change in the composition patterns across the study period.

Developing statistical models to improve the efficiency of techniques in the field of single molecule imaging

Emily Gribbin

Queen's University Belfast

Lung cancer is currently one of the leading causes of cancer deaths worldwide with Non-Small Cell Lung Cancer (NSCLC) accounting for as many as 85% of these cases. The development of NSCLC is in part due to mutations in the DNA of Epidermal Growth Factor Receptors (EGFRs) causing uncontrolled or erroneous clustering of this protein which is involved in normal cell growth and development processes. Through measuring the separations of populations of EGFR clusters, an understanding of the molecular processes in cancer can be obtained. Targeted therapeutic treatment can therefore be personalised for each patient by stratifying according to the individual's EGFR clustering profile.

Fluorescence Localisation Imaging with Photobleaching (FLImP) is a single molecule imaging technique which enables the resolution of fluorescently labelled EGFR molecules as close together as 3 nm, which is well below the diffraction limit of conventional single molecule light microscopy (~250 nm). FLImP relies on the successive photobleaching of fluorescent molecules, known as fluorophores, bound to the EGFR on the cell surface. Photobleaching is the result of applying laser light to the fluorophores, eventually causing them to bleach and stop emitting light. Modelling the number and location of these bleaching events can be used to estimate the number of fluorophores active in each observation frame. This information can then be used, alongside assumptions about the shape of the point spread function of the images obtained, to determine the separations between the EGFR clusters to which the fluorophores are bound. Following this, analysis is then carried out to determine the most likely type of EGFR cluster present.

FLImP generates approximately 1TB of raw data each day and is operating constantly. However, not all of this data is suitable for analysis. Thus, an additional purpose of modelling the number of fluorophores as described above is to rapidly identify subsets of the FLImP datasets which are suitable for the computationally expensive downstream analysis. Originally, this selection was completed manually and is now completed heuristically using a combination of k-means clustering, change-point algorithms, and extensive filtering, carried out using approximately twenty user-defined threshold parameters. However, this heuristic approach remains inefficient, in particular when fluorophores exhibit complex bleaching profiles.

In our research, we are focusing on using adaptive Markov chain Monte Carlo methods for changepoint problems to determine the location of the bleach events. Factorial hidden Markov models will then be used to directly model the state of fluorophores through time. This presentation will discuss the development of these approaches and future directions.

Asymptotic guarantees for Bayesian phylogenetic inference

Luke Kelly

University College Cork

Phylogenetics is the problem of reconstructing the evolutionary history of species descended from a common ancestor. We attempt to infer the phylogeny, typically a tree where the topology depicts relationships and edge lengths elapsed time, from data observed at its leaves. There are many models and algorithms for phylogenetic inference but few results on their consistency. For the Bayesian setting with two widely used choices of prior distributions on trees, we show under mild conditions that the posterior concentrates on the true data-generating tree as amount of data increases. The convergence rates match known frequentist rates obtained under stronger assumptions.

Navigating the publication system: Directions from an author, reviewer and editor

Brendan Murphy

University College Dublin

This talk gives an overview of what happens in the background when you submit a paper to a journal. The talk is based on my experience as an author, reviewer and editorial board member.

Ecological modelling for offshore renewable energy impact assessment

Ana Silva Couto

Biomathematics and Statistics Scotland, Edinburgh, UK

Due to the climate emergency and energy insecurity, many countries including the UK and Ireland, are increasing the pace of deployment of offshore wind farms (OWFs). Rapid expansion may lead to significant environmental impacts, creating an urgent need to better understand the ecological changes caused by the construction and operation of OWFs. This is of particular importance for protected species of seabirds, which may be adversely impacted by direct mortality from collisions or indirect harm such as disturbance (e.g., changes in prey availability). However, building an evidence base is challenging given the dynamic nature of both the marine environment and seabird behaviour.

The Offshore Renewables Group at Biomathematics and Statistics Scotland are currently involved in two large-scale projects aiming to improve our understanding of predator-prey interactions before, during, and after offshore wind farm construction. Both projects involve collecting and analysing detailed data on seabirds and fish in the Firth of Forth (east coast of Scotland). A variety of different data are being collected: boat-based surveys and unmanned autonomous vehicles (at the surface and underwater) are being used to collect information on fish presence and density using acoustic surveys, pelagic trawls, and grab samples; aerial surveys provide information on seabirds seen at-sea; and seabirds tagged with GPS devices collect data on individual movements over time. The complexity of the data and research questions require cutting-edge and robust modelling approaches that account for different ecological, spatial, and temporal scales, scalability (to other areas), and integration of different data types. Thus, we are developing spatial (inlabru, mgcv) and movement modelling (Hidden Markov Models) frameworks to address the research questions, while accounting for intrinsic issues such as missing data, sparseness, and autocorrelation. An increased understanding of how predators and prey and their interactions are impacted by offshore wind farms will allow us to improve the parameterisation of key impact assessment tools that have been developed. This will lead an improved quantification of uncertainty and has the potential to also lead a reduction in uncertainty in the assessment of ornithological constraints within the consenting process for wind farm developments.

The Role of Genomic Data in Stratifying Patients within Predictive Models for Breast Cancer Survival Outcome

Lydia King

University of Galway

Genomic instability (GI) is a common feature of cancers and can be defined as an increased tendency for genomic alterations to occur, such as base substitutions, indels and copy number alterations (CNAs). GI can initiate cancer, affect progression and influence patient prognosis. CNAs have been extensively profiled but due to the complexity of cancer genomes, frequent deviations from diploidy and the presence of both tumour and non-tumour cells, many studies have been limited to reporting total copy number, the sum of the copy numbers of the two homologous chromosomes. Determining the copy number landscape of each homologous chromosome, i.e. allele-specific copy number, is important for the characterisation of certain types of genomic aberrations and the inference of their clonal history.

Breast cancer is largely dominated by CNAs, rather than mutations in a single gene, and increasing evidence suggests that the genomic landscape of the tumour is associated with survival and incorporating this information into treatment decisions is beneficial to the patient. We therefore aim to use both total and allele-specific copy number data to explore the CNA landscape of breast tumours and their associations with survival.

This study focuses on observations from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) cohort. We define novel metrics for total CNA measurements, estimating the distribution of these metrics allowing for missing value presentation. ASCAT is applied to derive allele-specific copy number profiles making it possible to determine the CNA landscape of each homologous chromosome. We classify and model features of allele-specific CNA associated breakpoints via multivariate statistical modelling, investigate the relationship between derived CNA metric profiles and established molecular-based classifications, such as PAM50 intrinsic subtype and Integrative Cluster, and model survival outcome using the CNA metrics and relevant clinical variables.

Bayesian additive regression trees for genotype-by-environment interaction models

Estevao Batista Do Prado

Lancaster University

We propose a new class of models for the estimation of genotype by environment (GxE) interactions in plant-based genetics. Our approach, named AMBARTI, uses semiparametric Bayesian additive regression trees to accurately capture marginal genotypic and environment effects along with their interaction in a cut Bayesian framework. We demonstrate that our approach is competitive or superior to similar models widely used in the literature via both simulation and a real world dataset. Furthermore, we introduce new types of visualisation to properly assess both the marginal and interactive predictions from the model. An R package that implements our approach is also available at https://github.com/ebprado/ambarti

Identifying Risk Factors for Mental Health Disorders in Older People using Multi-View Latent Block Models

Joshua Tobin

Trinity College Dublin

The TUDA dataset is a large scale observational study investigating risk factors for poor mental health among older people. The TUDA study collected clinical, lifestyle, nutritional, and physiological data from patients, necessitating splitting the dataset into multiple views, each containing variables representing certain characteristics. To identify risk factors associated with poor mental health in older people, both the participants and survey features are to be summarized in groups. The simultaneous clustering of both the participants and the questions requires co-clustering methods. The Latent Block Model (LBM) is a prominent model-based co-clustering method, returning parametric representations of each block cluster. The LBM, while adapted in literature to handle varying feature types, cannot represent information which is spread heterogeneously between data views. We introduce the multi-view LBM, extending the LBM to multi-view data where each view marginally follows a LBM. In the case of two views, the dependence between them is captured by a cluster membership matrix and we aim to learn the structure of this matrix. To improve exploration of the model space, and to verify the connection between the multiple data views, we extend recent work developing hypothesis tests for the null hypothesis that the latent row-cluster memberships between the views are independent. The testing procedure is integrated into the model estimation strategy. Applying the new method to the TUDA dataset, we identify risk factors associated with poor mental health in older people using information spread across the different data modalities.

Acknowledgment

The Young-ISA would like to express our sincere thanks to everyone who contributed talks and for participating in this workshop. We acknowledge the Irish Statistical Association for funding this workshop through its short course fund.

We also thank the Sonraí Health Data Science Research Cluster and the admin office in the School of Mathematical and Statistical Sciences in University of Galway for their financial and organizational support.

We also thank the British and Irish Region (BIR) of the International Biometric Society (IBS) for cooperation and financial contribution toward organising our joint Young-ISA & BIR session.

Organising Committee:

University of Galway: Davood Roshan, Kishor Das, Nastaran Sharifian, Omid Khazaei and Pouyan Nejadi

Young-ISA: Shirin Moghaddam, Amirhossein Jalali, Silvia D'Angelo, Fatima Jaouimaa, Kate Finucane, Gabriel Rodriques Palma, Jonathan Henderson, Autumn O Donnell, Catherine Higgins, and James A Sweeney

Ollscoil na Gaillimhe
University of Galway

INTERNATIONAL BIOMETRIC SOCIETY
BRITISH AND IRISH REGION