I.S.A.

3rd Young-ISA meeting

Collaborative Workflow

In Statistics and Data Science

October 7, 2022

AD2-010, Analog Devices Building

University of Limerick

UNIVERSITY OF
LIMERICK
OLLSCOIL LUIMNIGH

MACSI
MATHEMATICS
APPLICATIONS CONSORTIUM
FOR SCIENCE & INDUSTRY

10:00 – 10:30	Registration	
10:30 - 10:40	Opening & welcome address	
Keynote talk 1 (Chair: Norma Bargary)		
10:40 - 11:30	Fernando de Pol Mayer	Case Studies in Data Science Workflow
Session 1 (Chair: Davood Roshan)		
11:30 - 11:50	Shubbham Gupta	MetaboVariation: Exploring individual variation in metabolite levels
11:50 - 12:10	Luke Kelly	Coupling Markov chain Monte Carlo phylogenetic inference
12:10 - 12:30	Hannah Comiskey	Estimating and Projecting Subnational Contraceptive Supply-Shares
12:30 - 12:50	Andrew McInerney	Feedforward neural networks as statistical models
Maria Doyle & Aedin Culhane		Introducing Bioconductor, R software for analysis of biological data
12:55 - 13:50	Lunch	
Keynote talk 2 (Chair: Kevin Burke)		
13:50 - 14:40	Isabella Gollini	Programming and collaborations: roles and tools when using R
Round table (Chairs: Silvia D'Angelo & Rafael de Andrade Moral)		
14:40 - 15:10	Isabella Gollini and Fernando de Pol Mayer	
15:10 - 15:30	Coffee break	
Session 2 (Chair: Amirhossein Jalali)		
15:30 - 15:50	Catherine Timoney	Network Analysis of a COVID-19 Outbreak in the West of Ireland
15:50 - 16:10	Jonathan Henderson	A Robust Mixed Effects Approach for the Early Diagnosis of Glaucoma
16:10 - 16:30	Kishor Das	Statistical Approaches for Method Comparison Studies involving Functional Responses with Applications in Elite Sports
16:30 - 16:50	Laura Byrne	Multivariate Statistical Models for Biodiversity Experiments
16:50 - 17:00	Closing and announcement of the Twitter poster conference winners	

Case Studies in Data Science Workflow

Fernando de Pol Mayer

Maynooth University & Federal University of Paraná, Brazil

In this talk, we will showcase different types of data science project outputs, and a general workflow to deal with them. We will show the importance of reproducibility and a set of software tools to help you manage this reproducible workflow. Some real practical examples will be shown thoroughly as case studies of this general practice.

MetaboVariation: Exploring individual variation in metabolite levels

Shubbham Gupta

University College Dublin

To date, most metabolomic biomarker research has focused on the identification of disease biomarkers. However, there is a need for biomarkers of early metabolic dysfunction to enable the identification of individuals who would benefit from lifestyle interventions. Concomitant with this, there is a need to develop strategies to model metabolomic data at an individual level. We propose in this paper to use repeated measurements on individuals to analyse fluctuations in metabolite levels at an individual level. We employ a Bayesian generalised linear model (BGLM) to detect individuals with considerable variation in metabolite levels. The BGLM models metabolite levels as a function of explanatory variables while accounting for intra-individual variation using repeated measurements. The posterior predictive distribution of metabolite levels at the individual level is returned, which can be used to flag individuals who have observed metabolite levels outside the 95 percent central credible interval at a given time point. Applying the proposed approach to the dataset containing metabolite levels for 20 metabolites measured once every four months in 164 individuals identified a notable percentage of individuals who had significant variation in three or more metabolites. In summary, MetaboVariation makes considerable progress in developing strategies for analysing data at the individual level, thus paving the way toward personalised nutrition.

Coupling Markov chain Monte Carlo phylogenetic inference

Luke Kelly

University College Cork

Phylogenetic inference attempts to reconstruct the ancestry of a set of observed taxa and is an intractable statistical problem on a complex, high-dimensional space. The likelihood function is an integral over unobserved evolutionary events on a tree and is often multimodal. Markov chain Monte Carlo (MCMC) methods are the primary tool for Bayesian phylogenetic inference but constructing sampling schemes to efficiently explore the associated posterior distributions or assess their performance is difficult.

Couplings have recently been used to construct unbiased MCMC estimators and compute bounds on the convergence of chains to their stationary distribution. We describe a procedure to couple a pair of Markov chains targeting a posterior distribution over a space of phylogenetic tree topologies, branch lengths, scalar parameters and latent variables such that the chains meet exactly at a random, finite time. We use the meeting times to diagnose convergence in total variation distance jointly across all components of the model on trees with up to 200 leaves.

Estimating and Projecting Subnational Contraceptive Supply-Shares

Hannah Comiskey[1], Leontine Alkema[2], Niamh Cahill[1]

[1]Hamilton Institute, Maynooth University, Ireland

[2]School of Public Health and Biosciences, University of Massachusetts Amherst

Assessing the longevity and sustainability of modern contraceptive availability and supply relies on quantifying the contributions of the public and private sectors to the method supply chain. In recent years, subnational estimation of the contraceptive supply share (public vs private sector) has become increasingly important as decision-making decentralises from the national level, making understanding variation across subnational regions (over time) essential for effective management of the contraceptive market. Large-scale national surveys, such as the Demographic and Health survey (DHS) provide valuable information for quantifying the contraceptive supply market. However, countries carry out a DHS approximately every 3-5 years, resulting in the need to bridge data gaps between survey years and project beyond the time of the most recent surveys.

Using DHS microdata for 35 countries and 248 subnational regions, we develop a statistical model to produce a set of related contraceptive supply-share outcomes (proportion of modern contraceptive methods supplied by the public and private sectors) at the subnational level. The model builds on a national-level model which relies on P-splines and a geographical hierarchical structure to estimate the country, method-specific contraceptive supply share over time. This national model utilises information about correlations between rates of change in method supply across all countries to help to inform supply-share estimates in the absence of data. However, disaggregating the model to the subnational level requires estimation of national-level correlations which poses a problem in data sparse settings when correlations cannot be estimated. To address this problem, we employ K-medoid clustering to cluster countries together using indicators of contraceptive prevalence and unmet need and replace national-level correlations with cluster-level correlations when required. Overall, the modelling approach offers an intuitive way to share information across subnational regions and countries and utilises information about relationships between rates of change in method supply to inform and produce a set of estimates that reflect the subnational changes in the method-specific contraceptive supply share over time, while accounting for data uncertainties.

Feedforward neural networks as statistical models

Andrew McInerney

University of Limerick

Feedforward neural networks (FNNs) are typically viewed as “pure prediction” algorithms, and their strong predictive performance has led to their use in many machine-learning applications. However, quite simply, FNNs are non-linear regression models, where the covariates are mapped to the response through a series of weighted summations and non-linear functions. Their success in predictivity can be attributed, at least in part, to their ability to capture complex relationships through the modelling of higher-order interactions. However, their flexibility comes with an interpretability trade-off; thus, FNNs have been historically less popular among statisticians, who tend to use more interpretable additive models. Nevertheless, classical statistical theory such as significance testing and uncertainty quantification is still relevant for FNNs. Supplementing FNNs with methods of statistical inference, model selection and the covariate-effect visualisations, can shift the focus away from “black-box” prediction and make FNNs more akin to traditional statistical models. This can pave the way towards more inferential analysis, and, hence, increase the utility of the FNN within the statistician's toolbox.

Programming and collaborations: roles and tools when using R

Isabella Gollini

University College Dublin

Programming is a crucial part of a statistician's job to initiate and carry out collaborative projects. We may collaborate for many different reasons such as new applications, or to develop new methods, improve efficiency in existing methods, apply or create some user-friendly package software or interface.

In this talk, I will go through some of the main aspects associated with the different roles that we can assume when collaborating in programming. I will then focus on the tools we can use to organise and share our projects in the context of R programming.

Network Analysis of a COVID-19 Outbreak in the West of Ireland

Catherine Timoney

HSE West

Notification and contact tracing systems for COVID-19 hold a vast amount of information on the transmission chains of the virus. Network analysis is very useful in visualising the spread of COVID-19, getting a better understanding of how cases are linked together and noticing how far reaching the initial infections were. Here the Irish infectious disease notification system, the Computerised Infectious Disease Reporting system (CIDR) and contact tracing system, the Covid Care Tracker (CCT), are used to create networks of the spread of the virus. These two databases are not linked together and no unique identifier for entries is present across the two. I will discuss a method of linking the datasets, issues that arise in doing so, and look at the resulting network of an outbreak in the student population of Galway, Ireland.

A Robust Mixed Effects Approach for the Early Diagnosis of Glaucoma

Jonathan Henderson

Queen's University Belfast

Glaucoma is a neurodegenerative disease caused by a loss of retinal ganglion cells (RGCs) which affects millions of people globally. Treatments to slow the rate of glaucoma related vision loss are most effective when given early in the disease process. An early diagnosis will therefore improve patient outcomes. Currently, diagnosis of glaucoma is a slow process relying on indirect measures of patient vision (retinal layer thickness or visual acuity). Emerging diagnostic techniques allow non-invasive visualisation of a small subset of the RGC population in patients. As the distribution of RGCs in the retina is heterogeneous, there is growing interest in determining whether patterns in this RGC point-pattern subset can be used to infer the presence of glaucoma and so facilitate early diagnosis. Using a replicated RGC point-pattern dataset derived from natural histories of well-established rodent models of glaucoma, this project sought to determine whether linear mixed models could infer disease status from RGC population subsamples earlier than conventional pooled summary statistic approaches. To account for outliers or other forms of contamination, a robust mixed effects model approach was adopted. The model covariates included the quadrant from which the RGC sample was derived, the distance to established retinal landmarks (the optic nerve head) and the size of the RGCs in the pattern. Likelihood ratio tests were carried out to obtain the best model and residual plots were used to detect outliers. Models were fitted to two longitudinal glaucoma datasets (OHT and pONT) and compared to a control (healthy) retinal dataset to determine how long after the degenerative process is induced that a significant difference in the parameters can be observed. The analysis revealed that a robust mixed effects model approach can be used to facilitate the diagnosis of Glaucoma earlier than conventional pooled summary statistic approaches. Future work will evaluate these models in patient derived RGC datasets.

Statistical Approaches for Method Comparison Studies involving Functional Responses with Applications in Elite Sports

Kishor Das

University of Galway

Method comparison study analyses agreement between two methods of measurement measuring the same quantity. This study could either involve a univariate response or a functional response. For the studies involving a functional response, this research proposes a novel approach to analyse agreement when the study administers a hierarchical study design with functional replicates and covariates adjustment is required.

Multivariate Statistical Models for Biodiversity Experiments

Laura Byrne[1], Rishabh Vishwakarma[1], John Connolly[2], Rafael de Andrade Moral[3], Forest Isbell[4], Caroline Brophy[1]

[1]Trinity College Dublin

[2]University College Dublin; [3]Maynooth University; [4]University of Minnesota

The relationship between the species biodiversity of an ecosystem and the outputs that the ecosystem produces (ecosystem functions) is often called the biodiversity and ecosystem function relationship. The Diversity-Interactions modelling framework is a regression-based approach used for modelling the biodiversity and ecosystem function relationship. It assumes that the initial proportional abundance of the species in an ecosystem is the primary driver behind the changes in its functioning, as opposed to the common usage of the number of species present (species richness). The models include the identity effect of each species, along with the interaction effects that may occur between species (which can take many forms). The form of the interactions may be altered through the inclusion of a non-linear parameter θ, forming a Generalised Diversity-Interactions model. This parameter is applied as a power to each interaction term, allowing deviation from the assumption that the interaction is directly proportional to the product of each species' proportions. Multivariate, logit, and mixed effect versions of these models have also been developed. The multivariate model is especially useful when multiple ecosystem functions are jointly of interest to study. The multivariate Diversity-Interactions model can account for covariance between differing ecosystem functions and identify any trade-offs that exist between functions in differing communities. It is also possible to compare and statistically test the relative performance of species (and their interactions) across the ecosystem functions using this method.

Following the publishing of the `DImodels` R package in 2020, we have been developing an add-on package `DImodelsMulti`. The original package enables the automatic fitting of a range of Generalised Diversity-Interactions models for a single ecosystem function studied at a single point in time. The new package can additionally model multiple responses and incorporates a variety of correlation structures to account for repeated measurements on a single experimental unit. The main function of the package is called `DIMulti()`, its purpose is to accept any relevant information from a user and fit the chosen styles of Diversity-Interactions models to the data provided using the internal modelling functions, returning an object of custom class `DIMulti`.

Current work involves the development of multivariate Diversity-Interactions models which can predict the change in the proportions of species present over time in parallel with their total ecosystem function value. These models, which will also be made available in the package, will enable a more specific prediction of the changes in an ecosystem caused by species interactions, both with each other and with invading species.

The new `DImodelsMulti` package provides considerable flexibility in model selection and estimation of Generalised Diversity-Interactions models and will be a useful tool in the study of multifunctional biodiversity and ecosystem function relationships studied over time.

Thanks

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 Prof. James Gleeson, Prof. Norma Bargary and Ena Brophy and the UL Department of Mathematics and Statistics (MACSI) for financial and organizational support.

Organising Committee:

UL: Shirin Moghaddam, Amirhossein Jalali, Fatima-Zahra Jaouimaa, Ena Brophy

Young-ISA: Rafael de Andrade Moral, Shirin Moghaddam, Amirhossein Jalali, Fatima Jaouimaa,
Lisa McFetridge, Davood Roshan, Silvia D'Angelo, Gabriel Rodriques Palma,
James Ng, Jonathan Henderson, Rabia Naqvi