2nd Young-ISA meeting

Communication & Reproducibility in Statistics

January 28, 2022

Virtual

UNIVERSITY OF LIMERICK
OLLSCOIL LUIMNIGH

10:00- 10:15	Opening & welcome address (Chair of the Young-ISA)	
Session 1 (Chair: Silvia D'Angelo)		
10:15-10:35	Edward Gunning	Processing and exploratory functional data analysis of a large biomechanical data set
10:35-10:55	Beatrice Charamba	Bayesian functional concurrent model for missing sensor data
10:55-11:15	Joshua Tobin	DCF: an efficient and robust density-based clustering method
11:15-11:35	Maeve Upton	Noisy input generalised additive models for relative sea level change along the east coast of North America
11:35-11:50	Break	
Keynote talk 1 (Chair: Davood Roshan)		
11:50-12:40	Professor Andrew Parnell	Reproduction as an enjoyable and important part of your research
12:40-13:30	Lunch	
Keynote talk 2 (Chair: Shirin Moghaddam)		
13:30-14:20	Professor Ailish Hannigan	From consultation to partnership: effective communication for statisticians
Round table (Chairs: Niamh Cahill & James Ng)		
14:20-14:50	Prof. Ailish Hannigan and Prof. Andrew Parnell	
14:50-15:00	Break	
Session 2 (Chair: Amirhossein Jalali)		
15:00-15:20	Sajal Kaur Minhas	Quantifying the degree of gait pathology in children with cerebral palsy
15:20-15:40	Kevin Brosnan	100 milliseconds to prevent fraud
15:40-16:00	Fengyun Gu	Quantitation of dynamic PET scans
16:00-16:20	Laura Boyle	Monitoring overcrowding in a network of hospital emergency departments
16:20-16:30	Closing	

Processing and exploratory functional data analysis of a large biomechanical data set

Edward Gunning

University of Limerick

Due to recent technological advances, laboratories, rehabilitation clinics, and wearable sensors now produce large amounts of data for biomechanics and human movement research. These data can improve understanding of sports performance, healthy and pathological movement, and the effects of treatments and therapies. Functional Data Analysis (FDA) has excellent applicability to biomechanical data that are measured continuously throughout a movement (e.g., forces and joint angles), because it assumes that the data are being generated by a smooth underlying curve, rather than treating them as a sequence of discrete measurements. In this talk I will present some preliminary work on a large kinematic dataset collected to investigate running-related injuries in recreational runners. These functional data are multivariate because the angles of the hip, knee and ankle in the three planes of motion are recorded simultaneously. The dataset also has a complex dependence structure – different subjects are measured for many strides, on both sides of the body.

Processing of functional data refers to the preparation of the "raw" data for analysis and modelling. It typically consists of three steps: time-normalisation to a common domain, smoothing or interpolation to convert the discrete measurements to functions, and registration to remove phase variation. After introducing the dataset, I will demonstrate how we have approached these steps. We combine time-normalisation and interpolation in the first step, and carefully choose the type and number of basis functions to achieve close approximation to the data, coupled with dimension reduction. Next, we apply landmark registration to remove phase variation in the data. It is well known that registration is application-dependent and there is no univocal set of criteria which can deem it to be successful. Therefore, I will present some exploration of the registered and unregistered data in an aim to understand the effects of registration and highlight how it might affect future analysis and results.

Bayesian functional concurrent model for missing sensor data

Beatrice Charamba

National University of Ireland, Galway

Introduction:

Functional data analysis (FDA) methods have recently been developed to analyse several variables measured repeatedly and concurrently over a domain such as time in a cohort of individuals. However, many FDA methods require data to be measured regularly, with data being collected at the same fixed times for all individuals. Often, with studies in humans, there tend to be missing data or data collected at different time points. To fit a functional regression model for such data, readily available software either use complete case data for modelling, perform single imputation or do not allow for inferences through confidence bands. In this study, we develop a Bayesian model for function-on-function regression in the situation of missing and irregular data which uses all the data for modelling and easily obtained inferences and can be used for imputation.

Methods:

A Bayesian functional concurrent model was developed and tested through a simulation study to compare it with other methods available. Four different functions for the functional coefficient parameter were considered, with increasing complexity. Methods were tested across a range of sample sizes, frequency of measurement, error structure and missingness. Models were then applied to a real world dataset involving concurrently measured glucose (every 5 minutes) and electrocardiogram (ECG) data (every 10 minutes) in a cohort of n = 17 type 1 diabetics.

Results:

The Bayesian model outperformed each of the other three functional concurrent models data in almost all scenarios considered. It is robust to changes in missingness and allows for easy inferences through automatically generated confidence bands.

Conclusion:

For irregular functional data with missingness, our novel Bayesian approach works well with missing data and can be used to perform multiple imputation. We recommend the use of Bayesian functional regression model for such data.

DCF: an efficient and robust density-based clustering method

Joshua Tobin

Trinity College Dublin

Density-based clustering methods have been shown to achieve promising results in modern data mining applications. A recent approach, Density Peaks Clustering (DPC), detects modes as points with high density and large distance to points of higher density, and hence often fails to detect low-density clusters in the data. Furthermore, DPC has quadratic complexity. We develop a new clustering algorithm, aiming at improving the applicability and efficiency of the peak-finding technique. The improvements are threefold: (1) the new algorithm is applicable to large datasets; (2) the algorithm is capable of detecting clusters of varying density; (3) the algorithm is competent at deciding the correct number of clusters, even when the number of clusters is very high. The clustering performance of the algorithm is greatly enhanced by directing the peak-finding technique to discover modal sets, rather than point modes. We present a theoretical analysis of our approach and experimental results to verify that our algorithm works well in practice. We demonstrate a potential application of our work for unsupervised face recognition.

Noisy input generalised additive models for relative sea level change along the east coast of North America

Maeve Upton

Maynooth University

The 2021 Intergovernmental Panel on Climate Change report highlighted how rates of sea level rise are the fastest in at least the last 3000 years. As a result, it is important to understand historical sea level trends at a global and local level in order to comprehend the drivers of sea level change and the potential impacts. The influence of different sea level drivers, for example thermal expansion, ocean dynamics and glacial – isostatic adjustment (GIA), has changed throughout time and space. Therefore, a useful statistical model requires both flexibility in time and space and have the capability to examine these separate drivers, whilst taking account of uncertainty.

The aim of our project is to develop statistical models to examine historic sea level changes for North America's and Ireland's Atlantic Coast. For our models, we utilise sea-level proxy data and tide gauge data which provide relative sea level estimates with uncertainty. The statistical approach employed is that of extensions of Generalised Additive Models (GAMs), which allow separate components of sea level to be modelled individually and efficiently and for smooth rates of change and accelerations to be calculated.

The model is built in a Bayesian framework which allows for external prior information to constrain the evolution of sea level change over space and time. The proxy data is collected from salt-marsh sediment cores and dated using biological and geochemical sea level indicators. Additional tide gauge data is taken from the Permanent Service for Mean Sea Level online. Uncertainty in dating is extremely important when using proxy records and is accounted for using the Noisy Input uncertainty method (McHutchon and Rasmussen 2011).

By combining statistical models, proxy and tidal gauge data, our results have shown that current sea level along North America’s east coast is the highest it has been in at least the last 15 centuries. The GAMs have the capability of examining the different drivers of relative sea level change such as GIA, local factors and eustatic influences. Our models have demonstrated that GIA was the main driver of relative sea level change along North America’s Atlantic coast, until the 20th century when a sharp rise in rates of sea level change can be seen.

This work is part of the larger nationally funded Irish A4 project (Aigéin, Aeráid, agus Athrú Atlantaigh — Oceans, Climate, and Atlantic Change), funded by the Marine Institute. It aims to examine ocean and climate changes in the Atlantic Ocean. The project targets three aspects of the Atlantic: its changing ocean dynamics; sea level changes; and Irish decadal climate predictions. In the future, we will apply this modelling technique to produce a long term historical record for relative sea level change in Ireland.

Reproduction as an enjoyable and important part of your research

Prof. Andrew Parnell

Hamilton Institute, Maynooth University

Making your work reproducible is widely considered to be a vital part of a modern researcher's toolkit. But the path to making code, data, and even full papers reproducible is surprisingly rocky and not just a simple task of learning some basic programming rules and styles. In this talk I'll go through some of the successes and challenges that I have had in trying to make my own work reproducible. There are a large number of resources and guides available online, and I will aim to distill these down to a set of useful lessons that should help you make your research more accessible to those who want to reproduce and adapt it

From consultation to partnership: effective communication for statisticians

Prof. Ailish Hannigan

School of Medicine, University of Limerick

Statistics is a collaborative discipline that frequently requires interactions with other disciplines and professions. Good communication skills (oral, writing and visual) are key to successful collaboration yet are rarely formally addressed in statistics education. This talk explores common communication challenges for statisticians in practice with examples of resources to develop communication skills in early career researchers. It also addresses new opportunities for communicating statistics with growing public engagement in research and citizen science.

Quantifying the degree of gait pathology in children with cerebral palsy

Sajal Kaur Minhas

University College Dublin

A typical gait analysis requires the analysis and interpretation of the kinematics of five segments or joints (trunk, pelvis, hip, knee and ankle/foot) in three planes. The quantity and complexity of the data necessitates the need to calculate the amount by which a subject's gait deviates from an average normal profile, and to represent this deviation as a single number. Such a measure can quantify the overall severity of a condition affecting walking, monitor progress, or evaluate the outcome of an intervention prescribed to improve the gait pattern.

The Gait Deviation Index (GDI) and Gait Profile Score (GPS) are the standard indices for measuring gait abnormality and work well on common gait pathologies such as cerebral palsy, rheumatoid arthritis and Parkinson's disease. The GDI is easy to interpret and is normally distributed allowing for parametric statistical testing. The GPS has the ability to decompose scores by individual joints/planes and altered indices without the need for a large control database, but it is not normally distributed. Neither indices accounts for the potential co-variation between the kinematic variables for any individual subject, i.e. the motions of one joint affect the motions of adjacent or remote joints. Additionally the intrinsic smoothness of the gait movement in each kinematic variable is not accounted for, i.e. the position of a joint at one time affects the positions at a later instant.

The aim of this work is to use techniques from multivariate functional principal components analysis to create an index that combines the advantages of the existing GDI and GPS. That is an index that is easy to interpret, is normally distributed, has the ability to decompose scores by individual joints and planes, and is easily adaptable. While also accounting for the intrinsic smoothness of the gait movement in each kinematic variable and the potential co-variation between the kinematic variables.

100 milliseconds to prevent fraud

Kevin Brosnan

Data Science Fraud and Financial Crimes, Fiserv

2020 was a record year for retail e-commerce sales worldwide with a total of $4.28 trillion spent, up almost a $1 trillion on 2019. However, with the adoption of e-commerce comes the opportunity for fraudsters to profit - with $17.5 billion confirmed as fraud in 2020, and an expected increase of over 14% in 2021 to approximately $20 billion, tools and techniques to detect and prevent fraudulent actors in real-time have never been so important. Data science lies at the core of making that possible, and this talk will discuss the problem, the challenges, the possible solutions, and the benefits of making accurate fraud predictions, while maintaining a frictionless experience for genuine consumers all in under 100 milliseconds.

Quantitation of Dynamic PET Scans

Fengyun Gu

University College Cork

Positron emission tomography (PET) scanning is an important diagnostic imaging technique used in the management of patients with cancer and in medical research. Dynamic PET scans give the possibility to recover metabolic information and the quantitation is concerned with the derivation of physiological parameters such as flow and flux. The main goal of this study is to apply an non-parametric approach for improving the diagnostic accuracy. Its performance has been examined and compared with the classical method – compartmental model in simulation studies. The results demonstrate that there are clear gains in mean square error (MSE) performance by the proposed approach.

Monitoring overcrowding in a network of hospital emergency departments

Laura Boyle

Queen's University Belfast

The COVID-19 pandemic has placed a novel strain on health systems internationally and changed the way that patients access medical care. At the start of the COVID-19 pandemic, the number of patients attending Australian emergency departments plummeted, with speculation that people were avoiding hospitals. Since then, the number of attendances has been increasing and EDs have been busier than their pre-pandemic operation since January 2021. This research uses near real-time data collected from a publicly available dashboard to monitor and predict the pressure on a network of hospital EDs in Australia. The talk will discuss models for predicting ED overcrowding and outline the challenges associated with causality and missing data.

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 are grateful to the admin support provided by the UL Department of Mathematics and Statistics admin office in particular, Peg Hanrahan.

Organising Committee:

Amirhossein Jalali

Shirin Moghaddam

Fatima Jaouimaa

Rafael De Andrade Moral

Davood Roshan

Niamh Cahill

Silvia D'Angelo

James Ng

Jonathan Henderson

Lisa McFetridge