Book of abstracts: Liverpool Early Career Researcher Conference 2019

Plenary Speakers

Prof. Nick Higham

Title: Exploiting Low Precision Arithmetic in the Solution of Linear Systems

Abstract: The landscape of scientific computing is changing, because of the growing availability and usage of low precision floating-point arithmetic, which provides advantages in speed, energy, communication costs and memory usage over single and double precisions. Of particular interest are the IEEE half precision (fp16) and bfloat16 arithmetics, the hardware support for which is primarily motivated by machine learning. Given the availability of these arithmetics, mixed precision algorithms that work in single or double precision but carry out part of a computation in half precision are now of great interest for scientific computing.

We consider solving a linear system $Ax = b$, with double precision A and b, by the use of a half precision LU or Cholesky factorization and mixed-precision iterative refinement. Among the points we discuss are

- how to avoid underflow and overflow, given the limited range of fp16 arithmetic,

- how to carry out error analysis for algorithms that use fp16 or bfloat16 arithmetic,

- how to simulate low precision arithmetic when hardware implementations are not available,

- the attainable speedups over state of the art solvers on current GPUs.

Dr. Richard Pinch

Title: Computing the future: mathematical challenges from computing and computational challenges from mathematics

Abstract: New ways of computing, not based on the further micro-miniaturisation of silicon circuits, are emerging as practical realities. The arrival of quantum computing is now a practical consideration for cybersecurity; and looking a little further ahead, data storage and processing based on the biological properties of DNA have been demonstrated in the laboratory. In this talk I will discuss some of the mathematics that is used to develop these novel computing paradigms, and some of the new mathematics that will be needed to make the most of them.

Dr. Maria Ferrario

Title: 'Measuring' human values in software production

Abstract: Human values, such as prestige, social justice, and financial success, influence software production decision-making processes. While their subjectivity makes some values difficult to measure, their impact on software - and of software on society - motivates my research. This talk makes the case for the study of human values in software production and offers two key principles in order to advance this research agenda. Firstly, the significance of values as distinguished from, though connected to, ethics; and secondly, the need for clear theoretical to values study. It then introduces a selection of tools and techniques that have been designed in accordance with these two principles and used with computing professionals from research, education, and industry. It concludes with discussion around lessons learnt, ongoing challenges, and future directions.

Dr. David Hughes

Title: Variational Models for Longitudinal Data

Abstract: Multiple longitudinal responses are now routinely collected in many areas of clinical research. Often a single outcome is considered and the effect of various covariates of interest on changes in the outcome over time is considered. However, many clinical outcomes are correlated and joint models for longitudinal outcomes not only allow a more efficient use of collected data, but also allow more complex questions about the data to be answered.

A common statistical tool for longitudinal data on multiple outcomes is a multivariate generalised linear mixed model. However, in large datasets these can be very computationally intensive if commonly used Bayesian techniques like MCMC are used. In this talk I will describe how a commonly used technique in the computer science and machine learning literature known as variational Bayes, can be applied to statistical problems to allow fast fitting of complex models.

Variational Bayes is an approximation technique that can be very accurate and fast in many scenarios. I will also show how we can utilise mathematical matrix algebra techniques to streamline our algorithm and make it faster.

I will illustrate and compare the use of MCMC and MFVB methods using simulation studies as well as a small Primary Biliary Cirrhosis dataset and a large Diabetes dataset to show that MFVB allows faster and comparably accurate inference to be made for MGLMMs with multiple markers.

Prof. Andy Jones

Title: An Introduction to the University of Liverpool's Computational Biology Facility and opportunities for ECRs in Data Science

Abstract: The University of Liverpool's Computational Biology Facility (CBF, http://cbf.liverpool.ac.uk/) provides collaborative research services for analysing large-scale data from health and life sciences. We are particularly focussed on data generated by "omics" techniques (genomics, proteomics and metabolomics amongst others). We use data science techniques extensively, particularly machine learning and computational statistics, for data integration, biomarker discovery and network analysis. Our current team comprises two Directors (Profs Falciani and Jones), two Associate Directors (Prof Rigden and Dr Antczak), six data scientists and two research software engineers. We also employee consultants from time-to-time to meet demand, and anticipate that we will continue to expand our staff base in the coming years.

In this talk, I will give an overview of the types of project we address in the CBF in areas ranging from disease research, through to crop sciences and environmental planning, as well as some of the software packages and web-based databases we have produced and actively maintain. I will also give some ideas as to how ECRs can get some experience working in data science applied to biological and biomedical data.

Dr. Hirbod Assa

Title: On some applications of machine learning in banking and insurance

Abstract: In this talk, we take a brief look at some applications of machine learning methods in banking and insurance. After reviewing the fundamentals, we discuss applications in factor analysis, credit risk valuation adjustment and applications in insurance conversion rate. In the end, we discuss to which extent machine learning can change the banking and insurance industry.

Dr. Edward Pyzer-Knapp

Title: Intelligent Simulations - Bigger, Faster, Smarter

Abstract: For a long time, computational scientists have relied on Moore's law (bigger, faster, cheaper) to realize their complex computational workflows on ever more powerful machines. Recently, however, this bastion of the digital world has started to show some cracks - the required speedups are no longer guaranteed simply by waiting a year or two, or investing in larger systems - clearly an alternative approach is needed.

We present the concept of *Intelligent Simulations* as the answer to this call. In an intelligent simulation, we ask the system to work smarter, not harder. This can include replacing expensive parts of the model with a cheaper surrogate, or iterating through batches of simulations in such as way that minimal redundant information is determined. A key player in this regime is a technique known as Bayesian optimization, and throughout this talk, I will demonstrate how we have built upon ancient mathematical foundations to build a new tool for the digital age.

Prof. Simon Maskell

Title: Big Data

Abstract: TBC

Dr. Lee Devlin

Title: Challenges in Space Situational Awareness: Detecting, Modelling and tracking

Abstract: In this talk I will discuss topics relevant to monitoring and characterising Earth-orbiting objects of human origin. Space is an exciting domain, utilised by many different disciplines which are important to modern civilisation. However, future missions and opportunities could be at risk from collisions either due to miscommunication or debris. I will first discuss the space domain and how it has evolved from Sputnik to the present day and highlight why debris has become such a hot topic. I will then present challenges in three areas where further work is required and seeing increased interest from both government and industrial sectors. These will be detection, modelling and tracking. Our understanding of the risk associated with a space collision (especially in the case of debris) is critically dependent on what we can detect. Soon-to-be online technology will mitigate the risk some-what, but this will ask some very challenging questions of satellite operators to ensure mission survival. Assuming we know what is up there, how can we model the dynamics of the system such that we can predict where an object will be at some future point in time. Finally, bringing the two together, how can we track an object given that our measurements are based on noisy data collected by telescopes observing objects potentially tens of thousands of kilometres away.

Dr. Mara Kozic

Title: Predicting Antimicrobial Peptides through Machine Learning

Abstract: Antimicrobial resistance within a wide range of infectious agents is a severe and growing public health threat. Antimicrobial peptides (AMPs) are among the leading alternatives to current antibiotics, exhibiting broad spectrum activity. Their activity is determined by numerous properties such as cationic charge, amphipathicity, size, and amino acid composition. In order to predict antimicrobials, the main focus hitherto has been on sequence features, and the models used were often very complex and obscure. In this work, we employed the support vector classifier method in order to develop a comprehensive model that can classify between AMPs and non-AMPs: our classifiers are based on sequence-based features as well as 3D features. The test set performance scores of our best 3-feature model were comparable to complex, state-of-the-art sequence-based deep learning models. Developing a simple classifier such as the one reported in this work are a step forward in protein design efforts towards better AMPs.

Contributed Talks

Wenyue Zhu

Title: Spatial modelling of retinal images---towards more accurate statistical inference

Abstract: In the analysis of retinal diseases, spatial context of the retinal images is a highly relevant information but its importance is not fully studied. For example, in ophthalmology, the thickness of macula is measured at nine circularly oriented locations with the goal to decide if the thickness is related to the disease. Despite the data being spatially collected the current approaches involve analysing each location separately in nine analyses or hence ignoring the possible spatial correlation. Moreover, correlation between eyes from the same patient are not taken into account which can make the existing statistical inference biased.

Our objective is to investigate appropriate statistical approaches to account for spatial correlations, specifically linear mixed effect model; the suitability of error specification.

We propose a statistical inference framework for retinal images. The framework is based on a linear mixed effect model with a spatial (Gaussian, autoregressive-1, exponential and spherical) error structure for the analysis of clinical imaging data. Correlation between eyes from the same patient is explained through nested random effects, and heteroscedasticity between groups is adjusted in the covariance. We evaluate the spatial framework on simulated and real data. A simulation study was inspired by a retinal thickness images from a prospective observational study Early Detection of Diabetic Macular Oedema (EDDMO). We compare our method with multivariate analysis of variance (MANOVA) to analyse the EDDMO dataset involving 89 eyes with maculopathy and 168 eyes without maculopathy from 149 diabetic participants at their baseline visit.

In the real dataset, MANOVA shows that maculopathy eyes are not different from no-maculopathy eyes in terms of retinal thickness over the nine locations ($p=0.11$), while our the mixed effect model with spatial error structure can detect the difference between maculopathy eyes and eyes without maculopathy ($p=0.02$). Based on information criterions, an exponential spatial correlation structure produce a better fit for this spatial data. We also find that age has a significant negative effect on the retinal thickness profile. In simulations, we illustrate how the spatial correlation (no correlation, low, moderate or high correlations) between different locations, as well as how sample size can change inferences about fixed effects.

Our model addresses the need of correct adjustment for spatial correlations in ophthalmic images. Such model is large potential to shed light into understanding the diseases and to be extended into prognosis.

Yiannis Simillides

Title: MLJ, Machine Learning in Julia

Abstract: In this talk, we present MLJ, Machine-Learning in Julia, a new alternative toolbox to compete against current offerings, such as MLR or Scikit-Learn, which allows for a pure-Julia matching language offering, while still allowing its integration with packages from other languages (such as Scikit-Learn) if one wishes. It has been designed at the Alan Turing Institute, London, alongside NeSI, New Zealand. We will discuss our interface design alongside our choices of API, and how these relate to machine-learning toolboxes in general. As part of this, we will discuss how we handle probabilistic predictions/models, something which is missing in offerings such as scikit-learn. We will also talk about how we handle categorical data types, thus providing safer and more accurate code, as MLJ models preserve categories throughout training and predicting. We then proceed to the handling of model metadata alongside our "task" pipeline, for efficient model evaluation. Finally, a quick demonstration of our package will be provided.

Sreelekshmy Sreeja

Title: A Semi Parametric Model for Relation Between Temperature and Ice Accumulation Rate

Abstract: In paleo-climatic studies chemical compositions of atmosphere which is preserved in the layers of ice sheets provide rich source of climate patterns. Ice-cores have been often drilled out and several paleo-climatic variables like temperature, carbon dioxide etc. were obtained by radio-active isotopes and/or air bubbles trapped between the layers of ice sheets. Scientists use the information obtained from ice cores to study the weather patterns, glacial-interglacial cycles, levels of green house gases etc. The relationship between temperature and ice accumulation is an important scientific study. We propose a semi-parametric model for the temperature and ice accumulation rate to describe their relationship and also an algorithm to estimate the parameters. The temperature reconstructed/proxy oxygen isotope) for the corresponding drilled ice cores (mostly equi-length) in the Antarctic regions as well as green land region is publically available at the website of National Centers for Environmental Information. We explore the temperature and ice accumulation behavior in the Antarctic region (EPICA Dome C, Vostok Lake and Dome Fuji) and Greenland region (Greenland Ice Core Project (GRIP)). The model is fitting all these data well.

Phillip Maffettone

Title: Deep learning from crystallographic representations of periodic systems

Abstract: While significant advances have been made in accelerating chemical discovery with machine learning, the use of these methods for crystalline materials usually requires restraints on the input structure or manually constructed feature vectors. The arbitrary size and periodicity of crystalline systems pose challenges as these systems need to be represented with a fixed dimensionality and retain translational, rotational, and permutation invariance. We present the use of crystallographically inspired transformations that are amenable to deep learning algorithms. By calculating a modified structure factor, a suite of representations are developed that are of fixed size irrelevant of the input dimensionality. The ability for these representations to capture the periodic structural information is demonstrated through classification and regression problems related to crystalline materials.

Bernadette Stolz

Title: Outlier-robust Subsampling Techniques for Persistent Homology

Abstract: The amount and complexity of biological data has increased rapidly in recent years with the availability of improved biological tools. Topological data analysis and more specifically persistent homology have been successfully applied to biological settings. When attempting to study large data sets however, many of the currently available algorithms fail due to computational complexity preventing many interesting biological applications. De Silva and Carlsson (2004) introduced the so called Witness Complex that reduces computational complexity by building simplicial complexes on a small subset of landmark points selected from the original data set. The landmark points are chosen from the data either at random or using the so called maxmin algorithm. These approaches are not ideal as the random selection tends to favour dense areas of the point cloud while the maxmin algorithm often selects outliers as landmarks. Both of these problems need to be addressed in order to make the method more applicable to biological data. Chawla (2013) developed a version of k- means that detects outliers while clustering data points. We show how this method can be used to select landmarks for persistent homology and also propose another new method specifically for the use in topological data analysis that can detect outliers based on the local persistent homology around data points. We show how both of these methods outperform the existing subsampling methods for persistent homology.

Yan Li

Title: Risk prediction models that use routinely collected electronic health data: generalisable and useful in heterogeneous settings?

Abstract:

Objective

To assess the extent of variability between practices on individual patients' risk of cardiovascular disease (CVD) that is not taken into account by the risk prediction model QRISK3.

Methods

Design: Longitudinal cohort study from 1st Jan 1998 to Jan 2015.

Setting: 392 general practices (including 3.6 million patients) from the Clinical Practice Research Datalink (CPRD)

Methods: Shared frailty model to incorporate QRISK3 predictors, practice variability and simulations to measure random variability.

Results

There was considerable variation in data recording between general practices. Practices on 5 th percentile of missingness of Body mass index have 18.7% patients with missing values and 60.1% on the 95th percentile (for ethnicity, these were 19.6% and 93.9%, respectively). The crude incidence rates also varied considerably between practices (from 0.4 to 1.3 CVD events per 100 patient-years, respectively). The estimates of individual CVD risks with the random effect model were inconsistent

with the estimated QRISK3 risk. For patients with a QRISK3 CVD risk of 10%, the 95% range of predicted risks were between 7.2% and 13.7% with the random effects model. Random variability only explained a small part of this. The random effects model was similar to QRISK3 for discrimination (C-statistic: 0.852 (95% CI: 0.850, 0.854)) and calibration (Brier score: 0.067 (95% CI: 0.067, 0.068)).

Conclusions

Risk prediction models that use routinely collected electronic health data can have limited generalisability and accuracy in predicting individual patient risks in heterogeneous settings. They need to be based on more robust evidence on causal risk factors.

Stephanie Shoop-Worrall

Title: Trajectories of disease activity over the first three years following juvenile idiopathic arthritis diagnosis

Abstract:

Background: The advent of biological therapies and early aggressive treatment strategies have drastically changed prognoses for children and young people (CYP) with juvenile idiopathic arthritis (JIA). Clinical trials and observational research have demonstrated improvements in disease for the majority, but not all, CYP over time. It is not currently known what the patterns of disease activity are in CYP with JIA and how these cluster over time.

Objectives: To explore latent patterns of clinical juvenile arthritis disease activity scores (cJADAS) following a diagnosis of JIA.

Methods: CYP with JIA were selected if enrolled in the Childhood Arthritis Prospective Study (CAPS), a UK multicentre inception cohort, before January 2015. cJADAS10 scores were calculated based on components (active joint count up to 10, physician global, patient/parent global) collected at diagnosis, six months, one year and then annually to three years. CYP were excluded if no cJADAS10 scores were available within this time frame.

Group-based trajectory models were constructed to model latent groups of cJADAS10 scores. Linear, quadratic and cubic polynomials were tested, with one to six trajectories tested within each polynomial group. An optimal model within each polynomial group was selected using Bayesian Information Criteria. The final model was then selected from this shortlist based on model parsimony and clinical plausibility.

Results: Of 1183 CYP selected, the majority were female (65%) and of white ethnicity (90%) with oligoarticular JIA the most common JIA category (45%).

The optimal model identified five cJADAS10 quadratic trajectories: Low-low (59%, initial cJADAS10 median: 6.1), moderate-low (16%, initial cJADAS10 median: 11.5) and three groups with high disease activity at initial presentation (initial median cJADAS10: 17.7 to 19.1). A high-low group experienced the greatest improvement (15%, median improvement 17.2 (IQR 13.7 to 20.1)), and a high-moderate group lesser improvement (5%, median improvement 7.3 (IQR 0.8 to 9.0)). A final high-low-high group experienced improvement to one year followed by disease relapse (5%).

Conclusions: Disease activity in CYP with JIA does not improve in a uniform manner following initial presentation to paediatric rheumatology. Five latent trajectory groups have been identified, with three of these displaying different patterns following initial high disease activity at diagnosis. Identifying distinguishing characteristics for each group may aid the stratification of different treatment strategies to facilitate personalised medicine in JIA.

Catherine Higham

Title: Deep Learning for Fast 3D Reconstruction with Compact Camera and Single Sensor Systems

Abstract: Real-time 3D scene reconstruction has applications in many areas including security, health and entertainment. Time-of-flight LiDAR systems are capable of depth estimation at the millimetre scale, but recovering the transverse spatial information requires bulky, expensive, laser scanning or detector array systems. Achieving 3D scene reconstruction with compact, relatively cheap camera devices and non-scanning single sensor hardware would considerably extend the bounds of possible applications.

We use deep learning models with sequential depth histograms acquired from a single sensor synchronised with RGB video, for spatial information, to learn compact representations of high resolution indoor and outdoor scenes and decoders to recover depth at video rates of 30Hz.

A recurrent neural network is trained on high resolution (480 x 640 pixels) RGB-depth datasets, comprising both indoor scenes (up to 10 metres) and outdoor scenes (up to 100 metres). We show that our approach, of fusing sequential single sensor laser depth histograms and a single channel from RGB images, improves the speed and accuracy of 3D reconstruction, in terms of quality and rate of reconstruction, compared with current methods based just on RGB.

We will discuss the novel algorithmic developments required in this work and illustrate performance on state-of-the-art RGB-Depth real-time datasets.

Gareth Jones

Title: Proof of Concept for Machine Learning Applications to Arterial Disease Detection

Abstract: Arterial disease is the name given to any disease effecting the arterial system. Two of the most common diseases are stenosis and aneurysm. A stenosis is a narrowing of an arterial vessel. The prevalance of stenosis has been recorded to be between 1.9% and 18.83% within different arterial vessels and different demographics. The second common form of arterial disease is aneurysm. An aneurysm is a localised weakening of an arterial vessel wall, causing the vessel

to bulge. The most common form of arterial aneurysm is the abdominal aortic aneurysm (AAA), with a prevalence of 4.8% [1]. It has been shown that changes to the cross sectional area of an arterial vessel cause a difference in the pressure-flowrate waveforms of the blood passing through that vessel [2-3]. This suggests that it should be possible to predict the presence of a stenosis or aneurysm within an arterial network using pressure-flowrate measurements. If a large database of pressure-flowrate measurements taken from patients of known arterial health is available, it should be possible for a machine learning (ML) classifier to be trained to distinguish between healthy and unhealthy patients. A direct prediction of a patient's health could then be made using pressure-flowrate measurements making the proposed method both inexpensive and near instantaneous.

This proof of concept will make a first step in assessing the possibility of using a ML algorithm to predict arterial disease. Two virtual patient databases containing healthy and unhealthy patients, similar to that presented in [4], are created as a surrogate to a real cohort. This virtual population is used to train a series of classifiers to detect arterial disease and then test their performance. It has been found that using virtual patients, a machine learning classifier could detect stenosis with a maximum accuracy of 77% for healthy patients and 61% for unhealthy patients . Aneurysm detection was worse with 62% of healthy patients and 61% of unhealthy patients classified correctly.

[1] Li, X., Zhao, G., Zhang, J., Duan, Z. and Xin, S., 2013. Prevalence and trends of the abdominal aortic aneurysms epidemic in general population-a metaanalysis. PLoS One, 8(12), p.e81260.

[2] May, Allyn G., James A. Deweese, and Charles G. Rob. "Hemodynamic effects of arterial stenosis." Surgery 53.4 (1963): 513-524.

[3] Bevan, R.L.T., Sazonov, I., Saksono, P.H., Nithiarasu, P., van Loon, R., Luckraz, H. and Ashraf, S., 2011. Patient-specific blood flow simulation through an aneurysmal thoracic aorta with a folded proximal neck. International Journal for Numerical Methods in Biomedical Engineering, 27(8), pp.1167-1184.

[4] Willemet, M., Chowienczyk, P. and Alastruey, J., 2015. A database of virtual healthy subjects to assess the accuracy of foot-to-foot pulse wave velocities for estimation of aortic stiffness. American Journal of Physiology-Heart and Circulatory Physiology, 309(4), pp.H663-H675.

Farhad Hatami

Title: Predicting Progression in Heterogeneous Neurodegenerative Diseases using a Joint Mixture Model Approach

Abstract: Much of the current research in neurodegenerative diseases focuses on identifying biomarkers and risk factors for clinical progression. However, these diseases are heterogeneous both in their biology and clinical phenotypes, and consequently predictive factors can vary between individuals. Therefore, identifying the predictive factors for a given patient is of great clinical importance to predict individual disease progression. The latter can be addressed by determining the underlying latent disease subtypes.

We develop a method that we name longitudinal joint cluster regression (L-JCR) to jointly estimate a predictive regression model and identify latent groups (or subtypes). Longitudinal dynamics are modelled using a mixed effects model, and estimated via restricted maximum likelihood (REML). The method can handle high-dimensional covariates by making sparsity assumptions via lasso penalization.

We apply our method to data from studies of Amyotrophic Lateral Sclerosis (ALS) patients, as an example of a heterogeneous neurodegenerative disease with very different progression profiles. We show that the accuracy of progression prediction improves taking the group structure into account, and that the inferred latent groups are biologically meaningful.