LASR 2019

University of Leeds, 24-26 June 2019

Book of Abstracts

Keynote Talks:

Alejandro Frangi (University of Leeds)

Statistical modelling in cardiac and bone imaging: from patient to population imaging

Statistical shape models are a technology that revolutionized computer vision and medical image analysis in the 1990s and 2000s and continues to exert its influence. SSMs allow computers to learn generative models that can explain the mean shape and shape variations of an object or organ of interest. This talk will update on a number of recent applications and direction in SSMs in the area of cardiology and musculoskeletal research, overviewing relevant algorithms, illustrating their application to medical image analysis.

Iuliana Ionita-Laza (Columbia University)

Integrative statistical approaches for predicting functional effects of variants in noncoding regions of the genome

Continuous advances in massively parallel sequencing technologies make large whole-genome sequencing studies increasingly feasible. The analysis of such data is challenging due to the large number of rare variants in noncoding regions of the genome and our limited understanding of their functional effects. In this talk I will discuss unsupervised and semi-supervised approaches to predict cell type/tissue specific regulatory function for variants in noncoding regions. I will also briefly discuss how to integrate a large number of functional predictions in sequence-based association tests for improved power to identify signals in noncoding regions. Throughout the talk I will show applications to several datasets.

Huiling Le (University of Nottingham)

Inference for Frechet means: Empirical Likelihood Methods versus CLT

The concept of Frechet means, as a generalisation of Euclidean means, is becoming an increasingly important fundamental tool in many statistical analyses of data of a non-Euclidean nature. At the same time, the non-standard probabilistic and statistical features of Frechet means give rise to many challenges, both practical and mathematical. A good understanding of these features is crucial for developing the appropriate statistical methods required, for example, for computation, estimation and inference. In this talk, I will discuss some recent progress in understanding the interplay between the probabilistic behaviour of Frechet means and the structure of the spaces on which they lie.

Kanti Mardia (University of Leeds and University of Oxford)

Recent advances in Directional Statistics with the focus on some cutting-edge applications

In Directional Statistics, like many other areas of Statistics, the computing power has led to solutions to many practical problems which were not tractable a decade ago. Also there is a growth of new data problems with advancing methods of collecting data. However, one of the major stumbling blocks in the Directional Statistics is difficulty in implementing the maximum likelihood estimation method for standard distributions, including the basic von Mises distribution for circular data. The root of the problem is that their normalizing constants are complicated; we show how this problem can be resolved by introducing score matching estimators into this area (Mardia, et al, 2016). Its use is illustrated by the data in tracking space debris (Kent et al, 2016). We then introduce Incremental estimation of von Mises mixtures for online inference for streaming data with our focus on human activities (Chinellato et al, 2017). As in the linear statistics, a need for PCA is obvious for high dimensional data on torus. Since a linear combination of angles does not make any sense, we have developed a torus PCA which looks into subspaces in a hierarchal way (Eltzner et al, 2018); we show how the torus-PCA helps in quantifying a hypothesised sub-structure in RNA . We will also discuss new circular piecewise regression and cell-cycle data arising in gene expression (Rueda, 2016), diffusion on torus and evolution through protein data bank (Golden et al, 2018), and finding peak times in Uber data through von Mises process (Navarro et al 2017).

One of the curious problems in this area is that there is no distribution like the normal distribution on line as the von Mises distribution belongs to the exponential family but the wrapped normal does not. However the wrapped normal has accessible moments and its sampling is straight forward. This has clearly led to two schools at least in the Machine Learning context (cf, Jona-Lasinio et al, 2018; Navarro et al, 2017). We will give a new general approximation “Score Matching Approximation” to connect the two distributions (Mardia, 2018).

All the above work is on continuous distributions but recently some challenging problems have appeared for discrete data such as in protein folding. We give a new discrete distribution which also helps in studying the classical problem of assessing bias in a roulette wheel which has found a new application in on-line gambling.

Most of the above my work is collaborative, and the papers have either appeared very recently or under the submission process. The references cited above will be available on my home page.

Virginie Rondeau (University of Bordeaux)

Multivariate Joint Frailty Model for the Analysis of Nonlinear Tumor Kinetics with Recurrent Progressions of Nontarget Progression and Dynamic Predictions of Death

The Response Evaluation Criteria in Solid Tumors are used as standard guidelines for the clinical evaluation of cancer treatments. The assessment is based on the anatomical tumor burden: change in size of target lesions and evolution of nontarget lesions (NTL). Despite unquestionable advantages of this standard tool, Response Evaluation Criteria in Solid Tumors are subject to some limitations such as categorization of continuous tumor size or negligence of its longitudinal trajectory. In particular, it is of interest to capture its nonlinear shape and model it simultaneously with recurrent progressions of NTL and overall survival. We propose different multivariate nonlinear joint frailty models for recurrent events, and a terminal event with a longitudinal data. In the model, the tumor size trajectory is described using an ordinary differential equation that accounts for the natural growth and treatment-induced decline. We perform a simulation study to validate the method and apply the model to a phase III clinical trial in colorectal cancer. In the results of the analysis, we determine on which component, tumor size, NTL, or death the treatment acts mostly and perform dynamic predictions of death. We compare the model with other models that consider parametric functions or splines for the tumor size trajectory in terms of goodness of fit and predictive accuracy.

Invited Talks

Nicole Augustin (University of Bath)

Modelling European spatio-temporal forest monitoring data on defoliation and mortality

Forests are economically, recreationally and ecologically important, providing timber and wildlife habitat and acting as a carbon sink, among many ecosystem services. Forest health is monitored in Europe by The International Co-operative Programme on Assessment and Monitoring of Air Pollution Effects (ICP Forests) in cooperation with the European Union. More recently climate change has contributed to the decline in forest health and monitoring data are increasingly being used to investigate the effects of climate change on forests in order to decide on forest management strategies for mitigation.

Here we model extensive yearly data on tree mortality and crown defoliation, an indicator of tree health, from a monitoring survey carried out in Southern Germany, which includes a part of the ICP transnational grid. On a changing grid, defoliation, mortality and other tree and site specific variables are recorded. In some cases the grid locations are no longer observed which leads to censored data, also recruitment of trees happens throughout when new grid points are added.

We model tree survival as a function of the predictor variables on climate, soil characteristics and water budget. We are interested in the process leading to tree mortality rather than prediction and this requires the inclusion of all potential drivers of tree mortality in the model. We use a smooth additive Cox model which allows for random effects taking care of dependence between neighbouring trees and non-linear smooth functions of time varying predictors and functional predictors. At each of 2385 locations 24 trees were observed between 1983 and 2016, with not all locations being observed yearly. Altogether a total of 57720 trees are observed making the analysis computationally challenging.

Joint work with Alice Davis, Axel Albrecht, Heike Puhlmann, Stefan Meining, Simon Wood and Karim Anaya-Izquierdo.

Jan Beran (University of Konstanz)

On aggregation in networks

Statistical analysis of networks is often based on aggregated series where aggregation is defined via routing matrices. Granger (1980) showed that aggregation of short-memory processes can imply long-range dependence. On the other hand, OD-flows (origin destination flows) often exhibit long memory. Thus, routing of OD-flows leads to cross-sectional aggregation of strongly dependent series. Asymptotically, dependence increases substantially, transforming a hyperbolic decay

of autocorrelations to a slowly varying rate. This makes statistical inference highly uncertain. The situation changes, when time-dependent aggregation is applied. Suitably chosen time-dependent routing schemes can preserve a hyperbolic rate or even eliminate autocorrelations completely. This is joint work with Haiyan Liu and Sucharita Ghosh.

Luisa Cutillo (University of Leeds)

An automated spectral clustering: theory and applications

Background: Data clustering can be reformulated in terms of a graph partitioning problem where the given set of data is represented as a graph. In this context, eigenvectors of the graph Laplacian are often used to obtain a new geometric representation of the original data set, which generally enhances cluster properties and improves cluster detection.

Graph Laplacian is a semi-definite positive matrix associated to a general graph, whose spectrum is widely used for analyzing graphs, in particular in graph partitioning and clustering.

Methodology: In this work, we apply a bootstrap Algebraic MultiGrid (AMG) method, which constructs a set of vectors associated with the graph Laplacian. These vectors, referred to as algebraically smooth ones, span a low dimensional Euclidean space, which we use to represent the data for cluster detection. We show how the computed smooth vectors employed in the construction of an AMG operator accurately approximate the space in the lower portion of the spectrum of a preconditioned version of the graph Laplacian.

Application and conclusion: We discuss the results of the application of our approach on simulated and real data, in particular we explore data from single-cell RNA sequencing (scRNA-seq). The results are very promising when dealing with networks with medium and high modularities. We show that our method can serve as valid alternative to modularity-based methods, when there is no “ground truth”. Indeed it exploits theory of AMG methods to decide how many smooth vectors have to be computed to well represent the data, while for the more standard spectral clustering it is not as clear how many eigenvectors to choose in advance in order to get satisfactory clustering results. We finally observe that our method has a linear computational complexity and can be very competitive for large-scale data sets.

Joint work with P. D’ambra, IAC-National Research Council, Italy and P. Vassivleski, Portland State University USA.

Sucharita Ghosh (Swiss Federal Research Institute WSL)

On trend estimation and related issues under Gaussian subordination

Time series data where the marginal distribution may assume arbitrary shapes over time have important consequences in many fields of research. These affect how the quantiles change, thus affecting future adaptation, planning etc. Consider the nonparametric regression model $Y_i = m(t_i) + e_i, i = 1,2, \dots, n, t_i = \frac{i}{n}$, with trend $\mathrm{m(t)} = \mathrm{E}(\mathrm{Y}_{[\mathrm{nt}]}),\ \mathrm{t} \in (0,1)$, and finite variance. Let the centered observations $\mathrm{e_i} = \mathrm{Y_i} - \mathrm{m(t_i)}$ be subordinated to a latent stationary Gaussian process $\mathrm{Z_i}$ via an unknown transformation $\mathrm{e_i} = \mathrm{G}(\mathrm{Z_i}, \mathrm{t_i})$. This means, $\mathrm{e_i}$ need not be Gaussian and $\mathrm{F}(\mathrm{x}, \mathrm{t}) = \mathrm{P}(\mathrm{e}_{[\mathrm{nt}]} < \mathrm{x})$, where $\mathrm{x} \in R$, may be an arbitrary function of time t. For this setup, we consider estimation of m and F using kernel smoothing. To motivate discussions, we consider some time series data from climate change research and illustrate estimation of rapid change points, time varying regression slopes and time-dependent distribution functions. In the sequel, we explore some strategies for bandwidth selection as well as discuss some theoretical properties of the estimators.

John Kent (University of Leeds)

The space environment for satellites orbiting the earth

There are currently about 2000 operational satellites orbiting the earth. The subject of space situational awareness deals with various hazards to these satellites ranging space weather to space debris. There are estimated to be over 30000 pieces of space debris and inactive satellites in orbit bigger than a grapefruit, which can be observed from earth. In this talk I will describe some recent work funded by the US Air Force to develop fast and accurate improved statistical methods to predict the path of the debris so that it can be avoided by active spacecraft. The methodology uses ideas from Kalman filtering, directional statistics and multivariate analysis.

Chuoxin Ma (University of Cambridge)

Multi-state Models for Multi-type Recurrent Events and Terminal Events with Feedback and Measurement Errors in Biomarkers

In cardiovascular disease study, one of the main interest is to investigate the association between risk factors and a series of multi-type recurrent events and terminal events, such as myocardial infarction, stroke and cardiovascular death. When the trajectories of some biomarkers contain past event feedbacks, existing approaches handling time-dependent covariates in event history analysis can be problematic. We propose a class of multi-state models for the analysis of multi-type recurrent events and terminal events when biomarkers contain past event feedback and are intermittently observed and subject to measurement errors. The competing risk structure and the progressive nature of the multiple events can be well captured

by state-specific intensity functions. Both time-varying and constant coefficients can be accommodated. Estimation procedure based on polynomial splines approximation and an extension to the corrected score approach is developed. The consistency and asymptotic normality of the proposed estimators are provided. Simulation shows that the naive estimators which either ignore the past event feedback or the measurement errors are biased. Our method achieves better coverage probability of the time-varying/constant coefficients, compared to the naive methods. An application to the data set from the Atherosclerosis Risk in Communities Study is presented.

Contributed Talks

Jacob Cancino-Romero (University of Leeds)

Causal joint models for the relationship between frailty, recurrent falls and mortality in the elderly

Joint modelling of longitudinal and time-to-event data is often used to explore the joint distribution of multiple outcomes, although fitting joint models is computationally challenging. This strategy has been an area of active research for description and prediction, but causal inference has received less attention. In this work we investigate the causal relationships between three outcomes in the elderly: geriatric frailty, recurrent falls and mortality. We formulate two causal diagrams and their corresponding statistical joint model with the aim of determining if the effect of frailty on mortality differs between the two alternatives: (1) a shared frailty joint model where the three outcomes are linked by random effects, and (2) a joint model for frailty and mortality treating frailty as an endogenous time-varying covariate subject to measurement error and falls as an exogenous time-varying covariate for mortality. Our contributions are the inclusion of exogenous time-varying covariates to the analysis and the comparison between two alternative causal joint models.

This work is motivated by the Yorkshire & Humber Community Ageing Research 75+ study (n=282) conducted in Northern England. The aim of the study is to understand the causal connections between frailty, recurrent falls and mortality in order to guide the early detection of treatable problems and intervene with the goals of improving quality of life and reducing the risk of mortality. Results of model (1) suggest effects of frailty on recurrent falls (p = 0:0004) and mortality (p = 0:03), but no effect of recurrent falls on mortality.

Joint work with Stuart Barber, Leonid V. Bogachev and Jeanne Houwing-Duistermaat.

Rachel Carrington (University of Nottingham)

Dynamic Word Embeddings Using Large Text Corpora

Word embeddings are a way of modelling language by representing words as low-dimensional vectors, which aim to capture some of the semantic information present in the data. There is a particular interest in developing word embeddings which have a temporal component, using a corpus in which each document or context is indexed by time. The goal of this is to discover trends in qualitative properties of language across time, for example detecting words that have changed in meaning or usage over the last few centuries.

Most current approaches to this problem use nonparametric models. Given the high dimensionality of the data, this often leads to overfitting, meaning that the results are

difficult to interpret or to draw inference from, and it is not always clear what the parameters represent. In addition, most word embedding models are non-identifiable -- that is, there exist multiple different sets of embeddings which satisfy the objective function of the model equally well. Most time-dependent methods rely on dividing the data up into time periods, with word embeddings being generated independently for each time period. Hence this non-identifiability causes problems when comparing word embeddings generated for different time periods, and although some attention has been given to this issue, it is often not fully addressed.

We present a parametric approach to this problem, based on low-rank matrix factorization, the properties of which are well understood. We show how the parameters of this model can be made precisely identifiable by imposing on them a set of conditions. We show how our model is related to the standard factor model, which allows us to use statistical properties of factor analysis. In particular, testing for time dependency in the model is equivalent to testing for the adequacy of the factor model. This allows us to detect words that may have significantly changed their meaning over time. We anticipate that the method developed should also have applicability for dealing with other forms of data, besides text, which is high-dimensional and has a time component.

Said el Bouhaddani (Utrecht University)

Probabilistic integration of omics data with PO2PLS

A key challenge in systems biology is understanding how several levels of omics data (e.g. genomics, glycomics, regulomics) interact. These data are characterized by high-dimensional measurements that are strongly correlated within and between datasets. Analysing all data simultaneously can provide more insight in the relationship between these data. Such approach is often termed “omics data integration”. We consider two examples: 300k genetic variants and 20 IgG1 glycomic measurements from the Korcula population cohort (885 participants), and 30k ChIP and 15k RNA seq measurements from 13 hypertrophic cardiomyopathy patients and 10 controls.

Traditionally, univariate association analyses are applied to analyse pairs of variables. These approaches ignore interaction effects and are computationally expensive. In contrast, we focus on multivariate latent variable models that can handle high dimensional data, while estimating a relationship in a low-dimensional latent space. Such methods should consider heterogeneity in data-specific characteristics as dimensionality, distribution and measurement platform. To take this into account, the O2PLS method has been developed that decomposes the data into joint, data-specific and residual components. The joint components are linear combinations of variables that best represent the relationship between the datasets.

However, these methods lack identifiable parameters and a probabilistic framework, which hampers interpretation of the results.

We propose Probabilistic O2PLS (PO2PLS), a latent variable model to estimate joint and data-specific components in two datasets. The probabilistic framework gives more insight in the model than the algorithmic O2PLS formulation, and facilitates inference. The PO2PLS model is identifiable under standard constraints on the parameters and an Expectation Conditional Maximisation algorithm is developed to obtain maximum likelihood estimates.

The PO2PLS performance is evaluated in terms of interpretation and prediction using a simulation study; PO2PLS is shown to outperform competing methods, especially when the noise level is high and the sample size is low. Finally, PO2PLS is applied to genetic and glycomic, and ChIP seq and RNA seq datasets. We find that the joint genetic-glycomic components reflect enzymatic reactions in glycosylation and the corresponding genes appear to be biologically relevant. Also, the ChIP-RNA seq joint components are biologically interpretable and can completely discriminate cases from controls, demonstrating the potential of PO2PLS in omics data integration.

Joint work with Hae Won Uh, Geurt Jongbloed and Jeanine Houwing-Duistermaat.

Agus Salim (La Trobe University)

DECENT: Differential Expression with Capture Efficiency adjustmeNT for single-cell RNA-seq data

Due to limitation in the molecule capture technology, molecule dropout is a common phenomenon in single-cell RNA-seq (scRNA-seq) data, and when left unaddressed it affects the validity of the statistical analyses. Despite this, few current methods for differential expression (DE) analysis of scRNA-seq data explicitly model the process that gives rise to the dropout events. We develop DECENT, a method for differential expression (DE) analysis of scRNA-seq data that explicitly and accurately models the molecule capture process in scRNA-seq experiments using a Beta-Binomial capture model that allows separating the variations due to biological conditions of interest from the technical variations. We show that DECENT demonstrates improved DE performance over existing DE methods that do not explicitly model dropout. This improvement is consistently observed across several public scRNA-seq datasets generated using different technological platforms. The gain in improvement is especially large when the capture process is overdispersed. DECENT maintains type I error well while achieving better sensitivity. Its performance without spike-ins is almost as good as when spike-ins are used to calibrate the capture model.

Joint work with Chengzhong Ye and Terence P Speed.

Posters

Jason Anquandah (University of Leeds)

Optimal Stopping in a Simple Model of Unemployment Insurance

Managing unemployment is one of the key issues in social policies. Unemployment insurance schemes are designed to cushion the financial and morale blow of loss of job but also to encourage the unemployed to seek new jobs more pro-actively due to the continuous reduction of benefit payments. A simple model of unemployment insurance is proposed with a focus on optimality of the individual's entry to the scheme. The corresponding optimal stopping problem is solved. The model and it solution are useful by illustrating approaches to optimal strategies of an individual seeking to get insured.

Ariadni Aravani (University of Leeds)

Flexible parametric survival analysis: Funnel plots from flexible parametric models to predict colorectal cancer survival in England

Colorectal cancer is the 4th most common cancer in the UK with significantly lower survival than in many other comparable countries. As a result, significant efforts are being made to understand the survival disparities and data are required comparing clinical outcomes between the multidisciplinary teams (MDT) that manage the disease within the National Health Service (NHS). Up-to-date relevant statistical methodologies are needed to robustly analyse the complex datasets available, to take into account the differences in characteristics, management and workload across MDT populations, handle missing data related to important factors such as stage of disease and produce robust and meaningful comparisons of care.

Flexible parametric approach was used to estimate survival of patients diagnosed with colorectal cancer. These models allow more complex modelling of relative survival than the more traditional Cox proportional hazards (PH) or other parametric models. They allow more accurate prediction and modelling of excess mortality alongside the ability to adjust for covariates and time-varying effects. The methods were compared using data from National Cancer Registration and Analysis Service (NCRAS). Additionally, other methods used, like multiple imputation (to impute missing stage information) and funnel plots (graphs used to identify outlier performance) will be presented.

Joint work with John Taylor and Eva Morris.

Iva Budimir (University of Bologna)

A stochastic neutral model for the gene lengths distribution

The omnipresence of long-tailed distributions in biology motivates the comprehensive study of the mechanisms that generate this type of distribution. This problem has been discussed in the context of ecological theories, starting with seminal studies by Fisher, Preston and others, and is known as Relative Species Abundance (RSA). One of the models proposed by Engen and Lande in 1996 describes a stochastic neutral model which generates Poisson-lognormal distribution. The approach used here is related to the so-called neutral theory that allows, by neglecting the interactions, to write a one-dimensional master equation and to compute its stationary solution. An interesting example of the long-tailed distribution is the distribution of gene lengths which we described in the context of population dynamics and fitted with Poisson-lognormal distribution.

To make an analogy between a genome and an ecosystem, we identified individuals with nucleotide bases and species with genes. In this way, gene length can be viewed as a number of individuals in a species (abundance) and modelled with RSA. Gene lengths data for more than 100 different species were obtained from Ensemble and fitted with Poisson-lognormal distribution. The fitting was performed using the Bayesian approach and MCMC (Monte Carlo Markov Chain) methods. Since Poisson-lognormal distribution doesn't have closed-form likelihood function, approximation developed by Bulmer (1974) was implemented in Python. We compared our fitting method with other standard distributions used for RSA models such as negative binomial. Poisson-lognormal showed to be the best fit thus supporting our theoretical model.

Joint work with Gastone Castellani, Claudia Sala and Enrico Giampieri.

Cheng Cai (University of Leeds)

Optimal hedging for American options with a single transaction

In the Black-Scholes model, an option seller constructs a self-financing stock-bond portfolio in order to Delta hedge a short position in a perpetual American put option. In contrast to the continuous trading of the Black-Scholes model, the option seller can only rebalance her portfolio once before the time $\tau_a \wedge \tau_b$ at which the underlying stock price S_t leaves an interval (a, b). Here the lower end-point a is the optimal exercise price of the put option and b(> a) is chosen arbitrarily. The goal is to determine the optimal time to rebalance the portfolio and the optimal hedge ratio (stock holding after the trade) that minimize the variance of the so-called tracking error (at time $\tau_a \wedge \tau_b$). First, we formulate the optimal hedging problem for a fixed initial stock holding as a one dimensional optimal stopping problem. This is solved by constructing three different free boundary problems depending on possible parameter choices. Second, we study analytically how the stopping boundaries

move in response to variations in the initial stock holding. Finally, we obtain an equation that must be satisfied by an optimal initial stock holding.

Samantha Crossfield (University of Leeds)

Variation in Methods, Results and Reporting in Electronic Health Record-based Studies Evaluating Routine Care in Gout

Objectives

A growing number of electronic health record (EHR)-based studies are being published, with wide variation in the methods, results, reporting and risk of bias. We performed a systematic review examining this in EHR-based studies evaluating management of a common musculoskeletal disease, gout.

Methods

Two reviewers systematically searched six literature databases and Google Scholar for all EHR-based studies published by February 2019 investigating gout pharmacological treatment. Information was extracted on study design, eligibility criteria, definitions, medication usage, efficacy and safety, comprehensiveness of reporting (RECORD), and Cochrane risk of bias (registered PROSPERO CRD42017065195).

Results

We screened 5,603 titles/abstracts, 613 full-text articles and selected 75 studies including 1.9M gout patients. Gout diagnosis was defined in 26 ways across the studies, most commonly using a single diagnostic code (n=31, 41.3%). 48.4% did not specify a disease-free period before ‘incident’ diagnosis. Studies with more stringent gout definitions reported higher medication use. Medication use was suboptimal and reported variably: 33 (44.0%) reported dosage and only 7 (9.3%) considered continuous or cumulative exposure. Effectiveness and safety were reported in 13 (17.3%) and 18 (24.0%) studies, with results being broadly similar despite variability in eligibility criteria. Comprehensiveness of reporting varied from 73% (55/75) appropriately discussing the limitations of EHR data use to 5% (4/75) reporting on key data cleaning steps. Risk of bias was generally low.

Conclusion

In EHR-based gout studies there was wide variation in case-definitions and medication-related analysis, which influenced reported medication use. Improvements in reporting and consideration of comprehensive assessment of EHR-pertinent biases are required.

Joint work with Lana Yin Hui Lai, Sarah R Kingsbury, Philip G. Conaghan and Mar Pujades-Rodriguez.

Zhujie Gu (Utrecht University)

Variable selection in integration of omics data using sparse O2PLS

Integration of multiple omics data to investigate the underlying system is an important topic in modern biostatistics. Various methods have been proposed for this task, in particular, PLS-related methods, which model the joint variation of two datasets, X and Y. One challenge is that each dataset has its specific characteristics. Ignoring this heterogeneity between datasets complicates the identification of the joint parts and makes interpretation difficult. To solve this problem, Two-way Orthogonal Partial Least Squares (O2PLS) was proposed, which decomposes both datasets into joint, specific and noise parts (Trygg 2003, Bouhaddani 2016). The joint latent variables in O2PLS are linear combinations of all observed variables. For interpretation, the variables with highest weights are investigated. Here, an arbitrary cut-off is usually chosen. In order to have a data-driven feature selection procedure, we propose sparse O2PLS (SpO2PLS).

O2PLS relies on the NIPALS algorithm (Wold 1975) to find the linear projections of the variables that yield maximal covariance in the joint space by iteratively regressing Y and X on projected X and Y, respectively. In SpO2PLS, we impose an L1 penalty on the weights of the projections in each regression step. Therefore, a sparse solution is obtained by retaining only features with a large contribution to the covariance. It can be shown that this is equivalent to applying soft-thresholding to the weights at each NIPALS step.

A simulation study is performed to evaluate the performance of SpO2PLS in terms of sparsity level (number of related features), accuracy (true positive rate & positive predictive value) and prediction performance (prediction error between joint latent variables of X and Y. The results show that SpO2PLS is able to determine the sparsity level correctly. In particular, SpO2PLS has higher accuracy and lower prediction error on an independent test set than O2PLS in most cases. O2PLS and SpO2PLS are applied to ChIP-seq (X) and RNA-seq (Y) data, measured in heart tissues from 13 hypertrophic cardiomyopathy patients and 10 controls. The top genes and regulatory regions are selected and interpretation of the results will be given.

Joint work with Said el Bouhaddani, Magdalena Harakalova, Jeanine Houwing-Duistermaat and Hae-Won Uh.

Nishant Ravikumar (University of Leeds)

Multi-dimensional registration for computer aided diagnosis and interventions

Multi-dimensional image registration is useful for a variety of computer aided diagnosis and intervention applications. Such approaches look to leverage the discriminative information contained within high-dimensional descriptors, for accurate pair-wise or group-wise registration of images. To this end, we formulate a

probabilistic registration framework using a hybrid mixture model, which is flexible and applicable to a variety of tasks in medical image analysis. The following use cases demonstrate the efficacy of the proposed approach: Group/population-wise analysis of DTI-derived data, and vessel-driven image registration for intraoperative brain shift compensation.