STOR-i
excellence with impact

Lancaster University

STOR-i Workshop 2015, 8th-9th January

Itinerary:

Day 1, Thursday 8th January

10:00-16:30 - Management School (Lecture Theatre 1 and breakout space, Building 52 on campus map)
(18:30-21:00 - Evening poster presentation in the LICA building, Building 4 on campus map)

Registration, refreshments and lunch will take place in the breakout space (opposite Management School reception). Talks will be held in Management School LT 1.

10:00- Registration and refreshments in the breakout space, Management School (External speakers, external attendees and staff only)
10:30- Arnoldo Frigessi, Statistics for Innovation, University of Oslo
11:30- Kaylea Haynes, STOR-i PhD student
12:00- Lunch
13:00- Ian Dryden, School of Mathematical Sciences, University of Nottingham
14:00- Hugo Winter, STOR-i PhD student
14:30- Refreshment break
15:00- Emanuele Ragnoli, IBM Research
15:30- Horst W. Hamacher, Department of Mathematics, Technische Universitat Kaiserslautern
16:30- Talks finish
18:30-21:00 Poster Session in the LICA Building with wine reception and buffet

Day 2, Friday 9th January

9:30-13:30- Management School (Building 52 on campus map, breakout space and Lecture Theatre 1)

Refreshments and lunch will be served in the breakout space (opposite Management School reception). Talks will be held in Management School LT 1.

09:30- Kyle Lin, Naval Postgraduate School
10:30- Jamie Fairbrother, STOR-i PhD student
11:00- Refreshment break
11:30- Ralph Mansson, DSTL
12:00- Christian Rohrbeck, STOR-i PhD student
12:30- Lunch
13:30- Close

Abstracts

Day 1

Bayesian Inference from Rank Data
Arnoldo Frigessi, Statistics for Innovation, University of Oslo

Modeling and analysis of rank data has received renewed interest in the era of big data, when recruited or volunteer assessors compare and rank objects to facilitate decision making in disparate areas, from politics to entertainment, from education to marketing. The Mallows rank model is among the most successful approaches, but for computational convenience its use has been limited to a particular form based on the Kendall distance. We develop computationally tractable methods for Bayesian inference in Mallows models with any right-invariant metric, allowing greatly extended flexibility. Our method allows inference on the consensus rankings of the considered items, also when based on data provided in the form of partial rankings, such as top-t or pairwise comparisons. If the assumption of an underlying common true ranking for all assessors is unrealistic, we can find by means of suitable clustering more homogeneous subgroups and consider consensus rankings within each. Our method allows making probabilistic predictions on the classification of assessors based on the ranking of some items, and on individual preferences based only on partial information. Finally, we construct a regression framework for ranks which vary over time. The performance of the approach is studied using several experimental and benchmark datasets, and on simulated data. This is work with Øystein Sørensen, Valeria Vitelli and Elja Arjas.

Efficient penalty search for multiple changepoint problems
Kaylea Haynes, STOR-i PhD student, Lancaster University

Advances in technology has allowed us to record and store more data than ever before in our day to day lives. The Big Data revolution has created many challenges within a number of different inference areas, one of which being changepoint detection. The main aim of this work was to develop an efficient algorithm for changepoint detection which scales to large data sets.

In the multiple changepoint setting, established search methods, which involve optimising either a constrained or penalised cost function over possible number and locations of changepoints using dynamic programming, can be computationally expensive. Recent work on pruning the penalised optimisation problem gives, under certain conditions, an improved computational cost which is linear in the number of data points. The main challenge with this method is it requires a penalty value to avoid under/over-fitting the model.

The talk will discuss our method which solves the changepoint detection problem for a range of penalty values instead of arbitrarily choosing one value. I will show that this method has a computational cost which is linear in the number of data points and linear in the difference between the optimal segmentations for the smallest and largest penalty values. Additionally I will show that this is quicker than solving the constrained optimisation problem for a range of segmentations. Finally I will show that in some situations the common penalty values in the literature can be sensitive to model specification whereas solving for a range of penalty values allows us to find values which result in reasonable segmentations.

Penalized Euclidean Distance Regression

Ian Dryden, School of Mathematical Sciences, University of Nottingham

We consider an alternative method for variable screening, variable selection and prediction in linear regression problems where the number of predictors can be much larger than the number of observations. The method involves minimizing a penalized Euclidean distance, where the penalty is the geometric mean of the l_1 and l_2 norms of the regression coefficients. This particular formulation exhibits a grouping effect, which is useful for screening out predictors in higher or ultra-high dimensional problems. Also, an important result is a signal recovery theorem, which does not require an estimate of the noise standard deviation. Practical performances of variable selection and prediction are evaluated through simulation studies and the analysis of a dataset of mass spectrometry scans from melanoma patients, where excellent predictive performance is obtained. This is joint work with Daniel Vasiliu (College of William & Mary) and Tanujit Dey (Cleveland Clinic).

Detecting changing behaviour of heatwaves with climate change

Hugo Winter, STOR-i PhD student

Extreme value models that can account for dependence are vital when modelling rare weather phenomena that can cause large damages and loss of life. An important question in the climate community regards how characteristics of future rare events might change with human induced climate change. In a worst case scenario, stronger and longer events coupled with an expanding global population could lead to greater problems in the future. In this work we shall focus on analysing temperature data from the Met Office's HadGEM2 global climate model for central France. The GCM is forced with climate forcing parameters consistent with the A2 climate change scenario which produces a rapid increase in global temperatures. The aim is to better understand how the duration and intensity of future heatwaves will change in a changing climate. For this purpose we develop methods for dealing with non-stationarity in our extreme value models by incorporating covariates.

Distributionally Robust Optimization

Emanuele Ragnoli, IBM Research

In many real-world optimization problems, one faces the dual challenge of hard nonlinear functions in both objective and constraints and uncertainty in some of the problem parameters. Often, samples for each uncertain parameter are given, whereas its actual distribution is unknown. We propose a novel approach for constructing distributionally robust counterparts of a broad class of polynomial optimization problems. The approach aims to use the given samples, not only to approximate the support of the unknown distribution or the first and second order moments, but also its density. We show that polynomial optimization problems with distributional uncertainty sets defined via density estimates are particular instances of the generalized problem of moments with polynomial data and employ Lasserre's hierarchy of SDP relaxations to approximate the distributionally robust solutions. As a result of using distributional uncertainty sets, we obtain a less conservative solution than classical robust optimization.

Operations Research Models in Evacuation Planning

Horst W. Hamacher, Department of Mathematics, Technische Universitat Kaiserslautern

Due to the variety of problems which need to be tackled, evacuation planning is an excellent field for the development of theory and implementations of Operations Research (OR) models. In this presentation we will focus on the fields of network flows and locational analysis.

Using dynamic network flow models, evacuees are associated with flow units which are sent over time from their homes (sources) to evacuation shelters (sinks). Objective values of this model help to predict evacuation parameters, for instance, the time to evacuate all evacuees, the minimization of the risk, or the determination of the main evacuation routes as input of subsequent simulations. Several location decisions, like the choice of shelters or the placement of the emergency units, influence the outcome of the network flow models. In this presentation we show how these decisions can be integrated in the network flow model.

In the **FlowLoc** model we consider the following situation: A decision on locating one or more facilities in a network has to be made which changes the capacities of the arcs. The quality of the location decision is measured by the change in the optimal objective value before and after the location of the new facilities. Using the **SinkLoc** model, we show how suitable evacuation shelters can be chosen for a given number of evacuees. This is achieved by

considering the number of evacuees as supplies in source nodes and finding a best possible set of locations as sinks which can take on all evacuees. We propose solution algorithms for FlowLoc and SinkLoc problems, discuss their theoretical complexity and present results of numerical tests.

References: H.W. Hamacher and S.A. Tjandra, Mathematical modeling of evacuation problems: a state of the art, In: *Pedestrian and Evacuation Dynamics*, Springer (2002), 227-266.

H.W. Hamacher, S. Heller, and B. Rupp, Flow location (FlowLoc) problems: dynamic network flows and location models for evacuation planning, Annals of Operations Research, (2013), 207, 161-180

P. Heßler and H.W. Hamacher, Sink Location to Find Optimal Shelters in Evacuation
Planning, Reports in Wirtschaftsmathematik (2014), University of Kaiserslautern, submitted

Day 2

Optimal Patrol on a Graph

Kyle Lin, Naval Postgraduate School

Abstract: Consider a patrol problem, where a patroller traverses a graph through edges to detect potential attacks at nodes. To design a patrol policy, the patroller needs to take into account not only the graph structure, but also the various attributes at each node, including the time needed to complete an attack, the probability of detecting an ongoing attack, and the consequence of a successful attack. The goal is to determine a patrol policy that minimizes the expected cost incurred when, and if, an attack eventually happens. Although it is possible to formulate a linear program to compute the optimal policy, such computation quickly becomes intractable as the problem size grows. Our main contribution is to develop index-based heuristic policies that typically achieve within 1% of optimality with computation time orders of magnitude less than what is required to compute the optimal policy.

Joint work with Michael Atkinson and Kevin Glazebrook.

Scenario generation for stochastic programs with tail risk measure - an application to portfolio selection

Jamie Fairbrother, STOR-i PhD student

Stochastic programming is a tool for making decisions in the presence of uncertainty. Its distinguishing feature is that it can explicitly model future decisions and costs based on investment decisions and realisations of a priori unknown parameters (such as demand or price of a product). This flexibility comes at a price: stochastic programs tend to only be tractable for problems where uncertain parameters are modelled by a finite number of possible future scenarios. How we generate these scenarios plays a key role on the quality of the solution a stochastic program yields.

The mean-risk approach in stochastic programming is to choose a decision which somehow balances expected profit against the risk of some investment. Tail risk measures are an important class of risk measures as they give one an idea of how much capital should be held to cover the most extreme losses. However, these are problematic as they typically only depend on a fraction of the scenarios we generate for a problem. This means that a scenario generation method will usually yield very unstable solutions unless we use a large and computationally expensive number of scenarios to model our uncertainty.

In this work we argue that we can gain better solutions with fewer scenarios by concentrating the scenarios in an area which we call the risk region. In particular, we characterise this region exactly for a class of portfolio selection problems, and we demonstrate numerically the improvements of our methodology over standard scenario generation methods for this problem.

Challenges with Big Data Analytics

Ralph Mansson, DSTL

Big Data provides challenges and opportunities to the Statistics and Computer Science communities. These range from storage and transmission of data to application of algorithms originally developed for smaller data sets. Parallel computing offers a solution in cases where the algorithms can be easily expressed in a form for divide and re-combine operations. In this talk we will consider some of these issues and discuss data analytics using open source data.

An approach for using monotonic regression in discrete spatial statistics

Christian Rohrbeck, STOR-i PhD student

Monotonic Regression is considered in both Statistics and Operational Research and has a wide range of application areas including economics and genetics. Statistical methodology has been developed in functional data analysis, Bayesian statistics, etc. and addresses several problems, e.g. high-dimensional regression. However, there is little research on its application to discrete spatial statistics. Such an approach should consider that the functional shape may vary spatially but with similarities of neighbouring spatial entities.

In this talk, we introduce a Bayesian method which fulfils this property and aims to apply monotonic regression to modelling problems in discrete spatial statistics. The method defines a Bayesian framework which accounts for potential similarities in the monotonic functions of neighbouring regions and allows for a high flexibility of the estimated functions. The model is fitted using a reversible jump Markov Chain Monte Carlo algorithm and cross-validation. Finally, we discuss the effectiveness of our approach by considering simulated data.