STOR-i Conference: 10th – 11th January 2019

Abstracts

Day 1

Modelling evolution in a spatial continuum

Alison Etheridge, Department of Statistics, University of Oxford

Since the pioneering work of Fisher, Haldane and Wright at the beginning of the 20th Century, mathematics has played a central role in theoretical population genetics. In turn, population genetics has provided the motivation both for important classes of probabilistic models, such as coalescent processes, and for deterministic models, such as the celebrated Fisher-KPP equation. Whereas coalescent models capture 'relatedness' between genes, the Fisher KPP equation captures something of the interaction between natural selection and spatial structure. What has proved to be remarkably difficult is to combine the two, at least in the biologically relevant setting of a two-dimensional spatial continuum

In this talk we describe some of the challenges of modelling evolution in a spatial continuum and then, as time permits, turn to some results concerning the interplay between natural selection and spatial structure.

Modelling Hypergraphs using a Latent Space Representation

Kathryn Turnbull, STOR-i PhD student

A range of models have been proposed for understanding complex interactions between a set of objects of interest. Typically these interactions are assumed to be pairwise, however this is not the case in many real world network datasets. For example, multiple authors may collaborate on a single paper or several people may occur simultaneously in a photograph. The statistical literature on modelling such higher-order interactions is relatively sparse and in this talk we will introduce and discuss a modelling framework for data of this type. Our approach will assume that nodes of the network can be represented in a low dimensional latent space and that the presence of an edge depends on the relative latent positions of each node.

Data Driven Product Development

Lisa Turner, Lubrizol (STOR-i Alumni)

Lubrizol are a global, market-driven chemical speciality company. Although you might have never heard of us, our products are likely in the things you use daily; whether that's through the car you drive, the shampoo you use or the clothes you wear. I will describe how data science plays an increasingly key role in product development at Lubrizol through examples of projects I have worked on. The level of impact and success some of those projects are having is due to collaboration between Lubrizol and STOR-i.

Using data science to build better digital products supporting researchers

Elisabeth Ling, Elsevier

Within the Research Products team at Elsevier, we aim to develop digital products to support researchers in their workflows. From reading recommendations to suggestions to invite to peer-review, data science helps us build great websites. It is highly rewarding to observe users getting a lot out of the features we launch. We work in multi-disciplinary teams, bringing together product managers, software engineers, user experience designers, data scientists and product analysts. It can be very demanding for each of these specialists to discover together the "sweet spot" between technical perfection and time to market. Rapidly building prototypes, iterating, wrestling with imperfect data and solving challenging technical problems are some of the challenges we face. A few examples drawn from 20 years in the trenches of digital product management…

STOR-i
excellence with impact

Lancaster University

Optimizing Prioritized and Nested Solutions

David Morton, Department of Industrial Engineering and Management Sciences, Northwestern University

A typical optimization model in operations research allocates limited resources among competing activities to derive an optimal portfolio of activities. In contrast, practitioners often form a rank-ordered list of activities, and select those with highest priority, at least when choosing an activity is a yes-no decision. Ranking schemes that score activities individually are known to be inferior. So, we describe a class of two-stage stochastic integer programs that accounts for structural and stochastic dependencies across activities and constructs an optimized priority list. We further discuss a class of optimization models, subject to a single "budget" constraint, that naturally leads to a family of optimal nested solutions at certain budget increments. Several applications both motivate the approach and illustrate results, ranging from a stochastic facility location model to a hierarchical graph clustering problem.

Meaningful data at scale: opportunities for delivering better treatments to the right patients faster and more safely

Chris Harbron, Roche

The pharmaceutical industry now has a range of multiple types and sources of data from genomic profiling to electronic health records to wearable devices, and at scales far greater than ever before to aid its objective of providing life changing medicines to patients. This growth of data and the application of analytic techniques is increasing the role of quantitative sciences within the pharmaceutical industry. This talk will explore some of these new opportunities and approaches, with a reminder that the basic underlying concepts of experimental design and data quality are still critical for meaningful analyses.

Too much environmental data or not enough?

E Marian Scott, School of Mathematics and Statistics, University of Glasgow

As statisticians, we value and learn from the data we observe. There have been substantial changes in the nature of data, with unprecedented growth in data availability and volume (as well as the speed with which they are generated). Our ability to measure almost every aspect of our daily lives, is linked to security and privacy concerns as well as economic aspects related to who benefits from our data.

In the environmental sciences communities, changes in technology have similarly seen a growth in data availability, and “the data deluge” is a phrase sometimes used. Deluge here “is the imminent flood of scientific data expected from the next generation of experiments, simulations, sensors and satellites” (Hay and Trefethen, 2003 in Grid Computing: Making the Global Infrastructure a Reality). Deluge suggest flood, so are too much data a possibility? Baraniuk wrote, in 2011, “The data deluge is changing the operating environment of many sensing systems from data-poor to data-rich—so data-rich that we are in jeopardy of being overwhelmed.”

In this presentation, using some environmental case studies, I will reflect on whether we can ever have too much environmental data.

Day 2

Last Mile Logistics

Arne Strauss, Warwick Business School

Last-mile logistics providers are facing a tough challenge in making their operations sustainable in the face of growing customer expectations to further decrease lead times to same-day or even same-hour deliveries, and/or to offer narrow delivery time windows. The providers respond to this challenge by investing in their analytic capabilities to make their last mile logistics are efficient and intelligent as possible.

In addition, innovative and disruptive business models are currently on trial, e.g. asset-lean start-ups use crowdsourced drivers or drivers on demand-dependent contracts. Several companies are experimenting with delivery drones or robots, and how to collaborate with each other (`shared economy').

Many of these developments entail exciting new challenges for operations researchers. In this talk, I will review some of the most recent developments and reflect on future research directions.

Why did the chicken affect my insurance premium?

Shreena Patel, dunnhumby (STOR-i Alumni)

Dunnhumby is a customer data science platform which works with retailers around the world to deliver exceptional customer experiences, personalised to their needs and expectations. The Tesco Bank team achieves this with Clubcard data, helping the bank to make better decisions concerning the services it offers to Tesco customers. The partnership offers many interesting applications of machine learning, including credit scoring, risk pricing and trigger-based marketing campaigns. In this talk, we will focus on the second of these cases: the use of shopper behaviour for insurance pricing. Two of the key challenges in this work centred around achieving model stability and developing the model independently, as an input to a 'meta model', whilst avoiding collinearity. I will take you through our key learnings from the project.

Forecasting the French electricity consumption at different time steps and at different scales

Audrey Lagache, EDF

Since electricity is not easily stored, EDF needs efficient consumption forecasting tools to maintain a balance between consumption and production. EDF's R & D consumption forecast group aims to set up forecasting models for various entities of the EDF group: the sales department, the department in charge of controlling production resources and Enedis, the distribution network manager. To meet their needs we set up models that can predict the national consumption therefore a single curve or consumption at local mesh of the network and in this case we seek to predict hundreds or even thousands of curves.

The models currently used to predict the electricity consumption are the generalized additive models (GAM). These models are semi-parametric models whose principle consists in explaining the variable to be predicted (here the electricity consumption) as a sum of regular functions of different explanatory variables. These regular functions are estimated by penalized linear regression after projection on a spline basis.

We are also exploring in recent years a wide range of methods of machine learning (boosting, random forests, deep learning…) and aggregation of experts in order to provide the best possible forecast in a context where the arrival of new uses such as electric vehicles can deform the load curve.

The increase in decentralized production in France also causes a greater difficulty to better size the network and to manage the balance between consumption and production. We are therefore working on the implementation of probabilistic forecasts in modelling the quantiles of electricity consumption.

STOR-i
excellence with impact

Lancaster University

Title: Speed or Accuracy? Optimal Search with Fast and Slow Speeds

Jake Clarkson, STOR-i PhD Student

A hidden object needs to be found in many real-life situations, some of which involve large costs and significant consequences with failure. Therefore, efficient search methods are paramount. Further, there is often a choice regarding the speed of the search. Area can be covered slowly, or quickly, with a faster search using less time but increasing the probability of missing the object. This trade-off is core to this research. We model a search for an object hidden in one of several discrete locations according to some known probability distribution. There are two available search speeds for each location, fast and slow, with the goal to discover the object in minimum expected time by successive searches of individual locations.

The same search problem with just one available search speed per location is well studied, with a simple optimal policy assigning each location a numerical value called a (Gittins) index and searching the location with the largest index. In the two-speed problem, each location has two indices, a fast index and a slow index, and any search policy, as well as which location to search, needs to determine which search speed to use.
We have two main results. For each location:

1. If the slow index is no smaller than the fast, slow is optimally always used for that location.
2. If the fast index is sufficiently greater than the slow, with 'sufficient' depending on the detection capability of the slow search, fast is optimally always used for that location.

Outside these results, the optimal mode for a location is complicated, sometimes depending on our current beliefs about the object's true hiding place and the search parameters of other locations. After examining the information gained about the object's true hiding place provided by each speed, we propose a threshold-type heuristic policy that demonstrates near-optimal performance in an extensive numerical study. In the future, we look to study the problem in a game-theoretic setting where the object is an intelligent hider who wishes to evade detection.