STOR-i Workshop 2014, 9th-10th January

Abstracts

Day 1

Professor Jim Berger, Duke University, USA
Working with Inexact Models: the World of Computer Modeling

A major activity in science and engineering is the development of simulation- or math-based computer models of processes. Such models are virtually always incomplete representations of reality. The models will be used, however, so the challenge is to understand how to do so effectively. Statistical challenges that arise in this area include the need to solve inverse problems and account for the resulting uncertainty, handling uncertainty in inputs, determination of model bias or discrepancy, and development of bias-adjusted predictions. The methodology to be discussed is a mix of Bayesian spatial, hierarchical and nonparametric techniques. After illustration on a simple pedagogical example, a series of increasingly involved real examples will be discussed.

David Hofmeyr, STOR-i PhD student, Lancaster University
Clustering high dimensional data streams by feature space partitioning

Data streams (high frequency data sequences) are becoming increasingly common in the digital age and so being able to handle these types of data is paramount. In many application areas (image sequences, sensor arrays, etc.) the acquisition of vast amounts of related data has become almost trivial, leading to data streams of very high dimension. The sparsity of high dimensional data leads to phenomena that render many classical methods ineffective.

In this talk we'll look at an approach to building clustering models based on a hierarchical partition of the feature space which avoids the pitfalls associated with high dimensional data. We'll investigate some of the challenges associated with building these models in the data stream context and how we can overcome them. Finally, we will look at some experimental results which illustrate the effectiveness of this approach.

Professor Sanjay Mehrotra, Northwestern University, USA
Optimizing Healthcare: Budgets, Operations, Policies, and Beyond

Healthcare, particularly in US, is a large and complex system. Policies are determined based on legislated priorities, and decisions are often made based on suboptimal algorithms. There is a growing interest in optimal resource utilization, while preserving the ethical equipoise between equity, justice and utility in healthcare. Solutions require a trans-disciplinary collaborative approach, where industrial and systems engineers, operations researchers, and management scientists can make significant contributions by developing realistic data-driven and model based approaches to promote evidence based decision making and informing policy changes. The need to bring greater realism to the decision models also motivates new methodological developments that can then benefit application in areas other than health. The central consideration in developing innovative strategies to improve the health system is to save patients' lives and to improve their quality of life. This must be balanced against risks and cost to individuals and society. This leads to problems with multiple objectives, and input from multiple experts weighing in on these objectives. The parameters of the functions modeling the objectives and constraints are uncertain as model recommendations have implications on future unknown.

Through a wealth of application problems, in this presentation we will first discuss the broader healthcare perspective, and the need and opportunities for the participation of our community. We will then focus on a few specific examples from our research illustrating the need for developing, and the use of recently introduced concepts such as robust Pareto optimality and risk adjusted decision making in the context of

addressing geographic disparity in budgeting US national diabetes prevention programs, stochastic scheduling, and a national policy analysis in kidney transplantation.

Philip Jonathan, Shell, UK
Blowing in the wind : remote sensing of airborne gases and particulates

It's increasingly important to quantify emissions of environmentally sensitive gases and particulate matter into the atmosphere. With some clever measurements, a bit of physics and statistics, the answer to "where are the emission sources, and how much of what are they emitting?" is literally blowing in the wind!

We describe a method for detecting, locating and quantifying sources of gas emissions to the atmosphere using remotely obtained gas concentration data; the method is applicable to gases of environmental concern. We demonstrate its performance using methane data collected from aircraft, and outline its application to ground-based line-of-sight monitoring. Atmospheric point concentration measurements are modelled as the sum of a spatially and temporally smooth atmospheric background concentration, augmented by concentrations due to local sources. We model source emission rates with a Gaussian mixture model and use a Markov random field to represent the atmospheric background concentration component of the measurements. A Gaussian plume atmospheric eddy dispersion model represents gas dispersion between sources and measurement locations. Initial point estimates of background concentrations and source emission rates are obtained using mixed L2-L1 optimisation over a discretised grid of potential source locations. Subsequent reversible jump Markov chain Monte Carlo inference provides estimated values and uncertainties for the number, emission rates and locations of sources unconstrained by a grid. Source area, atmospheric background concentrations and other model parameters are also estimated. We investigate the performance of the approach to aircraft sensing first using a synthetic problem, then apply the method to real airborne data from a 1600km2 area containing two landfills, then a 225km2 area containing a gas flare stack.

Ivar Struijker Boudier, STOR-i PhD student, Lancaster University
Scheduling under uncertainty

The National Nuclear Laboratory (NNL) operates a facility which processes radioactive materials. The scheduling of jobs passing through this facility presents a number of challenges: the duration of jobs is often not known in advance, there is equipment which may suffer breakdowns, and regular maintenance of equipment must also be scheduled. This talk will explore some ideas of how these forms of uncertainty can be integrated into the planning process.

Day 2

Professor Adrian Bowman, University of Glasgow
Surfaces, shapes and anatomy

Three-dimensional surface imaging, through laser-scanning or stereo-photogrammetry, provides high-resolution data defining the surface shape of objects In an anatomical setting this can provide invaluable quantitative information, for example on the success of surgery. Two particular applications are in the success of breast reconstruction and in facial surgery following conditions such as cleft lip and palate. An initial challenge is to extract suitable information from these images, to characterise the surface shape in an informative manner. Landmarks are traditionally used to good effect but these clearly do not adequately represent the very much richer information present in each digitised images. Curves with clear anatomical meaning provide a good compromise between informative representations of shape and simplicity of structure, as well as providing guiding information for full surface representations. Some of the issues involved in analysing data of this type will be discussed and illustrated. Modelling issues include the measurement of asymmetry and longitudinal patterns of growth.

Emma Ross, STOR-i PhD student, Lancaster University
Cross-trained workforce allocation for service Industries

In labour intensive service industries it is often difficult to match supply to demand. Services cannot be inventoried so their delivery must coincide with the timing of often unpredictable demand. Further, recent concerns over the increasing cost of living have put pressure on some organisations to reduce or freeze charges for their services whilst maintaining service levels and development.

The value of cross-trained workers, who can turn their hands to a number of different types of work, is a well-recognised staffing policy for coping with uncertain demand and improving workforce efficiency to reduce operating costs. The effective allocation of such a workforce across their different skills is essential to capitalising on the value of this added flexibility. Limited work of industrial applicability has been carried out in this area however.

In this presentation we explore a range of integer and mixed integer programming models for single and multi-period cross-trained workforce allocation. Through assessment of the `solvability' of these models and the practical relevance of the solutions they provide, we highlight the challenges of mathematical modelling - namely the trade-off between model simplicity and accuracy.

Duncan Elliott, Office for National Statistics
Time series research at the Office for National Statistics

The Office for National Statistics (ONS) publishes thousands of time series covering many areas of the economy and society. There is significant user interest and demand for frequent, timely and detailed time series. In order to help meet some of this demand, the Time Series Analysis Branch (TSAB), a small team of methodologists at ONS support other areas of the office producing outputs that involve time series methods. This talk will provide an overview of the sort of work carried out by TSAB, including areas of current research, issues that we would like to investigate further and discuss in more detail a project to improve the experimental monthly estimates of unemployment with a state space model.

Christopher Nemeth, STOR-i PhD student, Lancaster University
Particle MCMC: Getting the most out of your particle filter

Markov chain Monte Carlo (MCMC) has become highly popular in Bayesian statistics as a method for sampling from complex posterior distributions using samples from proposal distributions. For models such as state-space models, the likelihood is often intractable, which prohibits the use of MCMC. Recently, particle MCMC has been proposed as an extension to standard MCMC whereby the intractable likelihood

is replaced with an unbiased estimate from the output of a particle filter.

The efficiency of the MCMC sampler is highly dependent on the choice and tuning of the proposal distribution. In this talk we'll look at how we can use the output of a particle filter to not only estimate the likelihood, but to also improve the proposal distributions.

Professor Nedialko Dimitrov, Naval Postgraduate School, USA

Goal-oriented Design of Influenza Surveillance

The CDC employs a suite of data streams to achieve the multi-faceted goals of influenza surveillance: influenza-like-illness surveillance network (ILINet), IISP, WHO Labs, NRVSS. The data streams in many of these systems were largely assembled out of convenience or intuitive first principles. As a result, the main challenge is how to use the available data to achieve existing and emerging surveillance goals. Next generation data streams such as Google Flu Trends are now also available. It is unclear which data streams are best to incorporate: some are expensive, others provide noisy data, others yet are unreliable. In this presentation, we discuss a systematic process of design for influenza surveillance. Instead of constraining surveillance by convenience sampling, our process defines and selects the best available data through a four step process: 1) Formalize surveillance objectives 2) Specify candidate data sources 3) Simulate data where none exists 4) Select the most informative data sources. We present an example of this process by constructing a multi-objective influenza surveillance network in Texas. We also discuss ongoing efforts to introduce statistical guarantees into the national influenza surveillance system.