Model Fitting For Spatial Point Patterns On The Celestial Sphere

Thomas Lawrence

Most statistical research on spatial point patterns has focussed on data in two- and three-dimensional Euclidean space. In this presentation, we consider point patterns on a sphere such as the celestial sphere. We outline some basic structures for statistical models for point patterns on the sphere and methods for fitting these models. Models include the analogues of Neyman-Scott, Gibbs and Cox processes. We apply the methods to the spatial pattern of galaxies in the Revised New General Catalogue and Index Catalogue (NGC), a historically important dataset giving the locations (sky positions) of nonstellar astronomical objects as points on the celestial sphere.

Spatially Varying Dependence Parameter For Brown-resnick Max-stable Processes

Boris Beranger
Scott Sisson[1], Michel Broniatowski[2]
[1]School of Mathematics and Statistics, University of New South Wales, Sydney
[2]Theoretical and Applied Statistics Laboratory, University Pierre and Marie Curie

Changes in temperature are very important when predicting and assessing the damage caused by heatwaves. Calculating the 100-year quantile of the temperature regionally for particular days belongs to the study of spatial extremes, based on max-stable processes, and goes a long way to effectively managing a very 'hot' topic. An important criterion is the study of both the dependence between locations in space at high levels of the quantity of interest and the dependence between locations in space at moderate levels of that same quantity.

Buishand et al. (2006) proposed a method that simulates simultaneously extremes and non extremes. Extremes are constructed using Brown-Resnick max-stable processes where a spatial dependence parameter is incorporated. This dependence parameter is considered constant over the whole region estimated by the average of local dependence between pairs of stations. Such estimates rely on the expression of the bivariate marginal distribution of the process. This method is sufficient, although its major drawback is the assumption of equal elevation across an area, which is especially restrictive when the study area is large.

Here we present a fully multivariate extension of Buishand's method that innovates by allowing specific estimation of the dependence at each station. This is done by proving the n-dimensional marginal distribution of the Brown-Resnick process which, to our knowledge, has never been done for any other max-stable process. This method also has the advantage to take into account sudden changes in the characteristics of the region (e.g. altitude) and is using simultaneously information given by the surrounding neighbours.

We will demonstrate that this spatially varying dependence parameter yields to more accurate predictions on simulated and real datasets. We will apply our model to extremely large daily minimum temperature from Australia.

Methodology For Nonparametric Deconvolution When The Error Distribution Is Unknown

Aurore Delaigle
Peter Hall[1]
[1]University Of Melbourne

In nonparametric deconvolution problems, in order to estimate consistently a density or distribution from a sample of data contaminated by additive random noise it is often assumed that the noise distribution is completely known or that an additional sample of replicated or validation data is available. Methods have also been suggested for estimating the scale of the error distribution, but they require somewhat restrictive smoothness assumptions on the signal distribution, which can be hard to verify in practice. Taking a completely new approach to the problem, we argue that data rarely come from a simple, regular distribution, and that this can be exploited to estimate the signal distributions using a simple procedure, often giving very good performance. Our method can be extended to other problems involving errors-in-variables, such as nonparametric regression estimation. Its performance in practice is remarkably good, often equalling (even unexpectedly) the performance of techniques that use additional data to estimate the unknown error distribution.

Uncongeniality With Multiply-imputed Synthetic Data

Bronwyn Loong
Donald Rubin[1]
[1]Harvard University

Several statistical agencies have started to use multiply-imputed synthetic data to create public-use data for major surveys, that simultaneously (i) protect the confidentiality of respondents' identities and sensitive attributes, and (ii) allow standard analyses of microdata. A key challenge, faced by creators of synthetic data, is demonstrating that valid statistical inferences can be obtained from such synthetic data for many non-confidential questions. Large discrepancies between observed-data and synthetic-data analytic results for such questions may arise because of uncongeniality; that is, differences in the types of inputs available to the imputer and the analyst. In this talk we discuss some common examples of uncongeniality when using multiple imputation to create synthetic data. We describe the characteristics that lead to congenial imputation models in our examples. Our discussion highlights the competing objectives of preserving valid inferential conclusions and protecting confidentiality. We conclude that an understanding of uncongeniality is essential, at an intuitive level, by both imputers and analysts, in order to advance the apposite use of synthetic-data techniques for statistical disclosure control, across a broad range of survey data.

Measuring Income Mobility Using Pseudo-panel Data

Arturo Martinez
Mark Western[1], Michele Haynes[1], Wojtek Tomaszewski[1]
[1]Institute for Social Science Research, University of Queensland

Many developing countries are experiencing rapid economic development. However, this occurs within the context of widening (income) inequalities. There are two types of inequalities: inequalities of outcomes and inequalities of opportunities. Between these two types of inequalities, policy makers are often more concerned on increasing inequality of opportunities. Examining income mobility which measures changes in individuals' income over time provides a way of validating whether increasing inequality represents less equitable distribution of opportunities during episodes of economic growth. Longitudinal or panel data that tracks individuals' income is the appropriate data source for measuring income mobility. However, panel data is scarce in many developing countries. To reconcile the need of providing a more dynamic perspective of the evolution of income distribution with the lack of panel data, several techniques have been offered to construct pseudo-panel data from repeated cross-sectional surveys. Using actual panel data from the Philippines, this study evaluates the performance of four pseudo-panel techniques in measuring a wide array of income mobility indicators. Preliminary results suggest that methods with more flexible income model specifications perform better than those with highly parameterized models. More importantly, these flexible pseudo-panel procedures produced estimates of poverty dynamics and movement-based indices which are quite close to the estimates computed using the actual panel data. Nevertheless, further improvements are warranted to be able to develop a more satisfactory estimation procedure for indices measuring temporal dependence and the inequality- reducing effect of income mobility.

Weighted Gibbs Sampling For Mixture Modelling Of Massive Datasets

Clare McGrory
Clair Alston[1], Daniel Ahfock[2], Joshua Horsley[2]
[1]Queensland University of Technology
[2]University of Queensland

Massive datasets are becoming increasingly common in modern applications and this presents tremendous new challenges for statisticians. In a remote sensing problem where the aim is to produce a map showing categories of land use from satellite imagery, the large number of pixels in each image, and the number of images to be analysed can make fitting finite mixture models, or spatial mixture models to the data either infeasible, or too time-consuming to be of practical use. It has been shown that using a representative weighted subsample of the complete dataset to estimate model parameters can lead to much more time-efficient and yet still reasonable inference. These representative subsamples are called coresets. Naturally these coresets have to be constructed carefully as a naive approach of simply performing simple uniform sampling from the dataset could lead to smaller clusters of points within the dataset being severely undersampled which would result in unreliable inference. It has been shown that an adaptive sampling approach can be used to obtain a representative coreset of data points together with a corresponding set of coreset weights. In this talk we explore how this idea can be incorporated into a Gibbs sampling algorithm for mixture modelling of image data. We call the resulting algorithm a Weighted Gibbs Sampler. This algorithm takes account of the coreset weights in order to perform inference using just the coreset rather than all of the available data. We will illustrate this proposed approach through application to remote sensing of land use from satellite imagery.

Model Selection With Survey Data

Alastair Scott
Thomas Lumley[1]
[1]University Of Auckland

The analysis of survey data has expanded enormously in recent years, driven in particular by public access to the results of large medical and social surveys such as NHANES. Typically, researchers conducting such analyses want to use the same techniques that they would use with experimental data. This is now largely possible: all the main packages have survey versions for implementing standard techniques such as linear or logistic regression and some can handle arbitrary generalized linear models. There are still some widely-used quantities missing from these packages, however. Perhaps the most notable of these are standard criteria for model selection such as AIC and BIC. In this paper we show how to develop analogues of both criteria that can be applied to complex survey data.

A National Environmental Information Infrastructure

Andre Zerger
Andrew Woolf[1]
[1]Bureau Of Meteorology

The discipline of environmental modelling is often hampered by an inability for modellers to rapidly discover and access core environmental data. This is particularly pronounced in disciplines where data from multiple environmental domains needs to be accessed and coupled, for example in the area of integrated environmental modelling and assessment. The National Plan for Environmental Information (NPEI) Initiative is an Australian Government program intended to improve the quality and accessibility of environmental information, including for the modelling and science community. It is being jointly implemented by the Bureau of Meteorology and the Department of the Environment. The Bureau's role focuses on operational elements including implementation of technical components of a National Environmental Information Infrastructure (NEII). The NEII proposes a distributed platform to support the discovery, access and re-use of environmental information. It includes the open standards and specifications around the core IT components that will support improved discovery, access and re-use of environmental information. Key outcomes expected to emerge include: (a) an improved ability to discover, access and use national environmental information data resources through harmonised online services and web portals, (b) a sustainable standards-based distributed environmental information architecture that can support multiple application use cases; and (c) a governance and collaboration framework for coordination and environmental information standards adoption. The presentation provides an overview of the core architectural components of the NEII, an update around the current services offered through the NEII, and briefly provides an overview of our approach to engagement and collaboration. When sufficiently mature the NEII has the potential to significantly streamline discovery and provision of a core set of environmental information to support a number of continental-scale environmental modelling applications.

Three Approximate Laws Of Small Numbers Measured In The Wasserstein Distance

Aihua Xia

When we study the counts of rare events, we often choose a model according to the variance to mean ratio to be greater than one (over-dispersed) or less than one (under-dispersed). In this talk, I will look at three possible choices of models and their corresponding estimates when the approximation errors are measured in terms of the Wasserstein distance.

This talk is based on joint works with AD Barbour, H Gan and F Zhang.

Valid Inference After Selecting Predictors And Variable Transformations

Andreas Buja
Richard Berk, Lawrence Brown, Kai Zhang, Linda Zhao

Common practice in model fitting is (1) to select predictors based on one of the many popular predictor selection procedures, and (2) to transform the response and possibly the predictors with logarithms, square roots or more generally Box-Cox transforms. Recent research has attempted to address the problem of obtaining valid statistical inference after variable selection. In the present talk we address the larger and more realistic problem of allowing variable transformations as well. An issue that needs to be resolved before even attempting to address inference after (1) and (2) is the necessity of valid statistical inference in misspecified models. Thereafter we examine a procedure that has the promise of producing valid inference after predictor selection as well as selection from a finite vocabulary of variable transformations.

A Comparison Of Bayesian Models Of Heteroscedasticity In Nested Normal Data

Alan Herschtal
Kerrie Mengersen[1], Farshad Foroudi[2], Tomas Kron[2]
[1]Queensland University of Technology
[2]Peter MacCallum Cancer Centre

We consider the construction of Bayesian models for hierarchically structured data in which observed data points are normally distributed around group means that are themselves normally distributed around a global mean, with heteroscedasticity allowed for amongst the group variances. We consider the case in which the underlying distribution of the variances is unknown, and investigate how best to construct a Bayesian model that accommodates the heteroscedasticity. The most mathematically convenient representation of the group variances is to model them as inverse gamma distributed, since the inverse gamma distribution is the conjugate prior distribution for the unknown variance of a normal population. However this assumption may not reflect the physical reality of the data being modelled, as the true underlying distribution of the variances may conform to some other, possibly non-parametric, distributional form. In this work we demonstrate that for a wide class of underlying distributions of the group variances, the inverse gamma model displays extremely favourable goodness of fit properties, and hence may be used as the default assumption for modelling such data. This allows us to take advantage of the elegant mathematical property of prior conjugacy without compromising model fitness in a wide variety of contexts. We demonstrate our findings on nine real world publicly available datasets from different domains, as well as on an exhaustive suite of artificially generated datasets.

Some Asymptotic Properties Of Non-iterative Estimates Of Linear-by-linear Association Parameter

Sidra Zafar
Salman Cheema[1], Eric Beh[1], Irene Hudson[1]
[1]University Of Newcastle

Abstract

Ordinal log-linear models (OLLM) are among one of the most commonly used models to analyse the association in a contingency table with ordered categorical variables. An important aspect of the OLLM is the estimation of its linear-by-linear association parameter. Traditionally, this parameter is estimated by iterative methods; mainly the Newton-Raphson and the Iterative Proportional Fitting. However, a more direct approach is available to estimate the linear-by-linear association parameter of the OLLM. This approach is based upon a non-iterative framework and performs analogous to the iterative techniques, as shown by empirical and computational studies. Recently, the statistical properties of the non-iterative methods were explored and it was shown that two of these estimates were unbiased. This presentation will show that the non-iterative estimates are asymptotically consistent estimators of the linear-by-linear association parameters.

Keywords: *Ordinal log-linear models, linear-by-linear association, non-iterative estimation*

Random Forests And Statistical Data Editing

Alexander Hanysz

A random forest is an ensemble of decision trees, commonly used for regression or classification. Random forests contain a great deal of internal structure which can be used to interpret the model and "explain" the predicted values. This paper presents a novel application of forests to statistical data editing.

We address the problem of identifying anomalous observations in a dataset and finding measures of which variables are likely to be in error. Such measures can be derived from the internal structure of the forest. The resulting data can then be used as weights in a Fellegi-Holt type error localisation process. This is of special value in datasets with a large number of variables, where it could potentially replace much more labour-intensive methods of assuring data quality.

Adjustment To The Aggregate Association Index For Large Samples

Salman Cheema
Eric Beh[1], Irene Hudson[1]
[1]University Of Newcastle

The analysis of aggregated data has gained the attention of researchers from different fields over the last five decades. In the statistical and allied disciplines, different aspects of the analysis have been discussed and many others are currently under consideration. Recently, the development of the aggregated association index (AAI) provides the analyst with the opportunity to quantify the extent of association between two dichotomous variables of a 2x2 contingency table, rather than modelling the marginal data. Underlying the theory of the AAI is Pearson's chi-square statistic. The Pearson's chi-squared statistic is susceptible to changes in the sample size of the contingency table, therefore, as the sample size increases so does the AAI, even when the marginal proportions remain unchanged. Therefore, one salient feature of the AAI that deserves investigation is the effect that sample size has on the magnitude of AAI; This presentation proposes two adjustments to the AAI that help to overcome this problem of dependency of AAI on sample size. We consider a simple example using R.A. Fisher's classical criminal twin data to demonstrate the application of the AAI and its adjustments. This work is part of a development towards a more unified framework for testing association of cross-classified data.

Keywords: Aggregate association index (AAI), Pearson's chi-square statistic, extent of association.

Nonparametric Statistical Inference For Extensions Of Garch Models

Alexander Meister
Jens-Peter Kreiß[1]
[1]Institut für Mathematische Stochastik, TU Braunschweig

We consider extensions of the famous GARCH(1,1) model where the recursive equation for the volatilities is not specied by a parametric link but by a smooth autoregression function. Our goal is to estimate this function under nonparametric constraints when the volatilities are observed with multiplicative innovation errors. We construct an estimation procedure whose risk attains the usual convergence rates for bivariate nonparametric regression estimation. Furthermore, those rates are shown to be optimal in the minimax sense.

Learning Dags Based On Sparse Permutations

Caroline Uhler
Garvesh Raskutti[1]
[1]University of Wisconsin - Madison

Determining causal structure among variables based on observational data is of great interest in many areas of science. While quantifying associations among variables is well-developed, inferring causal relations is a much more challenging task. A popular approach to make the causal inference problem more tractable is given by directed acyclic graph (DAG) models, which describe conditional independence information and causal structure.

A popular way for estimating a DAG model from observational data employs conditional independence testing. Such algorithms, including the widely used PC algorithm, require faithfulness to recover the correct Markov equivalence class of a DAG. The main justification for imposing this assumption is that the set of unfaithful distributions has Lebesgue measure zero, since it can be seen as a collection of hypersurfaces in a hypercube. However, due to sampling error the faithfulness condition alone is not sufficient for statistical estimation, and strong-faithfulness has been proposed and assumed to achieve uniform or high-dimensional consistency.

We show that the strong-faithfulness assumption is in fact extremely restrictive, implying fundamental limitations for algorithms that require this assumption. We then propose an alternative approach based on finding the permutation of the variables that yields the sparsest DAG. In the Gaussian setting, this algorithm boils down to finding the sparsest Cholesky decomposition of a matrix. We prove that the constraints required for our sparsest permutation (SP) algorithm to recover the correct Markov equivalence class are strictly weaker than strong-faithfulness. In fact, our recovery conditions are necessary for consistency of any causal inference algorithm based on conditional independence testing. Through specific examples and simulations we also compare the SP algorithm to the PC algorithm in practice and show that the SP algorithm has better performance than the PC algorithm.

Risk Margin Quantile Function Via Parametric And Nonparametric Bayesian Quantile Regerssion

Xiaodan (Alice) Dong
Jennifer Chan, Gareth Peters

In this paper we propose to devleop quantile regression to derive risk margin and evaluate capital in non-life insurance applications. By utilizing the entire range of conditional quantile functions, especially higher quantile levels, we detail how quantile regression is capable of providing an accurate estimation of risk margin and overview of implied capital based on the historical volatility of a general insurers loss portfolio. Two modelling frameworks are considered based around parametric and nonparametric quantile regression models which we contrast specifically in this insurance setting. In the parametric quantile regression context, several models including the flexible generalized beta distribution family, asymmetric Laplace (AL) distribution and power Pareto distribution are considered which we detail how to develop under a Bayesian regression framework.

The Bayesian posterior quantile regression models in each case are studied via Markov chain Monte Carlo (MCMC) sampling strategies. In the nonparametric quantile regression models that we contrast to the parameteric Bayesian models we adopted AL distribution as a proxy and together with the parametric AL model, we expressed the solution as a scale mixture of uniform distributions to facilitate implementation. The models are extended to adopt dynamic mean, variance and skewness and applied to analyse two real loss reserve data sets to peform inference and disucss interesting features of quantile regression for risk margin calculations.

Topics In Hierarchical Functional Data

Haocheng Li
Raymond Carroll[1], Matthew McLean[1], David Ruppert[2], John Staudenmayer[3]
[1]Texas A&M University
[2]Cornell University
[3]University of Massachusetts Amhearst

We will review a series of problems and methods for hierarchical functional data. The common theme is that the problems of interest to us have non-standard correlation structures for the functional observations within a subject. Principal components approaches with algorithms incorporating mixed-model ideas will be used to illustrate the methodology.

Posterior Predictive Checking Of Multiple Imputation Models

Cattram Nguyen
Katherine Lee[1], John Carlin[1]
[1]Murdoch Childrens Research Institute

Multiple imputation (MI) is gaining popularity as a strategy for handling missing data, but there is a scarcity of tools for checking imputation models. Posterior predictive checking (PPC) has been recommended as an imputation diagnostic. PPC involves simulating "replicated" data from the posterior predictive distribution of the model under scrutiny. Model fit is assessed by examining whether the analysis from the observed data looks typical of results obtained from the replicates produced by the model. A proposed diagnostic measure is the PPC p-value, where an extreme p-value (i.e. a value close to 0 or 1) suggests a misfit between the model and the data.

The aim of this study was to evaluate the performance of the PPC p-value as an imputation diagnostic. Using simulation methods, we deliberately misspecified imputation models to determine whether PPC p-values were effective in identifying these problems. When the target parameter was the regression coefficient, the magnitude of the PPC p-values correlated with imputation model performance. However, when the target quantities were marginal means and medians, the PPC p-values did not flag the most problematic models. A shortcoming of the PPC method was its reduced ability to detect misspecified models with larger amounts of missing data. Despite the limitations PPC, it might still have a place in the imputer's toolkit. When checking models using PPC, we recommend imputers perform graphical checks and examine other numerical summaries of the test quantity distribution, rather than perform automated checking through PPC p-values.

What Do Learning French And Statistics Have In Common? They Are Both A Foreign Language To Many

Alice Richardson

All Australian universities have introductory statistics courses, with thousands of students in both statistical and non-statistical degrees exposed to statistical concepts annually. Before the statistics education reforms of the 1990s, such courses were heavily mathematical, focused on theory and providing little opportunity to discuss the results of data analysis. However the focus more recently has moved towards students being exposed to real data, using technology for analyzing it, and communicating results using appropriate statistical language.

However many words have specific meanings in statistics that are different to those of everyday English, and this ambiguity can make learning Statistics like learning a foreign language. In this talk I will describe experiences across three Australian universities in addressing the problem of students successfully engaging with the "foreignness" of the language of statistics. These experiences are based around adapting the techniques of language learning to the statistics classroom. I will describe some of the baseline information we collected from statistics students and tutors regarding language learning in statistics classrooms, and report on factors that affect the ability of students to comprehend ambiguous words. I will also describe some of the paper-based and online interventions we used to assist students.

The Effects Of Natural Selection In A Spatially Structured Population

Amandine Veber

One of the motivations for the introduction of the Fisher-KPP equation was to model the wave of advance of a favourable (genetic) type in a population spread over some continuous space. This model relies on the fact that reproductions occur very locally in space, so that if we assume that individuals can be of 2 types only, the drift term modelling the competition between the types is of the form $s.p(t,x)(1-p(t,x))$. Here, s is the strength of the selection pressure and $p(t,x)$ is the frequency of the favoured type at location x and time t. However, large-scale extinction-recolonisation events may happen at some nonnegligible frequency, potentially disturbing the wave of advance. In this talk, we shall address and compare the effect of weak selection in the presence or absence of occasional large-scale events. (Joint work with Alison Etheridge and Feng Yu.)

A Non-exchangeable Coalescent Process Arising In Phylogenetics

Amaury Lambert
Guillaume Achaz[1], Nicolas Lartillot[2], Todd Parsons[3]
[1]UPMC Univ Paris 06 and CIRB-Collège de France
[2]Univ Lyon 1 and CNRS
[3]UPMC Univ Paris 06, CIRB-Collège de France and CNRS

A popular line of research in evolutionary biology is to use time-calibrated phylogenies in order to infer the underlying diversification process. Most models of diversification assume that species are exchangeable and lead to phylogenetic trees whose shape is the same in distribution as that of a Yule pure-birth tree. Here, we propose a non-exchangeable, individual-based, point mutation model of diversification where interspecific pairwise competition (rate d) is always weaker than intraspecific pairwise competition (rate c), and is only felt from the part of individuals belonging to younger species. The only important parameter in this model is $d / c =: 1 - a$, which can be seen as a selection coefficient.

We show that as the initial metapopulation size grow to infinity, the properly rescaled dynamics of species lineages converge to a `shift-down/look-up coalescent' where lineages are given levels: the species at level i is the i-th most recent extant species. At constant rate, all lineages simutaneously ``shift down" their level by -1, while the lineage at level 1 ``looks up" to a geometrically distributed (with parameter a) level and coalesces with the lineage present there.

We propose a dimensionally-reduced version of this model allowing for fast simulation and likelihood computation of given trees. We use this algorithm and MCMC data augmentation methods to estimate a from real trees, and compare this estimate to classical measures of tree imbalance.

The Moderation Effect Of The Teachers Job Satisfaction

Ahmet Cezmi Savas

This study aims to determine the moderation effect of emotional labor competencies of teachers on the relationship between primary school principals' emotional intelligence and teachers' job satisfaction. The target population of this research, which is in causal-comparative model, consists of school administrators and teachers working in primary schools in the city center of Gaziantep in 2012. The sampling of the research consists of 27 school principals and 496 teachers who were selected randomly from the target population. Three research scales were used in order to collect data needed for the research: Emotional Labor Scale, Bar-On Emotional Quotient Inventory- EQ-i and Short Form Minnesota Satisfaction Questionnaire –MSQ. While analyzing the effects of emotional intelligence and emotional labor independent variables on the teachers' job satisfaction as dependent variable, hierarchical multiple linear regression model was used in which gender, age, seniority and education variables were controlled. The model to be tested is : "School principals' emotional intelligence levels affect teachers' job satisfaction both directly and indirectly". In the study, the moderation effect of teachers' emotional labor levels in predicting principals' emotional intelligence levels on teachers' job satisfaction levels was analyzed. After hierarchical regression analyses are made, in the analysis of moderation effect, ModGraph-I program that was developed by Paul Jose and is open to use in his website was utilized. As a result, emotional intelligence competencies of principals and emotional labor competencies of teachers significantly predict teachers' job satisfaction levels. Looking at the results of analysis of the moderation effect of the teachers' emotional labor on the relationship between school principals' emotional intelligence and teachers' job satisfaction; it can be clearly derived that teachers' emotional labor levels was found to have a moderating effect.

An Improved Exponentially Weighted Moving Average Chart For Monitoring Process Variance

Mohamed Razmy Athambawa
Peiris TSG[1]
[1]University of Moratuwa

The control charts for monitoring process variance were developed based on the charts developed for monitoring process mean such as Shewhart, exponentially weighted moving average (EWMA) and cumulative sum (CUSUM) control charts. In case of monitoring process variance, log transformation of the sample variance is used in these charts. Because of this log transformation the design procedure of this charts are complex and it is poorly understood in the industry. In this study, an EWMA chart for monitoring process variance is developed without the log transformation of the variance. Tables for the chart parameters were derived and a four step design procedure is explained with an industrial application. This new charting scheme has several advantages over the existing charts such as sample number free design, simplicity in using joint monitoring scheme of process mean and variance and it easily fits to multivariate monitoring. Using this chart, several variables can be monitored simultaneously on single display without worrying the scales of the variables.

Probabilistic Climate Model Evaluation

Amy Braverman
Noel Cressie[1], Snigdhansu Chatterjee[2]
[1]University of Wollongong
[2]University of Minnesota

Like other scientific and engineering problems that involve physical modeling of complex systems, climate models can be evaluated and diagnosed by comparing their output to observations. Though the global remote sensing data record is relatively short by climate-research standards, these data offer opportunities to evaluate model predictions in new ways. For example, remote sensing data are spatially and temporally dense enough to provide distributional information that goes beyond simple moments. For time periods during which remote sensing data exist, comparisons against multiple models can provide useful information about which models, and therefore which physical parameterizations and assumptions, most closely match reality. In this talk, we propose a method for evaluating the fidelity of a time series generated by a climate model to a comparable time series of remote sensing observations. We simulate the sampling distribution of the norm of the difference between the two series' periodograms, and "score" the climate model using the probability that this norm falls below a suitably chosen, fixed threshold. To simulate the sampling distribution, we use a modified version of wavestrapping (Percival, Sardy, and Davison, 2000) that accounts for the internal variability of both climate model and observational time series. We demonstrate our methodology by evaluating a set of atmospheric models from the Coupled Model Intercomparison Project (CMIP5), used by the Intergovernmental Panel on Climate Change, against observed time series from the National Aeronautics and Space Administration's (NASA) Atmospheric Infrared Sounder (AIRS) remote sensing instrument.

Self Excitation In Equity Indices: Investigation Using Historical S&p500 Returns And Option Prices

Andy McClelland
Stan Hurn[1], Ken Lindsay[2]
[1]School of Economics and Finance, Queensland University of Technology
[2]Department of Mathematics, University of Glasgow

In a ``self-exciting'' market, the occurrence of a crash has a destabilising effect and increases the probability of observing subsequent crashes in the near term. This analysis proposes a novel self-exciting extension of the Bates (1996) jump diffusion model and estimates its parameters using historical S&P500 returns and a sizeable panel of SPX options data (1991-2012). The estimation procedure adapts the particle filter-based method of Johannes and Polson (2009) to the specifics of the self-exciting model and implements the method in a parallelised fashion, achieving significant computational gains. As necessary intermediate steps, the pricing of intensity-related risk is explored using equilibrium arguments, and the pricing function for vanilla European options is derived using the transform methods of Duffie, Pan and Singleton (2000). The estimated parameter set reveals strong evidence of self excitation, and the self-exciting model is found to yield a superior fit to historical data when compared to the nested constant-intensity Bates (1996) model. The match to option prices over the sample period also enjoys a significant improvement.

Nonparametric Independence Screening For Ultra-high Dimensional Longitudinal Data

Ming-Yen Cheng
Toshio Honda[1], Jialiang Li[2], Heng Peng[3]
[1]Hitotsubashi University
[2]National University of Singapore
[3]Hong Kong Baptist University

Ultra-high dimensional longitudinal data are increasingly common and the analysis is challenging both theoretically and methodologically. The purpose of this paper is to offer an automatic procedure in hunting for a sparse semivarying coefficient model, which has been widely accepted in applications. The convention is to iterate between the screening and model estimation step. However, the computation burden is heavy and consistency is not guaranteed. We propose to first reduce the number of covariates to a moderate order by employing a screening method, and then identify both the varying and constant coefficients using a group SCAD estimator, and finally refine the group SCAD estimator by accounting for the within-subject covariance function. The screening procedure is based on working independence and B-spline marginal varying coefficient models. Under weaker conditions than existing ones, we show that with high probability only irrelevant variables will be screened out and the number of remaining variables can be bounded by a moderate order, which allows the sparsity and oracle properties of the subsequent variable selection step. Our group SCAD regularized B-spline estimator detects the constant and varying effects simultaneously. The refined semivarying coefficient model employs profile least squares, local linear smoothing and nonparametric covariance estimation, and is semiparametric efficient. We also suggest ways to implement the method and to select the tuning parameters and the smoothing parameters. An extensive simulation study is summarized to demonstrate its finite sample performance and the yeast cell cycle data are analyzed.

Forecasting Electricity Demand For Small Regional Towns

Anna Munday
John Henstridge[1], Ross Bowden[2], Yuichi Yano[1]
[1]Data Analysis Australia
[2]Horizon Power

Generating high quality and robust 20 year electricity demand forecasts for each town within their domain is a key strategic and asset planning requirement for Horizon Power – a State Government-owned, commercially-focused Corporation that provides high quality, safe and reliable power to approximately 100,000 residents and 10,000 businesses across more than 30 towns in regional and remote Western Australia. Since 2011, Data Analysis Australia has worked collaboratively with the Sales and Marketing team within Horizon Power to develop and implement a statistically sophisticated and best practice methodology to produce these forecasts on an annual basis.

The forecasting methodology is based on an integration of a statistical, data driven and evidentiary approach, balanced with a working knowledge of Horizon Power's business operation and practices, priorities and knowledge of growth prospects for each system. Key features of the methodology include the use of weather corrected data as the basis of forecasting; the separation of forecasting into two components, allowing major business customers to be forecast individually with the rest of customers being forecast using trended growth; estimation of demand both with and without the trend towards ever increasing installation of PV solar panels by individual customers; incorporation of medium and high case scenarios; and generation of P50 (median) forecasts, as well as P10 and P25 forecasts to provide an understanding of more extreme possibilities within each scenario. Implementation of the forecasting is via an online, fully automated and custom designed tool, which takes a number of forecast assumptions and generates not just forecasts but a suite of summary and diagnostic outputs to check and confirm reasonableness, in a highly repeatable manner.

Treatment Effect Estimation In Latent Variable Models With Structural Misspecification

Annette Kifley

Latent variable models are sensitive to misspecifications of the nature of the relationships between observed variables and unobserved underlying latent variables that may be of primary interest. However misspecifications of this type are likely to occur in practice. In this study, we evaluate the performance of reflective latent variable models in estimating treatment or exposure effects when presented with observed item measures that include a mixture of formative and reflective item types. Reflective models assume that observed items serve merely as indicators of the status of the underlying latent variables, while formative items in fact affect the latent variables directly. We explore the sensitivity of global treatment or exposure effect estimates to levels of direct, indirect and mediated effects of treatment. We consider a range of conditions with respect to the amount of formative information included in the assessment and the level of correlation among these items and the latent variable of interest. We find a tendency toward overestimation of treatment effects by the reflective model if, in truth, the treatment affects formative items present in the assessment with little or no direct treatment effect on the latent variable of interest. We find a weaker tendency toward underestimation of treatment effects by the reflective model if, in truth, the treatment directly affects the latent variable but does not affect potentially formative items that are included. Problems in estimation are substantially greater if the assessment is predominantly formative and the formative items share strong similarities with each other. Our simulation studies were motivated by issues arising in analysis of health-related quality of life data, but are relevant to many other applications of latent variable modelling.

A Set Of Characteristic Functions On The Space Of Signatures

Ilya Chevyrev
Terry Lyons[1]
[1]Mathematical Institute, University of Oxford

The expected signature of a probability measure on rough paths acts in many ways as the Laplace transform does for real random variables. As such, it has been asked if the law of a random signature can be determined by its expected value, and whether there exists in general an analogue for rough paths of the usual characteristic function. We study (geometric) rough paths of arbitrary roughness by introducing a separable, complete metric on their space of signatures. Using compact symplectic Lie groups, we define a set of characteristic functions on the space of signatures and show that two random variables are equal in law if and only if they agree on each characteristic function. We demonstrate how, under certain boundedness conditions, the law of a random signature can be completely determined by its expected value and apply this result to the classical Stratonovich signature, the expected signature of Brownian motion up to first exit time studied by Lyons and Ni, and the recent work of Friz and Shekhar on the expected signature of Levy processes.

Quenched Invariance Principle For Simple Random Walk In Correlated Percolation Models

Artem Sapozhnikov

Let S be a random subgraph of Z^d. I will discuss a set of conditions on the distribution of S under which the quenched invariance principle holds for the simple random walk on S. Examples of models satisfying the conditions include Bernoulli percolation, random interlacements and its vacant set, level sets of the Gaussian free field. This is a joint work with E. Procaccia and R. Rosenthal, based on an earlier joint work with A. Drewitz and B. Ráth.

Outlier Removal Using The Bayesian Information Criterion For Group-based Trajectory Modelling

Christopher Davies
Gary Glonek[1], Lynne Giles[2]
[1]Discipline of Statistics, The University of Adelaide
[2]Discipline of Public Health, The University of Adelaide

Attributes measured longitudinally can be used to define discrete paths of measurements, or trajectories, for each individual in a given population. Group-based trajectory modelling methods can be used to identify subgroups of trajectories within a population, such that trajectories that are grouped together are more similar to each other than to trajectories in distinct groups. Existing methods generally allocate every individual trajectory into one of the estimated groups. However this does not allow for the possibility that some individuals may be following trajectories so different from the rest of the population that they should not be included in a group-based trajectory model. This results in these outlying trajectories being treated as though they belong to one of the groups, distorting the estimated trajectory groups and any subsequent analyses that use them.

We have developed an algorithm for removing outlying trajectories based on the maximum change in Bayesian Information Criterion (BIC) due to removing a single trajectory. As well as deciding which trajectory to remove, the number of groups in the model can also change. The decision to remove an outlying trajectory is made by comparing the log-likelihood contributions of the observations to those of simulated samples from the estimated group-based trajectory model. In this talk the algorithm will be detailed and an application of its use will be demonstrated.

An Infinite-dimensional Skorohod Map And Continuous Parameter Priority

Rami Atar

We introduce a Skorohod map acting in the space of paths with values in the space of finite measures over the real line. The motivation is to treat queueing models in which tasks are scheduled according to a continuous parameter priority. Two such well-known models are (1) earliest-deadline-first and (2) shortest-remaining-processing-time-first disciplines. We apply the tool to obtain new formulations of fluid model equations and LLN-scale convergence results for both (1) and (2).

Obtaining Population Adjusted Relative Incidence Rates Using Conditional Likelihood

Christopher Aisbett
Lauren Jones[1]
[1]Lauren Jones Consulting

We consider the relative incidence rate of traumatic spinal cord injury between New Zealand and Australia drawing on hospital discharge data from five consecutive years.

The discordance between event time and capture in the discharge based dataset required proportional hazards assumptions as in D.R. Cox's treatment of Survival and additional assumptions to deal with the censoring of cases not yet discharged.

We exploit annual national census data to obtain the conditional likelihood of an event occurring in New Zealand given the age group, sex and year of the event and that the patient was discharge from hospital during the time spanned by our data.

The conditional likelihood function essentially reduces to the product of Binomial probabilities corresponding to the age-sex cells in the data. We incorporate methods to deal with single traumatic events affecting more than one person's spinal cord.

We find the incidence rate is higher in New Zealand.

Our approach is broadly applicable and robust in the sense that Cox's approach is robust.

Dynamic Analyses In Election Studies

Christian Hoops
Tobias Michael[1], Mark Davis[1]
[1]Ipsos Public Affairs

Our presentation argues for a design that has established itself in the field of scientific political research – the Rolling Cross-Section Design which allows unforeseeable media incidents to be dynamically researched at the daily level. By also taking into account mobile device samples, a better image of the general population can be obtained since young and highly mobile groups of voters are surveyed accordingly. This leads to authoritative analysis results that come close to the reality. Furthermore, the RCS-data allows even hard-to-reach populations to be taken into account appropriately in the random sample. Minimal deviations from the sample and general population will be corrected at the conclusion by a dual frame assessment system developed by the Arbeitskreis Deutscher Markt- und Sozialforscher e.V (Association of German Market and Social Researchers).

Additionally, landline and mobile device samples were used for the first time in conjunction with each other in this RCS-Design. This allows new knowledge to surface about relatively unknown groups such as the Mostly Mobiles. This group is mainly made up of middle-aged, well-educated men with a good income but a low level of political participation.

Posterior Uncertainty Quantification

Aad Van Der Vaart
Botond Szabo, Harry van Zanten

The spread of the posterior distribution supposedly gives an indication of remaining uncertainty of a Bayesian analysis. Unfortunately in nonparametric Bayesian statistics this spread is heavily dependent on a bias-variance trade-off induced by the combination of prior and true parameter (a function): if the prior models the true parameter as smoother than it is, then the spread will be too small. One may attempt to overcome this by an hierarchical or empirical Bayes approach that adapts the prior smoothness to the true value using the data. We show by example (pictures and calculations) that this is never completely successful, due to the fact that nonparametric estimation necessarily extrapolates into features of the truth that cannot be seen in the data. We also give generic conditions on the true parameter that prevent this undesirable phenomenon.

Bayesian Quantile Forecasting Using Realised Volatility

Christian Contino
Richard Gerlach[1]
[1]University Of Sydney

- A realised Volatility GARCH model is developed within a Bayesian framework for the purpose of forecasting Value at Risk (VaR) and Conditional VaR (CVaR). Gaussian and Student-t return distributions are combined with GARCH and E-GARCH volatility specifications to forecast tail risk in five international equity index markets over a four year period that includes the recent global financial crisis.
- Five realised volatility proxies are considered within this framework. Realised Volatility GARCH models show a marked improvement for Conditional Value at Risk forecasting, especially at the more extreme quantiles compared to standard GARCH models under a variety of formal and informal tests.
- This improvement is consistent across a variety of data, volatility model specifications and distributions, and demonstrates that Realised Volatility is superior when producing volatility forecasts

On Remote Analysis Servers And Virtual Data Centres

Christine O'Keefe

Vast amounts of data are now being generated from census and surveys, scientific research, observational projects, instruments and sensors of many kinds. The need to protect the privacy of individuals in the context of health, social and survey databases is widely-recognised, however confidentiality can also be an issue with business datasets of commercially sensitive information. The increasingly common practice of geo-coding datasets can substantially increase the confidentiality issues.

In this talk I will focus on confidentiality issues associated with the use of sensitive datsets for research. Traditional approaches to statistical confidentiality have involved applying confidentiality protection measures to datasets before releasing them to analysts. In response to increasing concerns about the amount of data released, and growing researcher demands for more and more detailed data, the alternatives of virtual data centres and remote analysis are under active investigation internationally. In virtual data centres, users have full remote access to sensitive data and can download outputs from statistical analyses. In remote analysis, users receive outputs from submitted statistical analysis requests but do not have direct access to data. The challenge, in both cases, is to ensure that the statistical outputs sufficiently protect the confidentiality of the data.

In this talk I will give an introduction to virtual data centres and remote analysis in the context of the statistical disclosure control literature. I will review the current methods for balancing data use with confidentiality protection, highlighting some recent and original advances.

Causal Mediation Analysis Without Sequential Ignorability

Xiao-Hua Andrew Zhou
Cheng Zheng[1]
[1]University Of Washinton

Mediation analysis is an important tool in social and behavirol sciences as it helps to understand why a behavioral intervention works. The commonly used approach given by Baron and Kenny requires the strong assumption ”sequential ignorability” to yield causal interpretation.

Tenhave proposed a rank preserving model to relax this assumption. However, the RPM is restricted to the case with binary intervention and single mediator. Also, it needs another strong assumption, ”rank preserving”. We proposed a new model that relax this assumption and our model can handle multi-level intervention and a multi-component mediator. As an estimating equation based method, our model can handle correlated data with the generalized estimating equation and handle missing data with an inverse probability weighting. Finally our method can also be used in many other research settings, which have a similar model as mediation analysis such as treatment compliance, post randomized treatment component analysis. For the proposed causal mediation model, we first showed identifiability for the parameters in the model. We then proposed a semi-parametric method for estimating the model parameters and derived the asymptotic results for the proposed estimators. Simulation showed the good performance of the proposed estimators in finite sample sizes. Finally we applied the proposed method to the two real-world clinical studies: (1) the college student drinking study (2) Improving Mood-Promoting Access to Collaborative Treatment for Late Life Depression.

Data Mining And Editing

Claire Clarke

ASC/IMS 2014 Conference

Abstract template

Data Mining and Editing

Claire Clarke

Assistant Director, Australian Bureau of Statistics, Adelaide

Data mining methods have long been used in the detection of anomalies, so applying them to editing data seems like a straight forward extension. As it turns out, the knowledge discovery capabilities of data mining methods can be employed in the detection of outliers/errors, but also in the preparation for editing by eg producing edit rules. Both supervised and unsupervised methods can be turned to these ends, and this paper will discuss the applicability and general usefulness of a number of methods.

Effects Of Different Prior Distributions On The Bayesian Predictive Inference

Azizur Rahman

Predictive inference is one of the oldest methods of statistical inference and it is based on the observable data. Prior information plays important role in the Bayesian predictive inference. Researchers in this field are often subjective to exercise non-informative prior distribution. This study tests the effects of a range of prior distributions on predictive inference for different modelling situations such as linear regression models under normal and Student-t errors. Findings reveal that different choice of priors not only provide different prediction distributions of the future response(s) but also change the location and/or scale or shape parameters of the prediction distributions.

A New Model To Study On Physical Behaviour Among Susceptible Infective Removal Population

Azizur Rahman
Abdul Kuddus[1]
[1]University of Rajshahi, Bangladesh

This paper is concerned about developing a susceptible-infected-removed (SIR) epidemic model
and to test its' various effects in studying a population for evaluation of policies to control the spread of disease. The SIR model has been divided into three disjoint groups of susceptible, infected and recovered populations and expressed by the differential equations. The theoretical
solutions of these equations are determined with empirical results. Findings reveal that there exists a large class of functions representing interaction between the susceptible and infective populations for which the model shows a very realistic behaviour. The rate of change of removal
population follows a fairly skewed t-distribution pattern with a very rapid increase to the peak,
but a slightly slower decreasing trend toward the right.

Assessing The Athority Of A Ranking Of Many Objects

Peter Hall

For better or for worse, rankings of institutions, such as universities, schools and hospitals, play an important role today in conveying information about relative performance. They inform policy decisions and budgets, and are often reported in the media. However, in many instances the authority of a ranking is particularly difficult to assess. In this talk we shall consider some of the statistical properties of rankings of many objects, particularly properties that can be described theoretically in relatively simple terms. We shall discuss the extent to which rankings are robust and interpretable, and address properties of bootstrap methods that are sometimes used to assess the authority of rankings.

Spectral Properties For Domains With Random Fractal Boundaries

Ben Hambly

A classical problem, dating back to the work of Weyl more than one hundred years ago, is to study the spectral asymptotics for domains in Euclidean space. In this talk we will consider a related quantity, the short time asymptotics of the heat content, for some examples of domains in which the boundary is a random fractal. Using techniques from the theory of general branching processes we will be able to establish the convergence of the rescaled limit of the heat content, as well as fluctuation results which show that there is a central limit theorem describing the short time behaviour of the heat content. This shows that the second order term in the short time expansion is determined by fluctuations in the leading order behaviour.

Optimal Weighting Strategies For Dual Frame Telephone Surveys In Australia

Bernard Baffour
Michele Haynes[1], Mark Western[1], Darren Pennay[2], Sebastian Misson[2]
[1]Institute for Social Science Research, University of Queensland, Brisbane
[2]The Social Research Centre, Melbourne

Traditional telephone surveys have typically relied on landline telephone numbers, but there has been a decline in the response rates and coverage of surveys, due to a wide variety of reasons. In addition, an increasing number of people are contactable by mobile phone and there is a surge in households that do not have landline connections. This has created a challenge for collecting suitably representative data. Dual frame telephone surveys that use both landlines and mobile phones sampling frames can overcome the incompleteness of the landline phone sampling as more and more people transition to being reliant on mobile phones. However, surveying mobile phones introduces new complexities in coverage and sampling, nonresponse measurement, weighting and legal and ethical issues that must be addressed. This paper illustrates some of the consequences of failing to include mobile phone users in telephone surveys, and examines and evaluates some of the different approaches to combining the results from these samples in dual-frame designs using data from Australia.

Generalising Aggregate Association Index.its Connection With Odds Ratio And Association Measurements

Duy Tran
Eric Beh[1], Irene Hudson[1]
[1]University of Newcastle

In many studies of categorical variables, only aggregate, or marginal, data is often available. This is largely due to confidentiality issues imposed by governments/corporations or the data collection process. Currently, there is a long-standing problem to investigate the association structure between two or more categorical variables where limited information is available. Such work lies in the world of Ecological Inference (EI) and has attracted a wide range of attention since the days of RA Fisher.

The main disadvantage of the various EI approaches is that the techniques make a variety of untestable assumptions that are directly related to individual level data. As an alternative to many of the EI techniques, one may consider the Aggregate Association Index (AAI) to obtain valuable information about the association between the variables of a single 2x2 table or stratified 2x2 tables. Since the AAI is a new approach, its connection with other measures of association is not yet well established.

Therefore, the objective of this presentation is to provide insight into the generalisations of the AAI and its relationship to a variety of classic association measurements including the odds ratio and those that may be expressed as a linear transformation of the data. As a demonstration of this new development, we shall analyse a unique record of New Zealand gendered election data from 1893 - the first time in the world where women were given the right to vote at a national election. Given only aggregate data and the generalisation of the AAI, the important nature of the association between gender and voting in the New Zealand election shall be examined and compared to the results where the individual level data is known.

Modeling The Changing Demographics Of Religion

Barry McDonald

A model based on ‘sociophysics’, using census data, recently predicted religion to become extinct in New Zealand, Australia and other countries. Contrariwise, a British demographer predicts that, due to higher birth rate, conservative religionists will eventually dominate their various religious traditions and overall religious numbers will grow, reversing the secular trend.

This talk will examine New Zealand evidence from census and independent survey data. It will explore the somewhat ambiguous category ‘no religion’. It will provide quantitative evidence on various drivers of change in religious demographics, and the likely effect on predictive models.

High Dimensional Semiparameric Inference

Han Liu

ASC/IMS 2014 Conference
Abstract template

High Dimensional Semiparametric Inference

Han, Liu
Professor, Operations Research and Financial Engineering at the Princeton University, Princeton

We introduce a unified semiparametric modeling famework for a large family of learning problems, including graphical models, sparse principal component analysis, topic models, covariance estimation, correlation screening, sparse regression, and classification. The key component of our framework is the transelliptical model, which is equivalent to the elliptical copula family but with different interpretation and identifiability conditions. Since the transelliptical model contains both finite- and infinite-dimensional parameters, it is more flexible in modeling. In terms of inference, we exploit a rank-based procedure which is highly robust to data contamination and directly estimates the parameter of interests by treating the infinite-dimensional component as nuisance. Theoretically, these estimators attain nearly optimal rates of convergence as the Gaussian based methods.

Variational Bayes Inference For Large Vector Autoregressions

Reza Hajargasht
Tomasz Wozniak

Variational Bayes provides an approximation to the joint posterior distribution of parameters of a model. The approximate posterior is usually accurate and of a tractable form. We show that when applied to large Bayesian Vector Autoregressions, proven to have excellent performance for forecasting of economic variables, Variational Bayes allows for fast and accurate computations of posterior distributions. The algorithms for the Variational Bayes estimation of VAR models with a variety of prior distributions, including hierarchical prior structures are derived. Based on a Variational Bayes measure of within sample fit, a procedure for choosing the optimal hyper-parameters of the prior distributions is also proposed. Finally, a new estimator of the marginal data density based on the output from both MCMC and Variational Bayes estimation is shown to have good properties.

Keywords: Large Bayesian VARs, Approximate Inference, Marginal Data Density, Hierarchical Prior Distributions, Optimal Hyper-parameters, Forecasting

Bayesian Age Reconciliation For The Auckland Volcanic Field

Emily Kawabata
Mark Bebbington[1], Shane Cronin[1], Ting Wang[2]
[1]Massey University
[2]Otago University

Bayesian age reconciliation for the Auckland Volcanic Field

The estimation of the spatio-temporal hazard from a monogenetic volcanic field, where each eruptive event creates a new volcanic centre, is critically dependent on a likely reconstruction of past events. The Auckland Volcanic Field has produced about 50 volcanic centres during its active phase over the last 250,000 years. Age data for many of these exist, from radiocarbon or other radiometric methods, thermoluminescence, and paleomagnetism. However, the results are often inconsistent. Moreover, the age order of some pairs is known due to the overlying of lavas (stratigraphy). We consider how best to reconcile this mess of data in a Bayesian paradigm, using informative priors, obtained via expert elicitation, on both the individual ages, and the reliabilities of the dating methods. We will also discuss how additional data, from ash layers deposited in five swamps within the field, can be incorporated. This requires an attenuation model, which links estimated eruption volumes with locations of source volcano and tephra deposits, to calculate the likelihood of any combination of volcano and tephra, all treated as an iteration within the Bayesian model.

Oracle Bayesian Variable Selection

E. Belitser

An important problem in statistical analysis is the choice of an optimal model from a certain set of models. In many cases, this reduces to the choice of which subset of variables should be included into the model. We introduce the notion of oracle set of variables and apply the Bayes approach to the problem of variable selection in the Gaussian white noise model. The proposed Bayesian procedure is shown to ``mimic the oracle''. We also study implications for the model selection problem, namely we propose a Bayes model selector and assess its quality in terms of the so called false selection probability.

Exchangeable Markov Models For Time-varying Complex Networks

Harry Crane

In fields as diverse as physics, biology, sociology and national security, complex networks are used to model structural relationships among individuals and variables. In many applications, the networks vary over time in a non-deterministic way and so are appropriately modeled by a graph-valued stochastic process. Motivated by this framework, we study Markov processes that evolve on the space of infinite networks. Natural statistical models for such networks are both exchangeable with respect to relabeling vertices and have the property that all restrictions to finite induced subgraphs are finite-state space Markov chains. Under the product-discrete topology, this class comprises all exchangeable Feller processes on graphs. Our main theorem provides a Levy-Ito-type characterization for processes in this class. Our approach also gives a straightforward recipe for simulating general processes of this type, which may be useful in a range of applications. Related topics, including Aldous-Hoover theory of partially exchangeable arrays and the Lovasz-Szegedy notion of graph limits, will be discussed in connection with this work.

This work is supported by grants from the US National Science Foundation and National Security Agency.

An R Implementation Of The Polya-aeppli Distribution

Conrad Burden

A Polya-Aeppli or geometric compound Poisson random variable is defined as the sum of a Poisson number of independent and identically distributed geometric random variables. Its distribution arises, for instance, as an approximation to the distribution of the number of occurrences of a given short word in a random Markovian sequence of letters from a finite alphabet. Thus the Polya-Aeppli distribution finds applications in bioinformatics in searches for over- or underrepresented words, which may signal functional elements within genomes.

The stats package in R contains implementations of many standard univariate probability distributions as functions for the density/mass function, cumulative distribution function, quantile function and random variate generation. We present here a fast implementation of these 4 functions for the Polya-Aeppli distribution. The implementation relies on iterative formulae for the log of the mass and cumulative distribution functions, which allow an accurate evaluation of the distribution in the extreme upper and lower tails.

Measuring Association And Finding Patterns In Correlation Matrices From Gene Expression Networks

Melanie Bahlo
Vesna Lukic[1], Karen Oliver[2], Natalie Thorne[1]
[1]Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research
[2]Epilepsy Research Center Heidelberg

Unlike DNA RNA quantities vary, depending on its tissue of origin. We can measure levels of RNA using gene expression microarrays or RNA-seq. Human diseases often show effects on RNA levels in specific tissues, for example brain tissue reflects expression changes for autistic patients in comparison to normal patients. We have been using publicly available gene expression data to infer gene expression networks with sets of known disease causing genes for a variety of brain-related disorders. Here we show results comparing different measures of association and how ordering of correlation matrices and the usage of partial correlation matrices can give insights into how these genes interact with each other. Finally we show these findings can help to identify new candidate disease causing genes for some of these diseases.

Generalizability In Causal Inference: Transportability, External Validity, And Meta-analysis

Elias Bareinboim
Judea Pearl[1]
[1]Computer Science Department, UCLA

Recent advances in graphical models and the calculus of actions have given rise to mathematical problems that are not easily formalized, let alone solved in the conventional language of probability and statistics. We exemplify this challenge through one such problem – transportability – which aims to determine when it is feasible to generalize experimental findings from one or several environments to another, potentially different from the rest. This problem is at the heart of every scientific investigation since, invariably, experiments performed in one environment (or population) are intended to be used elsewhere, where conditions may differ considerably. Using a graphical representation of differences and commonalities among two or more environments, we provide a formal characterization and complete algorithmic solution to the problem of whether a specific causal effect is transportable across environments and, if the answer is affirmative, what measurements need be taken in the various populations and how they ought to be combined to produce a consistent estimate of the causal effect in the target environment. Related problems concerning generalization across populations will be outlined (e.g., external validity, meta-analysis).

New Notions Of Tractability For Analytic Multivariate Problems

Henryk Wozniakowski

Standard notions of tractability are appropriate for problems with finite smoothness. They are defined in terms of (ε^{-1},d), where ε denotes an error threshold, and d is the number of variables.

For analytic multivariate problems we can demand more and express new notions of tractability in terms of (1+ log ε^{-1}, d). We verify new notions of tractability for multivariate integration and approximation defined over Korobov spaces. The talk is based on joint work with J. Dick, G. Larcher, P. Kritzer abd F. Pillichshammer.

Designing Large - Scale Nudge Algorithms

Chinmoy Mandayam
Balaji Prabhakar[1]
[1]Stanford University

In many of the challenges faced by the modern world, from overcrowded road networks to overstretched healthcare systems, large benefits for society come about from small changes by very many individuals. We survey the problems and the cost they impose on society, and describe a framework for designing, modeling and analyzing "nudge algorithms". In this talk we focus on reducing congestion during peak hours in transportation networks by shifting or nudging commuters from the peak hour using incentives. We find that the "cost of congestion" is intimately related to the Wasserstein distance of optimal transport theory, and that it is a convex function of the applied load. This allows us to determine the optimal amount by which to nudge peak-time commuters for a given budget. We present results from pilots conducted in Bangalore, at Stanford and in Singapore.

Building On Seifa: Finer Levels Of Socio-economic Summary Measures

Courtney Williamson
Phillip Wise[1]
[1]Analytical Services Unit, Australian Bureau of Statistics

Socio-Economic Indexes for Areas (SEIFA) seek to summarise the socio-economic conditions of an area using relevant information from the Census of Population and Housing. The SEIFA indexes are widely used measures of relative socio-economic advantage and disadvantage at the Statistical Area Level 1 (SA1) level. The indexes provide information about the area in which a person lives, but within any area there are likely to be households, families and individuals with different characteristics to the overall population of that area. Constructing socio-economic summary measures for finer units such as households would enable researchers and policy makers in Australia to better differentiate between areas with concentrations of advantage and disadvantage.

This paper proposes an experimental household level index as an addition to the current suite of SEIFA products. Using 2011 Australian Census of Population and Housing data, this paper focuses on an exploration into the development and dissemination of a socio-economic index for households. It would complement the area level rankings by adding more depth to the information given by SEIFA, as well as providing its own valuable insights.

A New Way Of Estimating A Density

Yannick Baraud
Lucien Birge[1]
[1]Université Paris 6

In the density estimation framework, we propose a new estimation procedure which allows to build on model S an estimator which is both robust (to misspecification) and optimal (or nearly optimal). When the model is exact, parametric and regular enough, we show that the estimator coincides with the maximum likelihood one with probability close to 1. In particular, the estimator is asymptotically efficient. For more general models, including some non-parametric ones, the estimator possesses the properties to be robust with respect to the Hellinger distance and to converge at optimal rate (up to a possible logarithmic factor) in all cases we know. □

Identifying Gene Interactions Using Sparse Canonical Correlation Analysis With Resampling

Haiyan Huang
Y.X. Rachel Wang[1], Keni Jiang[2], Lewis J. Feldman[2], Peter J. Bickel[1]
[1]Department of Statistics, University of California at Berkeley
[2]Department of Plant and Microbial Biology, University of California at Berkeley

Identifying gene interactions has been one of the major tasks in understanding biological processes. However, due to the difficulty in characterizing/inferring different types of biological gene relationships, as well as several computational issues arising from dealing with high-dimensional biological data, finding groups of interacting genes remains challenging. In this work we elucidate higher-level gene-gene interactions (i.e., gene group interactions) by evaluating conditional dependencies between genes, i.e., the relationships between genes after removing the influences of a set of other functionally related genes. The detailed technique involves performing sparse canonical correlation analysis with repeated subsampling and random partition. This technique is especially unique and powerful in evaluating conditional dependencies when the correct dependent sets are unknown or only partially known. When used effectively, this is a promising technique to recover gene relationships that would have otherwise been missed by standard methods. In addition, comparisons with other methods using simulated and real data show this method achieves considerably lower false positive rates.

Empirical Bayes In The Presence Of Exceptional Cases

Belinda Phipson
Gordon Smyth[1]
[1]Walter and Eliza Hall Institute of Medical Research

Empirical Bayes is a statistical approach for estimating a series of unknown parameters from a series of associated data observations. It provides an effective means to "borrow strength" from the ensemble of cases when making inference about each individual case. Such methods are ideally suited to genomic applications where data is collected for tens of thousands of genes simultaneously. Empirical Bayes methods can however be less effective when highly exceptional cases are present. I will discuss our solution for dealing with exceptional cases whereby these cases are identified and the amount of "learning" from the ensemble is assessed on a case-specific basis. Our particular application is for estimating gene-wise variances from microarray and sequencing data. The robust empirical Bayes procedure recognizes and protects against hyper-variable genes. The new procedure improves statistical power for most genes in many data sets. In the presence of hyper-variable genes, the robust method improves power to detect differential expression for the majority of genes that are not outliers, while down-weighting hyper-variable genes. The robust procedure is implemented in the limma software package, which is freely available from the Bioconductor repository.

Simulating Whole Genome Dna Methylation Data

Peter Hickey
Peter Hall[1], Terry Speed[2]
[1]Department of Mathematics and Statistics, University of Melbourne
[2]Bioinformatics Division, Walter and Eliza Hall Institute of Medical Research

DNA methylation is a biochemical process that plays an important role in regulating gene expression in a variety of scenarios, such as embryonic development, aging and cancer development. Developing statistical methods to analyse DNA methylation data, such as identifying differences between samples in the mean and variability of DNA methylation levels, is an active area of research in bioinformatics.

The development of statistical methods for the analysis of DNA methylation data, and assessing their performance, is complicated by the lack of large and biologically well-validated datasets. Therefore, it is very useful to be able to simulate data where the truth is both known and controllable. It is of course vital that the simulated data are similar enough to the real data, where "similar enough" is defined with respect to the statistical questions under consideration. Current software for simulating whole genome DNA methylation makes strong and unrealistic assumptions, such as spatial-independence of DNA methylation along the genome. This software, while useful for some purposes, does not generate sufficiently realistic data for problems such as comparing tests of differential methylation.

We have developed software to simulate DNA methylation data that includes the strong spatial dependence of DNA methylation. We do this by segmenting the genome into regions of “local similarity” and introducing spatial dependence by a Markov model where the transition probabilities depend on the genomic context. We also model the strong intra- and inter-sample heterogeneity found in DNA methylation data.

I will present data from our simulation software and show how it is useful in developing and evaluating the performance of methods designed to analyse DNA methylation data. I will also present some of the biological findings we made in our extensive exploratory analysis of tens of whole genome DNA methylation experiments, which were vital in developing our simulation methodology.

Semiparametric Bernstein Von-mises Theorem: Second Order Studies

Guang Cheng
Yun Yang[1], David Dunson[1]
[1]Duke University

Semiparametric Bernstein von-Mises Theorem has been successfully developed by Bickel and

Kleijn (2012) in a general setup among others. This talk mainly focuses on its second order extension with an attempt to figure out the influence of nonparametric Bayesian prior on the semiparametric inference, i.e., parametric component. Such results can provide us new theoretical insight in guiding the choice of objective prior in a general semiparametric setup.

Australian Teachers' Intent To Leave Teaching Profession Through Logistic Regression Analysis

Bo Cui
Alice Richardson

Teachers play a vital role in shaping the lives of our children, in Australia, teaching work force is experiencing teacher shortage especially in particular subject areas such as mathematics, science etc. This particular research project is utilizing Staff in Australia's Schools Survey (SiAS) 2010 data set as the data base to formulate Multiple logistic regression for the purpose of investigating the prominent factors would influence teachers' decision of leaving the teaching profession prior to retirement, the relevant research results would be informative and meaningful to address teacher shortage issues in particular subject areas across the states/territories of Australia.
Key words: Logistic regression, Australian teacher, Teaching workforce, teacher shortage, logit model, intent, attrition for school teachers, diagnostics for logistic regression.

Acknowledgement
We would like to thank the Department of Education, Employment and Workplace Relations (DEEWR, currently as Department of Education) Teacher Quality and Workforce Data Branch and staff members Paul Hunt, Catherine Quinn for the permission of the access to SiAS data set and the kind support.

Let Us Talk About Monotone Polynomials

Turlach Berwin

Due to their properties, monotone polynomials have, for some applications, a natural advantage over monotone smoothing techniques e.g. monotone polynomial are naturally strictly monotone. However, monotone polynomials have the reputation that they are hard to use in practice. In this talk we will review several approaches for fitting monotone polynomials to data and show that this reputation is completely undeserved. We also discuss various computational issues that arise when using monotone polynomials for statistical inference in either a Frequentist framework or a Bayesian framework.

Linear Models With Heteroskedastic And Correlated Errors

Han Xiao
Wei Biao Wu[1]
[1]The University of Chicago

We consider asymptotic inference of linear models with heteroskedastic and correlated errors. For the errors we propose a very natural dependence framework that is easy to work with. Under the assumption that the errors are short-range dependent, we show that the M-estimate is asymptotically normal. Furthermore, we establish consistency of the Bartlett-type estimate of the covariance matrix of the estimate. Our framework sheds new light on this classical problem and the imposed conditions for the central limit theorem and the heteroskedasticity and autocorrelation consistent covariance matrix estimate are natural and nearly optimal.

Forecasting Tea Auction Prices Capturing Common Seasonal Patterns

Dilshani Induruwage
Chandima Tilakaratne[1], Samithree Rajapaksha[1]
[1]Department of Statistics, University of Colombo

Tea auction prices around the globe show strong seasonality. Filtering out these seasonal components and use of seasonally adjusted data for the analysis may not only lead to loss of important information, but also true nature of the relationships between these series may not be revealed. So it is important to consider a procedure which is able to deal with seasonality. Therefore, this study attempts to capture seasonality in tea auction prices of eight auction centres around the world and as well as to account the seasonality in modelling the forecasting procedure.

In this study, long-run equilibrium relationships between tea auction prices (from January 2001 to January 2013) of eight auction centres around the world are estimated using seasonal cointegration and seasonal error-correction methods. Seasonal unit root tests provide evidence of presence of seasonal cycles in monthly tea auction prices of Colombo, Kolkata, Guwahanthi centres. This test is conducted using the procedure developed by Beaulieu and Miron in 1992 for monthly data. A long-run seasonal cointegration relationship was found in aforesaid three tea auction prices which make it possible to fit a seasonal error correction model (SECM) for them. This SECM model was developed based on the method proposed by Engle, Granger, Hylleberg and Lee in 1990.

The estimated SECM model includes six lags of each auction prices of Colombo, Kolkata and Guwahati and seasonal error correction term. The adequacy of the model was tested using mean square error (MSE) and the coefficient of determination (R^2) of the model of predicted values on actual values. MSE values of the models representing Colombo, Kolkata, Guwahanthi are 0.005, 0.34, 0.21 respectively while their R^2 values are 0.96, 0.87, 0.70. These results indicate that reasonably accurate forecasts can be obtained from the approach employed.

A Bayesian Dirichlet Process Mixture Model With Applications In Comparative Effectiveness Research

Chenguang Wang
Gary Rosner[1]
[1]Sydney Kimmel Comprehensive Cancer, Johns Hopkins University

Comparative effectiveness research (CER) is designed to synthesize evidence of the benefits and harms of a treatment option from disparate sources. Relevant data from randomized clinical trials, post-marketing surveillance trials, and registries are combined to inform healthcare decisions on what treatment is best for a patient. Thus, the task of addressing study-specific heterogeneities becomes one of the most difficult challenges in CER. Bayesian hierarchical models with non-parametric extension at certain levels of the hierarchy provide a powerful and convenient platform that formalizes the information borrowing across the studies. Müller et al. (2004) proposed a Bayesian hierarchical Dirichlet process mixture model to jointly model the response y and patient characteristic covariates x across related studies. The inference is then drawn based on the deduced conditional distribution y|x. In this paper, we show that by employing this approach a patient with specific characteristics may be "over-presented" by information from a study that had patients with similar characteristics. Alternatively, we propose a Bayesian Dirichlet process mixture model that explicitly models the regression of y|x, allowing the residual to have non-parametric Dirichlet process mixture priors. A simulation study is conducted to evaluate the model performance under different scenarios. Finally, the model is applied to a dataset that mimics real data available for CER study of cardiac resynchronization therapy.

Alan T. James' 1964 Paper And Its Implications For Modern Wireless Communications

Matthew McKay

Alan James' 1964 paper was an important milestone in multi-variate statistical theory; providing a useful expository treatment in addition to deriving new results. These have been applied extensively and have had a profound impact in diverse scientific areas, with new applications continually emerging. One prominent application in electrical engineering is that of multi-antenna and multi-user communication technologies, which are revolutionizing modern wireless systems such as 4G cellular and Wi-Fi. In this context, this talk will highlight numerous multi-variate statistical challenges which arise, and will discuss connections with the work of James and his colleagues.

Waiting Time Distributions In The Accumulating Priority Queue

Ilze Ziedins
David Stanford[1], Peter Taylor[2]
[1]University of Western Ontario
[2]University of Melbourne

In traditional priority queues, arrivals are assigned a fixed priority and served according to their initial assigned priority. This can lead to unacceptably long waiting times for low priority customers. We discuss a priority scheme, first introduced by Kleinrock (1964), where customers are assigned to a priority class, but their priority increases linearly with their waiting time in the queue. The higher the priority class of the customer, the faster their priority accumulates. When the server becomes free, the next customer served is the one with greatest accumulated priority, so that under this scheme a customer from a low priority class may be served before a customer from a higher priority class if they have been waiting sufficiently long. We derive the waiting time distribution for customers in the accumulating priority queue, and discuss how it may be used to meet performance measures in health care delivery, where such a priority scheme is particularly relevant.

Collisions Of Brownian Particles And Related Stochastic Analysis

Tomoyuki Ichiba

We examine colliding behaviors of Brownian particles on the real line with bounded, measurable drift and bounded, piecewise continuous diffusion coefficients of the current configuration of particles. The study of such collisions is closely related to the solvability of the corresponding stochastic differential equations that describe those particles. In this talk we discuss the connections of colliding behaviors to the filtration relations as well as to the reflected Brownian motions in polyhedral domains and the perturbed Tanaka equation.

On Variational Bayes Estimation And Model Selection For Linear Regression

Chong You
John T. Ormerod[1]
[1]University of Sydney

In applied statistics model selection is one of the most fundamental tasks. The selection of models in the linear regression context has been well-studied. However, the efficiency and effectiveness of these methods in high-dimensional settings is limited. We present a model selection method to addresses this.

Variational Bayes (VB) is known as a fast alternative to Markov chain Monte Carlo for performing approximate Bayesian inference. However, VB is often criticised, typically based on empirical grounds, for being unable to produce valid statistical inferences. In You et al (2013), asymptotic properties are proved for VB based estimators in Bayesian linear models, partially contradicting this criticism.

Encouraged by You et al (2013), we extend the VB approach to the more complicated spike and slab priors. We show under mild regularity conditions, that: (i) VB based estimators for the coefficients are consistent estimators of the true parameters; and (ii) the VB estimators of the model indicator variables shrink towards zero rapidly if the corresponding true value of the coefficient is zero and one otherwise. This property allows us to use VB estimates of indicator variables to select models.

Simulations results support that our method is competitive in terms of efficiency and effectiveness in comparison to various alternative model selection procedures considered.

Detecting Communities In Networks Using Latent Variable And Monte Carlo Simulations

Adrien Ickowicz

There has been considerable recent interest in algorithms for finding communities in networks - groups of vertex within which connections are dense (frequent), but between which connections are sparser (rare). Most of the current literature advocates an heuristic approach to the removal of the edges (i.e., removing the links that are less significant using a well-designed function). In this presentation, we will investigate a technique for uncovering latent communities using a modeling approach. It will prove to be easy to use, robust and scalable. It makes supplementary information related to the network/community structure (different communications, consecutive observations) easier to integrate.

The main idea is to consider a latent variable for each vertex in the network. This latent variable represents the ”location” of this vertex on a circle. The main assumption is that two vertex belong to the same community if the have they are close to each other on the circle. The probability that two vertex belong to the same community is then defined as a function of the distance between the two vertex. The advantage of this approach is that it both reduces drastically the parameter dimension and it is easy to capture.

We will show that using a well designed likelihood, the MCMC algorithm leads to an accurate estimation of the parameters, and finally using the popular modularity function, we automatically identify the communities. We extend the approach to the estimation of time-dependent networks, using the Sequential Monte Carlo principle.

We will demonstrate the efficiency of our approach by providing some illustrating real-world applications, like the famous Zachary karate club, or the Amazon political books buyers network.

Tensor Completion

Cun-Hui Zhang

Larry Shepp made fundamental contributions to medical imaging, finance and other areas where tensor data proliferate. Is it possible to recover a large low rank tensor when we observe only a small fraction of its entries? Unfolding tensor into matrix does not provide a satisfactory answer in terms of required sample size. We provide a solution to the tensor completion problem by minimizing a certain tensor nuclear norm, along with a sharper sample size condition.

□

Towards A New Statistical Computing System

Ross Ihaka
Brendan McArdle[1]
[1]Department of Statistics, University of Auckland

Systems like R and S provide useful vehicles for carrying out statistical computations. Such systems are very flexible but, unfortunately, this flexibility comes with a price—they are slow and demanding of machine resources. Some attempts have been made at overcoming these problems but the success in doing so has been limited.

This talk will examine the performance bottlenecks in R and will argue that the only way to truly overcome them is to change the underlying language semantics. By considering what can be done efficiently in collection-oriented languages, it is possible to get an idea of what an efficient statistical computing environment might look like and how it might be possible to implement one.

Hierarchical Integration Of Multi-layered Data For Classification And Biomarker Discovery.

Ellis Patrick
Samuel Mueller[1], Jean Yang[1]
[1]University of Sydney

Over the last decade, several statistical techniques have been proposed to tackle genome-wide expression data. However, with the advancement of many other high-throughput biotechnologies, the interest of researchers has been focusing on utilising multiple data sources together with the clinical data, to improve the prognosis of disease outcome. Integrating the features from different platforms has become a crucial step to better understand the relationships between clinical and -omics data and the information they provide to classify some response. The statistical task of preserving the stability and interpretability of the classifier has become more challenging in this framework. One major issue is that the large dimension of -omics data can completely dominate the modelling procedure and it is an open question how to best combine different types of variables. This talk will present our most recent results on improving upon standard classification procedures for Melanoma and Ovarian Cancer Data. We will use a hierarchical multi-stage framework to integrate large -omics datasets to improve the classification of different disease outcomes.

Subsequential Scaling Limits Of Simple Random Walk On The Two-dimensional Uniform Spanning Tree

David Croydon
Martin Barlow[1], Takashi Kumagai[2]
[1]Department of Mathematics, University of British Columbia
[2]Research Institute for Mathematical Sciences, Kyoto University

The joint work that I will describe establishes that the law of the simple random walk on the two-dimensional uniform spanning tree is tight under a particular rescaling of time and space. Whilst such a result immediately implies the existence of subsequential scaling limits for the random walks in question, our techniques further allow us to describe these limits as diffusions on random real trees embedded into Euclidean space, and derive various transition density estimates for them.

A Statistical Model For Estimating Isoform Expression Using Multi-sample Rna-seq Data

Agus Salim
Stefano Calza[1], Chen Suo[2], Yudi Pawitan[2]
[1]University of Brescia
[2]Karolinska Institute

RNA-sequencing technologies provide a powerful tool for expression analysis at isoform level, but accurate estimation of isoform abundance is still a challenge. Standard assumption of uniform read intensity would yield biased estimates when the read intensity is in fact non-uniform. The problem is that, without strong assumptions, the read intensity pattern is not identifiable from data observed in a single sample. We develop a statistical model that accounts for non-uniform read distribution and jointly estimate gene isoform expression. The main challenge is in dealing with the large number of isoform-specific read distributions, which potentially are as many as the number of splice variants in the genome. A statistical regularization via a smoothing penalty is imposed to control the estimation of read distribution. We develop a fast and robust computational procedure based on the iterated-weighted least-squares algorithm, and apply it to simulated data and two real RNA-Seq datasets with RT-PCR validation. Empirical tests show that our model performs better than competing methods in terms of reducing bias and increasing sensitivity in isoform-level differential-expression analyses.

Use Of Multiple-frame And Propensity Approaches To Bias Problems In The Integration Of Survey And Administrative Data

John Eltinge

In recent years, statistical organizations have expressed increased interest in the integration of standard sample survey data with information provided by other sources, e.g., administrative records or commercial transaction data. However, the quality of inferences based on the integrated data will depend on the implicit observational mechanisms that link the abovementioned sources with the underlying population(s) of interest. This paper presents a unified framework through which to analyze these mechanisms; multiple-frame and observational-propensity approaches are special cases within this framework. For this approach, auxiliary variables (e.g., domain membership indicators for multiple-frame analyses and regressors for propensity models) are of special interest. The proposed framework leads to several diagnostics for evaluation of data sources that are candidates for inclusion in the integration process.

Key words: Auxiliary data; coverage bias; design- and superpopulation-based inference; diagnostics; dual-frame survey; observational data; organic data; R-indicators; response propensity model; selection mechanism; unstructured data.

Stochastic Regression Clustering And Its Model Selection Using Mcmc

Ling Ding
Guoqi Qian[1]
[1]University of Melbourne

Regression clustering integrates cluster analysis and multiple regression to develop a new method for data mining. Instead of fitting a single regression hyperplane on the whole data set, it iteratively performs clustering of the data followed by fitting a regression hyperplane on each resultant cluster until certain optimality is achieved. In this article, we will develop such a new method using Gibbs sampling and least squares estimation techniques, where data partition and regression estimation are performed simultaneously in a cohesive way. Finally, a simulation study will be given to assess the performance of the method proposed. In addition, we will develop several new criteria to choose the optimal number of clusters in regression clustering.

A Version Of Mspe For Estimating Parameters In Spatial Regression Models

Hong-Ding Yang
Chun-Shu Chen[1]
[1]Institute of Statistics and Information Science, National Changhua University of Education, Changhua

Spatial regression models are often used to predict spatial variables of interest, where the model parameters are usually estimated by the maximum likelihood method or the restricted maximum likelihood method. Particularly, the parameters involved in the spatial correlation function usually cannot be well estimated even when increasing amounts of data are collected densely in a fixed domain. This would result in inaccurate variable selection and spatial prediction. Different from the likelihood-based methods, in this talk, we propose a new parameter estimation method based on L2 risk, in which the uncertainty of parameter estimation is considered via a data perturbation technique. Therefore, the resulting estimators have less bias, and accurate variable selection and spatial prediction can be obtained. Some statistical inferences regarding the proposed method will be justified theoretically and the advantages of our method will also be investigated numerically.

Key words: Estimation uncertainty, Infill asymptotics, Spatial prediction, Stein's unbiased risk estimate, Variable selection.

Natural Statistics For Spectral Samples

Elvira Di Nardo

We introduce spectral sampling in association with the group of unitary transformations, acting on matrices in much the same way that simple random sampling is associated with the symmetric group acting on vectors. A spectral sample is a vector of eigenvalues of a random matrix obtained by suitably weighting a random sample by pre and post multiplication with truncated rectangular Haar matrices. Then symmetric functions are introduced, paralleling the classical k-statistics and polykays, as unbiased estimators of cumulants of trace powers. The explicit expressions for these spectral k-statistics are obtained by using symbolic techniques, which have been successfully employed in speeding up the computation of classical univariate and multivariate polykays [1]. Behind their statistical properties, spectral samples turn to be useful in elucidating some of the concepts associated with freeness - free probability and free cumulants - in terms of spectral k-statistics. For this purpose, spectral sampling may be viewed as a restriction operation from a freely randomized Hermitian matrix of order n into a freely randomized Hermitian matrix of order m less than n and each spectral k-statistic is a class function depending only on the matrix eigenvalues. Finally, by considering the limit as n going to infinity, we show that the normalized spectral k-statistics are related to free cumulants in much the same way that polykays are related to ordinary cumulants. This contribution is a joint work with P.McCullagh and D. Senato [2].

References

[1] Di Nardo, E., Guarino, G., Senato D. (2008) A unifying framework for k-statistics, polykays and their multivariate generalizations. Bernoulli, vol. 14, 440 - 468.

[2] Di Nardo, E., McCullagh P., Senato D. (2013) Natural statistics for spectral samples. Annals of
Statistics. 41(2), 982-
1004.

Using Bayesian Models To Optimise Cure And Minimise Side-effects In Cancer Radiotherapy

Alan Herschtal
Luc Te Marvelde[1], Kerrie Mengersen[2], Farshad Foroudi[1], Tomas Kron[1]
[1]Peter MacCallum Cancer Centre
[2]Queensland University of Technology

Radiation therapy is one of three major modalities of treatment for solid tumours, and is applied to over 50% of cancer patients. Conventional radiation therapy divides the total treatment into 20-40 small doses, called "fractions", delivered one per day. Because the tumour moves from fraction to fraction relative to reference markers used for beam alignment, the beams must be directed to deliver a high dose not only to the tumour, but also to a surrounding margin for error such that the tumour remains within the high dose field despite its motion. Setting margins too narrow leads to parts of the tumour being underdosed and hence likely cancer recurrence. Setting margins too wide leads to excess healthy tissue damage and hence excess side effects, which can be devastating. Recent advances in cancer imaging allow the tumour to be imaged immediately before each fraction, and demonstrate that patient tumours vary significantly in their amount of day-to-day motion (their "stability"). Traditional margin recipes, however, provide a "one size fits all" margin which is too large for more stable tumours, and too small for the less stable ones. A hierarchical Bayesian model is capable of optimising the margin for each individual patient by optimising the trade-off between the traditional generic group margin, and a personalised margin based on displacement data that accumulates as the individual progresses through treatment. We have constructed a novel two tiered Bayesian model which can accurately calculate personal margins despite time trends, unexpected "spikes", and sudden shifts in tumour position, thus ensuring maximum cure rates while minimising side-effects. Testing on real patient motion data shows an average reduction of 20% in radiation dose to healthy tissue relative to current methods, with no loss of dose to the tumour itself and thus no increase in the risk of recurrence.

Optimal Estimation Of Generalized Fractional Processes With Garch Errors

Gnanadarsha Dissanayake
Shelton Peiris[1], Tommaso Proietti[2]
[1]PhD Supervisor, School Of Mathematics And Statistics, University Of Sydney
[2]PhD Co-Supervisor, Department of Economics and Finance, University of Rome

This paper focuses on the approximation of long memory Gegenbauer processes driven by heteroskedastic errors using finite order moving average (MA) and autoregressive (AR) processes. The corresponding state space form is used to estimate the parameters by pseudo maximum likelihood using the support of the Kalman Filter.

A comparative assessment of the two approximation techniques is performed using a variety of stationary Gegenbauer processes and benchmarks. An extensive Monte Carlo experiment is implemented to establish the optimal order of each AR and MA approximation. Optimal orders of estimating models are established using the minimum sum of mean square errors of the trace estimators of the variance-covariance matrix. This methodology is extended to non-stationary Gegenbauer processes and compared with similar estimation techniques and results available in the literature.

Finally, the better approximating option is applied to the famous Standard and Poor (S and P) 500 intraday volatility time series to illustrate a real application of this new process.

Understanding Dispersal Of Asian Visitors To Australia

Alethea Rea
John Henstridge[1], Donna Hill[1], Kathy Haskard[1], Carmel McGinley[2]
[1]Data Analysis Australia
[2]Tourism Research Australia

UNDERSTANDING DISPERSAL OF ASIAN VISITORS TO AUSTRALIA

Henstridge, John 1, Hill, Donna2, Haskard2, Rea, Alethea3, Ahn, Inja4, McGinley, Carmel5

1 Principal Consultant Statistician and Managing Director, Data Analysis Australia, Perth

2 Senior Consultant Statistician, Data Analysis Australia, Perth

3 Consultant Statistician, Data Analysis Australia, Perth

4 Senior Analyst, Outreach and Stakeholder Support, Tourism Research Australia, Canberra

5 Manager – Outreach and Regional Research, Tourism Research Australia, Canberra

Dispersal in tourism relates to the locations visited and can be investigated using visitor itineraries. In Australia there are several well-known tourist hotspots including some state capitals (like Sydney and Melbourne) and some scenic locations (such as Tropical North Queensland). Tourists visiting only these locations are considered to have less dispersed itineraries than those who also visit other capitals or regional Australia. Using data from Tourism Research Australia's International Visitor Survey, we investigated the dispersal of visitors from Asian countries of origin. Results are presented as visualisations and animations.

Level Set Percolation Of The Gaussian Free Field

Alexander Drewitz
Pierre-François Rodriguez[1]
[1]ETH Zürich

We consider the Gaussian free field on the Euclidean lattice in dimensions larger than two. It is known that there exists a percolation phase transition for its level set; i.e., there exists a critical level such that for values h smaller than this critical level, the level set (or also excursion set) above level h exhibits a unique infinite connected component, whereas for values h larger than the critical level, it consists of finite connected components only.

We investigate some properties of the critical level and give some ideas on their proofs.

Statistics Nz Navigates Online Data Collection, Planning An Internet-based Survey.

Emma Mawby
Helen Smith[1], Colin Hewat[1]
[1]Statistics New Zealand

Statistics New Zealand is New Zealand's national statistics office. Statistics 2020 Te Kāpehu Whetū is our transformation programme. It enables us to obtain more value from official statistics, transform the way we deliver our statistics, and create a responsive, customer-focused, influential, and sustainable organisation.

Within Statistics 2020 Te Kāpehu Whetū, the Transform Collections Programme manages modernisation of collection systems, business processes, and supporting tools. Part of this modernisation is to research, develop, and integrate an Internet mode of collection to create a multi-modal collection environment for business surveys. This could reduce collection costs, better meet the needs of respondents, and improve response rates.

Producing an Internet mode requires expertise from many divisions of a national statistics office.

We had to determine the requirements for the Internet solution for business surveys; prioritise outcomes; and provide evidence about how introducing an Internet mode to business surveys would affect data processing and estimation and induce measurement effects.

Eighteen months into the adventure, we evaluate our progress, acknowledging our successes, and identifying areas for development. However, success has many colours. On the palette of a do-learn-do approach our greatest success was discovering that although we were equipped to deal with mountains, we were actually stumbling over molehills. What we thought would be our greatest challenges, such as mode effects and method of first contact, were dwarfed by unexpected and potentially show-stopping issues, such as security, and editing and imputation capability.

Our lesson is to map the terrain with greater precision before embarking on any adventure, never assuming we know the way. Here we summarise our experiences of planning an Internet mode for an annual survey of businesses.

Efficient Method Of Moments Estimators For Integer Time Series Models

Andrew Tremayne
Vance Martin[1], Robert Jung[2]
[1]University of Melbourne
[2]University of Hohenheim

The parameters of integer autoregressive models with equi-dispersed Poisson, or over-dispersed negative binomial innovations can be estimated by maximum likelihood where the prediction error decomposition, together with convolution methods, is used to write down the likelihood function. When a moving average component is introduced this is not the case. In this paper we consider the use of efficient method of moment techniques as a means of obtaining practical estimators of relevant parameters using simulation methods. Under appropriate regularity conditions, the resultant estimators are consistent, asymptotically normal and under certain conditions achieve the same efficiency as maximum likelihood estimators. Simulation evidence on the efficacy of the approach is provided and it is seen that the method can yield serviceable estimates, even with relatively small samples. Estimated standard errors for parameters are obtained using subsampling methods. Applications are in short supply with these models, though the range is increasing. We provide two examples using well-known data sets in the time series literature that have hitherto proved difficult to model satisfactorily; these both require use of specifications with moving average components.

Meta-analysis Of Height And Risk Of Type 2 Diabetes Mellitus

Md. Erfanul Hoque
Belal Hossain[1]
[1]Dept. of Statistics, Biostatistics & Informatics, University of Dhaka

Though the association between height and type 2 diabetes mellitus (T2DM) has been investigated by several epidemiological studies, the relationship remains unclear and height as risk factor for T2DM remains uncertain. The current meta-analysis of published studies and extracted studies aims to investigate the relationship between height and risk of T2DM. Nine studies (4 cross-sectional and 5 cohort studies) with 1338 T2DM cases and 24,403 controls were included by searching online databases (1970-December, 1012) and the references of retrieved articles. A follow-up data set collected by Bangladesh Institute of Research and Rehabilitation in Diabetes, Endocrine and Metabolic Disorders (BIRDEM) consist of 2936 patients (1984-1998) including 2535 T2DM cases, was also used. Logistic regression was used to estimate the odds ratio (ORs) and 95% confidence intervals (CIs) and generalized estimating equations (GEE) were used to estimate risk ratio for BIRDEM data. A meta-analysis of 9174 T2DM cases and 31242 controls from 18 cross-sectional studies and 6982 T2DM cases and 45516 controls from 9 cohort studies including BIRDEM data was conducted to estimate the pooled relative risks. A fixed effects model for cross-sectional studies and a random effects model for cohort studies were used to estimate the pooled relative risks and 95% CIs. The summary estimates indicated the significance inverse association between height and T2DM for cross-sectional studies [RR, 0.79; 95% CI, 0.72-0.87] and cohort studies [RR, 0.90; 95% CI, 0.83-0.98]. Meta-analysis of pooled relative risks also showed the inverse association in both men and women. These results are consistent between cross-sectional and cohort studies. The inverse association between height and T2DM risk was significant both in men [RR, 0.85; 95% CI, 0.80-0.91] and women [RR, 0.92; 95% CI, 0.86-0.99] for cross-sectional studies. Finally, these findings indicated that persons with high stature are at substantially lower risk of developing T2DM.

Ballisticity Conditions For Random Walk In Random Environment (rwre)

Alexander Drewitz
Noam Berger[1], Alejandro RamÃrez[2]
[1]Hebrew University of Jerusalem and TU Munich
[2]Pontificia Universidad Católica de Chile

During the last decades a significant amount of work has been devoted towards a better understanding of trapping phenomena and the closely related goal of obtaining conditions for ballistic behavior of RWRE. While the one-dimensional situation is well-understood, in higher dimensions the situation is significantly more involved. In 2002 Sznitman introduced a family of conditions which are possible candidates for characterizing ballistic behavior and which have had a significant impact on research in RWRE. They require the stretched exponential decay of certain slab exit probabilities for the random walk under the averaged measure.

We show that in dimensions greater than or equal to two, in order to establish these conditions it is actually enough to check a corresponding condition of polynomial type on a finite box.

As a corollary of this result, the conjectured equivalence of the family of conditions introduced by Sznitman is extended to all dimensions larger than or equal to two.

Efficient And Accurate Imputation Of Kir Types From Snp Variation Data

Damjan Vukcevic
James Traherne[1], Sigrid Næss[2], Mark Lathrop[3], Tom Hemming Karlsen[2], Miriam Moffatt[4], William Cookson[4], John Trowsdale[1], Gil McVean[5], Stephen Sawcer[6], Stephen Leslie[7]
[1]Cambridge Institute for Medical Research, University of Cambridge, UK
[2]Research Institute of Internal Medicine, Oslo University Hospital Rikshospitalet
[3]McGill University and Genome Quebec Innovation Centre, Montreal, Quebec, Canada
[4]National Heart and Lung Institute, Imperial College London, UK
[5]Wellcome Trust Centre for Human Genetics, Oxford, UK
[6]Department of Clinical Neurosciences, University of Cambridge, UK
[7]Statistical Genetics, Murdoch Childrens Research Institute, Parkville, Australia

KIR (killer-cell immunoglobulin-like receptor) genes are of great interest in regard to resistance to viruses, autoimmune disease, reproductive conditions and cancer. They are highly variable, which makes measuring KIR gene variation expensive and time consuming. Thus, to date KIR has been understudied.

Another type of genetic variation, single nucleotide polymorphisms (SNPs), occurs prominently throughout the genome, and SNP variation is now measured routinely and cheaply in large cohorts of individuals. Imputation methods exist for unmeasured SNPs and for certain other genes, enabling the use of SNP data to directly study the effect of these genetic variants on, for example, disease risk and physical traits. However, currently no imputation methods exist for KIR genes.

We have developed an efficient and accurate statistical imputation method for KIR types using SNP variation data. We use a reference data set of 700 individuals from 350 families of European descent; measuring variation in both the KIR genes and at 300 SNPs located in the genome in the vicinity of the KIR genes. Our method is based on averaging of classification trees.

Validation experiments, using cross-validation and in a separate group of 1000 individuals, show accuracy for imputing KIR types is at least 95% for the majority of KIR genes, with better that 90% accuracy for the rest. Experience with other imputation of other genes indicates that accuracy is likely to improve with increased reference panel size, and that imputation is likely to perform well for non-European populations provided good reference data is available.

The high accuracy of the method will allow KIR data to be available for large cohort studies, meaning that disease association with KIR can be directly studied for the first time. This will facilitate significant insights into the role of KIR in human disease.

Structure Learning And Estimation Of Interventional Distributions In Causal Additive Models

Jan Ernest
Peter Buhlmann[1], JOnas Peters[1]
[1]Seminar for Statistics, ETH Zurich

We consider structure learning and estimation of causal effects in low- and high-dimensional causal (nonlinear) additive models. Our method is based on structural equation models that are additive in the variables and error terms, a natural extension of linear Gaussian structural equation models to the case of nonlinear additive functions. Recently, identifiability results have been proved for this model class (e.g. Peters et al., 2013). We present a novel algorithm based on the following main idea (Bühlmann et al., 2013): First, an order search among the variables is performed, which can be done with non-regularized maximum likelihood estimation, and then, feature selection is applied using sparse regression techniques. The decoupling of the two steps makes this approach very generic: We show high-dimensional consistency of the first step, and well-established methods and theory can be incorporated for the second part. We prove that an additional preliminary neighborhood selection permits us to use an unpenalized maximum likelihood approach even in the high-dimensional setting. The performance of the algorithm and its robustness against misspecification of the model, e.g. with nonlinear, non-additive functions in the model, are tested in various numerical experiments for the low- and high-dimensional setting. Finally, we present a new approach for estimating interventional distributions and causal effects in the given model class.

Bühlmann, P., Peters, J., Ernest, J. (2013). CAM: Causal Additive Models, high-dimensional order search and penalized regression. arXiv:1310.1533

Peters, J., Mooij, J., Janzing, D., Schölkopf, B. (2013). Causal discovery with continuous additive noise models. arXiv:1309.6779

Posterior Contraction In Sparse Bayesian Factor Models For Massive Covariance Matrices

Anirban Bhattacharya
Debdeep Pati[1], Natesh Pillai[2], David Dunson[3]
[1]Department of Statistics, Florida State University
[2]Department of Statistics, Harvard University
[3]Department of Statistical Science, Duke University

Sparse Bayesian factor models are routinely implemented for parsimonious dependence modeling and dimensionality reduction in high-dimensional applications. We provide theoretical understanding of such Bayesian procedures in terms of posterior convergence rates in inferring high-dimensional covariance matrices where the dimension can be larger than the sample size. To obtain the convergence rates, we construct test functions to separate points in the space of high-dimensional covariance matrices using insights from random matrix theory; the tools developed may be of independent interest. We also derive minimax rates and show that the Bayesian posterior rates of convergence coincide with the minimax rates up to a logarithmic term.

Improved Product Type Exponential Estimators Using The Linear Transformation To Auxiliary Variable

Mohammad Hanif
Mohammad Noor-ul-amin[1]
[1]National College Of Business Administration And Economics, Lahore, Pakistan

In survey sampling, the improved estimators for the population mean of study variable have been obtained using the transformations of auxiliary variables. In this context, Srivenkataramana (1980) established a transformation using the mean of non-sampled observations. Mohanty and Sahoo (1995) introduced linear transformation of auxiliary variable using the extreme value of auxiliary variable from the population. In this study, a linear transformation to the auxiliary variable has been introduced by combining the concepts of Srivenkataramana (1980) and Mohanty and Sahoo (1995). The transformation is based on the assumption of prior knowledge of population extreme values of auxiliary variable and the mean of non-sampled observations of auxiliary variable. The two product type exponential estimators have been proporsed using the proposed transformation. The mean square errors and biases have been obtained, up to first order of approximation. Theoretical comparison of proposed estimators has been made with some existing estimators. An empirical study has been conducted to verify the improvement in the proposed estimators.

Keywords: study variable, auxiliary variables, mean square error, bias, exponential estimators, percentage relative efficiency, simple random sampling without replacement

Bias And Estimation Issues In Cut-off Sampling

Dan Hedlin

Cut-off sampling is a specific situation where survey and administrative data are integrated. Although cut-off sampling is widely used in practice, for example in business surveys where small businesses are usually not sampled at all, there is little research on bias and estimation issues associated with cut-off sampling. Below the cut-off the data source is usually only administrative data, above it there are survey data on the variable of interest. This paper discusses different approaches to these issues.

Improving Meme Via A Two-tiered Significance Analysis

Emi Tanaka
Timothy Bailey[1], Uri Keich[2]
[1]Institute for Molecular Bioscience, The University of Queensland
[2]School of Mathematics and Statistics, The University of Sydney

With over 9000 unique users recorded in the first half of 2013 MEME is one of the most popular motif finding tools available. Reliable estimates of the statistical significance of motifs can greatly increase the usefulness of any motif finder. By analogy, it is difficult to imagine evaluating a BLAST result without its accompanying E-value. Currently MEME evaluates its EM generated candidate motifs using an extension of BLAST's E-value to the motif finding context. While we previously indicated the drawbacks of MEME's current significance evaluation we did not offer a practical substitute suited for its needs. Additionally, MEME relies on the E-value internally to rank competing candidate motifs.

Here we offer a two-tiered significance analysis that can replace the E-value in selecting the best candidate motif and in evaluating its overall statistical significance. We show that our new approach could substantially improve MEME's motif finding performance and would also provide the user with a reliable significance analysis. In addition, for large input sets our new approach is in fact faster than the currently implemented E-value analysis.

New Dependence Measures For Random Vectors

Artem Prokhorov
Ivan Medovikov

The objective of this paper is to develop a meaningful measure of vectorial dependence. First, we generalise the measure of Gaisser et al. (2010), which leads to a vector version of Hoeffding's multivariate Φ^2. We show that while it has several attractive properties, it is unable to completely separate dependence *between* random sets from dependence *within*.

We develop a new measure of association between random vectors which draws from the theory of linkage functions. Developed in Li et al. (1996), the concept of a linkage function is closely related to that of the copula in that it too specifies the relationship between the marginal distributions, but unlike the copula, it admits marginals of dimension greater than one. The proposed measure has the ability to "filter out" information about the association within vectors, and responds only to vectorial dependence. The measure is non-parametric, and can be computed directly from sample percentile ranks.

We derive the asymptotic properties of the new measure and compare them with those of the multivariate Hoeffding's Φ^2. We propose consistent estimators of these population quantities and look at the behavior of the estimators in finite samples. Finally, we consider several applications of the statistics, including a study of inter-regional financial contagion between the European, Asian and North American stock markets.

Change Point Inference For Time-varying Erdos-renyi Graphs

George Michailidis

We investigate a model of an Erdos-Renyi graph, where the edges can be in a set of finite states (e.g. present/absent). The states of each edge evolve as a Markov chain independently of the other edges, and whose parameters exhibit a change-point behavior in time. We derive the maximum likelihood estimator for the change-point and characterize its distribution. Depending on a measure of the signal-to-noise ratio present in the data, different limiting regimes emerge. Nevertheless, a unifying adaptive scheme can be used in practice that covers all cases. Finally, for appropriate choices of the parameters of the Markov kernels, the limiting distribution of the change-point exhibits long-range dependence. The model is illustrated on synthetic, as well as US Senate roll call data.

Aportioning Impacts Of Groundwater Abstractions Using Wavelets

Daniel Pagendam
Chris Turnadge[1], Catherine Moore[1]
[1]CSIRO Land and Water

We consider the problem of how to disentangle the impacts of various groundwater abstractions on groundwater levels at a given location in a landscape. Methods are developed using synthetic data obtained from a 3-dimensional MODFLOW groundwater model. Our study considers groundwater drawdown time series collected at four locations: (i) close to an irrigator; (ii) close to a coal seam gas (CSG) development; (iii) distant to groundwater abstractions; and (iv) at a central location, nearby to (i) and (ii). Using these time series, we present a method by which we infer the contributions of the activities at location (i) and (ii) to the drawdown dynamics at location (iv). This is achieved using the maximum overlap discrete wavelet transform (MODWT) to obtain multiple time series (called "details") at each of locations (i) - (iii) that correspond to dynamics operating within a number of frequency bands. We utilize these series within a Bayesian statistical model with a simple physical interpretation to draw conclusions about the contributions of irrigation and CSG abstractions to dynamics at location (iv). In addition to using these methods for inference, we also demonstrate how we can identify the optimal locations for monitoring wells to ensure that there is minimal confounding of the MODWT details from each location used in the analysis.

Many Species, Many Sites: The Spatial Block Bootstrap For Multivariate Ecology

Eve Slavich
Begona Peco[1], David Warton[2]
[1]Universidad Autonoma de Madrid
[2]University of New South Wales

Conservation biology is abundant with research questions involving spatial data. Spatial auto-correlation is a well-known problem to the field. Further, many problems involving multivariate species abundance data lend themselves to design based methods using resampling to account for the inter-species correlation. The spatial block bootstrap, where sites are resampled in blocks to account for spatial dependency, is already over 20 years old, yet it has seen virtually zero uptake in ecological fields and is not mentioned in ecological reviews of methods for dealing with spatial auto-correlation. Despite the block bootstrap being relatively easy to implement, selecting the block length parameter (the size of the blocks which are resampled) is less so. With the aim to encourage the addition of this idea to practitioner's toolbox, we explore the type of spatial multi-species problems it may be applied to and practical issues that need to be addressed such as block length selection guidelines.

Point Process Models On A Linear Netweork

Gopalan Nair
Adrian Baddeley[1]
[1]The University of Western Australia

Point patterns on a linear network are found in many applications. The ‘lines’ that form the network may be roads, rivers, railway lines, electrical wires, nerve fibres, airline routes or soil cracks. The ‘events’ or ‘points’ may be traffic accidents, vehicle thefts or street crimes; roadside trees or invasive species; urban green spaces, retail stores or roadside kiosks]; or insect nests. Ang, Baddeley and Nair (2012) gave a formal treatment of point patterns on linear networks.

The general theory of point processes easily handles the definition, construction and characterization of parametric point process models on a linear network, as well as space-time point processes. However, geometrical inhomogeneity of linear networks hampers the construction of models with desired properties. In this presentation we give method of construction of some models for point patterns on linear networks.

References:

Q.W. Ang, A. Baddeley, and G. Nair. (2012) Geometrically corrected second order analysis of events on a linearnetwork, with applications to ecology and criminology. Scandinavian Journal of Statistics,. doi: 10.1111/j.1467-9469.2011.00752.x

Simplifying Species Interactions With Covariance Selection

Gordana Popovic
David Warton

A major challenge in multivariate analysis is building a plausible but sufficiently parsimonious model for covariance that it can be estimated when the number of response variables is not large compared to the number of observations. In the context of ecology, this problem arises when modeling a community of species, where there are a large number of potential species interactions relative to the number of locations where species abundances have been observed. Modeling these covariances not only gives insight into how species interact with one another, but they also allow us to build models which take these interactions into account when making inferences about associations between a community and the environment, or potential environmental impacts. We propose covariance selection as an approach to modeling species interactions – that is, assuming that many species do not interact directly, hence the inverse covariance matrix is sparse. Discovering this sparsity pattern can lead to a parsimonious estimate of the covariance structure. We will discuss the application of covariance selection to the context of community ecology – requiring extension of the method to deal with high dimensional, overdispersed count data.

Practical Spatial Statistics: The Markov Property And Stochastic Partial Differential Equations

Daniel Simpson
Finn Lindgren[1], Geir-Arne Fuglstad[2], Havard Rue[2]
[1]University of Bath
[2]NTNU

In this talk I will show how the continuous Markov property, in concert with modern numerical methods, can be used to build computationally efficient models across various areas of spatial statistics, from smoothing splines to point processes. In particular, I will focus on the use of these techniques to develop practical, flexible models of non-stationary and multivariate phenomena.

Maxima Of Long Memory Stationary Alpha-stable Processes

Gennady Samorodnitsky
Takashi Owada[1]
[1]Technion, Israel Institute of Technology

We derive a functional limit theorem for the partial maxima process based on a long memory stationary alpha-stable process. The length of memory in the stable process is parameterized by a certain ergodic theoretical parameter in an integral representation of the process. The limiting process is no longer a classical extremal Frechet process. It is a self-similar process with alpha-Frechet marginals, and it has the property of stationary max-increments, which we introduce in this paper. The functional limit theorem is established in the space D[0,infty) equipped with the Skorohod M1-topology; in certain special cases the topology can be strengthened to the Skorohod J1-topology.`

Multivariate Normal Approximation By Stein's Method: The Concentration Inequality Approach

Xiao Fang
Louis H. Y. Chen[1]
[1]National University of Singapore

The concentration inequality approach for normal approximation by Stein's method is generalized to the multivariate setting. We use this approach to prove a non-smooth function distance for multivariate normal approximation for standardized sums of k-dimensional independent random vectors with an error bound of order square root of k times the sum of absolute third moments. We further study multivariate normal approximation for sums of locally dependent (unbounded) random vectors, where the induction approach and the recursive approach are not likely to be applicable. We obtain a fourth moment bound as well as a third moment bound with an additional log term. We apply our results to the graph coloring problem and the joint distribution of sums of partial products in an i.i.d. sequence.

Shape Constrained Nonparametric Smooth Estimators In The Cox Model

Gabriela Nane
Hendrik Lopuhaa[1], Geurt Jongbloed[1]
[1]Deflt Institute of Applied Mathematics, Delft University of Technology

In survival analysis, Cox proportional hazards model is the typical choice to account for the effect of covariates on the lifetime distribution. The model relates the hazard of each subject with a given covariate vector to the baseline hazard, that corresponds to the null covariate vector, and an exponential function of covariates. Even though the baseline hazard can be left completely unspecified, in practice, one might be interested in restricting it qualitatively. This can be done by assuming the baseline hazard to be monotone, for example, as suggested by Cox himself when proposing the model.

Furthermore, it is desirable to also account for the fact that the underlying baseline hazard function is usually assumed to be a smooth function. Therefore, we consider estimators of a baseline hazard function under the assumption that the baseline hazard is nondecreasing and smooth. We obtain these estimators by kernel smoothing two monotone estimators, the maximum likelihood and the Grenander-type estimator. The maximum likelihood estimator maximizes the likelihood function over all distributions with nondecreasing hazard functions. The Grenander-type estimator is defined as the left-hand slope of the greatest convex minorant of the Breslow estimator of the cumulative baseline hazard function.

Depending on the choice of shape constrained estimators and when the smoothing is performed, four different estimators are proposed. With this respect, we define a smoothed maximum likelihood estimator (SMLE) and a maximum smoothed likelihood estimator (MSLE), as well as a smoothed Grenander-type (SG) estimator and a Grenander-type smoothed (GS) estimator. We investigate the asymptotic properties of all four estimators.

Re-thinking Soil Carbon Modeling: A Stochastic Approach To Quantify Uncertainties

David Clifford
Dan Pagendam[1], Jeff Baldock[2], Noel Cressie[3], Ryan Farquharson[2], Mark Farrell[2], Lynne Macdonald[2], Lawrence Murray[1]
[1]CSIRO Computational Informatics
[2]CSIRO Land & Water
[3]National Institute for Applied Statistics Research Australia

The potential exists to sequester more carbon within the world's soils. There are multiple benefits of doing so, including improved productivity, reductions in atmospheric greenhouse gases, and (in some parts of the world) financial gains through the sale of carbon credits. For these reasons, many deterministic models of soil carbon dynamics have been developed but none of these models address uncertainty in a comprehensive manner. Uncertainty arises in many ways - around the model inputs, parameters, and dynamics, and subsequently around model predictions. In this paper, we address these uncertainties together by incorporating a one-pool model for carbon dynamics within a physical-statistical model using a Bayesian hierarchical modelling framework. Our model is based on the soil carbon dynamics in the Tarlee region of South Australia and uses a dataset from a twenty-year agricultural productivity field trial that recorded observations of soil carbon. We specify the model conditionally through its parameters, soil carbon input processes, soil carbon decay process, and observations of those processes. We use a particle marginal Metropolis-Hastings approach specified using the LibBi modelling language. We highlight how samples from the posterior distribution can be used to summarise our knowledge about model parameters, to estimate the chances of sequestering carbon, and to forecast changes in carbon stocks under crop rotations not represented explicitly in the original field trials. Having demonstrated that this approach is possible, we discuss what would be required to extend our approach to biogeochemical-statistical models that incorporate the latest multi-pool models of soil carbon dynamics.

Sparse Recovery Under Weak Moment Assumption

Guillaume Lecue
Shahar Mendelson[1]
[1]Technion - ANU

We prove that iid random vectors that satisfy a rather weak moment assumption can be used as measurement vectors in Compressed Sensing and high dimensional Statistics. In many cases, the moment assumption suffices to ensure that the number of measurements required for exact reconstruction is the same as the best possible estimate -- exhibited by a random gaussian matrix. We also prove that this moment condition is almost necessary. Applications to the Compatibility Condition and Restricted Eigenvalue Condition in the noisy setup and to the neighborliness property of random polytopes are also discussed

A Method For Perturbing Quantiles To Protect Respondent Confidentiality

David Dufty
Gwenda Thompson[1], James Chipperfield[1]
[1]Australian Bureau Of Statistics

The Australian Bureau of Statistics has developed a method for perturbing quantile boundaries so as to protect the confidentiality of the respondents whose records comprise the dataset.

The method is used in an online tool called *TableBuilder*, which offers users a much greater degree of flexibility in generating their own customised analyses than was previously available with much ABS data. However, such flexibility presented a confidentiality risk. Specifically, given a large number of quantile calculations, it may be possible for a user to reconstruct the underlying records. In order to protect against this risk, a suite of algorithms was developed to strategically perturb the results of various analyses. We are not presenting here the full suite of privacy-protecting techniques but only the specific technique used to confidentialise quantiles.

The underlying idea is straightforward: the location of the true quantile boundary is perturbed by a small random number that is unknown to the user. Without knowing the exact value, the obtained result cannot be used to reverse-engineer the underlying unit record values. The size of the perturbation decreases with increasing sample size so that for tables with very large numbers of records the perturbation is negligible.

We present our method for calculating perturbed quantiles, describe the effect on relative standard error, and present an analysis of the effectiveness of the respondent protection. We also discuss the cost-benefit tradeoffs with data accuracy.

Keywords: confidentiality, statistical computing, official statistics

High Dimensional Stochastic Regression With Latent Factors, Endogeneity And Nonlinearity

Bin Guo
Jinyuan Chang[1], Qiwei Yao[2]
[1]University of Melbourne
[2]Peking University

We consider a multivariate time series model which represents a high dimensional vector process as a sum of three terms: a linear regression of some observed regressors, a linear combination of some latent and serially correlated factors, and a vector white noise. We investigate the inference without imposing stationary conditions on the target multivariate time series, the regressors and the underlying factors. Furthermore we deal with the the endogeneity that there exist correlations between the observed regressors and the unobserved factors. We also consider the model with nonlinear regression term which can be approximated by a linear regression function with a large number of regressors. The convergence rates for the estimators of regression coefficients, the number of factors, factor loading space and factors are established under the settings when the dimension of time series and the number of regressors may both tend to infinity together with the sample size. The proposed method is illustrated with both simulated and real data examples.

Modelling And Optimisation Of Group Dose-response Challenge Experiments

David Price

Body of Abstract:

An important component of scientific research is the 'experiment'. Effective design of these experiments is important and, accordingly, has received significant attention under the heading 'optimal experimental design'. However, until recently, little work has been done on optimal experimental design for experiments where the underlying process can be modelled by a Markov chain. In this talk, I will discuss some background and methods for optimal experimental design, and some of the work that I have done in applying this theory to dose-response challenge experiments for the bacteria *Campylobacter jejuni* in chickens.

Spatiotemporal Modelling Of Patient Safety Indicators

Hassan Assareh
Jack Chen[1], Lixin Ou[1], Stephanie Hollis[1]
[1]Simpson Centre for Health Services Research, Australian Institute of Health Inn

Postoperative complications and mortality have been considered in development of patient safety indices (PSIs) which are now frequently used and evaluated across hospitals and regions. Study of PSIs enables policy-makers and clinicians to better understand successes and failures of quality initiatives at hospital and regional levels.

We conducted a population-based study on all surgical patients admitted to acute public hospitals in New South Wales (NSW), Australia between 2002 and 2009 to explore the geo-temporal variation of two PSIs: failure-to-rescue (FTR), defined as death among surgical patients with treatable complications, and postoperative deep vein thrombosis/pneumonia embolism incidents (DVT/PE) among included patients.

Within a Bayesian disease mapping framework, we developed spatiotemporal Poisson mixed models for FTR and DVT/PE incidents to obtain areal- and time-specific relative risks. Integrated nested Laplace approximation (INLA) methods were employed for computation. We chose local government area (LGA) as the areal unit. LGA-aggregated covariates were generated and applied for risk adjustment. Several models capturing different temporal and space-time patterns were constructed and tested. Significant spatial and temporal patterns were found in both PSIs. The results demonstrated potentials for regional interventions and further investigation of best and worst practices.

Statistical Distributions From Complex N-body Systems

David Ni

Contemporary models for N-body systems are mainly based on temporal, two-body, and mass point representation of Newtonian mechanics. Other mainstream models include 2D and 3D Ising models based on the lattice structures. These models, however, encounter several on-going debates in statistics. We are therefore motivated to develop a new construction directly from complex-variable N-body systems based on the extended Blaschke functions (EBF), which represent a non-temporal and nonlinear extension of Lorentz transformation with a term of phase evolution, $\exp(-i\psi)$ on the complex plane – the normalized momentum space. A point on the complex plane represents a normalized state of particle momentum observed from a reference frame in the theory of special relativity. The nonlinear representation extends the constant momentum to include other energy variables, such as acceleration.

An algorithm similar to Jenkins-Traub method is adopted for solving EBF iteratively.
From the results of the numerical analysis, we have observed that the parameter-dependent convergent domains demonstrating continuum-to-discreteness transitions, evolutional invariance of distributions, phase transitions with conjugate symmetry, etc., which manifest the construction as a potential candidate for the unification of statistics. We hereby classify the observed distributions, which depending on nonlinearity, momentum, and scale of the finite convergent domains. Continuous and discrete distributions exist and are predictable for given partition.
Further introduction of random scattering mechanisms into the computation will results in quasi-canonical distributions, such as normal distribution.

Estimating Smooth Structural Change In Cointegration Models

Degui Li
Peter Phillips[1], Jiti Gao[2]
[1]Yale University
[2]Monash University

Abstract:

This paper studies nonlinear cointegration models in which the structural coefficients may evolve smoothly over time. These time-varying coefficient functions are well-suited to many practical applications and can be estimated conveniently by nonparametric kernel methods. It is shown that the usual asymptotic methods of kernel estimation completely break down in this setting when the functional coefficients are multivariate. The reason for this breakdown is a kernel-induced degeneracy in the weighted signal matrix associated with the nonstationary regressors, a new phenomenon in the kernel regression literature. Some new techniques are developed to address the degeneracy and resolve the asymptotics, using a path-dependent local coordinate transformation to re-orient coordinates and accommodate the degeneracy. The resulting asymptotic theory is fundamentally different from the existing kernel literature, giving two different limit distributions with different convergence rates in the different directions (or combinations) of the (functional) parameter space. Both rates are faster than the root-nh rate for nonlinear models with smoothly changing coefficients and local stationarity. Hence two types of super-consistency apply in nonparametric kernel estimation of time-varying coefficient cointegration models. In addition, local linear methods are used to reduce asymptotic bias and a fully modified kernel regression method is proposed to deal with the general endogenous nonstationary regressor case. Simulations are conducted to explore the finite sample properties of the methods and a practical application is given to examine time varying empirical relationships involving consumption, disposable income, investment and real interest rates.

Bayesian Bandwidth Estimation For A Functional Nonparametric Regression Model With Mixed Types Of Regressors And Unknown Error Density

Han Lin Shang

We investigate the issue of bandwidth estimation in a functional nonparametric regression model with function-valued, continuous real-valued and discrete-valued regressors under the framework of unknown error density. Extending from the recent work of Shang (2013) ['Bayesian Bandwidth Estimation for a Nonparametric Functional Regression Model with Unknown Error Density', Computational Statistics & Data Analysis, 67, 185–198], we approximate the unknown error density by a kernel density estimator of residuals, where the regression function is estimated by the functional Nadaraya–Watson estimator that admits mixed types of regressors. We derive a likelihood and posterior density for the bandwidth parameters under the kernel-form error density, and put forward a Bayesian bandwidth estimation approach that can simultaneously estimate the bandwidths. Simulation studies demonstrated the estimation accuracy of the regression function and error density for the proposed Bayesian approach. Illustrated by a spectroscopy data set in the food quality control, we applied the proposed Bayesian approach to select the optimal bandwidths in a functional nonparametric regression model with mixed types of regressors.

The Crt Is The Scaling Limit Of Large Random Dissections

Benedicte Haas
Nicolas Curien[1], Igor Kortchemski[2]
[1]CNRS, Université Paris 6
[2]Ecole Normale Superieure

We are interested in the graph structure of large random dissections of polygons sampled according to Boltzmann weights, which encompasses the case of uniform dissections or uniform p-angulations. As their number of vertices n goes to infinity, we will show that these random graphs, rescaled by $n^{-1/2}$, converge in the Gromov--Hausdorff sense towards a multiple of Aldous' Brownian tree when the weights decrease sufficiently fast.

Inference On Self-exciting Jumps In Prices And Volatility Using High Frequency Measures

Gael Martin
Catherine Forbes[1], Worapree Maneesoonthorn[2]
[1]Monash University
[2]University of Melbourne

This paper investigates the dynamic behaviour of jumps in financial prices and volatility. The proposed model is based on a standard jump diffusion process for price and volatility augmented by a bivariate Hawkes process for the respective jump components. The latter process specifies a joint dynamic structure for the price and volatility jump intensities, with the intensity of a volatility jump also directly affected by a jump in the price. The impact of the dynamic intensity process on the higher-order (conditional) moments for returns is investigated. In particular, the differential impacts of the jump intensities and the dynamic process for latent volatility itself, are measured and documented. A state space representation of the model is constructed using both financial returns and nonparametric measures of integrated volatility and price jumps as the observable quantities. Bayesian inference, based on a Markov chain Monte Carlo algorithm, is used to obtain a posterior distribution for the relevant model parameters and latent variables, and to analyze various hypotheses about the dynamics in, and the relationship between, the jump intensities. An extensive empirical investigation using data based on the S&P500 market index over a period ending in early-2013 is conducted.

Functional Graphical Models

Xinghao Qiao
Gareth James[1], Jinchi Lv[1]
[1]University Of Southern California

Recently there has been a great deal of interest in constructing graphical models, or networks, from high-dimensional data. A graphical model is used to depict the conditional dependence structure among p different random variables, $X = (X_1, \ldots, X_p)$. Such a network consists of p nodes, one for each variable, and a number of edges connecting a subset of the nodes. The edges depict the dependence structure of the p variables.

In this talk we are interested in estimating a graphical network in a somewhat more complicated setting. Let $g_1(t), \ldots, g_p(t)$ represent p random functions. Our goal is to construct a functional graphical model depicting the conditional dependence structure among the p functions. Functional data of this sort can arise in a number of contexts. For example, rather than observing only a static set of p gene expression levels at a single point in time, it is now becoming common to observe multiple expression levels over time, so $g_{ij}(t)$ would represent the expression level for subject i, of gene j, at time t.

To fit our functional graphical model we propose a convex penalized criterion which has connections to both the graphical lasso and the group lasso. Minimizing our functional graphical lasso (fglasso) criterion provides a sparse estimate for the edge set. We also propose an algorithm for efficiently minimizing the criterion. Our theoretical results demonstrate that the estimated functional network asymptotically converges to the true network structure even for p > n, expanding previous work from the standard setting to include more complicated functional data. We also examine the finite sample performance of the fglasso through a series of simulation studies.

Detecting Anisotropy Of Line Segment Processes

Farzaneh Safavimanesh
Jesper Moller[1], Jakob Rasmussen, Gulddahl[1]
[1]Department of Mathematical Sciences, Aalborg University, Center For Stochastic Geometry and Bioimagi

Detecting a special kind of anisotropy caused by columnar structure of points has been addressed in Møller *et al.* (2013) by introducing the cylindrical *K*-function and PLCPP models in d-dimensional spaces. In some cases, having more coordinates indicating the location of objects in a d-dimensional space enables us to associate a line segment to each object. We introduce a new summary statistics for detecting columnar anisotropy of line processes extending the definition of marked and intensity re-weighted stationarity for spatial point processes to the setting of directed line segments. An unbiased estimator of this summary statistics is also introduced. Note that this summary statistics is a counterpart of the cylindrical *K*-function (Møller *et al.* (2013)) which is a directional counterpart of the *K*-function.

We apply this function to investigate the minicolumns hypothesis in Neouroscience using 3D datasets collected by the Bioimaging group at Aarhus University associated to the Center for Stochastic Geometry and Advanced Bioimaging (CSGB).

On Outlier Accommodation And Influence Diagnostics For P-splines

Felipe Osorio

It has been well documented that the presence of outliers and/or extreme data can strongly affect smoothing via splines. This work proposes an alternative for accommodating outliers in penalized splines considering the maximum penalized likelihood estimation under the class of scale mixture of normal distributions (SMN). This family of distributions include heavy-tailed distributions, such as Student-*t*, contaminated normal, slash, among others, and have been an interesting alternative to produce robust estimates, keeping the elegancy and simplicity of the maximum likelihood theory. The aim of this paper is to develop a variant of the EM algorithm for computing efficiently the penalized maximum likelihood estimates in the context of penalized splines. In order to highlight some aspects of the robustness of the proposed penalized estimators we consider the assessment of influential observations, as well as certain perturbation schemes in the model or data considering influence diagnostics through case deletion and local influence methods. Numerical experiments were carried out to illustrate the good performance of the proposed technique.

Nonparametric Estimation For Self-exciting Point Processes - A Parsimonious Approach

Feng Chen
Peter Hall[1]
[1]The University of Melbourne

There is ample evidence that in applications of self-exciting point process models, the intensity of background events is often far from constant. If a constant background is imposed, that assumption can reduce significantly the quality of statistical analysis, in problems as diverse as modelling the after-shocks of earthquakes and the study of ultra-high frequency financial data. Parametric models can be used to alleviate this problem, but they run the risk of distorting inference by misspecifying the nature of the background intensity function. On the other hand, a purely nonparametric approach to analysis leads to problems of identifiability; when a nonparametric approach is taken, not every aspect of the model can be identified from data recorded along a single observed sample path. In this paper we suggest overcoming this difficulty by using an approach based on the principle of parsimony, or Occam's razor. In particular, we suggest taking the point-process intensity to be either a constant or to have least differential entropy, in cases where there is not sufficient empirical evidence to suggest that the background intensity function is more complex than those models. This approach is seldom, if ever, used for nonparametric function estimation in other settings, not least because in those cases more data are typically available. However, our "ontological parsimony" argument is appropriate in the context of self-exciting point-process models.

Sample-based Uncertainty Quantification In High-dimensional Inverse Problems

Colin Fox

Inverse problems are a class of 'complex computer models' that have well-studied mathematical properties. For example, the forward map is typically compact, so can be approximated by a finite-dimensional operator to any desired accuracy. This property explains the extreme sensitivity to observation noise and model error displayed by inverse problems. These properties may also be exploited to build efficient MCMC algorithms for exploring the posterior distribution arising in inverse problems. Recently we have understood how to build efficient Gibbs samplers for important classes of inverse problems, and how to accelerate Gibbs samplers by drawing on the polynomial acceleration techniques from numerical computation.

Respiratory Disease Phenotypes In Early To Mid-childhood: A Latent Transition Analysis

Frances Garden
Judy Simpson[1], Craig Mellis[2], Guy Marks[3]
[1]Sydney School of Public Health, University of Sydney
[2]Central Clinical School, University of Sydney
[3]Woolcock Institute of Medical Research

Background: It is well known that asthma and related respiratory disease is a heterogeneous entity whose manifestations vary with age. Our aim was to describe this heterogeneity in terms of transitions between empirically-derived phenotypes over time based on observations of several manifestations of respiratory disease made repeatedly over time in a birth cohort of children at risk of asthma.

Methods: We used data on respiratory symptoms, health care utilisation, medications, lung function, airway hyper-responsiveness (AHR), exhaled nitric oxide concentration (eNO) and atopy from the Childhood Asthma Prevention Study to define the phenotypes. Data were acquired at ages 1.5, 3, 5, 8 and 11.5 years (AHR, eNO and lung function at 8 and 11.5 years only). Data were analysed using latent transition analysis.

Results: At all time-points we could classify subjects (n=370) into three phenotypes broadly described as: 1) "predominant wheeze" (prevalence:25%-29%); 2)"predominant cough and sneeze" or "minor symptoms with atopy" (prevalence:22%-31%); and 3)"minor symptoms without atopy" (prevalence:44%-53%). Before age five years, transition to different phenotypes over time was common, but phenotypes were more stable after age five. The most common trajectory path was to start and remain in the "minor symptoms without atopy" phenotype (20%). Only 40% of the sample classified as "predominant wheeze" at 1.5 years belonged to this class at 11.5 years.

Conclusion: This longitudinal analysis represents the first attempt to incorporate longitudinal patterns of several manifestations of chronic and recurrent respiratory disease in childhood into a single model. It provides quantitative support for the common observations that respiratory disease in children is a heterogeneous entity. It is clear that some children with wheeze and other respiratory symptoms in early life progress to asthma in mid-childhood, while others become asymptomatic.

The Binomial N Problem: A Consensus Approach

Frank Tuyl

The binomial N problem, based on multiple binomial counts and unknown population proportion θ as well as unknown sample size N, was considered by various authors, and quite recently by Berger et al. (2012) and Aitkin (2010). Kahn (1987) made an important two-page contribution, outlining the constraints, in the Bayesian approach, on the prior for (N, θ).

The usual interest is in N, θ being a nuisance parameter, but the above authors did not consider confidence/credible intervals. Two arguments are made to support a proposed consensus interval for N, based on a noninformative prior. First, the different priors for N derived by Berger et al. (2012), for the known θ and unknown θ cases, seem to lead to an inconsistency, an indication that their preferred Jeffreys prior for θ is inadequate and that the uniform or Bayes-Laplace prior for θ should be adopted instead. Second, given the marginal posterior for N, based on adding a less informative prior for N than is possible when combining with the Jeffreys prior for θ, it seems possible to improve on central intervals, or the highest posterior density intervals adopted by Rafterty (1988); these intervals were based on a slightly different marginal posterior than the one proposed here.

Aitkin, M. (2010), *Statistical Inference - An Integrated Bayesian/Likelihood Approach*, University Press, Cambridge.

Berger, J.O., Bernardo, J.M. and Sun, D. (2012), Objective priors for discrete parameter spaces, *Journal of the American Statistical Association*,107, 636-648.

Kahn, W.D. (1987), A cautionary note for Bayesian estimation of the binomial parameter *n*, *The American Statistician*, 41(1), 38-40.

Rafterty, A.E. (1988), Inference for the Binomial *N* Parameter: A Hierarchical Bayes Approach, *Biometrika*, 75, 223-228.

Sources Of Variability In The Estimation Of Malaria Parasite Density

Imen Hammami
ANDRE GARCIA[1], GREGORY NUEL[2]
[1]IRD UMR 216 Paris Descartes University
[2]MAP5 Paris Descartes University

Microscopic examination of stained thick blood smears (TBS) is the gold standard for routine malaria diagnosis. The level of infection, expressed as parasite density (PD), is classically defined as the number of asexual parasites relative to a microliter of blood. PD estimation methods usually involve threshold values; either the number of white blood cells counted or the number of high power fields read. In these methods, the number of parasites and the number of leukocytes per HPF are assumed to be Poisson-distributed. The objective of this paper is to explore two sources of variability in PD estimation methods: the sampling error and the overdispersion in the distribution of parasites and leukocytes in TBSs.

We studied the statistical properties (mean error, coefficient of variation, false negative rates) of parasite density estimators of commonly used threshold-based counting techniques depending on variable threshold values. Two sources of overdispersion in field-collected data are investigated: latent heterogeneity and spatial dependence. Unobserved heterogeneity in data was accounted for by considering more flexible models that allow for overdispersion. Of particular interest were the negative binomial model (NB) and mixture models. The dependent structure in data was modeled with hidden Markov models (HMMs).

We described the behavior of measurement errors according to varying threshold values through colormaps. We gave insights on what should be the optimal threshold values that minimize this variability. We showed that the Poisson assumption is inconsistent with parasite and leukocyte distributions per HPF. Among simple parametric models, the NB model is the closest to the unknown distribution that generates the data. On the basis of model selection criteria AIC and BIC, HMMs provided a better fit to data than mixtures. Ordinary pseudo-residuals confirmed the validity of HMMs.

Using Landscape Attributes To Predict Soil Carbon By Combining Random And Non-random Samples

Gavin Melville
Cathy Waters[1], Susan Orgill[1]
[1]NSW Department of Primary Industries

A study in western NSW is investigating the relationship between soil carbon and a number of key landscape attributes including ground cover, percentage of perennial species and distance to the nearest tree or shrub. Relating soil carbon to these types of auxiliary variables is best achieved using model-based prediction and non-random sampling. However in order to estimate carbon stock across a spatial region such as a paddock or geographic land unit we also require an estimate of the predictor variables in the soil carbon model and, in the absence of detailed vegetation maps, this is best achieved using random sampling. A survey design which incorporates both random and non-random components is discussed.

James (1964) And Spiked Models In Multivariate Analysis

Iain Johnstone

James' paper set out a 'five fold way' for classifying the distributions associated with many of the methods of classical multivariate analysis, such as principal components, canonical correlations and multiple response regression. With high dimensionality a popular theme in modern research, James' framework remains helpful. We will review some results that can be obtained by considering "low rank" alternatives to the traditional null hypotheses.

Imputing Rent For Australian Owner-occupied Dwellings: A New Methodology

Franklin Soriano
Cristian Rotaru[1], Sezim Dzhumasheva[1]
[1]Analytical Services Unit, Australian Bureau of Statistics

In 2008, ABS released its first experimental estimates of imputed rent for owner-occupied dwellings (OODs) and subsidised rentals in the ABS household income statistics. The availability of imputed rent estimates has significantly extended the range of analyses that can be undertaken using these statistics, both internally and externally, and has filled a significant data gap in the ABS household income measures.
In computing the gross imputed rent estimates, the current methodology employs a hedonic modelling procedure, with a Heckman correction and an extrapolation adjustment for the higher end values. The estimates were derived using data from the Household Income and Expenditure Survey (HIES) 2003-04 and the Survey of Income and Housing (SIH) 2005-06.
In this paper, the authors developed a stratification-based estimation procedure as an alternative to the current methodology using aggregated data from the 2006 Census linked to the 2005-07 Valuers-General (VGs) data at the collection districts (CD) level. The new methodology is based on a simple and a neat way of imputing the rental rate for owner-occupied dwellings during census years. The empirical results are promising.

Evaluation And Assessment Of Clustering Approaches For High-dimensional And Functional Data

Inge Koch

In supervised learning a decision rule can be assessed by error criteria which admit a comparison of different rules and therefore allow the selection of good or optimal rules for specific applications. In cluster analysis or unsupervised learning, this luxury is lost, and the exploratory nature of clustering approaches makes it difficult to compare and assess different approaches and thus arrive at an informed decision about which method to choose. This problem is exacerbated for high-dimensional data which may not naturally fall into tight clusters.

Motivated by problems arising in the clustering of high-dimensional or functional profiles of proteomics mass spectrometry data, which are measured on thousands of variables, we illustrate different cluster results that can be obtained with natural variations of the conventional k-means cluster algorithms. To obtain an insight into the differences and relationships of cluster assignments, we propose a number of criteria for assessing results of cluster analyses. These include a statistic based on chi-squared distances of profiles, the Jaccard distance and suitable modifications of the prediction strength approach of Tibshirani and Walther (2005). In addition we propose visualisation tools for summarising the statistics based on these assessment criteria and illustrate clustering algorithms and their assessment for proteomics mass spectrometry data.

Gpu Accelerated Mcmc For Bayesian Evaluation Of A Self-exciting Process With Non-constant Baseline For The Analysis Of Terrorist Activity, And The Identification Of Significant Parent Events

Gentry White
Michael Porter[1]
[1]University of Alabama

Hawkes process, or self-exciting models are used to model a variety of phenomenon that occur in clusters either in space or in time. The Hawkes process, or self-exciting models combine a baseline with self-exciting term. In the standard Hakes model the baseline is assumed to be a constant forcing all variation to be explained by the small-scale self-excitation effects. In practice it may be useful to identify both large and small-scale variation. In cases where the exact source or structure of the large-scale variation is unknown a non-parametric function such as kernel density estimation or splines may be used to estimate this variation. Implementing this can be difficult as the addition of these functions to the model can lead to weak identifiability.

While this can make maximisation of the likelihood difficult, in a Bayesian context it can make MCMC practically impossible due to poor convergence. A negative-binomial convolution model is used to address this issue and computational approaches using GPU parallel processing to accelerate MCMC are discussed, as well as a cluster model interpretation of results to identify significant parent events. The model results are demonstrated using an example for terrorist activity in Colombia from 2000 through 2010.

Estimation Of Varying Coefficient Models With Randomly Censored Data

Ingrid Van Keilegom
Seong J. Yang[1], Anouar El Ghouch[1], Cedric Heuchenne[1]
[1]Universite Catholique de Louvain

The varying coefficient model is a useful alternative to the classical linear model, since the former model is much richer and more flexible than the latter. We propose estimators of the coefficient functions for the varying coefficient model in the case where different coefficient functions depend on different covariates and the response is subject to random right censoring. Since our model has an additive structure and requires multivariate smoothing we employ a smooth backfitting technique, that is known to be an effective way to avoid "the curse of dimensionality" in structured nonparametric models. The estimators are based on synthetic data obtained by an unbiased transformation. The asymptotic normality of the estimators is established and a simulation study illustrates the reliability of our estimators.

The Optimal Fourth Moment Theorem

Ivan Nourdin
Giovanni Peccati[1]
[1]Luxembourg University

We will explain how to compute the exact rates of convergence in total variation associated with the Fourth Moment Theorem by Nualart and Peccati (2005), stating that a sequence of random variables living in a fixed Wiener chaos verifies a central limit theorem (CLT) if and only if the sequence of the corresponding fourth cumulants converges to zero. We will also provide an explicit illustration based on the Breuer-Major CLT for Gaussian-subordinated random sequences.

Multichannel Deconvolution With Long Range Dependence: Upper Bounds On The l^p-risk $(1 \le P < \infty)$

Justin Rory Wishart
Rafal Kulik[1], Theofanis Sapatinas[2]
[1]University of Ottawa
[2]University of Cyprus

We consider multichannel deconvolution in a periodic setting with long-memory errors under three different scenarios for the convolution operators, i.e., super-smooth, regular-smooth and box-car convolutions. We investigate global performances of linear and hard-thresholded non-linear wavelet estimators for functions over a wide range of Besov spaces and for a variety of loss functions defining the risk. In particular, we obtain upper bounds on convergence rates using the L^p-risk $(1 \le p < \infty)$. Contrary to the case where the errors follow independent Brownian motions, it is demonstrated that multichannel deconvolution with errors that follow independent fractional Brownian motions with different Hurst parameters results in a much more involved situation. An extensive finite-sample numerical study is performed to supplement the theoretical findings.

Using Genetic Sequences To Infer Population Dynamics: Phylodynamic Analysis Of Hiv Transmission

Edward Ionides

Advances in methods for analysis of population-level genetic variation of pathogens can potentially provide useful information about characteristics of donors of infections. This complements conventional epidemiological surveillance of infectious disease, which is focused on identifying recipients of infection. Pathogen genetic sequences are increasingly available for a growing number of infectious diseases. We discuss recent methodological developments that use both sequence data and conventional epidemiological data for inference on dynamic models of disease transmission. As a specific example, we estimate the fraction of HIV transmission that occurs in the first year of the donor's infection. We find that combining conventional and genetic data gives substantial improvement over each alone.

Kernel Principal Components Analysis For Remote Sensing.

Georgina Davies
Noel Cressie[1]
[1]Distinguished Professor, National Institute for Applied Statistics Research Australia, University of

Large datasets are becoming increasingly prevalent in remote sensing as our ability to collect data grows, outpacing our ability to analyse these datasets using traditional statistical methods. Thus statistical methods that can, in a computationally efficient manner, model and analyse large datasets are in demand. Kernel principal components analysis (KPCA) is a non-linear dimension-reduction technique that has been introduced in the geostatistics literature. Polynomial kernels of up to order 3, Gaussian kernels, and distance kernels have been explored in the KPCA literature. However, there are a large number of potential kernels (e.g., exponential and Laplacian) that have not yet been considered. The research presented investigates different kernels in terms of their methodological and computational consequences for KPCA.

Rough Flows

Ismael Bailleul

A simple method for constructing flows of maps on a Banach space from approximate flows was recently introduced in "*Flows driven by rough paths"*. Not only does it provide an elementary path to recover and extend most of the results on rough differential equations driven by finite and infinite dimensional rough signals, but it also happens to be tailor made to deal with path-dependent (potentially anticipative) rough evolutions, and for extending Le Jan-Watanabe-Kunita's framework for stochastic flows to a non-semimartingale setting. The talk will provide an account of the method and its use.

Kent Mixed Effects Models For Compositional Data

Janice Scealy
Alan Welsh[1]
[1]Centre for Mathematics and Its Applications, ANU

Compositional data are vectors of proportions defined on the unit simplex and this type of constrained data occur frequently in applications. It is also possible for the compositional data to be correlated due to the clustering or grouping of the observations. We propose a new class of mixed model for compositional data based on the Kent distribution, where the random effects also have Kent distributions. The advantage of this approach is that it handles zero components directly and the new model has a fully flexible underlying covariance structure. One useful property of the new Kent mixed model is that the marginal mean direction has a closed form and is interpretable. In compositional data settings the mean proportions are usually of primary interest and these are shown to be simple functions of the marginal mean direction. For estimation we apply a quasi-likelihood method which results in solving a new set of generalised estimating equations and these are shown to have low bias in typical situations. For inference we use a nonparametric bootstrap method for clustered data which does not rely on estimates of the shape parameters (shape parameters are difficult to estimate in Kent models). We analyse data from the 2009-10 Australian Household Expenditure Survey CURF (confidentialised unit record file). We predict the proportions of total weekly expenditure on food and housing costs for households in a chosen set of domains (small areas). The new approach is shown to be more tractable than the traditional approach based on the logratio transformation.

On Semiparametric Bernstein-von Mises Theorems

Ismael Castillo

In this talk I will discuss semiparametric estimators based on the Bayesian posterior distribution. In particular, the focus will be on Bernstein-von Mises (BvM) theorems for semiparametric functionals. I will review some recent results on this topic, including sufficient conditions for BvM for a variety of priors and models, as well as functional Donsker-type theorems for posterior distributions in i.i.d. sampling.

Nonparametric Independence Test Based On Copula Density

Gery Geenens

The concept of independence is central in statistics, and being able to test the assumption of independence between two random variables X and Y is evidently very important. In this work, such an independence test is proposed. It is able to detect any departure from the null hypothesis of independence (omnibus test) between two continuous random variables X and Y, without relying on any particular parametric assumptions on the underlying distributions (nonparametric test). The test is based on copula modelling, which has lately become a very popular tool for analysing the dependence structure of a random vector. Specifically, the test statistic is a Cramer-von Mises-type discrepancy measure between a (boundary-bias-corrected) kernel estimate of the copula density of (X,Y) and the independence copula density. Basing an independence test on the copula density has numerous advantages that will be discussed. In particular, this makes the test very powerful at detecting subtle departures from the null hypothesis of independence in any direction. This is explained through theoretical considerations and illustrated by substantial simulation studies.

Operationally Defined Effect Size Recommendations For The Odds Ratio And Relative Risk

Jake Olivier
Melanie Bell[1]
[1]University of Arizona

Sample size calculations are an important part of research to balance the use of resources and to avoid undue harm to participants. Effect sizes are an integral part of these calculations and meaningful values are often unknown to the researcher. General recommendations for effect sizes have been proposed for several commonly used statistical procedures, yet recommendations do not exist for the odds ratio or relative based on objective reasoning. Cohen proposed operationally defined recommendations for the correlation coefficient phi for binary data; however, it is well known that phi suffers from poor statistical properties. We will discuss odds ratio recommendations that are anchored to phi for fixed marginal probabilities. It will further be demonstrated that the marginal assumptions can be relaxed resulting in more general results, and the results will be extended to the relative risk and risk difference.

Bayesian Spatial Modelling Of Type Ii Diabetes In Queensland

Jannah Baker
Nicole White[1], Kerrie Mengersen[1]
[1]Queensland University of Technology

Type II diabetes mellitus (DM II) is an increasingly prevalent disorder in Australia and worldwide, leading to large healthcare and productivity costs. The identification of areas at excess risk of DM II prevalence is useful to guide allocation decisions around screening and service provision. We report findings from Bayesian spatial modelling of DM II prevalence across Queensland, accounting for missing data and spatial structure of outcomes and covariates. Covariates included in models include gender and age distribution, socioeconomic status, proportion overweight or obese*, proportion of daily smokers*, proportion with insufficient physical activity*, sufficient fruit consumption* and sufficient vegetable consumption* (*based on self-reported data). A cross-validation method was used to compare imputation methods for missing data and the most accurate method applied. Sensitivity analysis was performed to compare use of different prior distributions. This work was conducted in collaboration with Dr. Nicole White and Prof. Kerrie Mengersen.

Estimation Of The Treatment Effect On Mean And Dispersion

Gillian Heller
Dominique-Laurent Couturier[1], Stephane Heritier[1]
[1]Department of Statistics, Macquarie University

In clinical trials traditionally the effect of a treatment on the mean of an endpoint of interest is hypothesised. The underlying assumption in statistical models for the mean, is that the effect of the treatment on the response distribution is a location shift, with other aspects of the distribution (shape/dispersion/variance) remaining the same. We consider data from a clinical trial for a treatment hypothesised to reduce the mean number of falls, in which it is apparent that the treatment not only reduces the mean number of falls, but also the variability in falls. As the response is overdispersed, potential statistical models include Poisson mixture distributions such as the negative binomial and Poisson-inverse Gaussian (PiG), which are typically parametrised in terms of a mean and dispersion parameter. For our clinical data, the PiG was found to provide a good fit. The conventional analysis hypothesises a treatment effect on the mean, either adjusted or unadjusted for covariates, while assuming a constant dispersion parameter. On our data, this analysis yields a non-significant treatment effect. Mean and dispersion models, or more recently generalized additive models for location, scale and shape, allow linear models to be specified on the mean and dispersion parameter(s) of a broad range of distributions. We show that, on our data, if we model a treatment effect on both the mean and dispersion parameters, both effects are highly significant. In a simulation study we show that if a treatment effect exists on the dispersion parameter and is ignored in the modelling, estimation of the treatment effect on the mean can be severely biased. The question of what is meant by a treatment effect arises. This has implications in the planning of statistical analyses for clinical trials: should a treatment effect on the dispersion be prespecified?

On The Range Of Integration Of A Functional Linear Model

Giles Hooker

Body of Abstract:

This talk considers the problem of selecting the range of a functional covariate that is relevant to scalar-on-function regression. We consider a one-sided problem motivated by vehicular emissions data in which the end point of the integral in a functional linear model must be determined. We demonstrate that when the functional coefficient is defined non-parametrically, approaches such as cross validation do not yield satisfactory estimates. Instead, we propose a penalized criterion with a penalty chosen with reference to a parametric approximation to the model.

A Semiparametric Framework For Rank Tests

Jan De Neve
Olivier Thas[1], Jean-Pierre Ottoy[2]
[1]Ghent University and University of Wollongong
[2]Ghent University

We demonstrate how classical rank tests, such as the Wilcoxon-Mann-Whitney, Kruskal-Wallis, and Friedman tests can be embedded in a statistical modelling methodology and how our approach can be used for constructing new rank tests for more complicated designs. In particular, rank tests for unbalanced and multi-factor designs, and rank tests that allow for correcting for continuous covariates are included. The method also allows for the estimation of meaningful effect sizes. Our method results from two particular parametrizations of probabilistic index models (Thas et al., 2012).

Thas, O., De Neve, J., Clement, L. and Ottoy, J.P. (2012) Probabilistic index models (with discussion). Journal of the Royal Statistical Society - Series B. 74:623--671.

Trending Time Series Models With Non- And Semi-parametric Cointegration

Jiti Gao
Peter C. B. Phillips[1]
[1]Yale University

Abstract This paper studies a general class of nonlinear varying coefficient time series models with possible nonstationarity in both the regressors and the varying coefficient components. The model accommodates a cointegrating structure and allows for endogeneity with contemporaneous correlation among the regressors, the varying coefficient drivers, and the residuals. This framework allows for a mixture of stationary and nonstationary data and is well suited to a variety of models that are commonly used in applied econometric work. Nonparametric and semiparametric estimation methods are proposed to estimate the varying coefficient functions. The analytical findings reveal some important differences, including convergence rates, that can arise in the conduct of semiparametric regression with nonstationary data. The results include some new asymptotic theory for nonlinear functionals of nonstationary and stationary time series that are of wider interest and applicability and subsume much earlier research on such systems. The finite sample properties of the proposed econometric methods are analyzed in simulations.\ An empirical illustration examines nonlinear dependencies in aggregate consumption function behaviour in the US over the period 1960 - 2009.

Modelling Business Energy Consumption Using Agent-based Simulation Modelling

Jason Wong
Kay Cao[1]
[1]Australian Bureau of Statistics

Simulation techniques are becoming popular recently and have been applied to different research areas. An advantage of simulation over regression is that simulation assumes heterogeneity between individuals. It also allows interaction between individuals, and for the collective actions of agents to impinge on the individual actions. This enables more sophisticated models to be built to overcome the shortcomings of simplified ordinary least squares models.

This paper developed a prototype agent-based simulation model to estimate business energy consumption in the Australian manufacturing industry, taking into account business decision-making process in responding to government policies. The model developed in this study has been applied to a linked dataset of survey data, administrative data and synthetic price data to assess its feasibility.

Statistical Matching Using Fractional Imputation

Jae-Kwang Kim
Emily Berg

Statistical matching is a technique of integrating two or more data sets when information available for matching records for individual participants across data sets is incomplete. Statistical matching can be viewed as a missing data problem where a researcher wants to perform a joint analysis of variables that are never jointly observed. A conditional independence assumption is often used to create imputed data for statistical matching.

We consider an alternative approach of statistical matching based on an instrumental variable assumption. Parametric fractional imputation of Kim (2011) is applied to create imputed data under the instrumental variable assumption. Variance estimation is also discussed. The proposed method is directly applicable to the analysis of split questionnaire design and measurement error models.

Nonparametric Additive Models With Measurement Error

Jason Tran

There is growing interest in estimating nonparametric regression in the presence of measurement error. In this problem, strong theoretical foundations have been laid in the univariate setting, and a substantial understanding of the underlying difficulties has been achieved. When the true explanatory variable is measured with error, correcting for the error can reduce performance. As a result, we cannot expect that an estimator that needs to accommodate measurement error will perform competitively with an estimator in the error-free case. Reflecting this challenge, optimal rates of convergence in the measurement error setting are generally slower than their counterparts in the error-free case. Extending this problem to multivariate cases can reduce performance still further. However, if the true regression model is additive then we can potentially circumvent the curse of dimensionality. In this talk we discuss a simple and direct estimation procedure using orthogonal series representations, in cases where the true regression model is additive.

Contaminated Variance-mean Mixing Model

Joanna Wang
Thomas Fung[1], Seneta Eugene[2]
[1]Macquarie University
[2]University of Sydney

The Generalised Normal Variance–Mean (GNVM) model in which the mixing random variable is Gamma distributed is considered. This model generalises the popular Variance-Gamma (VG) distribution. This GNVM model can be interpreted as the addition of contamination to a (skew) VG base. We explore the possibility of using contamination as an instrument by specifying the level of contamination in the GNVM model. We demonstrate the idea using several simulated datasets. We show that the advantage of allowing for contamination is even more profound when the data has heavier tails. The discussion will be based on goodness of fit criteria but a more in-depth discussion of the criteria will be presented in another talk titled "Deviance Information Criterion in Comparison of Normal Mixing Models" by the same authors in this conference.

Is The Locally Best Invariant Test Uniformly Most Powerful For A Wider Class Of Invariant Tests?

Jahar Bhowmik

Abstract: In the context of a general regression model in which some regression coefficients are of interest and others are purely nuisance parameters, Bhowmik and King (2012) constructed the locally best invariant (LBI) test against one-sided alternatives. This paper investigates whether this LBI test is uniformly most powerful invariant (UMPI) or not through simulation results. A test that is locally best invariant against one-sided alternative hypotheses is found to be uniformly most powerful invariant (UMPI) in a wider class of tests than the invariant tests for the standard F test. The results of a simulation study conducted to prove that the LBI test is UMPI are presented.

Key words: Invariant; simulation; one-sided; nuisance parameters; F test.

Semiparametric Gee Analysis In Partially Linear Single-index Models For Longitudinal Data

Jia Chen
Degui Li[1], Hua Liang[2], Suojin Wang[3]
[1]University Of York
[2]George Washington University
[3]Texas A&M University

In this article, we study a partially linear single-index model for longitudinal data under a general framework which includes both the sparse and dense longitudinal data cases. A semiparametric estimation method based on the combination of the local linear smoothing and generalized estimation equations (GEE) is introduced to estimate the two parameter vectors as well as the unknown link function. Under some mild conditions, we derive the asymptotic properties of the proposed parametric and nonparametric estimators in different scenarios, from which we find that the convergence rates and asymptotic variances of the proposed estimators for sparse longitudinal data would be substantially different from those for dense longitudinal data. We also discuss the estimation of the covariance (or weight) matrices involved in the semiparametric GEE method. Furthermore, we provide some numerical studies to illustrate our methodology and theory.

Collapsed Variational Bayes For Model Selection

John Ormerod
Chong You[1], Matthew Stephens[1]
[1]University of Chicago

Collapsed variational Bayes (CVB) offers a potential approach for improving upon mean field variational Bayes (VB) approaches, both of which are approximate Bayesian inference alternatives to Markov chain Monte Carlo methods. In this talk we show how CVB is one approach for overcoming VB approaches which can fail for high-dimensional, low sample size problems. We will consider model selection for linear models, generalized linear models and linear models with missing covariates using a spike and slab prior for the coefficients. In our numerical examples we show how our CVB approaches are extremely efficient, and perform well in terms of variable selection and prediction in comparison to some popular alternative methods for the same task.

Testing For Serial Dependence In Binomial Time Series Regression

Jieyi He
William Dunsmuir

Detection and estimation of serial dependence in binomial response time series is considerably more difficult than it is in continuous response time series. This talk reviews various methods for detecting serial dependence in time series of binomial observations suitable for the regression setting. Methods include the traditional Box-Ljung-Pearce statistic based on autocorrelations of Pearson residuals, score tests against two classes of dependence models, Wald test and likelihood ratio tests in generalized linear autoregressive moving average models for binomial and binary counts and analogous results for parameter driven (latent process) models are presented. Performance of these methods is based on large sample asymptotic results, simulations and applications to real data sets. Power against various alternatives of the different methods is illustrated via simulation results. Examples include modeling binary time series of economic recessions, winners in completing sport events, and movements in financial series. We also illustrate the methods for screening time series for serial dependence in criminal data.

Network-based Approaches To Classification And Biomarker Identification In Metastatic Melanoma

Jean Yang
Rebecca Barter[1], Sarah-Jane Schramm[1], Graham Mann[1]
[1]University Of Sydney

Much current research in medicine and biology focuses on achieving a deep understanding of the biological processes underlying complex diseases such as diabetes, heart disease and cancer. These insights have been made possible by rapid advances in new biotechnologies that have generated a myriad of high-throughput data, including transcriptomics, proteomics, and genomics data.

Finding prognostic markers has been a central question in much of this research and in the last decade, approaches to prognostic prediction within a genomics setting are primarily based on changes in individual genes / proteins. Very recently, however, network based approaches to prognostic prediction have begun to emerge which utilise interaction information between genes. This is based on the belief that large-scale molecular interaction networks are dynamic in nature and changes in these networks, rather than changes in individual genes/proteins, are often drivers of complex diseases such as cancer.

In this talk, I use data from stage III melanoma patients provided by Prof. Mann from the Westmead Millennium Institute that comprises of clinical, mRNA, microRNA and protein data to discuss how network information can be utilised in the analysis of gene expression analysis to aid in biological interpretation. I will also present an R software package, Variability Analysis in Networks (VAN), that enables an integrative analysis of protein-protein or microRNA-gene networks and expression data to identify hubs (*i.e.* highly connected proteins/microRNAs in a network) that are dysregulated, in terms of expression correlation with their interaction partners.

Spectral Methods For Exploring Process Lags

John Henstridge
Karl Beidatsch[1]
[1]Data Analysis Australia

Industrial processes usually have complex systems that rely upon numerous sensors for control. When problems occur, these can provide an enormous amount of data that must be explored to understand what is happening. Simple correlations are not informative due to the time lags, whether they are lags as material moves through the system or lags induced by the control mechanisms themselves. Spectral principal component methods applied to high dimensional time series provide a means of exploring such relationships. The approach is illustrated with an application to a chemical plant.

Sparse Regression With Nonsparse Latent Features

Jinchi Lv
Zemin Zheng[1]
[1]University Of Southern California

Many modern statistical problems in such diverse areas as genetical genomics and financial econometrics can be cast in the framework of a multivariate regression model, where both the predictors and responses are of high dimensionality. A sparse singular value decomposition > (SVD) of the regression coefficient matrix, which is of low rank and with sparse singular vectors, can play a key role for simultaneous dimension reduction and variable selection. We introduce sparse orthogonal factor regression (SOFAR), a unified approach to regularized multivariate regression, to estimate such a sparse SVD structure. We formulate the regularization procedure as an orthogonality constrained optimization problem and employ sparsity-inducing penalties of a general, flexible form, which specializes to some important cases. The SOFAR methodology is connected to a variety of multivariate techniques, including biclustering, sparse principal components, factor analysis, and vector autoregression, and yields useful new methods in these contexts. We derive nonasymptotic error bounds for the regularized estimator under conditions that control the degrees of nonconvexity and nonidentifiability in the SVD problem. We develop an efficient optimization algorithm using the alternating direction method of multipliers, and investigate its convergence properties. Extensions to adaptively weighted penalties are also explored. Simulation studies show that the SOFAR methodology substantially outperforms some existing methods, and its usefulness is demonstrated by an analysis of yeast expression quantitative trait loci (eQTL) data.

A Functional Data Approach To The Analysis Of Gait Patterns And Kinematics Indicies

Julia Polak
Morgan Sangeux[1]
[1]Hugh Williamson Gait Analysis Laboratory, The Royal Children's Hospital, Melbourne

Recently attention in the field of gait analysis has been devoted to defining so-called "kinematics normalcy indices". These indices have several aims. First, to define a metric that allows a comparison between an arbitrary gait pattern of the particular patient of interest and a nominally "normal" gait pattern from the general population. Second, to detect and measure changes in the gait pattern of an individual before and after an intervention or over time. Such metrics are important in clinical cohort studies where investigators need to evaluate the effect of a treatment on the gait pattern. The majority of the indices commonly used in practice are constructed point by point, which ignores the functional nature of the data generated by gait analysis.

In this study, a typical dataset was collected by motion sensors connected to the hips, knees and ankles of the examined individual. At each of these three joints, the sensor records the joint movement in three-dimensions over time normalized by gait cycle. The nature of the dataset suggests a functional approach to the analysis.

We show that the index created from functional analysis tools takes into account the shape of the waveform of the gait pattern and, therefore, provides a more meaningful tool than the existing approaches. The functional analysis tools considered in the current study included data depth, functional quantiles and functional confidence bands. This talk presents the new proposed "functional" index and compares its performance to other commonly used indices. All illustrations presented use real data collected by the researchers at the gait analysis laboratory.

Deterministic Markov Chain Monte Carlo Algorithms

Josef Dick

Markov chain Monte Carlo requires as inputs a sequence of random numbers which are usually modeled as independent U(0, 1) random variables. In this talk we replace the sequence of random numbers by a deterministic sequence of points and discuss the convergence behavior of the discrepancy of the sample points from the target distribution. The main motivation is the search for quasi-Monte Carlo versions of MCMC.

A Flexible Model For Spatial Extremes

Jean-Noel Bacro
Carlo Gaetan[1], Gwladys Toulemonde[2]
[1]Universita Ca Foscari
[2]Universite Montpellier

Extreme events in many environmental or climate processes can induce strong damages on human lifes and/or material assets. In the last decade there has been a major effort to model extremes of spatially dependent data and max-stable processes appeared as natural models for processes of maxima. Max-stable processes arise as an infinite dimensional generalization of multivariate max-stable distributions. The multivariate extreme value theory offers various notions to capture the main characteristics of the underlying dependence structure. A particular one is asymptotic independence. Roughly speaking, the components of a random vector are asymptotically independent if an increasing number of independent copies of it tends to have their rescaled componentwise maxima independent. It is well known that Gaussian vectors are asymptotically independent. Extremal dependence related to max-stable processes is restricted to the notions of asymptotic dependence and exact independence. Some recent studies showed that environmental processes such as rainfalls or waves height ([2], [3]) exhibit asymptotic independence. A class of asymptotic independent models has been recently proposed by [3]. Here we propose an hybrid spatial model for extremes allowing to model asymptotic dependence at short distance, asymptotic independence at intermediate distance and possibly exact independence at greater distance. An application to extreme rainfalls data will be given.

References :
[1] Davison, A.C., Gholamrezaee (2012). Geostatistics of extremes. *Proc. Royal Soc. London*, A 468, 581–608.
[2] Davison, A.C., Huser, R., Thibaud, E. (2013). Geostatistics of dependent and asymptotically independent extremes. *Mathematical Geosciences*, to appear.
[3] Wadsworth, J. L., Tawn , J. A. (2012). Dependence modelling for spatial extremes. *Biometrika* 99, 253–272.

Optimal Rates For Estimating Functionals Of Covariance Matrices

Jianqing Fan
Philippe Rigollet[1], Weichen Wang[1]
[1]Princeton University

Covariance matrices are at the core of many statistical procedures such as principal component analysis or linear discriminant analysis. As such, they have been the focus of many recent studies, especially in a high-dimensional framework. Yet, in many applications, one only needs to estimate functionals of said covariance matrix rather than the matrix itself. Indeed these functionals arise over an over in the literature. They are known to characterize the performance of various estimators and are needed to construct test statistics. Motivated by such examples, we introduce minimax optimal estimators of several functionals of covariance matrices including their squared norm and other measures of sparsity. The performance and relevance of these estimators is illustrated on two-high dimensional testing problems, one arising in gene testing and another from financial economics.

Feature Selection For Varying Coefficient Models With Ultrahigh Dimensional Covariates

Jingyuan Liu
Runze Li[1], Rongling Wu[2]
[1]Dept. of Statistics and Methodology Center, Pennsylvania State University
[2]Dept. of Public Health Sciences, Penn State Hershey College of Medicine

This paper is concerned with feature screening and variable selection for varying coefficient models with ultrahigh dimensional covariates. We propose a new feature screening procedure for these models based on conditional correlation coefficient. We systematically study the theoretical properties of the proposed procedure, and establish their sure screening property and the ranking consistency. To enhance the finite sample performance of the proposed procedure, we further develop an iterative feature screening procedure. Monte Carlo simulation studies were conducted to examine the performance of the proposed procedures. In practice, we advocate a two-stage approach for varying coefficient models. The two stage approach consists of (a) reducing the ultrahigh dimensionality by using the proposed procedure and (b) applying regularization methods for dimension-reduced varying coefficient models to make statistical inferences on the coefficient functions. We illustrate the proposed two-stage approach by a real data example.

The Genealogy Of Populations Undergoing Selection

Jason Schweinsberg

Consider a model of a population of constant size N in which each individual dies at rate one and each individual experiences mutations at rate r. Mutations are assumed to be beneficial, so that the fitness of an individual with k mutations is $1 + s(k - m)$, where m is the mean number of mutations among individuals in the population. When an individual dies, a replacement is chosen at random from the population with probability proportional to the individual's fitness. We will discuss rigorous results concerning the rate of increase in the number of mutations, the distribution of the fitness values of individuals in the population, and the genealogy of the population.

Multivariate Generalized Poisson Geometric Process Model With Scale Mixture Distributions

Jennifer So-Kuen Chan
Wai-Yin Wan[1]
[1]Bureau of Crime Statistics and Research

This paper proposes a new model named as multivariate generalized Poisson log-t geometric process (MGPLTGP) model to study multivariate time-series of counts with overdispersion or underdispersion, non-monotone trends within each time-series and positive or negative correlation between pairs of time-series. This model assumes that the multivariate counts follow independent generalized Poisson distributions with an additional parameter to adjust for different degrees of dispersion including overdispersion and underdispersion. Their means after discounting the trend effect geometrically by ratio functions form latent stochastic processes and follow a multivariate log-t distribution with a flexible correlation structure to capture both positive correlation and negative correlation. By expressing the multivariate Student's t-distribution in scale mixtures of normals, the model can be implemented through Markov chain Monte Carlo algorithms via the user-friendly WinBUGS software. The applicability of the MGPLTGP model is illustrated through an analysis of the possession and/or use of two illicit drugs, amphetamines and narcotics in New South Wales, Australia.

Model Selection For Large Data By Subsampling

Jiayang Sun
Ethan Yifan Xu[1], Qiao Xinyge[2]
[1]Case Western Reserve University
[2]Binghamton University

Correct model selection is extremely important for making inferences about some characteristics of a population. Model selection includes (a) identification of the model form and (b) inclusion of the variables (features) so that the specific model of the form with these features fit the data well. If the model form is known, popular approaches for feature selection for large (sparse) data have been those using penalties and shrinkage. We propose a different approach, subsampling feature selection for feature and f selection based on a predictive measure, when f is unknown but belongs to a group of competing forms. Its analysis, numerical study and an application will be provided.

Multiple Imputation In The Presence Of Non-normal Data

Katherine Lee
John Carlin[1]
[1]Clinical Epidemiology and Biostatistics Unit, Murdoch Childrens Research Institu / Department of Pae

Although multiple imputation (MI) is becoming increasingly popular for handling missing data, there remains a scarcity of guidelines surrounding its practical application. The commonly applied approaches for imputing missing values assume normality for continuous variables (at least conditionally on the other variables in the imputation model). However, it is not clear how best to impute missing values for continuous variables that are non-normally distributed. One frequently applied approach is to transform such variables to improve normality prior to imputation, but it is unclear whether this approach is preferable to imputing missing values on the raw scale and also which transformation to use. We use a simulation study to compare the use of various transformations applied prior to imputation, including a novel non-parametric transformation, to using untransformed data when imputing a continuous variable from non-normal distributions.

We generated data from a range of non-normal distributions, and set 50% of values either missing completely at random or missing at random. We then imputed the missing values on the raw scale, on a zero-skewness scale (using either a zero-skewness log transformation, or the Box-Cox transformation), or following a non-parametric transformation where we equated percentiles of the variable to percentiles of the normal distribution. We compared inferences regarding the marginal mean of the incomplete variable and the association between the incomplete variable and a fully observed outcome to true parameter values used to simulate the data.

The results demonstrate that transforming a continuous variable with a non-normal distribution prior to imputation can introduce bias in the resulting parameter estimates, irrespective of the transformation used. This finding suggests that despite the normality assumption inherent in the commonly applied approaches to multiple imputation, it may be best to impute on the untransformed scale regardless of the distribution of the incomplete variable.

Stratification In Statistics New Zealand Social Surveys: Are Urban Areas Actually Cheaper For Fieldwork?

Julia Hall

Currently, Statistics New Zealand (SNZ) employs a layer of stratification for its social surveys that is designed to reduce fieldwork costs. Urban PSUs are over-sampled at a rate based on the square root of their assumed cost relative to rural PSUs (optimal allocation to minimise variance for a fixed cost – see Lohr (1999), Cochran (1977) and Kish (1965)). However, sampling at different rates has an impact on the precision of estimates for a given sample size. Surveys using disproportionate sampling need to utilise survey weights if they are to give unbiased cross-strata estimates. If there is no relationship between the survey variables of interest and the probability of selection into the survey then these survey weights will, on average, decrease precision. The aim of this project was to discover if the underlying assumption for this disproportionate sampling was correct. If not, could we reduce the level of over-sampling or get rid of this stratification layer altogether, freeing up room for other, more effective stratification (i.e. stratification more highly related to our survey variables of interest). We also investigated whether an alternative classification could provide a better definition of a 'high cost' strata to improve efficiencies.

Modelling Strategies For Elimination Of Infectious Diseases Of Livestock

Kathryn Glass
Belinda Barnes[1]
[1]Australian National University / Department of Agriculture, Fisheries and Forestry

When novel disease outbreaks occur, policy makers must respond promptly and are often called on to make control decisions before a detailed analysis of disease parameters can be undertaken. I will present a flexible meta-population model of disease spread that is designed to explore strategies for eliminating livestock infections such as avian influenza, classical swine fever, and equine influenza. The model includes variation in population density and allows for wide transmission through high-mixing locations or events. Using probability generating functions derived from the underlying branching process model, I will explore the likely success of reactive control strategies such as vaccination, quarantine and animal culling and a combination of these measures. High-mixing events are crucial to disease spread, but banning these events is often controversial. I will describe alternative control scenarios and the situations under which these can be effective alternatives.

Predictive Models In The Mining Sector

John Dickson
John Henstridge[1]
[1]Data Analysis Australia, 97 Broadway, Nedlands, WA 6009

Predictive models for recovery rates and grades are critical for both the ongoing operation of mineral processing plants and the proper evaluation of potential mineral deposits. Decisions worth hundreds of millions of dollars can depend on such models. Frequently these models are based on limited laboratory data. The building of such models requires an awareness of both the limitations of the data and the way in which the model will be applied, and necessitates statistical sophistication which might not always be apparent in the final result. This talk discusses experiences in developing such models in the context of several major ore types.

Creating Edit Rules Using Association Rule Mining

Kevin Mark
Claire Clarke[1]
[1]Australian Bureau of Statistics

Data editing is an important process to ensure good quality in data collections. Many editing techniques make use of edit rules. The creation of edit rules is usually done by hand. This can be a very time consuming exercise that requires subject matter specialists for the data. So the ability to automatically create edit rules could save a lot of time and resources. We show how association rule mining has the potential to achieve just that. We will review the statistical properties of association rules, which is a popular data mining method. This is then applied to automatically create data edit rules.

Using The Census As A Sampling Frame: A Rare Opportunity To Understand Non-response Bias

Kylie Maxwell

New Zealand's 2013 Census of Population and Dwellings was used as a sampling frame for two large household surveys: the Māori Social Survey (Te Kupenga) and the Household Disability Survey. One opportunity when using the census as a frame is having demographic information (age, sex, ethnicity, income) about all sample units, including non-respondents. This talk will discuss how we used this information to better understand non-response bias, and how we were able to use this information to adjust for non-response using calibration.

While using the census as a sampling frame has some advantages over alternative frames, it also presents some challenges. These include; drawing a sample from a list frame where it is possible to select multiple persons per household, which increases respondent burden, dealing with population groups that move often, and administrative challenges. This talk will discuss these challenges in more detail.

Biomarker Discovery In Metastatic Melanoma Via Multi-layered Data

Kaushala Jayawardana
Samuel Müller[1], Jean Yang[1]
[1]School of Mathematics and Statistics, The University of Sydney, Sydney

With the increased availability of multiple data sources through the advancement of high-throughput biotechnologies, interest has been increasingly focused on investigating which of these sources or the combination thereof, contains the most valuable prognostic information. However, integration of multiple data platforms is ever challenging as it requires understanding and processing many distinct platforms as well as dealing with the imbalance of the number of variables between clinical and high-throughput data. In this context, we evaluate and integrate multi-layered data derived from the same tumour specimens in stage III melanoma patients; gene, protein, and microRNA expression as well as clinical, pathologic and mutation information, to unravel biomarkers in stage III melanoma.

We explore an integrative classification framework which focuses on uncovering biomarkers in individual platforms as well as in various combinations of the data platforms using platform dependent methods, which render models with good prognostic power. The results using the method of pre-validation within our framework shows that different patient groups are correctly classified by different combinations of data platforms. However, devising a classifier that captures all the important prognostic information in multiple data sources to improve upon these observed classification errors is an open question. To address this, we investigate the effect of using platform dependent weights to guide the variable selection process. In this study, we use weights derived from the association between platforms such as correlation information, and investigate their effect on the successful construction of a classifier with good prognostic power. This approach shows great potential in prognostic modelling.

Hydrodynamic Limits Of Stochastic Networks With Load Balancing

Kavita Ramanan
Reza Aghajani [1]
[1]University of Louisville

We establish mean-field limits for a class of load balancing models in networks with many servers in the presence of general service distributions, and describe the insights that they provide into the performance of networks arising in applications. A key tool is the representation of the process in terms of interacting measure-valued processes that may be of independent interest.

Volume Doubling Property, Quasisymmetry And Heat Kernels

Jun Kigami

We consider time changes of the Brownian motions on the Sierpinski carpets. If the measure is volume doubling with respect to the restriction of the Euclidean metric, the existence of a metric which is quasisymmetric to the restriction of the Euclidean metric and

under which the heat kernel has a sub-Gaussian asymptotic estimate.

Statistical Computing In Protein Folding

Samuel Kou

Predicting the native structure of a protein from its amino acid sequence is a long standing problem. A significant bottleneck of computational prediction is the lack of efficient sampling methods to explore the configuration space of a protein. In this talk we will introduce a new statistical computing method to address this challenge: fragment regrowth via energy-guided sequential sampling (FRESS). The FRESS method combines statistical learning (namely, learning from the protein data bank) with sequential sampling to guide the computation, resulting in a fast and effective exploration of the configurations. We will illustrate the FRESS method with both lattice protein model and real proteins.

Pressing Statistical Challenges In Cancer Research

Kim-Anh Do

PRESSING STATISTICAL CHALLENGES IN CANCER RESEARCH

Do, Kim-Anh
Professor and Chair of Biostatistics, U.T.M.D. Anderson Cancer Center, Houston

For the past few years, the National Cancer Institute (NCI) has been leading, with input from the research community: the Provocative Questions project, which compiled a list of important but difficult questions that need to be addressed to drive cancer research forward. These are questions that address broad issues, build on recent advances, and take into consideration the likelihood of success. To inform NCI and statisticians who may be looking for important problems, I will describe, at a high level, the most pressing statistical questions posed by cancer research at MD Anderson. I will present a broad view of the quantitative challenges that cancer biologists and clinicians are facing. I will focus on the statistical problems geneated by personalized medicine research that involve the integration of multiplatform high-dimensional *omics* data with imaging and clinical outcomes.

Partial Results On Convergence Of Loop-erased Random Walk To Sle In The Natural Parametrization

Michael Kozdron

We outline a strategy for showing convergence of loop-erased random walk on the two-dimensional square lattice to SLE(2), in the supremum norm topology that takes the time parametrization of the curves into account. The discrete curves are parametrized so that the walker moves at a constant speed determined by the lattice spacing, and the SLE(2) curve has the recently introduced natural time parametrization. Our strategy can be seen as an extension of the one used by G. Lawler, O. Schramm, and W. Werner to prove convergence modulo time parametrization. The crucial extra step is showing that the expected occupation measure of the discrete curve, properly renormalized by the chosen time parametrization, converges to the occupation density of the SLE(2) curve, the so-called SLE Green's function. Although we do not prove this convergence, we rigorously establish some partial results in this direction including a new loop-erased random walk estimate. Based on joint work with Tom Alberts and Robert Masson.

Keeping Competitive Regression Models On Hand - Standardized Update And Its Application

Yoshinori Kawasaki
Masao Ueki[1]
[1]Tohoku University, Sendai

Abstract: Traditional model selection procedures aim at obtaining a single best model. Under severe multicollinearity, however, the selected single best model often shows poor reproducibility when applied to a new independent data set. In such a case, one may well keep competitive regression models on hand. This paper proposes a new method for choosing regression models which may produce multiple models with sufficient explanatory power and parsimony. The method ensures interpretability of the resulting models even under strong multicollinearity. The algorithm proceeds in the forward stepwise manner with two requirements for the selected regression models to be fulfilled: goodness of fit and the magnitude of update in loss functions. For the latter criterion, the standardized update is newly introduced, which is closely related with the model selection criteria including the Mallows' Cp, Akaike information criterion and Bayesian information criterion. Simulation studies demonstrate that the proposed algorithm works well with and without strong multicollinearity and even with many explanatory variables. Application to real data is also provided.

Heavy-tailed Features And Empirical Analysis Of The Limit Order Book Volume Profiles

Kylie-Anne Richards
Gareth W. Peters[1], William Dunsmuir[2]
[1]Boronia Capital Pty. Ltd.
[2]School of Mathematics and Statistics, University of NSW

Body of Abstract:

This paper investigates attributes of the stochastic structures of the volume profiles in each level of the Limit Order Book (LOB). The analysis is performed via statistically rigorous methods carried out to examine futures market Limit Order Book data. In particular, we investigate empirically three families of models: Alpha-stable, Generalized Pareto distribution (GPD) and Generalized Extreme Value (GEV) and find that there is statistical evidence that heavy-tailed sub-exponential volume profiles occur on the LOB bid and ask and on both intra-day and inter-day time scales. In futures exchanges, the heavy tail
features are not asset class dependent and they occur on ultra or mid-range high frequency data. Of the distributions and estimation methods considered, the GPD MLE provided the best fit for all assets. We demonstrate the impact of the appropriate modeling of the heavy tailed volume profiles on a
commonly used liquidity measure, XLM. In addition, we demonstrate that utilizing the GPD distribution to model LOB volume profiles allows one to avoid over-estimating the round trip cost of trading and also avoids erroneous estimations of volume leading to significant LOB imbalances in low count assets. We conclude that volume forecasting models utilized by a HFT strategy should account for heavy tails, time varying parameters and long memory present in the data.

Probabilistic Models For Detecting And Genotyping Structural Variation

Lachlan Coin
Evangelos Bellos[1]
[1]School of Public Health, Imperial College London

Motivation: Despite the prevalence of copy number variation (CNV) in the human genome, only a handful of confirmed associations have been reported between common CNVs and complex disease. This may be partially attributed to the difficulty in accurately genotyping CNVs in large cohorts using array-based technologies. Whole genome, exome and targeted regional capture sequencing are now widely being applied to case–control cohorts and presents an exciting opportunity to look for common CNVs associated with disease.

Results: We have developed a suite of tools for identifying and genotyping copy number variation from high-throughput sequence data. These tools include methodology for normalisation of signal to remove biases and technical artefacts; methodology for modelling the expected distribution of multiple observed features of high-throughput sequence data conditional on local copy number; as well as hidden Markov models for integrating information from multiple sources. We have demonstrated that these approaches are able to accurately identify and genotype copy number variation, and can re-capitulate known CNV-phenotype associations. Our methodology can be applied both to capture-based sequencing, as well as whole-genome sequencing. I will also describe extensions to apply these approaches to calling copy number variation from tumour-normal pairs.

Body Mass Index But Not Blood Pressure Appears Susceptible To Changes In Socioeconomic Position After Childhood

Katrina Scurrah
Anne Kavanagh[1], Rebecca Bentley[1], Lukar Thornton[2], Stephen Harrap[3]
[1]Centre for Women's Health, Gender & Society, University of Melbourne, Melbourne
[2]Centre for Physical Activity & Nutrition Research, Deakin University, Melbourne
[3]Department of Physiology, University of Melbourne, Melbourne

Low socieoeconomic position (SEP) measured at either the individual level or the area level is associated with increased cardiovascular disease risk, but the relative importance of SEP in childhood and adulthood, and changes in SEP between these two life stages, remains unclear. We aimed to assess whether no change, an improvement or worsening of SEP from childhood to young adulthood is associated with lower or higher cardiovascular disease risk measures in young adulthood, after adjustment for childhood SEP. Small area SEP (census collector district level, around 225 dwellings) was used as a proxy for individual SEP as this provides a good reflection of individual SEP. By using data from 286 adult Australian families from the Victorian Family Heart Study (VFHS), in which some offspring have left home (independent offspring, n=364) and some remained at home (dependent offspring, n=199), we were also able to adjust for shared but unmeasured childhood genetic and environmental effects between siblings. SEP was defined as the Index of Relative Socioeconomic Disadvantage and matched to addresses of offspring ("current SEP") and their parents ("parental SEP"). We fitted variance components models to test whether the change in SEP "(current SEP minus parental SEP)/100" was associated with body mass index (BMI) or systolic blood pressure (SBP), after adjustment for parental SEP and within-family correlation. Changing SEP was associated with BMI (b=-0.49kg/m^2, $p < 0.01$) but not SBP (b=-0.53 mmHg, p=0.4). Our findings suggest that young adults who live in low SEP areas with parents in high SEP areas had higher mean BMI than young adults whose SEP is higher than their parents'. These results suggest that a change in SEP in young adulthood is an important predictor of BMI, independent of childhood SEP, which may have implications for public health strategies.

Speeds Of Cyclists On Different Infrastructures In Nsw

Lloyd Flack
Roslyn Poulos[1]
[1]School of Public Health and Community Medicine, University of NSW

The Safer Cycling Study is a web based cohort study in which a large cohort of cyclists reported their cycling activity in six one week periods spread over a year.

As part of this they reported daily distances travelled, time spent cycling and proportions of time spent on different infrastructures.

Non linear regressions were used to estimate cyclist speeds on different infrastructures.

Data was collected on cyclist characteristics such as sex, age, risk propensity and experience. The effect of these on cycling speeds was analysed.

The effect of average speed of a rider on accident and injury rates per 1,000 hours and 1,000 kilometres ridden was also analysed.

Applications Of The Representation Points In Statistical Simulations

Kai-Tai Fang
Min Zhou[1], Wenjun Wang[2]
[1]BNU-HKBU United International College
[2]Hainan Normal University

Statistical simulation has played more and more important role in statistics research. The traditional simulation employs a sample taken from the population distribution by the Monte Carlo method. Then the sample statistic of the interest can be used for inference of the population. In this study we propose a new way for statistical simulation. Firstly we construct some approximate population distribution by the use two kinds of representative points. The first type of representative points is formed by the use of Quasi-Monte Carlo and related methods. The second kind of representative points is produce by the theory of quantization. Secondly, samples are generated from these approximate distributions. The simulation is based on these samples. Our results show that the new method can significantly improve the results by the use of Monte Carlo methods.

Input Substitution And Business Energy Consumption: Evidence From Abs Energy Survey Data

Kay Cao

This paper applies the system of equations approach to energy consumption modelling using the ABS 2008-09 Energy, Water and Environment Survey (EWES), Economic Activity Survey (EAS) and Business Activity Statement Unit Record Estimates (BURE) data. A system
of equations including a translog variable cost equation and an energy cost share equation is

estimated. Estimation results show that labour and energy are substitutes. Estimates of a range of elasticity measures, including Allen-Uzawa elasticity of substitution, own and cross price elasticities and Morishima elasticity of substitution, are also provided.

Key words: system of equations, energy consumption modelling, elasticity of substitution

Impact Of Misspecified Random Effects In Multilevel Models: Applications To Panel Data

Louise Marquart
Michele Haynes[1], Peter Baker[2]
[1]Institute for Social Science Research, University of Queensland, Brisbane
[2]School of Population Health, University of Queensland, Brisbane

Multilevel models, also known as hierarchical or mixed effects models, are commonly used to analyse longitudinal panel data in health and social sciences. Multilevel models capture the within-subject variability of multiple measurements per subject, as well as the between-subject variability due to observed and unobserved heterogeneity. The random effects characterize the heterogeneity among subjects, and are traditionally assumed to follow a Gaussian distribution due to computational and mathematical tractability.

Research has indicated mixed evidence for the impact of misspecified random effects distributions on the coefficients and standard errors of both fixed effects and random effects. The majority of work has been evaluated in biomedical settings where data is often collected in a controlled setting in daily or weekly intervals. Minimal research has been done in the longitudinal panel survey setting which is subject to more variability and higher rates of non-response due to the collection of self-reported data. Longitudinal panel survey data are generally collected six monthly or yearly, and hence attrition is often high, particularly in vulnerable populations.

The motivating example is an application to employment outcomes using the Household and Income Labour Dynamics in Australia (HILDA) panel survey. Gaussian distributed random effects may not be appropriate when modeling such binary outcomes. The results from a simulation study are presented to assess the impact of misspecified random effects distributions when the random effects are simulated from non-Gaussian distributions, including mixture of Gaussians. Multilevel models fitting Gaussian and non-Gaussian random effect distributions are implemented. The simulation study replicates the complexities inherent in panel survey data, including attrition and non-response, as well as misspecified fixed effect and random effect models. The results will enhance interpretation of longitudinal panel data by helping users of multilevel models be aware of the consequences of misspecified random effect distributions.

The Empirical Geodesic Graphs And Their Deformation For Data Analysis

Kei Kobayashi
Henry Wynn[1]
[1]London School of Economics

Abstract:
This talk is largely about empirical metrics and empirical geodesics and their application to data analysis. We introduce a novel deformation of a metric called alpha, beta, gamma metric and give a way of computation via the metric graphs. This new deformation of a metric relates to a curvature of the data space and robustness of the estimation based on it. The means by alpha and beta metrics correspond to an intrinsic and an extrinsic mean, respectively, usually used when the data is on a specific manifold though our method is not necessarily for such cases. We show some theoretical results on the properties of the metric and give some examples to show how the metric works for data analysis.

Heat Kernel Estimates And Local Clt For Random Walk Among Random Conductances With A Power-law Tail Near Zero

Takashi Kumagai
Laurent Saloff-Coste[1]
[1]Cornell University

We study on-diagonal heat kernel estimates and exit time estimates for continuous time random walks (CTRWs) among i.i.d. random conductances with a power-law tail near zero. For two types of natural CTRWs, we give optimal exponents of the tail such that the behaviors are ‘standard’ (i.e. similar to the random walk on the Euclidean space) above the exponents. We then establish the local CLT for the CTRWs. We will also compare our results to the recent results by Andres-Deuschel-Slowik.

New Statistical Tools To Study Heritability Of The Brain

Pierre Lafaye De Micheaux
Benoit Liquet[1], Perminder Sachdev[2], Anbupalam Thalamuthu[2], Wei Wen[2]
[1]University of Queensland
[2]University of New South Wales

Diffusion tensor imaging (DTI) allows in vivo, non-invasive mapping of the diffusion process of water molecules in biological tissues. This can help us to can obtain axonal pathways within the brain. Neuroscientists are interested in heritability of such structures. Heritability analysis helps us to know if the development of axons is ruled by genetic or environmental factors. In this talk, we will describe the kind of data involved in such studies. We will then present a new model we have developed that take into account the geometric information of axons fiber bundles for a group of elderly twins. A parallel sliding coordinate system will be built, using a tool called the projection median of a set of points. We will then use genetic information collected on these people to investigate heritability of such structures.

Bayesian Analysis Of Structural Credit Risk Models With Micro-structure Noises And Jumps

Kwok Wah Ho
Sau Lung Chan[1], Hoi Ying Wong[2]
[1]Hong Kong Exchanges and Clearing Limited
[2]Chinese University Of Hong Kong

There is empirical evidence that structural models of credit risk significantly underestimate both the probability of default and credit yield spread. We propose a Bayesian approach to simultaneously estimate jump-diffusion firm value process and micro-structure noise from equity prices based on common structural credit risk models. The proposed Bayesian approach has the advantage of producing posterior distributions of model parameters and latent variables for financial analysis. The focus of our study lies on investigating whether the bias of structural credit risk model is mainly caused by the firm value distribution, the option-theoretic method or the micro-structure noise of the market. We apply our method to real data in Hong Kong market and our results indicate that the effects of micro-structure noises are often not significant when jumps are included in the credit risk model. A simulation study is also conducted to ascertain the effectiveness of our method.

A Comparison Of Item Response Theory And Latent Class Methods For Combining Data For Regression

Ken Beath

When a number of outcomes or responses are observed on each subject they may be combined using a latent variable method and the latent variable used as a response in a regression model, forming a simple structural equation model, with the advantage of increased power and avoiding fitting a large number of regression models. For binary outcomes a latent class model is commonly used, assuming an underlying categorical latent variable, which determines the outcome probabilities for each class. The predictors are introduced through use of a multinomial logit model, with the effect of the predictors as odds ratios for membership of each class compared to a reference class. There are two difficulties with this method. Firstly, is that as the sample size increases the number of classes will increase leading to difficulties in interpretation, as each additional class after the first introduces an odds ratio for each predictor. Secondly, if the underlying latent variable is actually continuous, we are discarding information by assuming that the latent variable is categorical and therefore there is a loss of power. When the relationship between the outcomes can be modeled using a continuous latent variable a better model is based around an Item Response Theory (IRT) model, which assumes a normally distributed latent variable with the outcome probabilities determined using a logistic function. The predictors are incorporated into the model as part of a regression model including the latent variable. The advantages of this method are increased power, as well as easier interpretation. The methods are compared using simulation, demonstrating the increased power of the IRT based model, and an example using the effect of predictors on physical functioning.

An L P-theory For A Class Of Spdes With Non-local Operators

Kyeong-Hun Kim

The non-local operators which we will discuss in this talk are infinitesimal generator of d-dimensional subordinate Brownian motions. An L_p-theory will be introduced for a class of stochastic partial differential equations having such operators. The key estimate of our theory is the Parabolic Littlewood-Paley inequality related to such operators.

A Comparison Of Survey Quality Indicators For Improving Data Collection Efficiency

Rosslyn Starick
Rosslyn Starick[1]
[1]Australian Bureau Of Statistics

In an environment where it is increasingly difficult to achieve high survey response rates, responsive survey designs have been proposed to maintain the quality of survey outcomes through more targeted follow-up strategies, while reducing data collection costs.
In addition to response rates, responsive design uses indicators of survey data collection quality and cost effective targeting of sample units to monitor and reduce non-response bias.

The Australian Bureau of Statistics (ABS) is in the process of implementing responsive design strategies to its household survey program. A number of different survey data collection quality indicators and targeting methods which have been proposed in the literature were compared using data from the National Health Survey conducted in 2007-8. This talk will outline the advantages and disadvantages of each method and their suitability for use at the ABS.

Some Recent Progress Relating Pinned And Unpinned Variants Of Sle

Laurence Field

The Schramm-Loewner evolution (SLE) is a conformally invariant family of random planar curves that describes the scaling limit of interfaces in several critical models from 2D statistical physics.

A measure on discrete paths can be restricted to those paths that pass through a marked point. If you add together these "pinned measures" over all possible marked points, you obtain the unpinned measure weighted by the path's length.

For continuous fractal curves like SLE, this idea presents some difficulties, because it is not immediately clear what the pinned measure should be, and what is meant by the curve's length.

In the case of SLE, the pinned measure is called two-sided radial SLE and its total mass is the normalised probability of passing near the marked point, called the SLE Green's function. For the curve's length we use the so-called natural parametrisation of SLE, developed by Lawler, Rezaei, Sheffield and Zhou, which is a purely geometric quantity. We prove that for SLE, the integral of the pinned measures is the unpinned measure weighted by the curve's length in the natural parametrisation.

Modeling High-dimensional Time Series Via Low-dimensional Groups Structure

Jinyuan Chang
Bin Guo[1], Qiwei Yao[2]
[1]Peking University
[2]London School of Economics

We propose an approach to modeling high-dimensional time series via low-dimensional groups structure, which can be viewed as a dimension reduction method for high-dimensional time series. Both theoretical results and algorithm for this approach are established. We also provide a pre-grouped method in practice to decrease the computation cost, which can be used to split a high-dimensional optimization problem into several lower-dimensional subproblems. The proposed methodology is illustrated with both simulated and real data sets.

An Optimization Approach To Placing Hard Bounds On Solutions To Free Boundary Problems.

Louis Bhim
Reiichiro Kawai[1]
[1]The University of Sydney

We approach the problem of placing hard bounds on the price of American options by re-formulating this free boundary problem as an optimization problem. To this end, we implement techniques and results from semidefinite programming theory, sums of squares polynomial representations and stochastic analysis in order to derive an optimization formulation of the problem that is solvable by available semidefinite programming problem solving softwares. In particular, we utilise the well-known Dynkin formula in order to place deterministic bounds on the price of an American option and then use results on the sum of squares representations for non-negative polynomials to relax the constraints of our problem and make the problem computationally tractable. We will then discuss how these techniques, having been successfully applied in this particular problem setting, can then be applied to obtain upper and lower bounds for solutions to more general free boundary problems.

Confidence Sets For Variable Selection

Davide Ferrari
Yuhong Yang[1]
[1]School of Statistics, University of Minnesota

We introduce the notion of variable selection confidence set (VSCS) for linear regression based on F-testing. The VSCS extends the usual notion of confidence intervals to the variable selection problem: A VSCS is a set of regression models that contains the true model with a given level of confidence. For noisy data, distinguishing among competing models is usually very difficult and the VSCS will contain many models; if the data are really informative, the VSCS will contain a much smaller number of useful models. We advocate special attention to the set of lower boundary models (LBMs), which are the most parsimonious models that are not statistically significantly inferior to the full model at a given confidence level. Based on the LBMs, variable importance and measures of co-appearance importance of predictors can be naturally defined.
Up to date, an almost exclusive emphasis has been on selecting a single model or two. In the presence of a number of predictors, especially when the number of predictors is comparable to (or even larger than) the sample size, the hope of identifying the true or the unique best model is often unrealistic. Consequently, a better approach is to select a relatively small set of models that all can more or less adequately explain the data at the given confidence level. This strategy identifies the most important variables in a principled way that goes beyond simply trusting the single lucky winner based on a model selection criterion.

Semilinear Stochastic Differential Equations In Infinite-dimensional Spaces

Kevin Mark

This talk discusses stochastic processes with values in a separable Hilbert space and have their evolution described by a semilinear stochastic differential equation (SDE) driven by a cylindrical Wiener process. For square-integrable processes (with finite second moment), unique mild solutions exists under certain restrictions (see, e.g., the monograph by Da Prato & Zabcyzk (1992)). We look at relaxing these conditions to enable study in a more general setting.

We present a general semilinear SDE for stochastic processes that are not necessarily square-integrable, but are in a Kondratiev space of Hilbert space-valued stochastic distributions. We use techniques from white noise analysis to define the stochastic integral by using Hitsuda-Skorohod integration, rather than the traditional Ito integration. We provide conditions for when a unique mild solution exists for such a semilinear SDE.

The semilinear SDE is applied to forward interest rate modelling in the Heath-Jarrow-Morton (HJM) framework.

Link Between The Default Rate And The Economic Situation

Lao Kenao
Komla Mawulom Agudze[1], Mahamadou Tankari[2]
[1]School of Economic and Management of Venice
[2]IFPRI, West and Central Africa Office

We analyze the financial situation of enterprises in the economic environment of France. Previous studies are based on the microeconomic analysis through the probability of default and the economic situation has shown that the numbers of default increased in case of recession. This study tries to explore a macro-economic analysis with the default rate. Thus, the goal of this study is to explain the default rate evolution by macro-economic factors which highlight the sector fragility of enterprises in front of extreme macro-economic shocks. The tests of hypothesis are relative to: the evolution of default rate is contra-cyclic of the economic activity; to seek whether the effect of the economic situation on the default rate happens with lateness; to see if there is the contagion effect between the default rates of the different finalities. Two approaches have been used through the descriptive statistic and the econometric modeling by using a Vector Error Correction Model (VECM) with exogenous variables. The results indicate that in short term, the defaults in the finalities of development, creation and financial restructuration can be explained by worsening of defaults in those same finalities but, at long term, the effects of contagion between finalities are noticed. Therefore, it is necessary to set up, structural policy to increase economic activity, conjunctural policy to create employment to stimulate domestic demand and competitiveness. Finally, to bring down the average rate of interest on the monetary market so as not to harm the banks.

Selfish Routing And Network Games: Does More Information Help Or Hinder?

Ilze Ziedins
Heti Afimeimounga[1], Lisa Chen[2], Niffe Hermansson[1], Mark Holmes[1], Wiremu Solomon[1]
[1]University of Auckland
[2]Harmonic Analytics

It is well-known that adding extra capacity to queues in networks where individuals choose their own route can sometimes severely degrade performance, rather than improving it. We will discuss examples of queueing networks where this is the case under probabilistic routing, but where under state-dependent routing the worst case performance is no longer seen. This raises the question of whether giving arrivals more information about the state of the network leads to better performance more generally.

Linear Regression Analysis In Non-parametric Populations

Lawrence Brown

Linear regression analysis is often applied to populations that do not satisfy the conventional assumptions of linearity, homoscedasticity and normality of residuals. Furthermore, the classical theory considers the covariates as fixed constants, whereas in most applications in the social sciences (and many elsewhere) they are actually random variables. This talk surveys a comprehensive theory of linear regression that is free of any of these classical assumptions. Much of the talk is "expository". However, some new results will be included related to the structure of statistical errors in such situations. In particular, alternate forms of the familiar Huber-Eicker-White "sandwich" estimator are described as well as a coordinate specific test for conventional assumptions. Results concern both inference about the linear regression coefficients and about best linear predictive inference.

The basic perspectives thus involve assumption-lean statistical populations including random covariates having an unspecified distribution, and then applying assumption-rich statistical models as approximations. If time permits, some further developments built on this approach will be discussed. These include a simple estimator for the Average Treatment Effect in randomized trials. Another expansion of the perspective involves settings often described as semi-supervised learning in which additional observations are available on the covariates unaccompanied by the response of interest.

This is joint work with R. Berk, A. Buja, E. George, E. Pitkin, K. Zhang and L. Zhao.

Heat Kernels Of Non-local Operators

Zhen-Qing Chen

The study of discontinuous Markov processes has attracted lots of attentions recently due to their importance in theory and in applications. In contrast with diffusion processes, the infinitesimal generator of a discontinuous Markov process is a non-local operator. In this talk, I will survey some recent development in the study of heat kernels of non-local operators.

District Level Child Nutrition Status In Bangladesh: An Application Of Area-level Sae Method

Mossamet Kamrun Nesa
Sumonkanti Das[1]
[1]PhD Student, NIASRA, SMAS, UOW

National level indicators of child undernutrition often hide the real scenario across a country. In order to construct a child nutrition map, accurate estimates of undernutrition are required at very small spatial scales, typically the administrative units of a country or a region within a country. Although comprehensive data on child nutrition are collected in national surveys, the small scale estimates cannot be calculated using the standard estimation methods employed in national surveys, since such methods are designed to produce national or regional level estimates, and assume large samples. Small area estimation (SAE) method has been widely used to find such micro-level estimates. Due to lack of unit level data, area level SAE methods (e.g., Fay-Herriot method) are widely used to calculate small-scale estimates. In Bangladesh, a few works have been done to estimate district level child nutrition status. The recent survey Bangladesh DHS covers all districts but district wise sample sizes are very small to get consistent estimates. So Fay-Herriot Model (Fay and Herriot, 1979) has been developed to calculate district wise estimates with efficient mean squared error (MSE). The recent Bangladesh DHS 2011 and Population Census 2011 will be utilized for this study.

Scaling Properties For Generators Of Jump Processes

Moritz Kassmann
Ante Mimica[1]
[1]University of Zagreb

We employ techniques from stochastic analysis and study properties of jump processes and of corresponding harmonic functions. Our emphasis is on cases of jump processes which are not scale invariant, e.g. the geometric stable process and nonhomogeneous versions of it.

Generating A Dynamic Synthetic Population - Using An Age-structured Two-sex Model

Mohammad-Reza Namazi-Rad

Generating a reliable computer-simulated synthetic population is needed for knowledge processing and decision making analysis in agent-based systems in order to measure, interpret, and describe each target area and the human activity patterns within it. In this paper, both synthetic reconstruction and combinatorial optimisation techniques are discussed for generating a reliable synthetic population for a certain geographic region (in Australia) using aggregated- and disaggregated-level information available for such area. Combinatorial optimisation algorithm using the quadratic function of population estimators is presented in this paper in order to generate a synthetic population for the study area based on the Australian Census. The baseline population in this study is generated from the Confidentialised Unit Record Files (CURFs) and 2006 Australian Census tables. The dynamics of the population is then projected over five years using an age-structured two-sex model for household dynamics. This projection is then compared with the 2011 Australian Census. A prediction interval is provided for the population estimates obtained by bootstrapping method by which the variability structure of a predictor can replicate in bootstrap distribution.

Key words: *Combinatorial Optimisation; Bootstrap Confidence Interval; Hierarchical Structure; Household Projection; Multi-Agent Systems; Population Synthetiser.*

Accommodating Missingness When Assessing Surrogacy Via Principal Sratification

Michael Elliott
Yun Li[1], Jeremy Taylor[1]
[1]University of Michigan

When an outcome of interest in a clinical trial is late-occurring or difficult to obtain, good surrogate markers can reliably extract information about the effect of the treatment on the outcome of interest. Surrogate measures are obtained post-randomization, and thus the surrogate-outcome relationship may be subject to unmeasured confounding. Thus Frangakis and Rubin (2002) suggested assessing the causal effect of treatment within principal strata defined by the counterfactual joint distribution of the surrogate marker under the treatment arms. Li, Taylor, and Elliott (2010) elaborated this suggestion for binary markers and outcomes, developing surrogacy measures that have causal interpretations and utilizing a Bayesian approach to accommodate non-identifiability in the model parameters. Here we extend this work to accommodate missing data under ignorable and non-ignorable settings, focusing on latent ignorability assumptions (Frangakis and Rubin 1999; Peng, Little, and Raghunathan 2004, Taylor and Zhou 2009). We also allow for the possibility that missingness has a counterfactual component, one that might differ between the treatment and control due to differential dropout, a feature that previous literature has not addressed.

Fast Approximate Inference For Bayesian Multilevel Models

Yuen Yi Lee
Matt Wand[1]
[1]School of Mathematical Sciences, University of Technology

Exact inference for semiparametric regression models that use spline basis functions with penalization is typically intractable, requiring approximate inference methods for use in practice.

Markov Chain Monte Carlo (MCMC) is the most commonly used approximate inference method in this setting, but can be computationally intensive and often suffers from poor convergence in complex models. A faster, deterministic alternative to MCMC is Mean Field Variational Bayes (MFVB). We derive MFVB algorithms for a variety of Bayesian multilevel semiparametric regression models. In order to overcome the computational cost of the direct naïve approach to the underlying MFVB calculations for models, we introduce a novel, streamlined approach which involves matrix block decomposition. Through a series of numerical studies, we demonstrate that the MFVB algorithms achieve a good level of accuracy compared to that of the MCMC. Furthermore, our developed streamlined calculations are shown to be linear in the number of groups, representing a two orders of magnitude improvement over the naïve approach.

The Partial Linear Model In High Dimensions

Patric Mueller
Sara van de Geer[1]
[1]ETH Zurich

Partial linear models have been widely used as a flexible method for modeling linear components in conjunction with non-parametric ones.

In this talk we consider a partial linear model with a high-dimensional linear part. That is, the observed response variable is the sum of a linear (parametric) high-dimensional part and a non-parametric nuisance function and an error term.

In a high-dimensional dataset we often have strong reasons to believe that some variables can not be linearly modeled. The yield of some genetically modified plants, for example, depends in a non-linear way on factors like 'water' and 'temperature', whereas the high-dimensional gene expression data may be modeled linearly.

Prediction and estimation for such models is a challenging problem. Neither the naive solutions of considering the entire model as non-linear nor neglecting the nuisance function do not lead to satisfying results.

We combine LASSO and non-parametric techniques in order to estimate both the parametric and the nonparametric part.

We show that the model can be estimated with oracle rates, using the LASSO penalty for the linear part and a smoothness penalty for the nonparametric part.

Asymptotically, we have the same rate as if the nuisance function were known.

Our theoretical results are supported with simulation studies.

Larry Shepp, A Consummate Problem Solver

Lawrence Brown

Abstract: Larry Shepp took great pleasure in accessing, attacking and solving difficult and important stochastic problems. Other speakers in this session will describe some of these varied and successful activities. I will try to survey the broad scope of Larry's interests, and fill in a few specifics about some that the other speakers have not mentioned in detail.

Comparing The Efficiency Of Logistic Regression Estimators

Lyle Gurrin
Elizabeth Williamson[1]
[1]Melbourne School of Population & Global Health, The University of Melbourne

Data from individually-matched case-control studies, cohorts of twins and other paired designs provide a powerful resource that can be used estimate the magnitude of exposure-outcome associations free from confounding by shared factors. For binary outcomes, these data are typically analysed using conditional logistic regression (CLR), which uses only data from outcome-discordant pairs and, for binary exposures, also requires that pairs are exposure-discordant. An alternative is to fit an ordinary (unconditional) logistic regression (OLR) model that includes terms for between-pair and within-pair regression effects. Estimates of the within-pair regression coefficient from the OLR are potentially more efficient that the corresponding estimate from CLR since all exposure-discordant pairs contribute to the within-pair estimate regardless of whether they are outcome concordant or not. We compare closed-form expressions and variances for estimators of the within-pair effect based on CLR and three OLR models where the pair-mean exposure is (i) assumed not to be associated with the outcome; (ii) assumed to be linearly associated with the log-odds of the outcome; (iii) included as a categorical variable to allow for an unstructured between-pair relationship. The within-pair estimator from all four regression models are special cases of a formula based on weighted counts from a 2 x 2 exposure-outcome contingency table for exposure-discordant pairs, generalising the results of Sjolander et al. (Stat. Sci., 2012). Results from a simulation study show that the efficiency gains from using OLR over CLR for binary exposures are measurable but modest, consistent with published results, but that the approach based on OLR is superior for continuously-valued exposures. We conclude by re-visiting the analysis of the cord-blood erythropoietin (EPO) and birthweight from Carlin, Gurrin et al. (IJE, 2005) and presenting some results from analyses, using data from sib-pairs, of the association between environmental factors and the risk of allergic disease in childhood.

Stochastic Analysis For Gaussian Rough Paths And A Multilevel Monte Carlo Algorithm

Sebastian Riedel
Christian Bayer[1], Peter Friz[2], Benjamin Gess[3], Archil Gulisashvili[4], John Schoenmakers[1]
[1]WIAS Berlin
[2]TU Berlin / WIAS Berlin
[3]University of Bielefeld
[4]Ohio University

We present a novel criterion for the existence of Gaussian rough paths in the sense of Friz–Victoir. It is formulated in terms of a covariance measure structure together with a classical condition due to Jain–Monrad. It turns out that this condition is easy to check in many examples which allows for a stochastic calculus for a large class of Gaussian processes, ranging from (bi-)fractional Brownian motion to processes given as random Fourier series. We discuss several implications in stochastic analysis such as Itō-type exponential integrability for stochastic integrals and non-Markovian Hörmander theory. In the second part of the talk, we present a multilevel Monte Carlo algorithm that reduces the computational complexity considerably when calculating the mean of diffusion functionals when the SDE is driven by a Gaussian process.

The first part of the talk is joint work with Peter Friz (TU and WIAS Berlin), Benjamin Gess (Universität Bielefeld) and Archil Gulisashvili (Ohio University). The second part is joint work with Christian Bayer (WIAS Berlin), Peter Friz (TU and WIAS Berlin) and John Schoenmakers (WIAS Berlin).

A Smooth Test Of Goodness Of Fit For The Quantile-based Skew Logistic Distribution

Robert King
Paul van Staden[1]
[1]Department of Statistics, University of Pretoria

The Quantile-based Skew Logistic distribution features a constant 4th L moment and a constant value of any skewness invariant measure of kurtosis.
The parameters of this distribution can be estimated via the method of L-Moments.

Based on this estimation method, we develop a semiparametric test of goodness-of-fit for the distribution. This is achieved by adapting smooth tests based on moments, to work instead on L-moments and quantile functions.

After fitting the skew logistic distribution using the first 3 L-moments, the null hypothesis for the goodness-of-fit test is equality of the next 3 L-moments between the true and hypothesised distribution.

Graphical Methods For Latent Structure Testing

Robin Evans

Many models which include latent structure can be considered as semi-algebraic sets, and have recently begun to be studied from this perspective; this has shed much light on the dimension, identifiability, and asymptotic statistical properties of these models. Though most of the attention has been on equality constraints, some progress has also been made on evaluating inequalities which might be used to test such models.

However, the mathematical complexity of these approaches seems to have led to a gap between our theoretical understanding and the manner in which these models are applied in practice. In this talk we plead for a focus on finding simpler (in particular more graphical) and more computationally feasible ways to express such constraints, even at the cost of a loss of statistical power. Recent advances for directed acyclic graph models with latent variables and phylogenetic models are given as illustrations.

Statistical Challenges In Nanoscale Fluorescence Microscopy

Axel Munk
Timo Aspelmeier[1], Alexander Egner[2]
[1]Georg-August University Goettingen
[2]Laser Lab Goettingen

Modern super-resolution microscopy techniques are important tools for investigating biological structure and function in living cells. These require several data processing steps in order to obtain sharp high resolution images from the measurement process. We will discuss some statistical challenges and methods for different state of the art fluorescence microscopy techniques, such as confocal microscopy, stimulated emission depletion microscopy, stochastic marker switching microscopy and fluorescence microscopy with excitation by polarized light. This includes optical deconvolution, time resolved imaging by physical sparsity, and motion blurring and estimation.

Call For Frank And Fearless Statisticians

Robyn Attewell

Do you have an eye for detail?
Do you search for patterns in lists?
Does inefficiency frustrate you?
Does ambiguity frustrate you?
Have you studied statistics at university?
Have you passed a second year statistics course?
Are you comfortable summarising numerical data?
Do you see the sense in coding non-numerical data?
Do you understand the concept of independence?
Do you enjoy the challenge of presenting data in graphical format?
Do you enjoy the challenge of interrogating a data set to find the story behind it?
Do you question what information justifies a particular point of view?
Can you write in sentences?

If you have answered yes to all of the above, then please apply for a job in the public service.

Even if you are not in the job market, please consider coming to my talk where I will discuss the opportunities for statistically-minded individuals within the public sector. I will illustrate the need for statistical literacy in an environment of tightening budgets and calls for greater accountability. I will use examples from my own experience in law enforcement where there is a growing acceptance of the benefit of having an evidence base for policy decisions.

Bayesian Nonparametric Alternative To Multiple Imputationmissing Covariates

Murray Aitkin
Irit Aitkin[1]
[1]Department of Mathematics and Statistics, University of Melbourne

This talk describes an extension of multiple imputation which provides a fully Bayesian analysis without requiring a parametric model for the covariates with missing values, or the analysis of multiply imputed samples. It has two innovations:

i) the MCMC computational approach for multiple imputation is used, but to provide the full posterior distribution of the regression model parameters;

ii) the distribution of the incomplete covariates is modelled nonparametrically by the multivariate multinomial distribution.

The extension is illustrated with simple normal regression models; the computational approach can be extended to GLMs and GLMMs.

Generalized Arma Models -extensions And Applications

Rong Chen
Tingguo Zheng[1], Han Xiao[2]
[1]Wang Yanan Institute for Studies in Economics, Xiamen Univ. China
[2]Dept. of Statistics, Rutgers University

Time series data comes in various forms, whether being integer valued, proportions, strictly positive

or heavily skewed. Generalized ARMA model of Benjamin et al (2003) assumed that the

observations follow an exponential distribution conditional on the past. With an approach similar to

generalized linear model, the conditional mean, with a link function, assumes an autoregressive and

moving average form. In this talk we expand the GARMA approach to accommodate more flexible

structure, easier estimation procedure, deeper understanding of probabilistic property, and better

model validation procedures. Theoretical results and empirical applications will be presented.

Confidence Distribution, A Unifying Concept For Statistical Inference

Min-Ge Xie

A confidence distribution is a sample-dependent distribution function that can represent confidence intervals of all levels for a parameter of interest. It provides "simple and interpretable summaries of what can reasonably be learned from data (and an assumed model)", and it can provide meaningful answers for all sorts of questions related to statistical inference. A major theme emerging from recent developments is: *Any statistical approach, regardless of being frequentist, fiducial or Bayesian, can potentially be unified under the concept of confidence distributions, as long as it can be used to build confidence intervals of all levels, exactly or asymptotically*. In this talk, we will discuss the logic behind these developments and demonstrate that confidence distributions can potentially serve as a unifying platform for Bayesian, fiducial and frequentist inferences in all areas including estimation, hypotheses testing and prediction. We will also present practical examples, demonstrating that the developments can provide us useful statistical inference tools for statistical problems where methods with desirable properties were previously unavailable or could not be easily obtained.

Teaching Statistics Through Baseball Data And Excel

Youyu Phillips

Statistics help recap baseball routine and assess players' performances. In teaching statistics, the author used examples of baseball data to demonstrate statistical concepts of qualitative or quantitative, and continuous or discrete variables; the players' ERA (Earned Run Average) was used to demonstrate descriptive statistics, while league, batting average and homeruns were used to identify level of measurements. Students were taught to use EXCEL to sort baseball data by making classes/calculating the sum, mean, variance and standard deviation, graphing histograms, pie charts, stem-and-leaf plots, frequency polygons, Ogives, Box Plots, etc. Sample work will be presented.

Detecting Non-monotonic Trends And Testing For Synchronism In Multiple Time Series

Yulia Gel
Vyacheslav Lyubchich[1]
[1]University Of Waterloo

In this talk, we explore a new statistical test for synchronism of trends exhibited by multiple time series, i.e., testing whether two or more time series follow the same common trend. The core idea of our new approach is based on employing the local regression goodness-of-fit test statistic of Wang, Akritas and Van Keilegom (2008), which allows to detect possibly non-monotonic (non)linear trends. The finite-sample performance of the test statistic is enhanced by employing robust data-driven bootstrap approach and m-out-of-n selection algorithm to choose an optimal window in local regression. We illustrate the proposed methodology by simulations and case studies on assessing joint dynamics of various climatic variables in space and time.

What Is The Recipe For A Long Life: Good Diet, Exercise, Or Both?

Lyle Gurrin
Julia Polak[1], Andrew Forbes[2], Elizabeth Williamson[1], Julie Simpson[2], Allison Hodge[3], Julie Basset[3], Michael Fahey[2], Dallas English[1], Graham Giles[3]
[1]Melbourne School of Population & Global Health, The University of Melbourne
[2]Department of Epidemiology & Preventive Medicine, The Alfred Centre, Monash Uni.
[3]Cancer Epidemiology Centre, Cancer Council Victoria

Abstract:

Poor diet is an acknowledged risk factor for life-threating conditions such as cancer and cardiovascular disease. The association between improvement in diet and mortality risk has, however, received little attention. Changes in diet are likely to be associated with other lifestyle choices, therefore, may influence or be influenced by other risk factors in a time-dependent manner.

To study the effect on all-cause mortality of a specific dietary regimen we used the Parametric G-formula. This standardisation procedure allows us to contrast the marginal mean risk between multiple dietary regimens at the population level. It provides unbiased estimates of parameters that have a causal interpretation in the presence of time-dependent confounders, such as physical activity, alcohol consumption, BMI and energy intake. The Parametric G-Formula is a simplification of G-Computation, which becomes computationally impractical in the presence of multiple covariates and dynamic interventions. A system of regression models predicts individual level covariates (continually reset to ensure consistency with the intervention) and participants' mortality risk on comparative dietary regimens at successive follow-ups. Standard errors were computed simulation and bootstrapping. We also estimated the sensitivity of conclusions to the specification of the regression models.

Preliminary data from the Melbourne Collaborative Cohort Study (MCCS) (41,514 participants assessed 3 times over 15 years) were used to calculate a Mediterranean Diet Score from MCCS participants' responses to a Food Frequency Questionnaire and to estimate model parameters. The predicted mortality risk matched the observed mortality risk. The effect of the entire population adopting a Mediterranean diet appeared greater than that for the adoption of increased physical activity alone. There are threats to validity that need to be considered in this modelling approach include unmeasured confound and exposure measurement error. We outline our attempts to address these issues.

Comparison Of Different Approaches In Applying Multiple Imputation To Skewed Distruptions

Nargess Saiepour
Gail Williams[1]
[1]The University of Queensland

The purpose of multiple imputation (MI) is to generate possible values for missing values. But when responses are isolated in a small number of categories (very skewed distributions) with a limited range of possible values, MI can produce out-of-range imputations. As well as this, normality violations can introduce bias to the result. Different parameter estimates show different sensitivity to normality violations. The effects can be reduced through a normalizing transformation or other corrective procedures. Other ways of dealing with out-of-range values are 'Rounding the out-of-range values to the nearest possible score' and 'repeating the imputation for the out-of-range values, until values within the required range are obtained'.

Data from the Mater-University Study of Pregnancy (MUSP) are used for this study. Baseline information was collected for 7718 individuals during 1981 to 1983 in Brisbane, Australia. Follow-up waves occurred at 5 days, 6 months, 5 years, 14 years and 21 years and 27 years and data on demographics, lifestyle and mental health were collected. Mental health scores are in range of 0 to 4 but the imputation process produces some negative values.

Data sets with Missing values on mental health were generated by sampling 30%, 40% and 50% of the original sample and assigning values to missing. Different approaches such as normalizing transformation, rounding and truncated regression are used to avoid out-of-range values of mental health during the imputation procedure.

The results show that in this case all of these techniques give biased estimates in comparison with the estimates from the original sample. The reasons for this are investigated.

Dealing With Deaths In Longitudinal Surveys

Nicole Watson

Longitudinal surveys follow people over time and some deaths will occur during the life of the panel. Through fieldwork efforts, some deaths will be known but others will go unobserved due to sample members no longer being issued to field or having inconclusive fieldwork outcomes (such as a non-contact not followed by a contact at a later wave). The coverage of deaths identified amongst sample members has flow-on implications to non-response correction. Using the Household, Income and Labour Dynamics in Australia (HILDA) Survey, we examine the extent of missing death reports through two methods. The first method extrapolates the expected number of deaths in the original sample (selected in 2001) over the first 11 years using life expectancy tables. The second method matches the sample to the National Death Index. Both methods have challenges in their implementation and these are explored. We further examine the impact of missing deaths on assumptions in the construction of weights for the balanced panel and subsequent population inference.

Pattern Recognition Based On Naive Canonical Correlations In High Dimension Low Sample Size

Kanta Naito
Mitsuru Tamatani[1], Inge Koch[2]
[1]Graduate School of Science and Engineering, Shimane University, Matsue, Japan
[2]School of Mathematical Science, University of Adelaide, Adelaide, Australia

In multi-class pattern recognition for High Dimension Low Sample Size settings it is not possible to define Fisher's discriminant function, since the sample covariance matrix is singular. For the special case of two-class problems, the naïve Bayes rule has been studied, and combined with feature selection, this approach yields good practical results. We show how to extend the naive Bayes rule based on the naive canonical correlation matrix to a general setting for K≧2 classes, and we propose variable ranking and feature selection methods which integrate information from all K-1 eigenvectors. Provided the dimension does not grow too fast, we show that the K-1 sample eigenvectors are consistent estimators of the corresponding population parameters as both the dimension and sample size grow, and we give upper bounds for the misclassification rate.For real and simulated data we illustrate the performance of the new method which results in lower errors and typically smaller numbers of selected variables than existing methods.

Understanding Missing Data: The Use Of Classification And Regression Trees (cart), And Boosted Regression Trees (brts)

Nicholas Tierney
Jegar Pitchforth[1], Kerrie Mengersen[1]
[1]Queensland University of Technology, Brisbane

Missing data is essentially guaranteed in applied research studies, so it is of great import to understand:

- The nature and characteristics of missing data
- How the data is missing (randomly or systematically)
- How the missingness of data may bias analysis
- The steps required to handle missing data

With access to data becoming easier, and the emergence of Big Data, situations arise where researchers may not easily understand the context in which the data was measured, collected, and collated. Consequentially, researchers may find themselves confused as to what they should do with an 'unusable' dataset containing a high number of missing values. Although multiple imputation (MI) methods can provide accurate prediction of missing values, they are not always the answer to a researcher's missing data problem, and rely heavily on the researcher having a complete understanding of how data were generated. We demonstrate a method for approaching missing data when using linked health data, and propose a set of steps that could be generalised to applied research.

Rationale

We evaluate different approaches for evaluating missing data, from contextual, empirical and modelling perspectives.

Context. From the contextual perspective we explore the provenance of the data through its life from collection to compilation and presentation to the analyst. In doing so, we hope to

understand the reasons for missing data and therefore how it may be collected or imputed appropriately.

Empirical. From the empirical perspective we apply standard statistics tests to evaluate whether the presence or absence of important dependent variables is related to a significant change in other variable. Suggestions for appropriate handling of missing data are then made.

Modelling. From the modelling perspective we apply a CART and Boosted Regression Tree (BRT) to explore variables predicting missingness.

Note: This work was conducted in collaboration with Fiona Harden (QUT) and Maurice Harden (Hunter Industrial Medicine)

A Note On Orthogonal Array Composite Design

Yongdao Zhou

A note on orthogonal array composite design

Yongdao Zhou
Sichuan University, China

Abstract

The orthogonal-array composite design is a new useful type of composite design for practical application. It has more benefits than the popular used center composite design. In this paper, some theoretical result and discussion of orthogonal-array composite design are shown. For decreasing the number of runs, the two-level portion with resolution III or higher is permitted and the number of center point can be smaller than that of center composite design.

Fluctuation Versus Fixation In The Constrained Voter Model

Nicolas Lanchier
Stylianos Scarlatos[1]
[1]University of Patras

The constrained voter model describes the dynamics of opinions in a population of individuals located on a connected graph. Each agent is characterized by her opinion, where the set of opinions is represented by a finite sequence of consecutive integers, and each pair of neighbors, as defined by the edge set of the graph, interact at a constant rate. The dynamics depends on two parameters: the number of opinions and a so-called confidence threshold. If the opinion distance between two interacting agents exceeds this threshold then nothing happens, otherwise one of the two agents mimics the other one just as in the classical voter model. The main question about this process is whether the system fluctuates or fixates, i.e., whether the number of opinion changes at each vertex is infinite or finite. In this talk, we prove necessary and/or sufficient conditions for fluctuation and fixation of the one-dimensional system that extend and contradict previous conjectures based on numerical simulations.

Semi-parametric Garch Via Bayesian Model Averaging

Wilson Ye Chen
Richard Gerlach[1]
[1]University Of Sydney

In a standard GARCH model, the next period variance is driven by a quadratic function of the current shock. A semi-parametric GARCH model that makes fewer assumptions on the functional form of the news impact function is proposed by the authors, where the relationship between the next period variance and the current shock is modelled by a spline. The knots of the spline are determined using a Bayesian variable selection approach, where a Metropolis-Hastings algorithm is used to generate samples from the joint posterior distribution of the model parameters and the knots. The next period variance is then obtained by model averaging using the generated samples. In a simulation study, the performance of the proposed approach is compared to parametric GARCH models in cases where the parametric models are correctly specified and where they are misspecified. In an empirical study, the proposed approach is applied to the returns of stock indices and individual stocks, where the accuracy of the one-step-ahead volatility forecasts are assessed.

Spatial-temporal Modelling Of Disease Patterns When The Data Contain Many Zero Counts

Oyelola Adegboye
Denis Leung, You-Gan Wang

Leishmaniasis is a serious health concern in Afghanistan with about 250,000 estimated new cases of cutaneous infection nationwide and 67,000 cases in Kabul alone. This paper studies the spatio-temporal pattern of Leishmaniasis incidence in Afghanistan using data from 2003-2009. The data is characterized by a high percentage of zero disease counts. Since the data covers a period that overlaps with the US invasion of Afghanistan, the zero counts may be the result of no disease incidence or lapse of data collection. To resolve this issue, we use a model truncated at zero. Our approach is built on a foundation of the generalized estimating equations, which has the advantage of producing consistent regression parameter estimates under mild conditions due to separation of the processes of estimating the regression parameters from the modelling of the correlation, and therefore, estimates of the regression parameters are consistent under mild conditions. To account for the spatial-temporal nature of the data, we propose a method that decouples the two sources of correlations. Specifically, we model spatial and temporal effects separately and then combine them optimally. Our approach circumvents the need of inverting the full covariance matrix and simplifies the modelling of complex relationships such as anisotropy, which is known to be extremely difficult or impossible to model in analyzing spatio-temporal data.

KEY WORDS: Generalized method of moments; Generalized estimating equations; Overdispersion; Poisson; Spatio-temporal

Non-existence Of Stable Policies For Critical Queueing Networks With Infinite Supplies

Yoni Nazarathy
Leonardo Rojas-Nandayapa[1], Thomas Salisbury[2]
[1]School of Mathematics and Physics, The University of Queensland
[2]Department of Mathematics and Statistics, York University

Multi-class queuing networks with infinite supplies are networks where servers can either serve jobs from queues or generate new arrivals to the system. The simplest interesting example is known as the push-pull network. For such networks there often exist policies that allow servers to be fully utilized while keeping queue sizes stochastically stable. Stabilizing, fully-utilizing policies often depend on the parameters of the network, yet they are known to exist for a variety of settings. This is not the case for critical networks. For certain such cases, we show that there do not exist stabilizing policies.

Snapshot Of Biometry Frontiers

Olena Kravchuk

Expertise in statistics is essential in any research-intensive bio-science environment where appropriate experimental design and statistical analysis underpin high-quality publications and the ability to meet expectations of funding bodies. The Biometry Hub is part of the School of Agriculture, Food and Wine at the Waite campus in the University of Adelaide. The Waite campus hosts many research organizations and centres, including Australian Centre for Plant Functional Genomics, the Plant Accelerator, and the Wine Innovation Cluster, and is renowned for outstanding research. The Hub's involvement in the biological frontiers at Waite and a well-established collaboration with the group of Prof. Brian Cullis (NIASRA, University of Wollongong) shapes the group's current research profile and further growth. In this talk, I give an overview of our research directions.

- Advanced linear mixed models for designed experiments – developing powerful statistical software and computational platforms (ASREML) enables plant scientists to establish and analyse complex relationships in spatial and temporal trials across multiple field and controlled environments.
- QTL and genetic analysis – developing and improving efficient and robust algorithms and tools to handle complex genetic data (WGAIM R package) enhances the current plant breeding programs at Waite and in Australia.
- Training in biometrics and statistics capacity building – designing and delivering project-based advanced statistics learning methodologies makes the wealth of statistical theory and methods accessible for biologists.
- Statistical designs for Food Sensory and Nutrition – developing advanced mixture designs enhances the understanding of the combined effects of several food ingredients (e.g. fat, sugar and protein in a diet or a food product) in the food sensory and nutrition research.
- Robust inference for ratio indices in plant sciences – investigating distributional properties of various productivity, efficiency and biomass partitioning indices and setting up robust statistical methodologies of analysis allows biologists to accurately use such traits in their decision making.

Analysis Issues In Perinatal Trials With Multiple Births

Lisa Yelland

Perinatal trials including infants from both single and multiple births present unique statistical challenges. Cluster randomisation is common in this setting to ensure that all infants from the same birth receive the same treatment. While methods for analysing clustered data are widely available, perinatal trials are somewhat different to the usual clustered data settings due to the small cluster sizes. Many clusters consist of only a single infant, which results in a mixture of independent and clustered data. Conflicting recommendations have been made regarding if and how clustering due to multiple births should be taken into account in the analysis of perinatal trials, particularly when the multiple birth rate is low. The potential for informative cluster size to occur in perinatal trials, where the outcome is related to the size of the cluster, has also recently been recognised and this has implications for the choice of analysis method. In this presentation, I will discuss the different analysis approaches that may be used in perinatal trials including infants from both single and multiple births, and will explore some of the factors that influence whether clustering due to multiple births should be taken into account in the analysis.

Diagnostic Tests For The Location-shift Assumption

Olivier Thas
John Rayner[1], Jan De Neve[2]
[1]University of Wollongong
[2]Ghent University

The Wilcoxon rank-sum test is often considered as the nonparametric version of the t-test for comparing means. However, the hypotheses of the Wilcoxon test can only be expressed in terms of means under restrictive distributional assumptions. The most common assumption demands that the distributions of the two populations belong to a location-shift family, i.e. the distributions agree in shape except for a location shift. We present tests for testing the location-shift assumption. A first test is based on the Wasserstein distance between the two empirical quantile functions. A second class of tests is based on the set of semiparametric efficient score statistics. In an empirical power study we compare the tests. Finally, we present some meaningful graphical diagnostic tools for assessing the location-shift assumption.

Planning Accelerated Life Testing With Two Experimental Factors When Heteroscedasticity

Xiaojian Xu
Mark Krzeminski

We present methods for obtaining optimal designs for time-censored accelerated life tests with two experimental factors. The data collected from designed experiments are used for estimating percentiles of product life at usage conditions in the presence of heteroscedasticity. We assume a Weibull distribution and log-linear life-stress and scale-stress relationships. Both cases of the parameters in the scale-stress relationship being prespecified and not being prespecified are considered. We take as the primary optimality criterion the minimization of asymptotic variance of the maximum-likelihood estimator of life percentiles at usage conditions. In addition, we discuss D-optimality and A-optimality as secondary criteria. Our methods are illustrated by finding optimal designs for two practical examples and a comparison study is also presented.

Recent Advances In Multivariate Multiple Imputation For Mar Missing Data In Health

Lucy Leigh
Irene Hudson[1]
[1]University of Newcastle

Missing data are a common problem in behavioural and medical sciences - for example, in both randomised control trials and observational studies patient reported outcomes are common, but are also prone to missing data due to item and survey non-response. Inappropriate handling of missing data, such as complete case analysis, pairwise deletion or single imputation methods, will often lead to biased results unless the missing data is assumed to be missing completely at random (MCAR). However, an MCAR assumption in behavioural and medical data sets is generally unrealistic, and a missing at random (MAR) assumption is more plausible. Under the MAR assumption the missing data are related to one or more observed variables (such as previous values of the outcome, or other observed covariates). Multiple imputation (MI) is a flexible and appropriate method for dealing with missing data when the missing is assumed to be missing at random (MAR). The first aim of this presentation is to inform researchers new to MI of the benefits of commonly available multivariate MI procedures, such as joint modelling MI and sequential MI, as well as limitations that can arise from complex data sets, such as can be commonly found in health and behavioural research. These complications include a large number of mixed variable types with missing data, non-normal variables, non-monotone patterns of missing data, a longitudinal design, hierarchical design or clustering of the data, and missing values on the covariates as well as the response variable. The second aim of this presentation is to present recent advances in the literature for dealing with some of these limitations. These advances include latent class MI, semi-parametric MI using predictive mean matching or splines, MI for meta-analytic studies, longitudinal MI using functional mixed effects models or quasi-imputation, and MI diagnostics.

New Results On Sobol' Sensitivity Indices

Art Owen

Many problems in science and engineering are addressed via deterministic computational models. These embed known physics to predict one or more outcomes based on a set of input factors. Even in deterministic functions it is challenging to know which variables are most important, which interact, and so on.

Sobol' indices describe the relative importance of input variables and subsets thereof in terms of a functional ANOVA decomposition. Sobol's identities let one evaluate his indices without actually estimating, squaring, integrating and summing any interactions. His identities are like a form of tomography, where quadrature alone reveals internal structure.

For some Sobol' identities there are multiple ways to compute estimates, and these can have markedly different efficiencies. This talk reviews Sobol' indices and presents some new results on their estimation: new Sobol' style quantities, new estimates geared to estimating small indices, and higher order Sobol' indices (joint with Josef Dick of UNSW).

When Should Matching Be Used In The Design Of Cluster Randomised Trials?

Patty Chondros
Obioha C Ukoumunne[1], Jane M Gunn[2], John B Carlin[2]
[1]University of Exeter Medical School
[2]The University of Melbourne

A matched-pair (MP) cluster randomised trial (CRT) may be adopted in an effort to minimise imbalance on known prognostic factors, add study credibility and increase efficiency provided the analysis recognises the matching. We lack evidence to guide decisions about when to use the MP design, especially for trials with few clusters. This study **aimed** to develop practical guidance for researchers and statisticians about when to use the MP design.

Methods: In a simulation study the efficiency of the MP design was compared with the stratified and simple randomisation designs using the mean confidence interval width of the estimated intervention effect. A matched and unmatched analysis was used for the MP design, a stratified analysis for the stratified design, and an analysis with and without post-stratification adjustment for potential matching risk factors for the simple design.

Results: The MP design was most efficient for CRTs with 10 or more pairs when the correlation between paired cluster-level outcomes was greater than 0.1. For trials with fewer clusters and matching correlation less than 0.5, simple and stratified designs were more efficient than the MP design because greater degrees of freedom were available for estimation, but an unmatched analysis of the MP design recovered precision. The stratified and simple design with post-stratification analysis had similar efficiency, provided the number of strata was small relative to the number of clusters.

Conclusions: The simple design is recommended for CRTs with matching correlations less than or equal to 0.1. For stronger matching correlations, the MP or stratified design (or simple design with post-stratification analysis) is recommended for CRTs with 10 or more clusters/arm. For CRTs with fewer clusters, the stratified design is a good compromise between the MP design, which may lack efficiency and the simple design where the chance of imbalance on risk factors is greater.

Measuring Change Overtime: Recommendations On Comparisons Of Survey Estimates

Lujuan Chen
Anura Amarasinghe[1]
[1]Australian Bureau of Statistics

Abstract: Despite their widespread application, statistical hypothesis significance tests may add little value to the product of research and studies; they are the most widely accepted and frequently used tools of statistical inference in measuring progress over time in the report of The Council of Australian Governments. The criticisms against them have grown dramatically including a sensitivity to sample size, unacceptable Type II error rates, inability to reflect the practical significance, misunderstanding and abuse, and limitation on the usage of all available information. The ABS has been invited to provide advice on the reliability of comparisons of survey estimates as well as comparisons based on administrative data.

This paper outlines a proposal to provide more meaningful statistical alternatives on the measurement of change including reporting effect size and its confidence interval, undertaking power analysis, decision theory for guiding actions in the face of uncertainty, applying Bayesian approaches to hypotheses, and using multiple observations to generate growth curves.

Key words: Statistical hypothesis test, confidence interval, power analysis, statistical power, Bayesian analysis.

Likelihood-based Estimators For Meta-analysis And Meta-regression Analyses

Luke Prendergast
Michael Malloy[1], Robert Staudte[2]
[1]Victorian Cytology Service, Melbourne
[2]Department of Mathematics and Statistics, La Trobe University, Melbourne

In this paper we present the distribution for standardized difference of means estimators under the assumption that the data is sampled from a normal distribution which consists of a normally distributed random effect component. This distribution, a rescaled non-central t and which is not conditional on the random effect, can be used to formulate the joint distribution for standardized mean difference estimates collected from a number of independent studies. We show how this leads to likelihood-based estimators for fixed and random effect parameters for both the meta-analysis and meta-regression contexts. Additionally, we explore the use of normalization transformations on the estimates that can be used to obtain approximate joint distributions based on normal densities. This can be used to reduce the computational effort in finding the maximum of the log-likelihood function. An advantage of these two approaches is that they do not require the studies to consist of large sample sizes which is commonly assumed to be the case. Our simulations show that our methods can produce excellent results in both small and large sample scenarios and we provide comparisons with other popular estimators. We also highlight how simple these estimates and associated Wald-type and profile-based confidence intervals are to obtain using existing functionality within the R statistical package. Several examples are also presented using data collected from the scientific literature.

Heat Kernel Of Non-symmetric Levy Operators

Xicheng Zhang

We construct the heat kernel of non-symmetric Levy operators, and prove two-sided sharp estimates and gradient estimate of the heat kernel. The lower bound estimate is based on a probabilistic approach. This is partly a joint work with Zhenqing Chen.

A Comparison Of Bayesian And Frequentist Interval Estimators In Regression That Utilize Uncertain Prior Information

Paul Kabaila
Gayan Dharmarathne[1]
[1]University of Colombo

Consider a linear regression model with normal errors. Suppose that the parameter of interest θ is a specified linear combination of the regression parameters. Also suppose that we have uncertain prior information that τ, a distinct specified linear combination of these parameters, is zero. Our aim is to utilize this uncertain prior information in the construction of an interval estimator for θ. This can be done in two ways: Bayesian and frequentist. A Bayesian $1-\alpha$ credible interval that utilizes this uncertain prior information is obtained by using an informative prior distribution for τ, combined with noninformative prior distributions for the other parameters. We assess a frequentist $1-\alpha$ confidence interval J for θ by its scaled expected length, defined to be (expected length of J)/(expected length of the standard $1-\alpha$ confidence interval for θ). A $1-\alpha$ frequentist confidence interval for θ utilizes the uncertain prior information that $\tau=0$ if it has the following properties: (a) it has scaled expected length that is substantially less than 1 when $\tau=0$, (b) the maximum value of the scaled expected length is not too large and (c) this confidence interval reverts to the standard $1-\alpha$ confidence interval for θ when the data happen to strongly contradict the prior information. Kabaila and Giri, *JSPI*, 2009, describe a new frequentist $1-\alpha$ confidence interval for θ that utilizes the uncertain prior information that $\tau=0$. We compare this confidence interval with Bayesian $1-\alpha$ credible intervals for θ that result from a prior density for τ that is a mixture of an infinite "slab" and a Dirac delta "spike", combined with noninformative priors for the other parameters. These Bayesian and frequentist interval estimators depend on the data in very different ways. We also consider some variants of this prior distribution that lead to greater similarity between these estimators.

Two-sample Covariance Matrix Testing And Support Recovery

Yin Xia
Tony Cai[1], Weidong Liu[2]
[1]University of Pennsylvania
[2]Shanghai Jiao Tong University

This talk proposes a new test for testing the equality of two covariance matrices in the high-dimensional setting and investigates its theoretical and numerical properties. The limiting null distribution of the test statistic is derived. The test is shown to enjoy certain optimality and to be especially powerful against sparse alternatives. When the null hypothesis of equal covariance matrices is rejected, recovering the support of the difference of two covariance matrices is also studied.

Governing Equations For The Reflected Spectrally Negative Process

Peter Straka
Boris Baeumer[1], MihÃ¡ly KovÃ¡cs[1], Mark Meerschaert[2], RenÃ© Schilling[3]
[1]University of Otago, Department of Mathematics and Statistics
[2]Michigan State University, Department of Probability and Statistics
[3]Technische UniversitÃ¤t Dresden, Institut fÃ¼r Mathematische Stochastik

We show how to explicitly compute the transition densities of a spectrally negative stable process with index greater than one, reflected at its infimum. First we derive the forward equation using the theory of sun-dual semigroups. The resulting forward equation is a boundary value problem on the positive half-line that involves a negative Riemann-Liouville fractional derivative in space, and a fractional reflecting boundary condition at the origin. Then we apply numerical methods to explicitly compute the transition density of this space-inhomogeneous Markov process, for any starting point, to any desired degree of accuracy. Finally, we discuss an application to fractional Cauchy problems, which involve a positive Caputo fractional derivative in time.

A Geostatistical Approach For Landscape Image Classification

Ronny Vallejos
Adriana Mallea[1], Myriam Herrera[1], Silvia Ojeda[2]
[1]Universidad Nacional de San Juan
[2]Universidad Nacional de Cordoba

In this talk we describe a methodology to deal with the problem of classification of images that have been acquired from remote sensors. The classification method consists in reducing the dimensionality of the spectral bands associated with a multispectral satellite image. Such dimensionality reduction is accomplished by using the divergence of a modified Mahalanobis distance. Instead of using the covariance matrix of a multivariate spatial process we consider the codispersion matrix, which under very precise conditions have some desirable asymptotic properties. The consistency and asymptotic normality hold for a general class of processes that is a natural extension of the well known one-dimensional spatial processes for which the asymptotic properties were first established. The results allow one to select a set of spectral bands that produce the highest value of divergence. Then a supervised maximum likelihood method using the selected spectral bands is considered to perform landscape classification. An application with a real LANDSAT image is introduced to explore and visualize how our method works in practice.

Inverse Problems For Regular Variation

Jan Rosinski
Ewa Damek[1], Thomas Mikosch[2], Gennady Samorodnitsky[3]
[1]University of Wroclaw
[2]University of Copenhagen
[3]Cornell University

Regular variation is one of the basic concepts used to model heavy tailed phenomena in one dimension as well as in the multivariate case. An important feature of regular variation is that it tends to be preserved by various linear operations on random structures (such as weighted sums, products, integrals, etc.) An inverse problem for regular variation aims at understanding whether the regular variation of a transformed random object is caused by regular variation of components of the original random structure. It is somewhat surprising that such inverse problem can be very sensitive to change of parameters in a random structure (e.g., change of coefficients in a weighted sum). We give complete answers to the inverse problems for weighted sums, products and stochastic integrals in one dimension, and for some special types of weighted sums and products in the multivariate case. Our results in the multivariate case cover the situation where regular variation is not restricted to any one particular direction or quadrant.

Forecasting Coconut Yield At Bandirippuwa Estate By Exploring Annual Seasonal Cycles

Samithree Rajapaksha
Chandima Tilakaratne[1]
[1]Department of Statistics, University of Colombo, Sri Lanka

The environmental factors play an important role in either to increase or decrease coconut pick yield in any country. Therefore it is recommended to consider the behavioral cyclic patterns of the environmental factors when modeling the coconut yield. Incorporating environmental factors in modeling coconut yield will also help to take precautionary actions for a worst case scenario which can be happen in the future.However the past studies engrossed on forecasting coconut yield in Sri Lanka does not consider the effect of predominant environmental factors much accurately even with the seasonally adjusted data. Therefore this study models the association between seasonally unadjusted bi-monthly coconuts pick yield and lagged environmental factors from January 1990 to July 2010 of the Bandirippuwa Estate, Sri Lanka.HEGY testing procedure introduced to identify the seasonal cycles in quarterly data was extended in order to identify the availability of seasonal cycles in bi-monthly data. All the meteorological factors as well as the coconut yield evident the seasonal cyclic patterns and it was found that the critical period for the current coconut yield is three picks prior to the current harvest. Nonlinear Autoregressive Neural Network with Exogenous input variables (NARX models) were fitted with and without incorporating seasonal cycles of the metrological factors and the coconut yield. The coefficient of determination (R^2) of predicted values on actual values of test cases was 99.5% for the model which incorporated with the seasonal cyclic behaviors. Therefore it is evident that the identified seasonal behaviors expedite the NARX model for forecast the bi-monthly coconut yield in Bandirippuwa Estate.

Use Of Observational Data To Examine Treatment Effects Of Medicare Subsidised Mental Health (bas) Services.

Xenia Dolja-Gore
Deborah Loxton, Cate D'Este, Julie Byles

Introduction: Observational data is often used to examine the relationship between explanatory variables and an outcome; however this type of analysis does not allow a causal interpretation. Methods have been proposed to investigate causal relationships in observations data, which would allow assessment of the effect of a particular treatment or intervention of interest. The aim of this study was to compare mental health outcomes of Australian women with declining mental health who do and do not uptake BAS counseling services, using propensity analysis.

Methods: The Australian Longitudinal Study on Women's Health data provided baseline measures to derive propensity scores. These scores estimate each participant's predicted probability of using a BAS counseling service. Participants were stratified[1] into quintiles based on their propensity scores and pooled estimates were used to evaluate the average treatment effects (ATE) and average treatment effect on the treated (ATT).

Results: Stratification eliminated a large proportion of the bias leaving a well-balanced model between those participants that had and had not used the BAS counseling services. Stratum-specific estimates of the use of the BAS counseling services suggest a greater proportion of participants with poorer mental health using the counseling services had improvement in their mental health outcome score.

Conclusions: Propensity score modelling allowed estimation of the marginal treatment effects of use of BAS counselling services on mental health similar to randomised control trials. Large cohort datasets are rich sources of information which can be resourcefully used to address causal inference. Propensity score modelling techniques are timely and cost efficient and can be used in situations where randomised control trials are unethical or infeasible.

- Word Count: 287
- Must be submitted by 30 October 2013 via the conference website:
  - http://www.ims-asc2014.com

[1] Rosenbaum PR & Rubin DB (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70, 41-55.

The Structure Of Interdependence Between Stock Market Movement And Interest Rate Dynamics: A Nigerian Situation.

Chibuzo Gabriel Amaefula

This paper investigates the structure of interdependence between stock market movement and interest rate dynamics in Nigeria. The data sets on yearly transactions at the Nigerian Stock Exchange (NSE) and deposit rate used cover the period of 1961 to 2012. The Diagonal BEKK parameterization of a bi-variate GARCH(1, 1) model was employed to capture the correlation structure between stock market movement and interest rate dynamics. The method of Least Square (LS) was used to estimate the nature of the trend. And the result shows that the correlation between stock market movement and interest rate varies over time and the trend exhibits an upward movement and it is significant at 1% level. The average value of correlation is -0.27. The observable structural breaks in the time varying correlation were attributed to high interest rate regimes in 1997 and 2007-2009 due to interest rate deregulation and global financial crises respectively. This result indicates that there exit inverse relationship between stock market movement and interest rate changes and any structural change due to Government policy in either of the variable will affect the other. So, the Government should keep the money market interest rates in a level that will spur the growth of the stock market condition as well as the economy at large.

The Effects Of Deregulation And Global Financial Crises On Stock Market Returns And Volatility: Evidence From Nigeria

Chibuzo Gabriel Amaefula

This paper examines the effects of deregulation of capital market and global financial crises on stock market returns and volatility in Nigeria. The data set on monthly All Shares Index covers the period of 1985 to 2012. The periods of deregulation and global financial crises were respectively represented using dummy variables. GARCH (1, 1) was adopted to measure the volatility of stock market returns and the method of Least Squares was used to estimate the effects of deregulation and global financial crises on both stock market returns and volatility. The result reveals that deregulation has no significant effect on both stock market returns and volatility but, the global financial crises exerts significant negative effect on stock market returns and significant positive effect on the volatility of stock market returns. This result implies that during the period of global financial crises there was a fall in stock market returns and a rise in volatility. It thus, becomes imperative for the Government and policy makers to develop proactive measures that will enhance capital market development and also hedge it against external shocks.

Hpd Regions For Computer Experiments

Robert Wolpert

Many physicists like to think about uncertainty in data and models in a different way than most statisticians do--- they regard the evidence from experiments as a "confinement" of parameter values, a separation of the parameter space into one (hopefully small) region in which the "true" parameter value might lie, and the region's complement whose values are untenable in light of the experiment. It is hoped that identifying and visualizing the boundary of the tenable regions may lead to further physical insight and suggest ways to improve existing physical models. In support of this perspective, we are developing methods intended to identify "highest posterior density" or "hpd" regions of parameter space from MCMC output streams.

For one-dimensional parameter spaces this is a well-understood problem solved in 1999 by Ming-Hui Chen and Qi-Man Shao. In two or more dimensions the problem is more delicate.

As a step toward finding suitable approximate HPD regions, we generate convex polygons in the plane and polyhedra in higher dimensions and mount a stochastic search for the smallest-volume polygons containing a prescribed fraction of the MCMC points.

Correlation Of Intracellular Components Due To Limited Processing Resources

Ruth Williams

A major challenge for systems biology is to deduce the molecular interactions that underlie correlations observed between concentrations of different intracellular components. Of particular interest is obtaining an understanding of such effects when biological pathways share common elements that are limited in capacity. Here we use stochastic models to explore the effect of limited processing resources on correlations when these resources are positioned downstream or upstream of the molecular species of interest. In both situations, we observe a correlation resonance phenomenon: correlations tend to have an extremum slightly beyond the point where the systems transition from underloading to overloading of the processing resources, although the sign of the correlation is different in the two cases. As time permits, related experimental work will be described.

This presentation is based on joint work with current or former members of the UCSD Biodynamics lab and in particular with William H. Mather, Natalie A. Cookson, Tal Danino, Octavio Mondragon-Palomino, Jeff Hasty and Lev S. Tsimring.

Feature Extraction For Characterising Cancer In Spatial Proteomics Mass Spectrometry Data

Lyron Winderbaum
Inge Koch[1], Peter Hoffmann[1]
[1]University of Adelaide

Recent technological advances in Matrix Assisted Laser Desorption Ionisation (MALDI) imaging mass spectrometry have facilitated the acquisition of spatially resolved proteomics data on tissue. We consider data collected from samples that include regions of cancerous and healthy tissue from patients with ovarian cancer. Our exploratory analysis is part of a comprehensive approach which aims to gain a better understanding of the cancer proteome, including identifying proteomic differences between patients who respond to a treatment and those who do not. We discretise the functional proteomics data into binary data which shows the presence or absence of molecular masses.

Our feature extraction approach exploits the spatial aspect of the data, which have been collected at thousands of grid points on the tissue sample, by including a spatial smooth based on cellular automata. The spatial smooth is combined with k-means clustering of the binary data, and leads to cluster maps which show the cluster membership at each grid point. The cluster membership correlates well to the histology visible in the tissue. Feature extraction of masses which characterise the cancer cluster yields masses the biologist can identify with appropriate follow-up analyses, and leads to dramatic variable reduction. The quantitative results are complemented by visualisations of the distribution of masses as spatial maps that enhance the cluster maps. Comparing the spatial maps with annotations by histopathologists provides insight into the pathology of the cancer as it enables the important link between molecular information and histology.

Matrix Completion Via Max-norm Regularized Approach

Wen-Xin Zhou
T. Tony Cai[1]
[1]University of Pennsylvania

In this talk, we consider both standard and 1-bit noisy matrix completion problems under a general sampling model using the matrix max-norm as a convex relaxation for the rank of the matrix. The max-norm constrained minimization method is introduced and studied. The rate of convergence for the estimate is obtained in both cases. Information-theoretical methods are used to establish a minimax lower bound under the general sampling model. It is shown that the max-norm regularized method is rate-optimal and it yields a stable approximate recovery guarantee with respect to the sampling distribution. Computational effectiveness of this method is also discussed, based on a first-order algorithm for solving convex programs involving a max-norm constraint. This talk is based on a joint work with Professor Tony Cai.

Hunting For Significance: Nonparametric Bayesian Rules

Linda Zhao

Hunting for significance: Nonparametric Bayesian rules

In modern data analyses, we often run into the situation that a large number of hypotheses need to be tested simultaneously. Yet only a few of the alternatives hypotheses are believed to be true. We propose to use Bayesian nonparametric schemes to tackle the problem. The FDR criterion is used.

Bayesian Dynamic Time-frequency Estimation

Wen-Hsi Yang
Scott Holan[1], Christopher Wikle[1]
[1]Department of Statistics, University of Missouri

We propose a novel approach to model-based time-frequency estimation using time-varying autoregressive models. In this context, we take a fully Bayesian approach and allow both the autoregressive coefficients and innovation variance to vary over time. Importantly, our estimation method is cast within the partial autocorrelation domain. The majority of the marginal posterior distributions are of standard form and, as a convenient by-product of our estimation method, our approach avoids undesirable matrix inversions. As such, estimation is computationally efficient and stable. We conduct a comprehensive simulation study that compares our method with other competing methods and find that, in most cases, our approach performs superior in terms of average squared error between the estimated and true time-varying spectral density. Lastly, we demonstrate our methodology through several real case studies.

Feature Sensitive And Automated Curve Registration

Radhendushka Srivastava
Dibyendu Bhaumik[1], Debasis Sengupta[2]
[1]Reserve Bank of India, Mumbai, India
[2]Indian Statistical Institute, Kolkata, India

Paleo-climatic variables, e.g., temperature and atmospheric concentrations of greenhouse gases like carbon dioxide and methane, are measured from chemical analysis of the ice itself or of air-bubbles trapped in it. These observations are recorded against age, which is determined by dating techniques involving radioactive isotopes. Such ordered pairs of data can be regarded as functional data sampled at irregularly spaced time-points. These dating techniques have certain imperfections. When observations on a particular variable obtained from different ice cores from a common geographic region are available, it seems reasonable to assume that the data sets arose from the same underlying process. Collation and joint analysis of these comparable sets of data present difficulties, as imperfections in dating may lead to non-uniformity of the time scales of different data sets. This problem is generally resolved by identifying one data set as reference and aligning others with it, through a task called curve registration.

Given two sets of functional data having a common underlying mean function but different degrees of distortion in time measurements, we provide a new estimator of the time transformation necessary to align (or 'register') them. Instead of relying on smoothing, the proposed method aims to harness the information contained in the sharp features of the data, which are often found to be useful in manual alignment. Our automated procedure is in sharp contrast with land-mark based techniques that seek to exploit such features after manually identifying them. The consistency of the estimator is proved under general conditions. Simulation results show superiority of the performance of the proposed method over its competitors. The method is illustrated through the analysis of some paleo-climatic data sets.

A Multidimensional Stochastic Fluid Model For An Ad Hoc Mobile Network

Peter Taylor
Guy Latouche[1], Nikki Sonenberg[2]
[1]Departement d'Informatique, Universite Libre de Bruxelles.
[2]Department of Mathemtaics and Statistics, University of Melbourne

In conventional mobile telephone networks, users communicate directly with a base station, via which their call is

transferred to the recipient. In an ad hoc mobile network, there is no base-station infrastructure and users need to

communicate between themselves, either directly if they are close enough or via transit nodes if they are not.

A number of interesting questions immediately arise in the modelling of ad hoc mobile networks. One concerns the

`amount of resource' that a network needs in order to be able to operate with a reasonable quality of service. We

shall consider this question by modelling each user's battery energy as a uid, the rate of increase or decrease of

which is modulated by the network occupancy. This results in a network of stochastic uid models, each modulated

by the same background process.

Using this model, we can calculate the background rate of re-charge that is necessary and su_cient to guarantee that

no connections are lost. For recharge rates less than this, we propose a reduced-load approach to the calculation of

dropout probabilities

Conditional Moment Restrictions Estimation: Finite Sample Theory

Valentin Patilea
Vladimir Spokoiny[1]
[1]Weierstrass Institute, Humboldt University, Moscow Institute of Physics and Tech

We are interested by statistical models where parameters are identified by a set of conditional estimating equations, also called moment restrictions. Such statistical models are semiparametric in the sense that one aims at estimating a finite dimensional vector of parameter without specifying entirely the distribution of the variables of interest. Nonlinear mean or quantile regressions, econometric instrumental variables models, are only few examples of statistical models that fit into this framework. We consider the smooth minimum distance (SMD) estimation method introduced by Lavergne and Patilea (2013, "Smooth Minimum Distance Estimation and Testing with Conditional Estimating Equations: Uniform in Bandwidth Theory", *Journal of Econometrics*) for inference on the parameters of the model. We investigate the statistical properties of the SMD estimates using modern tools as proposed by Spokoiny (2012, "Parametric Estimation. Finite sample Theory", *Annals of Statistics*). The main feature of the new approach is the non-asymptotic nature of the results. Moreover, the model could be misspecified. We deduce concentration and confidence sets, risk bounds and local expansions of the minimum of the empirical criterion and the corresponding estimate. The asymptotic results of Lavergne and Patilea can be easily derived as corollaries of the new nonasymptotic statements. At the same time, the new approach works well in the situations with large or growing parameter dimension in which the previous parametric theory fails. The results apply for any dimension of the parameter space and provide a quantitative lower bound on the sample size yielding the root-n accuracy.

False Discovery Control In Large-scale Spatial Multiple Testing

Wenguang Sun
Brian Reich[1], Tony Cai[2], Michele Guindani[3], Armin Schwartzman[1]
[1]North Carolina State University
[2]University of Pennsylvania
[3]UT MD Anderson Cancer Center

This talk discusses a unified theoretical and computational framework for false discovery control in multiple testing of spatial signals. We consider both point-wise and cluster-wise spatial analyses, and derive oracle procedures which optimally control the false discovery rate, false discovery exceedance and false cluster rate, respectively. A data-driven finite approximation strategy is developed to mimic the oracle procedures on a continuous spatial domain. Our multiple testing procedures are asymptotically valid and can be effectively implemented using Bayesian computational algorithms for analysis of large spatial data sets. Numerical results show that the proposed procedures lead to more accurate error control and better power performance than conventional methods. We demonstrate our methods for analyzing the time trends in tropospheric ozone in eastern US.

Functional Regression Models Checks

Valentin Patilea

Given a response variable taking values in a Hilbert space and covariates that could be of finite or infinite dimension, the problem of testing the effect of the functional covariates on the response variable is addressed. This problem occurs in many situations as for instance significance testing for functional regressors in nonparametric regression with hybrid covariates and scalar or functional responses, testing the effect of a functional covariate on the law of a scalar response, testing the goodness-of-fit of regression models for functional data. We propose a new testing approach based on univariate kernel smoothing. The test statistic is asymptotically standard normal under the null hypothesis provided the smoothing parameter tends to zero at a suitable rate. The one-sided test is consistent against any fixed alternative and detects local alternatives approaching the null hypothesis at suitable rate. In particular we show that neither the dimension of the outcome nor the dimension of the functional covariates influences the theoretical power of the test against such local alternatives. Simulation results and real data applications illustrate the new approach.

Survey Design Within The International Development Sector

Mark Griffin

Survey design at the best of times involves a wide range of educated guesses about the nature of the target population. This field becomes more complex still within the field of international development where a target population may be migratory in nature, the definition of a "household" may be ill-defined, and there may be scarce information available from previous surveys and censuses. In addition the technical capacity for designing, conducting, and analyzing surveys may be very limited within some geographical regions.

Within this presentation Dr Griffin will discuss some of his achievements and challenges in designing and analyzing surveys within this context. Some of the surveys that he has recently been involved with include (a) a survey of 5000 households in Cambodia, China, Laos, Myanmar, Thailand and Vietnam exploring residents' knowledge and attitudes towards human trafficking, (b) a survey of 20,000 households in Sri Lanka in the field of human trafficking, and (c) a survey of store owners in PNG evaluating store owners attitudes towards the distribution of condoms and its potential effect on the AIDS epidemic in PNG.

Network Tomography For Integer-valued Traffic

Martin Hazelton

Volume network tomography is concerned with inference about route traffic flow characteristics based on traffic measurements at fixed locations on the network. A classic example is estimation of the traffic flow between origin and destination nodes using traffic counts obtained from a subset of the links of the network. The observed data provide only indirect information about the target variables, giving rise to a challenging type of statistical linear inverse problem. In principle maximum likelihood and Bayesian inference can be implemented through stochastic EM and MCMC methods, but implementation requires an efficient method of sampling route flows conditional on the observed pattern of traffic counts. None of the algorithms proposed in the literature have proved reliable for integer-valued traffic. In this talk I shed light on why this is the case by examining the geometry of the space of feasible route flows. I then describe a modified MCMC sampler with much improved mixing behaviour, and illustrate its application on some simple real-world examples.

An Outlier Robust Small Area Predictor For Count And Binary Data

Payam Mokhtarian

Count and binary data are often of interest in surveys, particularly when the aim is small area estimation. Plug-in predictors of small area characteristics based on estimated parameters of small area models have been used to estimate small area totals for count and binary data. Generalized linear mixed models (GLMMs) are widely used to fit such non-normal response data when over-dispersion due to small area effects is believed to be present. If the data are outlier free, then the generalized maximum quasi-likelihood (GMQL) approach fitting a GLMM works well in practice and can be used to obtain consistent as well as efficient estimates for model parameters and consequently small area estimates. However, this approach can be highly influenced by the presence of outliers in the data. In this paper, we first examine the effect of the presence of outliers on GMQL estimation for the parameters in a GLMM as well as the corresponding small area estimator. We then develop a robust block bootstrap technique for finding outlier robust estimates of the GLMM parameters, which appears to be useful in down-weighting influential data points when estimating the model parameters. Using this outlier robust approach we can obtain corresponding robust small area estimates for count and binary data. An extensive simulation study is conducted to examine the performance of the proposed robust approach. Moreover, we illustrate the method by using it in an analysis of real data set.

Key words: Bootstrap, Generalised Linear Mixed Model, Outlier, Robustness, Small Area Estimation.

Cluster Analysis: Unsupervised Learning Via Supervised Learning With A Non-convex Penalty

Wei Pan
Xiaotong Shen[1], Binghui Liu[1]
[1]University Of Minnesota

Clustering analysis is widely used in many fields. Traditionally clustering is regarded as unsupervised learning for its lack of a class label or a quantitative response variable,

which in contrast is present in supervised learning such as classification and regression.

Here we formulate clustering as penalized regression with grouping pursuit.

In addition to the novel use of a non-convex group penalty and its associated unique operating characteristics in the proposed clustering method, a main advantage of this formulation is its allowing borrowing some well established results in classification and regression, such as model selection criteria to select the number of clusters,

a difficult problem in clustering analysis. In particular, we propose using the generalized cross-validation (GCV) based on generalized degrees of freedom (GDF) to select the number of clusters. We use a few simple numerical examples to compare our proposed method with some existing approaches, demonstrating our method's promising performance.

Large-scale Multiple Testing Of Correlations

Weidong Liu

Multiple testing of correlations arises in many applications including gene coexpression network analysis and brain intelligence analysis. In this paper, we propose a large scale multiple testing procedure for correlations. A new test statistic is introduced and a bootstrap method is proposed for estimating the proportion of the nulls falsely rejected among all the true nulls.

The properties of the proposed procedure are investigated both theoretically and numerically. It is shown that the procedure asymptotically controls the overall false discovery rate at the nominal level. Simulation results show that the procedure performs well numerically in terms of both the size and power of the test and it significantly outperforms two alternative methods.

This work is jointed with Tony Cai.

Population-level Relevance Of Risk Factors For Cancer In The Presence Of Competing Risk Of Death

Maarit Laaksonen
Karen Canfell[1], Claire Vajdic[1], Robert MacInnis[2], Emily Banks[3], Graham Giles[2], Paul Mitchell[4], Robert Cumming[4], Barbara-Ann Adelstein[1], Julie Byles[5]
[1]University of New South Wales, Sydney
[2]Cancer Council Victoria, Melbourne
[3]Australian National University, Canberra
[4]University of Sydney, Sydney
[5]University of Newcastle, Newcastle

Quantification of the impact of exposure to different risk factors on mortality and morbidity at the population level is a fundamental public health issue. The Population Attributable Fraction (PAF) is a statistical measure which integrates the strength of the association between the exposure and the outcome and the prevalence of the exposure in the population, to assess the fraction of the outcome in the population that could be avoided if the exposure was eliminated or reduced. Such population-level information is important in planning preventive strategies.

PAF is most validly estimated from well-designed cohort studies. If the cohort is representative of the target population, exposure prevalence can be estimated from cohort data; otherwise, representative secondary sources, such as population-level surveys, can be used. The presence of potential competing risks may alter the PAF estimates, as the risk factor modification is likely to affect them as well, and thus needs to be accounted for to avoid misleading conclusions.

We will evaluate and compare the cancer burden attributable to lifestyle-related risk factors and their combinations in Australia by applying our recently developed PAF method, accounting for competing risks, to data from five established Australian cohort studies, linked to cancer and death registries, and representative national health surveys. The times until the occurrence of cancer or death are assumed to follow a parametric proportional hazards model with piecewise constant baseline hazard function. Maximum likelihood estimation will be used to obtain the parameter estimates and their estimated covariance matrices. The asymptotic variance estimate of PAF will be obtained using the delta method. We will pool the homogeneous PAF estimates and analyse reasons for heterogeneity. We expect to provide more accurate estimates of the lifestyle-related avoidable cancer burden, the most harmful risk factor combinations and the most vulnerable sub-groups, essential in prioritising preventive interventions in Australia.

Causal Discovery With Additive Noise Models

Jonas Peters
Peter Buhlmann[1], Jan Ernest[1]
[1]Seminar for Statistics, ETH Zurich

We consider the problem of learning causal directed acyclic graphs from an observational joint distribution. These graphs can be used to predict the outcome of interventional experiments, from which data are often not available. If the observational distribution follows a structural equation model with an additive noise structure, the directed acyclic graph becomes identifiable from the distribution under mild conditions. Additive noise models therefore constitute an interesting alternative to constraint-based methods. The latter can only identify the Markov equivalence class of the graph, i.e., they leave some edges undirected.

We discuss recent advances in identifiability results of additive noise models and provide statistical guarantees and new algorithms that efficiently estimate the potentially high-dimensional causal graph from a finite amount of data.

Meta-analysis Of Incidence Rate Data In The Presence Of Sparse Data

Matthew Spittal
Jane Pirkis[1], Lyle Gurrin[1]
[1]Melbourne School of Population and Global Health, The University of Melbourne

Abstract:

When summary results from studies of counts of events over time contain zeros, the study-specific incidence rate ratio and its standard error cannot be calculated because the log of zero is undefined. This poses problems for inverse-variance methods of pooling data for meta-analysis. We conducted a simulation study comparing and contrasting the standard methods of conducting meta-analysis (with and without the continuity correction) with several alternative methods based on the Poisson distribution: fixed-effects Poisson regression and mixed-effects Poisson regression implemented using the "poisson" and "xtmepoisson" procedure in Stata, and empirical Bayes Poisson regression using the JAGS implementation of the BUGS language accessed via the R package "rjags". We manipulated the sparseness of the data (from no zeros to approximately 70 percent zeros) and the heterogeneity of the rate ratios resulting from between-study variability in rates for both control and intervention groups. Our results show that as the sparseness of the data increases, the standard method of pooling data increased the bias of parameter estimates and reduced the coverage of the confidence intervals for the corresponding population parameter below the nominal value of 95%. Estimates from fixed-effects Poisson regression also display bias and poor coverage, due to the presence of heterogeneity which was not accounted for in the analysis. Effect size estimates from mixed-effects methods were unaffected by the sparseness of the data or the magnitude of the heterogeneity. These findings have important implications for studies in public health and other disciplines that seek to evaluate the impact of rare events (e.g., suicide, HIV). We conclude with recommendations for undertaking meta-analyses of sparse rate data.

A Semiparametric Mixture Method For Local False Discovery Rate Estimation

Woncheol Jang
Seok-Oh Jeong[1], Dongseok Choi[2], Justine Smith[2]
[1]Hankuk University of Foreign Studies
[2]Oregon Health & Science University

We propose a seminparametric mixture model to estimate local false discovery rates in multiple testing problems. The two pillars of the proposed approach are Efron's empirical null principle and log concave density estimation for alternative distribution. Compared to existing methods, our method can be easily extended to multivariate cases. Simulation results show that our method outperforms other existing methods and we illustrate its use in ophthalmologic gen expression data analysis. □

Oracally Efficient Inference For Time Varying Garch Model

Jiangyan Wang

We propose a two step estimator for the coefficients of a time varying GARCH model. After preliminary estimation of the time varying scale trend, approximations to the latent stationary GARCH series X_t are obtained, which are then used for maximum likelihood estimation for the GARCH coefficients in X_t. Under mild assumptions, we establish oracle efficiency of the proposed estimator, i.e., the GARCH coefficients of X_t are estimated asymptotically as efficient as the maximum likelihood estimator (MLE) based on the unobserved X_t. Simulation studies corroborate the asymptotic theory.

Confirmatory Vs. Exploratory Methods For The Estimation Of Complicated Factor Analysis Models

Walt Davis

Factor analysis models are generally estimated either based on a fairly large set of *a priori* restrictions on the model parameters (a "confirmatory" approach) or based on a largely unrestricted factor rotation (an "exploratory" approach). Because of these quite different model structures and the difficulty of identifying highly complicated confirmatory models, comparisons of these two approaches have been limited.

This presentation will briefly summarize a new approach to the identification of complicated factor analysis models under minimal a priori assumptions, allowing for the estimation of confirmatory models with an equivalent overall fit to any exploratory factor rotation. This equivalence of overall fit allows for a proper comparison of the approaches in terms of bias and efficiency in the estimation of the model parameters for both sparse and saturated models. The results show that a confirmatory approach is sensitive to model misspecification while an exploratory approach is highly sensitive to the choice of rotation and, in some cases, produces high Type I and Type II error rates.

Multiple Break-points Detection In Biological Sequences Via The Cross-entropy Method

Madawa Priyadarshana
Georgy Sofronov[1]
[1]Department Of Statistics, Macquarie University

Detection and characterization of genomic structural variations are essential in identifying disease causing genes that have functional importance in exemplifying genome wide complex diseases. These multifarious genetic variations in the human genome have been identified as a pivotal driving force behind tumour development and its progression. Copy number variation (CNV) is one of the major and common types of structural variation in the human genome that has been discussed broadly in different perspectives. For human autosomes the normal copy number is two, whereas at the sites of oncogenes it increases and at the tumour suppressor genes it decreases. CNV is defined as a DNA segment that is 1kb or larger and present at variable copy number in comparison with a reference genome. Generally CNVs can be identified by analysing the data obtained through microarray-based technologies as well as next-generation sequencing technologies. We consider CNV detection as a multiple break-point problem and propose a framework based on a variant of the popular Cross-Entropy method, which is a model based stochastic optimization technique as an exact search method to estimate both the number as well the locations of the break-points in biological sequences of continuous and discrete measurements. The proposed framework uses a mean and variance break-point detection methodology in order to accurately account for the random noise inherited in data preparation stages of different sequencing technologies. We model the continuous scale log-ratio data obtained through array comparative genomic hybridization (aCGH) technique and DNA read count data obtained through next generation sequencing platforms. The proposed framework is compared with publicly available methods using both artificially generated data and real data to illustrate the usefulness of the methodology. Results show that the proposed procedure is an effective approach of estimating number as well as the locations of break-points with high level of precision.

Distributed Statistical Inference In High-dimensional Settings

Martin Wainwright

In the modern era of "big data", many data sets are so large that they cannot be stored on a single computer. In such settings, it is necessary to develop distributed methods for performing statistical inference, in which each machine is permitted only to manipulate a subset of the full data set, and convey summary messages to other machines. What is the price of this decentralization from the statistical perspective? When can a distributed method perform yield estimates that are as good as any centralized procedure? We describe some recent results on such questions, including discussion of non-parametric regression, as well as techniques for obtaining minimax lower bounds in the distributed setting.

Homogeneity Analysis In Rainfall Data In Peninsular Malaysia

Wan Zawiah Wan Zin
Abdul Aziz Jemain[1], Marina Zahari[1]
[1]Universiti Kebangsaan Malaysia

The inhomogenities rainfall data in Peninsular Malaysia were analyzed by using daily rainfall data for 75 stations from 1975 to 2007. Three indicators were selected for this study are the annual rainfall amount, annual number of wet days and annual number of dry days. In this study, the extent of 1 mm threshold used to define the wet days. Descriptive analysis was done by finding the mean and the annual standard deviation of all indicators. The results are mapped to see the rainfall distribution and related reasons of it. Four homogeneity tests namely standard normal homogeneity test, Buishand range test, Pettitt test and Von Neumann ratio test were analyzed on all three indicators. Further results of these homogeneity tests be categorized into class 1, class 2 and class 3 according to predetermined criteria. Results of analysis found that 13.3% of the stations is not homogeneous based on rainfall amount data while for the number of wet days and dry days are 21.3% and 20.0%. Indicator number of wet days can be seen more sensitive in detecting inhomogenity of rainfall data while rainfall amount indicator were better to show the quality of data. Much more break focused on the middle of year 1975-2007. Inhomogenity also applies to stations that are near or adjacent to the same location. Industrial areas, development and economic activity are seen to affect the level of homogeneity of rainfall data compared to the less exposure and less developed areas. The results indicate that changes in environmental factors influence the level of homogeneity in the station data.

Inference For Arma Models With Unknown-form And Heavy-tailed Arch-type Noises

Shiqing Ling

ASC/IMS 2014 Conference

Abstract

Inference for ARMA Models with Unknown-Form and Heavy-tailed ARCH-type Noises

ShiqingLing
Department of Mathematics, Hong Kong University of Science and Technology, Hong Kong

Abstract:

Recent research reveals that the traditional inference is invalid for the ARMA model with unknown- form and heavy-tailed ARCH-type noises. This paper is to develop a systematic procedure for statistical inference of this model. We first study the Hill's estimator for the tail index and show that this estimator is consistent and asymptotic normal. We then investigates the least absolute deviation (LAD) estimator and the self-weighted LAD estimator for the model. Both estimators are shown to be strongly consistent and asymptotically normal when the noise has a finite variance and infinite variance, respectively. The random weighting approach is proposed for statistical inference under this nonstandard case. We further develop a score-based good-ness-fit test for model adequacy and a Wald test for structural change of models. The limiting distributions of both test statistics are obtained. Simulation study is carried out to access the performance of our procedure and a real example is given to illustrate our procedure.

Sensitivity Analysis Within The Multiple Imputation Framework To Assess Departures From The Missing Data Assumption

Panteha Hayati Rezvan
Lyle C Gurrin[1], Julie A Simpson[1]
[1]Melbourne School of Population and Global Health, University of Melbourne

Multiple imputation (MI) is a flexible approach for handling missing data. MI assumes that the data are 'Missing At Random' (MAR), where the probability of missingness does not depend on the missing values after conditioning on the observed data. This assumption is unlikely to be true in practice since it's often plausible that differences in the data distribution between individuals missing items and those with complete data can't be explained by the observed data alone, in which case the data are 'Missing Not At Random' (MNAR). Distinguishing between MAR and MNAR based on the analysis of data is possible only in simulated examples where the values of the missing observations are known.

A number of approaches have been proposed in the statistical literature to investigate the sensitivity of the MI results to departures from the MAR assumption. We performed a simulation study to evaluate two types of sensitivity analyses: *Re-weighting* and *Pattern-Mixture model* approach. We generated 1000 datasets of size 100 from a bivariate normal distribution, and assigned 50% of the observations of the outcome variable to missing, under a MNAR mechanism. Estimates of the marginal mean of the outcome and the association between the exposure and outcome, derived from (i) complete case analysis; (ii) MI_{MAR} (under MAR); and (iii) MI_{MNAR} (under MNAR using re-weighting or pattern-mixture model) were compared to the true values used to simulate the data.

The preliminary findings show that, as expected, the complete case and MI_{MAR} estimates are biased. The MI_{MNAR} estimates obtained from the re-weighting method were highly unstable across varying numbers of imputations, and the distribution of the weights tended to be dominated by a few large values. Using the pattern-mixture method, the MI_{MNAR} estimates do not substantially change over the number of imputations and are close to the true parameter values.

A Framework For Targeted Follow-up In The Australian Bureau Of Statistics

Philip Bell
Damian Garrard[1]
[1]Australian Bureau of Statistics

In a statistical agency such as the Australian Bureau of Statistics, a major cost in the collection process is the follow-up required to obtain cooperation from data providers. In household surveys this can involve repeated face-to-face visits to dwellings in an attempt to make contact with the residents, or multiple phone calls in cases where a phone number has been obtained. In business surveys there can also be repeated attempts to contact a business by phone to elicit response. These follow-up processes are resource-intensive, and it is important that follow-up is targeted appropriately to cases where it is likely to have the most benefit.

A responsive design is a plan for follow-up that allows for different follow-up strategies for different types of units selected in the collection. It can include plans for monitoring the sample as follow-up proceeds and changing strategies used for particular groups that would otherwise be over or under-represented among the responding sample. The objective of responsive design is to obtain a representative responding sample while controlling costs and quality of survey outcomes.

This talk presents a framework for applying a responsive design approach to follow-up of an arbitrary collection. A case study is presented of the application of the framework to a large household survey, the National Health Survey of 2007 (NHS 2007). The case study shows that savings can be made by applying a less intensive follow-up strategy in geographic areas of types that are well-represented among the responding sample, with minimal loss in estimate quality as measured against the actual strategy applied in NHS 2007.

Bayesian Estimation Of Low-rank Matrices

Pierre Alquier

The problem of low-rank matrix estimation recently received a lot of attention due to challenging applications: recommender systems, quantum statistics, cointegration... A lot of work has been done on rank-penalized methods and convex relaxation, both on the theoretical and applied sides. However, only a few papers considered Bayesian estimation.

The objective of this work is twofold. First, I will review the different type of priors considered on (approximately) low-rank matrices. I will also prove that the obtained Bayesian estimators, under suitable assumptions, enjoy the same optimality properties as the ones based on penalization (in the minimax sense).

Minimal Thinness With Respect To Symmetric Levy Processes'

Panki Kim

Minimal thinness is a notion that describes the smallness of a set at a boundary point depending on each Levy process on an open subset. In this talk, we discuss how to test whether given set is minimally thin or not both at finite point and infinite point. We show that the same test for the minimal thinness is valid for all processes in the considered class.

An Sis Epidemic In Large Population With Individual Variation

Philip Pollett
Ross McVinish[1]
[1]The University of Queensland

We consider a model similar to the stochastic SIS (Susceptible-Infectious-Susceptible) model in continuous time, which accounts for variation amongst individuals. We prove that, as the population size grows, the process can be approximated by a deterministic process. The equilibrium points of the limiting process are identified and their stability is examined.

Latent Variable Graphical Model Selection Via Convex Optimization

Venkat Chandrasekaran
Michael Jordan[1]
[1]University of California at Berkeley

Modern massive datasets create a fundamental problem at the intersection of the computational and statistical sciences: how to provide guarantees on the quality of statistical inference given bounds on computational resources such as time or space. Our approach to this problem is to define a notion of "algorithmic weakening," in which a hierarchy of algorithms is ordered by both computational efficiency and statistical efficiency, allowing the growing strength of the data at scale to be traded off against the need for sophisticated processing. We illustrate this approach in the setting of denoising problems, using convex relaxation as the core inferential tool. Hierarchies of convex relaxations have been widely used in theoretical computer science to yield tractable approximation algorithms to many computationally intractable tasks. In this talk we show how to endow such hierarchies with a statistical characterization and thereby obtain concrete tradeoffs relating algorithmic runtime to amount of data.

Renewable Energy And Energy Efficiency Practices: Evidence From 2008-09 Energy Water Environment Survey (ewes)

Maria Balogh
Kay Cao[1]
[1]Australian Bureau Of Statistics

The objective of this report is to examine the relationship between energy demand, Energy Management (EM) and Renewable Energy (RE) practices using the 2008-09 EWES survey. The modelling techniques utilised in this paper include log linear, multivariate, and logistic regression methods. The log linear model is motivated by Cao et al. (2012) in which an economic model was developed to estimate energy demand in the survey gap years. When including RE and EE measures into the model; estimation results suggest a significant positive correlation between energy demand and energy saving initiatives (i.e.: improving EM practices and/or investing in RE operating systems). This positive correlation may indicate the presence of the rebound effect which was found in numerous previous studies. Due to the rebound effect, improvements in energy efficiency lead to increases – rather than decreases - in the overall demand for energy. Theory suggests that there is a greater need for energy intensive firms to uptake energy saving activities. Reasons for this include preserving long term price competitiveness, complying with regulation, gaining better reputation, and being in a better position to finance innovations. Contrarily to expectations, this study found that firms in the Manufacturing and Transport sectors were not proactive in operating Renewable Energy (RE) systems, whilst a high portion of firms in the Education and Training sectors were found to have invested in energy innovations, especially in solar panels. Logistic regression results suggest large firms are more likely and medium firms are less likely to invest in RE systems compared with small firms. For large firms this can be explained by greater savings brought about by reduced cost when economies of scale are higher; small business owners on the other hand are often found to be environmental entrepreneurs driven by personal values.

Minimising Sample Overlap In Snz Household Surveys

Vic Duoba
Terry Moore[1]
[1]Statistics NZ

For very practical and obvious reasons, statistical agencies often practice some form of control over the overlap of surveys. Usually the overlap is intended to be zero, or at least minimised, using acceptable statistical techniques.

Statistics NZ has until recently controlled for overlap for household surveys via synchronised sampling using a PRN (Permanent Random Number) technique. Each frame member (primary sampling unit – PSU) is allocated a uniform random number and then the frame is sorted by PRN within strata. For any survey we allocate a block of contiguous PRN's per stratum. Selected PSU's are then not available for selection in subsequent surveys during a quarantine period.

Although this method has worked well in the past, it does have some operational deficiencies. First, all surveys must use the same stratification, and PSU's must have equal inclusion probabilities within each stratum. This means that it is not possible to use options such as probability proportional to size (PPS) sampling. Also, we cannot guarantee that SNZ surveys do not overlap with those for other government departments which do not subscribe to a common PRN method because they want their own stratification and/or PPS. However, by moving to a more flexible and widely-used overlap minimisation methodology we raise the likelihood of exhausting the frame of PSU's. If that should occur, we would like the overlap to reflect some defined preferences.

SNZ investigated the conditional sampling methods of Chowdhuri et al. (2000), and Bell (2011), then fused the main ideas, and extended the system to deal with frame reformation (changing the PSU boundaries) as in Lu (2012). The operationalised result is a SAS EG application.

Triple-goal Estimation Using The U.s. Current Population Survey Data

Parthasarathi Lahiri
Daniel Bonnery[1], Yang Cheng[2]
[1]University of Maryland
[2]The U.S. Census Bureau

In this talk, we introduce a triple-goal small area estimation methodology, motivated by the well-known paper by Shen and Louis (1998), for simultaneous estimation of small area means using complex survey data. The main goal of the proposed methodology is to produce a set of small area estimates that are good in simultaneously meeting three different goals of producing estimates for individual small area means, histogram of true small area means and ranking of the small areas by true small area means. Using a Monte Carlo simulation experiment, we compare our proposed simultaneous estimation of small area means with those obtained by the posterior means and constrained empirical/hierarchical estimators, developed earlier by Louis (1984), Lahiri (1992) and Ghosh (1992). We implement our methodology using Monte Carlo Markov Chain (MCMC) and demonstrate the impact of our proposed method in estimating unemployment rates using Current Population Survey and administrative data.

Statistics And The Environment: What Is On The Horizon?

Marian Scott
Campbell Gemmell[1]
[1]South Australia Environment Protection Authority

Much environmental regulation has been "silo-based", focussing on single issue, environmental media or legal regime components. In recent years, more integrated and now more harms-based thinking is emerging. State of the environment reporting is more data rich and crowd-sourced and cross-cutting issues are getting more attention, e.g. the European Water Framework Directive where the classification of water body status includes chemistry, morphology and biology. Global assessments such as the Millennium Ecosystem Assessment described a more multi-dimensional, holistic approach and driven by this, official environmental statistics have rolled out an extensive array of indicators and indices as communication and visualization tools.

The ICSU (2010) identified 5 grand earth system science challenges, the first two being forecasting and observing, which are strongly statistical in nature. In the former, the drive is to improve the functionality of forecasting, with a focus on the consequences for people, and in the latter, to improve our earth observation systems. New sensor technology is generating increasingly large volumes of earth observation data and challenging more traditional statistical monitoring strategies in space and time, to incorporate mobile and adaptive virtual sensors, thus enabling "citizen science" as part of a new environmental monitoring framework.

This evolution in environmental regulation, operationalisation of the ecosystem model and reductions in public budgets for monitoring and therefore the increased need for low cost/high value and low maintenance data sourcing and the increasing availability of high volume data streams mean that data linkage across environmental, social and political domains is essential to deliver a more holistic ecosystem view. Integrated models are becoming key tools linking humans, societal systems and the ecosystems they depend on (Royal Society, 2013) but present challenges spanning the statistical dimensions of data acquisition, modeling, inference and uncertainty. This presentation will consider a number of case studies illustrating these points.

Adjusted Density Estimates And Approximants

Serge Provost

Improved density estimates and approximations are obtained by means of moment-based polynomial adjustments applied to the widely used saddlepoint approximation.The support of an initial approximation is determined from the Lugannani-Rice formula. Approximate percentiles as evaluated from the original saddlepoint formula and its adjusted counterpart are compared numerically and graphically to the exact values in several illustrative examples. Known density functions shall also be utilized as initial approximations and smoothed averaged shifted histograms shall be introduced as alternative density estimates. The bivariate case is addressed by applying a polynomial adjustment involving both variables to the product of the approximated marginal densities of the standardized variables. Furthermore, extensions to the context of density estimation are formulated and applied to several univariate and bivariate data sets. In this instance, the sample moments and empirical cumulant-generating functions are utilized in lieu of their theoretical analogues. Interestingly, the proposed methodology for approximating bivariate distributions not only yields density estimates whose functional forms readily lend themselves to algebraic manipulations, but also gives rise to copula density functions that prove much more flexible than the conventional functional type.

Efficiency Transfer For Regression Models With Missing Responses

Ursula U Muller

An accepted way of analyzing missing data is by imputing the missing values. This often gives better results than a "complete case analysis", which is the fastest and simplest method of dealing with missing data since it uses only cases that are completely observed. Although this may seem to be a wasteful approach, there are in fact many situations where a complete case analysis turns out to be (asymptotically) optimal and should therefore be used.

In this talk I focus on i.i.d. data (X,Y), where the response Y is missing at random and where the covariate vector X is always observed. I demonstrate that general functionals of the conditional distribution of Y given X can be estimated efficiently by a complete case version of an efficient estimator. This is a very general and useful result since it tells us that in such situations we can simply omit incomplete cases and work with some familiar efficient estimator without losing efficiency.

This result applies to homoscedastic and heteroscedastic regression, with the conditional expectation of Y given X being modeled by a general semiparametric regression function that involves a finite and an infinite-dimensional parameter. This includes the fundamental linear and nonparametric regression model, but also more complex models, e.g. partially linear additive regression and the single-index random coefficient model. I will discuss estimation of various functionals of the conditional distribution, e.g. of regression parameters and of the error distribution.

This talk is based on joint work with Anton Schick.

Estimation Of Extreme Quantiles For Functions Of Dependent Random Variables

Qiwei Yao
Jinguo Gong[1], Yadong Li[2], Liang Peng[3]
[1]Esouthwestern University of Finance and Economics
[2]Barclays
[3]Georgia Institute of Technology

We propose a semiparametric method for estimating extreme quantiles for functions of several dependent variables. The method is designed for the extreme cases such that the corresponding empirical quantiles are not observable with given sample sizes. Both theoretical results and numerical illustration will be presented.

Flexible Error Distributions To Model Duration Data

Rasika Yatigammana
Richard Gerlach[1]
[1]Supervisor, University of Sydney

Some financial time series exhibit changes in its properties over different ranges of values within its support. In the case of trade duration data, it is often due to the impact of various economic and other related information. Empirical evidence suggests that duration data generally has a unimodal distribution. It has a very long right tail and the majority of transactions have durations close to zero. Frequently, the conditional distribution has been modelled by employing Exponential and Weibull distributions while Log-Normal, Generalised Gamma and Burr have also been considered. However, these distributions have been rejected due to lack of goodness of fit to real data. Ability to capture some of the features, especially the long right tail may have significance for forecasting and trading strategies. The literature has shown that distributions with either constant or strictly increasing or decreasing hazard functions do not capture the characteristics of durations that well. Hence it is proposed here to use partitioned distributions with different parameters for each section and unknown partition points for added flexibility. Combinations of different distributions such as Exponential and Weibull are considered in the context of Autoregressive Conditional Duration (ACD) models. A Bayesian approach is used in the estimation of distribution parameters employing an independent Gaussian proposal within a Markov chain Monte Carlo framework. After a simulation study on multiple series, the estimation is carried out on real data employing Log predictive likelihood ratios for model selection.

Assessment Of Biosimilarity Based On A Tolerance Interval Approach

Hsiao-Hui Tsou
Chin-Fu Hsiao[1], Yi-Hsuan Lai[1]
[1]National Health Research Institutes

The development of follow-on biologics products has received much attention from both sponsors and regulatory authorities while more biologic innovator products are going to lost patent protection in the next few years. Unlike the chemically synthesized drugs, the development of biologic products is much different and more complicated because of the fundamental differences in functional structure and manufacturing process. The European Medicines Agency (EMEA) of the European Union (EU) has published a guideline on similar biological medicinal products for approval of these products in 2005. On February 9, 2012, the US Food and Drug Administration (FDA) issued three draft guidance documents on biosimilar product development to assist industry in developing such products in the United States. In both guidance, however, no specific statistical methods for assessment of biosimilarity were mentioned. As indicated by Chow et al. (2010), current regulation for assessment of bioequivalence may be too loose to be applied for assessment of biosimilarity. Other statistical methodologies for evaluation of biosimilarity from different approaches are recommended. In this study, we focused on the evaluation of the consistency of the treatment effects from the reference drug and the biosimilar. We adopted two confidence interval approaches for the assessment of the consistency and illustrated the application by examples. We also proposed method for sample size determination to ensure enough probability (eg., 80%) of consistency between the biosimilar and the innovator biologic based on the two confidence interval approaches. Numerical examples were given to evaluate the performance of the proposed method.

Key Words: Biosimilars; innovator biologic; consistency.

Bootstrapping Two-phase Sampling With Applications To General Semiparametric Models

Takumi Saegusa

We develop a bootstrap procedure for two-phase stratified sampling without replacement. We establish the weak convergence of our bootstrap Inverse Probability Weighted (IPW) empirical processes with several variants of calibration. Difficulties of this problem are that the original IPW empirical processes involve dependent observations due to sampling without replacement and that its limiting processes are linear combinations of independent Brownian bridge processes. Our proof adopts the conditional argument at the different phases by exploiting the structure of our bootstrap weights as the product of two weights corresponding to variations from each phase. We apply our bootstrap to weighted likelihood estimation and establish two Z-theorems for a general semiparametric model where a nuisance parameter is estimable either at a regular or a non-regular rate.

A Semiparametric Locally Stationary Arch Model

Valentin Patilea[1]
Lionel Truquet
[1]CREST-Ensai

We consider a ARCH(p) model with a time-varying constant and fixed, but unknown, coefficients of the lag variables. This model is a compromise between the stationary ARCH model of Engle and the time-varying ARCH models introduced by Dahlhaus and Subba Rao (2006). The time-varying constant allows capturing volatility non stationarity. Meanwhile, assuming constant coefficients for the lag variables could be appealing for ARCH models users.

The time-varying constant is estimated using a kernel estimator, while the lag variables coefficients are estimated using weighted marginal (unconditional) moment conditions. Our procedure could be interpreted as an extension to nonlinear time-series framework of the classical least-squares approach for semiparametric partially linear models as considered by Speckman (1988) and Robinson (1988). We derive the parametric rate of convergence and the asymptotic normality for the lag variables coefficients estimates when the weights belong to a class of functions depending on the past. Moreover, for Gaussian inputs, we construct an asymptotically efficient estimator in the semi-parametric framework. We also prove the uniform convergence of the kernel estimator of the time-varying constant. The results are derived under minimal moment conditions on the non stationary process, in particular without additional restrictions on the lag variable coefficients with respect to the classical ARCH(p) modeling. The problem of testing whether the coefficients of the lag variables are constant or not is also investigated. Simulation and real data applications illustrate the new modeling approach.

Statistical And Computational Trade-offs In Estimation Of Sparse Principal Components

Richard Samworth
Tengyao Wang[1], Quentin Berthet[2]
[1]Department of Operations Research and Financial Engineering, Princeton University
[2]Princeton University

Recent months have seen exciting advances in understanding the limitations in terms of statistical efficiency of (randomised) polynomial time algorithms in high-dimensional problems. While previous work has focused on hypothesis testing (detection) problems, we show that similar phenomena may occur in estimation problems, in particular for the problem of estimating a sparse principal component.

The Generalised Likelihood Ratio Test For Sparse Normal Mixtures

Thomas Porter
Michael Stewart[1]
[1]The University Of Sydney

Sparse Normal mixtures have many applications for detecting, estimating, and classifying signals embedded in noise. The Neyman-Pearson Likelihood Ratio Test (LRT) would be the optimal test to use. However, it requires a full specification of the mixture, rendering it unusable for practical situations.
Instead, we can consider an extension of the LRT, the Generalised Likelihood Ratio Test (GLRT), that tests against the maximum likelihood estimate under the alternative model. The asymptotic behaviour of the GLRT is only partially understood, and much more needs to be done to enable a proper (local power) comparison with non-parametric methods such as the Donoho-Jin Higher Criticism test.

This talk will summarise current developments in detection of both homogeneous and heterogeneous Sparse Normal Mixtures, and demonstrate that the GLRT attains the LRT optimal detection boundary, a lower-order local power property.

Combining Isotonic Regression And The Em Algorithm To Predict Genetic Risk Under A Monotonicity Constraint And Unknown Genotypes

Tanya Garcia
Jin Qin[1], Yanyuan Ma[2], Ming-Xin Tang[3], Karen Marder[3], Yuanjia Wang[3]
[1]National Institute of Allergy and Infectious Diseases
[2]Texas A&M University
[3]Columbia University

In certain genetic or nutritional epidemiology studies, estimating the cumulative distribution of the age-at-onset of a disease can be characterized as a mixture model with known mixing proportions. For example, research interest may lie in estimating the cumulative risk of a disease for individuals with and without a rare deleterious mutation. These estimates provide crucial information to assist clinicians, genetic counselors and subjects carrying a mutation in making important decisions such as mastectomy. Estimating cumulative risk can be complicated, however, because the genetic mutation status in many family members of study participants may be unknown and onset ages are subject to right censoring. Earlier methods provide estimation at individual time points rather than over a range of time points, and are not guaranteed to be monotonic, nor non-negative. Here we develop a novel method that combines Expectation-Maximization and isotonic regression to estimate the cumulative risk across the entire support. Our estimator is monotonic, satisfies self-consistent estimating equations, and has high power in detecting differences between the cumulative risks of different populations. We perform extensive simulation studies to compare proposed estimator with the existing ones. Application to a Parkinson's disease (PD) study provides the age-at-onset distribution of PD in PARK2 mutation carriers and noncarriers, and reveals a significant difference between the distribution in compound heterozygous carriers compared to non-carriers, but not between heterozygous carriers and non-carriers.

Comparison Of Methods For Non-response Adjustment Of Longitudinal Weights

Benedict Cusack
Ryan Defina[1]
[1]Australian Bureau of Statistics

This paper investigates the weighting methods applied to the Longitudinal Survey of Australian Children (LSAC). Various approaches are appraised and tested with regards to their robustness, associated statistical variability and general 'fit for purpose' qualities. LSAC implements a two-step weighting approach, first using logistic regression to adjust for non-response, followed by a 'standard' generalised regression step. Analysis in this talk identifies drawbacks of the current approach and makes recommendations for improvements, along with future directions for research. Particular focus is placed on quantifying the understatement in variance currently being quoted by data users, and the implications on meaningful longitudinal inference.

Keywords: longitudinal surveys, weighting, regression, modelling error, official statistics

Branching Interlacements

Balazs Rath
Omer Angel[1], Qingsan Zhu[1]
[1]University of British Columbia

We consider critical branching random walk (BRW) with geometric offspring distribution and uniform starting point on the d-dimensional torus of side length n, and condition the total number of offspring to be equal to the integer part of the volume of the torus multiplied with a fixed parameter u.

We look at the limit of the law of the trace of this BRW as n goes to infinity for some fixed value of u and d (the latter has to be greater than equal to 5), and find that it is a random subset of the d-dimensional lattice which can be constructed as the trace of a Poisson point process on the space of infinite trees embedded in the lattice. Our construction relies on the notion of contour process of a plane tree. Inspired by similar results about random interlacements, we study the connectivity properties of this random subset of the lattice (the branching interlacement at level u) and its complement. The talk is based on joint work in progress with Omer Angel and Qingsan Zhu.

Variational Inference For Heteroscedastic Nonparametric Regression

Marianne Menictas
Matt Wand[1]
[1]School of Mathematical Sciences, University of Technology

A standard assumption in regression analysis is homogeneity of the error variances. In many applications this assumption does not hold and the variance is a function of the regressors. Standard Markov chain Monte Carlo (MCMC) methods can be used to perform Bayesian fitting and inference, but are computationally costly. This impedes MCMC-based fitting, especially for high volume/velocity data.

We develop a fast deterministic approach to simultaneous nonparametric estimation of the mean and variance functions based on mean field variational Bayes (MFVB). MFVB produces approximate inference, rather than 'exact' inference produced by MCMC, but we show that it achieves good to excellent accuracy in this context. We also describe extensions such as a bivariate predictor case and real-time processing.

Application Of Multivariate Scale Mixture Models

Thanakorn Nitithumbundit
Jennifer Chan[1]
[1]University of Sydney

Abstract:
Mean Variance Mixture models contain an extensive number of common distributions, and are widely applied to analyze the dynamics in financial stock market. We demonstrate how these models can simplify and improve the efficiency of models implementation especially under the Bayesian approach. In addition, these multivariate specifications allow us to investigate the dependence structure of these distributions. Discussion of these different mixture models and their performance with simulated and real data are considered.

Impulse Controls: Explicit Solutions And Regularity Properties

Xin Guo

One of Larry's favorite research topics is to find explicit solutions for optimal stopping or control problems. This talk shows how explicit solutions for a class of impulse control problems provide critical insight for characterizing the regularity properties of their corresponding value functions.

An Efficient Algorithm For Detecting Change Boundaries In Random Fields With Heavy-tailed Distributions

Tsung-Lin Cheng

In this talk, we propose a computationally efficient algorithm to detect the change curves of random fields with distributional changes outside some unknown curve (closed or not closed).

When the change curve is star-shaped (e.g. a circle or an elliptic), some well-known methods which deal with random fields in Cartesian coordinate cannot be directly applied to solve this problem. We consider a polar-coordinated model to overcome the difficulties. The simulation study shows that our method works well especially for heavy-tailed distributional models.

Some Results Useful For Analysis Of Survey Data

Stephen Haslett

Abstract:

When regression-type models are fitted to survey data, and inverse selection probabilities are used as weights, full efficiency of the estimated parameters requires assuming that population errors (or rather those of them that are used in the survey based model) are uncorrelated. When they are not, this assumption can lead to estimates that remain essentially unbiased but which may not be fully efficient. Methods of adjusting for non-zero correlation in the population (which occurs, for example, when the population itself is clustered) will be discussed, along with their advantages and shortcomings. Maintaining full efficiency of parameter estimation in linear models by using aggregation to lower computational burden will also be considered, along with its application to analysis of survey data where model errors are correlated. A key idea used, both for analysing data with correlated model errors and for aggregation, is model equivalence, i.e. models that differ in form but which have the same parameter estimates. The results given are extendable in principle to linear mixed models and to generalized linear models.

Key words: Aggregation, bias, correlated model errors, superpopulation models, general linear models, generalized linear models, joint selection probabilities, model equivalence, regression, selection probabilities, survey data.

A Performance Comparison Of Three Phase I Nonparametric Control Charts

Marien Graham
Margarethe Coelho[1], Subhabrata Chakraborti[2]
[1]University of Pretoria
[2]University of Alabama

Phase I analysis is a vital part of an overall statistical process control (SPC) and monitoring regime and control charts play an important role in this analysis. Several Shewhart-type Phase I control charts for monitoring the location of a process have been proposed in the literature. Three of the popular ones are examined and compared in this study. These are the parametric Xbar-chart, based on the normality assumption, and two nonparametric charts, one proposed by Jones-Farmer et al. (2009) based on the subgroup mean-rank and the other proposed by Graham et al. (2010) based on the subgroup precedence counts from the pooled median. It is emphasized that to date, a head to head performance comparison of the mean-rank chart and median chart is not available and it is our intent to fill this gap. An extensive simulation study was performed and it is seen that the mean-rank chart, as well as the median chart performed similarly to the parametric Xbar-chart for normally distributed data, but for heavy-tailed or skewed data they both outperformed the parametric Xbar-chart, with the mean-rank chart showing the best results overall. The results make a strong case for using nonparametric Phase I charts in practice.

A Research On The Non-parametric Evaluation Method For Genetic Differences

So Young Park
Sojeong Lee[1], Taeseon Yoon[1]
[1]Hankuk Academy of Foreign Studies

Past studies have employed parametric for arithmetic calculation of distance between proteins. However, as proteins synthesized from genes cannot reveal parameter value, parametric method is not appropriate in calculating the data. Therefore, this research estimates distance between proteins through non-parametric method by first measuring the distances, assuming similarities among them, and then determining reject through Hypothesis Test in significance level to speculate the distance between proteins. As a consequence, circumstantial modeling value appeared more distinctly and the method was more realistic for proteins without parameters compared to the parametric method.

Wavelet Methods For Estimating Erratic Regression Means In The Presence Of Measurement Error

Spiridon Penev

There is a large and extensive literature about nonparametric regression with errors in the explanatory variable. The regression function in this setting has almost exclusively been assumed to be relatively smooth. The methodology typically uses modified kernel estimators, the performance of which deteriorates when the regression function becomes more erratic.

We suggest relatively adaptive wavelet-based methods instead. The approach is non-standard due to the difficulty of estimating wavelet coefficients in the presence of measurement error. To the best of our knowledge, there are, as yet, no competing approaches to constructing wavelet estimators in nonparametric errors-in-variables regression. Our proposal is to minimize an "explained sum of squares" (ESS). Then we use matrix regularization to reduce noise.

For the implementation, we employ compactly supported wavelets from the Daubechies family. Our simulated examples demonstrate that for sufficiently smooth regression curves, conventional techniques and our wavelet method perform about equally well. However, in the presence of jump discontinuities in the regression curve or in cases of highly oscillatory signals, the wavelet approach performs noticeably better than the conventional techniques. The results, although not as impressive as the landmark performance in the errorless case, are still very good.

We also apply our method to a gross domestic product data which is known to be prone to significant measurement error.

At present, we are able to establish consistency of the proposed procedure based on the minimization of the ESS.

Deviance Information Criterion In Comparison Of Normal Mixing Models

Thomas Fung
Joanna Wang[1], Eugene Seneta[2]
[1]University of New South Wales
[2]University of Sydney

To perform model selection using the Deviance Information Criterion (DIC) with BUGS software, a focus has to be selected for each fitted model. This is particularly important if one can specify a model in various mixing representations as for the normal variance-mean mixing distribution commonly occurring in a financial context. In this presentation, we indicate which focus is appropriate in a comparison of several models in respect of goodness of fit. This talk is a continuation to another talk titled “Contaminated Variance-Mean mixing model” by the same authors in this conference, as that discussion is based on the same goodness of fit criteria for model comparison.

Current Issues In Analysis Of Weight Loss Trials: In Search Of The Perfect Diet.

Marijka Batterham
Linda Tapsell[1], Karen Charlton[1]
[1]School of Health Sciences, University of Wollongong

This presentation discusses analytical issues in research on the effectiveness of manipulating dietary composition to enhance fat metabolism and weight loss. Our studies have demonstrated short term beneficial effects of dietary protein and polyunsaturated fats on fat metabolism compared with control diets. However, the effects are not sustained in long term studies. Analytical issues which confound the longer term studies include smaller than expected between group effects and study attrition.

Small between group differences agree with recent systematic reviews questioning the effectiveness of these dietary manipulations for additional benefit over a standard energy restriction. Attrition is common in weight loss studies and this may result in substantial missing data. We have previously demonstrated that frequently used methods of accounting for study attrition such as baseline observation carried forward, last observation carried forward and complete case analysis can provide inaccurate estimates in weight loss trials. Our current research suggests maximum likelihood based approaches represent the preferred approach when attrition rates are low and that multiple imputation should be implemented when attrition rates are high and effect sizes small. The assumptions about the type of missing data, whether it is missing at random or not missing at random are important in analysis decisions. These issues will be discussed using our data and simulations.

Bayesian Semi-parametric Daily Tail-risk Forecasting Employing Intra-day Databayesian

Richard Gerlach
Cathy Chen

Bayesian methods have proven effective for estimation in distributional tails, including for financial Value at Risk and Expected Shortfall (ES) forecasting. Intra-day measures, such as realized volatility, realized range and intra-day range, have shown promise to add efficiency and accuracy to return volatility forecasts and potentially to tail risk measures as well. In this paper, the class of conditional autoregressive expectile (CARE) models is extended to directly incorporate intra-day measures as an input to forecast related tail risk measures. Adaptive Markov chain Monte Carlo sampling schemes are employed for estimation and forecasting. Models employing intra-day measures are favoured in an empirical study forecasting multiple financial return series, in particular during the recent global financial crises, relative to a raft of popular competing time series models and methods.

The 1964 Paper Of A. T. James And Its Influence On Statistics, Mathematics, Chemistry, And The Analysis Of Wireless Communication Systems

Donald Richards

Alan T. James' 1964 paper was, in part, a survey of aspects of the non-central distributions that arise in multivariate statistical analysis. At the time of its appearance, the paper may have been viewed as narrowly focused, but it is a tribute to James' vision that the paper has received thousands of citations since then and has fostered profound results in the theory and applications of statistics (distribution theory, design of experiments, analysis of variance), numerous areas of mathematics (harmonic analysis on Lie groups, random matrix theory, analytic number theory), theoretical chemistry (Gaussian macromolecules), and the design of wireless communications systems (multiple-input multiple-out, or MIMO, models for cell-phone systems). In this talk, I will survey some aspects of the influence of James' paper on each of these four fields.

Further, I will describe some results obtained in work with S. Kuriki (Institute of Statistical Mathematics, Japan), A. Takemura (University of Tokyo, Japan), and C. Siriteanu (University of Tokyo, Japan). This research is motivated by problems arising in MIMO analysis, and addresses the problem of deriving the exact distribution of the Schur complement of a non-central complex Wishart-distributed random matrix. We will show that a solution to this problem begins with an intensive study of James' seminal 1964 paper.

A New Approach To The Interpolation Of Complex Spatial Data

SHEN LIU
Vo Anh[1], James McGree[1], Erhan Kozan[1], Rodney Wolff[2]
[1]Mathematical Sciences School, Queensland University of Technology / CRC-ORE
[2]WH Bryan Mining and Geology Research Centre, the University of Queensland / CRC-ORE

Interpolation techniques for spatial data have been applied frequently to various fields of geosciences. Although most traditional methods assume that it is sufficient to use the first- and second-order statistics to characterize spatial random fields, researchers have now realized that these methods cannot always provide reliable interpolation results, since geological and environmental phenomena tend to be very complex, presenting non-Gaussian and/or non-linear features. This study concentrates on such an issue and proposes a new approach to the interpolation of complex spatial data, which does not rely on the Gaussianity and linearity assumptions, and can be applied with great flexibility. To characterize the non-Gaussian and non-linear features, suitable cross-variable higher-order spatial statistics are developed to measure the spatial relationship between the random vector at an unsampled location x0, denoted Z(x0), and the observations in its neighbourhood. On this basis, the conditional probability density function (CPDF) of Z(x0) is approximated. The introduction of the cross-variable higher-order spatial statistics noticeably improves the quality of the approximation of the CPDF since it enriches the information that can be extracted from the observed data, and this benefit is substantial when working with sparse and/or complex data. The approximated CPDF can be utilized in three different ways: a) to determine E[Z(x0)] as the interpolated value at x0; b) to obtain inferences such as prediction intervals and hypothesis tests; and c) to carry out sequential conditional simulations which are of much interest in geosciences. The proposed method is applied to a mineral deposit dataset, and the results demonstrate that it provides good approximations of the CPDFs at unsampled locations, and it outperforms widely used kriging methods as its interpolated values are closer to the real observations than those of its competitors.

Comparative Study Of Human Development Indicators: India And Australia

Tamoghna Halder
KHYATI SHARMA

India is ranked 136th in terms of Human development Index by the UNDP where Australia is ranked 2nd. Our topic of interest is to study and compare the changes in different human development indicators in the two nations and how each factor has changed over time from 1980 to present. Our indicators include life expectancy at birth, literacy rate, GNI per capita, gender inequality index and various other factors like sustainability, demography, composite indices, innovation and technology, trade, economy and income. This would enable us to understand where India is lagging behind and what immediate measures can be taken. This study of ours would involve extensive time series analysis and from there drawing valid inferences. For instance we would study the data on child mortality rate of the two nations over the years and then infer what has led to substantial decrease in this rate in Australia and where India is still struggling. Thus from this analysis of ours our aim would be to learn an appropriate development model from Australia in Indian context so that Indian ranking can be improved. This can ultimately lead to framing of policies accordingly.

Adjusted Risk Difference Estimation Using A Stable And Flexible Method For Additive Binomial Models

Mark Donoghoe
Ian Marschner[1]
[1]Macquarie University, Sydney

Risk difference is an important measure of effect size in both randomised and observational studies. The natural way to adjust risk differences for potential confounders is to use an additive binomial model, which is a binomial generalised linear model with an identity link function. However, implementations of the additive binomial model in commonly used statistical packages can fail to converge to the maximum likelihood estimate (MLE), necessitating the use of approximate methods involving misspecified or inflexible models. We propose a novel method that retains the additive binomial model but uses the multinomial-Poisson transformation to convert the problem into an equivalent additive Poisson fit. Combined with a stable method for fitting additive Poisson models, this allows reliable computation of the MLE, as well as allowing for semi-parametric monotonic regression functions. We use our method to analyse datasets from two clinical trials in acute myocardial infarction.

Fast Global Convergence Of Gradient Methods For High-dimensional Statistical Recovery

Sahand Negahban
Alekh Agarwal[1], Martin Wainwright[2]
[1]MSR
[2]University of California

Many statistical M-estimators are based on convex optimization problems formed by the combination of a data-dependent loss function with a norm-based regularizer. We analyze the convergence rates of gradient methods for solving such problems, working within a high-dimensional framework that allows the data dimension d to grow with (and possibly exceed) the sample size n. This high-dimensional structure precludes the usual global assumptions---namely, strong convexity and smoothness conditions---that underlie much of classical optimization analysis. We define appropriately restricted versions of these conditions, and show that they are satisfied with high probability for various statistical models. Under these conditions, our theory guarantees that proximal gradient descent has a globally geometric rate of convergence up to the \emph{statistical precision} of the model, meaning the typical distance between the true unknown parameter θ^* and an optimal solution $\hat{\theta}$. This result is substantially sharper than previous convergence results, which yielded sublinear convergence, or linear convergence only up to the noise level.

On The Cover Time Of A Random Walk With Reflection Barriers

May-Ru Chen
Zong-Yi Liou[1]
[1]Depart. of Appl. Math., National Sun Yat-sen University

Imagine that a particle starts from the origin of the x-axis and moves at times $t=0, 1, \ldots$ one step to the right with probability p or one step to the left with probability $q=1-p$, which is usually called a simple random walk.

For a given positive integer n, define the cover time to be the time when the number of points visited has just increased to the given number n. In this talk, we first review the cover time of a simple random walk starting from 0. Next, we consider the simple random walk with reflection barriers and then give the expression of the probability generating function of its cover time.

Identification Of Important Regressor Groups, Subgroups, And Individuals Via Regularization Methods

Samuel Mueller
Tanya P Garcia[1], Raymond J Carroll[1], Rosemary L Walzem[1]
[1]Texas A&M University

This presentation is motivated through a dietary treatment study in mice for which we measured fecal microbial diversity. The study used an obesity reversal paradigm and consisted of 30 obese, male mice equally and randomly assigned to one of three diets. For each mouse, data consisted of relative mRNA expression of CD68 in adipose, and microbial percentages from 51 microbes classified at the phylum, family, and genus levels. Strategies to use changes in microbiota composition to effect health improvements require knowing at which taxonomy level interventions should be aimed. Identifying these important levels is difficult, however, because most statistical methods only consider when the microbiota are classified at one taxonomy level, not multiple. Using L1 and L2 regularizations, we present a new variable selection method that identifies important features at multiple taxonomy levels. The regularization parameters are chosen by a new, data-adaptive, repeated cross-validation approach which performed well. In simulation studies, our method outperformed competing methods: it more often selected significant variables, and had small false discovery rates and acceptable false positive rates. Applying our method to the motivating data, we found which taxonomic levels were most altered by specific interventions or physiological status.

A Preliminary Investigation Of Survival Following Ovarian Cancer

Md Jamil Hasan Karami
Thomas Fung[1], Kehui Luo[1]
[1]Department of Statistics, Macquarie University

Ovarian cancer is one of the most malignant gynaecological cancer for women. Women with ovarian cancer generally have poor prognosis with short survival. In our study an exploratory analysis was carried out using the data routinely collected on patients with ovarian cancer, who were diagnosed and treated in a large hospital in Sydney. Survival following the diagnosis of ovarian cancer was examined in relation to several important prognostic factors, including FIGO stage, age, residual disease and histologic type, aiming to evaluate the effect size of each factor and build up a predictive model for survival. The predicted survival provides valuable information for patients and their families, and is particularly important in helping clinician and/or hospital with the management of each patient with ovarian cancer.

Zones Issues For Small Area Health Data

Sandy Burden
David Steel[1]
[1]NIASRA, University of Wollongong

Statistical analysis is often carried out using aggregated data for a population partitioned into geographical areas or zones. For a fixed number of zones, or scale, there are many different ways these zones could be formed, and the results of the analysis can vary for different sets of zones. The distribution of a statistic or parameter estimate over all the different sets of zones that could be formed is called the zoning distribution.

This talk considers how zoning distributions can be interpreted and how they may be used. Since data are usually available for only a single set of areas, a simulation approach to rezoning data at multiple higher scales is used to identify the zoning distribution for aggregated data. Moreover, by considering the effect of zoning as an additional source of error, a model for the bias and variance of parameter estimates is derived. An example of the zoning distribution for a regression-parameter estimate is presented.

Ordered Subset Algorithms For Penalized Likelihood Image Reconstructions In Medical Imaging

Jun Ma

Since the work of Hudson and Larkin (1994), ordered subset (OS) (aka block-iterative) algorithms have been widely studied and adopted in the medical imaging field, mainly due to their fast convergence rate when compared with the traditional algorithms where full data are used to update the reconstruction in each iteration. In this talk we focus on penalized likelihood based image reconstructions in medical imaging and summarize existing OS algorithms. We then develop a new non-negatively constrained OS algorithm using a specially designed multiplicative iterative algorithm. The performance of this new OS algorithm will be investigated by a simulation study.

Local Composite Likelihood For Spatial Point Processes

Adrian Baddeley

In the analysis of spatial point pattern data, an important and difficult challenge is to deal with spatial inhomogeneity of the pattern. This includes spatial variation in the abundance of points, or in the scale or spacing between points; abrupt or gradual transitions between different spatial textures; the presence of clusters or hot spots; and spatial variation in the effect of covariates. Here we propose a general approach to spatial inhomogeneity which combines the ideas of local likelihood (or `geographically weighted regression') with the composite likelihoods which are commonly used for spatial point processes. We develop local versions of Besag's pseudolikelihood for Gibbs point processes, and of Ogata and Katsura's "Palm likelihood" for Cox processes. Bandwidth selection, inference, and computational strategies are developed. There are connections with existing methods such as scan statistics, LISA (local indicators of spatial association), and point process residuals. We will end with a live software demonstration of the methodology applied to geological and ecological data.

Functional Central Limit Theorem For Heavy Tailed Stationary Infinitely Divisible Processes Generated By Conservative Flows.

Takashi Owada
Gennady Samorodnitsky[1]
[1]Cornell University

We establish a new class of functional central limit theorems for partial sum of certain symmetric stationary infinitely divisible processes with regularly varying Levy measures. The limit process is a new class of symmetric stable self-similar processes with stationary increments, that coincides on a part of its parameter space with a previously described process. The normalizing sequence and the limiting process are determined by the ergodic theoretical properties of the flow underlying the integral representation of the process. These properties can be interpreted as determining how long the memory of the stationary infinitely divisible process is. We also establish functional convergence, in a strong distributional sense, for conservative pointwise dual ergodic maps preserving an infinite measure.

Optimization In Data Analysis And Learning

Stephen Wright

Optimization techniques are proving to be vitally important in formulating and solving problems in data analysis and machine learning, and computational statistics generally. The challenges posed by applications in these areas, and by "big data," are stimulating a great deal of new and interesting research in optimization.

We will start this talk by surveying a variety of canonical problems in data analysis and machine learning, including support vector machines for regression and selection, variable selection, sparse covariance estimation, and matrix completion. We will then discuss a range of optimization tools that have been applied in one or more of these fields, describing the key properties of each approach and the features that make them particularly suitable for these applications. Our discussion will include dual formulations, accelerated gradient methods, prox-linear approaches (including mirror descent), stochastic gradient methods, coordinate descent methods, and higher-order methods.

OBLIQUELY REFLECTED BROWNIAN MOTIONS IN BOUNDED PLANAR DOMAINS

Kavita Ramanan[1]

Krzysztof Burdzy[2], Zhen-Qing Chen[2], Donald Marshall[2]

[1]Brown University

[2]University of Washington, Seattle

We provide a definition and classification of obliquely reflected Brownian motions (ORBMs) in bounded planar domains. ORBMs in smooth planar domains can be characterized in terms of their reflection vector field on the boundary. A conceptual difficulty in defining ORBMs in domains with rough boundaries is that the normal direction at a point on the boundary has no meaning in the classical sense. Instead, we use conformal maps and excursion theory to establish an alternative characterization of ORBMs in smooth domains, and show that this notion can be suitably extended to classify ORBMs in simply connected planar domains. Our analysis also includes certain discontinuous processes such as excursion reflected Brownian motions or, equivalently, Brownian motions with darning, which arise in the boundary theory of Markov processes and can be viewed as extremal elements in the family of ORBMs.

Gaussian Processes, Self-normalized Sums, Optimal Stopping, And Singular Stochastic Control

Tze Leung Lai

This talk gives a brief review of Larry Shepp's seminal contributions to four areas in probability and their impact on subsequent developments, including some of my own work.

Openness Of Individuals To Migrate And Job Mobility: An Application Of A Multiprocess Multilevel Model

Sergi Vidal
Johannes Huinink[1], Stefanie Kley[2]
[1]University Of Bremen
[2]University of Hamburg

In this article we extend the scope of the interdependence between migration and job mobility: We investigate whether an individual's openness to migrate not only increases the probability of migration but also the likelihood to conduct a job search and exhibit job mobility. Using data from a three-wave panel study, which allows the analysis of temporal links between decision-making and subsequent events regarding migration and job mobility, a joint estimation of multiple equations is performed. We show that considering migration as an option for the future, which is our indicator of individuals' openness to migrate, is positively associated with both migration and job mobility. It even increases job mobility independently of whether migration takes place or not. These findings contribute significantly to the existing body of knowledge on the interdependence of migration and job mobility. Additionally, they enhance our understanding of the mechanisms behind a common selectivity of migrants and job mobile individuals.

Latent Supervised Learning For Estimating Treatment Effect Heterogeneity

Susan Wei

Michael Kosorok[1]

[1]University of North Carolina at Chapel Hill

It is oft observed in medicine that what works for one patient may not work for another. Determining when and for whom a treatment works and does not work is of great clinical interest. We propose a methodology to estimate treatment effect heterogeneity, i.e. to ascertain for which subpopulations a treatment is effective or harmful. The model studied assumes the relationship between an outcome of interest (e.g. blood pressure, cholesterol, survival) and a set of covariates e.g. treatment, age, gender) is modified by a linear combination of a set of features (e.g. gene expression). Specifically a threshold on the linear combination divides the population into two subpopulations with different responses to treatment. Techniques from Latent Supervised Learning, a novel machine learning idea, is applied for model estimation. Consistency of the estimator is established. In simulations the proposed methodology demonstrates high classification accuracy in a wide array of settings. Three data analysis examples are presented to illustrate the efficacy and applicability of the proposed methodology.

Estimating Relative Survival Using Bayesian Flexible Parametric Models With Spatial Frailties

Susanna Cramb
Kerrie Mengersen[1], Peter Baade[2]
[1]Queensland University of Technology
[2]Cancer Council Queensland

Recommended approaches for modeling patient survival include using parametric models and incorporating frailties. Despite this, concern remains over the generally poor baseline model fit obtained under a parametric approach.

To overcome this poor fit, flexible parametric survival models were developed by Royston and Parmar (2002) and extended to relative survival by Nelson (2007). Here, the baseline hazard function is fit using cubic restricted splines on log time, which enables great flexibility in the shape.

Flexible parametric models are becoming increasingly popular, and have several advantages over more standard approaches. One advantage is the ability to model all-cause survival, cause-specific survival and relative survival within the same framework. Another is the computational efficiency, as no splitting of the time scale is required. However, we are not aware of any studies incorporating frailties into these models.

We present a Bayesian flexible parametric relative survival model applied to population-based breast and colorectal cancer data in Queensland. Models were adjusted for cancer stage at diagnosis, age group, gender, remoteness and socioeconomic status. Spatial frailties were included to allow for correlation between neighbouring small regions, as well as uncorrelated frailties. Modelling was conducted via Stata and WinBUGS software.

Incorporating frailties within a Bayesian flexible parametric survival framework to model relative survival, while novel, is easy to implement and recommended.

Spatial Regression With Covariate Measurement Error: A Semi-parametric Approach

Md Hamidul Huque
Howard Bondell[1], Raymond Carroll[2], Louise Ryan[3]
[1]North Carolina State University
[2]Texas A&M University
[3]University Of Technology, Sydney

Spatial data have become increasingly common in epidemiology and public health research due to the rapid advances in GIS (Geographic Information Systems) technology. In health research, for example, it is common for epidemiologists to incorporate geographically indexed data into their studies. In practice, however, the spatially-defined covariates are often measured with error. The classical measurement error theory is inapplicable in the context of spatial modeling because of the spatial correlation among the observations. The naïve estimator of regression coefficients are attenuated if measurement error is ignored. We proposed a semi parametric regression approach to obtain the bias corrected estimates of the regression parameter and derived the large sample properties of the estimates. We evaluate the performance of the proposed method through simulation studies and illustrate using real examples.

Key words: Measurement error, spatial correlation, penalized likelihood, splines, thin plate spline regression basis.

Computing Statistics Of A Linear Network: A Study Of Western Australia's Road Network

Suman Rakshit
Adrian Baddeley[1], Gopalan Nair[1]
[1]School of Mathematics and Statistics, The University of Western Australia

One of the basic problems in point pattern analysis on a linear network is the computation of distance based statistics of the network. Ang, Baddeley and Nair (2012) proposed a K-function for second-order analysis of point process data on a linear network. However, computation of the K-function involves calculation of shortest path distances between points and computation of circumference of points in the network. Another important statistics is circumradius of the network. General statistical software packages used for these tasks are inadequate if the network graph consists of large number of vertices and edges. We propose an efficient way of computing these statistics of the network. The results in this paper are derived after applying proposed methodologies on Western Australia's road network. The R code used for the computation will be made available to public.

Assessing The Impact Of Task-switching On Task Length In The Presence Of Length Bias

Scott Walter
William Dunsmuir[1]
[1]Department of Statistics, School of Mathematics and Statistics, UNSW.

Clinical work is characterised by frequent interjection of events that cause clinicians to switch from their primary task to deal with the incoming secondary task, before then returning to complete the primary task. This type of task-switching has been associated with several negative effects, including modification of the time taken to complete a task in the presence of one or more instances of task-switching. An increase in task length due to task-switching implies reduced efficiency, while decreased length suggests hastening of tasks to compensate for the increased workload brought by the unexpected secondary tasks, which is a potential safety issue. Tasks that are naturally longer are more likely to have one or more task-switching events. This is a manifestation of length bias. In order to assess the effect of task-switching on task length it is necessary to estimate task lengths had they not experienced any task-switching, while also accounting for length bias. We review an existing method and propose two new approaches. The three methods are shown to be equivalent under simple assumptions. Their performance under departures from these assumptions is compared via a simulation study. The methods are also applied to observational data from a hospital emergency department. Each method necessarily uses lengths for tasks unaffected by task-switching to generate estimates. Some approaches rely more heavily on numerical estimation than others, which makes sample size a relatively more important consideration in these cases. The reliance of each method on particular parametric assumptions means that their application is limited to data that satisfies the corresponding assumptions. A fully non-parametric solution would be more widely applicable to real data, however, such a solution may not be possible. Hence, checking of assumptions prior to analysis is important to select which method, if any, is the most appropriate.

Robust Estimation In Negative Binomial Regression

Stephane Heritier
Eva Cantoni[1], William Aeberhard[2]
[1]University of Geneva
[2]Macquarie University

Negative binomial (NB) regression is commonly used in multiple sclerosis trials where the primary outcome is the number of active lesions in the brain. It is also the preferred model in intervention studies aiming at reducing the number of falls in patients with degenerative disease. Difficulties arise when unexpected large counts are observed like the so-called "multiple fallers" in Parkinson's disease sufferers. Another issue is the well-documented bias to the overdispersion parameter in small samples. Motivated by these problems, we extend two approaches for building robust M-estimators developed for generalised linear models to the NB model. The first approach achieves robustness by bounding the Pearson residuals that appear in the maximum likelihood estimating equations, while the second bounds the unscaled deviance components. An auxiliary weighted maximum likelihood estimator is introduced for the overdispersion parameter. Simulations show that redescending bounding score functions yield estimates with smaller biases under contamination while keeping high efficiency at the assumed model, and this for both approaches. Exponential tilting estimators based on the previous score functions are also considered and sensibly reduce the small sample bias to the overdispersion parameter. An application to recent fall trial data will also be presented.

Testing The Equality Of The Covariance Functions Of Several Functional Populations

Jin-Ting Zhang

With modern recording equipments, functional data are collected frequently in many scientific fields.

To analyze such functional data, more and more techniques have been proposed and studied.

In this talk, we discuss an L2-norm –based global test for comparing the covariance functions of several functional populations. We show that the proposed statistic has an asymptotic distribution of chi-square-type mixtures and enjoys good asymptotic powers. Two approximate methods are proposed to approximate the underlying null distributions. The methodologies are illustrated via a real data application.

A Simulation Study To Compare Contemporary Unit Level Small Area Estimation Methods For Poverty Mapping

Sumonkanti Das

In the recent time, small area estimation (SAE) method has been used as an indirect procedure for preparing improved geographic profile of poverty indicators. During the last decade, three unit-level SAE techniques: ELL method of Elbers, Lanjouw and Lanjouw (2003) known as World Bank method, Empirical Bayes (EB) method of Molina and Rao (2010) and M-Quantile (MQ) method of Tzavidis *et al.* (2008) are being widely used to estimate the micro-level FGT poverty indicators (Foster, Greer and Thorbecke, 1984). These methods vary in terms of their underlying model assumptions specifically differences in the consideration of random effect and perform better when the real data set follow the respective underlying assumptions but no one knows the reality. In this talk, I will compare the performance of the mentioned methods in terms of poverty estimates with their accuracy measures considering different empirical situations such as no random area/cluster effect, either cluster or area or both random effects. The comparison will be based on a simulated data set which represents the picture of developing countries like Bangladesh. On the other hand, all the three methods are based on two-level nested error regression model instead of three-level model due to practical consideration. As an alternative three-level nested-error regression model is aimed to develop and compare with the existing methods in this presentation.

Key Words: Empirical Bayes method, M-Quantile Method, Small Area Estimation, Poverty Mapping, World Bank Method

Fitting And Diagnosing Generalized Linear Mixed Models Using A Partially Noncentered Parametrization

Linda Tan
David Nott[1]
[1]Department of Statistics & Applied Probability, National University of Singapore

We present a nonconjugate variational message passing algorithm for fitting generalized linear mixed models (GLMMs) under a Bayesian framework. In addition, we show that diagnostics for prior-likelihood conflict, which are useful for identifying divergent units, can be obtained from nonconjugate variational message passing automatically as an alternative to simulation-based MCMC methods. We consider a partially noncentered parametrization for the GLMM, which is able to adapt to the quantity of information in the data and determine automatically a parametrization close to optimal. Reparametrization techniques have been used to improve convergence in MCMC and EM algorithms, and we demonstrate that partial noncentering can also accelerate convergence in the context of variational Bayes and produce more accurate posterior approximations than centering or noncentering.

Bayesian Methods For Learning Relations In High Dimensional Models

Subhashis Ghosal

Sayantan Banerjee[1]

[1]North Carolina State University

In modern statistical applications, it is very common to encounter high dimensional observations. In spite of the very high complexity of the data, a key feature that allows valid statistical analysis is sparsity. Classical statistical methods of structure learning are typically based on finding sparse structures using penalization techniques, and they generally do not give assessment of uncertainties in the decision process. Bayesian methods, on the other hand, give assessment of uncertainty of each conceivable submodel arising out of all possible sparse structures. In a Gaussian model, intrinsic relations between variables are neatly summarized by the precision matrix, given by the precision matrix. Motivated by the so called graphical lasso, which is the maximum likelihood estimator subject to an L1-norm penalty, a natural prior on sparse precision matrix is given by imposing independent exponential priors on diagonal elements, mixture of point mass at zero and double exponential priors on off-diagonal elements and forcing positive definiteness condition on the resulting matrix. We derive the posterior convergence rate at a sparse true precision matrix under the Frobenius norm, and show that it agrees with the oracle convergence rate. Bayesian computation in the high dimensional situation is very challenging as traditional approaches, typically based on Markov chain Monte Carlo methods, do not scale well. We device an approximate computing technique based on Laplace approximation avoiding Markov chain Monte-Carlo methods completely. The Bayesian approach is shown to perform very well in simulated and real data in terms of accuracy and computing speed.

Minimal Spectral Representations Of Infinitely Divisible And Max-infinitely Divisible Processes

Stilian Stoev
Zakhar Kabluchko[1]
[1]Ulm University

We introduce the notion of minimality for spectral representations of sum- and max-infinitely divisible processes and prove that the minimal spectral representation on a Borel space exists and is unique. This fact is used to show that a stationary, stochastically continuous, sum- or max-i.d. random process on Rd can be generated by a measure-preserving flow on a sigma-finite Borel measure space and that this flow is unique. As a particular case, we characterize stationary, stochastically continuous, union-infinitely divisible random subsets of Rd. We introduce several new classes of max-i.d. random fields including fields of Penrose type and fields associated to Poisson line processes.

Drug-likeness In The Drug Industry: What Role Is There For Statisticians In Obtaining So-called Structural Fingerprints?

Shanjeeda Shafi
Irene Hudson[1], Sean Hudson[2]
[1]University Of Newcastle
[2]University of California San Francisco

Drug-likeness', a qualitative property of chemicals assigned by experts committee vote, is widely integrated into the early stages of lead and drug discovery. The discrimination between 'drugs' (represented by a collection of pharmaceutically relevant small molecules, some of which are marketed drugs) and 'nondrugs' (chemical reagents) is possible using different statistical tools and chemical descriptor systems. As recently reported by Hudson (2013) (ANZIAM 2013) "optimization of a large number (*N*) of variables in two different domains, the chemical and biological, is fundamental to successful drug discovery. Relevant variables concern both physico-chemical properties of ligand, such as molecular weight (MW), to more complex measures related to its bioavailability and toxicity and its affinity towards the target. In drug discovery the challenge is to identify regions of chemical space that contain biologically active compounds for given biological targets (i.e. proteases, calpains etc). Lipinski and Hopkins (2004) suggested that within the continuum of chemical space, there should be discrete regions occupied by compounds with specific affinities towards particular biological targets. The question of which variables (or coordinate systems) would facilitate such segregation, however, was not delineated in their seminal paper – is as yet not determined. How can we best, if at all identify regions of chemical and biological space with a higher probability of yielding clinical target-specific (druggable) drugs? Can a mixture of conceptual, numerical and visual representations of (drug) databases of inhibitors and commercial drugs lead to a more effective drug discovery process? We shall discuss visualization, data reductive, clustering and discrimination methods as used more and more by the drug industry. Can we estimate oral versus non-oral drug-likeness? Can we distinguish drugs from non-drugs? Can we create better version of recent drug-like filters which go beyond Lipinski's Ro5? What role is there for statisticians in obtaining so-called *structural fingerprints*?

Kernel Additive Sliced Inverse Regression

Heng Lian

In recent years, nonlinear sufficient dimension reduction (SDR) methods have gained increasing popularity. However, while semiparametric models in regression have fascinated researchers for several decades with a large amount of literature, parsimonious structured nonlinear SDR has attracted little attention so far. In this paper, extending kernel sliced inverse regression, we study additive models in the context of SDR and demonstrate its potential usefulness due to its flexibility and parsimony. Theoretically we clarify the improved convergence rate using additive structure is due to faster rate of decay of the kernel's eigenvalues. Additive structure also opens the possibility of nonparametric variable selection. This sparsification of the kernel however does not introduce additional tuning parameters, in contrast with sparse regression. Simulated and real data sets are presented to illustrate the benefits and limitations of the approach.

Using Bayesian Shrinkage Methods To Detect Genetic Associations

Evangelina Lopez De Maturana
Noelia Ibáñez-Escriche[1], Oscar González-Recio[2], Gaelle Marenne[3], Hossein Mehrban[4], Stephen Chanock[5], Mike Goddard[2], Núria Malats[3]
[1]Research Institute and Agricultural technology (IRTA). Genética i Millora Animal
[2]Biosciences Research Division, Department of Environment and Primary Industries
[3]Spanish National Cancer Research Center (CNIO)
[4]Department of Animal Science. University of Shahrekord
[5]Division of Cancer Epidemiology and Genetics, National Cancer Institute

The continuous advancement in genotyping technology has not been accompanied with the application of innovative statistical methods, as multi-marker methods, to unravel genetic associations with complex traits. Although the performance of multi-marker methods has been widely explored in a prediction context, little is known on their behaviour in the quantitative trait loci (QTL) detection framework. We try to shed light on this open question by comparing two Bayesian shrinkage methods, Bayes A (BA) and Bayesian LASSO (BL), coupled with a within Markov chain Monte Carlo (McMC) permutation technique to declare as QTL a given marker, and the standard method used in most genome wide association studies (GWAS), the single marker regression (SMR), both in simulated and real data. Simulated data were generated in the context of six scenarios differing on effect size, minor allele frequency (MAF) and linkage disequilibrium (LD) between QTLs using real genotypes of SNPs in chromosome 21 from the EPICURO/Spanish Bladder Cancer Study. Those methods were also applied to real data to detect associations with urothelial carcinoma of the bladder (UCB) risk.

We show how the genetic architecture dramatically affects the methods'behaviour in terms of power, type I error and accuracy of estimates. Markers with high MAF are better detected, mainly those with large effect. A high LD between QTLs with heterogeneous effects differently affects methods' power: it impairs QTL detection by BA, irrespectively of the effect size, although boosts that of small effects by BL and SMR.

We demonstrate the convenience of applying multi-SNP methods rather than SMR because of their over performance. Results from real data suggest novel associations with genes in

chromosome 21 not detected by SMR in previous studies. Our findings should encourage future GWAS to use multi-marker methods to detect associations with complex traits.

Transmuted Weibull Distribution For Lifetime Modeling

Muhammad Shuaib Khan
Robert King[1], Irene Hudson[1]
[1]School of Mathematical and Physical Sciences, The University of Newcastle

In this article, we present the transmuted Weibull distribution with applications to lifetime data. We investigate the potential usefulness of the transmuted Weibull distribution for modeling survival data from biomedical studies. Many other generalizations of the two-parameter Weibull distribution are compared using maximum likelihood estimation. We obtain the analytical shapes of density and reliability functions. Explicit expressions are derived for the moments, moment generating function, entropy, mean deviation and order statistics.

Keywords: Reliability functions; moment estimation; order statistics; maximum likelihood estimation.

Getting Published In The Australian And New Zealand Journal Of Statistics. An Editor Perspective.

Michael Martin

As one of the Theory and Methods editors for the Australian and New Zealand Journal of Statistics, I see a lot of submitted papers that I know from the start will struggle to get published in the Journal. Many obstacles stand in the way of successfully getting a paper published. The quality of the idea behind the work is, of course, of primary importance, but many other factors figure in the editorial decision process. In this session, I will speak about some ways in which authors can materially improve the chance that their paper will survive the review process and appear in print.

Mixture Detection For Exponential Families

Michael Stewart

The simple normal location mixture test between 1 and 2 components originally studied by Hartigan (1985) has enjoyed renewed interest in recent times for use in signal detection and multiple testing problems, thanks in part to the work of Ingster (Math. Methods Stat. 1997,2001,2002) and also that of Donoho and Jin (Ann. Stat. 2004) on their higher criticism method. In this work we extend theoretical results concerning both the (generalised) likelihood ratio test and higher criticism in the normal mixture model to general one-parameter exponential families. If time permits we shall discuss some interesting mathematical features of the general problem.

References

J. A. Hartigan. A failure of likelihood ratio asymptotics for normal mixtures.
In Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, 1985.

Why Performance Indexes Fail

Stephen Horn

Government programmes are said to be driven by evidence: evidence that spending money in a certain way is likely to bring about a certain outcome; evidence that a programme is working as intended (rather than not); evidence that a programme has worked and can thus be emulated (or not and should be withdrawn). Evidence rests on observation; an observation is transportable as data, transforming data on stochastic phenomena (and all implementation of policy is subject to laws of chance) into useful inference is the province of statistics. This characterisation has reinforced statistics as a nominal guarantor for the knowledge base of government in the same way that physiology was for practised medicine. As an orthodoxy with procrustean applications it needs to be questioned.

Statistical methods in government – whether used to craft policies, to monitor the implementation of programmes or to evaluate their success - are not value free. Methods authority rests on fragments of theory developed in diverse contexts: experimental design (plant breeding, pharmaceutics); quality control (industrial process engineering); survey quality assurance (operations research; cognitive methods); index theory (economics). The fragmentation of survey methods is an unfolding phenomenon among national statistical agencies, now challenged by revolutions in information architecture and analytics. Clouding, a byproduct of linked up administration, dilutes the reporting authority leaving evidence as an orphan. I comment on a frame of inference to match this enlarging field, using insights from engineering but staying within the discipline and drawing on Smee's upstream analogy for its role.

This endeavour does of course fail; but an examination of its failure should be useful for those concerned with a more rigorous foundation to program evaluation. I draw on cases within the social welfare field – payments system health; panel survey design and estimation; COAG agreement processes.

Influence Diagnostics In The Elliptical Linear Regression Model With Stochastic Restrictions

Shuangzhe Liu
Shi Lei[1], Victor Leiva[2], Claudia Navarro[2], Francisco Cysneiros[3]
[1]Yunnan University of Finance and Economics, China
[2]Universidad de ValparaÃso, Chile
[3]Universidade Federal de Pernambuco, Brazil

In this paper, we propose diagnostic tools for a regression model with stochastic restrictions. Specifically, we study how a minor perturbation may impact on the mixed estimation procedure of parameters in a linear regression model with errors following an elliptically contoured distribution. The normal curvatures for assessing local influence under usual perturbation schemes, including case-weight, response variable and explanatory variable perturbations, are derived. A numerical example is presented to illustrate these results.

Bayesian Spatially Varying Autoregresive Process Models For Large Space-time Data

K. Shuvo Bakar
Philip Kokic[1], Huidong Jin[1]
[1]CSIRO, Computational Informatics

Statistical modelling of large space-time data is challenging due to the computational complexity that arises from both the spatial and temporal dimensions of the data. In climate change research combining different global climate model (GCM) products with observed point level measurement data is important for reducing uncertainty in the model predictions. This is also referred to as data assimilation and downscaler techniques in the statistical literature. In this paper we develop a Bayesian spatially varying autoregressive model that has the ability to address the spatial misalignment of different data products through spatially varying parameters, and temporal dynamics using the autoregressive part of the model. To handle large data sets, a reduced rank approach is utilised in the model for both the spatially varying coefficients and for the spatio-temporal random effects. The resulting model is presented through a Bayesian hierarchical setting, and the Markov chain Monte Carlo algorithm is used to make inferences.

We apply this model to observed summer season daily maximum temperature data from 1960 to 2013, monitored at a large number of locations from a study region in the south-eastern part of Australia. The deterministic grid based model output from the NCEP reanalysis data and CSIRO CCAM regional climate model are used as covariates in the model. The proposed model generates accurate spatial prediction of aggregated maximum temperatures, as evidenced by cross-validation results. We use the model to understand the distribution and extent of extreme heat events that occurred in the study region. The trends in the heat events and their spatial patterns are also investigated through model based posterior predictive distributions.

Estimating Granulated Change By Combining Panel And Adminstrative Data

Stephen Horn
Raymond Czaplewski

Public policy analysis works with available official data, presented as data series, at root survey or administratively obtained estimates within an accounting frame or estimated from repeated population surveys or register extract.

Even as access to transactional data has been transformed by harnessing electronic flows, use of satellite imagery, research access to linked customer level records, and harmonised collections across jurisdictions, official statisticians are under pressure to detect significant turning points within response times and resolutions that cannot be handled by present estimation methods.

Kalman filters are a state space technique, first developed to correct satellite positioning in real time where direct measurement was not possible but found to have a simple Bayesian interpretation in process control and other statistical measure problems.

We describe two applications where state space methods are used to combine sources of data efficiently while respecting quality demands in advising government decision making. Specifically we phrase the measure problem as how to combine high quality, high cost unit level information obtained from a sparse sample with continuous (or quasi continuous) 'short' granular population views in an optimal manner with calculable error structure.

By way of illustration the survey data may comprise a restricted number of interviews or site monitors amassing per sample unit a high dimensional observation, repeated infrequently. The auxiliary data may be constructed from linked administrative records of the population framing the sample, or satellite images of the region of interest within which the monitoring sites are selected forming a continuous or high frequency low dimensional image.

While Kalman filters are widely used outside official statistics, the application to this problem in its broad generality would seem to be new.

Bayesian Estimation For Diagnostic Testing In Biosecurity Risk Analysis

Sandy Clarke
Stuart Jones[1]
[1]Centre of Excellence for Biosecurity Risk Analysis, University of Melbourne

Typically associated with medicine, diagnostic testing is also used routinely by quarantine officers for the detection and identification of animal and plant pathogens. The results of such tests are used to inform high-consequence decisions such as whether to restrict entry of a traded commodity. As in medicine, it is rare that a test can be considered to be a gold standard and the error rates of such a test – diagnostic sensitivity and specificity – are usually uncertain, particular for new, environmental samples. In an area of historically poor statistical protocols, test performance on samples of known status, the results of other independent tests and prior knowledge from expert judgement can be used to improve results through quantifying and reducing uncertainty.

This talk presents an extension to the approach of Joseph et al. (1995) for Bayesian estimation in the absence of a gold standard test, which allows for the use of incomplete test data. This approach is applied to a case study of myrtle rust (*Puccinia psidii* sensu lato) involving samples from potential exposure pathways into Australia. The testing was done at two independent laboratories using polymerase chain reaction assays and prior estimates for pathway prevalence were obtained by expert elicitation. The results of the Gibbs sampler show that test sensitivity and specificity, and pathogen prevalence can be estimated using the available data even where some samples have been subject to only one of two tests and these tests have demonstrated fallibility. The findings enable diagnostic testing laboratories to make use of all available results and to explicitly incorporate prior knowledge to estimate pathogen prevalence and test accuracy.

Robust Exponential Tilting Alternative To Generalized Method Of Moments Estimation

Serigne Lo

Procedures based on the Generalized Method of Moments (GMM) are basic tools in a broad range of statistical modeling. In most cases, the theory available for making inference with these procedures is based on first order asymptotic theory. It is well known that the (first order) asymptotic distribution does not provide accurate p-values and confidence intervals in moderate to small samples. Moreover, in the presence of small deviations from the assumed model, p-values and confidence intervals based on classical GMM procedures can be drastically affected (nonrobustness). Several alternative techniques have been proposed to improve the accuracy of GMM procedures. These alternatives address either the first order accuracy of the approximations (information and entropy econometrics (IEE)) or the nonrobustness (Robust GMM estimators and tests). This work proposes a new procedure which combines robustness properties and accuracy in small samples. Specifically, IEE techniques are combined with robust methods obtained by bounding the original orthogonality function. This leads to new robust estimators and tests in moment condition models with excellent finite sample accuracy. Finally, the accuracy of the new statistic is illustrated with Monte Carlo simulations for three models on overidentifying moment conditions.

Mechanistic Nonlinear Mixed Effects Modelling Of Parasite Counts Measured In Severe Malaria Patients

Sophie Zaloumis
Julie Anne Simpson[1], PKPD-IVARS Study Group .
[1]Melbourne School of Population and Global Health, The University of Melbourne, M

There were an estimated 660,000 deaths from severe malaria in 2010, mostly young children living in sub-Saharan Africa. Early diagnosis and treatment with highly effective anti-malarial drugs have been shown to reduce the risk of death substantially. The World Health Organization now recommends intravenous artesunate (IV-ARS) as the first line treatment of severe malaria in adults and children. The current weight adjusted dosing regimen may be suboptimal for some study populations as it was extrapolated from studies of patients with mild malaria. Furthermore, resistance has developed to artesunate in South East Asia, and therefore, new dosing regimens need to be investigated in clinical efficacy trials. We aim to aid optimization of severe malaria treatment by developing a mechanistic model that can be used to help select dosing regimens of IV-ARS for further investigation.

The largest pooled dataset of drug concentration and parasite count data collected from severe malaria patients receiving IV-ARS has been assembled. A mechanistic model relating antimalarial drug concentration to the clearance of parasites in the body over time was developed and fitted to the pooled parasite count data using non-linear mixed-effects (NLME) modelling to allow for between patient variability in the clearance of parasites over time. Drug concentrations for each patient at a particular time point, corresponding to when a parasite count was recorded, were predicted using parameters that govern the within patient processes of drug absorption, distribution and elimination obtained from NLME modeling of the observed drug concentration data. Bayesian inference and Markov chain Monte Carlo methods were used to derive point and interval estimates for the parameters of interest and for simulation of parasite count time profiles for hypothetical patients receiving different IV-ARS dosing regimens.

Goodness Of Fit For Network Models: Dynamic Markov Bases

Sonja Petrovic
Despina Stasi[1], Elizabeth Gross[2]
[1]Illinois Institute of Technology
[2]North Carolina State University

Social networks and other large sparse date sets pose significant challenges for statistical inference, as many standard statistical methods for testing model/data fit are not applicable in such settings. Algebraic statistics offers a theoretically justified approach to goodness-of-fit testing that relies on the theory of Markov bases and is intimately connected with the geometry of the model as described by its fibers.

Current practices require the computation of the entire basis, which is infeasible in many practical settings. We present a dynamic approach to explore the fiber of a model, which bypasses this issue. Our algorithm is based on the toric geometry of hypergraphs. The running example is the p_1 model for social networks, a statistical model of random directed graphs with reciprocation.

Quantile-based Classifiers

Cinzia Viroli
Christian Hennig[1]
[1]UCL

A distance based classifier, using component-wise quantiles, is define by classifying an observation according to a sum of appropriately weighted component-wise distances of the components of the observation to the within-class quantiles. The method is inspired by the recent median-based classifiers (Hall et al., 2009) that represent a robust version of the conventional Euclidean distance-based classifiers for potentially high-dimensional data. The optimal percentage for the quantiles can be chosen by minimizing the misclassification error in the training sample.
It is shown that this is consistent, as the sample size increase, for the classification rule with asymptotically optimal quantile. Moreover, under some assumptions, as the dimensionality increases, the probability of correct classification converges to one. The role of skewness of the involved variables is discussed, which leads to an improved classifier.
The optimal quantile classifier performs very well in a comprehensive simulation study and a real data set from chemistry (classification of bioaerosols) compared to other classifiers.

Massively Parallel Inference For Massively Parallel Data

Gordon Smyth

Modern genomic technologies continue to present statisticians with large data structures that push the boundaries of statistical theory and require novel solutions. Gene expression technologies such as microarrays or RNA-seq are examples of this. These technologies generate data for thousands or millions of genomic features simultaneously. The data for each genomic feature (a gene for example) might be viewed as a classical statistical problem, typically with a very small sample size, but the data as a whole has an extremely high dimensional nature that requires new solutions. This talk will review statistical methods for differential expression analysis of RNA-seq data, with a focus on estimating biological and technical variation and on borrowing strength between genes.

On High-dimensional Robust Regression, Penalized Robust Regression And Glms

Noureddine El Karoui

I will discuss the behavior of widely used statistical methods in the high-dimensional setting where the number of observations, n, and the number of predictors, p, are both large. I will present limit theorems about the behavior of the corresponding estimators, their asymptotic risks etc...

Many surprising statistical phenomena occur: for instance, maximum likelihood methods are shown to be (grossly) inefficient, and loss functions that should be used in regression are shown to depend on the ratio p/n. This means that dimensionality should be explicitly taken into account when performing simple tasks such as regression. Interesting questions about the bootstrap also arise.

Mathematically, the tools needed mainly come from random matrix theory, measure concentration and convex analysis. The proximal mapping, widely used in optimization, plays in particular a central role in our analyses.

Random Maps And 2-dimensional Random Geometries

Gregory Miermont

A map is a gluing of a finite number of polygons, forming a connected orientable topological surface. It can be interpreted as assigning this surface a discrete geometry, and the theoretical physics literature in the 80-90's argued that random maps are an appropriate discrete model for the theory of 2-dimensional quantum gravity, which involves ill-defined integrals over all metrics on a given surface. The idea is to replace these integrals by finite sums, for instance over all triangulation of the sphere with a large number of faces, hoping that such triangulations approximate a limiting "continuum random surface".

In the recent years, much progress has been made in the mathematical understanding of the latter problem. In particular, it is now known that many natural models of random planar maps, for which the faces degrees remain small, admit a universal scaling limit, the Brownian map. Other models, favoring large faces, also admit a one-parameter family of scaling limits, called stable maps. The latter are believed to describe the asymptotic geometry of random maps carrying statistical physics models, as has now been established in some important cases (including the so-called rigid O(n) model on quadrangulations). On the other hand, there are many conjectured links between random maps and conformally invariant processes in the plane, that remain widely unexplored so far.

In this talk, I will review some of the recent progress on these topics, partly based on joint work with Jérémie Bettinelli, Nicolas Curien and Jean-François Le Gall.

Limits Of Minimum Spanning Trees

Louigi Addario-Berry

Give the edges of the complete graph K_n iid continuous edge weights, and let M_n be the resulting minimum spanning tree. We discuss existence and universality, and dimensional bounds, for local and scaling limits of M_n.

Impact Of Regularization On Spectral Clustering

Bin Yu
Antony Joseph[1]
[1]University of California Berkeley, CA USA

The performance of spectral clustering is considerably improved via regularization, as demonstrated empirically recently in a paper by Amini. et. al. (AoS, 2013). In this paper, we attempt to quantify this improvement through theoretical analysis. Under the stochastic block model (SBM) (and its extensions), previous results on spectral clustering relied on the minimum degree of the graph being sufficiently large to prove its good performance. By analyzing the spectrum of the Laplacian of an SBM as function of the regularization parameter, we provide bounds for the perturbation of the regularized eigenvectors that potentially do not depend on the minimum degree. Moreover, we demonstrate the usefulness of regularization in situations where not all nodes can be clustered accurately. As an important application of our bounds, we propose a data-driven technique DK-est (standing for estimated Davis-Kahn bounds) for choosing the regularization parameter. This technique is shown to work well through simulations and on real data sets.

Inference Of Network Summary Statistics Through Network Denoising

Eric D. Kolaczyk
Prakash Balachandran, Edoardo Airoldi[1]
[1]Department of Statistics, Harvard University

Consider observing an undirected network that is 'noisy' in the sense that there are Type I and Type II errors in the observation of edges. Such errors can arise, for example, in the context of inferring gene regulatory networks in genomics or functional connectivity networks in neuroscience. Given a single observed network then, to what extent are summary statistics for that network representative of their analogues for the true underlying network? Can we infer such statistics more accurately by taking into account the noise in the observed network edges? In this talk I will describe work in which we answer both of these questions. In particular, we develop a spectral-based methodology using the adjacency matrix to 'denoise' the observed network data and produce more accurate inference of the summary statistics of the true network. We characterize performance of our methodology through bounds on appropriate notions of risk in the L2 sense, and conclude by illustrating the practical impact of this work on synthetic and real-world data.

Community Detection In Networks With Node Features

Elizaveta Levina
Yuan Zhang[1], Ji Zhu[1]
[1]Department of Statistics, University of Michigan

Many methods have been proposed for community detection in networks, but most of them do not take into account additional information on the nodes that is often available in practice. We propose a new joint community detection criterion that uses both the network and the features to detect community structure. One advantage our method has over existing joint detection approaches is the flexibility of learning the impact of different features, which may differ across communities. Another advantage is the flexibility of choosing the amount of influence the feature information has on communities. The method is asymptotically consistent under the block model with additional assumptions on the feature distributions, and performs well on simulated and real networks.

Exploring Genomic Data With Bcmi

Chris Pardy
Susan R Wilson[1]
[1]Australian National University, Canberra; University of New South Wales

Large and high-dimensional datasets are increasingly common in genomics. There is a need for exploratory approaches that can identify novel, possibly complex nonlinear, associations in these data which often contain different types of variables (say, continuous and categorical). As it is infeasible to directly inspect plots of all pairs of variables a single measure that can identify a wide class of associations can be of great use. We propose the use of a bias corrected mutual information measure (BCMI) which can be applied to all kinds of variables measured while being comparatively quick and easy to calculate. The use of BCMI provides a novel exploratory approach that can be applied in a wide variety of settings. These association scores can also be used as a basis for clustering and network construction. We demonstrate our approach using genomic data from a mouse model of obesity that contains clinical measurements, microarray gene expression levels (continuous variables) and single nucleotide polymorphisms (SNPs) (categorical variables). We compare our approach with the recently proposed maximal information coefficient (MIC), in particular we show BCMI to have equal or better power than MIC for several common functional relationships. An associated R package “mpmi” is available on CRAN.

Validation Of Ten Non-verbal Personality Extremities

John Magnus Roos

Personality research has established a five-factor model of personalities, constituted by the dimensions of agreeableness, conscientiousness, extraversion, neuroticism and openness (McCrae and Costa, 1992). Each one of the dimensions have two extremities; (1) agreeable versus antagonistic, (2) conscientious versus spontaneous, (3) extravert versus introvert, (4) neurotic versus emotionally stable, (5) open-minded versus close-minded. Together with designers at Veryday in Sweden we have personified each one of the ten extremities (Roos, 2013). The present study aims to validate this non-verbal scale of the five-factor model of personality dimensions.

Each non-verbal extremity [cartoon-like character] was validated through 156 undergraduate students at Stockholm University. The validation process included content validity, criterion validity and construct validity (Laurans, 2011). The content validation of the non-verbal scale is based on tag clouds of adjectives from top-of- mind responses. The tag clouds have been compared to the general framework of the five-factor model of personalities. The criterion validation of the non-verbal scale is based on the equivalence between this scale and an established verbal scale of the five-factor model of personalities, the HP5i (Gustavsson, Jönsson, Linder and Weinryb, 2008; Holmberg and Weibull, 2010). Construct validity refers to how the extremities are constructed and how they differ from each other. Ideally, the items used to measure one non-verbal extremities should be as different as possible but have high correlation between themselves (convergent validity) while the correlations between different extremities (and items) should be as low as possible (divergent validity). Also the construct validation of the non-verbal scale is based on convergent validity and divergent validity in relation to HP5i.

Predicting The Spatial Distribution Of Seabed Hardness Based On Preence/absence Data Using Random Forest

Jin Li
Justy Siwabessy[1], Maggie Tran[1], Zhi Huang [1], Andrew D. Heap [1]
[1]GEOSCIENCE AUSTRALIA

Spatial information on seabed biodiversity is important for marine zone management in Australia and is often predicted using spatially continuous data of seabed biophysical properties. Seabed hardness is one of the important properties for predicting biodiversity and is often inferred from multibeam echo-sounder backscatter data. Seabed hardness can also be inferred based on underwater video footage that is however only available at a limited number of sampled locations. To generate spatially continuous data of seabed hardness from point samples, spatial prediction methods are essential. Random forest (RF) is one of the top performing methods in predictive modeling. Because of its high predictive accuracy, it was introduced into spatial statistics by applying it to continuous environmental data (Li et al., 2011). Such applications have significantly improved the prediction accuracy and opened an alternative source of methods for spatial prediction. Given that it has only been applied to continuous data, a few questions remain, namely: is it data type- specific? How reliable are its predictions for presence/absence data type? To address these questions, in this study we used RF to predict the spatial distribution of seabed hardness based on presence and absence data derived from video classification and 15 seabed property predictors. The prediction accuracy was assessed using a 10-fold cross validation. We tested the effects of various predictor sets on the accuracy of predictive models. We also illustrated the effects of 10-fold cross-validation methods including a new cross-validation algorithm for RF on selecting the most optimal predictive model. In this paper, we discuss the research findings, visually examine the spatial predictions, and compare the results with the findings in previous publications. This study provides an example of predicting the spatial distribution of environmental variables of presence/absence data type using RF.

Applying Mixture Models To High-throughput Data

Geoffrey John McLachlan

In this talk we consider some recent advances in mixture models for undertaking the supervised and unsupervised classification of some high-throughput data in bioinformatics. We also look at their application in performing multiple hypothesis testing. The focus is on the use of mixtures of factor and of linear mixed models. Various examples are to be given for gene expression data.

Wild Binary Segmentation For Multiple Change-point Detection

Piotr Fryzlewicz

We propose a new technique, called Wild Binary Segmentation (WBS), for consistent estimation of the number and locations of multiple change-points in data. We assume that the number of change-points can increase to infinity with the sample size. Due to a certain random localisation mechanism, WBS works even for very short spacings between the change-points, unlike standard Binary Segmentation. On the other hand, despite its use of localisation, WBS does not require the choice of a window or span parameter, and does not lead to significant increase in computational complexity. WBS is also easy to code. We provide default recommended values of the parameters of the procedure and show that it offers very good practical performance. In addition, we provide a new proof of consistency of Binary Segmentation with improved rates of convergence, as well as a corresponding result for WBS.

Applications Of Sequentially Adaptive Bayesian Learning Using Graphics Processing Units

Huaxin Xu
John Geweke[1], Bin Peng[1]
[1]University of Technology Sydney

This paper extends sequential Monte Carlo methods that have been applied to state space models to the more general tasks of integration and optimization. In statistics a leading example of integration is Bayesian inference and leading examples of optimization are maximum likelihood and method of moments estimation. It implements these procedures using graphics processing units, which provide massively parallel desktop computing at less than one dollar per core, in a single Matlab toolbox that makes application of the methods to different models simple and straightforward. Compared with other approaches, these algorithms are highly robust with respect to irregular functions, make significantly fewer demands in implementing new models, and reduce computation time by factors of 10 to 100.

A Massively Parallel Algorithm For Simulated Annealing With Adaptive Temperature Reduction

Bart Firshknecht
John Geweke[1], Garland Durham[2]
[1]University of Technology Sydney
[2]Quantos Analytics

Simulated annealing is a well-established approach to function optimization that can be highly competitive with other methods when the objective function is irregular and/or multimodal. This paper improves this approach using massively parallel sequentially adaptive Bayesian learning algorithms implemented on graphics processing units. It makes three contributions. (1) Applications of simulated annealing typically rely on ad-hoc approaches to finding an effective and efficient temperature reduction schedule. The paper shows how to determine this schedule algorithmically,, thereby greatly reducing required investigator time. (2) Conventional applications of simulated are single-stream and likely to become trapped in local modes. The massively parallel approach in this paper greatly reduces the risk of "solutions" that are local but not global in cases where the number of modes is on the order 10"2 to 10^4. (3) The algorithm is self-tuning, freeing the investigator from the need to approximate gradients or construct local search methods.

Robust Bayesian Inference, Maximum Likelihood And Method Of Moments Estimation By Means Of Simulated Annealing

John Geweke
Garland Durham[1]
[1]Quantos Analytics

Sequential Monte Carlo approaches to Bayesian inference usually proceed by the successive introduction of observations. Practitioners of these methods sometimes use the alternative approach of "powering up" the likelihood function, which has the same theoretical foundations. The method is quite similar to simulated annealing methods for optimization, and inherits the need to design application-specific temperature reduction schedules. This paper uses recently developed massively parallel simulated annealing algorithms to solve the temperature reduction problem, and extends the approach to general systems of estimating equations. It focuses specifically on maximum likelihood and method of moments estimators. It shows how to compute the relevant asymptotic variance matrices as a by-product of the algorithm, thereby obviating the need for analytical or numerical first and second derivatives.

Using Sequentially Adaptive Bayesian Learning And Graphics Process Units For Bayesian Inference When The Likelihood Function Must Be Simulated

Garland Durham
John Geweke[1]
[1]University of Technology Sydney

In models with latent variables exact computation of the likelihood function is often impossible, but in these same situations one can often generate simulations of the likelihood function that are unbiased. This has proved very useful for Bayesian inference, most notably in the context of sequential Monte Carlo algorithms for non-linear state space models. Going back over a decade the literature has noted that such algorithms are close to embarrassingly parallel, this fact has been exploited little in practice. In fact there are a number of theoretical and practical problems that arise in such attempts. This paper enumerates and solves these problems. It implements the resulting adaptation of sequential Monte Carlo in a massively parallel algorithm using C/CUDA and graphics processing units. It demonstrates the effectiveness of the solution in a number of applications.

Delayed Acceptance Particle Mcmc For Exact Inference In Stochastic Kinetic Models

Andrew Golightly
Daniel Henderson, Chris Sherlock

Recently proposed particle MCMC methods provide a flexible way of performing Bayesian inference for parameters governing stochastic kinetic models defined as Markov jump processes (MJPs). Each iteration of the scheme requires an estimate of marginal likelihood calculated from the output of a sequential Monte Carlo scheme (also known as a particle filter). Consequently, the method can be extremely computationally intensive. We therefore aim to avoid most instances of the expensive likelihood calculation through use of a fast approximation. We consider two approximations: the chemical Langevin equation diffusion approximation (CLE) and the linear noise approximation (LNA). Either an estimate of marginal likelihood under the CLE, or the tractable marginal likelihood under the LNA can be used to calculate a first step acceptance probability. Only if a proposal is accepted under the approximation do we then run a sequential Monte Carlo scheme to compute an estimate of the marginal likelihood under the true MJP and construct a second stage acceptance probability that permits exact (simulation based) inference for the MJP. We therefore avoid expensive calculations for proposals that are likely to be rejected. We illustrate the method by considering inference for parameters governing a Lotka-Volterra system, a simple epidemic model and a model of gene expression.

Estimating The Probability Of Simultaneous Rainfall Extremes Within A Region

Lee Fawcett
Dave Walshaw [1], Simone Padoan [2]
[1]Newcastle University
[2]Bocconi University

We investigate the impact of model mis-specification, in terms of the dependence structure in the extremes of a spatial process, on the estimation of key quantities that are of interest to hydrologists and engineers. For example, it is often the case that severe flooding occurs as a result of the observation of rainfall extremes at several locations in a region simultaneously. Thus, practitioners might be interested in estimates of the joint exceedance probability of some high levels across these locations. It is likely that there will be spatial dependence present between the extremes, and this should be properly accounted for when estimating such probabilities. We compare the use of standard models from the geostatistics literature with max-stable models from extreme value theory. We find that, in some situations, using an incorrect spatial model for our extremes results in a significant under-estimation of these probabilities which - in flood defence terms - could lead to substantial under-protection.

Risk Aversion In Equilibrium Modeling Of Cap-and-trade System

Juri Hinz

According to theoretical arguments, a properly designed emission trading system should helpreaching pollution reduction at low social costs. Based on the theoretical work of environmental economists,cap-and-trade systems are put into operations all over the world. However, the practice from emissionstrading yields a real stress test for the underlying theory and to reveals some of its weak points.This paper aims to fill this gap. We extend and generalize existing approaches, by establishing an equilibrium modelingfor risk averse market players. We show how of the architecture of an environmental market can be optimizedunder realistic assumptions or risk aversion and which assumptions and approximations must be made therefore.

Spatial Analysis Of Point Patterns On A Linear Network

Adrian Baddeley

The analysis of spatial patterns of events that occur on a network of lines, such as traffic accidents on a street network, poses many difficult problems for statistical methodology. The network is not a homogeneous space, so that most of the existing methodology for spatial point process cannot be applied. This talk presents some recent advances, including the correct analogues of Gaussian kernel smoothing and Ripley's K-function on a network. Applications include traffic accident research, neuroscience, criminology and ecology.

Modeling Evolving Phylogenies By Means Of Marked Metric Measure Spaces

Sandra Kliem

In this talk, a model for evolving phylogenies, incorporating branching, mutation and competition is introduced. The state-space consists of marked tree-like metric measure (mmm)-spaces. The model arises as the limit of approximating finite population models with rates dependent on the individuals' traits and their genealogical distances. The main focus of the talk will be on presenting the notion of mmm-spaces and to highlight their advantages in the given context. In particular, necessary and sufficient conditions for relative compactness of sets in mmm-spaces are explained. The route to verify these conditions to conclude the tightness of the approximating models from above is given. A similar approximating model and its limit is treated in the framework of nonlinear historical superprocess approximations. In the framework of mmm-spaces, work of [Depperschmidt, Greven, Pfaffelhuber and Winter, 2012-2013] introduces and studies tree-valued Fleming-Viot dynamics. During this talk, new ideas and challenges that arise from working with mmm-spaces in the context of evolving phylogenies are put into context of the above.

Mean Field Games With A Dominating Player: Theory, Examples, And Hysteresis

Phillip Yam

In this talk, I first introduce a class of mean field games between a dominating player and a group of agents, each of them acts similarly and also interacts with each other through a mean field term being substantially influenced by the dominating player. I then provide the general theory and discuss the necessary condition for the optimal controls and game condition via an adjoint equation approach. The special case in the context of linear-quadratic framework will be discussed, in which a necessary and sufficient condition can be asserted by stochastic maximum principle; further connection with hysteresis effect in the population will be indicated.
This is a joint work with Alain Bensoussan and Michael Chau.

Taming Non-response In Finite Population Sampling With Auxiliary Information

Siu-Ming Tam

Non-response in surveys is a fact of life. To reduce non-response bias, traditional responses from National Statistical Offices (NSOs) are to "throw money at the problem" by converting as many non-respondents into respondents as possible. With budget pressure facing NSOs, including the ABS, the traditional approach is becoming non-affordable, and a different approach to managing non-response is required. In this talk, Siu-Ming Tam will explain that if the non-response mechanism is non-informative, given the benchmarking variables, the GREG estimator has an automatic adjustment feature for non-response bias. However, if it is only non-informative, given the benchmarking variables, AND an additional X variable known only for those units in the sample, then we can (a) still use the GREG estimator provided that we follow up a supplementary sample of the non-respondents to achieve "balance" on the X variable; or (b) adjust the GREG estimator for this balance if supplementary sample is not followed up. Siu-Ming will show, in either case, the quantifiable increase in the uncertainty of the GREG estimator. The statistical implications of these new findings in managing non-response in the ABS will also be explained in the talk.

Joint Estimation Of Quantile Planes

Surya Tokdar

Quantile regression offers direct inference on non-central parts of the response distribution and captures dependence beyond changes to the mean. These attributes have serious practical benefits in applications to ecology, economics, epidemiology and climate change studies. Most scientific applications of quantile regression require inference over a dense grid of response quantiles. Often primary interest focuses on identifying the predictors that affect only a subset of quantiles, e.g., the extremes. However, the current practice of drawing composite inference by stitching together single quantile model fits has several shortcomings: the estimated quantile curves may cross violating laws of probability, a poor borrowing of information across quantiles may lead to unreliable composite inference on predictor relevance and often uncertainty is undermined near the data boundary. In this talk I will present recent work where we develop a logically coherent statistical model that facilitates joint estimation of conditional quantile planes over multivariate, convex predictor domains of arbitrary shapes, facilitating a likelihood based joint estimation of quantile planes with a Bayesian semi-parametric method. I will present theoretical results on asymptotic guarantees and empirical studies comparing the new method against single quantile fits and also an existing semi-parametric approach that offers a partial solution to the joint estimation problem.

On Various Confidence Intervals Post-model-selection

Hannes Leeb

We compare several confidence intervals after model selection in the setting recently studied by Berk et al. (2013), where we focus on the PoSI- intervals that are proposed in that paper, and on the 'naive' confidence interval, where the presence of model selection is ignored and the interval is constructed as if the selected model were correct and fixed a-priori. Overall, we find that the actual coverage probabilities of all these intervals are moderately close to the desired nominal coverage probability. This finding is in stark contrast to several papers in the existing literature, because Berk et al. (2013) consider confidence intervals for a non-standard quantity of interest that depends on the selected model.

What Makes A Good Reviewer?

Peter Hall

First name, last name, title: PETER, HALL PROF.
Affiliation/company: UNIVERSITY OF MELBOURNE
Address, phone nr: DEPT OF MATHEMATICS & STATISTICS, UNIVERSITY OF MELBOURNE, VIC 3010; 03 8344 9682

Theme as per list on the conference website: YOUNG STATISTICIANS

15 MINUTE TALK INVITED BY SUSANNA CRAMB; TITLE AND ABSTRACT: "What makes a great reviewer?"

Refined Limit Theorems For Random Walks Among Random Conductances

Nina Gantert

*Nina Gantert, Prof.
*Technische Universität München
*Fakultät für Mathematik, Boltzmannstr. 3, 85748 Garching, phone
49 89 289-18780

*Short Bio: Nina Gantert was born in Zürich and got her undergraduate diploma in mathematics from ETH Zürich.
She did her Ph. D. 1991 in Bonn. Then she spent one year as a Post-Doc at ETH, 6 years at TU Berlin, interrupted by a one-year visit to Technion Haifa and a visit to Paris VI. In 2000, she moved to Universität Karlsruhe as an associate professor.
In 2004, she became a full professor at Universität Münster. Since 2011, she holds a chair of probability at Technische Universität München. Her field is probability, in particular, she is interested in stochastic processes, random media and large deviations.

*Theme: Stochastic processes
I am speaking in the session "Random walks in random environments"
organized by Mark Holmes.
(I will also give a medallion lecture in the conference).

Title: Refined Limit Theorems for Random Walks among Random Conductances We introduce the Random Conductance Model which has attracted a lot of recent interest. We survey limit theorems for this model and explain why refined limit theorems are needed. Then, we present some results in this direction which are based on joint work with Omar Boukhadra, Christophe Gallesco, Serguei Popov and Marina Vachkovskaia.

Mean Field Limits Of Symmetric Stochastic Differential Games With Common Noise

Daniel Lacker

A general characterization is derived for the limits of approximate equilibria of large-population symmetric stochastic differential games as the number of agents tends to infinity. It is shown that the equilibrium empirical measures admit limits in distribution, and every limit is a weak solution of the mean field game (MFG).
Conversely, every weak MFG solution can be obtained as the limit of a sequence of approximate equilibria in the finite-player games. In other words, the MFG precisely characterizes the possible limits of the finite-player games, formalizing the well-known intuition. The proofs use relaxed controls to provide the compactness needed to obtain limits under quite general assumptions, and it is then shown how to sharpen the characterization of the limit under various additional assumptions. In particular, under modest convexity assumptions, versions of the main theorems are stated with no mention of rela xed controls. Stronger assumptions yield uniqueness of the weak MFG solution and thus a full convergence result.

Consistent Model Selection Criteria For Quadratically Supported Risks

Yongdai Kim

"Consistent model selection criteria for quadratically supported risks"
We consider a class of model selection criteria and provide sufficient conditions for a given model selection criterion to be consistent for a class of loss functions so called "the quadratically supported risks." The class of loss functions considered in this talk includes the squared loss, Huber loss, logistic loss and quantile loss. We also discuss the pathwise-consistent of the thresholded Lasso and SCAD (or MCP) estimators with the quadratically supported risks.

Visualising Bigger Data In R

Hadley Wickham

Title: Visualising bigger data in R

In this talk I'll discuss the perceptual and computational challenges of visualising 10-100 million observations on commodity hardware fast enough for interactive exploratory data analysis (e.g. <5 seconds per plot). This is challenging, but achievable, and I'll demonstrate a reference R + C++ implementation that demonstrates the key ideas (which are equally applicable to other programming environments). The visualisations are built around a process of group, summarise,
(smooth) and visualise, and I'll discuss how this process fortuitously makes sense from statistical, computational and data analytic perspectives.

Alternate Constructions Of The Gaussian Free Field And Fast Simulation Of Schramm-loewner Evolutions

Brent Werness

The Schramm--Loewner evolutions (SLE) are a family of stochastic processes which describe the scaling limits of curves which occur in two-dimensional critical statistical physics models. SLEs have had found great success in this task, greatly enhancing our understanding of the geometry of these curves. Despite this, it is rather difficult to produce large, high-fidelity simulations of the process due to the significant correlation between segments of the simulated curve. The standard simulation method works by discretizing the construction of SLE through the Loewner ODE which provides a quadratic time algorithm in the length of the curve.

Recent work of Sheffield and Miller has provided an alternate description of SLE, where the curve generated is taken to be a flow line of the vector field obtained by exponentiating a Gaussian free field. In this talk, I will describe a new method of approximately sampling a Gaussian free field, and show how this allows us to more efficiently simulate an SLE curve.

Stochastic Control Problems With Singular Value Functions -- Regular Solutions And Applications To Optimal Portfolio Liquidation

Paulwin Graewe

We study stochastic optimal control problems with singular terminal values arising in models of optimal portfolio liquidation under market impact when traders can simultaneously submit active orders to a primary market and passive orders to a dark pool. For the benchmark Markovian control model we show that the value function can be characterized by associating the HJB equation with an appropriate singular terminal condition, and establish existence and uniqueness of a classical solution. If the market impact or cost function is non-Markovian, then the value function can be described by a BSPDE rather than a PDE. We show that the BSPDE with singular terminal condition has a unique, sufficiently regular solution.

The talk is based on joint work with Ulrich Horst, Jinniao Qiu (both Humboldt University of Berlin) and Eric Séré (Paris-Dauphin).

Stochastic Price Dynamics Of Emission Permits: Theory And Empirical Evidence

Steffen Hitzemann

We analyse the stochastic price dynamics of emission permits induced by the design of state of-the-art emission trading systems from both a theoretical and an empirical perspective. Based on a theoretical equilibrium model accounting for all important design features of today's cap-and-trade systems, we characterize an emission permit as a strip of European binary options written on economy-wide emissions. These option characteristics especially imply (i) a state- and time-dependent volatility structure, (ii) a futures price curve that is partly in contango and partly backwardated, and (iii) a downward-sloping option smile. Empirical evidence from existing emissions markets confirms the theoretical predictions. Furthermore, we show that reduced-form models for emission permit prices that capture these features in a simplified way outperform classical approaches for asset price dynamics both regarding their historical model fit to futures prices and their option pricing performance.

Pathwise Solutions To Fully Nonlinear Spdes

Paul Gassiat

Stochastic viscosity theory is concerned with the solution to (nonlinear, parabolic) equations driven along a Brownian trajectory. Although these equations only makes classical sense when the driving signal is Lipschitz-continuous, it is sometimes possible to extend the notion to paths with less regularity.

This was achieved in the original articles of Lions--Souganidis for the “constant coefficient” case, along with complete existence/uniqueness theorems, as well as pathwise stability in the driving signal. In the general case, some results were recently obtained by Caruana, Friz and Oberhauser for some specific cases (e.g. linear Hamiltonians). Even in these special cases, pathwise stability may be lost and the best one can get here is stability in rough path sense à la T. Lyons.

The general case however still offers room for better understanding. We present in this talk new results in this direction, in particular we will discuss the case of quadratic Hamiltonians, using the key fact that they can be related to a Riemannian structure.

Synthesizing Genetic Markers For Incorporation Into Clinical Risk Prediction Tools

Sonja Grill
Mahdi Fallah[1], Donna P. Ankerst[2]
[1]German Cancer Research Centre
[2]Technical University Munich

Clinical risk prediction tools built on standard risk factors are important devices for many different diseases. Newly discovered genetic and high-dimensional-omic markers, such as single nucleotide polymorphisms (SNPs) and gene expressions, have the potential to increase the practical utility of clinical risk prediction tools. Typically these markers are assessed among multiple large case-control studies but not in the original cohorts used to build the existing risk prediction tools, making their incorporation into those tools complicated. We provide an intuitive Bayesian method for updating an existing clinical risk prediction tool with external marker information via the use of likelihood ratios to transform the prior odds of a disease to posterior odds. We illustrate the method with two applications, the first incorporating SNPs from multiple published genome-wide association studies via a random-effects meta-analysis and the second, detailed family history of cancer from the nationwide Swedish Family-Cancer Database (the world's largest of its kind), into the Prostate Cancer Prevention Trial Risk Calculator.

Measuring The Adaptive Capacity And Vulnerability Of Agricultural Communities To Climate Change

Philip Kokic

Vulnerability is a term frequently used to describe the potential threat to rural communities posed by climate variability and change. Despite a growing use of the term, analytical measures of vulnerability that are useful for prioritising and evaluating policy responses are yet to evolve. Demand for research capable of prioritising adaptation responses has grown rapidly with the increasing awareness of climate change and its current and potential future impacts on rural communities.

We show how hazard/impact modelling can be combined with an holistic measure of adaptive capacity to analyse the vulnerability of Australian rural communities to climate variability and change. Bioeconomic modelling was used to predict the exposure and sensitivity of Australian rural communities to climate variability and change. Rural livelihoods analysis was used as a conceptual framework to construct a composite index of adaptive capacity. We illustrate how to construct this index using a combination of existing data sources including sample survey data, biophysical data and environmental data. Bringing these data together in a manner that does not distort the analysis is important.

Relying on hazard/impact modelling alone can lead to entirely erroneous conclusions about the vulnerability of rural communities, with potential to significantly misdirect policy interventions. Here we present a preliminary assessment of which Australian rural communities are vulnerable to climate variability and change. Our results reveal a complex set of interacting environmental, economic and social factors contributing to vulnerability.

Generalized Gamma Approximation With Rates For Urns, Walk, And Trees

Adrian Roellin
Erol Peköz[1], Nathan Ross[2]
[1]Boston University
[2]University of Melbourne

We present recent limit theorems for certain generalised Polya urns which arise in preferential

attachment processes and which converge to generalised gamma distributions. We give applications to various statistics in random walk and trees.

On An Extension Of Aldous' Standard Multiplicative Coalescent

Amarjit Budhiraja

On an Extension of Aldous' Standard Multiplicative Coalescent.

Abstract: We introduce a two component Markov process on an infinite dimensional state space for which the first component is Aldous' standard multiplicative coalescent while the second component takes values in $\mathbb{N}^{\infty}$.
Distribution at any fixed time instant is given in terms of excursions of a reflected Brownian motion with drift and an independent standard Poisson process. We show that the process is `nearly Feller' in a suitable sense. Using this property we characterize the joint asymptotic behavior of the vector of component sizes and surpluses in the critical scaling window for a general family of random graphs corresponding to bounded sized rules. This is joint work with S.Bhamidi and X.Wang.

Generalized Fiducial Inference For High Dimensional Problems

Jan Hannig

In recent years the ultrahigh dimensional linear regression problem has attracted enormous attention from the research community. Under the sparsity assumption most of the published work is devoted to the selection and estimation of the predictor variables with non-zero coefficients. This paper studies a different but fundamentally important aspect of this problem: uncertainty quantification for parameter estimates and model choices. To be more specific, this paper proposes methods for deriving a probability density function on the set of all possible models, and also for constructing confidence intervals for the corresponding parameters. These proposed methods are developed using the generalized fiducial methodology, which is a variant of Fisher's controversial fiducial idea. Theoretical properties of the proposed methods are studied, and in particular it is shown that statistical inference based on the proposed methods will have correct asymptotic frequentist property. In terms of empirical performance, the proposed methods are tested by simulation experiments and an application to a real data set. Lastly this work can also be seen as an interesting and successful application of Fisher's fiducial idea to an important and contemporary problem. To the best of the authors' knowledge, this is the first time that the fiducial idea is being applied to a so-called "large p small n" problem.

Two-sample Thresholding Tests For High Dimensional Means

Song Chen

We propose two tests for the equality of two population mean vectors under high dimensionality and column-wise dependence by thresholding.
They are designed to obtain better power performance when the mean vectors of two populations differ only in a small number of coordinates.
The first test is constructed based on the original data and achieves a power improvement by reducing the level of variance of the test statistics with thresholding.
When the data are column-wise dependent, the second test based on transformed data by the inverse of the linear combination of two covariance matrices produces further power improvement by not only reducing the variance but also enhancing the signal strength. The asymptotic distributions of test statistics are established and the power of two tests are analyzed. It is shown that the second test is particularly powerful by incorporating the correlations among the coordinates of the variables. Simulation studies are conducted to confirm the theoretical findings and to offer practical performance of the tests.

Keywords: Large deviation; Large p small n; Sparse signals; Threshold.

Statistical Inference By Crowd-sourcing

Di Cook

Plots of data often provoke the response ``is what we see really there''. In this talk we will discuss ways to give visual statistical methods an inferential framework. Statistical significance of ``graphical discoveries'' is measured by having the human viewer compare the plot of the real dataset with collections of plots of null datasets: plots take on the role of test statistics, and human cognition the role of statistical tests, in a process modeled after the ``lineup'', popular from criminal legal procedures. This is a simple but rigorous protocol that provides valid inference, yielding p-values and estimates of the test power, for graphical findings. Amazon's Mechanical Turk is used to implement the lineup protocol and crowd-source the inference. Turk is a resource where people are employed to do tasks that are difficult for a computer, in this case, evaluating structure in plots of data. With a suite of experiments, the lineup protocol was run head-to-head against the equivalent conventional test, yielding results that mirror those produced by classical inference. This talk will describe these results, and show how the lineup protocol is used for assessing graphical findings and designing good data plots.

From Compressed To Corrupted Sensing

Rina Foygel

In the compressed sensing problem, a small number of linear measurements can be used to accurately recovery a high dimensional signal, as long as the signal is sparse or has some other low-dimensional structure that allows us to undersample. We extend these ideas to the problem of corrupted sensing, where linear measurements of a structured signal are also obscured by some kind of structured corruption - for example, a sparse set of measurements may be completely unreliable. Our method allows arbitrary structure in both the signal and corruption, as long as these structures can be captured by a convex norm or penalty, such as using the L1 norm for sparse signal recovery. We relate the corrupted sensing problem to the geometric notions of the Gaussian complexity of a tangent cone and the Gaussian distance to a subdifferential. Our analysis covers both constrained and penalized convex programs, in each case giving guarantees of exact signal recovery from structured corruption and stable signal recovery from structured corruption with added unstructured noise. We also show simulations in a range of settings that confirm the phase transitions found in the theory. (Joint work with Lester Mackey.)

Some Existence Results On Locating And Detecting Arrays

Yu Tang

Proposed in Colbourn and McClary (2008), locating and detecting arrays are of interest in generating software test suites to cover all t-way component interactions and locate or detect interaction faults in component-based systems. In this talk, I will introduce some lower bounds on the size of locating or detecting arrays with specific parameters, and then prove that optimal locating arrays meeting these bounds can be equivalently characterized in terms of orthogonal arrays or other configurations with prescribed properties. Moreover, using these characterizations, we develop several construction methods and obtain some infinite series of optimal locating arrays and detecting arrays

Learning In Sequential Decision Problems

Peter Bartlett

Many stochastic optimization problems that arise in robotics, control, and economics can be formulated as sequential decision problems in which a strategy's current state and choice of action determine its loss and next state, and the aim is to choose actions so as to minimize the sum of losses incurred. We consider three problems of this kind: Markov decision processes with adversarially chosen transition and loss structures; approximate policy optimization for large scale Markov decision processes; and linear tracking problems with adversarially chosen quadratic loss functions.

The key challenge in these problems is to develop methods that are effective with large state spaces.

We aim to incur a total loss that is not too much worse than the best in some comparison class. Since optimality with respect to the class of all policies is unachievable in general for large scale problems, we consider more restricted comparison classes. We present algorithms and optimal excess loss bounds for these three problems. We show situations where these algorithms are computationally efficient, and others where hardness results suggest that no algorithm is computationally efficient.

Joint work with Yasin Abbasi-Yadkori, Varun Kanade, Alan Malek, Yevgeny Seldin and Csaba Szepesvari.

Multivariate Regression Model With An Underlying Network

Simon Cheung

Multivariate Regression Model with an underlying network

Simon K.C. Cheung[1], Tommy K.Y. Cheung[2]

[1]The Open University of Hong Kong, Hong Kong

[2]Hang Seng Management College, Hong Kong

Regression analysis is a statistical method for the investigation of relationships between variables of interests. Values of these variables are typically observed from individuals that belong to a population. It is often the case that there is an underlying network connecting the individuals in the population. This underlying network can be directed or un-directed. Empirical studies confirmed that such a network possesses many properties such as growing number of vertices over time, power law degree distribution, existence of hubs, small or decreasing diameter, high clustering coefficient, and robustness against random removal of vertices. A form of preferential attachment scheme has been proved to exhibit many of the said properties. The underlying complex network is also dynamic in nature, either through re-wiring or natural growth, which may affect the estimation of the prediction values. We introduce a multivariate regression model in which the underlying complex network assumes a preferential attachment scheme. This extends the idea of Pinkham and Imbens (2001), who examined regression on a friendship network. The proposed model simultaneously allows both inferences on relationships between variables and for dynamic configurations of the complex network structure. It is also well suited to both directed and un-directed networks. With this multivariate regression model, a Gibbs Sampler has been developed to generate random numbers from the joint posterior distribution. At each of the simulation steps in Gibbs Sampling, the underlying complex network is regenerated with the updated posterior probabilities. Information such as the posterior distributions of parameters can be obtained to provide insight into regression analysis.

On The Asymptotics Of Nadaraya-watson Estimator: Toward A Unified Approach

Qiying Wang

This paper investigates the asymptotics of Nadaraya-Watson estimator, providing a framework and a unified approach for stationary and non-stationary times series. This paper also establishes an extension to multivariate regression with non-stationary time series and provides a brief overview on the asymptotic results in relation to non-linear cointegrating regression.

Distributional Limits For Generalized Polya Urn Models, Stein's Method And The Beta-gamma Algebra

Nathan Ross

Since their conception 90 years ago, Polya urn models and generalizations have been of great interest to mathematicians, statisticians, biologists, and more recently, computer scientists. The basic model is that an urn contains balls of different colors and at sequential steps a ball is randomly chosen from the urn, its color noted, and then the contents of the urn are altered based on this color; in the classical Polya urn the ball drawn is returned to the urn along with another of the same color. This talk will focus on the limiting distribution of the (scaled) composition of the urn in a collection of these models, some of which (both the limits and the models) do not appear to have been studied previously. The limiting distributions are characterized as unique fixed points of certain distributional transformations and are explicitly written as products of powers of independent beta and gamma variables. Our methods suggest some conjectures for limiting distributions in further models and these conjectures suggest a bigger picture that we don't yet have. Joint work with Erol Pekoz and Adrian Roellin.

Moderate Deviations For Studentized Two-sample U-statistics With Applications

Wen-Xin Zhou

Abstract: In this talk, we derive Cram_er type moderate deviation theorems in a general Studentized two-sample U-statistics framework, including two-sample t-statistic and Wilcoxon test as prototypical examples. A re_ned moderate deviation theorem with second-order accuracy is also established for two-sample t-statistic. One-sample tests are appropriate when a sample is being compared to the population from a hypothesis, where the population information is known as a prior or can be calculated from the population. Two-sample tests, on the other hand, are appropriate for comparing two groups, typically experimental and control groups from scienti_cally controlled experiments in the _eld of economics and medical research, for example, in microarray analysis for comparing two groups and in multi-center clinical trials where researchers seek to summarize the di_erence between the treatments and investigate whether treatment-by-center interaction exist. In particular, we are interested in the applications of our results to multiple-hypothesis testing problems, where we use a regularized Bootstrap calibration method, that was recently proposed by Liu and Shao (2013), in large scale two-sample t-tests with false discovery rate control. This is a joint work with Qi-Man Shao and Jinyuan Chang.

AMS 2000 subject classi_cations: Primary62E20, 60F10.

Keywords and phrases: Moderate deviation, two-sample U-statistics, Studentized statistics, Bootstrap, false discovery control.

Multiple Event Incidence And Duration Analysis For Credit Data Incorporating Non-stochastic Loan Maturity

John Watkins

Applications of duration analysis in economics and finance exclusively employ methods for events of stochastic

duration. In application to credit data, previous research incorrectly treats the time to predetermined maturity

events as censored stochastic event times. The medical literature has binary parametric 'cure rate' models that

deal with populations that never experienced the modelled event. We propose and develop a multinomial

parametric incidence and duration model, incorporating such populations. In the class of cure rate models, this is

the first fully parametric multinomial model and is the first framework to accommodate an event with

predetermined duration. The methodology is applied to unsecured personal loan credit data provided by one of

Australia's largest financial services organizations. This framework is shown to be more flexible and predictive

through a simulation and empirical study that reveals: simulation results of estimated parameters with a large

reduction in bias; superior forecasting of duration; explanatory variables can act in different directions upon

incidence and duration; and variables exist that are statistically significant in explaining only incidence or

duration.

Inference For Population Dynamics In The Neolithic Period

Richard Boys

Inference for population dynamics in the Neolithic period

We consider parameter estimation for the spread of the Neolithic incipient farming across Europe using radiocarbon dates. We model the arrival time of farming at radiocarbon-dated, early Neolithic sites by a numerical solution to an advancing wavefront. The model allows for (technical) uncertainty in the radiocarbon data, lack-of-fit of the deterministic model and uses a Gaussian process to smooth spatial deviations from the model. Inference for the parameters in the wavefront model is complicated by the computational cost required to produce a single numerical solution. We therefore employ Gaussian process emulators for the arrival time of the advancing wavefront at each radiocarbon-dated site. We validate our model using predictive simulations. This work appeared recently in a special issue of the Annals of Applied Statistics on the Mathematics of Planet Earth.

Adaptive Piecewise Polynomial Estimation Via Trend Filtering

Ryan Tibshirani

Trend filtering is a recently proposed tool of Kim et al. (2009) for nonparametric regression. It enjoys both strong computational and statistical properties. In particular, it can be fit in nearly linear time with specialized optimization techniques, and converges at the minimax rate over a broad class of underlying functions (functions whose kth derivative is bounded in total variation). In addition to covering the properties of univariate trend filtering, this talk discusses an extension of the trend filtering framework to graphs.

Paracontrolled Distributions And Singular Pdes

Massimiliano Gubinelli

We use the idea of para-products to introduce a class of random generalised functions and a calculus of non-linear operations on them which allows us to understand few examples of singular random PDEs in a reasonably direct way. We will explain how to use these techniques to handle the KPZ equation, the stochastic quantization equation in 3 dimension and a parabolic Anderson model in two dimensions.

Tight Convex Relaxations For Sparse Matrix Factorization

Guillaume Obozinski

We consider statistical learning problems in which the parameter is a matrix which is the sum of a small number of sparse rank one (non-orthogonal) factors, and which can be viewed as generalizations of the sparse PCA problem with multiple factors. Based on an assumption that the sparsity of the factors is fixed and known, we design a matrix norm which provides an tight although NP-hard convex relaxation of the learning problem. We consider also a natural variant of that norm related to the planted clique problem. We study the sample complexity of learning the matrix in the rank one case and show that considering a computationally more expensive convex relaxation leads to an improvement of the sample complexity by an order of magnitude as compared with the usual convex regularization considered, like combinations of the ℓ_1-norm and the trace norm. We also propose an algorithm, relying on a rank-one sparse PCA oracle to solve the convex problems considered and illustrate that, in practice, when state-of-the-art heuristic algorithms for rank one sparse PCA are used as surrogates for the oracle, our algorithm outperforms other existing methods.

Modelling Patient Flow In Hospitals

Mark Fackrell

In order for hospitals to meet the competing demands of emergency and elective surgical patients, with limited resources, it is essential for managers and administrators to know how patients are flowing through the hospital. Modelling patient flow enables the source of bottlenecks to be detected and gives insight into how they can be alleviated. In this talk I will discuss how patient flow at a major metropolitan hospital is modelled.

Comparing Singular Value Decomposition And Non-negative Matrix

Kumer Pial Das

In statistical theory and application, two-way tables of numeric data are often analyzed using dimension reduction methods like the principal component analysis (PCA), singular value decomposition (SVD), and non-negative matrix factorization (NMF).The e ect of PCA, SVD and NMF on the large data set has been represented in di erent ways. This study is designed to compare PCA, SVD, and NMF using mortality data in USA since 1968 to 2010.

Assessing Changes Over Time In Health Care Provider Performance

Jessica Kasza

Assessing changes over time in health care provider performance

It is often important to monitor the changes in the performance of health care providers over time. Such analyses can be used to determine if there are any providers showing significant improvements, deteriorations, unusual patterns, or systematic changes in performance. Studies which monitor health care provider performance in this way have typically been limited to comparing performance in the most recent period with performance in the previous period, ignoring any antecedent performance indicators. It is also important to consider a longer-term view of performance, and assess changes over more than two periods. For example, if a provider exhibits a deterioration in performance in the most recent period, this deterioration may be of greater cause for alarm if it was preceded by deteriorations in performance in previous periods than if it was preceded by improvements in performance. Accounting for a longer run of previous performance indicators when testing for a recent change in performance can provide additional information about the significance of the change.

We present new test statistics that account for variable numbers of prior performance indicators, accounting for the clustering of performance indicators within providers, and show that these are particularly useful for assessing consecutive improvements or deteriorations in performance. The new test statistics have gains in power over previously available test statistics, and allow an understanding of changes in performance of health care providers in the context of longer-term trends.

We apply the test statistics to data from Australian and New Zealand intensive care units over the period 2006-2010, assessing the changes in risk-adjusted mortality levels undergone by each intensive care unit.

On The Anomalous Fluctuations For The Number Of Blocks In The Bolthausen-sznitman Coalescent

Jean Bertoin

Jason Schweinsberg has recently established a sharp limit theorem for the fluctuations of the number of blocks in the Bolthausen-Sznitman coalescent, where the limit involves a Cauchy distribution. We shall explain these anomalous fluctuations by analyzing the effects of percolation at different stages of the construction of a random recursive tree. Specifically, we shall point at a first phase where the germ of these anomalous fluctuations appears, and a second phase of regular growth. The argument relies on the construction (due to Goldschmidt and Martin) of the Bolthausen-Sznitman coalescent based on random recursive trees, and on a coupling (due to Iksanov and Möhle) relating the destruction of random recursive trees to a remarkable random walk.

Branching Processes And Epidemics.

Andrew Barbour

Branching processes and epidemics.

A. D. Barbour, Universitaet Zuerich

Many models of epidemic spread have a common qualitative structure. First, the numbers of infected individuals during the initial stages of an epidemic can be well approximated by a branching process, a fact exploited by Whittle (1955) for approximating the probability that an initial infection leads to a large outbreak of disease. Thereafter, the proportion of individuals that are susceptible follows a more or less deterministic path (Kendall, 1956); for instance, that obtained by solving Kermack & McKendrick's (1927) equation. In this talk, we show that the deterministic path can also be determined from the distribution of a random variable derived from the backward, susceptibility branching process associated with the epidemic. Examples that can be treated in this way include a stochastic version of the Kermack & McKendrick model, the Reed--Frost model, and the Volz configuration model.

(Joint work with Gesine Reinert, Oxford University)

Ph: 03 9482 2740, a.d.barbour@math.uzh.ch
Presentation theme as per list on the conference website –

perhaps `Stochastic models in biology'? Joshua Ross's section, anyway.

Practical Approaches To Sample Design Using Imperfect Design Information

Robert Clark

A well-designed sampling plan can greatly enhance the information that can be produced from a survey. Once a broad sample design is identified, specific design parameters such as sample sizes and selection probabilities need to be chosen. This is typically achieved using an optimal sample design, which minimizes the variance of a key statistic or statistics, expressed as a function of design parameters and population characteristics, subject to a cost constraint. In practice, only imprecise estimates of population characteristics are available, but the effects of this variability are usually ignored. Two approaches to sample design using imprecise design data are proposed. The first is based on the availability of two sets of design data, which can act as a check on each other. Design parameters are numerically optimised to minimise a variance estimator based on statistical learning principles. The second approach involves fitting regression models to design data, and can be easily implemented using existing allocation software. Simulation results based on real data show useful gains in a hypothetical farm survey, business survey, and household survey of a subpopulation.

Items For A Simple Step-stress Model With Progressive Hybrid Censoring From The Exponential Distribution

Indrani Basak

In this article we consider the problem of predicting times to failure of units from the Exponential distribution which are censored under a simple step-stress model. We discuss two kinds of predictors - the Maximum Likelihood Predictors (MLP) and the Conditional Median Predictors (CMP) in the context of progressive Type I and progressive Type II hybrid censoring scheme. In order to illustrate the prediction methods we use some numerical examples. Furthermore, mean squared prediction error (MSPE) and prediction intervals are generated for these examples using simulation studies. MLP and the CMP are then compared with respect to MSPE and the prediction interval for each type of censoring.

Key Words: Accelerated Testing; Conditional Median Predictor; Maximum Likelihood Predictor; Mean Squared Prediction Error; Order Statistics; Prediction Interval; Progressive Type I Hybrid censoring; Progressive Type II Hybrid Censoring.

On The Convergence Of Particle Markov Chain Algorithms

Pierre Del-Moral

Particle Markov chain Monte Carlo methodologies (PMCMC) is a new class of Markov chain Monte Carlo (MCMC) type algorithms equipped with sophisticated particle type proposal probabilities to sample from high dimensional probability distributions. The two most popular classes of algorithms developed in the literature are the particle version of the Metropolis-Hasting algorithm (PMH) and the particle Gibbs sampler (PGS). By construction, these particle population models are defined on extended product state spaces. Nevertheless their common feature is that their marginal distributions doesn't depend on the number of particles, and they need to coincide with some prescribed target distribution. This talk is concerned with some convergence properties of the PMH and the PGS models.

Generalized Method Of Moments Estimator Based On Semiparametric Quantile Regression Imputation

Cindy Yu

We propose an imputation method to handle missing response values using semiparametric quantile regression method. In the proposed method, the missing response values are generated based on the estimated conditional quantile regression function at given values of covariates. Then generalized method of moments is used to estimate model parameters defined through a general estimation equation. We establish the consistency and the asymptotic normality of our estimator based on the proposed imputation method. A simulation study is provided to show the adequacy of the proposed method.

The Asymptotic Covariance Of Multi-dimensional Renewal Reward Processes

Brendan Patch
Yoni Nazarathy[1], Thomas Taimre[1]
[1]The University of Queensland

We consider a sequence of independent and identically distributed random vectors with possibly dependent coordinate elements, where the first coordinate has non-negative support. The first coordinate represents the time between occurrences of events, which we call *renewals*, and the other coordinates represent *rewards*, that accumulate at the time of the associated renewal. We provide a second order approximation for the covariance matrix of cumulative rewards in terms of the moments and cross moments of the coordinate random variables. Our expression becomes exact as time goes to infinity and extends a classic result of Brown and Solomon for the variance of an individual reward coordinate.

Cluster Randomized Crossover Trials In Clinical Research: Design And Analysis Issues For Practical Implementation

Andrew Forbes

Cluster randomised crossover trials are a class of multiple-period cluster designs that have been increasingly used in clinical and public health research. These trials gain efficiency by incorporating treatment crossover across observation periods within each cluster. However, the development and assessment of these designs to date has been limited. In this presentation we report on our recent design and analysis work: We present expressions for the variance of treatment effect estimators and the resulting sample size formulae which take into account period effects, within- and between-period intracluster correlations, as well as within-and between-period cluster sizes, and discuss an extension for use with binary outcomes. Using an underlying marginal model for binary outcomes, we present results of a simulation exercise with binary outcomes, varying sample sizes and outcome prevalences, and discuss problematic parameter configurations. We illustrate all methods with an application involving a proposed large cluster randomised crossover trial to evaluate interventions to reduce mortality in the intensive care research setting. We also discuss the potential for extension to multiple-period-multiple treatment designs and conditions for their feasibility in particular research settings.

On Dimensioning Intensive Care Structures

Nico van Dijk

Clearly, Intensive care units (ICU's) in conjunction with operating theatres are of vital importance within hospitals. For a number of natural reasons, data on the probability for an ICU being congested are highly limited and largely underscore the real congestion. To also capture the interaction with the Operating theatre (OT) for necessary postoperative care, a combined OT-ICU system will be studied.

It will be argued as well as formally proven that the critical congestion probability can well be approximated as well as be bounded by standard Erlang loss queues. These results are of particular real-life interest for dimensioning the number of beds and associated resources. Some extending structures, as with specialized ICU's, more flexible allocation, step down units and after care departments (e.g. nursing homes), will be mentioned as of remaining challenging research interest.

Diffusion Approximations And Lookdown Constructions For Moran Models

Thomas G. Kurtz

Moran models of population genetics are just one example of models involving finite collections of similar "particles" whose types and/or locations evolve in time. Similar models arise in finance, queueing, and many other areas. The problem of approximating these models when the number of particles is large leads naturally to diffusions, measure-valued processes, and stochastic partial differential equations. The approximations are typically derived by normalizing the empirical measure of the set of type/location values and passing to the limit as the number of particles in the model goes to infinity. Here, we keep the particles discrete and obtain a limiting model given by a countable collection of discrete particles. A simple "lookdown" construction for a Moran model makes this convergence obvious, although identifying exactly what the limit is may not be so obvious.

In many other settings, limit arguments can be simplified and/or additional insight obtained by first obtaining a limiting model with a countably infinite collection of particles. The more familiar limits are then defined in terms of this countable collection. The second and third lectures will discuss some of the technical tools needed and provide additional examples.

References

Peter Donnelly and Thomas G. Kurtz. A countable representation of the Fleming-Viot measure-valued diffusion. Ann. Probab., 24(2):698-742, 1996.

Peter Donnelly and Thomas G. Kurtz. Particle representations for measure-valued population models. Ann. Probab., 27(1):166-205, 1999.

Matthias Birkner, Jochen Blath, Martin Möhle, Matthias Steinrücken, Johanna Tams. A modified lookdown construction for the Xi-Fleming-Viot process with mutation and populations with recurrent bottlenecks. ALEA Lat. Am. J. Probab. Math. Stat. 6 (2009), 25–61.

Exchangeable And Conditionally Poisson Processes

Thomas G. Kurtz

From the example of the lookdown construction of the Moran model, we see the central role that exchangeability plays. Each particle has an integer-valued "level", and the types and locations associated with the particles form an exchangeable sequence indexed by the levels. The de Finetti measure associated with the sequence then gives a measure-valued approximation to the empirical measure determined by the finite population model.

Similar constructions for a larger class of models can be given if we assign real-valued rather than integer-valued levels. For constructions of this type, the analog of exchangeability in the finite population case is that conditioned on the collection of types and locations, the levels are independent uniform random variables, and in the infinite population limit, the point process given by the set of type-location-level triples is a conditionally Poisson process with a Cox measure given as the product of a random measure on type/location space times Lebesgue measure on the nonnegative half-line. The random measure on type location space then gives the state of the measure-valued process that approximates the empirical measure determined by the finite population model.

References

Andreas Greven, Vlada Limic, and Anita Winter. Representation theorems for interacting Moran models, interacting Fisher-Wright diffusions and applications. Electron. J. Probab., 10:no. 39, 1286-1356 (electronic), 2005.

Thomas G. Kurtz and Eliane R. Rodrigues. Poisson representations of branching Markov and measure-valued branching processes. Ann. Probab., 39(3):939-984, 2011.

Amandine Veber and Anton Wakolbinger. The spatial lambda-Fleming-Viot process; an event-based construction and a lookdown representation. To Appear in Ann. Inst. H. Poincare Probab. Stat., 2013.

Alison M. Etheridge and Thomas G. Kurtz. Genealogical constructions of population models. Preprint http://arxiv.org/abs/1402.6724

Particle Representations For Stochastic Partial Differential Equations

Thomas G. Kurtz

Many stochastic partial differential equations arise as limits of finite particle models, and particle representations can be constructed for many of these. In a sense these constructions are lookdown constructions without the lookdowns. Natural examples include stochastic versions of McKean-Vlasov models. Not so natural examples include stochastic partial differential equations with boundary conditions. For many of these constructions, the empirical measures that give the solutions are weighted. Examples with weights include the classical filtering equations, and applications of the constructions include derivation of consistent numerical schemes.

References

Thomas G. Kurtz and Philip E. Protter. Weak convergence of stochastic integrals and differential equations. II. Infinite-dimensional case. In Probabilistic Models for Nonlinear Partial Differential Equations (Montecatini Terme, 1995), volume 1627 of Lecture Notes in Math., pages 197–285. Springer, Berlin, 1996.

Thomas G.Kurtz and Jie Xiong. Particle representations for a class of nonlinear SPDEs. Stochastic Process. Appl. 83 (1999), no. 1, 103–126.

Peter M. Kotelenez and Thomas G. Kurtz. Macroscopic limits for stochastic partial differential equations of McKean-Vlasov type. Probab. Theory Related Fields, 146(1-2): 189–222, 2010.

Einstein Relation And Homogenization Of Random Media

Nina Gantert

Many applications, such as porous media or composite materials, involve heterogeneous media which are modeled by random fields. These media are locally irregular but are "statistically homogeneous" in the sense that their law has homogeneity properties. Considering random motions (random walks or diffusions) in such a random medium, it turns out often that they can be described by their effective behaviour. This means that there is a deterministic medium, the effective medium, whose properties are close to the random medium, when measured on long space-time scales. In other words, the local irregularities of the random medium average out over large space-time scales, and the random motion is characterized by the "macroscopic" parameters of the effective medium. How do the macroscopic parameters depend on the law of the random medium?
As an example, we consider the effective diffusivity (i.e. the covariance matrix in the central limit theorem) of a random walk among random conductances. It is interesting and non-trivial to describe this diffusivity in terms of the law of the conductances. The Einstein relation gives a different interpretation of the effective diffusivity as mobility. The mobility measures the response of the diffusing particle to a constant exterior force: Consider the perturbed process obtained by imposing a constant drift of strength lambda in some fixed direction. The perturbed motion satisfies (as one can show in many
examples) a law
of large numbers with effective velocity v(lambda). The mobility is the derivative of v(lambda) as lambda goes to 0. The Einstein relation says that the mobility and the diffusivity of a particle coincide.
The Einstein relation is conjectured to hold for a variety of models, but it is proved insofar only for particular cases. We explain some of the ideas of the proof for reversible diffusions in random environment, random walks among random conductances and random walks on percolation clusters.

The talk is based on joint work with Pierre Mathieu and Andrey Piatnitski, and on work in progress with Noam Berger, Jan Nagel and Xiaoqin Guo.

Directed Polymer And Percolation Models On The Plane

Timo Seppalainen

The limiting shapes of percolation models and limiting free energies of polymer models are basic applications of subadditive ergodic theory. A long-standing challenge has been to find descriptions of these limits. This talk describes two types of variational formulas for directed last-passage percolation and directed polymer models. One kind of variational formula maximizes over measures, another type minimizes over stationary cocycles. For explicitly solvable models on the planar lattice, such as the corner growth model with exponential weights and the log-gamma polymer, the cocycles that solve the variational formula arise as Busemann functions which are limits of gradients of free energy. These cocycles can also be used to derive fluctuation exponents for the explicitly solvable models.

Regularity Structures

Martin Hairer

Many of the stochastic partial differential equations that arise naturally in the context of statistical physics are ill-posed: their solutions are so singular that it is not even clear what it means to be a solution to these equations. Several approaches have been developed over the years to tackle this problem.
The recently developed theory of regularity structures provides a unified framework which relates many of the previous results and allows to develop robust solution theories for a number of equations that were previously not amenable to mathematical analysis. In this talk, we will survey some of the main ideas of the theory and give a number of its applications to date.

Removing Unwanted Variation Using Controls

Terry Speed

About 15 years ago I became aware of the need for background adjustment and normalisation in the field of microarray gene expression analysis. Not long after this I saw my first batch effects, which is a term that describes many related phenomena that can affect our data, all undesirable. Dealing with these issues was what I used to call low-level analysis. It seemed to be necessary before we got on to the statistical analyses with which we are all familiar, which I called high level analysis. In the years since then, I have seen the same issues arise with other technologies that generate the data we analyse. I've changed my name for them to unwanted variation, and changed mind on how we should deal with them. These days I think the low-level analyses should be tailored to the high-level analyses that follow, though this sometimes causes problems as they are not specified in advance. In this talk I'll sketch this history, describe what I think we should now do, and in the process explain why I think statisticians should pay much more attention to controls.

False Discovery Rates - A New Deal

Matthew Stephens

False Discovery Rate (FDR) methodology, first put forward by Benjamini and Hochberg, and further developed by many authors - including Storey, Tibshirani, and Efron - is now one of the most widely used statistical methods in genomics, among other areas of application. A typical genomics workflow consists of i) estimating thousands of effects, and their associated p values; ii) feeding these p values to software (e.g. the widely used qvalue package) to estimate the FDR for any given significance threshold. In this talk we take a fresh look at this problem, and highlight two deficiencies of this standard pipeline that we believe could be improved. First, current methods, being based directly on p values (or z scores), fail to fully account for the fact that some measurements are more precise than others. Second, current methods assume that the least significant p values (those near 1) are all null – something that initially appears intuitive, but will not necessarily hold in practice. We suggest simple approaches to address both issues, and demonstrate the potential for these methods to increase the number of discoveries at a given FDR threshold. We also discuss the connection between this problem and shrinkage estimation, and problems involving sparsity more generally.

Statisticians In The Era Of Big Data

Bob Rodriguez

During 2013 we celebrated the International Year of Statistics, a global campaign whose goal was to raise the visibility of the field of statistics. This is also a time when the demand for statistical skills is unprecedented in areas of business, government where value, competitiveness, and efficiency are driven by an abundance of data – and increasingly Big Data. And yet the distinctive value of our professional contributions is often overlooked in emerging areas of practice where others, such as data scientists, are identified as those who learn from data.

This presentation considers the challenges of our journey into the area of Big Data. We can prepare for a successful journey by strengthening our professional participation, and by developing our skills for communication and leadership. The rewards of the journey are increased visibility for our field and greater roles for statisticians within their organisations.

Statistical Challenges In Genomic Discovery

Peter Donnelly

IMS Neyman Keynote Wednesday 9th July 2014 13:40-14:35pm

Biostatistician Behind Bars: By Design And On Trial

Sheila Bird

The UK's Medical Research Council Biostatistics Unit celebrates its centenary in 2014. Over its history, the Unit has tackled epidemics from tuberculosis to HIV, cigarette smoking to heroin injection, and their related causes of death. Heroin injectors are, of course, frequently incarcerated for acquisitive crimes and vulnerable to blood-borne viruses. I shall describe how a quarter century of surveillance designs (with associated biological sample), record-linkage studies, and bespoke "questionnaires" - a phrase coined from Hill and Doll - have improved prisoners' access to harm reduction (eg Hepatitis B immunization) , contributed to changed policy in prisons (eg put an end to random mandatory drugs testing), got us barred, yet quantified a 7 times higher risk of overdose death soon after prison-release, and eventually enabled three musketeers to mount the pilot N-ALIVE Trial in England, which tests whether those randomized to receive naloxone-on-release have 30% fewer opioid-related deaths in the 4-weeks post-release than controls (prior estimate: 1 in 200). Even before the N-ALIVE Trial's first randomization, however, Scotland became the first country to introduce take-home-naloxone as a funded public health policy. Wales followed in May 2011. I shall describe the trials and tribulations of Scotland's closely-monitored evaluation of its 2011-15 take-home naloxone policy, which is complicated because Scotland's policy was introduced against a still-rising trajectory of age-related opioid-deaths. I ask: is it impact for an RCT to be overtaken by the policy it seeks to inform? I also commend asking the right questions.

From Le Cam 1973 To Optimal Estimation

Harrison Zhou

Lucien Le Cam's ideas have had a great influence on the development of modern statistical theory. In this talk, I will survey two lines of recent research that grew out of the hypothesis testing idea of Le Cam (1973): minimax estimation and Bayesian posterior contraction. The convex hull testing method in Section 2 of Le Cam (1973) and its recent extensions have helped establish minimax lower bounds for estimation of large covariance and precision matrices. In particular, I will discuss rate-optimal estimation of Toeplitz, bandable, low rank and sparse covariance matrices, as well as structured Gaussian graphical models. I will also discuss applications of the ideas in Section 4 of Le Cam (1973) to some recent work on adaptive rate-optimal posterior contraction for nonparametric estimation and high dimensional models.

Rough Paths, Signatures, And The Modelling Of Functions On Streams

Terry Lyons

Rough path theory is focused on capturing and making precise the interactions between highly oscillatory and non-linear systems. The techniques draw particularly on the analysis of LC Young and the geometric algebra of KT Chen. The concepts and theorems, and the uniform estimates, have found widespread application; the first applications gave simplified proofs of basic questions from the large deviation theory and substantially extending Ito's theory of SDEs; the recent applications contribute to (Graham) automated recognition of Chinese handwriting and (Hairer) formulation of appropriate SPDEs to model randomly evolving interfaces.

At the heart of the mathematics is the challenge of describing a smooth but potentially highly oscillatory and vector valued path x_{t} parsimoniously so as to effectively predict the response of a nonlinear system such as $% dy_{t}=f(y_{t})dx_{t}$, $y_{0}=a$. The Signature is a homomorphism from the monoid of paths into the grouplike elements of a closed tensor algebra. It provides a graduated summary of the path x. Hambly and Lyons have shown that this non-commutative transform is faithful for paths of bounded variation up to appropriate null modifications. Among paths of bounded variation with given Signature there is always a unique shortest representative. These graduated summaries or features of a path are at the heart of the definition of a rough path; locally they remove the need to look at the fine structure of the path. Taylor's theorem explains how any smooth function can, locally, be expressed as a linear combination of certain special functions (monomials based at that point). Coordinate iterated integrals form a more subtle algebra of features that can describe a stream or path in an analogous way; they allow a definition of rough path and a natural linear "basis"\ for functions on streams that can be used for machine learning. The expected signature gives a description of measures on rough paths that is analogous to the martingale problem except that it does not require an a priori Markovian assumption.

‘measurement Of’ And ‘adjustment For’ Census Coverage, A Uk Perspective

James Brown

The UK has been measuring census coverage using surveys since 1971. However, the failure of the 1991 Census Validation Survey to estimate the level of under-coverage led to a fundamental change in the approach to measurement of census coverage. It also created a desire amongst users for the output database to be adjusted. In this paper we review briefly the 1991 history for context behind the 2001 Census. We describe the 2001 Census Coverage Survey in terms of its design and estimation strategy, which involves dual-system estimation at a local level with adjustments for dependence that is then combined with ratio estimation. We then cover the process used to adjust the full output database for the estimated coverage errors. We reflect on the lessons learnt from 2001 and discuss how these were integrated into the broadly successful coverage assessment and adjustment for the 2011 Censuses of England and Wales. □

Spatial Statistical Stream-network Models: Overview, Tools, And Data

Erin Peterson
Jay Ver Hoef[1]
[1]NOAA National Marine Mammal Laboratory Seattle

Streams and rivers form dendritic networks, which are connected by spatially and temporally variable directional flow. In addition, locations have a dual spatial representation; as points within the network and as points in geographical space. Consequently, some analytical methods used to quantify relationships in other types of networks, or in 2-D space, may be inadequate for studying the influence of structure and connectivity on biological and physical processes within streams. Recently, a new class of autocovariance functions has been developed for stream networks based on a moving-average construction. These models account for the dendritic structure of the network, flow volume, and flow direction. Furthermore, these models may be combined with traditional autocovariance models based on Euclidean distance to account for complex multi-scale patterns of spatial autocorrelation in streams data. An overview of spatial statistical stream-network models and a brief introduction to the SSN (SpatialStreamNetwork) package for R will be provided. Links to freely available streams datasets will also be provided, with the goal of encouraging other statisticians to develop additional methods that account for the unique spatial and spatio-temporal characteristics of stream networks.

Statistical Methods In Life Course Epidemiology

Gita Mishra

Life course epidemiology examines biological, behavioral and social pathways that link exposures during gestation, childhood, adolescence, and adult life, as well as across generations, to influence health and health inequalities in later life. The life course approach represents a relatively new concept in epidemiology that began to emerge in the 1980s and 1990s.

We discuss and provide illustrative examples of study designs, the current life course theoretical framework, and statistical methods that can test the hypotheses related to the main life course models. Relevant statistical methods include dynamic path models, marginal structural models, and nested regression models. In particular, we propose a model-building framework that can be used to formally compare alternative hypotheses on the effect of multiple exposure measurements collected across the life course. We describe how simple life course models can be applied to elaborate the independent, cumulative, and interactive effects of exposures. We highlight the assumptions underlying these models and discuss future directions for life course methodology.

Cluster Randomised Trials Across The Lifespan

Judy Simpson

Cluster randomised trials are becoming increasingly common in public health and are almost de rigueur in general practice research. As more researchers become aware of the need to take the clustering into account in the design and analysis, the quality of CRTs is improving but many challenges remain in terms of their design, implementation and analysis. Sample size calculations may be hampered by lack of good estimates of the intracluster correlation coefficient and sometimes difficulty in gauging the cluster size when not all members of the clusters will be eligible or willing to participate. Because the number of clusters is often small, achieving balance at baseline across treatment groups is difficult, so that methods such as pairing of clusters may be considered but have their disadvantages. There are often difficult ethical issues to consider, such as whether consent is required at both the cluster and individual level. Maintaining the intensity of the intervention and the participation of members of control clusters may also be problematic, but are essential to the success of the CRT. Nevertheless, loss of one or more clusters is not uncommon. These issues and more will be discussed and illustrated with examples from my own experience, ranging from preventing malnutrition in infants in Tanzania to preventing falls in the frail elderly in aged care hostels in Australia.

Ergodic And Mixing Properties Of The Boussinesq Equations With A Degenerate Random Forcing

Geordie Richards

The Boussinesq equations play an important role in the analysis of buoyancy driven fluid convection problems. We will discuss the existence, uniqueness and attraction properties of an ergodic invariant measure for the Boussinesq equations in the presence of a degenerate stochastic forcing acting only in the temperature equation and only at the largest spatial scales. Here the central challenge is to establish time-asymptotic smoothing properties of the Markovian dynamics corresponding to this system. Towards this aim we encounter a Lie bracket structure in the associated vector fields with a complicated dependence on solutions. This leads us to develop a novel Hörmander-type condition for infinite-dimensional systems. Demonstrating the sufficiency of this condition requires new techniques for the spectral analysis of the Malliavin covariance matrix.

Combining Graph-theoretic And Statistical Approaches To Explore Key Genes In Sea Urchin Development

Dario Strbenac
Jean Yang[1], Nicola Armstrong[1]
[1]University of Sydney

The key challenges of applying high-throughput sequencing technologies for time-series gene expression analysis to organisms without a quality genome is determining what genes were measured and the quantities of the inferred genes. To infer the sequences of transcripts, the Trinity algorithm, which employs traversal of a de Bruijn graph, is used. Once the inferred sequences are available, their abundance can also be estimated by expectation-maximisation. The RNA-seq by Expectation Maximization (RSEM) software was employed for this purpose. Sequencing reads which map to multiple sequences are probabilistically weighted amongst the locations. Comparison to experimentally validated abundances has previously shown this is a better approach than simply ignoring the multi-mapping sequence reads. Once both the objects measured and their abundances have been inferred, one biological point of interest is to cluster gene expression profiles. Through a simple simulation, we highlight the problems in clustering absolute values, rather than zero mean-centered profiles, over a large dynamic range. We find that Mahalanobis distance with standard clustering algorithms does not work well for this situation, and propose a new approach. The combination of techniques from computer science and statistical inference is successful for determining genes, and sets of genes, which are changing at key stages of sea urchin development.

Location, Location, Location: Econometric Analysis Of Asset Pricing With Spatial Interaction.

Xianhua Peng
Steven Kou, Haowen Zhong

Spatial interaction is well-known to be important in modeling real estate assets, as housing prices are significantly affected by neighborhood prices. Although spatial econometrics has been applied to empirical studies of housing markets, there is as yet little theoretical work that studies the risk and return of real estate securities. We attempt to fill this gap by proposing a spatial capital asset pricing model (S-CAPM) and a spatial arbitrage pricing theory (S-APT) that extend the classical asset pricing models by incorporating spatial interaction among asset returns. Furthermore, we give rigorous econometric analysis of the models by deriving identifiability conditions for the parameters and asymptotic properties of estimators and studying test statistics needed for implementing the models. Finally, an empirical study of the futures contracts on S&P/Case-Shiller Home Price Indices shows that the S-APT is not rejected and the spatial interaction parameter is statistically significant.

A Bayesian Hierarchical Model For Estimation Of Rare Mutation Levels In Sperm Using Deep Targeted Next Generation Sequencing

Eleni Giannoulatou

The RAS proto-oncogene Harvey rat sarcoma viral oncogene homolog (HRAS) encodes a small GTPase that transduces signals from cell surface receptors to intracellular effectors to control cellular behavior. Although somatic HRAS mutations have been described in many cancers, germline mutations cause Costello syndrome (CS), a congenital disorder associated with predisposition to malignancy. Based on the epidemiology of CS and the occurrence of HRAS mutations in spermatocytic seminoma, we proposed that activating HRAS mutations become enriched in sperm through a process akin to tumorigenesis, termed selfish spermatogonial selection. To test this hypothesis, we have developed a Bayesian hierarchical model to quantify the levels, in 7 blood and 89 sperm samples, of mutations occurring in the HRAS gene, using ultra deep targeted next generation sequencing data. The data were generated by an experimental protocol that combines restriction enzyme digestion, PCR amplification and massively parallel sequencing. Our model accounts for sequencing errors, noise of different sequencing libraries, and artifacts introduced during PCR rounds and digestion. For model inference we used a Metropolis-within-Gibbs sampling scheme. In order to assess the performance of our method, we estimated mutation levels in a titration-reconstruction experiment using biological replicates containing control blood DNA supplemented with dilution series of genomic DNA from four CS patients heterozygous for HRAS mutations. We found a good correlation between the amount of input DNA and the mutation levels estimated by our model. Our model was therefore used to quantify the mutation levels at the p.G12 codon, where causative mutations occur, and compared the results to changes at the p.A11 codon, at which activating mutations do not occur. Our results strongly support the role of selection in determining HRAS mutation levels in sperm, and hence the occurrence of CS.

Extending Approximate Bayesian Computation (abc) Methods To High Dimensions Using Gaussian Copula

Scott Sisson
Jingjing Li[1], David Nott[1], Yanan Fan[2]
[1]NUS
[2]UNSW

Approximate Bayesian computation (ABC) refers to a family of inference methods used in the Bayesian analysis of complex models. These methods do not rely on numerical evaluation of the likelihood function, and so are often used for models where likelihood evaluation is difficult. Conventional ABC methods suffer from the curse of dimensionality. A "marginal adjustment" was proposed by Nott et al.(2014) to improve performance in high dimensional analyses, however this suffered from poor estimation of the joint parameter distribution. In this talk I will introduce an extension to the marginal adjustment that better estimates the dependence between parameters in high dimensional models through the use of a Gaussian copula. This will be illustrated on an 85 dimensional parameter model, which is comfortably well beyond current ABC practice.

Do Early Life Experiences Matter And How Will We Know?

John Lynch

David Barker's seminal studies in the late 1980s are often recognized as sparking the rise of modern lifecourse epidemiology, although it's scientific roots can be traced back much earlier. Indeed, the core ideas behind lifecourse epidemiology have been around for centuries. Studies examining 'programming' of later life outcomes explicitly or implicitly during sensitive periods or during periods of developmental plasticity continue to appear in the literature. The core of lifecourse epidemiology lies in causally linking early life exposures with outcomes sometimes occurring many decades later, and understanding the mechanisms through which these effects are mediated. The 1980s also witnessed the rise of methodological advances in causal inference by scientists such as James Robins, Sander Greenland, and Donald Rubin. These methodological developments had implications for how to deal with complex exposure, confounding and outcome associations in a longitudinal data structure of repeated exposures and time dependent confounding – the classic data structure for lifecourse epidemiology. Methodological and empirical studies since the 1980s have highlighted the value of the potential outcomes model of causation, Directed Acyclic Graphs (DAGs), careful consideration of assessing direct and indirect effects, imputation of missing data, quantitative bias analysis, instrumental variables analysis and other techniques to improve causal inference. This presentation will trace some of the history of lifecourse epidemiology and it's achievements, and then reflect on some of the challenges that the crucial questions being posed by lifecourse epidemiology face within the context of methodological advances in causal inference.

Estimation For Single-index And Partially Linear Single-index Nonstationary Models

Chaohua Dong

Estimation in two classes of popular models, single-index models and partially linear single-index models, is studied in this paper. Such models feature nonstationarity. Orthogonal series expansion is used to approximate the unknown integrable link function in the models and a profile approach is used to derive the estimators. The findings include dual convergence rates of the estimators for the single-index models and a trio of convergence rates for the partially linear single-index models. More precisely, the estimators for single-index model converge along the direction of the true parameter vector at rate of $n^{-1/4}$, while at rate of $n^{-3/4}$ along all directions orthogonal to the true parameter vector; on the other hand, the estimators of the index vector for the partially single-index model retain the dual convergence rates as in the single-index model but the estimators of the coefficients in the linear part of the model possess rate n^{-1}. Monte Carlo simulation verifies these theoretical results. An empirical study on the dataset of aggregate disposable income, consumption, investment and real interest rate in the United States between 1960:1-2009:3 furnishes an application of the proposed estimation procedures in practice.

Stat Wars Episode Vi: Return Of The Fiducialist

Keli Liu
Xiao-Li Meng[1]
[1]Harvard University

Priors are the path to the dark side. Fisher developed the Fiducial argument to obtain prior free posterior inferences. But his solutions were restricted to special examples. Attempts at generalization often lead to posterior inferences lacking confidence calibration, the heart of the Fiducial promise. Without it, what claim has Fiducial to objectivity? If one lusts after posteriors, a turn to the dark side would seem inevitable. Can fiduciary responsibility be restored to this inferential force? The missing data perspective lays bare the mechanism by which the Fiducial method buffers posterior inferences from prior influence: a *two-phase* structure. Importantly, the prior for the parameter plays a role only in Phase I. The Fiducial method exploits the separation of phases to minimize the leaking of prior influence from Phase I to Phase II. We can truthfully say, "The prior is weak with this one." This insensitivity is achieved by a very special mode of information passing between phases. The multi-phase model thereby illuminates how posterior inferences under the Fiducial construction continue to be robust to prior misspecification even when they cease to supply exact Frequentist answers. As Skywalker redeemed Vader, so we say, "There is still good in it. We can save it. We can turn Fiducial back to the good side. We have to try." Let the Fiducial order be reborn.

[*Episode IV: A New Hope (For Objective Inference)* and *Episode V: Ancillarity Paradoxes Strike Back (At Fiducial)* played respectively at the ICSA Applied Symposium in Portland (June 15-18, 2014) and the IMS Asia Pacific Ram Meeting in Taipei (June 30-July 3, 2014).]

Detecting Trends In Time Series Of Functional Data: A Study Of Antarctic Climate Change

Regina Liu

The Spanish Antarctic Station Juan Carlos I has been registering surface air temperatures with the frequency of one reading per ten minutes since the austral summer 1987-88. Although this data set contains valuable information about the climate patterns in and around Antarctica, it has not been utilized in any existing climate studies

thus far. This is due to the concern of the substantial missing data caused by the difficulty in collecting data in the extreme winter weather conditions there. Such data sets do not fit the standard setting covered by the existing times series techniques. However, by treating the temperature readings for each summer as a function, the temperature data can be viewed as a time series of functional data. We introduce new notions of trends for general time series of functional data based on the so-called record functions, and also develop useful nonparametric tests for such trends. Following our analysis, the data collected from Juan Carlos I Station exhibit an increasing trend in the Antarctic

temperature.

Spatio-temporal Modelling Of Pollutant Loads In Great Barrier Reef Catchments

Dan Gladish

Physical processes can typically be quite complex, exhibiting dependences in space and time. Difficulties arise in modelling such processes due to uncertainties associated with observations, high dimensionality of the data, and the general dynamics of the system. One such process is pollutant load runoff from catchments into the Great Barrier Reef (GBR) lagoon. Due to the potential ecological impact pollutant runoff has on the GBR, it is critical to develop statistical models that accurately quantify sediment loads. However, observational data are sparse due to difficulties in monitoring catchments. Further complications arise with large changes in magnitude of water flow. We use the Bayesian hierarchical modelling framework utilizing the two-tiered dimension reduction approach in space and time as a basis for our modelling procedure in assimilating data and sediment concentration, erosion, and flow processes from catchments into the GBR lagoon. The result is a model that provides a level of confidence in the estimation of pollutant loads and exceedance probabilities that highlight problem areas in the catchment where loads are persistently high.

The New Zealand Health Survey – An Ongoing National Survey That Can Be Linked To Health System Records

Deepa Weerasekera

The NZ Health Survey has been recently re-designed so that it is in continuous operation, with content that consists of a core set of questions supplemented by an evolving set of topic modules. Explicit consent is requested from respondents to allow linkage, at an individual level, to routinely collected health records. The routinely collected data includes records of events that occur in hospitals, prescriptions of pharmaceuticals, laboratory tests, mortality records etc.

The positioning of an on-going survey at the centre of a comprehensive health information system, allows us to assess quality of health service use data collected by recall in the survey, to assess quality of the demographic data in the administrative collections, to re-prioritise the content of the survey and to develop new analyses based on combining the data records.

Efficiency Transfer For Regression Models With Missing Responses

Ursula U Muller

An accepted way of analyzing missing data is by imputingthe missing values. This often gives better results thana "complete case analysis", which is the fastest andsimplest method of dealing with missing data since it usesonly cases that are completely observed. Although this mayseem to be a wasteful approach, there are in fact manysituations where a complete case analysis turns out to be(asymptotically) optimal and should therefore be used. In this talk I focus on i.i.d. data (X,Y), where the responseY is missing at random and where the covariate vector X isalways observed. I demonstrate that general functionals ofthe conditional distribution of Y given X can be estimatedefficiently by a complete case version of an efficientestimator. This is a very general and useful result sinceit tells us that in such situations we can simply omitincomplete cases and work with some familiar efficientestimator without losing efficiency. Our result applies to homoscedastic and heteroscedasticregression, with the conditional expectation of Y given Xbeing modeled by a general semiparametric regression functionthat involves a finite and an infinite-dimensional parameter.This includes the fundamental linear and nonparametricregression model, but also more complex models, e.g.partially linear additive regression and the single-indexrandom coefficient model. I will discuss estimation of variousfunctionals of the conditional distribution, e.g. ofregression parameters and of the error distribution.

Looking Better: Using Images To Improve Images

Kerrie Mengersen
Matthew Moores[1], Catriona Hargrave[1], Fiona Harden[1]
[1]QUT

Consider a patient who is due to undergo daily radiotherapy. At commencement, a fan-beam CT may be used to establish a treatment plan, and a cone-beam CT is then used each day to determine any changes in the image and hence in the plan. Whereas the fan-beam CT is less subject to artifacts induced by X-ray scatter or metal implants, the cone-beam CT can be undertaken *in situ*. At present, all images are separately and manually inspected.

This presentation examines ways of combining the two technologies. More generally, it suggests a method for automatically updating the use of spatial prior information to improve identification of features in images with low contrast-to-noise ratio. The prior is represented as an external field in a hidden Potts model of the image lattice. In particular, the prior distribution of the latent pixel labels is a mixture of Gaussian fields, centred on the positions of the objects in a previous point in time. The model is thus particularly applicable in longitudinal imaging studies, where the manual segmentation of one image can be used as a prior for automatic segmentation of subsequent images. Substantive computational issues and solutions, notably for sampling the temperature parameter in the Potts model, are also described. In addition to the automated feature of the approach, we show that the external field prior results in a substantial improvement in segmentation accuracy, reducing the mean pixel misclassification rate on test images from 87% to 6%.

The presentation will also touch briefly on our group's encounters with other methods for analysing images and their application to real problems.

Stochastic Simulation Of Biological Regulatory Networks: From Stochastics Via Time Delay To Memory

Tianhai Tian

Stochastic modelling and simulation is a very important research area in systems biology. The stochastic simulation algorithm (SSA) is a powerful numerical method to describe the intrinsic noise in molecular systems. Due to the complexity of biological networks, more sophisticated modelling methodologies are still needed to describe various biological behaviours. Among them, memory is a ubiquitous phenomenon in biological systems in which the present system state is not entirely determined by the current conditions but also depends on the time evolutionary path of the system. These conditional chemical reactions contradict to the extant stochastic approaches for modeling chemical kinetics and have increasingly posed significant challenges to stochastic modelling. To tackle the challenge, the delay-SSA and memory-SSA have been designed to study the effects of multi-step reactions and conditional chemical reactions, respectively, in an efficient way. This talk will discuss these three numerical methods and demonstrate how to use the memory-SSA to realize bursting gene expression that has been observed in single cells recently.

Bayesian Methods For Density And Regression Deconvolution

Raymond Carroll

Deconvoluting kernel methods for density and regression deconvolution have undergone vigorous development in the past 10 years. I will review recent work on flexible Bayesian methods for these problems, including the multivariate case and multivariate problems with excess zeros.

Total Quality Management In Teaching An Introductory Statistics Course: Know Your Customer

Peter Howley

Teaching a compulsory introductory statistics service course to primarily non-quantitative and relatively disinterested Business students presents serious challenges. Key among these are how to appeal to such students and win them over and finding ways to assist students to unlock their abilities to understand key concepts and core principles. The key elements of Total Quality Management - such as systems thinking, process focus, know your customer, systematic methods - as well as scaffolding, the zone of proximal development, constructive alignment and cognitive load theory were utilised in the redesign of a first year service course to Business students following continued undesired outcomes such as high failure rates (25% to 40%) and poor Student Feedback (3.5 to 3.8 on a 5-point Likert Scale). The newly-designed course was delivered in Semester 2, 2013, with a remarkable turnaround - Failure rate of 7% and average Student Feedback at 4.7. Unsolicited student emails after the course confirmed this success (e.g., "... the course was actually enjoyable and really well structured, it was my favourite course this semester to my surprise!" and "...without your methods I would not have been able to attain that elusive HD...".) The presentation will describe the structure and teaching methods used in achieving the improved results and provide examples of the application of the methods used towards reducing students' resistance, including efforts at being appealing, enthusing students and generating interest, facilitating students' self-learning, and assisting recall and understanding.

How Do We Create The Next Generation Of Statisticians?

Michael Martin
Peter Howley[1]
[1]School of Mathematical and Physical Sciences/Statistics, The University of Newcastle

Statistical education faces many challenges. How do we as a society hope to engender increased interest in the area? What activity would you like to see from your Society's Statistics Education Section?
Over to you! The co-Chairs of the SSAI's Statistics Education Section invite you to attend and discuss desires and ideas for future Statistics Education Section activity. Prof Michael Martin will chair the session, list activities and ideas to-date and seek your thoughts on where to from here.

Joint Clustering And Matching Of Multivariate Samples Across Objects

Geoffrey John McLachlan

We present a widely applicable mixture model-based framework for the simultaneous clustering of multivariate samples observed on objects in a class and the matching of the object clusters. This statistical approach obviates the post-clustering need to match the object clusters since the matching is done during the fitting of the overall mixture model, which can be used as a template for the class distribution.

It thus provides a basis for discriminating between different classes in addition to the identification of anomalous events within a sample and a class.

Key applications include image analysis and the automated analysis of data in flow cytometry.

Recent Advances In Nearest-neighbour Classification

Richard Samworth
Timothy Cannings[1]
[1]University of Cambridge

Nearest neighbour classifiers are perhaps the simplest and most intuitively appealing nonparametric classifiers. We will present recent results on the asymptotically optimal weighting scheme for a weighted nearest neighbour classifier, as well as the attainable improvement over the unweighted k-nearest neighbour classifier. Interestingly, asymptotically this depends only on the dimension of the feature vectors. We will also discuss the error rate in the tails of the feature vector distribution and ongoing work on high-dimensional extensions.

Don't Fall For Tuning Parameters

Johannes Lederer

Lasso is a seminal contribution to high-dimensional statistics, but it hinges on a tuning parameter that is difficult to calibrate in practice. A partial remedy for this problem is Square-Root Lasso, because it inherently calibrates to the noise variance. However, Square-Root Lasso still requires the calibration of a tuning parameter to all other aspects of the model. In this talk, we discuss TREX, an alternative to Lasso with an inherent calibration to all aspects of the model. This adaptation to the entire model renders TREX an estimator that does not require any calibration of tuning parameters. We show that TREX can outperform cross-validated Lasso in terms of variable selection and computational efficiency. We also introduce a bootstrapped version of TREX that can further improve variable selection. We illustrate the promising performance of TREX both on synthetic data and on recent high-dimensional biological data sets.

Maximum Smoothed Likelihood Component Density Estimation In Mixture Models With Known Mixing Proportions

Tao Yu
Pengfei Li[1], Jing Qin[2]
[1]University of Waterloo
[2]National Institute of Allergy and Infectious Diseases, National Institutes of Health

In this paper, we propose a maximum smoothed likelihood method to estimate the component density functions of mixture models, in which the mixing proportions are known and may differ among observations. The proposed estimates maximize a smoothed log likelihood function and inherit all the important properties of probabili- ty density functions. A majorization-minimization algorithm is suggested to compute the proposed estimates numerically. In theory, we show that starting from any initial value, this algorithm increases the smoothed likelihood function and further leads to estimates that maximize the smoothed likelihood function. This indicates the conver- gence of the algorithm. Furthermore, we theoretically establish the L1 consistency of our proposed estimators. An adaptive procedure is suggested to choose the bandwidth- s in our estimation procedure. Simulation studies show that the proposed method is more efficient than the existing method in terms of integrated square errors. A real data example is further analyzed.

Principal Flows

Tung Pham

We revisit the problem of extending the notion of principal component analysis (PCA) to multivariate data sets

that satisfy non-linear constraints, therefore lying on Riemannian manifolds. Our aim is to determine curves on

the manifold that retain their canonical interpretability as principal components, while at the same time being

flexible enough to capture non-geodesic forms of variation. We introduce the concept of a principal flow, a curve

on the manifold passing through the mean of the data, and with the property that, at any point of the curve, the

tangent velocity vector attempts to fit the first eigenvector of a tangent space PCA locally at that same point, subject

to a smoothness constraint. That is, a particle flowing along the principal flow attempts to move along a path of

maximal variation of the data, up to smoothness constraints. The rigorous definition of a principal flow is given by

means of a Lagrangian variational problem, and its solution is reduced to an ODE problem via the Euler-Lagrange
method.

On The Prediction Performance Of The Lasso

Johannes Lederer[1]

[1]Cornell University

Although the Lasso has been extensively studied, the relationship between its prediction performance and the correlations of the covariates is not fully understood. We show that the incorporation of a simple correlation measure into the tuning parameter leads to a nearly optimal prediction performance of the Lasso even for highly correlated covariates. However, we also reveal that for moderately correlated covariates, the prediction performance of the Lasso can be mediocre irrespective of the choice of the tuning parameter.

Becoming An Editor

Louise Ryan

The editors perspective - publishing, reviewing and joining the ranks (panel discussion)

Getting Published

Michelle Haynes

The editors perspective - publishing, reviewing and joining the ranks (panel discussion)

Addressing The Analytical Skill Shortage Via The Cloud

James Enoch
Olivera Marjanovic

The way we educate students is changing rapidly, while coinciding with a time where industry is seeking significant growth in analytically trained talent. This presentation looks into the opportunities and challenges of teaching students analytics via the cloud, focusing on a current example at the University of Sydney where the use of cloud hosted solutions and real-life like stories are helping to educate students within the Business School, focusing on the business rather than IT or data science perspectives. This presentation also offers an example of new types of value networks being formed among university and industry, to enable industry-relevant, up-to-date and an educationally-sound model of teaching necessary for the-yet-to-be discovered future of business analytics.

Growing Random Trees, Maps, And Squarings.

Louigi Addario-Berry
Nicholas Leavitt

We use a growth procedure for binary trees due to Luczak and Winkler, a bijection between binary trees and irreducible quadrangulations of the hexagon due to Fusy, Poulalhon and Schaeffer, and the classical angular mapping between quadrangulations and maps, to define a growth procedure for maps. The growth procedure is local, in that every map is obtained from its predecessor by an operation that only modifies vertices lying on a common face with some fixed vertex. The sequence of maps has an almost sure limit G; we show that G is the distributional local limit of large, uniformly random 3-connected graphs.

A classical result of Brooks, Smith, Stone and Tutte associates squarings of rectangles to edge-rooted planar graphs. Our map growth procedure induces a growing sequence of squarings, which we show has an almost sure limit: an infinite squaring of a finite rectangle, which almost surely has a unique point of accumulation. We know almost nothing about the limit, but it should be in some way related to "Liouville quantum gravity".

Parts joint with Nicholas Leavitt.