1998

THIRTY-FOURTH GREGYNOG STATISTICAL CONFERENCE PROGRAMME

All talks will take place in Seminar Room 1 (Floor 2, far end)

Friday
24 April

16.00 *Tea*
17.00 Professor Ray Chambers (Southampton)
Likelihood - based inference for complex survey data I (3 lecture series)
19.00 *Dinner*
20.00 Dr Frank C Duckworth (Editor, RSS News)
One-day cricket: a mathematical solution to a mathematical problem

Saturday
25 April

08.00 *Breakfast*
09.30 Dr Malcolm Faddy (University of Queensland, visiting Open University)
Extended Poisson process modelling and analysis of count data
11.00 *Coffee*
11.30 Professor Ray Chambers
Likelihood - based inference for complex survey data II
13.00 *Lunch*

Afternoon free (walks, etc)

16.00 *Tea*
17.00 Professor Peter Green (Bristol)
Exact sampling from a continuous state space

Sunday
26 April

08.00 *Breakfast*
09.30 Professor Anatoly Zhigljavsky (Cardiff)
Analysis and forecasting of time series based on their principal component expansions
11.00 *Coffee*
11.30 Professor Ray Chambers
Likelihood - based inference for complex survey data III
13.00 *Lunch*
14.00 Professor Gad Nathan (Hebrew University, Jerusalem, visiting LSE)
Multilevel modelling for longitudinal analysis of complex survey data
15.30 *Tea and finish*

GREGYNOG STATISTICS CONFERENCE 1998
PARTICIPANTS

SPEAKERS

Professor Ray Chambers (University of Southampton)
Dr Frank C. Duckworth (Editor, *RSS News*)
Dr Malcolm Faddy (University of Queensland, visiting Open University)
Professor Peter Green (University of Bristol)
Professor Gad Nathan (Hebrew University, Jerusalem, visiting LSE)
Professor Anatoly Zhigljavsky (University of Wales, Cardiff)

STAFF **STUDENTS**

Aberystwyth

Dr John Basterfield
Mr Alan Jones
Dr John Lane
Miss Sylvia Lutkins
Professor Dennis Lindley

Bangor

Mr Chris Whitaker

Birmingham

Professor Tony Lawrance
Dr R A Atkinson
Mr R L Holder
Professor Henry Daniels
Mr David Kinns
Mr Muhammed Ali

Cardiff

Mr T C Iles
Dr S K Sahu

Swansea

Professor Alan Hawkes
Dr Mark Kelbert
Mr Alan Sykes
Dr John Pemberton

University of Wales College of Medicine

Dr Frank Dunstan

Warwick

Professor John Copas
Dr D Taneva
Dr J Q Shi
Mr P Corbett
Mr R Motsoeneng
Miss M Al-Awadi
Miss Z Mohl-khalid
Miss S Myonaki
Dr M Pucci
Miss V Teng
Mrs N W Yeh
Mr Yeh
Mr C K Tze

Gregynog Statistics Conference 1998 Abstracts

Likelihood-Based Inference for Complex Survey Data

Ray Chambers (Department of Social Statistics, University of Southampton)

The traditional approach to analysis of sample survey data is to focus on inference about finite population parameters. These are functions of the actual population values, for example the average value of a survey variable over the population, or the ratio of two such averages. In many cases, however, we are not interested in analysing sample survey data in this way. Rather, we view the sample data as providing a "window" on the underlying stochastic process that gave rise to the population values in the first place. Our interest then is in using the survey data to make inferences about the parameters of a statistical model for this "superpopulation". Within mainstream statistics, the method of maximum likelihood is probably the most widely used tool for such "analytic inference". However, the standard assumptions underlying many ML methods are typically inappropriate when the data are obtained via a sample survey with a complex design and with the usual complications (nonresponse, measurement error) that accompany a survey-based data collection methodology.

In the first of three lectures making up my contribution to the Gregynog Conference, I will outline a general methodology for ML inference with sample survey data, and compare this approach with the "pseudo-likelihood" and "complete data likelihood" methods of inference that have been suggested for data of this type. In particular, I will discuss the results presented in Breckling et al (1994), including the expressions for the ML score and information functions given in that reference, and then show how these apply to a variety of complex sample survey scenarios. Application to regression analysis with complex survey data will be briefly discussed, including the role of survey weights.

In my second lecture I will look at the important problem where limited information is available for this inference, essentially consisting of the sample data together with first order sample inclusion probabilities for the sample units, and where the sampling process is informative, in the sense that these inclusion probabilities potentially depend on the analysis variables of interest. Results obtained in Chambers, Dorfman and Wang (1998) will be used to show that, although strict maximum likelihood appears analytically impossible in this case, a close approximation seems possible.

Finally, in my third lecture I will discuss recent research on maximum likelihood under informative sampling. The Breckling et al (1994) approach outlined in lecture 1 is based on application of the Missing Information Principle (MIP; Orchard and Woodbury, 1972), and does not specifically model the distribution of the sample data. Consequently, it is not obvious that maximum likelihood based on it leads to exactly the same inference as a more direct approach which builds upon this sample distribution. In order to throw some light on this issue, I shall demonstrate the equivalence of "MIP-based" and "sample-based" approaches to maximum likelihood for data obtained via a particular method of informative sampling called array sampling.

Exact sampling from a continuous state space
Peter Green (Bristol, UK)

Propp and Wilson (Random structures and Algorithms, 1996) described a protocol, called coupling from the past, that allows us to organise a Markov chain Monte Carlo simulation so that it yields EXACTLY a sample from its limiting distribution (after a random but finite time). Current applications of this idea are to large but discrete physical systems: what possibilities does this idea open up for Bayesian computation? Potentially, this could completely solve the problem of deciding if your MCMC method has converged.

I will present methods for extending coupling from the past to various MCMC samplers on a continuous state space; rather than following the monotone sampling device of Propp and Wilson, our approach uses methods related to gamma-coupling and rejection sampling to simulate the chain, and direct accounting of sample paths. I will touch on the possibilities for automating the process to avoid the cumbersome algebra currently needed. This is joint work with Duncan Murdoch (Queens University, Canada)

Analysis and Forecasting of Time Series Based on their Principal Component Expansions
Anatoly Zhigljavsky (University of Wales, Cardiff)

A non-parametric technique of time series analysis and forecast is described and illustrated on different data sets. It is based on the, principal component analysis applied to a 'delay matrix' constructed from the original time series. The main objective is to get an expansion of the time series into a sum of a small number of 'independent' and 'interpretable' components. Quadratic trend, seasonalities and other harmonic components with perhaps variable amplitude are typical examples of these components. Stationarity of the time series is not assumed. The technique can be used for multivariate time series and even for image processing.

One-day Cricket: a Mathematical Solution to a Mathematical Problem

by Frank Duckworth, Editor *RSS NEWS*

One-day cricket has always been intolerant of rain because, unlike in the normal game, a result *must* be produced even if the match is shortened after it has started. Various methods of setting revised targets in shortened games have been used in the past and these have often produced a gross injustice, usually to the side batting first.

The year 1997 saw the introduction of the 'Duckworth/Lewis' method for providing a revised target which maintains the balance of the game as it was when play was suspended. This is based on a simple mathematical formula which relates the further runs that may be scored in an innings to the two run-scoring 'resources' possessed by the batting side, the overs remaining and the wickets in hand.

The method is now used by the England & Wales Cricket Board for all its one-day competitions including home internationals, and also by Zimbabwe and New Zealand, and by the ICC (International Cricket Council) for some of their international competitions.

Despite the enormous number of one-day matches that have been played throughout the world since the 1960s, very little over-by-over data are available in readily accessible form. Hence the parameters of the formula used for the application of the D/L method so far have been based on a superficial analysis of the data. However, systems have now been set up to record detailed match data on disk and a useful database is growing rapidly. The methods by which the parameters should be refined as data accumulate will provide an interesting challenge to statisticians.

Multilevel modelling for longitudinal analysis of complex survey data
Professor Gad Nathan (Hebrew University, Jerusalem visiting LSE)

Recently there has been increasing interest in the longitudinal analysis of data from complex panel sample surveys. These surveys are characterised by repeated observations for the same units on characteristics that change over time. The interest is in measuring the dynamics of the changes (e.g., gross flows). The situation is complicated by the fact that the population may be hierarchically structured (e.g., persons within households within localities) and that the sampling design may be complex. (e.g., varying selection probabilities). An example of this situation is the longitudinal analysis (rather than cross-sectional analysis) of Labour Force Surveys. These surveys are carried out, in many countries, as rotating panel surveys of complex design.

As background, we shall examine the effects of each of these elements on the analysis separately and in combinations. Thus multilevel modelling is often used to model complex hierarchical population structures for cross-sectional analyses. This is a general class of mixed linear models, which takes into account the hierarchical population structure. The models allow fixed affects at different levels, as well as random effects, in a way which enables the modelling of non-independent error structures. If the structure of the covariance matrix is known, Generalised Least Squares estimation can be used to estimate the model parameters. In general, the covariance matrix will not be known and Iterative Generalised Least Squares estimation will be required.

When multistage sampling with unequal selection probabilities is used, standard estimation methods are only appropriate if all the sampling effects are included as covariates in the model. If this is not the case, or if information is not available on all relevant sample design variables, unweighted estimation may lead to biases. A possible solution is based on 'pseudomaximum likelihood' estimation. This considers the 'census' likelihood, based on the data for the whole population, were they available. If this likelihood can be estimated consistently from the survey data, then the values that maximise this estimate can be shown to be consistent estimators of the model parameters, under fairly general conditions. For multilevel models this has to be adapted by first obtaining standard iterative generalised least square census estimators of the model parameters and then replacing them by their weighted sample estimators. This can be done if components of the census estimators can be expressed as sums over levels of squares and products of the variables.

In order to deal with the longitudinal aspect of the surveys, state-space models on the individual level can be used. These model separately the observations in terms of unobserved 'state' variables, which are time dependent. The state variables are then modelled by an appropriate time-series model. Thus we consider a multilevel mixed effects model, where the random effects parameters are themselves treated as time dependent random variables. These can then be modelled, say, by first-order autoregression models. By using many similar time series models for short periods, rather than a small number of long-term series, the model parameters can be estimated. Thus, Kalman filtering and appropriate iterative generalised least square estimation can be used jointly to estimate the model parameters.

Finally, the various elements – the multilevel model, the state space model and the weighted estimation are to be combined in order to produce consistent estimates of the model parameters and in order to allow the testing of alternative models. In order to do this, 'census' likelihoods for the combined state-space model must be expressed in terms of sums of squares and products over levels and over time periods. These will be estimated by appropriate weighting. Initial attempts to set-up suitable models and the estimation of their parameters will be described. If available, we shall present preliminary results of their application to simulated data and to empirical data from the Israel Labour Force Survey.