TE WHARE WĀNANGA O TE ŪPOKO O TE IKA A MĀUI
VICTORIA
UNIVERSITY OF WELLINGTON

Third Wellington Workshop in Probability and Mathematical Statistics

28 – 29 November 2011
Victoria University of Wellington

Presenters, Titles and Abstracts

(Ordered alphabetically, by presenters' last names)

History-dependent Volcanic Eruptions and Triggering by Earthquakes

Mark **Bebbington**
Volcanic Risk Solutions, Massey University

We show how a stochastic (point process) version of a general load-and-discharge model for volcanic eruptions can be implemented. Long-time DVJ observers may notice some familiar elements, but as they say, "lesser artists borrow, great artists steal". The model tracks the history of the volcano through a quantity proportional to stored magma volume. Thus large eruptions can influence the activity rate for a considerable time following, rather than only the next repose as in the time-predictable model. Applied to flank eruptions of Mount Etna, it exhibits possible long-term quasi-cyclic behavior, and to Mauna Loa, a long-term decrease in activity. When augmented by a time-, distance-, and magnitude-dependent triggering term, this enables us to detect triggering of eruption onsets by earthquakes over varying temporal and spatial scales. This effect is not detectable when using a renewal process as the basic model. We apply this to volcanic eruptions (VEI 2+) and great earthquakes (Mw 7+) in the Indonesian arc since 1900. Of 35 volcanoes with at least three eruptions in the study region, seven (Marapi, Talang, Krakatau, Slamet, Ebulobo, Lewotobi and Ruang) show statistical evidence of triggering over varying temporal and spatial scales.

Sensitivity Problems for Boundary Crossing

Konstantin **Borovkov**
University of Melbourne
(Joint work with A.N. Downes and A.A. Novikov)

Computing the probability for a given diffusion process to stay under a particular boundary is crucial in many important applications including pricing financial barrier options. It is a rather tedious task that, in the general case, requires the use of some approximation methodology. One possible approach to this problem is to approximate given (general curvilinear) boundaries with some other boundaries of a form enabling one to relatively easily compute the boundary crossing probability. In our talk, we discuss the problem of the sensitivity of the boundary crossing probabilities with respect to small perturbations of the boundaries, both in the case of the Brownian motion process and that of general time-homogeneous diffusions.

Point Processes: Book-Writing and a 'Discontinuity'

Daryl J. **Daley**
University of Melbourne and Australian National University

I shall reflect on my period of experience working with David Vere-Jones spanning about forty years. I then describe a family of point-process models exhibiting within their range of definition a limiting point-discontinuity property. The work is based on a formula from our writing period giving the variance of $N(0,t]$, the number of points of a stationary orderly point process N on a half-open interval of length $t > 0$, in terms of the Palm expectation function, loosely $U(t) = E(N(0,t] \mid \text{point at } 0)$, namely

$$\operatorname{var} N(0,t] - E(N(0,t]) = 2 \int_0^t [U(u) - \lambda u]\, \lambda \, \mathrm{d}u,$$

where $\lambda = E(N(0,1])$. Then when the integrand here has a limit for large u, as holds for a renewal process with generic lifetime X (and the limit equals $\frac{1}{2} \operatorname{var} \lambda X$ when X has finite second moment), we can conclude that the variance is asymptotically like $[\operatorname{var} \lambda X]\lambda t$. With a finite third moment for X, we can find the constants for this right-hand side to be equal to $At + B + o(1)$.

Mimicking Self-similar Markov Martingales

Kais **Hamza**
Monash University
(Joint work with Fima Klebaner and Jie Yen Fan)

While it is well known that stochastic processes are not uniquely defined by their one-dimensional marginals, the question of how to construct martingales with given marginals has recently become the focus of a substantial body of work. In this talk we will review some of the constructions found in the literature. In particular, we construct a family of Markov martingales having the marginal distributions of a self-similar Markov martingale. This construction relies on the self-similarity property of the processes we "mimic". We present two approaches to the construction, a transition-randomisation approach and a time-scale-change approach. We compute the infinitesimal generators and obtain some path properties of these processes. We also give some examples.

Relationships Between the Discrete and Continuous Time Hidden Markov Models

David **Harte**
Statistics Research Associates Ltd.
(Joint work with Mark Bebbington)

We compare two hidden Markov models. The first is a simple model, denoted here by HMM, in the class described by MacDonald & Zucchini (1997). It has an exponential random outcome that is observed at each time point. The rate parameter is determined by the current (discrete) state of a discrete time Markov chain. The second is the Markov modulated Poisson process (MMPP) described by Ryden (1996), whose exponential interevent times parallel the observations in the HMM.

Our empirical studies indicate (our conjecture) that the estimated exponential rates in the HMM case converge asymptotically to the eigenvalues of the $Q - \Lambda$ matrix in the MMPP case. Our empirical studies also indicate that, given a sample realisation from the MMPP case, one cannot determine statistically whether this sample sequence is drawn from the HMM or MMPP cases based on the respective likelihood functions. This can be shown to follow from our conjecture.

Warranty Models with Non-zero Repair Time

Yu **Hayakawa**
Waseda University
(Joint work with Stefanka Chukova)

This talk revisits modelling of warranty servicing costs assuming perfect repairs with non-zero repair times. For both non-renewing and renewing free replacement warranty policies, the alternating renewal process is used to model operating and repair times. We present some theoretical results on these processes in the case of a finite horizon and use them to derive the expected servicing cost over the product warranty coverage as well as the product life cycle. Numerical examples illustrate the ideas.

The Survival Probability for Statistical-Physical Models in High Dimensions

Mark **Holmes**
University of Auckland

The probability that a critical Galton-Watson branching process survives until time n is asymptotic to C/n, where C is a known constant that depends on the variance of the offspring distribution.

We'll discuss more challenging models in statistical physics (including some models for the spread of infection in space and time) where some spatial interaction means the elements of the population do not have independent offspring. We'll discuss joint work with Remco van der Hofstad in which we prove that the survival probability for each of these models (in high dimensions and at criticality) is asymptotically C/n, where C is some model-dependent constant.

If time permits, we will also briefly discuss how this result (when combined with other results in the literature) verifies a weak version of the "central limit theorem" for these measure-valued processes.

Markov Chain Properties in Terms of Column Sums of the Transition Matrix

Jeff **Hunter**
Auckland University of Technology

Questions are posed regarding the influence that the column sums of the transition probabilities of a stochastic matrix (with row sums all one) have on the stationary distribution, the mean first passage times and the Kemeny constant of the associated irreducible discrete time Markov chain. Some new relationships, including some inequalities, and partial answers to the questions, are given using a special generalized matrix inverse that has not previously been considered in the literature on Markov chains.

Iterative Methods in Model Fitting and Diagnostics

Murray **Jorgensen**
University of Waikato

Behind much software for the fitting of statistical models lie iterative methods such as Newton-Raphson, Fisher Scoring, the EM algorithm and iteratively re-weighted least squares. By an *iteration* I will mean a function $g : \Theta \to \Theta$, where $\Theta \subseteq \mathbb{R}^p$ may be thought of as a parameter space, and a point $\theta^0 \in \Theta$ which may be thought of as a starting value. From here one goes on to define a sequence $(\theta^k : k = 0, 1, 2, \ldots)$ by

$$\theta^{k+1} = g(\theta^k) \qquad k = 0, 1, 2, \ldots.$$

If we are clever or lucky enough to make good choices of g and θ^0 our sequence $(\theta^n : n = 0, 1, 2, \ldots)$ may converge to a limit θ^*. Further if g happens to be continuous at θ^* we will have that $g(\theta^*) = \theta^*$ in which case we say that θ^* is a *fixed point* of the iteration. In many statistical applications θ^* will be the parameter estimate that we are seeking.

I will consider questions of convergence and I will also show how the function g may be used to construct influence curves for the parameters of the fitted model.

As an example I look at the Gauss-Newton method for the fitting of nonlinear regression models. I construct influence curves for a 4-parameter nonlinear growth curve model fitted by Gauss-Newton to some data on the growth of nematodes.

Distribution Free Tests for Discrete Distributions and the Corollary for Continuous Distributions

Estáte V. **Khmaladze**
Victoria University of Wellington

Let $p_1, \dots, p_m$ be a discrete probability distribution; all $p_i > 0$ and $\sum_{i=1}^m p_i = 1$, which we consider as a hypothesis. Denote $\nu_{1n}, \dots, \nu_{mn}$ the corresponding frequencies in a sample of size n and consider the vector Y_n of components of the chi-square statistic

$$Y_{in} = \frac{\nu_{in} - np_i}{\sqrt{np_i}},\ i = 1, \dots, m.$$

How many sensible statistics do we know from the vector Y_n of these components, which are asymptotically distribution free? Probably just one:

$$\langle Y_n, Y_n \rangle = \sum_{i=1}^m \frac{(\nu_{in} - np_i)^2}{np_i}.$$

For example, the discrete version of the Kolmogorov–Smirnov statistic, $\sup_i |\sum_{j \le i} Y_{jn}|$, is not asymptotically distribution free.

However, we show that any test based on the vector Z_n with coordinates

$$Z_{in} = \frac{\nu_{in} - np_i}{\sqrt{np_i}} - \frac{1}{\sqrt{m}} \sum_{j=1}^m \frac{\nu_{jn} - np_j}{\sqrt{np_j}} \frac{1}{1 + \sum_{j=1}^m \sqrt{p_j/m}} \left(\frac{1}{\sqrt{m}} + \sqrt{p}_i \right)$$

will be asymptotically distribution free and that the relationship between Y_n and Z_n is one-to-one: hence, they contain the same amount of "statistical information".

The continuous time version of this statement is: if F is an absolutely continuous distribution on $[0,1]^d$ and v_{nF} is the empirical process based on a sample from F, then

$$du_n(x) = \frac{1}{\sqrt{f(x)}} dv_{nF}(x) - \int_{[0,1]^d} \frac{1}{\sqrt{f(y)}} dv_{nF}(y) \frac{1 + \sqrt{f(t)}}{1 + \int_0^1 \sqrt{f(s)} ds}\, dx$$

converges weakly to a standard Brownian bridge.

Goodness-of-Fit Tests for Long Memory Moving-Average Marginal Density

Hira **Koul**
Michigan State University

In this talk we will discuss the problem of fitting a known distribution function or density to the marginal error density of a stationary long memory moving-average process when its mean is known and unknown. When the mean is unknown and estimated by the sample mean, the first-order difference between the residual empirical and null distribution functions is known to be asymptotically degenerate at zero. Hence, it cannot be used to fit a distribution up to an unknown mean. However, we shall show that by using a suitable class of estimators of the mean, this first order degeneracy does not occur. We also present some large sample properties of the tests based on an integrated squared-difference between kernel-type error density estimators and the expected value of the error density estimator based on errors. The asymptotic null distributions of suitably standardized test statistics are shown to be chi-square with one degree of freedom in both cases of known and unknown mean. This is totally unlike the i.i.d. errors set-up where suitable standardizations of these statistics are known to be asymptotically normally distributed.

Empirical Process Central Limit Theorems Needed for Survey Data

Thomas **Lumley**
University of Auckland
(Joint work with Alastair Scott)

Empirical process central limit theorems are fundamental to modern semiparametric statistics, giving a tractable and general approach to asymptotics in models with infinite-dimensional parameters. For sequences and triangular arrays of independent random variables, and for stationary sequences under mixing conditions, reasonably sharp conditions are known for sets of functions $\mathcal{F}$ to ensure $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} f(X_i)$ converges weakly to a tight Gaussian process indexed by f.

For data from complex survey designs much less is known. Stratified element sampling with a fixed number of strata can be handled using Praestegaard's exchangeable bootstrap, giving similar results to i.i.d. sampling. Convergence of the one-dimensional classical empirical process has been shown under weak conditions for element sampling designs (Wang, in press) using brute-force fourth-moment bounds, but this does not extend easily to larger classes of functions. We will review the existing literature, explain why we think this problem is non-trivial and important, and appeal for help.

Entropy Estimation for Multivariate Distributions

Robert M. **Mnatsakanov**
West Virginia University
(Joint work with J.E. Harner and S. Li)

Problems in molecular science, fluid mechanics, chemistry, physics, etc., often require the estimation of unknown differential entropy $H(f)$ of the density f when the data are circular or spherical. Three new constructions of unknown entropy $H(f)$ are obtained using the histogram, nearest neighbor, and moment-recovered approaches. In this talk the consistency and L_1-rate of convergence of the proposed entropy estimates will be discussed. The performances of our estimates are demonstrated via a simulation study.

Point Processes and Patch Survival in Metapopulations

Phil **Pollett**
University of Queensland
(Joint work with Ross McVinish)

We consider a model for the presence/absence of a population in a collection of habitat patches, which assumes that colonisation and extinction of patches occur as distinct phases. Since the local extinction probabilities are allowed to vary between patches, our model permits an investigation of the effect of habitat degradation on the persistence of the population. The limiting behaviour of the model is examined as the number n of habitat patches becomes large. When the initial number of occupied patches increases at the same rate as n a law of large numbers ensues. However, here we focus attention on the case where the initial number occupied is *fixed*, and thus our aim is to determine conditions under which a metapopulation that is close to extinction may recover. By treating the patch survival probabilities of occupied patches at time t as a point process S_t^n on $[0, 1)$, we are able to exploit the probability generating functional techniques described in Section 9.4 of Daley and Vere-Jones Vol. II to show that S_t^n converges weakly to a point process S_t. Since extinction of the metapopulation by time t corresponds to the event that S_t is the empty set, this probability can be calculated from the probability generating functional of S_t in much the same way that the probability of extinction of a branching process can be calculated from the probability generating function of its offspring distribution.

Is Testing a Tree Easier than Finding It?

Mike **Steel**
University of Canterbury

A fundamental problem in evolutionary biology is to reconstruct a tree that has a set X of present-day species as its leaves, using just sequence data sampled from the species in X. This tree represents how the species in X descended from a common ancestral species. The standard approach is to assume that sequence sites evolve i.i.d. on the tree according to a Markov process (or a mixture of such processes). This leads to some fundamental information-theoretic questions: How long do the sequences need to be so that the reconstructed tree is correct with high probability? And in particular, how fast does the sequence length need to grow as a function of the number n of vertices of the tree? We contrast this 'reconstruction' question with a corresponding 'testing question': Is the required rate of sequence length growth with n lower if we are given a candidate tree and wish to merely 'test' it by asking: "Is this the tree that produced the sequences?" How about if we are given two trees along with the promise that one of the two trees produced the sequences – can we 'tease' out the true tree with even shorter sequences? Using an interplay of combinatorial and stochastic arguments, we find the answers to these questions can sometimes be surprising, depending on the properties of the Markov process.

Poisson Process Approximation for Dependent Superposition of Point Processes

Aihua **Xia**
University of Melbourne
(Joint work with L.H.Y. Chen)

Grigelionis (1963) proved that the superposition of independent sparse point processes converges weakly to a Poisson process on the carrier space $\mathbb{R}+$ and the result was subsequently extended to more general carrier spaces by Goldman (1967) and Jagers (1972) and to the superposition of dependent sparse point processes by Banis (1975, 1985), Kallenberg (1975), Brown (1979) and Banys (1980). In this talk, we use the Palm theory and Stein's method to study the rate of convergence of Poisson point processes to the locally dependent superposition of point processes. As examples, we'll discuss the rate of Poisson process approximation to the superposition of thinned point processes and of renewal processes.