Seventh Wellington Workshop in Probability and Mathematical Statistics

5 – 7 December 2019

Victoria University of Wellington

Titles, Presenters and Abstracts

Ordered alphabetically, by presenters' last names

Editor: **John Haywood**

Uniform Spanning Tree and Loop Erased Random Walk

Martin **Barlow**, University of British Columbia
Joint work with David Croydon and Takashi Kumagai

5 December, Session 2, Presentation 1

The two models, loop erased random walk (LERW) and uniform spanning tree (UST), have a close connection via algorithms due to Aldous, Broder, Wilson and Hutchcroft, which give the construction of the UST from LERW paths. In this talk we go in the opposite direction, and use properties of the UST to obtain tail bounds on the length of LERW paths.

Local Limit Theorems for Occupancy Models

Peter **Braunsteins**, University of Melbourne
Joint work with Andrew Barbour and Nathan Ross

6 December, Session 2, Presentation 2

Local central limit theorems for general sums of independent integer valued random variables are well understood; however, for sums of dependent random variables much less is known. In this talk we establish a local central limit theorem for two sums of dependent random variables: the number of degree d vertices in an Erdős–Rényi graph, and the number of germs with d neighbours in a germ-grain model. Our approach relies on Stein's method for distributional approximation.

Arcsine Laws for Permutation Generated Random Walks

Han Liang **Gan**, University of Waikato (from 2020)

5 December, Session 3, Presentation 1

One of the classic arcsine laws for the simple symmetric random walk is that as the length of the walk increases, the proportion of time the walk spends above zero converges to the arcsine distribution. We consider a random walk generated by a permutation in the following manner: given a random uniform permutation of length n, we generate a random walk by the ascents and descents of the sequence. It turns out that we still have the same arcsine law for the amount of time this walk spends above zero. In this talk we will discuss why this problem is important in genomics, some ideas of the proof and an interesting conjecture.

The Signature of a Path: Uniqueness and Beyond

Xi **Geng**, University of Melbourne
Joint (ongoing) work with Horatio Boedihardjo *et al.*
5 December, Session 3, Presentation 2

Iterated path integrals, also known as the signature of a path, plays a fundamental role in the analytic approach to stochastic differential equations (rough path theory). The significance of path signature is revealed by the uniqueness theorem which asserts that every (rough) path is determined by its signature up to tree-like pieces. In this talk, we discuss two important open problems for path signature both in the deterministic and probabilistic contexts. One is concerned with the relationship between the tail behaviour of signature and geometric properties of the underlying path, and the other is concerned with the image of the signature map. The core of our study is the methodology which involves developing paths onto certain Lie groups and relating the signature with geometric/algebraic properties of the group.

Extreme Values, Couplings and Graphs

Jesse **Goodman**, University of Auckland
5 December, Session 1, Presentation 3

Given a sample of n i.i.d. copies of a random variable, the first k order statistics can be approximated using the points of a Poisson process, in the limit when $n \to \infty$. This talk will describe a coupling approach that quantifies the error in this approximation, even when the number of approximated points, k, is large compared to n. The coupling will be applied to give a modified look at the Extreme Value Theorem, and an extended version of the coupling allows us to find a "strong disorder" property in a random graph problem with exotic tail behaviour.

Noise Sharing and Mexican-Hat Coupling in a Stochastic Neural Field

Priscilla **Greenwood**, University of British Columbia
Joint work with Peter H. Baxendale and Lawrence M. Ward
5 December, Session 1, Presentation 1

A diffusion-type operator biologically significant in neuroscience is a difference of Gaussian functions used as a spatial-convolution kernel (Mexican-Hat operator). We are interested in the dynamics inherent in a neural structure such as visual cortex modelled by stochastic neural field equations, a class of stochastic differential-integral equations using the Mexican-Hat kernel. We find that spatially smoothed noise, in a field of Ornstein-Uhlenbeck processes, without direct spatial coupling, causes pattern formation. Our analysis of the interaction between coupling and noise-sharing yields optimal parameter combinations for the formation of spatial pattern.

Bibliography

Baxendale, Peter H., Greenwood, Priscilla E. and Ward, Lawrence M. (2019) Noise sharing and Mexican-Hat coupling in a stochastic neural field. *Phys. Rev. E* 100, 022130.

Random KNN Classification and Regression

E. James **Harner**, West Virginia University
Joint work with Shengqiao Li
6 December, Session 1, Presentation 1

Random KNN (RKNN) is a generalization of traditional nearest-neighbour modelling, which consists of an ensemble of base k-nearest neighbour models, each constructed from a random subset of the input variables. Random KNN can be used to select important features using the RKNN-FS algorithm. Empirical results on microarray data sets with thousands of variables and relatively few samples show that RKNN-FS is an effective feature selection approach for high-dimensional data. RKNN is similar to Random Forests (RF) in terms of classification accuracy without feature selection. However, RKNN provides much better classification accuracy than RF when each method incorporates a feature-selection step. RKNN is significantly more stable and robust than RF. Further, RKNN-FS is much faster than the Random Forests feature selection method (RF-FS), especially for large scale problems involving thousands of variables and/or multiple classes. Random KNN and feature selection algorithms are implemented in an **R** package `rknn`, which supports both classification and regression. We show how to apply the Random KNN method to high-dimensional genomic data using `rknn`.

Keywords: High dimensional data, k-nearest neighbour, parallel computing, statistical learning.

Bibliography

Li, Shengqiao (2015) `rknn`: Random KNN Classification and Regression.
`https://cran.r-project.org/package=rknn`

Inference in Population-Size-Dependent Branching Processes

Sophie **Hautphenne**, University of Melbourne
Joint work with Peter Braunsteins and Carmen Minuesa
6 December, Session 2, Presentation 1

Population-size-dependent branching processes (PSDBPs) are models which describe the evolution of populations where individuals in the same generation give birth independently according to a probability distribution which depends on the current population size. One important class of PSDBPs are branching processes with a carrying capacity; these are appropriate for modelling populations that exhibit logistic growth, where the population size tends to fluctuate, for a long period of time, around a threshold value corresponding to the maximum number of individuals that an ecosystem can support.

We propose an estimator for the mean of the offspring distribution at each population size in a discrete-time PSDBP, based on the observation of the total population sizes up to some generation. Our main challenge is the fact that branching processes with a carrying capacity eventually become extinct with probability one (after a long time). We propose a way to derive asymptotic properties of the estimator in this setting. This leads to a number of questions about desired properties of estimators in branching processes that almost surely become extinct. Our proofs rely on coupling arguments and the analysis of the Q-process associated with (or Doob h-transform of) the original branching process.

Using Copulas to Model and Simulate Two-Dimensional Warranty Data

Yuichi **Hirose**, Victoria University of Wellington

6 December, Session 1, Presentation 2

We focus on the simulation of automotive warranty data, using a two-dimensional (2D) distribution based on copulas. First we build a model by selecting a 2D parametric model for warranty prediction among symmetric and asymmetric copulas. In the second part, we build a nonparametric estimate of a copula function using a neural network. The task of parameter fitting and measuring the quality of the fit is also addressed.

Distribution-free Testing for Markov Sequences

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

6 December, Session 4, Presentation 3

Karl Pearson could have done what we will do here, if only he had chosen not just an invariant, which his χ^2 statistic is, but the maximal invariant under the group that he consciously or sub-consciously encountered in his famous 1900 paper. True, he would have needed the notion of Markov processes, the theory of linear spaces and linear operators, the notion of empirical processes and their asymptotic theory, none of which existed in 1900. But in principle, at the root of what follows is the difference in the choice of invariant: it was an invariant in his great paper, and it is a maximal invariant here.

Given a Markov sequence $X_1, X_2, \ldots$, taking values in some measurable space $(\mathbb{S}, \mathcal{S})$ with transition probabilities $(P_x(B), x \in \mathbb{S}, B \in \mathcal{S},)$ consider

$$v_t(f) = \frac{1}{\sqrt{T}} \sum_{s \le t} (f(X_{s-1}, X_s) - E[f(X_{s-1}, X_s)|X_{s-1}]), \quad t \le T,$$

and call it a *function parametric empirical process* for the Markov sequence. Functionals, especially omnibus functionals from $v_t(f)$ (in f) are a direct analogue of goodness of fit statistics from classical empirical processes.

The limit distribution of the process $v_t(f)$, and therefore statistics based on it, depends on the set of transition probabilities $(P_x(B), x \in \mathbb{S}, B \in \mathcal{S})$, which makes the theory heavy. However, we consider a group of transformations that map $v_t(f)$ into any other empirical processes with transition probabilities $(Q_x(B), x \in \mathbb{S}, B \in \mathcal{S})$ within its equivalence class. The maximal invariant under this group will have a limit distribution, independent from the particular choice of $(P_x(B), x \in \mathbb{S}, B \in \mathcal{S})$, just as the χ^2 statistic has a limit distribution independent from the underlying hypothetical distribution. The approach extends to parametric classes of transition probabilities, for which distribution free testing is also possible.

Bibliography

Khmaladze, Estáte (2013) Note on distribution free testing for discrete distributions. *Annals of Statistics* 41(6), 2979–2993.

Khmaladze, Estáte (2016) Unitary transformations, empirical processes and distribution free testing, *Bernoulli* 22, 563–599.

Khmaladze, Estáte (2019) Testing hypothesis on transition matrix of a Markov chain, draft exists.

Khmaladze, Estáte (2019) Distribution free testing for parametric regression in $\mathbb{R}^p$, under submission.

Equality between Quenched and Annealed Rate Functions of Random Walks in Random Environments

Alejandro **Ramírez**, Catholic University of Chile
Joint work with Rodrigo Bazaes, Chiranjib Mukherjee and Santiago Saglietti
5 December, Session 2, Presentation 2

We consider random walks in i.i.d. uniformly elliptic random environments in dimensions $d \geq 4$. In 2004 Varadhan established both an annealed and a quenched large deviation principle with a rate function which is finite on the unit l^1 ball $\mathbb{D}$. In 2010, Yilmaz proved the equality between the annealed and the quenched rate functions on a neighbourhood of the velocity whenever the so called ballisticity condition (T) is satisfied. Here we prove that actually there is equality between these rate functions, whenever the disorder is low, and at points on the boundary of $\mathbb{D}$, without assuming condition (T).

Bibliography

Bazaes, R., Mukherjee, C., Ramírez, A. and Saglietti, S. (2019) Equality and difference of quenched and averaged large deviation rate functions for random walks in random environments without ballisticity. `https://arxiv.org/pdf/1906.05328.pdf`

On the Use of the second Khmaladze Transform to Test Goodness of Fit in Mixtures of Hidden Markov Models

Leigh **Roberts**, Victoria University of Wellington
5 December, Session 1, Presentation 2

This paper applies the second Khmaladze transform to test goodness of fit of mixtures of hidden Markov models. Mixing probabilities are assumed to be logistic, and may depend on known covariates. The methodology illustrated in this paper applies equally well to more general mixture models when the emphasis in modelling is on the mixture probabilities.

Keywords: Empirical process, Kolmogorov-Smirnov statistic, logistic mixing probabilities, Markov chain, rotation, unitary transform.

Dirichlet and Poisson-Dirichlet Approximations for some Wright-Fisher Models with Mutation

Nathan **Ross**, University of Melbourne
6 December, Session 3, Presentation 1

In population genetics, the Wright-Fisher genealogy is a fundamental model of finite population structure. Given the genealogy, each individual has a type, which is either inherited from their parent, or randomly chosen according to a mutation rule. The stationary distribution of the Markov chain of type-counts in each generation is an important quantity in population genetics, since it governs the sampling distribution of types. Even for nice mutation rules, such as parent independent mutation (PIM), these stationary distributions are analytically intractable. However, for PIM, as the size of the population becomes large, the stationary distributions are well-approximated by the Dirichlet (finite number of types) and Poisson-Dirichlet (infinite types) distributions. We discuss the error in these approximations, including the approximation of the discrete sampling formula by that of the limit, which in the case of infinite types is the Ewens Sampling Formula.

The Friendship Paradox and a Probabilistic Friendship Model

Sheldon M. **Ross**, University of Southern California

6 December, Session 2, Presentation 3

The friendship paradox states that "your friends tend to have more friends than you do". We explain exactly what is meant by this, and give some extensions and generalizations. We then discuss a probability model for a friendship network that assumes that each individual has a value, with these values being independent and identically distributed, and that individuals with values x and y are friends with probability $p(x, y)$.

Finding Optimal Solutions by Stochastic Cellular Automata (SCA)

Akira **Sakai**, Hokkaido University
Joint work with Satoshi Handa, Katsuhiro Kamakura and Yoshinori Kamijima

5 December, Session 2, Presentation 3

Finding a ground state of a given Hamiltonian is an important but hard problem. One of the potential methods is to use a Markov chain Monte Carlo (MCMC) to sample the Gibbs distribution whose highest peaks correspond to the ground states. We use SCA and see if it is possible to find a ground state faster than the conventional MCMCs, such as the Glauber dynamics. In my presentation, I will explain that, if the temperature is high enough, it is possible for SCA to have on average more spin-flips per update than Glauber and, at the same time, to have an equilibrium distribution "close" to the one for Glauber, i.e., the Gibbs distribution.

Sparse Principal Component Analysis with Preserved Sparsity Pattern

Karim **Seghouane**, University of Melbourne

6 December, Session 4, Presentation 1

Principal component analysis (PCA) is widely used for feature extraction and dimension reduction in pattern recognition and data analysis. Despite its popularity, the reduced dimension obtained from PCA is difficult to interpret due to the dense structure of principal loading vectors. To address this issue, several methods have been proposed for sparse PCA, all of which estimate loading vectors with few non-zero elements. However, when more than one principal component is estimated, the associated loading vectors do not possess the same sparsity pattern. Therefore, it becomes difficult to determine a small subset of variables from the original feature space that have the highest contribution in the principal components. To address this issue, an adaptive block sparse PCA method is proposed. The proposed method is guaranteed to obtain the same sparsity pattern across all principal components. Experiments show that applying the proposed sparse PCA method can help improve the performance of feature selection for image processing applications. We further demonstrate that the proposed sparse PCA method can be used to improve the performance of blind source separation for functional magnetic resonance imaging (fMRI) data.

Exit Time Distributions of the Finite Mixture of Markov Jump Processes: Properties and the EM Estimation

Budhi **Surya**, Victoria University of Wellington

6 December, Session 4, Presentation 2

In this talk, I will discuss properties and the EM estimation of the distribution of exit time to an absorbing state of the finite mixture of right-continuous Markov jump processes moving on a given finite state space. When the process is observed in a state or making a transition from one state to another, there is uncertainty associated with which underlying Markov process drives the movement of the process. Unlike its underlying processes, the mixture process does not have the Markov property. When conditioning on its past observations, Bayesian update of the exit time distribution is given explicitly in terms of the states the process has visited, the number of transitions between the states, the length of time the process stays in each state, and the intensity matrices of the underlying Markov processes. The prior (unconditional) distribution forms a generalized mixture of phase-type distributions. Explicit and closed form maximum likelihood estimates of the distribution parameters are presented. Under incomplete observations where either sample paths of the process or the exit times are available, the estimation is performed using the EM algorithm. Some Monte Carlo simulations are discussed to exemplify and validate the proposed algorithm.

A Multivariate Chain Rule for Derivatives with Applications to Studentized Functions

Kit **Withers**, Wellington

6 December, Session 3, Presentation 3

We extend Faa di Bruno's chain rule for the derivatives of a function of a function to the case where functions and arguments are multivariate. We use this to find the derivatives of a Studentized function of a parametric estimate. These are needed to obtain its cumulant coefficients, the basis of higher order analytic inference.

The Distribution of Zeros of the Derivative of the Riemann Zeta Function via Random Unitary Matrices

Nicholas **Witte**, Massey University, Palmerston North

6 December, Session 3, Presentation 2

In 1935 Speiser formulated an equivalent statement of the Riemann hypothesis in terms of the zeros of the derivative of the Riemann zeta function: that $\zeta'(s)$ has no zeros to the left of the critical line, $\mathrm{Re}(s) < 1/2$. We accept this statement and pose the question of the distribution of such zeros to the right of the critical line $\mathrm{Re}(s) > 1/2$, located far up the imaginary axis. To do this we model $\zeta'(s)$ by the derivative of the characteristic polynomial of a random matrix from $U(N)$, and deduce the distribution of its zeros within the unit circle, which corresponds to the region on the right-hand side of the critical line. Because the distribution of eigenvalues of a random unitary matrix sampled according to Haar measure is rotationally invariant the distribution we seek has only radial dependence. Formulating this distribution using tools from random matrix theory we uncover a bi-orthogonal polynomial system with an integrable measure on the unit circle and proceed to employ approximation theory to evaluate this for all ranks N. The correspondence of our distribution with data from the Riemann zeta function is remarkable and predicts some curious features observed in the latter.