

UNIVERSITY of OTAGO
Te Whare Wānanga o Otāgo
NEW ZEALAND

SAPERE AUDE

NEW ZEALAND STATISTICAL ASSOCIATION CONFERENCE 2019

NZSA 2019

26 - 28 November

University of Otago, Dunedin, New Zealand

The New Zealand Statistical Association

harmonic analytics

ORITAIN™
PROVING ORIGIN, PROTECTING REPUTATIONS

The open LaTeX template, AMCOS_booklet, used to generate this booklet is available at
https://github.com/maximelucas/AMCOS_booklet

Contents

Greetings! **4**
- Sponsors . . . 4
- Organising committee . . . 4

Useful Information **5**
- Venues . . . 5
- Analytics Forum . . . 6
- Speaker uploads . . . 6
- Session chairs . . . 6
- Wifi/Internet . . . 7
- Code of conduct concerns . . . 7
- Just for fun . . . 7

Timetable **9**
- At-a-glance . . . 9
- Tuesday 26/11 . . . 10
- Wednesday 27/11 . . . 11
- Thursday 28/11 . . . 12

List of Abstracts – Talks **13**

List of Abstracts – Posters **43**

List of Participants **48**

Greetings!

On behalf of the organisers, welcome to the 70th annual New Zealand Statistical Association (NZSA) conference.

The main conference activities from Tuesday, 26th November to Thursday, 28th November are detailed in the ensuing content. They comprise keynote presentations, contributed presentations, and posters from a variety of statisticians and data analysts, from academia, industry, and government. We are proud to host a breadth of both local and international delegates.

The general goal of this conference is to celebrate Kiwi statistics in academia and industry, and in particular the rapidly growing body of students studying statistics at Honours and Masters/PhD levels.

Sponsors

We are grateful for the generous support of our sponsors: Harmonic Analytics (student talk prizes), Oritain (student poster prize), and the NZSA (student travel grants). A special presentation by Lisa Chen and Glenn Thomas of Harmonic Analytics will take place at 1420 on 27/11 in the St. David lecture theatre.







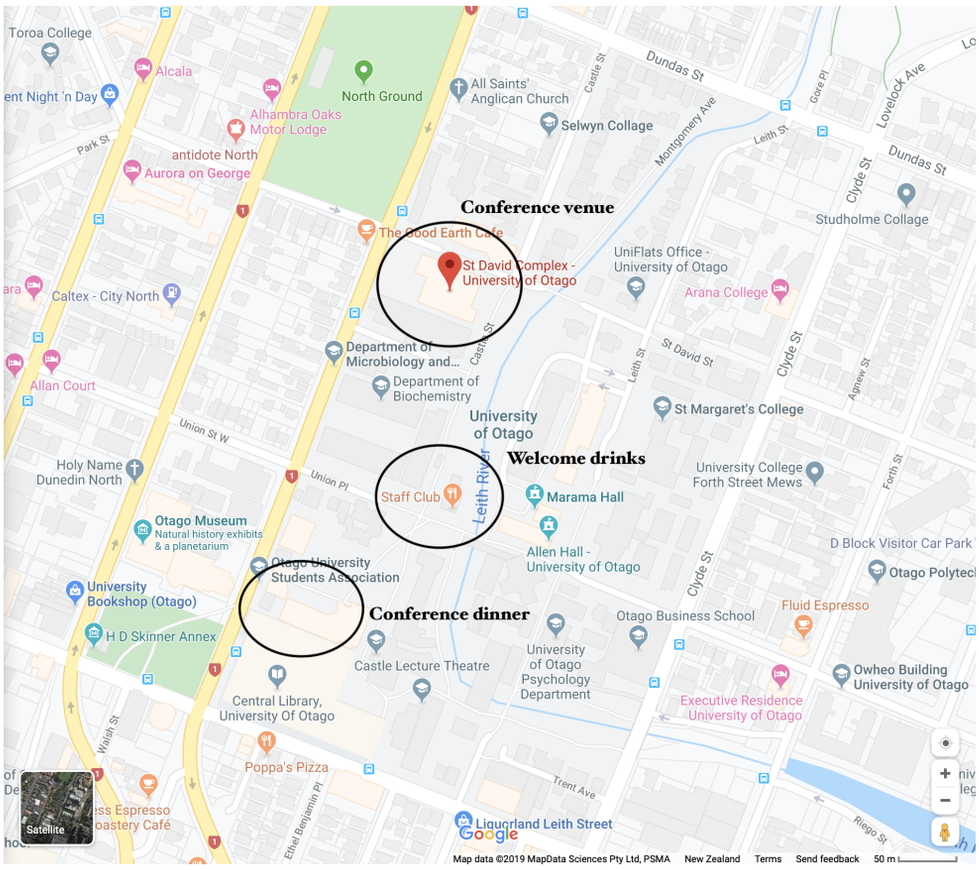
Organising committee

Dr Tilman Davies (Chair)
A/Prof David Fletcher (Co-chair)
Dr Matthew Parry
Dr Ting Wang
Ms Deanna McLennan (Events Specialist, Otago Events Hub)
Ms Jasmine Millar (Events Specialist, Otago Events Hub)

Useful Information

Venues

- The conference proceedings from 26/11 to 28/11 will take place in the **St. David lecture theatre complex** on campus. We make use of the main St. David lecture theatre (STDAV), as well as two of the adjacent seminar rooms for parallel sessions: St. David seminar room A (SDAVA) and St. David seminar room B (SDAVB). The poster session attached to lunch on 27/11 will take place in the foyer of the St. David complex.
- All morning/afternoon teas and lunches will be in the foyer of the St. David complex.
- The welcome drinks from 5:30pm on 25/11 are on the upper level of the University of Otago staff club, approximately 200m south of the conference venue along the Leith River.
- The conference dinner from 6:30pm on 27/11 will take place in the University of Otago Union building, another 100m in a southerly direction from the staff club.



Analytics Forum

As in previous years, the NZSA is also hosting a NZ Data Science + Analytics Forum during the conference. The forum is scheduled for 3:10-5:30pm on Wednesday, November 27, and all conference delegates are welcome to attend.

This year the Forum highlights the growing role of Analytics and AI, especially around Agritech. It will also showcase a number of Dunedin-based firms that are having an impact nationally and globally. The keynote address will given by **Christopher Laing, Head of Xero AI** on the topic of

"The opportunity for NZ-grown AI"

This will be followed by presentations from:

- John Enlow, Chief Product Officer, TracMap
- Michael Lee, Lead Scientist, Beef + Lamb New Zealand
- Paul Hansen, Co-founder, 1000minds
- Benoit Auvray, Director & Lead Scientist, Iris Data Science
- Michael Macknight, Director, ADInstruments
- Aaron Furrer, Co-founder, Next Farm
- Jonah Duckles, Consultant, AbacusBio
- Callum McKenzie, Lead IIOT, Scott Technology
- Esther Meenken, Senior Statistician, AgResearch

Speaker uploads

- We ask that all speakers bring their presentation on a portable flash drive and upload them to the designated Dropbox folder on the computer in the room in which they are presenting **prior to their session**. This can be done at any time in the morning before the daily keynote or during a break (morning/afternoon tea or lunch). It is helpful if the file is labelled with the presenting author's surname.

- Those presenting posters may mount their posters on the stands provided in the St. David foyer at any time prior to the session on 27/11. Pins/velcro dots will be supplied. We encourage early mounting!

Session chairs

- In each session, note that unless otherwise pre-warned by the organising committee, the **last speaker in the group** is assumed to be the chair. "Be aware, if you're the chair!"

- The rooms will have 3- and 1-minute warning signs provided for the chair's use.

Wifi/Internet

Delegates have two network options for connecting to wifi during the conference:

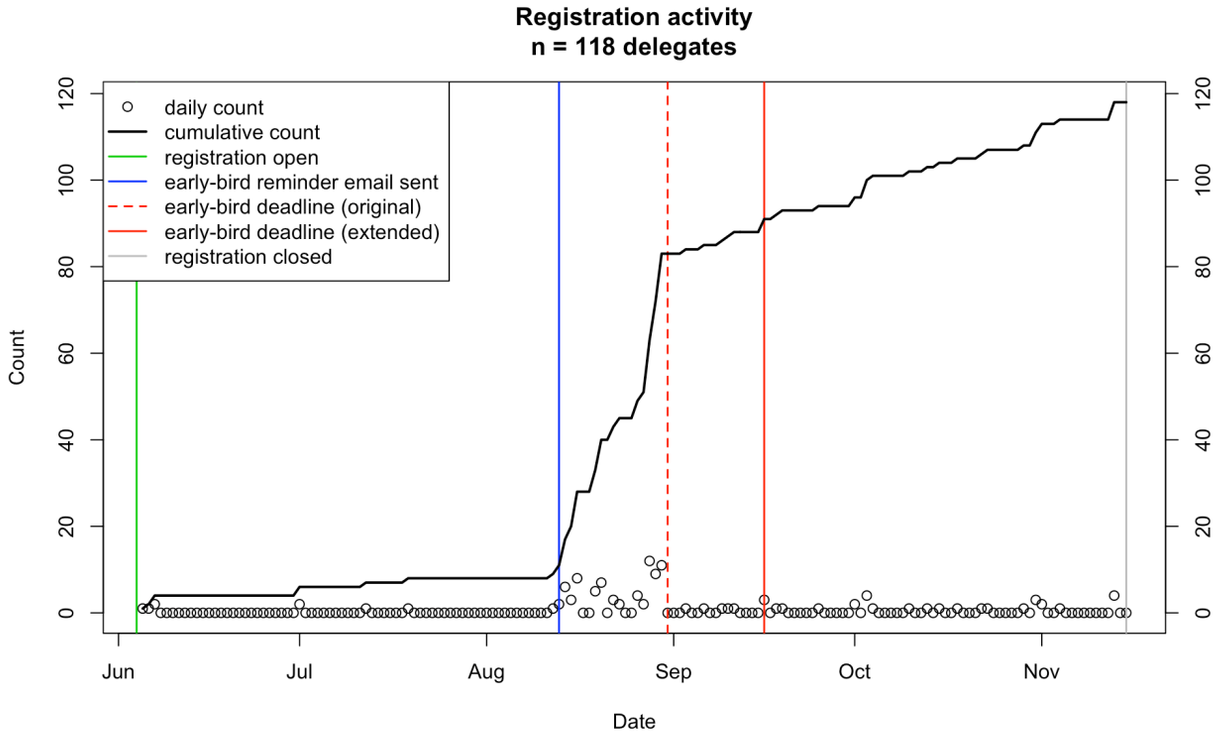
UO_Guest Connect to the UO_Guest wireless network (SSID) on your device; log in if you have an account or click on the "Don't have an account?" link to register; fill in your details and sign on. A UO_Guest account provides access to the Internet for up to 14 days/500Mb per day/2 devices.

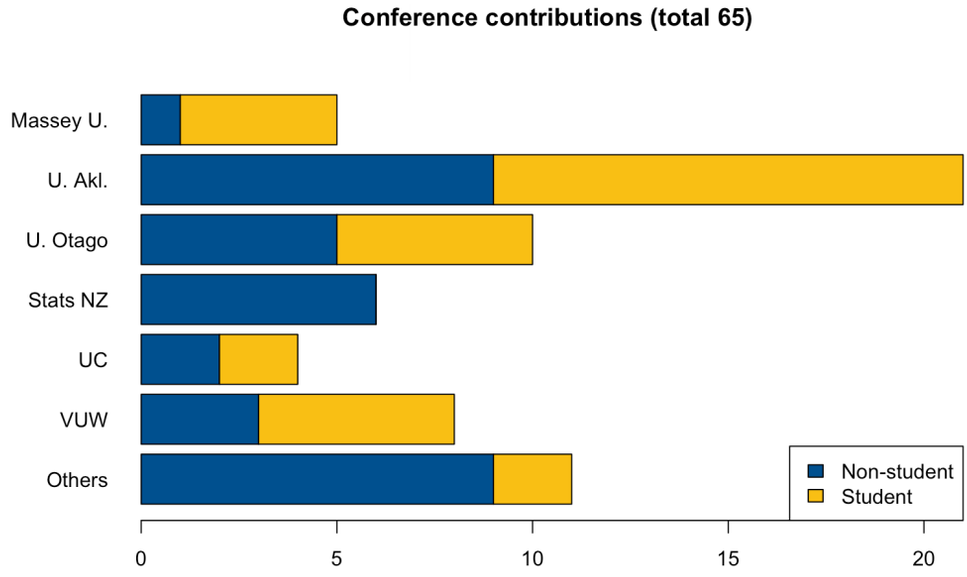
eduroam Delegates from an institution that subscribes to the eduroam network can use Otago's eduroam access as usual.

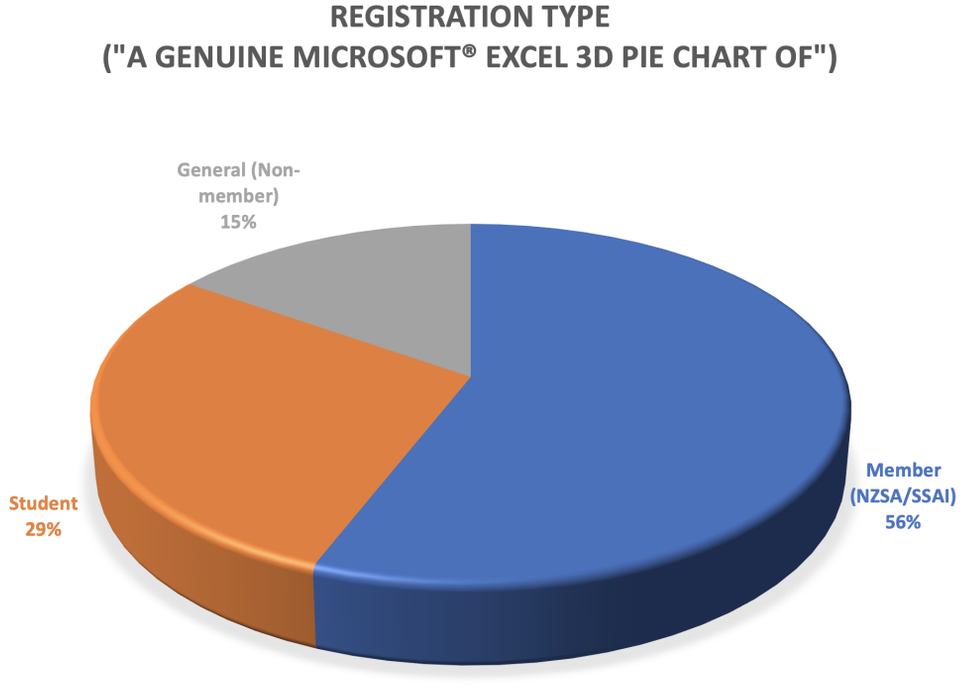
Code of conduct concerns

As part of registration, all delegates have agreed to a code of conduct to ensure an inclusive and supportive environment `https://www.otago.ac.nz/nzsa/conduct/`. If you have any concerns or questions related to conduct during the conference please approach Matthew Parry or Ting Wang of the organising committee as a first step; alternatively you may email `nzsa2019@otago.ac.nz`. Any concerns raised will be treated confidentially.

Just for fun




Conference contributions (total 65)
Massey U.
U. Akl.
U. Otago
Stats NZ
UC
VUW
Others
Non-student
Student
0
5
10
15
20


REGISTRATION TYPE
("A GENUINE MICROSOFT® EXCEL 3D PIE CHART OF")
General (Non-member) 15%
Student 29%
Member (NZSA/SSAI) 56%

Timetable

At-a-glance

	Monday 25/11	Tuesday 26/11	Wednesday 27/11		Thursday 28/11
0845 - 0915		Registration			
0915 - 0930		Housekeeping	Notices		Notices
0930 - 1030		Keynote 1	Keynote 2		Keynote 4
1030 - 1100		Morning tea	Morning tea		Morning tea
1100 - 1220		Parallel sessions	Parallel sessions		Parallel sessions
1220 - 1320		Lunch	Lunch, Posters		Lunch, Conf. Close
1320 - 1440		Parallel sessions	Keynote 3, Sponsor talk		
1440 - 1510		Afternoon tea	Afternoon tea		
1510 - 1630		Parallel sessions	Parallel sessions	Analytics forum	
1630 - 1730					
1730	Welcome drinks, Registration	Early career event	NZSA AGM		
1830			Conference dinner		

Tuesday 26/11

FOY: St. David foyer; STDAV: St. David lecture theatre; SDAVA: St. David seminar room A; SDAVB: St. David seminar room B
* Student presentation

0845	**Registration (FOY)**		
0915	**Welcome/Housekeeping/Notices (STDAV)**		
0930	Keynote 1 (STDAV) Prof. Adrian Baddeley, Curtin University, Australia (p. 15) *"There's gold in them thar coefficients!": The strange story of mineral prospectivity analysis*		
1030	**Morning tea (FOY)**		
	Contributed 1A (STDAV)	Contributed 1B (SDAVA)	Contributed 1C (SDAVB)
1100	Matthieu Vignes (p. 40) *Stochastic simulations of gene regulatory networks with* `sismonr`	Christine Bycroft (p. 16) *Administrative records in the 2018 Census dataset*	Oliver Stevenson* (p. 38) *Finding your feet: a Gaussian process model for estimating the abilities of Test cricket players*
1120	Olivia Angelin-Bonnet* (p. 13) *Causal inference of gene regulatory networks*	Tanjila Tabassum* (p. 38) *Factors influencing termination of defined benefit pension plans in the US*	Azam Asanjarani (p. 15) *Some results on stationary Markovian arrival processes*
1140	Xiaqiong Wang* (p. 41) *Integrative association analysis for multi-omics data*	Sarah Hall (p. 23) *Dual systems estimation and linkage error*	Lindsay Morris* (p. 33) *Goodness-of-fit of Bayesian spatial models using pivotal discrepancy measures*
1200	Yang Hai* (p. 22) *Bayesian linear mixed model with multiple random effects for risk prediction of complex traits*	Andrew Richens (p. 35) *Using Bayesian dual systems estimation to produce a population record file*	Patricio Maturana-Russel (p. 30) *Spectral density estimation using P-spline priors*
1220	**Lunch (FOY)**		
	Contributed 2A (STDAV)	Contributed 2B (SDAVA)	Contributed 2C (SDAVB)
1320	Yijia Qian* (p. 34) *Predicting breeding values after culling*	Emil Mendoza* (p. 31) *Nonparametric estimation of the random coefficients model using regularized maximum likelihood*	Andrew Robinson (p. 36) *A simple model for size-at-detection of an invasive species under surveillance*
1340	Roy Costilla (p. 19) *The need for speed in Genomics: Comparing Bayesian algorithms to estimate polygenic effects in tropically adapted beef cattle*	John Siryj* (p. 37) *Quantifying uncertainty in marginal likelihood estimation via the stepping-stone sampling algorithm*	David Chan* (p. 17) *Spatial capture-recapture – applications to acoustic surveys of cetacean populations*
1400	Jie Kang* (p. 26) *Construction of multi-allelic genomic relationship matrices using Genotyping-by-Sequencing (GBS) data*	Kate Lee (p. 27) *Calibration procedures for approximate Bayesian credible set*	Rodelyn Jaksons* (p. 25) *Modelling the emergence dynamics of the western corn rootworm beetle*
1420	Luis Apiolaza (p. 14) *A Bayesian multivariate evaluation of mānuka flowering*	Victor Miranda (p. 31) *Vector generalized linear models and two-parameter links with applications to quantile regression and the NBD*	Richard Arnold (p. 14) *Biclustering in capture-recapture experiments*
1440	**Afternoon tea (FOY)**		
	Contributed 3A (STDAV)	Contributed 3B (SDAVA)	
1510	Fangyao Li* (p. 28) *On the selection of predictors by using greedy algorithms and information theoretic criteria*	Tong Chen* (p. 18) *Optimal multi-wave sampling for regression modelling*	
1530	Zhijian Wen* (p. 42) *Identifying firing pin impressions by using machine learning*	Yujin Kim* (p. 26) *Constrained estimation using group-summary level information for correlated data*	
1550	Kien Tran* (p. 39) *Fuzzy clustering of categorical data using finite mixtures*	Omer Ozturk (p. 34) *Post-stratified probability-proportional-to-size sampling from stratified populations*	
1610	Matthew Parry (p. 34) *Archetypal analysis as an autoencoder with application to data imputation*	Claudia Rivera-Rodriguez (p. 36) *Conditional likelihood for two-phase designs: continuous responses and relative risk*	
1730	**Early career/Young statisticians event (STAFF CLUB – SEE PG. 5 FOR DIRECTIONS)**		

Wednesday 27/11

FOY: St. David foyer; STDAV: St. David lecture theatre; SDAVA: St. David seminar room A; SDAVB: St. David seminar room B
* Student presentation

0915	**Notices (STDAV)**		
0930	Keynote 2 (STDAV) Prof. Martin Hazelton, University of Otago, New Zealand (p. 23) *Dynamic polytope samplers for linear inverse problems*		
1030	**Morning tea (FOY)**		
	Contributed 4A (STDAV)	Contributed 4B (SDAVA)	Contributed 4C (SDAVB)
1100	Kevin Cheng* (p. 18) *Assessing heterozygous balance in NGS DNA profiles*	Lingyu Li* (p. 28) *Comparison of stopping criteria for the EM algorithm*	Elena Moltchanova (p. 32) *Hydrothermal time seed germination model: Non-linear regression vs. survival analysis*
1120	John Holmes (p. 24) *Recycling genome-wide test statistics: Can you really find more information?*	Amir Bashir* (p. 16) *Decoupled shrinkage and selection of sparse factor models*	Khandoker Akib Mohammad* (p. 32) *Efficient estimation for the joint model of survival and longitudinal data*
1140	Maarten Kruijver (p. 27) *Using a mixture model to estimate the frequencies of Y-chromosomal DNA profiles*	Yongshi Deng* (p. 20) *Multiple imputation through denoising and variational autoencoders*	Beatrix Jones (p. 25) *Longitudinal analysis of dietary patterns*
1200	James Curran (p. 19) *Estimation of Y haplotype frequencies with lower order dependencies*	Zach Goeden (p. 22) *Likelihood-based clustering of commercial fisheries data*	
1220	**Lunch (FOY)** **+ Poster Session** (pp. 43-47)		
1320	Keynote 3 (STDAV) Prof. Barbara Holland, University of Tasmania, Australia (p. 24) *Statistical phylogenetics: An introduction and some open problems*		
1420	Sponsor presentation (STDAV) Lisa Chen and Glenn Thomas, Harmonic Analytics, New Zealand (p. 17) *Harmonic Data Science Team Building Experience and University Engagements*		
1440	**Afternoon tea (FOY)**		
	Analytics Forum (STDAV)	Contributed 5A (SDAVA)	
1510		Christine Bycroft (p. 16) *Dual system estimation combining census responses and an administrative population*	
1530		Jamas Enright (p. 21) *From Near and Far: Use of alternative sources and imputation for item non-response in 2018 Census*	
1550		Tadhg Daly (p. 20) *New Zealand Crime and Victims Survey: Filling the knowledge gap*	
1610		Diane Ramsay (p. 35) *Measuring child poverty – Unleashing the power of data to make New Zealand the best place in the world to be a child*	
1730	**NZSA Annual General Meeting (STDAV)**		
1830	**Conference dinner (UNIVERSITY UNION BUILDING – SEE PG. 5 FOR DIRECTIONS)**		

Thursday 28/11

FOY: St. David foyer; STDAV: St. David lecture theatre; SDAVA: St. David seminar room A; SDAVB: St. David seminar room B
* Student presentation

0915	**Notices (STDAV)**	
0930	Keynote 4 (STDAV) A/Prof. Ivy Liu, Victoria University of Wellington, New Zealand (p. 28) *The development of stereotype models for ordinal data*	
1030	**Morning tea (FOY)**	
	Contributed 6A (STDAV)	Contributed 6B (SDAVA)
1100	Thomas Lumley (p. 29) *Pairwise likelihood for multilevel models in multistage samples*	Harel Lustiger (p. 29) *R user-friendly functions*
1120	Paul Smith (p. 37) *Application of discriminant and principal component analyses to analysis of metabolomics data obtained from brain tissue*	Robin Turner (p. 39) *Teaching biostatistics to clinical researchers*
1140	Rolf Turner (p. 40) *Model selection via cross validated log likelihood*	Roger Marshall (p. 30) *Enhancing regression nomograms to visualise regression models*
1200	David Fletcher (p. 21) *Recent developments in model averaging*	Louise McMillan (p. 30) *Directional measures and visualization for population differentiation*
1220	**Lunch (FOY)** **+ Conference Close**	

List of Abstracts – Talks

Student presentations are labelled with S*; keynote presentations with* K.

Causal inference of gene regulatory networks

Olivia Angelin-Bonnet[1], **Matthieu Vignes**[1], **Patrick J. Biggs**[1,2], **Susan Thomson**[3], **Samantha Baldwin**[3] S

[1] School of Fundamental Sciences, Massey University, New Zealand

[2] School of Veterinary Science, Massey University, New Zealand

[3] Plant & Food Research, Christchurch, New Zealand

Multi-omic datasets enable us to quantify the molecular content of cells at different scales, e.g. transcriptome, proteome and metabolome, with ever-increasing precision. From these datasets, network inference methods aim at extracting information about the regulatory interactions that shape the behaviour of cells and influence their phenotype. By including genotypic information, it becomes possible to decipher the flow of information from the DNA to the phenotype of organisms. This knowledge is invaluable in a large array of biological problems, from understanding the influence of genotype on diseases to informing crop selection in plant breeding.

Recently, we developed the `sismonr` R package for simulating *in silico* biological systems, to generate benchmark datasets of gene expression for genetically diverse samples [1]. `sismonr` provides a biologically accurate model of gene expression regulation, being the first simulation tool to integrate the presence of noncoding genes, post-transcriptional regulation, genetic variation, and to explicitly represent the ploidy (definable by the user) of the system.

We used this simulator to compare the performance of existing network inference methods on simulated datasets that mirror the complexity of biological systems. In particular, we are focusing on statistical and computational methods dedicated to causal structure learning, from the prominent PC algorithm [3] to more recent methods, such as the LRpS-GES algorithm [2]. During this talk, we will highlight the strengths and weaknesses of these methods and the factors influencing their results. In particular, we will discuss the ability of existing methods to detect post-transcriptional changes and the effect of polyploidy on their precision. Building on these methods and the results of these comparisons, we are aiming to develop improved models for causal inference, that integrate different omics datasets and can be used for organisms of any ploidy.

References

[1] Olivia Angelin-Bonnet, Patrick J Biggs, and Matthieu Vignes. “The sismonr Package: Simulation of In Silico Multi-Omic Networks in R". In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2018, pp. 2729-2731.

[2] Benjamin Frot, Preetam Nandy, and Marloes H. Maathuis. “Robust causal structure learning with some hidden variables". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* (2019).

[3] Peter Spirtes et al. *Causation, prediction, and search*. MIT press, 2000.

Keywords

Causal inference; Gene regulatory networks; Systems Biology

A Bayesian multivariate evaluation of mānuka flowering

Luis Apiolaza

School of Forestry, University of Canterbury, New Zealand

Mānuka honey is a popular New Zealand food product, where both bees and plants can be bred for improved productivity. This research deals only with characterisation of mānuka breeding. Besides plant survival and growth, flowering attributes are important selection criteria for mānuka honey production. I analysed data from a mānuka progeny trial growing in a single environment, which included plant survival, total height, and three flowering traits: cover, density and openness.

The analyses comprised a Bayesian multivariate model that included intercept, plot and additive genetic effects for each of the traits, with a combination of Gaussian and Threshold models for the link functions. Model fitting relied on the MCMCglmm R package.

Posterior distributions for the degree of genetic control (heritability, h_2) and association between traits (genetic correlations, r_g) suggest that flowering traits are under strong genetic control (0.7 [0.5-0.8]) and are highly correlated to each other. Meanwhile, plant height (0.12) and survival (0.33) displayed lower genetic control. There was a positive genetic association between plant height and flowering ($r_g \approx 0.6$ for all flowering traits).

These genetic parameters will help breeders select genotypes with superior survival and flowering characteristics.

Keywords

Bayesian analysis; Breeding; Quantitative genetics

Biclustering in capture-recapture experiments

Richard Arnold[1], **Yu Hayakawa**[2]

[1] School of Mathematics and Statistics, Victoria University of Wellington, New Zealand

[2] School of International Liberal Studies, Waseda University, Japan

In capture-recapture experiments the size of a population is estimated from multiple capture-mark-release events of population members. This approach has been used in reliability to estimate the number of undetected faults in a piece of software following testing by a panel of reviewers.

Appropriate treatment of heterogeneity is critical to the creation of unbiased estimates of the total number of faults, both detected and undetected. There is variation in the intrinsic detectability of the faults, and variation in the skill and experience of the reviewers.

Finite mixture models have been a successful means of accounting for heterogeneity, and recently Manrique-Vallier (2016) used Dirichlet Process mixtures (DPM) to create non-parametric estimates of population size allowing for heterogeneity between individuals (faults), and fixed effects for capture occasions (reviewers). Unlike the finite mixture modelling approach, there is no need in the DPM method to specify the number of clusters in advance.

In this talk we extend the DPM approach to simultaneous clustering not only of the faults but also the reviewers. This biclustering technique is useful not only in the analysis of capture-recapture experiments, but in the exploration of structure in large $n \times m$ arrays of binary and count data. We also address a method to resolve a deficiency of DPM models in which the number of detected clusters increases with sample size.

Keywords

Capture-recapture; Clustering; Bayesian nonparametrics

Some results on stationary Markovian arrival processes

Azam Asanjarani[1], **Yoni Nazarathy**[2]

[1] Dept. of Statistics, University of Auckland, New Zealand

[2] School of Mathematics and Physics, University of Queensland, Australia

A Markovian arrival process (MAP) is a versatile modelling tool which is characterised by two matrices C and D. The matrix D represents the rates of observable transitions and C represents the rates of hidden transitions of the underlying Markov chain. Here, we consider two classes of irreducible MAPs; the Markov modulated Poisson process (MMPP) and the Markov switched Poisson process (MSPP). The former exhibits a diagonal D while the latter exhibits a diagonal C.

We show that for general MSPPs and two-state MMPPs, the following statements hold:

(I) The counting process is overdispered.

(II) The hazard rate of the event-stationary interarrival time is non-decreasing.

(III) The squared coefficient of variation of the event-stationary process is greater than or equal to unity.

(IV) There is a stochastic order showing that the time-stationary interarrival time dominates the event-stationary interarrival time.

Then, for general MMPPs, we show that it is easy to establish overdispersion of the counting process, while (II) is false due to a counter-example. For general simple point processes, (III) follows from (IV). For MMPPs we conjecture and numerically test that (IV) and thus (III) hold. Importantly, modeling folklore has often treated MMPPs as “bursty" and implicitly assumed that (III) holds. However, this is still an open question.

Keywords

Markovian arrival processes (MAPs); Markov modulated Poisson process (MMPP); Markov switched Poisson process (MSPP); Overdispersion; Hazard rate; Squared coefficient of variation; Interevent times; Stationary process

“There’s gold in them thar coefficients!”: The strange story of mineral prospectivity analysis

Adrian Baddeley K

Department of Mathematics and Statistics, Curtin University, Australia

Mineral prospectivity analysis is a technique used by geoscientists to identify the places where it is most likely to find deposits of valuable minerals such as gold. It is intimately connected to statistical methods including generalized linear models and spatial point process models, and to other fields including spatial ecology. This talk tells the strange story of prospectivity analysis, including a brilliant early insight from John Tukey, great practical successes and failures, some very confusing misinterpretations and digressions, and the exciting modern challenges.

Keywords

Spatial statistics; GLMs; Mineral prospectivity

Decoupled shrinkage and selection of sparse factor models

Amir Bashir[1], **Adam Smith**[1], **Beatrix Jones**[2]

[1] School of Natural and Computational Sciences, Massey University, New Zealand

[2] Department of Statistics, University of Auckland, New Zealand

Bayesians desiring a sparse model frequently formulate priors that enhance sparsity. However, in most settings, this leads to sparse posterior samples, not a sparse posterior mean. We address this issue in the context of sparse factor analysis. Decoupled shrinkage and selection post-processes the posterior mean covariance matrix to produce a sparse factor model whose expected fit lies within the upper 95% of the posterior over fits. In the Gaussian setting, simulation studies show this approach is competitive with *fanc* (factor analysis using non-convex penalties, Hirose and Yamamoto, 2015) in terms of sensitivity, specificity, and picking the correct number of factors. Decoupled shrinkage and selection is also easily applied to models where a latent multivariate normal underlies non-Gaussian marginals, e.g., multivariate probit models. We illustrate our findings with moderate and high dimensional data examples from modelling of food frequency questionnaires and fish abundance.

Keywords

Factor loadings; Posterior distribution; Sparse

Administrative records in the 2018 Census dataset

Christine Bycroft, Nathaniel Matheson-Dunning, Megan Gath, Abby Morgan

Statistics New Zealand

For the first time, the 2018 New Zealand Census includes administrative records to count people who were missed by the census field collection, replacing the use of 'substitute' imputed records in previous censuses. This innovative approach was developed because of a lower than expected response rate to the 2018 census field collection. The final New Zealand 2018 Census dataset is made up of census forms received through field collection, supplemented by high-quality administrative records for people we have not received a response from. These admin enumerations are drawn from a New Zealand resident population that has been derived from administrative data. New methods have been developed specifically for the 2018 Census application to account for known limitations of the administrative sources. The final 2018 Census dataset consists of 89 percent census responses and 11 percent admin enumerations.

This presentation provides an overview of the statistical methods used to determine high quality administrative records for inclusion in the 2018 New Zealand Census dataset and illustrates some results.

Keywords

2018 Census; Administrative data; Non-response

Dual system estimation combining census responses and an administrative population

Sabrina Rowlatt, Christine Bycroft, Nathaniel Matheson-Dunning

Statistics New Zealand

Dual system estimation (DSE) combining 2018 census responses and an administrative population has been used to derive a new estimate of the NZ resident population at the time of the 2018 Census. Application

of DSE based on these two large, but incomplete lists poses different challenges from those usually encountered by national statistics offices when applying DSE in the context of a census and coverage survey. The most critical assumptions to address were erroneous inclusions in the administrative population, and missed linkages between the two data sources.

This presentation describes how dual system estimation has been applied in this novel context, with a focus on meeting key assumptions that underpin the method.

Keywords

Dual system estimation; 2018 Census; Administrative data

Spatial capture-recapture—applications to acoustic surveys of cetacean populations

David K. Chan[1], **Ben C. Stevenson**[1], **Regina A. Guazzo**[2], **Tyler A. Helble**[2] S

[1] Department of Statistics, University of Auckland, New Zealand

[2] Naval Information Warfare Center Pacific, USA

Acoustic surveys are rapidly becoming one of the most common ways to assess cetacean populations, and can be vastly cheaper than other alternatives, like visual surveys. This is because visual surveys have low detection rates, even under ideal conditions. However, acoustic surveys can detect large numbers of vocalising individuals, if the monitoring devices and survey design are appropriate for the target population. Acoustic spatial capture-recapture (ASCR) can estimate call density from these surveys. ASCR has not been widely applied to acoustic surveys of cetacean populations although it is acknowledged in the literature.

In this talk I describe an application of ASCR to estimate call density of eastern North Pacific gray whales, *Eschrichtius robustus*, and how it varies with environmental covariates like ocean depth and distance from coast. Our findings are broadly consistent with those of Guazzo et al. (2017), who described the data set and estimated homogeneous call density within the hydrophone array using a method that analysed a subset of calls that could be accurately localised. Our analysis highlights two considerable advantages of ASCR over other established methods in the literature: calls do not have to be localised to be included in the analysis, and it allows estimation of an inhomogeneous call density surface.

References

Guazzo, R. A., Helble, T. A., Gerald, L. D., Weller, D. W., Wiggins, S. M., & Hildebrand, J. A. (2017). Migratory behavior of eastern North Pacific gray whales tracked using a hydrophone array. *PloS one*, *12*(10), e0185585.

Keywords

Spatially explicit capture recapture; Acoustic survey; Passive acoustic monitoring; Call density; Eastern North Pacific gray whales

Harmonic Data Science team building experience and university engagements

Lisa Chen, **Glenn Thomas**

Harmonic Analytics, New Zealand

Harmonic is an NZ-based Data Science company operating since 2003. We are currently active across many sectors. In recent years, there has been an increasing interest from the industry seeking to build their own Data Science teams. This leverages both the technical expertise of Harmonic Data Scientists and the

contextual expertise from clients. Some of these engagements further extend to even sourcing budding Data Scientists from universities. In this talk, we will share our experience in helping organisations develop their own Data Science capabilities through customised training, professional mentoring programmes, internships and other university-related engagements.

Keywords

Two-phase design; Optimal sampling; Multi-wave sampling; Prior; Efficiency

Optimal multi-wave sampling for regression modelling

Tong Chen, Thomas Lumley, Claudia Rivera-Rodriguez S

Department of Statistics, University of Auckland, New Zealand

Two-phase designs involve measuring extra variables on a subset of the cohort where some variables are measured. The goal is to choose a subsample of people from the sampled sub-cohort and analyse that subsample efficiently. There is a large body of literature on statistical inference for two-phase designs. However, compared with estimation methods, there is much less attention focused on the design aspect. It is desirable to obtain an optimal design which ends up with the most efficient estimation. In this talk, I will firstly introduce two-phase sampling and optimal sampling strategies. After that, I will talk about multi-wave sampling and focus on design-based estimation without making strong assumptions about the model. Finally, I will show that efficiency can be further improved using priors.

Keywords

Two-phase design; Optimal sampling; Multi-wave sampling; Prior; Efficiency

Assessing heterozygous balance in NGS DNA profiles

Kevin Cheng[1,2], Jo-Anne Bright[2], John Buckleton[1,2], James M. Curran[1] S

[1] Department of Statistics, University of Auckland, New Zealand

[2] Institute of Environmental Science and Research, New Zealand

In forensic casework, DNA is typically extracted from crime samples. Specific loci in the extracted DNA containing short tandem repeats (STRs) are amplified using the polymerase chain reaction (PCR) technique. The amplified products are separated by specific florescent markers and a capillary-electrophoresis (CE) technique. This separation of amplified products are visualized as an electropherogram (epg). Peaks visualized in an epg are typically analyte signals corresponding to one or more alleles at a locus and the fluorescence of these signals are measured in relative fluorescent units (RFUs).

In a perfect system, a pair of heterozygote alleles present in an individual at a given locus would demonstrate equal amounts of fluorescence. This can be expressed as having a heterozygous balance (*Hb*) of one. However, there is increasing variation introduced at each stage of the process from sampling to the visualization of a DNA profile. This results in some stochastic variation between the analyte signals of two alleles of a heterozygote and variation in *Hb*.

Through years of experience the forensic community have developed a good understanding of *Hb* to assist with their interpretation of crime samples. Within literature, Kelly et al. (2012) investigated the relationship between the heterozygous balance (*Hb*) and the average peak height (*APH*). They also demonstrated that the average peak height has minimal impact on the heterozygous balance, but the variance of the heterozygous balance can be modelled as inversely proportional to the average peak height.

There has been increasing interest from the forensic community to utilize new sequencing technologies, known as next-generation sequencing (NGS) or massive parallel sequencing (MPS), to amplify similar STR loci. This technique incorporates additional processes and the resulting analyte signals are measured as read counts. Whilst it has been theorized that read counts are equivalent to RFUs, there are more steps within this process that could result in increasing variability.

Using similar techniques as Kelly et al. (2012), we assess the variability of *Hb* in NGS DNA profiles. As expected, we demonstrate that NGS profiles have greater variability in *Hb* and could be modelled as inversely proportion to average read count.

Keywords

NGS; MPS; STRs; Forensics; Heterozygous balance

The need for speed in Genomics: Comparing Bayesian algorithms to estimate polygenic effects in tropically adapted beef cattle

Roy Costilla

Centre for Animal Science, Queensland Alliance for Agriculture and Food Innovation University of Queensland, Australia

With the advent of new genomic technologies comes the need to develop new statistical and computational algorithms that can handle large amounts of data in Genetics. Within the Bayesian paradigm, current methods to estimate polygenic effects for complex traits rely mostly on Markov chain Monte Carlo (MCMC) methods and Gibbs sampling. These approaches are not necessarily scalable to big datasets as the computation time grows more than linearly with sample size. This means that huge computational resources, in terms of RAM memory and/or computing time, need to be used to fit such models.

Here, we compare the performance of alternative algorithms when estimating polygenic effects for complex traits. In addition to MCMC using Gibbs sampling in BayesR, two algorithms are compared: Hamiltonian Monte Carlo (HMC) and Automatic Differentiation Variational Inference (ADVI) implemented in the STAN software. The proposed algorithms are compared using simulated data and real data. In particular, we use data from over 4,500 animals and 50,000 SNPs in a case study of animal temperament for tropically adapted beef cattle in Australia.

Keywords

Genetics; Variance components estimation; BayesR; STAN; Animal science

Estimation of Y haplotype frequencies with lower order dependencies

Mikkel Meyer Andersen[1,2], Amke Caliebe[3], Katrine Kirkeby[1], Maria Knudsen[1], Ninna Vihrs[1], James M. Curran[4]

[1] Department of Mathematical Sciences, Aalborg University, Denmark

[2] Section of Forensic Genetics, Department of Forensic Medicine, University of Copenhagen, Denmark

[3] Institute of Medical Informatics and Statistics, Kiel University, University Hospital Schleswig-Holstein, Germany

[4] Department of Statistics, University of Auckland, New Zealand

Estimating Y haplotype population frequencies is a demanding task in forensic genetics. Despite the suggestion of various methods, it has not been successfully solved yet. At the basis of this problem is the

complex dependency structure between the involved STR loci. Here, we approximate this structure by the use of specific graphical models, namely t-cherry junction trees.

We apply trees of order three by which dependencies between three STR loci can be taken into account, thereby extending the Chow-Liu method which is restricted to pairwise dependencies. We show that the t-cherry tree method outperforms the Chow-Liu method as well as the well-established discrete Laplace method in estimation accuracy.

Keywords

Frequency estimation; Y haplotypes; Graphical models; Forensic genetics; Chow-Liu algorithm; Discrete Laplace method; t-cherry junction trees

New Zealand Crime and Victims Survey: Filling the knowledge gap

Tadhg Daly, Michael Slyuzberg, Tianying Chu, Kate Preston, James Swindells

Research and Evaluation, Sector Group, New Zealand Ministry of Justice

The New Zealand Crime and Victims Survey (NZCVS) is a nationwide, face-to-face annual survey. It asks 8,000 randomly selected New Zealanders aged 15 years and over, about incidents of potential crime that they experienced in New Zealand over the last 12 months. This includes both incidents reported to the Police and unreported incidents. These incidents are later assessed and categorised by expert coders. The NZCVS provides information for researchers, policy makers and the public about the nature and extent of crime and victimisation in New Zealand.

This presentation will discuss the history and formation of NZCVS, our survey methodology and our results from the first year of data collected in 2018. The results presented will include statistics on a range of areas including crime rates, victim demographics, family violence, highly victimised people, drivers of crime and reporting to police.

Keywords

Crime statistics; Survey design; Victims

Multiple imputation through denoising and variational autoencoders

(Agnes) Yongshi Deng S

Department of Statistics, University of Auckland, New Zealand

Missing values are ubiquitous in clinical and social science data. Incomplete data not only leads to loss of information but can also introduce bias, which poses a significant challenge for data analysis. Various imputation procedures were designed to handle incomplete data under different missingness mechanisms. Rubin (1977) introduced multiple imputation to attain valid inference from data with ignorable nonresponse. Some techniques and `R` packages are developed to implement multiple imputations, such as MICE, Amelia and MissForest. However, the running time of imputation using these methods can be excessive for large datasets. We propose a scalable multiple imputation method based on variational and denoising autoencoders. Our `R` package `misle` is built using the `tensorflow` package in `R`, which enables fast computation and thus provides a scalable solution for missing data.

In this presentation, I will demonstrate some features of the `R` package `misle` and compare the performance of several commonly used multiple imputation techniques. Multiple imputation inference will also be discussed.

Keywords

Multiple imputation; Missing data; Variational autoencoders; Denoising autoencoders; `misle`

From Near and Far: Use of alternative sources and imputation for item non-response in 2018 Census

Jamas Enright

Statistics New Zealand

Previous censuses in New Zealand included only limited adjustment for non-response to questions (item non-response), with the core demographic variables age, sex and place of usual residence the only variables where all missing data was imputed. Rates of missingness for variables in the 2013 Census ranged between about 5 percent and 10 percent, with some groups disproportionately affected. The 2018 Census strategy included use of alternative sources as a mitigation for item non-response and statistical imputation for more variables than we had in earlier Censuses, with the aim of improving data quality.

The use of alternative sources and imputation for census characteristics has taken on greater significance than originally anticipated given the lower than expected response rate to the census field collection. The final 2018 census dataset includes 4 percent partial census responses and 11 percent of people added through administrative records, for whom we have no information directly from census individual forms.

In this talk, I will discuss the methods we used to fill in the missing variable information. In order of priority, these were historical (2013 Census) data for variables which were expected to remain fairly stable over time, and good quality administrative data, based on the findings of the longer-term administrative data research. For variables which have good imputation properties, we also made use of nearest neighbour imputation, using the software Canadian Census Editing and Imputation System (CANCEIS), to find donors to copy information from.

For CANCEIS, variables were grouped into modules for easier processing and maintaining correlation. This talk will explain the logic of these groups and how CANCEIS handles them. For variables where no alternative sources are available, the levels of imputation are higher than usual for a census, especially for population groups with the lowest response rates to the census field collection. This talk will also explore the implications of the degree of imputation.

Keywords

Census 2018; Imputation; CANCEIS

Recent developments in model averaging

David Fletcher

Department of Mathematics & Statistics, University of Otago, New Zealand

Model averaging is an approach to allowing for model uncertainty in estimation. I will provide an overview of interesting developments in this area, from both a frequentist and Bayesian perspective.

Keywords

Bayesian model averaging; Frequentist model averaging; Model uncertainty

Likelihood-based clustering of commercial fisheries data

Zach Goeden, Louise McMillan, Shirley Pledger, Richard Arnold

School of Mathematics and Statistics, Victoria University of Wellington, New Zealand

Commercial trawl data present a valuable opportunity for fisheries stock assessment and for measurement of changes over time, such as may be seen as a result of climate change, when species may move. The data include trawl location, gear type, and biomass of fish caught by species, which we have initially coded as presence/absence for the 25 top occurring species. We use a likelihood-based method which clusters the trawls into groups with similar species patterns. The method uses finite mixture models, estimated by the EM algorithm, returning fuzzy membership indicators. The method is implemented in the new R package clustglm, which allows for interactions between gear, species type and cluster membership. It is, however, computationally intensive given the expected large number of clusters, and we have been investigating ways to make the method feasible at scale. These include an improved choice of starting points for the EM algorithm, and hierarchical top down splitting of clusters. This latter method allows for reallocations of events after splits to reduce the risk of misallocation. We will present some initial results from clustering of a trawl data set from NZ commercial fisheries.

Keywords

Clustering; EM algorithm; Fisheries

Bayesian linear mixed model with multiple random effects for risk prediction of complex traits

Yang Hai S

Department of Statistics, University of Auckland, New Zealand

Risk prediction capitalizing on emerging human genome discoveries holds great promise for improved health-care and precision medicine (Ashley et al., 2015). Most existing methods for genetic risk prediction (e.g., gBLUP) make an assumption on the effect size distribution, which is usually unknown in advance for most complex traits (Chatterjee et al., 2016). In this talk, I will present a Bayesian linear mixed model with multiple random effects (BLMM-MR) to model disease risk based on high-dimensional genetic data, where the shape of true effect size distributions of genetic variants can be resembled. BLMM-MR first groups predictors based on genomic annotations, and then uses a hybrid of linear mixed model and sparsity regression to select and model their predictive effects. Through extensive simulations, we have demonstrated that the BLMM-MR obtained better prediction performance than most of the existing methods. We further illustrated the BLMM-MR method through an application to a whole-genome sequencing data obtained from Alzheimer's Disease Neuroimaging Initiative study.

Keywords

Hybrid model; Multiple random effects; Fixed effects selection; Random effects selection; Risk prediction

Dual systems estimation and linkage error

Patrick Graham, Rebecca Green, Sarah Hall, Matt Hendtlass, Abby Morgan, Andrew Richens, Lucianne Varn

Statistics New Zealand

Dual systems population estimation is equivalent to a capture-recapture analysis of a closed population using two incomplete listings of the population. Variants of the idea have been widely used in official statistics for some time, but generally only for population estimates.

We are investigating methods for using a Bayesian Gibbs sampler with census responses and administrative data to produce a unit record file of the New Zealand usually resident population. A key assumption of dual systems estimation is perfect linking between the two lists.

In this presentation I outline our Bayesian dual systems estimation model. I talk to the challenge of linkage error, and the linkage validation study we carried out to help adjust for linkage error. I will also delve into our results.

Keywords

Dual systems estimation; Population unit record files; Record linkage; Linkage error

Dynamic polytope samplers for linear inverse problems

Martin L. Hazelton K

Department of Mathematics and Statistics, University of Otago, New Zealand

Statistical inverse problems occur when we wish to learn about some random process that is observed only indirectly. Inference in such situations typically involves sampling possible values for the latent variables of interest conditional on the indirect observations. This talk is concerned with linear inverse problems for count data, for which the latent variables are constrained to lie on the integer lattice within a convex polytope (a bounded multidimensional polyhedron). An illustrative example arises in transport engineering where we observe vehicle counts entering or leaving each zone of the network, then want to sample possible interzonal patterns of traffic flow consistent with those entry/exit counts. Other areas of application include inference for contingency tables, and capture-recapture modelling in ecology.

Such sampling can be conducted using Markov chain Monte Carlo methods, through a random walk on the lattice polytope. A major challenge is finding a set of basic moves (sampling directions) so as to ensure that the walk can connect any two feasible points without leaving the polytope. In principle this can be done by computing a Markov basis of potential moves, but in practice the resulting sampler can be hugely inefficient even when such a basis is computable. In this talk I will describe some current work on developing a dynamic Markov basis that generates moves on the fly. This approach can guarantee irreducibility of the sampler while gaining efficiency by increasing the probability of selecting serviceable sampling directions.

Keywords

Inverse problem; Lattice polytope; Markov basis; MCMC

Statistical phylogenetics: An introduction and some open problems

Barbara Holland K

School of Physical Sciences, University of Tasmania, Australia

Molecular phylogenetics is concerned with estimating the evolutionary relationships amongst species on the basis of molecular sequence data. Statistical approaches to this problem have been developed where evolution is modelled as a Markov process acting on a tree. Given a specific tree along with parameter choices in our Markov model, these models allow us to calculate the probability of seeing any particular pattern of nucleotides at the tips of the tree (i.e. in the modern-day species). We can then calculate a likelihood score for a sequence alignment given our model; from here we can use maximum likelihood or Bayesian approaches to find choices of tree and model parameters that give the best explanation of the sequence data. This provides a nice framework for inference, but there are some interesting challenges that arise.

Our model is a mix of combinatorial structure (the branching pattern in the tree) and continuous parameters (e.g. rates of change from one particular nucleotide to another, lengths of edges in the tree), this makes it difficult to assess the uncertainty in our estimates, i.e. what is 95% CI for a tree?

It is a common maxim that just because a model fits "best" doesn't mean it fits "well". However, it is surprisingly difficult to do anything analogous to residual diagnostics in a phylogenetic context. Our models give us a multinomial distribution on site patterns. Even for just 10 species there are $4^{10} \approx 1$ million possible site patterns, and in any particular sequence alignment most will not be observed. We need more tools to help us visualise how well our models fit and where they break down.

How sure are we that evolution is tree-like anyway? There are many biological processes (hybridisation, recombination, lateral gene transfer) that cannot be well modelled by a tree. As soon as we move away from the tree assumption we need to think harder about model selection issues as statistical identifiability can become a problem.

This talk will give an introduction to statistical phylogenetics and survey some of the above open problems.

Keywords

Molecular phylogenetics; Model selection, Residual diagnostics

Recycling genome-wide test statistics: Can you really find more information?

John Holmes[1], **David Balding**[1], **Doug Speed**[2]

[1] School of Mathematics and Statistics, University of Melbourne, Australia

[2] Aarhus Institute of Advanced Studies, Aarhus University, Denmark

Considerable resources have been placed into genome-wide studies for a wide variety of traits over the last decade. To date, the results obtained from these studies have been disappointing, with fewer variants found to be associated than expected. In this paper, we will discuss some techniques that have been developed to tackle this 'missing heritability' problem. We will firstly highlight the close relationship these techniques have to ANOVA based methods of variance component estimation. Secondly, we will demonstrate both mathematically and with simulation how sensitive the resulting estimates are to a) different types of discrepancy between the true and fitted model, and b) participant selection.

Keywords

Genome-wide association studies; Model mis-specification; Test statistics; Heritability estimation

Modelling the emergence dynamics of the western corn rootworm beetle

Rodelyn Jaksons[1,2], **Katharina Falkner**[3], **Elena Moltchanova**[2]

[1] Plant and Food Research, New Zealand

[2] School of Mathematics and Statistics, University of Canterbury, New Zealand

[3] University of Natural Resources and Life Sciences, Austria

The Gompertz curve was originally used to model human population mortality. It is a sigmoidal curve where the beginning and end of a period shows the slowest time for growth, and adequately describes observed dynamics of many phenomena. Since its inception it has been used in a wide range of applications such as population biology, metabolic curves, growth of tumours etc. We propose the use of the Gompertz function in a Bayesian Hierarchical framework to model the emergence dynamics of the western corn rootworm beetle, which is an invasive species in Europe. Furthermore, I will also discuss how we can also include spatial effects to capture spatial autocorrelation, to infer how these dynamics vary over space. We apply the model to Austrian monitoring data collected in 2014.

Keywords

Gompertz curve; Spatial analysis; Emergence dynamics

Longitudinal analysis of dietary patterns

Beatrix Jones[1], **Clare Wall**[1], **John Thompsom**[1], **Larisa Morales**[2]

[1] Department of Statistics, University of Auckland, New Zealand

[2] Universidad Naciónal Autonoma de México

When study participants complete a food frequency questionnaire, it is common to summarise their diet with a few principal components. We look at a method to extend this to track dietary consumption over time, in the context of the Auckland Birth Cohort Study. The children in this study had their diet assessed at ages 3.5, 7, 11, and 16. There are some changes to the questionnaire used over this time period, as well as to the typical foods eaten (e.g. coffee and tea figure much more prominently for 16 year olds than 3.5 year olds). We consider JIVE (Joint and Individual Variance Estimation), which was developed for looking at high throughput molecular data from different sources, to summarize diet variability that is shared across the four ages. We augment this strategy with visualisations to understand to what extent diets follow a predictable trajectory through childhood and adolescence.

Keywords

Principal components; Food frequency questionnaire; Joint and individual variance estimation

Construction of multi-allelic genomic relationship matrices using Genotyping-by-Sequencing (GBS) data

Jie Kang[1], **Phillip Wilcox**[1], **Michael Black**[1], **Dan Milbourne**[2], **Stephen Byrne**[2], **Ken Dodds**[3]

[1] Department of Mathematics and Statistics, University of Otago, New Zealand

[2] Teagasc, Ireland

[3] AgResearch, New Zealand

Advances in sequencing technologies enable us to characterise variation in the genome of non-model but agriculturally important species. Approaches such as Genotyping-by-Sequencing (GBS) can produce abundant markers at relatively low cost. This has encouraged implementation of Genomic Selection (GS) to accelerate genetic gains in animal and plant breeding, but has also raised the challenge of how to make better use of the genomic information.

Unlike animal breeding, where high-quality reference genomes and well-developed modelling strategies already exist, a versatile analysis pipeline is needed for out-breeding plant species, such as perennial ryegrass (*Lolium perenne*). In addition, we want approaches that take the highly polymorphic nature of ryegrass into consideration when analysing (low-depth) GBS data. We thus hypothesise that existing GS models can be enhanced by accounting for short haplotypes or 'ShortHaps', that is, multiple variants in small genomic segments such as those captured within a GBS read.

In this study, we explored different ways to construct multi-allelic genomic relationship matrices using ShortHaps, and evaluated the performance of these methods within the genomic best linear unbiased prediction (GBLUP) framework.

Keywords

Short haplotypes (ShortHaps); Genomic relationship matrices (GRMs); Genotyping-by-Sequencing (GBS); GBLUP; Genomic delection (GS)

Constrained estimation using group-summary level information for correlated data

Yujin Kim

Department of Statistics, University of Auckland, New Zealand

In resource-limited settings, a cost-efficient approach is to implement a two-phase design and make use of big data information available from the first phase. This information is generally available from large data sources such as census data and disease registries and it is also obtainable from public websites as a form of grouped-level information.

Some existing methods for the the analysis of two-phase samples such as calibration or estimation of sampling weights have been shown to yield large improvements in efficiency, as well as a recently proposed method called constrained maximum likelihood. This method builds regression models based on individual-level data from an internal study while using summary-level information from an external big data source (first phase). A set of general constraints are proposed to link internal and external models. The methodology is cleaner because the investigator does not intervene in the process. The approach is well established, but only when participants are assumed independent. In situations when participants are naturally clustered (families, clinics), existing methods can yield concerning and invalid inference and conclusions. Thus, we aim to fill this gap by developing novel statistical methods that

efficiently analyze clustered data from two-phase studies. This work aims to develop constrained maximum likelihood estimation for the case where the first phase is cluster-correlated and when the external model is a cluster-level model.

Keywords

Two-phase designs; Cluster-correlated data; Constrained maximum likelihood

Using a mixture model to estimate the frequencies of Y-chromosomal DNA profiles

Maarten Kruijver[1], **Duncan Taylor**[2]

[1] Institute of Environmental Science and Research, New Zealand

[2] Forensic Science South Australia, Australia

A common problem in forensic genetics is to estimate the frequency of a DNA profile in a population. This problem is especially challenging for Y-chromosomal profiles, because father and son have the same profile (barring mutations), which precludes the use of models based on independence of alleles between markers. The profile is inherited completely. Although one can estimate the frequency of a profile in a population by the observed frequency in a sample of the population, this is inefficient because most profiles in the population sample will be observed only once in the sample, at least for contemporary profiles on 17 DNA markers. In recent years, model-based approaches have been suggested that estimate of the frequency of the 17-dimensional data either by exploiting conditional independence of markers (e.g. using Chow-Liu trees) or by using a mixture model. The discrete Laplace model is a mixture model with a small number of central profiles that models the frequency marker-by-marker conditionally on each central profile with a discrete Laplace distribution around the central allele at each marker in the central profile. We discuss how to apply this model and extend the approach to handle missing data and duplicated data. The model is demonstrated on a published dataset of profiles of Australian males.

Keywords

Forensic genetics; Y-chromosomal profiling; Haplotype frequency; Multivariate analysis

Calibration procedures for approximate Bayesian credible set

Kate Lee

Department of Statistics, University of Auckland, New Zealand

We develop and apply calibration procedures for checking the coverage of Bayesian credible intervals estimated using Monte Carlo methods. We focus on estimating the bias in coverage caused by approximations adopted in implementing an otherwise asymptotically exact Monte Carlo method. For example, in Metropolis Hastings MCMC, the likelihood may have been replaced by an estimate in the Hastings ratio. In our setting the coverage is asymptotically exact when the user's exact prior and likelihood are used in an asymptotically exact Monte Carlo method, and otherwise it may be distorted.

Keywords

Bayesian inference; Approximation; Coverage; Diagnostic test

On the selection of predictors by using greedy algorithms and information theoretic criteria

Fangyao Li, Ciprian Doru Giurcăneanu, Chris Triggs S

Department of Statistics, University of Auckland, New Zealand

Over the last three decades, much work has been focused on the problem of high-dimensional (HD) statistical modelling, where the sample size (n) is much smaller than the number of covariates (p_n). It has been shown that the greedy algorithms efficiently generate candidate models in the HD case as well as in the big data (BD) case ($n \gg p_n$). In this talk, we will present five greedy algorithms: Matching Pursuit Algorithm (MPA), Orthogonal Matching Pursuit (OMP), Relaxed Matching Pursuit (RMP), Frank-Wolfe Algorithm (FWA) and Constrained Matching Pursuit (CMP).

In the previous literature, either cross-validation (CV) or information theoretic (IT) criteria have been applied for selecting the "best" model from the family of models produced by a greedy algorithm. The IT criteria are less computationally intensive than CV; the key part in their evaluation is the calculation of the degrees of freedom. We will briefly discuss our theoretical results concerning the properties of the hat matrix for MPA and RMP. They enable the calculation of the degrees of freedom and help to alter the IT criteria used in classical linear regression such that to become compatible with these algorithms.

A comparison of the computational complexity of the five greedy algorithms will be provided. If the time will permit, we will discuss the performance of the algorithms in experiments with air pollution data.

Keywords

High-dimensional modelling; Greedy algorithms; Information theoretic criteria

Comparison of stopping criteria for the EM algorithm

Lingyu Li, Ivy Liu, Richard Arnold S

School of Mathematics and Statistics, Victoria University of Wellington, New Zealand

This talk discusses different stopping criteria for Expectation-Maximisation algorithm. Our model is model-based clustering using adjacent-categories logit model via finite mixture models for ordinal data. The EM algorithm is guaranteed to converge through iterations of calculation. While we use the EM algorithm to estimate the parameters of the model, convergence means all the parameters estimates become stable, and no further improvements can be made to the likelihood value. However, in reality, to avoid the long computation time, we need to set the stopping criteria to stop the algorithm before it completely converges. In this talk, we propose several stopping criteria for the EM algorithm and then use a simulation study to illustrate the performance of different stopping criteria.

Keywords

Clustering; EM algorithm; Finite mixture model; Convergence

The development of stereotype models for ordinal data

Ivy Liu K

School of Mathematics and Statistics, Victoria University of Wellington, New Zealand

The collection and use of ordinal variables are common in many studies. Since about 1980, several models have been proposed to analyse ordinal response variables, such as proportional odds models,

adjacent-categories logit models, and stereotype models. This talk focuses on the advantages of the ordered stereotype model, which is more flexible than the most common ordinal models. The additional score parameters added in the ordered stereotype model provide the spacing information for the response categories. It might be interesting to incorporate the fitted spacing information for further model diagnostics or analyses. The recent development of the use of the spacing information will be discussed.

Keywords

Ordinal; Score; Stereotype model

Pairwise likelihood for multilevel models in multistage samples

Thomas Lumley, Xudong Huang

Department of Statistics, University of Auckland, New Zealand

Reweighting the log-likelihood for a linear mixed model to account for complex sampling is difficult: it is not clear where to put the weights, and the log-likelihood involves the population precision matrix, which is not (in general) computable from sample data. For the special case where the clusters in the model and in the sampling design are the same, J.N.K. Rao and co-workers have proposed an approach based on pairwise composite likelihood. Motivated by a genetics application in the Hispanic Community Health Study, we show that this method can be used for general models and designs. We present a new R implementation for the case of multilevel linear models in multistage samples, evaluate it in simulations, and describe how we plan to extend it to the motivating genetic application.

Keywords

Mixed model; Composite likelihood; Complex sampling; Genetics

R user-friendly functions

Harel Lustiger

Harmonic Analytics, New Zealand

Statisticians are typically not Software Engineers, nor are they taught good practices of software engineering. And yet, Statisticians often take the role of Software Engineers when they program functions in R.

This talk covers major computer programming principles and good practices that help to make R functions intuitive to other R users and our future selves.

As Statisticians try a function from any R package, they face two gaps: first, they try to figure out how to use the function; and second, they try to understand whether their actions with the function get them to their goal.

The role of R programmers is to help Statisticians bridge these two gaps. For example, the second gap is small when the function provides informative feedback about its failure, execution or success. This feedback should be easy to interpret and should match the way the Statisticians think about the function.

The takeaway of the talk is actionable principles which can be applied to design of R functions.

Keywords

R; NZSA2019; Programming

Enhancing regression nomograms to visualise regression models

Roger J. Marshall

Department of Epidemiology and Biostatistics, University of Auckland, New Zealand

Statistical analysis is now dominated by regression modelling, in its many guises. Since models are multidimensional, their visualisation is difficult but one way to appreciate their nature is with "regression nomograms". For example, the R `rms` package has a `nomogram` command. Regression nomograms were intended, as the name suggests, for graphical calculation, though since manual calculation is still required, an app would be a lot easier. Nomograms do, however, provide a useful way to "see" a regression model. I will discuss their use and present some enhancements to basic nomograms. Enhancements include superimposing data values and covariate distributions of different types (density, box plots, etc) - and making on-the-fly changes to these - showing random effects, and showing spline and other functions. I will introduce an R package (`regplot`) that accepts many classes of R regression objects, including `rms`, and incorporates these enhancements. Examples will include survival models.

Keywords

Regression nomogram; Visualisation; Covariate distribution

Spectral density estimation using P-spline priors

Patricio Maturana-Russel, Renate Meyer

Department of Statistics, University of Auckland, New Zealand

Useful information about a stationary time series is encoded in its spectral density function. We propose a Bayesian approach to estimate it based on a mixture of P-spline distributions. Our proposal is motivated by the B-spline Dirichlet process prior in combination with the Whittle's likelihood and aims at reducing the high computational complexity of its posterior computations. The strength of the B-spline Dirichlet process prior over the Bernstein-Dirichlet process prior lies in its ability to estimate spectral densities with sharp peaks and abrupt changes due to the flexibility of B-splines with variable number and location of knots. Here, we suggest to use P-splines that combine a B-spline basis with a discrete penalty on the basis coefficients. We demonstrate in a simulation study and two real case studies that this approach retains the flexibility of the B-splines, achieves the same ability to accurately estimate peaks due to a new data-driven knot allocation scheme but significantly reduces the computational costs.

Keywords

P-spline prior; B-spline prior; Spectral density estimation; Whittle likelihood

Directional measures and visualization for population differentiation

Louise McMillan[1], Rachel Fewster[2]

[1] School of Mathematics and Statistics, Victoria University of Wellington, New Zealand

[2] Department of Statistics, University of Auckland, New Zealand

McMillan & Fewster (2017) proposed GenePlots, a visualization method for genetic assignment of individuals. We now advance these techniques further by proposing methods for quantifying population genetic structure, and associated tests of significance. The measures we propose are closely related to

GenePlots, and enable visual features obvious from the plots to be expressed more formally. They are derived using the saddlepoint method to approximate the genetic distribution of a population, based on the posterior allele frequencies.

One measure is the interloper detection probability: for two random genotypes arising from populations A and B, the probability that the one from A has the better fit to A and thus the genotype from B would be correctly identified as the 'interloper' in A. Another measure is the correct assignment probability: the probability that a random genotype arising from A would be correctly assigned to A rather than B.

Using permutation tests, we can test two populations for significant population structure. These permutation tests are sensitive to subtle population structure, and are particularly useful for eliciting directional features of the population structure, e.g. where one population has undergone extensive genetic drift but the other population has remained large enough to retain greater genetic diversity. We illustrate the new methods using microsatellite and SNP data.

References

McMillan, L. F. and Fewster, R. M. (2017), Visualizations for genetic assignment analyses using the saddlepoint approximation method. *Biometrics*, 73:1029–1041. doi:10.1111/biom.12667

Keywords

Population genetics; Genetic assignment; Saddlepoint approximation; Population differentiation

Nonparametric estimation of the random coefficients model using regularized maximum likelihood

Emil Mendoza, Fabian Dunker, Marco Reale

School of Mathematics and Statistics, University of Canterbury, New Zealand

The random coefficients model is written as $Y_i = \beta_{0i} + \beta_{1i} X_{1i} + \beta_{2i} X_{2i} + \ldots + \beta_{di} X_{di}$, with $\mathbf{X}_i$ and Y_i i.i.d, and $\mathbf{fi}_i$ independent of $\mathbf{X}_i$. This model is often used to capture unobserved heterogeneity in a population. In standard analyses using random coefficient models it is assumed that these random coefficients come from a normal distribution. To do away with this assumption, we employ a nonparametric method of estimation to reconstruct the joint density of the random coefficients, f_β. We propose a quasi-maximum likelihood method to estimate the joint density distribution of the random coefficient model. This method implicitly involves the inversion of the Radon transformation in order to reconstruct the joint distribution, and hence is an ill-posed problem. To add stability to the solution, we apply regularization methods. To show the quality of the results, we apply the developed algorithm on simulated data and real data.

Keywords

Random Coefficients; Non-parametric; Regularized maximum likelihood; Inverse problem

Vector generalized linear models and two-parameter links with applications to quantile regression and the NBD

Victor Miranda[1], **Thomas Yee**[2]

[1] Department of Mathematical Sciences, Auckland University of Technology, New Zealand

[2] Department of Statistics, University of Auckland, New Zealand

Vector generalized linear models (VGLMs) are extensions of GLMs and accommodate multiple linear predictors which individually involve a link function applied to a parameter. While VGLMs have shown to

offer advantages for different data types, we loosen the restriction of one-parameter linear predictors by developing the two parameter case. For instance, estimating the negative binomial distribution with its canonical link (called the NB-C2-2 model) had previously been unsatisfactory in the VGAM R package but has now been implemented correctly. In this talk we present details of the new methodology. The key is to use total derivatives when deriving the score vector and the working weight matrices, quantities which are essential requirements for the iteratively reweighted least squares/Fisher scoring algorithm. In addition to a correctly implemented NB-C2-2 model, we also develop direct quantile regression based on the normal distribution. By 'direct' it is meant that at least two specified quantiles are modelled by new link functions which enable the covariates to be more directly related with the quantiles. Coupled with a parallelism constraint, the quantile-crossing problem is overcome.

Keywords
Link function; VGLM; NBD; Quantile regression

Efficient estimation for the joint model of survival and longitudinal data

Khandoker A. Mohammad S

School of Mathematics and Statistics, Victoria University of Wellington, New Zealand

In survival studies it is important to record the values of key longitudinal covariates until the occurrence of event of a subject. For this reason, it is important to study the association between longitudinal and time-to-event (which can be assessed using Cox PH model) outcomes using joint models. Here profile likelihood approach has been used to estimate the cumulative hazard and regression parameters from the joint model of survival and longitudinal data. Moreover, we show the asymptotic normality of the profile likelihood estimator via asymptotic expansion of the profile likelihood and obtain the explicit form of the variance estimator. The estimators of the regression parameters from the joint model are shown to be semi-parametric efficient. The numerical result of our proposed method is shown by the 'aids' data from JM R-package where we have computed the standard errors of the estimated parameters from the efficient score function and efficient information matrix.

Keywords
Cox PH model; Efficient score function; Longitudinal data; Profile likelihood; Survival data

Hydrothermal time seed germination model: Non-linear regression vs. survival analysis

Elena Moltchanova[1], **Shirin Sharifiamina**[2], **Derrick J. Moot**[2], **Ali Shayanfar**[3], **Mark Bloomberg**[4]

[1] School of Mathematics and Statistics, University of Canterbury, New Zealand

[2] Field Research Centre, Lincoln University, New Zealand

[3] Plant Production, Gorgan University of Agricultural and Natural Resources, Iran

[4] School of Forestry, University of Canterbury, New Zealand

Hydrothermal time (HTT) models describe the time course of seed germination for seeds germinated under specific seedbed temperature and water potential conditions. In doing so, they represent the cumulative germination proportion $g(t)$ as a function of time since sowing. The germination experiment usually consists

of a number of identical dishes with n seeds each, which are subject to factorial combinations of seedbed water potential and temperature. Germination of seeds in the dishes is monitored at pre-specified time points. The cumulative proportion of seeds germinated in each dish at each time point is then recorded, and the parameters of $g(t)$ are usually estimated using the non-linear regression model. In order to account for repeatedly measuring the dishes, jacknife or bootstrap are used to obtain the standard errors for the parameter estimates.

The above approach, however, tends to disregard the data at the beginning (since no seeds will germinate during the initial "lag" phase of germination) and the end of the germination process, (since the maximum number of seeds possible will have germinated, and any remaining seeds are either dormant or dead). In either case there are no observations of germination to record. Alternatively, one may treat the data generating process within the time-to-event framework and apply survival analysis to estimate the parameters of interest and the associated uncertainty.

In this study, we use experimental germination data for two species of clover (white clover and subterranean clover), to illustrate the performance of the above methods. We also discuss the estimation in case of non-differentiable $g(t)$ and outline advantages of placing the problem within the Bayesian inference framework.

Keywords
Survival analysis; Bayesian modelling; Seed germination modelling

Goodness-of-fit of Bayesian spatial models using pivotal discrepancy measures

Lindsay Morris, Nokuthaba Sibanda S

School of Mathematics and Statistics, Victoria University of Wellington, New Zealand

Within the field of geo-statistics, Gaussian processes are a staple for modelling spatial and spatio-temporal data. Statistical literature is rich with definitions and estimations of the mean and covariance structures of these processes (in both frequentist and Bayesian contexts). However, considerably less attention has been paid to developing goodness-of-fit tests that allow for assessment of model adequacy. Jun et al. (2014) introduced a statistical test that uses pivotal discrepancy measures to assess the goodness-of-fit in the Bayesian context. We present a modification and generalization of their statistical test. The initial method involves spatial partitioning of the data, followed by evaluation of a pivotal quantity at each posterior draw to obtain a posterior distribution of pivotal statistics. Order statistics from this distribution are used to obtain approximate p-values. Jun et al. (2014) use arbitrary partitions based on pre-existing spatial boundaries. The partitions are made to be of equal size.

Our method uses a novel approach to partitioning based on K-means clustering and generalises Jun et al.'s approach to incorporate unequal partition sizes. Observations from a spatial or spatio-temporal process are partitioned using an appropriate feature vector that incorporates the geographic location of the observations into subsets (not necessarily of the same size). The method's viability is illustrated in a simulation study, and in an application to hoki (Macruronus novaezelandiae) catch data from a survey of the sub-Antarctic region.

References Jun, M., Katzfuss, M., Hu, J., & Johnson, V. (2014). Assessing fit in Bayesian models for spatial process. *Environmetrics*, **25**(8), 584–595.

Keywords
Spatio-temporal; Bayesian; Pivotal discrepancy measures

Post-stratified probability-proportional-to-size sampling from stratified populations

Omer Ozturk

Department of Statistics, Ohio State University, USA

This paper develops statistical inference based on a post-stratified probability-proportional-to-size (*pp*) sample from a finite population. A *pp* sample selects the sample units with selection probabilities proportional to their size and measures them for the characteristic of interest. For each measured unit, the *pp* sample further creates position information (rank) in a comparison set of size M. The sample is then post-stratified into ranking classes based on their position information in the comparison set. A *pp* sample is expanded to stratified populations by selecting a *pp* sample from each stratum population to form the stratified *pp* sample. Using this stratified *pp* sample we construct unbiased and Rao-Blackwell estimators for the mean of the stratified populations. Different sample size allocation procedures for stratum sample sizes are investigated. The new sampling design is applied to apple production data to estimate the total apple production in Turkey.

Keywords
Rao-Blackwell estimator; Stratified sampling; Post-stratified sample; Neyman allocation; Probability-proportional-to-size sampling

Archetypal analysis as an autoencoder with application to data imputation

Matthew F. Parry[1], **Pórtya P. Calvacanti**[2]

[1] Department of Mathematics and Statistics, University of Otago, New Zealand

[2] Universidade de São Paulo - Escola Superior de Agricultura "Luiz de Queiroz", Brazil

The idea of an autoencoder comes from machine learning but it is implicit in a number of statistical techniques. I give a brief review of autoencoders, focusing mainly on their use as a dimension reduction technique. I then pivot to the archetypal analysis of Cutler and Breiman (1994), which is a form of cluster analysis given in terms of extremal points, i.e. archetypes. Following Bauckhage et al. (2015), I show how archetypal analysis can be viewed as an autoencoder. I finish with an application of archetypal analysis to data imputation.

Keywords
Archetypal analysis; Autoencoder; Data imputation; Dimension reduction

Predicting breeding values after culling

Yijia Qian[1], **Daniel Gerhard**[1], **Luis Apiolaza**[2], **Peter Jaksons**[3]

[1] School of Mathematics and Statistics, University of Canterbury, New Zealand

[2] School of Forestry, University of Canterbury, New Zealand

[3] Plant and Food Research, New Zealand

Culling is the process of removing individuals from a group according to a group of desired or undesired

traits. It is a common process in breeding trials. In first-stage kiwifruit breeding, around 80% of the vines are removed for pre-harvest reasons such as PSA secondary symptoms, low/no floral yield, floral bud rot, unattractive size or shape. Another 10% are removed at harvest due to failing to meet fruit weight or dry matter content thresholds. The remaining 10% of vines will be selected further according to storage time and sensory characteristics. Bias can be introduce to breeding value estimation as the remaining vines are a non-representative sample of the initial population planted out due to culling.

Based on Monte Carlo Simulations, we are investigating the effect of culling on the accuracy of predicting breeding values and selecting promising genotypes. Trial management and sampling strategies are assessed for bias in the predictions of genotypic effects, using a multivariate mixed-effects model framework with missing values.

Keywords

Culling; Breeding values; Simulation

Measuring child poverty — Unleashing the power of data to make New Zealand the best place in the world to be a child

Diane Ramsay

Statistics New Zealand

The current Government has a vision to make New Zealand the best place in the world to be a child. This has involved development of a Child Wellbeing Strategy, launched in August 2019, and the passing in December 2018 of the Child Poverty Reduction Act. This Act requires Government to set 3- and 10-year targets for the level of child poverty according to 4 measures of poverty defined in the Act. The Act also requires the Government Statistician to report annually on the level of child poverty to monitor progress towards these targets. The Government Statistician's first report is due to be published in February 2020. This talk will discuss Stats NZ's work to develop data sources to measure child poverty, some of the measurement issues that needed to be addressed and the methodological decisions that have been made. Some innovative work involving the use of administrative data in combination with more traditional survey data will be described.

Keywords

Child poverty; Measurement; Surveys; Administrative data

Using Bayesian dual systems estimation to produce a population record file

Patrick Graham, Rebecca Green, Sarah Hall, Matt Hendtlass, Abby Morgan, Andrew Richens, Luciannce Varn

Statistics New Zealand

Producing a population unit record file using dual system estimation (DSE) contains some interesting challenges, because DSE is generally only used to produce population estimates. We use a Bayesian Gibbs sampler with census responses and administrative data to produce a unit record file of the New Zealand usually resident population.

A key assumption of DSE is independence of inclusion in both datasets conditioning on certain covariates. This presentation will cover modelling how these covariates effect inclusion in the data. Further

complications are that some covariates in administrative data are prone to mis-classification (e.g. location), as well as how to deal with missingness in the key variables.

Keywords

Bayesian; Dual systems estimation; Administrative data; Gibbs sampler

Conditional likelihood for two-phase designs: continuous responses and relative risk

Claudia Rivera-Rodriguez[1], **Gustavo Amorim**[2]

[1] Department of Statistics, University of Auckland, New Zealand

[2] Department of Biostatistics, Vanderbilt University, USA

Several methods can be used for the the analysis of two-phase samples. A popular approach is weighted likelihood, which can be improved by adjusting the sampling weights using information available at first phase. The best choice of adjusting variables are the influence functions, which are usually unknown and therefore need to be estimated. For adjusted weights, the method has been shown to be efficient, but it is prone to error and implementation can be complex. An alternative method is conditional maximum likelihood (CML) estimation, which is computationally more demanding, but does not require the imputation step. Another advantage of it is that non-experts can utilize it without interfering in the estimation process while still obtaining gains in efficiency. This paper presents a CML approach for the analysis of two-phase samples and for distributions that belong to the exponential family- continuous and discrete. It allows for modeling of the sampling probabilities through calibration(raking) or estimation. A large simulation study is conducted to evaluate the proposed method under different sampling designs for i) a continuous outcome and ii) a model for the relative risk. Lastly, the methods are applied to model the number of vaccines in a costing study of immunization in Brazil and a log-model for the relative risk from a subset of Breast Cancer Surveillance Consortium data (BCSC). Software and examples for the implementation is provided in github.

Keywords

Conditional maximum likelihood; Calibrated weights; Two-phase designs; Missing data

A simple model for size-at-detection of an invasive species under surveillance

Andrew P. Robinson[1], **John Kean**[2], **Melissa Welsh**[3]

[1] School of BioSciences, University of Melbourne, Australia

[2] AgResearch, New Zealand

[3] Scion Research, New Zealand

The impact of invasive species is affected by a range of factors, many of which can be anticipated in advance — for example, the prevalence of host material, climate suitability, the size of affected agriculture and so on. One factor that cannot be anticipated is the size of the incursion at the time of its detection. The impact of an incursion is tightly tied to its maturity at detection, ranging from a single seed, for example, to a 50,000 hectare infestation of plants.

We propose a simple probability model for the detection of an invasive species that can either capture or integrate out the consequent uncertainty of the maturity of the incursion. We capture the relationship between surveillance and the detection of the organism using *survival analysis*: the detection of the incursion

is analogous to the survival event; it is a binary occurrence that happens at some point in time, and once it has happened it does not happen again.

Under such a model, we can easily connect the distribution of the size of the infestation at the time to detection to the probability of detecting the incursion given that it has not already been detected, namely, the hazard function. For example, a popular model for the detection of an incursion of size x with number of traps t and probability of detecting a single pest p is

$$h(x, t, p) = 1 - (1 - p)^{tx}.$$

Algebra leads us to a size-at-detection pdf. Other corrections are also applied as needed. The outcome is a pdf that is a function of process parameters, enabling straightforward assessment of different surveillance choices. Parameter estimates for the distribution can be derived from first principles, field experiments, or expert elicitation.

In this presentation we will derive and demonstrate the use of the survival-based incursion size at detection pdf and its implications and challenges.

Keywords
Surveillance; Invasive species; Survival analysis

Quantifying uncertainty in marginal likelihood estimation via the stepping-stone sampling algorithm

John Siryj S

Department of Statistics, University of Auckland, New Zealand

Critical to any statistical model is some assessment of its strength. Under the Bayesian framework the marginal likelihood is often used. In principle independent marginal likelihood calculations could be drawn to ascertain its uncertainty, however, this proves impractical in most settings. Therefore in order to have reasonable confidence in the probable range for the marginal likelihood, resampling techniques are encouraged instead. In the context of dependent samples the usual independent bootstrap underestimates the uncertainty in our model evidence calculations. In order to overcome this a block bootstrap approach is employed. The block bootstrap approach requires specification of a block length parameter, which is expected to depend on the magnitude of the underlying dependence structure. We explore an empirical method aimed at finding the optimal block length for a given dependence structure under a variety of conditions. We first define what properties an optimal block length is expected to have before using our derived empirical approach on various models. In the case of the AR(1) where this problem has an analytical solution we show that our approach correctly approximates this. Finally, an application to simulated gravitational wave data is considered using one of our models.

Keywords
Marginal likelihood; Block bootstrap; Optimal block length; Stepping-stone sampling; Gravitational waves

Application of discriminant and principal component analyses to analysis of metabolomics data obtained from brain tissue

Paul F. Smith, Yiwen Zheng

Department of Pharmacology and Toxicology, University of Otago, New Zealand

The study of molecular biomarkers which can predict disease or disorders in humans is an important area of medical science. However, data analysis in this area requires multivariate analyses due to the number

of independent and dependent variables. In this study, we analysed over 100 metabolites in 12 different brain regions collected from animals subjected to a sham procedure or to acoustic trauma, and profiled them using a gas chromatography mass spectrometry (GC/MS)-based metabolomics platform. Acoustic trauma is the most common cause of hearing loss and tinnitus in humans. After deconvolution of mass spectra and identification of the molecules, the metabolomic data were processed using multivariate statistical analysis. Principal component analysis showed that metabolic patterns varied among different brain regions; however, brain regions with similar functions had a similar metabolite composition. Acoustic trauma did not change the metabolite clusters in these regions. When analyzed within each brain region using partial least squares discriminant analysis (PLS-DA), 17 molecules showed distinct separation between control and acoustic trauma groups in the auditory cortex, inferior colliculus, superior colliculus, vestibular nucleus complex, and cerebellum. Further metabolic pathway impact analysis and enrichment overview with network analysis suggested the primary involvement of amino acid metabolism, including the alanine, aspartate and glutamate metabolic pathways, the arginine and proline metabolic pathways and the purine metabolic pathway. Our results provide the first metabolomics evidence that acoustic trauma can induce changes in multiple metabolic pathways.

Keywords

Principal component analysis; Partial least squares discriminant analysis; Metabolomics

Finding your feet: a Gaussian process model for estimating the abilities of Test cricket players

Oliver G. Stevenson, Brendon J. Brewer S

Department of Statistics, University of Auckland, New Zealand

In the sport of cricket, player ability is traditionally measured using batting and bowling averages. However, these averages fail to measure both short-term changes in ability that occur during a match, and long-term changes that occur between matches, due to the likes of recent form, age and experience. We derive and fit a Bayesian parametric model that employs a Gaussian process to measure and predict how the abilities of cricket players vary and fluctuate over the course of entire playing careers. The general findings suggest players do not begin their careers playing to the best of their ability. Rather, it takes players a number of matches to 'find their feet' at the international level. The results allow us to better quantify and predict player ability, compared with both traditional cricket statistics – such as batting and bowling averages – and more complex models, such as the official International Cricket Council ratings. In a practical sense, the model may have significant implications in the likes of player comparison and team selection policy, as coaches and selectors will have a more direct means of understanding the risks and real life impacts of selecting one player over another.

Keywords

Cricket; Gaussian process; Bayesian analysis; Forecasting; Ranking

Factors influencing termination of defined benefit pension plans in the US

Tanjila Tabassum, Eric Ulm, Robert Kirkby S

School of Economics and Finance, Victoria University of Wellington, New Zealand

Defined benefit (DB) pension plans in US participate in the Pension Benefit Guaranty Corporation (PBGC) program. Funded by the premiums and also from recoveries of the terminated pension plans-

PBGC insures the benefit payments of retirees when it's due. The negative net position of PBGC since the last two decades have raised question on the threat of burdening the tax payers if huge underfunded plans get terminated. Although voluntary termination of overfunded plans does not affect the PBGC directly, but it poses the risk of miscalculation in the projection and future plans of PBGC. In this paper we investigate the influences behind both voluntary and involuntary termination of overfunded and underfunded pension plans. We also study the transition of pension plans between the states of being insured and uninsured in consecutive years along with the likelihood of termination of uninsured plans. Unlike previous studies we are considering association of both pension fund and sponsor assets with termination decision. We found significant but very small effect of pension fund variables on the termination decision. There is positive marginal effect of economic factors on termination. We have also found that the plans insured by both insurance companies and PBGC are less likely to terminate. This means that paying more premiums does not necessarily result in termination of the pension plans.

Keywords
DB pension; PBGC; Overfunded; Underfunded; Termination; Marginal effects; Insurance

Fuzzy clustering of categorical data using finite mixtures

Kien Tran

School of Mathematics and Statistics, Victoria University of Wellington, New Zealand

The direct analysis of very large data sets using techniques such as regression can be computationally infeasible. Instead, the researcher may want to reduce the dimension of the data sets first, for example, by finding observations that responds similarly to a set of questions (row clustering), attributes that sees similar responses (column clustering) or groups of observations that answers similarly to a group of questions (two-mode clustering).

The clustering of non-independent categorical variables in particular poses distinct difficulties due to its lack of a distance metric and a measure of correlation (in the nominal data case). In this work, a clustering approach for categorical data based on finite mixtures is proposed, where the correlation structure of columns is handled by either pairwise modelling or joint modelling. The use of composite likelihood methods are also used to limit the number of parameters used. If applicable, this would provide a parsimonious, likelihood-based fuzzy clustering approach suitable for statistical inference; as well as the potential for extension to more general mixed type data.

Keywords
Dimension reduction; Model-based clustering; Nominal variables; Finite mixture models; Composite likelihood

Teaching biostatistics to clinical researchers

Robin Turner[1], **Nicole Mealing**[2]

[1] Centre for Biostatistics, University of Otago, New Zealand

[2] NSW Health, Australia

Teaching biostatistics to clinical researchers is a challenging endeavour. The aim with this type of teaching is to introduce and embed statistical thinking that is relevant and useful for their research and clinical work. There are a range of students attending these courses including those with a fear of statistics, those who have not used mathematics for many years and some who have completed first year statistics. This latter group may have some statistical understanding but often struggle to relate this to their health sciences research. We have developed a short course that overcomes people's fear of biostatistics using fun

and interactive examples. This presentation will describe the pedagogic approach used including how we made use of the innovative teaching developed by others in New Zealand at the secondary level and applied this to the tertiary environment.

Keywords
Biostatistics; Education; Pedagogy

Model selection via cross validated log likelihood

Rolf Turner[1], **Kate Richards**[2]

[1] Department of Statistics, University of Auckland, New Zealand

[2] Plant and Food Research, New Zealand

The problem being addressed arises in the context of fitting dose-response models to data on insect survival. These models take the form of generalised linear mixed models. The practical objective is to determine an upper bound on the "lethal dose" (at a specified probability level) which is as small as possible. The upper bound is taken to be the upper end point of a 95% confidence interval for the lethal dose. The confidence interval is (usually) calculated using Fieller's formula. The "dose" may be a literal dose level, as in the concentration of fumigant used, or time or temperature, amongst other possibilities.

A substantial number of different model choices are possible. Choices to be made may include the link function, whether to log transform the numeric predictor, and whether to include 'controls" (observations for which the value of the "dose" is zero) in the data. There is no general consensus as to which is the "correct" set of choices, and no subject-area based criterion for making these choices. Different choices give substantially different upper bounds on the lethal dose.

A promising technique for making the necessary choices is cross validated log likelihood. In this talk I will explain this technique and present some results of applying it. I will also point out some caveats and considerations that require further investigation.

Keywords
Dose-response models; Lethal dose; Generalised linear mixed models; Fieller's formula

Stochastic simulations of gene regulatory networks with `sismonr`

Olivia Angelin-Bonnet[1], **Matthieu Vignes**[1], **Patrick J. Biggs**[1,2], **Susan Thomson**[3], **Samantha Baldwin**[3]

[1] School of Fundamental Sciences, Massey University, New Zealand

[2] School of Veterinary Science, Massey University, New Zealand

[3] Plant & Food Research, Christchurch, New Zealand

In this talk, I present `sismonr` (available from the CRAN at `https://CRAN.R-project.org/package=sismonr`, with a detailed tutorial found at `https://oliviaab.github.io/sismonr/`), a tool we developed to generate and simulate *in silico* biological systems. I decompose the methodological description of our approach in three steps:

- the gene network generation, i.e. the structure of the network, the properties of the regulatory relationships and kinetic parameter values,
- the creation of a population of individuals, with their genetic characteristics, and

- the Stochastic Simulation Algorithm [2], which generates the abundance of the different biological entities/units in the system over the simulation time.

A first application of the package is to generate benchmark datasets for gene network reconstruction methods [1], as `sismonr` provides a biologically accurate model of gene expression regulation. To the best of our knowledge, it is the first to encompass the presence of non-coding genes, post-transcriptional regulation, genetic variation, and to explicitly represent the ploidy of the organism.

A second application of the package is that of model checking. Given a model for a real biological system under study, we can generate the distribution of biological unit abundance under a chosen structure/parametrisation. It allows to validate (or not) the model by comparing simulated data to real data obtained from -omics measurements. In particular, it can give a range of plausible values for regulation parameters.

Lastly, `sismonr` can assist biologists when intervening (e.g. gene Knock-Out or over-expression) on the model to anticipate the accurate consequences of their experiment design.

References

[1] Olivia Angelin-Bonnet, Patrick J Biggs, and Matthieu Vignes. "The sismonr Package: Simulation of In Silico Multi-Omic Networks in R". In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2018, pp. 2729-2731.
[2] Darren J. Wilkinson. *Stochastic Modelling for Systems Biology*. 2nd edition. CRC Press, 2012.

Keywords

Systems Biology; Gene regulatory networks; Stochastic Simulation Algorithm

Integrative association analysis for multi-omics data

Yalu Wen, Xiaqiong Wang

Department of Statistics, University of Auckland, New Zealand

With high-throughput technologies, multi-layer omics data (e.g., genomic, transcriptomic, epigenomic and proteomic) have emerged and provide unprecedented opportunities for investigating their interplay underlying complex disease at various molecular levels. However, associations between disease and omics data are usually tested on a single layer. Therefore, the integrative association of multi-omics data with outcomes should be used to provide a comprehensive view of biological system. In this paper, we use four approaches including the Brown method, the Omnibus-Fisher method, the truncated product method, and the harmonic mean p-value method to combine individual P-values which are calculated by kernel machine regression (KMR) model into an overall significance, and then compare these four combining P-values methods with a projection-based KMR method (denoted as KPCA) to test their performance of association analysis. In the simulation, the Omnibus-Fisher method and the harmonic mean p-value method perform better than the projection-based KMR method with the consideration of power results and type I error results. Finally, we illustrate these five existing methods by analysing multi-omics data from a study of CLL patients.

Keywords

Multi-omics data; Integration; Association analysis

Identifying firing pin impressions by using machine learning

Zhijian (Jason) Wen[1], **James Curran**[1], **SallyAnn Harbison**[2], **Douglas Elliot**[2], **Gerhard Wevers**[2]

[1] Department of Statistics, University of Auckland, New Zealand

[2] Institute of Environmental Science and Research, New Zealand

In serious crimes committed, such as robberies and murders, are often involving firearms. In order to solve the crime cases, firearm examiners are required to identify whether the cartridge cases found from crime scenes are fired from a suspected firearm. This identification is conducted based on the unique marks left on the surfaces of cartridge cases and firing pin impression is one of the most vital marks. In this study, a total of 10 auto-loading rifles are used. Fifty bullets are fired from each rifle. The cartridge cases are collected and photographed under the optical microscope. Two machine learning methods, the convolutional neural network (CNN) and the histogram of orientated gradient – support vector machine (HOG-SVM) are employed to identify the cartridge cases based on the firing pin impressions.

Keywords

Firing pin impression; Machine learning; Convolutional neural network; Histogram of orientated gradient; Support vector machine

List of Abstracts – Posters

Traffic modelling and inference using bluetooth tracking data

Ghazaleh Aslani, Geoff Jones, Katharina Parry

School of Fundamental Science, Massey University, Palmerston North, New Zealand

Data collected by Bluetooth detectors is becoming a more widely spread method for collecting travel information for use in transportation studies.

Each Bluetooth device uses a unique identifier code known as its Media Access Control (MAC) address. When an active Bluetooth-equipped device travels past a Bluetooth sensor, its MAC address and time of detection are recorded. This allows us to monitor movements between sensors. This kind of data often includes duplicate or multiple detections of MAC addresses, which is generally considered a source of noise.

We take another point of view! In our work, we consider multiple detections as a source of information in their own right. We are looking at the distributions of gap times between consecutive detections, and how we can use these distributions for classification. We will present some initial results on our clustering of MAC addresses, Bluetooth sensors and time interval of days.

Keywords
MAC addresses; Multiple detections; Cluster analysis

Unweighted regression is preferred for RDS data

Lisa Avery[1,2], Michael Rotondi[2]

[1] Department of Mathematics and Statistics, University of Otago, New Zealand

[2] Faculty of Health, York University, Canada

Background: It is unclear whether weighted or unweighted regression is preferred in the analysis of data derived from respondent driven sampling. Our objective was to evaluate the validity of various regression models, with and without weights and with various controls for clustering in the estimation of the risk of group membership from data collected using respondent-driven sampling (RDS). Methods: Twelve networked populations, with varying levels of homophily and prevalence, based on a known distribution of a continuous predictor were simulated using 1000 RDS samples from each population. Weighted and unweighted binomial and Poisson general linear models, with and without various clustering controls and standard error adjustments were modelled for each sample and evaluated with respect to validity, bias and coverage rate. Population prevalence was also estimated. Results: In the regression analysis, the unweighted log-link (Poisson) models maintained the nominal type-I error rate across all populations. Bias was substantial and type-I error rates unacceptably high for weighted binomial regression. Coverage rates for the estimation of prevalence were highest using RDS-weighted logistic regression, except at low prevalence (10%) where unweighted models are recommended. Conclusions: Caution is warranted when undertaking regression analysis of RDS data. Even when reported degree is accurate, low reported degree can unduly influence regression estimates. Unweighted Poisson regression is therefore recommended.

Keywords
Respondent driven sampling; Weighted regression; Poisson regression

Modelling social dilemmas and treatment scheduling in tumours using evolutionary game theory

Sarah Croft, Michael Black S

Department of Mathematics and Statistics, University of Otago, New Zealand

Cancer is the second highest cause of mortality in New Zealand. In treating a patient's tumour, the strategy is generally to administer a "maximum tolerated dose" therapy, in an attempt to completely eradicate the tumour. While this is successful in many cases, for heterogeneous tumours, an initial shrinkage in tumour volume may be followed by renewed growth driven by a population of treatment-resistant cells. An alternative approach to treatment is use an "evolutionary" strategy, adjusting the treatment regime to the tumour profile to exploit the treatment sensitive cells and maintain the disease in a non-lethal or chronic state. This involves considering the cancer cell populations in the tumour, and the presence and maintenance of intra-tumour heterogeneity. Intra-tumour heterogeneity of aerobic and anaerobic cells is known to persist within a tumour, despite the increased energy efficiency of aerobic metabolism in non-oxygen limited environments. Here we use the replicator equations of Evolutionary Game Theory to define the maintenance of the intra-tumour heterogeneity by frequency-dependent selection, and a double goods game is defined to couple the fitness of the aerobic and anaerobic cells in the tumour environment. The double-goods game is thus used to describe the environmental conditions supporting a mixed tumour population. The proportions of each cell type in the tumour and the tumour environment influence the response to treatment due to the assumptions of frequency dependent selection. We show that tumour composition should be considered when deciding treatment scheduling to efficiently produce recovery. Additionally, if error-prone measurements of the cell proportions are used to decide treatment scheduling in a clinical setting, adjustment for the uncertainty must be made to reliably reach recovery.

Keywords

Public Goods Game; Intra-tumour Heterogeneity; Treatment Scheduling

Role of time variant factors in determining subpopulation heterogeneity in life course outcome: A case study on number of sexual partners

Bhubaneswor Dhakal, John Horwood

Christchurch Health and Development Study, University of Otago, New Zealand

Studies on the role of time variant factors in determining subpopulation heterogeneities in life course outcomes are rare. This study utilized longitudinal data on the sexual partnership histories of a birth cohort of over 1000 individuals to demonstrate the use of time variant factors in determining unobserved heterogeneity in a population. The data were gathered as part of the Christchurch Health and Development Study (CHDS). A latent class trajectory analysis framework was used to model heterogeneity in sexual partnership histories from age 16-35 years. The analysis utilized the 3-step Bolck, Croon and Hagenaars (BCH) method to characterize latent classes, incorporating both fixed and time dynamic factors as auxiliary variables. This approach accounts for misclassification errors associated with distal outcome and covariate effects in the assignment of individuals to latent classes. The latent class analysis identified a five-class model as the preferred solution, with each class representing a sub-group with a different life course trajectory of sexual partners. These trajectories were characterized as: very low partnership frequency (16.6%); typical or majority class (61.0%); young adult peak frequency (7.6%); adult peak frequency

(13.2%); and consistently high frequency (1.6%). The following time variant factors were identified as discriminating between the groups: duration of cohabitation; number of dependent children; history of alcohol or illicit drug use disorders; and history of depression. The findings illustrate the ways in which normal adult role transitions (partnership formation, having children) and lifestyle/life course factors (substance abuse, depression) impact in different ways to influence sexual partnership trajectories.

Keywords
Latent-class; Unobserved-heterogeneity; Time-variant-factor; Sexual-partner

Evaluating time delays in parallel backpropagation

Seong-Yun Hong

Department of Geography, Kyung Hee University, South Korea

Over the past years, artificial neural networks have been successfully applied to various geographic problems, such as spatial interpolation and regression. Several recent studies have shown that even a simple feedforward neural network can provide more accurate results than traditional methods when there are enough resources for training it. In general, a neural network is trained through an iterative process, in which each iteration consists of forward and backward propagation. To make this iterative learning process successful, it requires a considerable amount of training data and numerous hidden layers between the input and output layers. The use of parallel computing becomes increasingly important as well because it can significantly reduce the calculation time for each iteration. However, the backward propagation, or backpropagation, in a parallel computing environment may not be as efficient as expected due to the so-called locking problem. If the calculation at one node in a hidden layer is delayed, all the nodes in the subsequent layer become idle to wait for the missing parameter. Theoretically, this problem gets worse when there are more hidden layers, but its practical significance has not been evaluated in a practical setting. The main purpose of this paper is, therefore, to address this gap in the literature. In this work, I built a feedforward neural network for spatial interpolation of real estate transactions data in Seoul, South Korea and estimated time delays during the backpropagation process. The results from this experiment can reveal the extent of the locking problem and its relationship with the number of hidden layers in a neural network.

Keywords
Time delays; Parallel backpropagation; Feedforward neural network; Spatial interpolation

Two-phase subsampling design for DNA sequencing with application in the relatedness of endangered species

Zoe Luo S

Department of Statistics, University of Auckland, New Zealand

Whole-genome sequencing has been completed for the entire kākāpō species. However, this sort of effort is not feasible in most situations and only some individuals can be sequenced. Despite the decreasing cost of DNA sequencing, budget remains a substantial problem for most research funders. A cost-saving strategy is to resequence a small subsample from the original sample that participated in the genome-wide association study with higher resolution, then use the data from the subsample to infer the rest of the sample. This strategy is called two-phase sampling, because the initial sampling of the cohorts is followed by a subsampling of individuals to be resequenced. The aim of this project is to design a sampling method and

investigate the optimal way to choose a subsample based on known relatedness, phenotypes, or genotypes at a small number of sites. The kākāpō genomes will provide a basis for simulating subsampling strategies in an endangered species.

Keywords
Genotype imputation; Genome-wide association study; Conservation genetics

Detecting slow slip events in GPS time series data by using a change-point detection method

Yiming Ma, Fabien Montiel, Ting Wang S

Department of Mathematics and Statistics, University of Otago, New Zealand

Slow slip events (SSEs) is a type of slow earthquakes, which plays an important role in releasing strain in subduction zones, with duration ranging from days to years, much longer than ordinary earthquakes of comparable seismic moment. SSEs recur frequently and some have been found to precede large natural earthquakes. Recently several studies focused on using SSEs to infer the probability of a damaging earthquake in the same subduction zone. Geodetic measurements have been widely used to detect SSEs along several subduction zones worldwide, especially Global Positioning System (GPS) because of its high precision and cheap cost. However, lots of SSEs remain undetected due to noises in GPS time series, and the start time of SSEs detected so far has large errors. Therefore, detecting SSEs in GPS time series accurately is the first step to forecast damaging earthquakes. Linear regression, sometimes combined with other methods, is the most common method for detecting SSEs, but it can work well only in simple cases (such as high signal-to-noise ratio). A more effective detection method is needed for noisy data. To develop a reliable SSE detection technique (change-point detection method), we constructed a physical SSE simulator able to generate artificial GPS time series. This approach allows us to know in advance the underlying SSE-generated signal and therefore validate the change-point detection method for a range of conditions. Preliminary results will be presented in this poster.

Keywords
Slow slip events; Change-point detection; GPS time series

Reliable Gaussian process approximations for dynamic traffic models

Ahmad Mahmoodjanlou[1], Martin Hazelton[1,2], Katharina Parry[1] S

[1] School of Fundamental Science, Massey University, Palmerston North, New Zealand

[2] Department of Mathematics and Statistics, University of Otago, New Zealand

Stochastic models are attractive for representing traffic dynamics over a network because of the unpredictable fluctuations in traffic flow from day-to-day. Nonetheless, it is easier to work with deterministic models, which seek to describe the behaviour of the system "on average". In particular, there is now a large body of theory on these deterministic models.

A potentially attractive hybrid approach is to approximate discrete-flow stochastic models by a Gaussian random process, in which the dynamics of the mean follow one of these well understood deterministic models. This is known to work well for stable traffic networks, but we have recently shown that such

approximations can break down badly for highly unstable systems. The question remains, how can one tell if a system is sufficiently stable for the Gaussian approximation to be serviceable?

The presented work shows that this question can be examined through a measure called the coefficient of reactivity. This is illustrated through a range of numerical experiments, including a variety of different road networks.

Keywords
Stochastic models; Coefficient of reactivity; Stability

Hierarchical modelling for spatial muscle fibre configurations

Anna K. Redmond, Tilman M. Davies, Matthew R. Schofield

Department of Mathematics and Statistics, University of Otago, New Zealand

Muscle fibre type is an important factor in determining muscle function and performance. The spatial distribution of fast and slow twitch fibres in fascicles and the muscle as a whole can vary in complex ways which affect the properties of the muscle. Being able to quantify these distributions is vital to allow comparisons to be made between different muscles or one muscle over time, which can be used to understand progressive muscle weakening due to age or disease. We model the fibre type data with a hidden conditional autoregressive model to each fascicle including a covariate measuring distance from muscle centre. We also allow the parameters for each fascicle to vary spatially across the muscle. The use of the Stan software to fit this model in a Bayesian framework allows changes to be made to the model more easily than using the standard Metropolis-within-Gibbs algorithm.

Keywords
Hierarchical modelling; Binary data; Stan

Recovery-test sampling size for stunned poultry in industrial halal meat production

Maryann Staincliffe, Mustafa M. Farouk

AgResearch, New Zealand

Poultry/birds meat is only halal if the birds are alive at the time of slaughter. 100% assurance that birds are alive at slaughter can only be provided when birds are hand slaughtered without stunning and the speed the birds are presented was optimal to allow tasmiyya (invocation) to be uttered on each bird by the slaughterer. In the industrial production of halal poultry meat, birds are mostly electrically stunned in waterbath stunners (ES) and either hand or machine slaughtered. A recovery-test (birds are stunned, not slaughtered and allowed to recover) of the birds should be performed before any stunner is accepted as meeting the requirements for halal. The recovery-tests (RT) are currently done too subjectively with no sampling plans to provide confidence that stunning processes are halal compliant. The aim of this study is to determine statistically the sampling size for RT needed to objectively validate a new electrical stunner for halal compliance and for harmonization of halal standards.

Keywords
Recovery-test; Binomial probability; Stunning; Poultry; Halal; Rule of three

List of Participants

Sawsan Al-Shamaa	Auckland Institute of Studies
Olivia Angelin-Bonnet	Massey University
Luis Apiolaza	University of Canterbury
Richard Arnold	Victoria University of Wellington
Azam Asanjarani	University of Auckland
Ghazaleh Aslani	Massey University
Lisa Avery	University of Otago
Adrian Baddeley	Curtin University
Gerard Barbalich	Harmonic Analytics
Amir Bashir	Massey University
Timothy Bilton	University of Otago
Michael Black	University of Otago
Robert Borotkanics	Auckland University of Technology
Jennifer Brown	University of Canterbury
Christine Bycroft	Statistics New Zealand
Alana Cavadino	University of Auckland
Vanessa Cave	AgResearch
David Chan	University of Auckland
Lisa Chen	Harmonic Analytics
Tong Chen	University of Auckland
Kevin Cheng	University of Auckland
Len Cook	Victoria University of Wellington
Roy Costilla	University of Queensland
Sarah Croft	University of Otago
James Curran	University of Auckland
Tadhg Daly	Ministry of Justice
Tilman Davies	University of Otago
Brandon de Graaf	University of Otago
Yongshi Deng	University of Auckland
Bhubaneswor Dhakal	University of Otago
Ken Dodds	AgResearch
Fabian Dunker	University of Canterbury
Jamas Enright	Statistics New Zealand
David Fletcher	University of Otago
Yuki Fujita	Harmonic Analytics
Nick Garrett	Auckland University of Technology
Daniel Gerhard	University of Canterbury
Zachary Goeden	Victoria University of Wellington

Peter Green	AgResearch
Hongbin Guo	Harmonic Analytics
Yang Hai	University of Auckland
Sarah Hall	Statistics New Zealand
John Harraway	University of Otago
Jill Haszard	University of Otago
John Haywood	Victoria University of Wellington
Martin Hazelton	University of Otago
Harold Henderson	AgResearch
Barbara Holland	University of Tasmania
John Holmes	University of Melbourne
Seong-Yun Hong	Kyung Hee University
Amanda Hughes	WorkSafe New Zealand
Takayoshi Ikeda	Auckland University of Technology
Rodelyn Jaksons	Plant and Food Research
Stéphane Janicot	Auckland University of Technology
Beatrix Jones	University of Auckland
Tim Jowett	University of Otago
Jie Kang	University of Otago
Yujin Kim	University of Auckland
Maarten Kruijver	Institute of Environmental Science and Research
Alastair Lamont	University of Otago
Kate Lee	University of Auckland
Fangyao Li	University of Auckland
Lingyu Li	Victoria University of Wellington
Ivy Liu	Victoria University of Wellington
Qing Lu	University of Florida
Thomas Lumley	University of Auckland
Dongwen Luo	AgResearch
Zoe Luo	University of Auckland
Harel Lustiger	Harmonic Analytics
Yiming Ma	University of Otago
Ahmad Mahmoodjanlou	Massey University
Roger Marshall	University of Auckland
Patricio Maturana-Russel	University of Auckland
Louise McMillan	Victoria University of Wellington
Emil Mendoza	University of Canterbury
Poppy Miller	AgResearch
Victor Miranda	Auckland University of Technology
Khandoker Akib Mohammad	Victoria University of Wellington

Elena Moltchanova	University of Canterbury
Lindsay Morris	Victoria University of Wellington
Omer Ozturk	Ohio State University
Matt Parry	University of Otago
Richard Penny	Statistics New Zealand
Shirley Pledger	Victoria University of Wellington
Yijia Qian	University of Canterbury
Diane Ramsay	Statistics New Zealand
Anna Redmond	University of Otago
Kate Richards	Plant and Food Research
Andrew Richens	Statistics New Zealand
Claudia Rivera-Rodriguez	University of Auckland
Andrew Robinson	University of Melbourne
Matthew Schofield	University of Otago
Katrina Sharples	University of Otago
Dalice Sim	University of Otago (Retired)
John Siryj	University of Auckland
Paul Smith	University of Otago
Maryann Staincliffe	AgResearch
Oliver Stevenson	University of Auckland
Tanjila Tabassum	Victoria University of Wellington
Glenn Thomas	Harmonic Analytics
Dale Townsend	Hamilton City Council
Kien Tran	Victoria University of Wellington
Chris Triggs	University of Auckland
Robin Turner	University of Otago
Rolf Turner	University of Auckland
Martin Upsdell	AgResearch
Alain Vandal	University of Auckland
Matthieu Vignes	Massey University
Ting Wang	University of Otago
Xiaqiong Wang	University of Auckland
Zhijian Wen	University of Auckland
Ian Westbrooke	Department of Conservation
Shanika Wickramasuriya	University of Auckland
Jennifer Wilcock	University of Canterbury
Graham Wood	University of Otago
Xun Xiao	Massey University
Ploi Yibmantasiri	Harmonic Analytics
Rachel Youngson	Victoria University of Wellington