Honours in Mathematical Statistics

Detailed Guide for the 2018 academic year

THE UNIVERSITY OF
SYDNEY

School of Mathematics and Statistics

Contents

1 Entry requirements

Preliminary entrance into the honours program is through the Faculty of Science. The Faculty requirements which must be met include:

- qualifying for the pass degree with a relevant major;
- having a SCIWAM of at least 65.

In addition, the School of Mathematics and Statistics has some extra criteria:

- 24 credit points in relevant Senior units of study, (some of which are compulsory and/or should be completed at the Advanced level: see the appropriate detailed guides for listings of these);
- of these relevant units, those at the
  - Advanced level should have an average mark of at least 65;
  - Normal level should have an average mark of at least 75;
- prospective student should actively seek a supervisor.

For Mathematical Statistics we require in addition that

- the "relevant major" is in either Mathematical Statistics or Financial Mathematics and Statistics;
- completed units of study include STAT3911 and STAT3912.

All acceptances into Honours in Mathematical Statistics are ultimately at the discretion of the School, however a student meeting all of the above criteria (or the equivalent from another institution) should be confident of acceptance.
Please note the Faculty of Science Honours **application deadline** (for Honours commencement in Semester 1, 2018) is Thursday 30 November 2017.

2 Structure of Honours

An Honours year in Mathematics and Statistics involves six courses (worth 60% of the final mark) and a project (worth 40%). Formally, each student is administered by one of the three main areas of Applied Mathematics, Pure Mathematics and Mathematical Statistics; this is determined by the project topic and supervisor.

2.1 The Honours project

The Honours project centres around an essay/thesis consisting of 50-60 pages written on a particular topic from your chosen area. It need not contain original research (although it might) but it should clearly demonstrate that you have understood and mastered the material. The assessment of the honours thesis is based on the mathematical/statistical content and its exposition, including the written english. The thesis is due at the end of your second semester, specifically on Monday of week 13.

Toward the end of the second semester (Friday week 10), each student gives a 25 minutes talk on their thesis project. The aim of the talk is to explain to a broader audience the purpose and nature of the project. The talk is followed by 5 minutes dedicated to questions from the audience which includes staff members and fellow students.

2.2 Course work

Full-time students normally attend three lecture courses each semester, for a total of six courses. All six courses will count towards the student's final assessment. If a student takes more than six courses in total then the top six results will count towards the student's final assessment.

Students are expected to select a mixture of applied and theoretical course courses and their selection has to be *explicitly approved* by their supervisor as well as by the Honours coordinator at the start of each semester. Please note that the course on Probability Theory is *mandatory* for all stats honours students.

In practice our statistics honors courses are often made of two "half-courses" suggesting a recommended pairing of the two halves although in principle students can substitute one "full" course with two half-courses which are not paired.

A tentative list of the stats honour-level course offering is available in Section 5. Subject to the approval of your supervisor and the Honours coordinator the following courses can also be taken for credit toward your required coursework:

- Honours Applied Mathematics and Pure Mathematics courses available at our School. Note in particular the courses in financial mathematics offered by the Applied Mathematics group. Please contact the respective coordinators for more details.
- Third year advanced courses offered at our School (obviously only those not taken before)
- Courses available through the Advanced Collaborative Environment (ACE)
- Up to one course offered at the AMSI Summer School (January 2018)

2.3 Writing proficiency

As mentioned above your essay is also assessed based on the quality of the writing. This does not mean we look for the next Shakespeare however you should make sure you express your ideas in an organized manner using a clear and grammatically correct English. The university offers several resources that can help you achieve this goal. The Learning Centre offers workshops for students that need help with extended written work, and a trove of online resources for improving your writing skills is also available. Make sure you make use of these resources as early as possible as writing skills develop slowly over time and with much practice.

3 Program Administration

The Statistics Honours Program Coordinator is

A/Prof. Uri Keich,
Carslaw Building, Room 821, Phone 9351 2307,
Email: `uri.keich@sydney.edu.au`

The director of the Statistics teaching program is

Dr. Michael Stewart,
Carslaw Building, Room 818, Phone 9351 5765,
Email: `michael.stewart@sydney.edu.au`

The Program Coordinator is the person that students should consult on all matters regarding the honours program. In particular, students wishing to substitute a course from another Department, School or University must get prior written approval from the Program Coordinator. Matters of ill-health or misadventure should also be referred to the Program Coordinator

Students **must select their courses after consulting the Honours supervisor and the Honours Coordinator**.

4 Academic Staff and their Research Interests

Doctor Lamiae Azizi
Bayesian nonparametrics, Graphical modelling, Variational methods, probabilistic learning, Analysis of biomedical data, image processing and engineering problems.

Associate Professor Jennifer Chan
Generalised Linear Mixed Models, Bayesian Robustness, Heavy Tail Distributions, Scale Mixture Distributions, Geometric Process for Time Series Data, Stochastic Volatility models, Applications for Insurance Data.

Professor Sally Cripps
Construction of Markov Chains for Efficient Marginalisation, Bayesian Variable Selection, Finite and Infinite Mixture Models, Nonparametric Bayesian Regression, Spectral Analysis of Time Series, Flexible Methods for Longitudinal and Panel Data, Computational Statistics.

Doctor Ray Kawai
Numerical Methods in Probability, Statistical Inference for Stochastic Processes, Stochastic Analysis, Mathematical Finance, Partial Differential Equations.

Associate Professor Uri Keich
Statistical Methods for Bioinformatics, Statistical Analysis of Proteomics Data, Computational Statistics, Analysis of False Discoveries in Multiple Hypotheses Testing.

Associate Professor Samuel Muller
Model Selection, Robust Methods, Applied Statistics, Extreme Value Theory.

Doctor John Ormerod
Variational Approximations, Generalised Linear Mixed Models, Splines, Data Mining, Semiparametric Regresssion and Missing Data.

Associate Professor Shelton Peiris
Time Series Analysis, Estimating Functions and Applications, Statistics in Finance, Financial Econometrics, Time Dependent Categorical Data.

Doctor Michael Stewart
Mixture Model Selection, Extremes of Stochastic Processes, Empirical Process Approximations, Semiparametric Theory and Applications.

Doctor Emi Tanaka
Applied Statistics in Agriculture and Bioinformatics, Linear Mixed Models, Experimental Design, Computational Statistics.

Doctor Garth Tarr
Applied statistics, Robust Methods, Model Selection, Data Visualisation, Biometrics.

Associate Professor Qiying Wang
Nonstationary Time Series Econometrics, Nonparametric Statistics, Econometric Theory, Local Time Theory, Martingale Limit Theory, Self-normalized Limit Theory.

Doctor Rachel Wang
Statistical Network Models, Bioinformatics, Markov Chain Monte Carlo Algorithms, Machine Learning, Distributed Inference.

Emeritus Professor Neville Weber
U-statistics, Exchangeability, Generalized Linear Models, Asymptotic Approximations.

Professor Jean Yang
Applied Statistics, Statistical Bioinformatics, Integrative Analysis of Microarray, Sequence and Protein Data, Statistical Computing.

Doctor Pengyi Yang
Signalling Network Reconstruction, Transcription Network Reconstruction, Statistical Learning in Omics, Omic Data Visualisation, Decipher Embryogenesis.

Recent publications of these members are available on the School's website. See the individual staff member for any reprints of their published papers.

5 Stats honours courses

The following stats honours topics are *expected* to be on offer in 2018. Please note that some, like Probability Theory will be offered as full courses while others will be offered as paired half courses.

1. **Advanced Bayesian Inference**

   First we will cover key concepts of Bayesian inference including: choices of priors, point estimates, credible intervals and model selection. We will also discuss philosophical differences, compare and contrast Bayesian and frequentist paradigms. Secondly, we will consider different methods for generating random variables from a desired distribution including: methods for generating uniform (pseudo-)random variables, transformation of random variables and rejection and adaptive rejection sampling. Thirdly, we will next consider methods for approximating integrals including Laplace's approximation and various types of importance sampling. The fourth group of topics concern Markov chain Monte Carlo (MCMC) including the justification of MCMC methods, different flavours of MCMC and the use the software package STAN in order to perform MCMC inference. Finally, we will introduce approximate Bayesian inference methods including and variational Bayes. Many different models will by considering including linear models, linear mixed models, generalized linear mixed models, models for missing data, models which automatically incorporate model selection and survival models to name a few.

2. **Extreme Value Theory**

   The course consists of 12 lectures, divided into two parts, Classical extreme value theory (EVT) and Special topics.

   Classical EVT

   - Limiting distribution of sample extremes in given cases.
   - Approximating tail probabilities.
   - Extreme-value distributions (EVDs).
   - The Fisher-Tippet theorem.
   - Maximum domains of attraction.
   - Generalised EVDs.

   Special topics

   These change each year. In previous years we have had **one** of

   - Rates of convergence in the Fisher-Tippet theorem.
   - Extremes of Gaussian processes.
   - Applications of EVT to mixture detection.
   - Applications of EVT to tests based on the lasso.

3. **Introduction to Stochastic Calculus**

   We will explore some basic theories of stochastic calculus, such as

- Conditional expectations, filtrations, martingales, stopping times,
- Brownian motion, stochastic integral, semimartingales,
- Ito's lemma, Lévy's characterization theorem, martingale representation property,
- Stochastic differential equations, Feynman-Kac formula, Girsanov's theorem, first passage times.

Note that this is an Assumed Knowledge course for "AMH4 Advanced Option Pricing".

Assumed Knowledge: It is *essential* that students have a very good command of the contents of STAT2911 (Probability Models), STAT3911 (Stochastic Processes) and MATH3969 (Measure Theory).

Beneficial Knowledge: It is *beneficial* that students have a very good command of the contents of MATH3961 (Metric Space), MATH3975 (Financial Mathematics) and MATH3963, MATH3974, MATH3978 (PDEs).

4. **Advanced Time Series Analysis and Forecasting Methods**

The course covers advanced methods of modelling and analysing of time series data with emphasis on theoretical development. The material includes review of linear time series models and properties, an introduction to spectral analysis of time series, generalized AR and MA Models and their properties, an introduction to fractional differencing and long memory time series modelling, generalized fractional processes, Gegenbaur processes, topics from financial time series/econometrics: ARCH, GARCH and other related volatility models, duration models in finance (ACD, Log-ACD and SCD), analysis of multiple time series, an introduction to state-space modelling and Kalman filtering in time series

Assumed knowledge: Mathematical Statistics (Advanced knowledge at Intermediate and Senior Levels) including a course on Time Series Analysis or equivalent.

References:

- Brockwell, P. J. and Davis, R. (1991). Time Series: Theory and Methods.
- Priestley, M. B. (1981). Spectral Analysis and Time Series.
- Tsay, R.S. (2005). Analysis of Financial Time Series.

5. **Fundamentals of Statistical Consulting**

This course is designed to assist students to develop effective consulting strategies and skills for dealing with real world data. Students will work on existing case studies and on data and analysis problems arising with real statistical consulting clients. There will be a mixture of lectures on consulting as well as discussion on general ways of handling challenges that occur in the consulting process. Learning outcomes of this course include

- Ability to formulate questions and appropriate hypotheses in a consulting context.
- Experience and exposure to a variety of data and questions.
- Identify and perform appropriate data analysis using a range of statistical procedures.
- Communicate via verbal and written consulting report how an appropriate analysis was performed and how it supports the research questions being tested.

6. **Generalized Linear Models**

   The topics include maximum likelihood inference, Newton-Raphson and Fisher Scoring methods, expectation maximization (EM) methods including Monte Carlo EM and expectation conditional maximization (ECM) methods, scale mixtures presentation, state space models, exponential family, generalized linear models, weighted least squares, quasi-likelihood, BLUP estimator, generalized estimating function, random effects models, Akaikes information criterion, Bayesian information criterion, logistic regression, two way contingency tables, Poisson regression, negative binomial and generalized Poisson distributions, over and under-dispersion, mixture models, log-linear models for categorical data, decomposable models, incomplete tables, quasi-independence, survival Analysis, Kaplan-Meier estimator, proportional hazards models and Coxs proportional hazards model.

7. **Asymptotics in Statistics and Econometrics**

   This course will introduce fundamental convergence concepts that are used in Statistics, Econometrics and other related fields. The topics include Slutsky's theorem, Delta method, continuous mapping theorem, central limit theorems for martingale, independent and dependent random variables, weak convergence (functional central limit theorem) and convergence to stochastic integrals. This course will consider the asymptotics in relation to Kernel, Quantile and least squares estimation, unit root testing and cointegration.

   References: 1. TS Ferguson: A course in large sample theory, Chapman and Hall, 1996 2. Anirban Das Gupta: Asymptotic Theory of Statistics and Probability, Springer, 2008. 3. A. W. Van der Vaart : Asymptotic Statistics, Cambridge University Press, 1998. 4. Qiying Wang: Limit theorems for nonlinear cointegrating regression, World Scientific, 2015

8. **Probability Theory**

   This is a rigorous course on probability with a measure theoretic basis.

   Contents: Measure spaces: properties and construction of measures; Borel-Cantelli Lemmas; Measurable functions and random variables; Independence; Tail sigma-algebra and Kolmogorov's 0-1 Law; Lebesgue Integral; Fatou's Lemma, DCT and MCT; Expectation of a random variable; Notions of convergence; Jensen's Inequality; Product space / sigma-algebra / measure; Fubini's Theorem; Conditional Expectation; Martingales: Betting Strategies, Doob's Upcrossings Lemma and Forward Convergence Thm, Stopping Times, Doob's Optional Stopping Thm, Uniformly Integrable Martingales, Backwards (reversed) Martingales and the SLLN; Exchangeable sigma-algebra and Hewitt-Savage 0-1 Law; Wald's Identities; Expected Exit Time of a Random Walk; Doob's Decomposition; Kolmogorov's 3 Series Thm; Doob's Submartingale's Inequality; Kolmogorov's Inequality; Doob's L^p Inequality; Martingale convergence in L^p; Helly's Selection Thm; Tightness; Characteristic Functions: basic properties, Levy's Inversion and Continuity Theorems.

   Assumed knowledge: STAT 2911 Probability and Statistical Models (Advanced) + real variable analysis. A knowledge of Measure Theory would be an advantage.

   References:

   - Rick Durrett. Probability: theory and examples, Fourth Edition.
   - David Williams. Probability with martingales.

9. **Statistical Methods in Bioinformatics**

Contents: Bioinformatics is a field that applies ideas from computer science, mathematical modeling, and statistics in order to make sense of the huge datasets that typify current research in biology.

This half course offers an in depth study of a few fundamental topics in bioinformatics while concentrating on the statistical point of view. Topics include the pairwise alignment problem, the construction of substitution matrices, the significance analysis of similarity searches, and hidden Markov models.

Assumed Knowledge: STAT 2911 Probability and Statistical Models

References:

- Warren Ewens, Gregory Grant. Statistical methods in bioinformatics: an introduction.
- Richard Durbin *et al.* Biological sequence analysis: probabilistic models of proteins and nucleic acids.

6 Project

6.1 General information on projects

Each student is expected to have made a choice of a project and supervisor well before the beginning of the first semester (or the beginning of the second semester for students starting in July).

Students are welcomed to consult on this matter with the Head of the statistics program and or the Honours Coordinator. At any rate, the latter should be informed as soon as a decision is made.

Work on the project should start as soon as possible but no later than the start of the semester. The break between the semesters is often an excellent time to concentrate on your research but you should make sure you make continuous progress on your research throughout the year. To ensure that, students should consult their appointed supervisor regularly, in both the researching and writing of the work.

A list of suggested project topics is provided in Section 6.2 below. Prospective students interested in any of these topics are encouraged to discuss them with the named supervisors as early as possible. Keep in mind that this list is not exhaustive. Students can work on a project of their own topic provided they secure in advance the supervision of a member of staff of the Statistics Research Group (including emeritus staff) and provided they receive the approval of the Program Coordinator.

Three copies of the essay typed and bound, as well an electronic copy must be submitted to the Honours Coordinator before the beginning of the study vacation at the end of your last semester. The exact date will be made known.

It is recommended that you go through the following checklist before submitting your thesis:

- Is there an adequate introduction?
- Have the chapters been linked so that there is overall continuity?
- Is the account self-contained?
- Are the results clearly formulated?
- Are the proofs correct? Are the proofs complete?
- Have you cited all the references?

6.2 Proposed project topics in Mathematical Statistics

For additional projects see the Section 12 at the end of this document.

1. **Automated Bayesian statistical Machine learning models evaluation**
   Supervisor: Dr. Lamiae Azizi

   Project description: Probabilistic modeling is a flexible approach to analyzing complex real data. Three steps define the approach. First we specify the model. Then, we infer the hidden structure. Last we evaluate the model. How do we evaluate the models?. A number of various techniques have been proposed for model checking, comparison and criticism in the recent years with one ultimate goal: the desire to generalize well. In Machine Learning, two complimentary tools are usually used to evaluate models: predictive accuracy and cross-validation. However, both measures do not tell us the whole story and the design and criticism of probabilistic models is still a careful, manual craft. The goal of this project is twofold: 1) exploiting the new advances in decision theory and information theory to propose new general ways of evaluating a Bayesian model and 2) making these tools automated to make it easier for practitioners to use them efficiently.

2. **Time series models using variance gamma distribution with an application to Bitcoin data**
   Supervisor: A/Prof. Jennifer Chan

   Project description: This project will investigate properties of high frequency data which often display high kurtosis. Popular heavy tail distributions like Student t and exponential power may still be inadequate to provide high enough level of kurtosis. Recent studies have considered variance gamma distribution in which the shape parameter can be made sufficiently small to provide unbounded density around the centre and heavy tails at the two ends of the distribution. As gamma variance distribution can be expressed as scale mixtures of normal, it facilitates model implementation in the Bayesian approach via some Bayesian software such as OpenBUGS and Rstan. We will consider long memory, stochastic volatility and leverage effect modelling to describe features of the data. For the application, we will adopt the recently emerged Bitcoin data which display extremely high kurtosis. Currently, not much studies have been directed to study properties of Bitcoin data and so this study will be pioneering, interesting and important.

3. **Volatility models for high frequency data**
   Supervisor: A/Prof. Jennifer Chan

   Project description: Volatility forecast is important in risk management. However since volatility is unobserved, most volatility models like the GARCH models are based on daily return and model volatility as a latent process. This unavoidably leads to the loss of intraday market information. In recent years, high frequency data in financial markets have been available and *realized volatility*, being the sum of squared intraday returns, is taken as a proxy and an unbiased estimator for actual volatility.

   An alternative measure of volatility is the daily range which is the difference between the daily highest and lowest prices. The daily range is also an unbiased estimator of daily volatility and is shown to be five times more efficient than the squared daily return. Moreover the Conditional Autoregressive Range (CARR) model, proposed for analyzing range data, provides better volatility forecast than the traditional GARCH model. Hence the *realized range* defined as the sum of high-low range for intraday interval is also shown to be more efficient than the realized volatility.

Sampling frequency related to the intraday interval is very important to the realized range and five-minutes frequency is suggested as the best way to avoid microstructure error of the market. This project compares different volatility models based on a range of volatility measures from high frequency data and proposes some guidelines in choosing volatility models to analyze high frequency data.

4. **Random Priors over Spatial-Temporal Partitions for Non-stationary Gaussian Processes**
   Supervisor: Prof. Sally Cripps

   Project Description: Gaussian Process Priors over functions have been extensively used for the development of nonparametric regression models. Typically a function is decomposed into a polynomial of degree, m, say, and deviations from this polynomial are assumed to follow a Gaussian process, with zero mean and stationary covariance function. The assumption of a stationary covariance can be too restrictive in many applied fields such as in geology and FMRI images, where "smooth" regions are partitioned by abrupt changes or discontinuities. The goal in this project is to develop priors over possible partitions of the covariate space, where the number and shape of these regions are assumed finite, but unknown, so that estimates of the unknown function can capture both the relatively smooth regions as well as the sudden changes. The space of possible partitions can be very large and the challenge is to develop priors which restrict this space, and computationally efficient algorithms, which explore this restricted space, that result in function estimates that are very flexible yet computationally feasible.

5. **Bayesian Variable Selection in high dimensional observational data**
   Supervisor: Prof. Sally Cripps

   Project Description: In today's data rich world, researchers have access to measurements on a very large number of factors which may predict a particular outcome. Ironically, this ease of data capture makes the task of selecting important factors very challenging. Often the number of factors available for prediction on a given individual is larger than the number of individuals on whom we measurements, making the identification of important factors statistically challenging. For example inference in a frequentist procedure usually relies on the assumption of asymptotic normality of the sample estimates. While this assumption is generally correct for situations where the true number of factors, p, in the model is small relative to the number of observations n, i.e. p ¡¡ n, it is unlikely to hold as p ! n and for p ¿ n the maximum likelihood estimates (MLEs) do not even exist. Another related issue is that good predictive performance does not necessarily equate with the identification causal factors; many different combination of factors may be equally good in predicting. However in many situations policy makers need to know what factors are likely to be causal so that appropriate intervention strategies are used. To address the joint issues of high dimensionality and casual inference we take a Bayesian approach. Specifically we propose to reduce the dimensionality by using a horse shoe prior over the regression coffecients. This regularlization may result in biased estimates of the regression coefficients and to address this we develop a series of conditional models by dividing the covariates into two groups, those which have the potential to be changed by an intervention, and those which do not. These conditional models are typically based on very small sample sizes, n ¡ p, making variable selection important.

6. **Unbiased probability density estimation of multidimensional time-changed diffusion processes using Malliavin calculus and its error analysis**
   Supervisor: Dr. Ray Kawai

   Project Description: Probability density estimation of multidimensional time-changed diffusion processes [2010, 2017] is unbiased by employing the Malliavin calculus (stochastic calculus of variation), whereas this unbiased estimation method invites infinite estimator variance. In this project, we aim to develop error analysis of a perturbed version of the unbiased method (hence, biased with a finite estimator variance) along the lines of [2009].

   - Kohatsu-Higa, A., Yasuda, K. (2009) Estimating multidimensional density functions using the Malliavin-Thalmaier formula, *SIAM Journal on Numerical Analysis*, **47**(2) 1546-1575.
   - Kawai, R., Kohatsu-Higa, A. (2010) Computation of Greeks and multidimensional density estimation for asset price models with time-changed Brownian motion, *Applied Mathematical Finance*, **17**(4) 301-321.
   - Carnaffan, S., Kawai, R. (2017) Solving multidimensional fractional Fokker-Planck equations via unbiased density formulas for anomalous diffusion processes, *SIAM Journal on Scientific Computing*, in press.

7. **Minimization of finite sums with stochastic gradient methods**
   Supervisor: Dr. Ray Kawai

   Project Description: There has been an explosion of interest in stochastic gradient methods for computing a minimizer of a finite sum of functions measuring misfit over a large number of data points. The goal of this project is to get a good understanding of this very quickly-evolving area and explore many possible variants on the existing algorithms for further improvements. This project benefits from new creative ideas and good coding skills.

   - Schmidt, M., Le Roux, N., Bach, F. (2017) Minimization of finite sums with the stochastic gradient, *Mathematical Programming*, **162**(1) 83-112.
   - Robbins, H., Monro, S. (1951) A stochastic approximation method, *The Annals of Mathematical Statistics*, **22**(1) 400-407.

8. **False Discovery Rate (FDR)**
   Supervisor A/Prof. Uri Keich

   Project Description: The multiple testing problem arises when we wish to test many hypotheses at once. Initially people tried to control the probability that we falsely reject at least one true null hypothesis. However, in a ground breaking paper Benjamini and Hochberg suggested that alternatively we can control the false discovery rate (FDR): the expected percentage of true null hypotheses among all the rejected hypotheses. Shortly after its introduction FDR became the preferred tool for multiple testing analysis with the original 1995 paper garnering over 35K citations. There are several related problems in the analysis of false discoveries that would be intriguing to explore.

9. **Fast exact tests**
Supervisor A/Prof. Uri Keich

   Project Description: Exact tests are tests for which the statistical significance is computed from the underlying distribution rather than, say using Monte Carlo simulations or saddle point approximations. Despite of their accuracy exact tests are often passed over as they tend to be too slow to be used in practice. We recently developed a technique that fuses ideas from large-deviation theory with the FFT (Fast Fourier Transform) that can significantly speed up the evaluation of some exact tests. In this project we would like to explore new ideas that we allow us to expand the applicability of our approach to other tests.

10. **Ultra-high variable screening**
Supervisor: A/Prof. Samuel Muller

    Project description: This project will review recent literature in ultra-high variable screening, computationally fast and ingenious methods to sort out the 'good from the ugly'. Ultra-high means that there are million's of variable, too many so that any regression procedure can be run with the full data. The task is to safely eliminate that part of the design matrix which is guaranteed to be uninformative. Such variable screening is an essential pre-step for successful model selection methods.

11. **Robust model selection criteria, specific examples and R package**
Supervisor: A/Prof. Samuel Muller

    Project description: Mueller and Welsh (2005;09) introduced methods to robustly select variables in a regression type model using the bootstrap. This project would revisit their methods and special additional cases will be identified first and then investigated. One aim of the project could be to make available an R-package or at least an R-function. There are also additional algorithms that could be explored that do not require to have to consider all possible submodels, i.e. how to robustly reduce the powerset of models with fast and robust methods before turning attention to more computationally expensive but more efficient model selectors is a potential important question as well.

12. **Vector Autoregressive Fractionally Integrated Moving Average (VARFIMA) Processes and Applications**
Supervisor: A/Prof. Shelton Peiris

    Project description: This project extends the family of autoregressive fractionally integrated moving average (ARFIMA) processes to handle multivariate time series with long memory. We consider the theory of estimation and applications of vector models in financial econometrics.

    - Tsay, Wen-Jey (2012). Maximum likelihood estimation of structural VARFIMA models, *Electoral Studies,* **31,** 852-860.
    - Sela, R.J. and Hurvich, C.M. (2008). Computationally Efficient Gaussian Maximum Likelihood Methods for Vector ARFIMA Models.
    - Wu, Hao and Peiris, S. (2017). Analysis of Vector GARFIMA Processes and Applications (Working paper).

13. **Theory of Bilinear Time Series Models and Applications in Finance**
Supervisor: A/Prof. Shelton Peiris

Project description: This project associated with employing the theory and applications of bilinear time series models in finance. Various extensions including the integer valued bilinear models and their state space representations are considered. Sufficient conditions for asymptotic stationarity are derived.

- Rao, T.S. (1981), On the Theory of Bilinear Time Series moedls, *J.R.Statist.Soc. B,* **43,** 244-255.
- Doukhna, P., Latour, A., Oraichi, D.(2006), A Simple Integer-Valued Bilinesr Time Series Model, *Adv. Appl. Prob.,* **38,** 559-577.

14. **Using orthonormal series for goodness of fit testing and mixture detection**
Supervisor: Dr. Michael Stewart

Project description: Suppose X has density $f(\cdot)$ and the (infinite) collection of functions $\{g_j(\cdot)\}$ is such that the random variables $g_1(X), g_2(X), \ldots$ all have mean 0, variance 1 and are uncorrelated. Then we say the g_j's are *orthonormal* with respect to $f(\cdot)$.

If $X_1, \ldots, X_n$ are a random sample from $f(\cdot)$ then the *normalised sample averages* $\bar{G}_1, \bar{G}_2, \ldots$ given by

$$\bar{G}_j = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_j(X_i)$$

give a sequence of statistics, any finite subset of which are asymptotically standard multivariate normal with covariance the identity. These can be used to construct goodness-of-fit statistics for f. For instance for any fixed k, $\bar{G}_1^2 + \cdots + \bar{G}_k^2$ is asymptotically χ_k^2 and indeed the smooth tests of Neyman (1937) and chi-squared tests of Lancaster (1969) are of this form. More recently work has been done using *data-driven* methods for choosing k, for example Ledwina (1994) using BIC.

The project will involve two things:

(a) surveying the literature on the use of (normalised) sample averages of orthonormal functions for testing goodness of fit;

(b) the implementation (using R) and theoretical study of some new tests of this type with special interest in their performance under certain mixture alternatives, that is densities of the form $(1-p)f + pg$ for some $g \neq f$ and p positive but close to zero.

- H.O. Lancaster. *The chi-squared distribution.* Wiley, 1969.
- T. Ledwina. Data-driven version of Neyman's smooth test of fit. *J. Amer. Statist. Assoc.*, 89(427):1000–1005, 1994.
- J. Neyman. "Smooth" test for goodness of fit. *Skandinavisk Aktuaristidskrift*, 20:149–199, 1937.

15. **Robust regression for the analysis of crop breeding trials**
Supervisor: Dr. Emi Tanaka and A/Prof. Samuel Müller

Project description: Crop breeding trials are conducted routinely across locations and years (environment) with an aim to select varieties or breeding lines (genotype) with improvement in economic traits of interest. The analysis of these data are often conducted using linear mixed models for the predictions of genotype × environment (G×E) effects. Such predictions are, however, sensitive to the presence of outliers and thus a robust prediction of the G×E effects will be beneficial. In this project we explore the use of robust estimation of linear mixed models by Koller (2013) to the analysis of crop breeding trials.

- Koller (2013) Robust Estimation of Linear Mixed Models. PhD Thesis.

16. **Extensibility of linked block designs for boutique arrays**
    Supervisor: Dr. Emi Tanaka and Prof. Jean Yang

    Project description: Gene expression profiling is the determination of the pattern of genes expressed under specific conditions/circumstances and is a major tool for discovery in biological sciences, in particular medicine. Boutique arrays, where the technology only examine a subset of genes (approx 200-800 genes) as opposed to the whole genome (25,000 genes), are becoming increasingly common with such arrays typically providing a higher signal to noise ratio. Data generated from these boutique arrays tend to have a high proportion of changes between samples, which makes the design of replications a critical component. Traditional designs are based on a fixed known structure, however, this limits the flexibility and extensibility of designs to accommodate for new samples. In this project, we explore the effectiveness of extensible linked block designs for a particular type of boutique array known as NanoStrings nCounter.

17. **Finite sample performance of robust location estimators**
    Supervisor: Dr. Garth Tarr

    Project description: Consumer sensory scores are typically constrained within bounded intervals, for example when asked to give a score out of 100, however the measurements often exhibit outliers within that bounded interval. This project will investigate finding an optimal robust location estimator for bounded data with a focus on small sample performance. This project will consider various univariate and multivariate robust location estimators and assess their small sample performance. You will have access to an extensive sensory database with which to compare and contrast various techniques and put forward recommendations that will help shape the future of consumer sensory evaluation of lamb and beef. Development of more efficient processes and protocols for conducting and summarising consumer sensory scores will lead to substantial savings for future research experiments and/or enable more research to be done with the same amount of funding.

18. **Outlier identification and classification in functional data**
    Supervisor: Dr. Garth Tarr

    Project description: Functional data is where we observe a curve for each sample. Examples of functional data include growth curves, brain electrical activity and colour spectra measurements. It is important to be able to identify any unusual sample curves that do not align closely with the other observations so that they can be dealt with appropriately in any subsequent analysis of the data. This project will look at existing (and perhaps new) approaches to outlier identification in functional data. We will also consider classification of colour spectra into multiple categories and using colour spectra to predict consumer colour appreciation scores.

19. **Nonlinear cointegrating regression with latent variables**
    Supervisor: A/Prof. Qiying Wang

    Project description: Using the estimation theory currently developed in nonlinear regression with nonstationary time series, this topic will consider the links between untraded spot prices (such as DJIA index, S & P 500 index), traded ETFs, and traded financial derivatives, the traded Volatility index (VIX), and other derivatives.

20. **Testing for nonlinear cointergation**
    Supervisor: A/Prof. Qiying Wang

    Project description: This topic intents to develop residual-based test for various nonlinear cointergation models. Some empirical applications in money demand and other real time series data will be considered.

21. **Place Holder**
    Supervisor: Dr. Rachel Wang

    Project description: Please contact for potential projects.

22. **Classification and statistical network.**
    Supervisor: Prof Jean Yang

    Project description: Classical approaches in classification are primarily based on single features that exhibit effect size difference between classes. In omics data, this is equivalent to finding differential expression of genes or proteins between different treatment classes. Recently, network-based approaches utilising interaction information between genes have emerged and our recent work (Barter et al., 2014) further reveals that simple network based methods are able to classify alternate subsets of patients compared to gene-based approaches. This suggests that next-generation methods of gene expression signature modeling may benefit from harnessing data from external networks. This project will further explore the strength and weaknesses of utilizing statistical network as a feature in classification. The project will also extend Barter et al, 2014 by examining the effect of robust networks obtain from external databases or complementary datasets and evaluate its effect in classification (prognostic) setting.

    - Barter RL, Schramm SJ, Mann GJ, Yang YH. Network-based biomarkers enhance classical approaches to prognostic gene expression signatures. BMC systems biology. 2014 Dec;8.

23. **Methods towards precision medicine**
    Supervisor: Prof Jean Yang

    Project description: Over the past decade, new and more powerful -omic tools have been applied to the study of complex disease such as cancer and generated a myriad of complex data. However, our general ability to analyse this data lags far behind our ability to produce it. This project is to develop statistical method that deliver better prediction of response to drug therapy. In particular, this project investigates whether it is possible to establish the patient or sample specific network based (matrix) by integrating public repository and gene expression data.

24. **Rare cell type discovery using AdaSampling**
    Supervisor: Dr. Pengyi Yang and Prof. Jean Yang

    Single-cell RNA sequencing (scRNA-seq) is a revolutionary technique that enables the gene expression profiling of thousands of cells. One of the key task in scRNA-seq data analysis is to identify rare cell types that are hiding in the tissue samples such as various cancer stem cells in tumour tissues. AdaSampling is a semi-supervised machine learning approach that we developed recently for detecting noisy samples in a given dataset. In this project, we will be look into transferring the AdaSampling technique for identifying rare cell types that are previously unknown and therefore labeled incorrectly in the initial dataset. Requirements: good programming skill (preferentially in R) and basic understanding of statistical learning (desirable).

25. **Estimation of transcription networks based on epigenetic data**
    Supervisor: Dr. Pengyi Yang and Dr. Ashnil Kumar (School of Information Technologies)

    Project description: Predicting transcription factor binding sites using deep learning The advance of ultrafast sequencing (ChIP-seq) allows the profiling of transcription factor (TF) binding sites genome-wide in a cell. The massive amount of sequencing data generated from these genome-wide profiling of TF requires sophisticated computational algorithm to be developed for accurately identifying TF binding sites. Deep learning is the latest development in machine learning that has been successfully utilised to address many bioinformatics applications. In this project, we aim to develop and apply deep learning models for predicting TF binding sites by integrating ChIP-seq data with other biological knowledge. This project will allow you the opportunity to develop and apply cutting-edge deep learning algorithms for solving a key biological problem. You will get involved in all aspects of the development including algorithm design, implementation and testing. Requirements: good programming skill (essential) and experience in deep learning (desirable).

7 Assessment

7.1 The Honours grade

The examiners' recommendation to the Faculty of the student's Honours grade is based on the average mark achieved by each student, over the 6 best courses and the project. Courses account for 60% of the assessment and the project for the remaining 40%.

According to the Faculty of Science guidelines, the grade of Honours to be awarded is determined by the Honours mark as follows:

Grade of Honours	Faculty-Scale
First Class, with Medal	95–100
First Class (possibly with Medal)	90–94
First Class	80-89
Second Class, First Division	75-79
Second Class, Second Division	70-74
Third Class	65-69
Fail	0-64

The Faculty has also given the following detailed guidelines for assessing of student performance in Honours.

95–100 Outstanding First Class quality of clear Medal standard, demonstrating independent thought throughout, a flair for the subject, comprehensive knowledge of the subject area and a level of achievement similar to that expected by first rate academic journals. This mark reflects an exceptional achievement with a high degree of initiative and self-reliance, considerable student input into the direction of the study, and critical evaluation of the established work in the area.

90-94 Very high standard of work similar to above but overall performance is borderline for award of a Medal. Lower level of performance in certain categories or areas of study above.

Note that in order to qualify for the award of a university medal, it is necessary but not sufficient for a candidate to achieve a SCIWAM of 80 or greater and an Honours mark of 90 or greater. Faculty has agreed that more than one medal may be awarded in the subject of an Honours course.

The relevant Senate Resolution reads: "A candidate with an outstanding performance in the subject of an Honours course shall, if deemed of sufficient merit by the Faculty, receive a bronze medal."

80-89 Clear First Class quality, showing a command of the field both broad and deep, with the presentation of some novel insights. Student will have shown a solid foundation of conceptual thought and a breadth of factual knowledge of the discipline, clear familiarity with and ability to use central methodology and experimental practices of the discipline, and clear evidence of some independence of thought in the subject area.

Some student input into the direction of the study or development of techniques, and critical discussion of the outcomes.

75-79 Second class Honours, first division student will have shown a command of the theory and practice of the discipline. They will have demonstrated their ability to conduct work at an independent level and complete tasks in a timely manner, and have an adequate understanding of the background factual basis of the subject. Student shows some initiative but is more reliant on other people for ideas and techniques and project is dependent on supervisor's suggestions. Student is dedicated to work and capable of undertaking a higher degree.

70-74 Second class Honours, second division student is proficient in the theory and practice of their discipline but has not developed complete independence of thought, practical mastery or clarity of presentation. Student shows adequate but limited understanding of the topic and has largely followed the direction of the supervisor.

65-69 Third class Honours performance indicates that the student has successfully completed the work, but at a standard barely meeting Honours criteria. The student's understanding of the topic is extremely limited and they have shown little or no independence of thought or performance.

0-64 The student's performance in fourth year is not such as to justify the award of Honours.

7.2 The coursework mark

Students are required to attend a minimum of 6 courses during the academic year. Only the best 6 results will be included in the overall assessment. These 6 results are weighted equally.

Student performance in each honours course is assessed by a combination of assignments and examinations. The assignment component is determined by the lecturer of each course and the examination component makes up the balance to 100%. The lecturer converts the resulting raw mark to a mark on the above mentioned Faculty scale, which indicates the level of Honours merited by performance in that course alone.

7.3 The project mark

The project's mark is split 90% for the essay and 10% for the student's presentation. The presentation mark is determined by the stats staff attending the presentation.

The essay is assessed by three members of staff (including the supervisor). The overall final mark for the essay is a weighted mean of all three marks awarded. A weighting of 50% is attached to the supervisor's original mark, while a weight of 25% is attached to each of the two marks awarded by the other examiners.

The criteria which the essay marks are awarded by each examiner include:

- quality of synthesis of material in view of difficulty and scope of topic, and originality, if any.
- evidence of understanding.
- clarity, style and presentation.
- mathematical and/or modelling expertise and/or computing skills.

The student's supervisor will also consider the following criteria:

- Has the student shown initiative and hard work which are not superficially evident from the written report?

- Has the student coped well with a topic which is too broad or not clearly defined?

7.4 Procedures

All assessable student work (such as assignments and projects) should be completed and submitted by the advertised date. If this is not possible, approval for an extension should be sought in advance from the lecturer concerned or (in the case of honours projects) from the Program Coordinator. Unless there are compelling circumstances, and approval for an extension has been obtained in advance, late submissions will attract penalties as determined by the Board of Examiners (taking into account any applications for special consideration).

Appeals against the assessment of any component of the course, or against the class of Honours awarded, should be directed to the Head of School.

Note: Students who have worked on their projects as Vacation Scholars are required to make a declaration to that effect in the Preface of their theses.

8 Seminars

Mathematical Statistics seminars are usually held fortnightly on Friday afternoons. These seminars are an important forum for communicating ideas, developing critical skills and interacting with your peers and senior colleagues. Seminars are usually given by staff members and invited speakers. All honours students are encouraged to attend these seminars. Keep in mind that attending these seminars might help develop your presentation skills.

9 Entitlements

Mathematical Statistics 4 students enjoy a number of privileges, which should be regarded as a tradition rather than an absolute right. These include:

- Office space and a desk in the Carslaw building.
- A computer account with access to e-mail and the WorldWideWeb, as well as LaTeX and laser printing facilities for the preparation of projects.
- Photocopy machine for any of your work related material.
- After-hours access to the Carslaw building.
- A pigeon-hole in room 728 please inspect it regularly as lecturers often use it to hand out relevant material.
- Participation in the Schools social events.
- Class representative at School meetings.

10 Scholarships, Prizes and Awards

University of Sydney Honours Scholarships
These $6,000 Honours Scholarships are awarded annually on the basis of academic merit and personal attributes such as leadership and creativity.

The following prizes may be awarded to statistics honours students of sufficient merit. Students do not need to apply for these prizes, which are awarded automatically. The complete list is available here.

The Joye Prize
Awarded annually to the most outstanding student completing fourth year Honours in Applied Mathematics, Pure Mathematics or Mathematical Statistics (provided the work is of sufficient merit).

George Allen Scholarship
This is awarded to a student proceeding to honours in Mathematical Statistics who has shown proficiency in all Senior units of study in Mathematical Statistics.

University Medal
Awarded to Honours students who perform outstandingly. The award is subject to Faculty rules, which require a mark of at least 90 in Mathematical Statistics 4 and a SCIWAM of 80 or higher. More than one medal may be awarded in any year.

Ashby Prize
Offered annually for the best essay, submitted by a student in the Faculty of Science, that forms part of the requirements of Honours in Pure Mathematics, Applied Mathematics or Mathematical Statistics.

Barker Prize
Awarded at the fourth (Honours) year examination for proficiency in Pure Mathematics, Applied Mathematics or Mathematical Statistics.

Norbert Quirk Prize No IV
Awarded annually for the best best entry to the SUMS Competition by an honours student.

Veronica Thomas Prize
Awarded annually for the best honours presentation in statistics.

Australian Federation of University Women (NSW) Prize in Mathematics
Awarded annually, on the recommendation of the Head of the School of Mathematics and Statistics, to the most distinguished woman candidate for the degree of BA or BSc who graduates with first class Honours in Applied Mathematics, Pure Mathematics or Mathematical Statistics.

11 Life after Fourth Year

Students seeking assistance with post-grad opportunities and job applications should feel free to ask lecturers most familiar with their work for advice and written references. The Head of Statistics Programme, the Program Coordinator and the course lecturers may also provide advice and personal references for interested students.

Students thinking of enrolling for a higher degree (MSc or PhD) should direct all enquiries to the Director of Postgraduate Studies:

`pg-director@maths.usyd.edu.au`

Students are also strongly encouraged to discuss potential research topics with individual staff members.

Students who do well in their honours studies may be eligible for postgraduate scholarships, which provide financial support during subsequent study for higher degrees.

Last but not least, there is a number of jobs for people with good statistical knowledge. Have a look here.

12 Additional proposed project topics in Mathematical Statistics

1. **Volatility models using flexible range information**
   Supervisor: A/Prof. Jennifer Chan

   Project description: Volatility forecast is important in risk management. Since volatility is unobserved, most volatility models like the GARCH and stochastic volatility models are based on daily return and model volatility as a latent process. This unavoidably leads to the loss of intraday market information.

   In recent years, high frequency data in financial markets have been available and *realized range*, being the sum of squared range over many short, say 5-minutes, intervals of a day, is an unbiased estimator of daily volatility. As it can capture the intraday market information, it was shown to be five times more efficient than the squared daily return for the realized volatility.

   Other range measures such as interquartile range is robust and hence should provide a favorable alternative to the realized range measure. However this kind of range measures is still incapable for measuring the volatility dynamic when the distribution is asymmetric. Subsequently, half range, upper and lower, measures are proposed for more general distributions.

   This project will compare the efficiency of modeling volatility using the Conditional Autoregressive Range (CARR) model based on different types of realized range measures. It involves searching over high frequency data, calculation of various range measures, model implementation and forecast. Hopefully, some guidelines in choosing range measures to analyze high frequency data will be provided after the study.

2. **Parametric quantile regression models for Value-at-risk forecast**
   Supervisor: A/Prof. Jennifer Chan

   Project description: Quantile regression is emerging as a comprehensive tool to the statistical analysis of linear and nonlinear response models for value-at-risk calculation in risk management. By supplementing the exclusive focus of least squares based methods on the estimation of conditional mean functions with the estimation on the conditional quantiles of a distribution, a parametric quantile regression model provides great flexibility in the model structure. However, the general technique for estimating families of conditional quantile functions under a parametric approach is to first build a mean regression model and then calculate quantile functions based on the mean regression model.

   This project considers models that directly regress on the quantiles of distributions and hence they can reveal the change of covariate effects across quantile levels as the nonparametric quantile regression but they are free from the problem of crossover of quantile functions in the nonparametric approach. Distributions on the real and positive domains will be adopted and the Bayesian and classical likelihood methods of inference will be applied to estimate the model parameters.

3. **Singular Fisher information for stochastic processes under high frequency sampling**
   Supervisor: Dr. Ray Kawai

   Project Description: High frequency sampling has attracted much attention due to increasingly availability of high-resolution data, for example, of asset price dynamics in finance and individual animal movement in ecology. After understanding the concept of normal asymptotic normality, we investigate the asymptotic behavior of MLE under high frequency discrete sampling of some continuous time stochastic processes, in terms of the corresponding Fisher information matrix. Strong numerical experiment skill is essential.

   - Kawai, R., Masuda, H. (2011) On the local asymptotic behavior of the likelihood function for Meixner Lévy processes under high-frequency sampling, *Statistics and Probability Letters*, **81**(4) 460-469.
   - Kawai, R., Masuda, H. (2013) Local asymptotic normality for normal inverse Gaussian Lévy processes with high-frequency sampling, *ESAIM: Probability and Statistics*, **17**, 13-32.

4. **Exact simulation of stochastic differential equations**
   Supervisor: Dr. Ray Kawai

   Project Description: The exact method enables us to simulate a hitting time, and other functionals of a one-dimensional jump diffusion with state-dependent drift, volatility, jump intensity, and jump size. This acts as an alternative to the discretization-based approximate methods and eliminates the need to control the bias of a discretization-based simulation estimator. In this project, we will explore a variety of exact simulation methods with a view towards applications, including unbiased estimation of security prices, transition densities, hitting probabilities, and other quantities arising in jump-diffusion models. Strong numerical experiment skill is essential.

   - Beskos, A., Roberts, G.O. (2005) Exact simulation of diffusions, *Annals of Applied Probability*, **15**(4) 2422-2444

- Giesecke, K., Smelov, D. (2013) Exact sampling of jump diffusions, *Operations Research*, **61**(4) 894-907.

5. **FDR in mass spectrometry**
   Supervisor A/Prof. Uri Keich

   Project Description: In a shotgun proteomics experiment tandem mass spectrometry is used to identify the proteins in a sample. The identification begins with associating with each of the thousands of the generated peptide fragmentation spectra an optimal matching peptide among all peptides in a candidate database. Unfortunately, the resulting list of optimal peptide-spectrum matches contains many incorrect, random matches. Thus, we are faced with a formidable statistical problem of estimating the rate of false discoveries in say the top 1000 matches from that list. The problem gets even more complicated when we try to estimate the rate of false discoveries in the candidate proteins which are inferred from the matches to the peptides thus this project is really a framework for several different projects that involve interesting statistical questions that are critical to the correct analysis of this promising technology of shotgun proteomics. *No prior understanding of proteomics is required.*

6. **Generalizing Fisher Exact Test**
   Supervisor A/Prof. Uri Keich

   Project Description: Young et al. (2010) showed that due to gene length bias the popular Fisher Exact Test should not be used to study the association between a group of differentially expressed (DE) genes and a conjectured function defined by a Gene Ontology (GO) category. Instead they suggest a test where one conditions on the genes in the GO category and draws the pseudo DE expressed genes according to a length-dependent distribution. The same model was presented in a different context by Kazemian et al. (2011) who went on to offer a dynamic programming (DP) algorithm to exactly compute the significance of the proposed test. We recently showed that while valid, the test proposed by these authors is no longer symmetric as Fisher's Exact Test is: one gets different answers if one conditions on the observed GO category than on the DE set. As an alternative we offered a symmetric generalization of Fisher's Exact Test and provide efficient algorithms to evaluate its significance. After reviewing that work we will look into other approaches for testing enrichment and the question of how should one choose the "right" kind of enrichment test.

   - Majid Kazemian, Qiyun Zhu, Marc S. Halfon, and Saurabh Sinha. Improved accuracy of supervised crm discovery with interpolated markov models and cross-species comparison. *Nucleic Acids Research*, 39(22):9463–9472, Dec 2011.
   - MD Young, MJ Wakefield, GK Smyth, and A Oshlack. Gene ontology analysis for rna-seq: accounting for selection bias. *Genome Biology , 11:R14*, 11, 2010.

7. **Regularization methods, does the sign of correlation coefficients matter?**
   Supervisor: A/Prof. Samuel Muller

   Project description: Regularization methods such as Ridge regression, Lasso or the adaptive Lasso aim to deal with both, high correlations in the predictor variables and when there are more variables, p, than observations, n, i.e. the currently very popular large p small n problem. There are rumours that the way negatively correlated variables impact these regularization methods is different to how positive variables do. This project will investigate that rumour with the aim to find either supporting or contradicting empirical and maybe even theoretical evidence.

8. **Learning from changing slopes to identify the better classifier**
   Supervisor: A/Prof. Samuel Muller

   Project description: A receiver operating characteristic curve (ROC curve), is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. To compare and test different ROC curves is an ongoing challenge. There are various tests, which will be studied first and then the project aims to explore new ROC curve tests that are based on learning how the slopes in the various ROC curves differ. For example, in a single sample problem, testing whether or not the ROC curve is better than flipping a coin, any test measures essentially how much the curve deviates from the identity, for example through studying the maximum/minimum of the derivative (slope) of the ROC curve.

9. **Efficient calculation of data depth**
   Supervisor: A/Prof. Samuel Muller

   Project description: Data depth is a concept from robust statistics that can be used to measure the "depth" or "outlyingness" of a given multivariate sample with n observations in $\mathbb{R}^d$ with respect to its underlying distribution. In the last 40 years there were hundreds of different suggestions to define such depth measures. Therefore, in the first part of the project, different depth measures will be revised and classified according to their computing cost. The second part of the project will aim to investigate to what extent repeated subsampling of k out of n points can be more efficient than working with the full sample to calculate data depth. This is motivated by the simple observation that Bk^d can be considerably less than n^d, where B is the number of resamples.

10. **Measuring instability of model selection methods**
    Supervisor: A/Prof. Samuel Muller

    Project description: Nan and Yang (2014; Journal of Computational and Graphical Statistics, 23:636-656) introduce the notion of variable selection deviation, a new concept to assess the stability of selected models. This notion is particularly useful for high-dimensional data settings where the number of variables can be much larger than the sample size. Such situations require more modern fitting and selection methods than what was covered in the undergraduate courses, where variables were selected using stepwise procedures that are based on p-values or on information criteria such as AIC or BIC. This project is on implementing the methods and concepts of Nan and Yang and on exploring how well they perform on real data and when considering traditional model selection methods.

11. **Robust monotone curve estimation**
    Supervisor: A/Prof. Samuel Muller

    Project description: In many regression settings it is known, e.g. from some underlying physical or economic theory, that the regression curve is monotone. Further examples include, but are not limited to, calibration problems, estimation of monotone transformations (e.g. to transform a variable to normality), growth curves, and dose-response curves. The objective of this project is to robustify using smoothing splines via the smooth.monotone function which is part of the R-package fda (Ramsay et al, 2012). The initial motivation for this work is to fit a monotone decreasing function into the measured motor evoked potentials (TST amplitude) as a function of stimulation delay on more than 40 different data sets, one from each participating patient in a recent health study and with published results in Firmin, Müller and Rösler (2011,2012; Clinical Neurophysiology)

12. **Are uncorrelated variables more likely to be selected?**
    Supervisor: A/Prof. Samuel Muller

    Project description: This project will explore empirically for general situations and possibly theoretically in toy examples whether or not adding a non-correlated / independent variable which is known to be redundant is selected more often than other redundant variables but correlated with important features when using criteria such as AIC in situations where AIC is known to choose models that are slightly too large.

13. **Fractional Differencing and Long Memory Time Series Analysis with Stochastic Variance: Applications to Financial Statistics**
    Supervisor: A/Prof. Shelton Peiris

    Project description: In recent years, fractionally-differenced processes have received a great deal of attention due to their flexibility in financial applications with long-memory. This project considers the family of fractionally-differenced processes generated by ARFIMA (Autoregressive Fractionally Differenced Moving Average) models with both the long-memory and time-dependent innovation variance. We aim to establish the existence and uniqueness of second-order solutions. We also extend this family with innovations to follow GARCH and stochastic volatility (SV). Discuss a Monte Carlo likelihood method for the ARFIMA-SV model and investigate finite sample properties. Finally, illustrate the usefulness of this family of models using financial time series data.

    - Peiris, S. and Asai, M. (2016). Generalized Fractional Processes with Long Memory and Time-Dependent Volatility Revisited, *Econometrics,* **4(3)**, No 37, 21 pages.
    - Bos, C., Koopman, S.J., Ooms, M. (2014). Long memory with stochastic variance model: A recursive analysis for US inflation, *Computational Statistics & Data Analysis,* **76,** 144-157.
    - Ling, S., Li, W.K. (1997). On fractionally integrated autoregressive moving average time series with conditional heteroscedasticity, *Journal of American Statistical Association,* **92,** 1184-1194.

14. **Second-order least-squares estimation for regression with autocorrelated errors**
    Supervisor: A/Prof Shelton Peiris

    Project description: In their recent paper, Wang and Leblanc (2008) have shown that the second-order least squares estimator (SLSE) is more efficient than the ordinary least squares estimator (OLSE) when the errors are iid (independent and identically distributed) with non zero third moments. In this paper, we generalize the theory of SLSE to regression models with autocorrelated errors. Under certain regularity conditions, we establish the consistency and asymptotic normality of the proposed estimator and provide a simulation study to compare its performance with the corresponding OLSE and GLSE (Generalized

Least Square Estimator). In addition we compare the efficiency of SLSE with OLSE and GLSE in estimating parameters of such regression models with autocorrelated errors.

- Wang, L and Leblance (2008), Second-order nonlinear least squares estimation, *Ann. Inst. Stat. Math.,* 883-900.
- Rosadi, D. and Peiris, S. (2014), Second-order least-squares estimation for regression models with autocorrelated errors, *Computational Statistics,* **29,** 931-943. (su

15. **On the estimate of heritability and its uncertainty**
Supervisor: Dr. Emi Tanaka

Project description: Heritability is a key measure commonly used in genetics and plays a central role in many practical decision for selective breeding. It is a summary of the proportion of the phenotypic variability that is attributed to the average effects of the genes and determines the degree of resemblance between relatives. There are a number of existing methods for estimating heritability, however (1) estimates may be out of boundary with little attention given on how to deal with such situation and (2) very little work is done in quantifying the precision of these estimates. In this project we investigate various measures of heritability and their uncertainty based on specific applications to selective breeding.

16. **Spatial model selection for field trials**
Supervisor: Dr. Emi Tanaka and Dr. Garth Tarr

Project description: With an exponentially increasing population, there is a pressing need to increase the rate of genetic gain to meet the demand for food. Selective breeding plays a crucial role in meeting such demand. Crop breeding trials are routinely conducted with the aim of predicting genetic performance, or the so-called breeding values, of the test lines. Linear mixed model is the prevailing method used to analyse crop breeding trials owing to their flexible framework to accommodate the analysis of complex data. However, the economic traits of crops, such as yield, are largely affected by the spatial variation and trend within the field, stemming from environmental factors, such as soil fertility and management practices. A data-driven approach may be used to select these variation via the approach of Gilmour et al (1997) and Stefanova et al (2009). This step motivates an application-specific model selection. In this project, we aim to develop best statistical practice for model selection and diagnostic for the analysis of a real wheat breeding trial.

- Gilmour et al. (1997) Accounting for Natural and Extraneous Variation in the Analysis of Field Experiments. *Journal of Agricultural, Biological, and Environmental Statistics* **2** (3) 269–293
- Stefanova et al. (2009) Enhanced diagnostics for the spatial analysis of field trials. *Journal of Agricultural, Biological, and Environmental Statistics* **14**(4) 392–410

17. **On the computing strategies for linear mixed models**
Supervisor: Dr. Emi Tanaka and Dr. John Ormerod

Project description: Linear mixed models offer a general, flexible framework that fits the structure of many complex, often correlated data and are widely used in a variety of disciplines in the physical, biological and social sciences. Modelling of the so-called "big data" often offers computational challenges peripheral to the aim of the analysis. Johnson and

Thompson (1995) introduced what is now widely known as the AI-REML algorithm that largely elevated the speed to fit a general linear mixed model which is used in the R package asreml-R (Butler et al. 2009) motivated by analysis of animal and plant breeding trials. However, with large data, it is not feasible to use asreml-R to fit a model in a timely manner. The biggest bottleneck of the algorithm is solving the augmented mixed model equations (Henderson, 1949) – which is equivalent to the fundamental problem: "solving a system of linear equations". Many advancements have been made in solving this fundamental problem with an implementation available in optimised LAPACK routines with some exploiting parallel computing to achieve high performance. In this project, we aim to explore various computing strategies for linear mixed models with speed performance testing on different scenarios motivated by crop breeding trials. Strong linear algebra background is necessary for this project. This project is only offered to those with strong mathematical and computational skills.

- Henderson (1949) Estimation of changes in herd environment (Abstract)
- Johnson & Thompson (1995) Restricted Maximum Likelihood Estimation of Variance Components for Univariate Animal Models Using Sparse Matrix Techniques and Average Information. *Journal of Dairy Science.* **78** (2) 449–456
- Bulter et al. (2009) Mixed models for S language environments ASReml-R reference manual.

18. **Ordering strategies for linear mixed models with genetic relatedness matrix**
Supervisor: Dr. Emi Tanaka and Dr. John Ormerod

Project description: The algorithm underlying the fitting of linear mixed models generally involves solving a sparse set of linear equations. The linear equations are solved using a form of gaussian elimination which in its simplest form is adding a multiple of one row to another. During this step, some zero elements may become non-zero which is termed *fill-in*. For computational efficiency, we ideally would reorder the equations to minimise fill-in. One permutation scheme is the minimum degree, which effectively eliminates rows with the fewest non-zero elements first. This scheme works well in general however can cause issues for particular models with pedigree information in the analysis of crop breeding trials. In this project we explore other ordering strategies for linear mixed models which incorporate genetic relatedness matrix. This project is only offered to those with strong mathematical and computational skills.

19. **Outlier detection for complex linear mixed models**
Supervisor: Dr. Emi Tanaka

Project description: Outlier detection is an important preliminary step in the data analysis often conducted through a form of residual analysis. A complex data, such as those that are analysed by linear mixed models, gives rise to distinct levels of residuals and thus offers additional challenges for the development of an outlier detection method. A mean (variance) shift outlier model assumes an i-th observation has a shifted location (inflated variance) and test this assumption. This can be easily incorporated into a standard linear mixed model software, and is computationally efficient for routine use. This method of outlier detection, however, is susceptible to swamping and masking of outliers. In this project we explore methods for a group outlier detection. The method development will be based on a set of real wheat yield trials.

20. **Optimal model-based designs for plant improvement programs**
Supervisor: Dr. Emi Tanaka

    Project description: Plant improvement programs aim to identify new varieties or breeding lines with enhanced traits (e.g. yield) for use as parents or for commercial release. The identification is generally based on data from designed experiment, often conducted on a number of fields across years and location. These fields are naturally fixed in size thus it is critical to consider treatment composition (i.e. the distribution of the lines across location) and replication schemes (i.e. which lines to replicate) for optimal genetic gain. A class of design that is widely adopted is optimal design owing to its flexibility to accommodate a number of factors and constraints via a model-based approach. A particular advantage of optimal designs, for example, is the ability to take into account correlated treatment and random effects. This is crucial for plant breeding experiments where often test lines are genetically related and traits exhibit spatial trends. Taking into account the latter trend, however, presents some experimental design issues. In this project, we investigate this issue and present an alternate solution.

21. **Improved model averaging through better model weights**
Supervisor: Dr. Garth Tarr

    Project description: Model averaging seeks to address the issue post model selection inference by incorporating model uncertainty into the estimation process. This project will investigate different weighting approaches used to obtaining model averaged estimates. Existing approaches will be compared to a new method where model weights are obtained through bootstrapping.

22. **Modelling money demand by nonlinear cointegration**
Supervisor: A/Prof. Qiying Wang

    Project description: A classical application of linear cointegration analysis is the long term money demand modelling. The assumption of linearity, however, is often too restrictive in practice. Using the estimation theory currently developed in nonlinear coinegrating regression, this topic will consider the money demand function estimation by nonlinear cointegration.

23. **Nonlinear cointegration regression with endogeneity**
Supervisor: A/Prof. Qiying Wang

    Project description: A general framework on the convergence to stochastic integrals is currently established in Peng and Wang (2016). In this topic, we consider the applications of this framework to nonlinear cointgerating regression. Some empirical applications in money demand and other real time series data will be considered.

24. **Functional limit theorem for a class of martingales with applications**
Supervisor: A/Prof. Qiying Wang

    Project description: A new martingale limit theorem was established in Wang (2015, ET). This new result is vital in the development of non-linear cointegrating regression. In this topic, we investigate the extension of Wang (2015) to functional limit theorems with applications in non-linear cointegrating regression.

25. **Mixture modelling for high throughput data**
    Supervisor: Prof. Jean Yang and Dr. Michael Stewart

    Project description: In recent years, single cell RNA-Sequencing (scRNA-Seq) has become a key technology in understanding the variability and complexity of individual cells within a tissue. This technology however has raised issues in analysis of the arising data due to distinct characteristics including a large proportion of exactly zero expression (dropout) as well as bimodal or multimodal distributions of the non-zero gene expression values. An approach is to use a two-component gamma-normal mixture modelling to classify each individual cell for each gene into multiple transcriptional states (components) for each gene.This works well for genes that exhibit a biomodal distribution but there remains a large proportion of genes with unimodal or multimodal distributions that are not well-explained by the model. This project aims to examine an approach to select and fit the most appropriate mixture model for each gene in a high-dimensional setting and visualise and assess the model fit.

26. **Tests for publication bias, and their applicability to variance-based effect sizes.**
    Supervisor: Dr. Alistair M Senior and Prof. Jean Yang

    Project description: Meta-analysis is now considered the gold standard for quantitatively assessing the evidence for a given phenomenon in a range of fields. To date meta-analysis has largely been concerned with evaluating differences in central tendency between groups, or the magnitude of correlations. More recently however, a newly defined set of effect sizes related to variance are increasing in popularity. The behavior of these new statistics in standard meta-analytic tests for publication bias remains questionable, yet these tests represent an important component of any meta-analysis. This project aims evaluate the behavior of variance-based effect sizes in common meta-analytic tests for publication bias, using simulated and/or real data, and if necessary to develop new tests of publication bias suitable to these statistics.