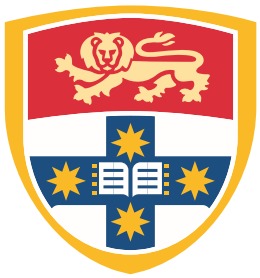
Honours in Mathematical Statistics

Detailed Guide for the 2017 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, 2017) is Wednesday the 30th November 2016.

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 on 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 honours 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 Access Grid
- Up to one course offered at the AMSI Summer School (January 2017)

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

For the second semester of 2016 the acting director of the Statistics teaching program is

Dr. John Ormerod,
Carslaw Building, Room 815, Phone 9351 5883,
Email: john.ormerod@sydney.edu.au

The Statistics Honours Program Coordinator in 2017 is

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

The current director of the Statistics teaching program is

Dr. Jennifer Chan,
Carslaw Building, Room 817, Phone 9351 4873,
Email: jennifer.chan@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

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

Dr 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, Analysis of DNA replication Origins, Computational Statistics, Statistical Analysis of Proteomics Data

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

Dr John Ormerod
Variational Approximations, Generalised Linear Mixed Models, Splines, Data Mining, Semiparametric Regression, Missing Data, Model Selection.

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

Dr Michael Stewart
Mixture models, Extremes of Stochastic Processes, Empirical Process Approximations, Density Estimation, Feature Selection, Applied and Computational Statistics.

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

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.

Dr Lamiae Azizi
Graphical modelling, Variational methods, Bayesian nonparametrics, Clustering, Classification; Spatial and spatio-temporal modelling, Applications of statistics to complex diseases (e.g. Cancers), fMRI and Genomics data.

Dr 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 2017. Please note that some, like Probability Theory will be offered as full courses while others like will be paired as two 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. **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.

3. **Fundamental 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.

4. **Generalized Linear Models**

The following topics will be covered:

- **Maximum Likelihood Inference**

  Newton-Raphson and Fisher Scoring methods, Expectation Maximization (EM), Monte Carlo EM and Expectation Conditional Maximization (ECM) algorithms. Scale mixtures presentation.

- **Exponential Family**

  Generalized linear models; Exponential family, Weighted least squares, Quasi-likelihood, BLUP estimator, Generalized estimating function, Random effects models.

- **Model Selection**

  Deviance for Likelihood Ratio Tests, Wald Tests, Akaike's information criterion (AIC) and Bayesian information criterion (BIC), Examples.

- **Logit models for binary data**

  Logistic regression; Binary response designs; Two way contingency tables; Exact tests; Matched case control data.

- **Poisson models for count data**

  Distributions for count data with equidispersion and overdispersion.

- **Log-linear models for categorical data**

  Hierarchical log-linear models; Decomposable models; incomplete tables; Quasi-independence; Tests for symmetry and marginal homogeneity.

- **Survival Analysis**

  Kaplan-Meier estimator; Proportional hazards models; Cox's proportional hazards model.

5. Introduction to Stochastic Calculus with applications

This course covers elementary stochastic processes, Brownian motion, stochastic integrals, Ito's formula, simple stochastic differential equations. Applications to mathematical finance, biology and engineering will be given.

Assumed knowledge: STAT 3911 Stochastic Processes and Time Series (Advanced) or equivalent.

References:

- Hui-Hsiung Kuo: Introduction to Stochastic integration. Springer, 2006.
- Klebaner, F. C.: Introduction to stochastic calculus with applications. Imperial College Press, 1998.
- Oksendal, B.: Stochastic differential equations (Sixth Edition). Springer, 2005.

6. Probability Theory

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

Contents: Axiomatic probability: probability space; continuity of probability measures; independence; product spaces; conditional probability and conditional expectations with respect to a given *sigma*-field; inequalities. Modes of convergence: almost sure convergence; convergence in probability; convergence in distribution. Characteristic functions: properties; inversion theorem and continuity. The Helly-Bray lemma; convergence via characteristic functions. Limit theorems: Laws of Large Numbers; Central Limit Theorem (Lindeberg); infinitely divisible distributions.

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

References:

- Billingsley, P. *Probability and Measure*, 1995.
- Chung, K. L. *A Course in Probability*, 1974

7. Statistical Methods in Bioinformatics

Contents: Bioinformatics is a field that applies ideas from computer science, mathematical modelling, and statistics in order to make sense of the huge datasets that typify current research in biology. 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, 2005.
- Richard Durbin et al. Biological sequence analysis: probabilistic models of proteins and nucleic acids, 1999.

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

1. Volatility models for high frequency data

Supervisor Dr 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 analysing 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 minute 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 analyse high frequency data.

2. Volatility models using flexible range information

Supervisor Dr 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 favourable 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 modelling 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 analyse high frequency data will be provided after the study.

3. **Parametric quantile regression models for Value-at-risk forecast**
Supervisor Dr 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.

4. **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, developed in [2010], 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.

5. **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 behaviour 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 behaviour of the likelihood function for Meixner L´evy processes under high frequency sampling, *Statistics and Probability Letters*, **81**(4) 460469.
- Kawai, R., Masuda, H. (2013) Local asymptotic normality for normal inverse Gaussian L´evy processes with high-frequency sampling, *ESAIM: Probability and Statistics*, **17**, 13-32.

6. 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.

7. Sample path generation of COGARCH processes

Supervisor Dr. Ray Kawai

Project Description: COGARCH (continuous-time GARCH) processes play an increasingly important role in the representation of stationary time series with continuous time parameter, excellent blend of continuous-time stochastic processes and discrete-time time series. In this project, we go through their theoretical properties and construct methods for sample path generation. This project assumes strong background of both stochastic processes and time series analysis, along with strong numerical skills.

- Brockwell, P., Chadraa, E., Lindner, A. (2006) Continuous-time GARCH processes, *Annals of Applied Probability*, **16**(2) 790-826.
- Kawai, R. (2016) Sample path generation of L´evy-driven continuous time autoregressive moving average processes, *Methodology and Computing in Applied Probability*.

8. Advanced versions of stochastic approximation algorithms

Supervisor Dr. Ray Kawai

Project Description: In this project, we will explore recent advanced versions of stochastic approximation algorithms with convergence and error analysis. We examine a few versions, including a multi-step Richardson-Romberg extrapolation version and a multi-level Monte Carlo version. Strong numerical experiment skill is essential.

- Frikha, N., Huang, L. (2015) A multi-step Richardson-Romberg extrapolation method for stochastic approximation, *Stochastic Processes and their Applications*, **125**(11) 4066-4101.
- Frikha, N. (2016) Multi-level stochastic approximation algorithms, *Annals of Applied Probability*, **26**(2) 933-985.

9. 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 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 25K citations. After covering the basics of this important area we can choose to focus on one of several important extensions. For example, more recently it was noted that if the hypotheses naturally divide into groups then by performing separate FDR analysis in each group typically increases the number of discoveries. The problem is, in general control of the FDR is no longer guaranteed in such a setting. We can look into strategies that try to guarantee some control of the FDR while taking advantage of the group structure. Alternatively, we can for example study the question of the level of confidence we have in our FDR estimation.

10. 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.

11. Statistical analysis of biological sequence similarity search tool

Supervisor A/Prof. Uri Keich

Project Description: Since the advent of new genomic sequencing techniques our sequencing capacity grows more rapidly than our computing power. One promising approach to address this problem has been the introduction of tools for compressed genomics that take advantage of the genome level similarities between species. In particular versions of compressed BLAST, an extremely popular tool for searching genomic databases has been proposed. However, the statistical analysis BLAST utilizes relies on a model which is incompatible with the same database redundancy that the novel compressed BLAST is relying on. In this project we will explore alternative idea for assigning significance for compressed genomics tools.

12. 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.

13. 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.

14. Regularization methods, does the sign of correlation coefficients matter?

Supervisor: A/Prof. Samuel Müller

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.

15. Learning from changing slopes to identify the better classify

Supervisor: A/Prof. Samuel Müller

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.

16. Efficient calculation of data depth

Supervisor: A/Prof. Samuel Müller

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 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.

17. Measuring instability of model selection methods

Supervisor: A/Prof. Samuel Müller

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.

18. Robust monotone curve estimation

Supervisor: A/Prof. Samuel Müller

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).

19. Robust model selection criteria, specific examples and R package

Supervisor: A/Prof. Samuel Müller

Project Description: Müller 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.

20. Are uncorrelated variables more likely to be selected?

Supervisor: A/Prof. Samuel Müller

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.

19. **Model Selection for Generalised Additive Models**

Supervisor: Dr John Ormerod

Project description: Generalised additive models are a flexible class of models which extend linear regression models to incorporate both linear and nonlinear effects of covariates on the response. Model selection in this setting is more challenging in this setting because (i) not only do the correct variables need to be selected, but also (ii) whether the selected variable has a linear or a nonlinear effect on the response and (iii) the roughness/smoothness of the nonlinear effects. Extensions to non-Gaussian responses, interactions, heteroscedasticity, spatially adaptive smoothing and missing covariate values may also be considered. Data mining applications are envisaged.

21. **Bayesian approaches to Differential Distribution**

Supervisor: Dr John Ormerod

Project description: The distribution of genes is potentially informative when trying to distinguish between health samples and diseased samples. Traditionally this has been performed via a hypothesis testing approach which tests for differences in the mean gene expression levels between healthy samples and diseased samples, which is called differential expression. In this project we will perform analogous Bayesian test for differences across the whole distribution of gene expression levels between two states. A multiple testing approach will be developed to take into account false discoveries. This work will be motivated by real gene expression data from melanoma patients where it is hoped that this new approach will be able to uncover new biomarkers for the disease.

22. **Mixtures of Survival Models**

Supervisor: Dr John Ormerod

Project description: Suppose we wish to estimate the survival time for a particular cancer with two subtypes, less aggressive and more aggressive, say. However, it is possible that we do not know what subtype of cancer a particular patient has. Inferring the cancer subtype from the survival times can be important in determining what medical treatment the patient receives. Mixture models can be used in such situations. In this project we will develop fast and effective methods for fitting mixtures of survival data.

23. **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^2_k 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:

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

References

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.

24. Classification and statistical networks.

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 modelling 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.

Reference: 11. 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.

25. Methods towards personalize medicine

Supervisor: Prof Jean Yang

Project description: Over the past decade, new and more powerful genomic 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 investigate whether it is possible to establish the patient or sample specific network based (matrix) by integrating public repository and gene expression data.

26. Is integrative-omics the new currency for solutions to complex diseases? Exploring correlations in high-dimensional data.

Supervisor: Prof Jean Yang

Project description: In the surge of large volumes of high-throughput biological data being generated, more researchers are looking to integrate data of different types to inform hypotheses. For example, in complex metabolic diseases such as T2D and obesity, it is crucial to interrogate multiple data types to gain a comprehensive picture of the system defects and may eventually lead to identification of T2D or obesity markers. In this project, we aim to apply multivariate statistical approaches to integrate the data and build better predictors from multiple data sources. In this project, we will explore ways of weighting the relatively sparse proteomics data with information borrowed from the transcriptomics data. This involves, exploring or developing methods to correlate multiple high dimensional datasets to identify common and differentiating patterns.

27. High dimensional data visualization and cluster computing.

Supervisor: Prof Jean Yang and Dr. Michael Stewart

Project description: High dimensional data refers to data anywhere from a few dozen to many thousands of dimensions and they are found in many disciplines including finance, medicine and biology. These include next generation sequencing or brain imaging data where a large number of measurements are produced simultaneously. We will examine some modern techniques that enable us to visualize high-dimensional data and enable to better formulate questions and refine modelling strategies. Furthermore, when the data becomes too large for a single computer, divide and recombine strategies are often required. Example of such platform includes Trelliscope within the Tessera computational environment. Trelliscope is backed by datadr, scales Trellis Display, allowing the analyst to break potentially very large data sets into many subsets, apply a visualization method to each subset, and then interactively sample, sort, and filter the panels of the display on various quantities of interest (Reference: http://tessera.io). In this project, we will examine how we can effectively visualize large dimensional data from functional MRI (fMRI) as well as its corresponding connectivity network (estimate inverse sparse covariance matrix) and potentially utilise it to guide modelling strategies.

28. Dimension reduction in resting state fMRI data

Supervisor: Prof Jean Yang and Dr. John Ormerod

Project description: Anatomical, functional and effective networks within the brain are currently being elucidated at fine temporal and spatial resolution using magnetic resonance imaging, via both functional MRI (fMRI). The concepts behind local region clustering such as superpixels are becoming increasingly popular for use in computer vision applications, data visualization and dimensional reduction strategies. This project involves exploring ideas and models for segmenting fMRI imaging data by borrowing information across multiple samples. Specific applications of this information sharing may be to improve the identification of interesting biologically features or improve sample classification in large p small n datasets.

29. Tests for publication bias, and their applicability to variance based effect sizes.

Supervisor: Dr. Alistair M Senior and P 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 behaviour 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 behaviour 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.

30. **Embryonic stem cell (ESC)-specific pathway identification and annotation using multi-layered omics data and statistical learning**

Supervisor: Dr. Pengyi Yang

Project description: Supervisor: While all cells from a given organism have the same DNA sequence that codes for the same genes, different cell types of that organism only have a subset of genes "turned on".

Genes are commonly annotated into pathways for summarising their collective effect in the biological systems. One of the main drawbacks in current pathway annotation is that they are NOT cell type-specific.

We propose to identify and curate cell type-specific pathways for embryonic stem cells (ESCs) using our multi-layered omics data. The key assumption is that genes within a pathway should have correlated expression profile changes when perturbed.

We have collected ESC differentiation data profiled in a time-course on both proteome and transcriptome levels. Following the above assumption, we aim to address the following points:

- Identify pathways that are regulated specifically in ESC differentiation. This can be done using clustering-based approach (http://www.ncbi.nlm.nih.gov/pubmed/26252020) where genes that exhibit similar temporal profiles will be clustered using e.g. k-means algorithms and pathways that are over-represented in each cluster can be identified.

- Curate ESC-specific pathways using statistical learning. A statistical learning model such as positive-unlabelled learning (http://www.ncbi.nlm.nih.gov/pubmed/26395771) can be used to predict "novel" genes that are not being previously associated with an ESC-specific pathway. The predictive model can also be used to remove genes that are non-specific to ESC in a pathway.

This project will expose honours student to the development and application of cutting-edge statistical learning methods to the state-of-the-art bio-molecular applications. It sits at the heart of interdisciplinary research.

Dr. Lamiae Azizi has asked that any students interested in doing an honours project with her to contact her directly in order to tailor a project for the student.

Dr. Uri Keich will be overseas in the weeks leading up to the deadline for honours. He has asked that students interested in his projects to email him to arrange for a Skype meeting.

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 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.