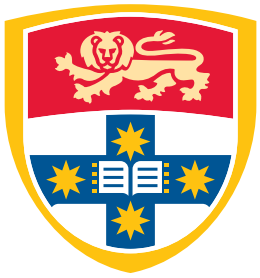
Honours in Statistics

Honours in Data Science

Detailed Guide for the 2022 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 application portal. The Faculty requirements which must be met include:

- qualifying for a degree in a major which is cognate to the proposed honours stream, that is, a major which provides a suitable background for the honours stream. Note that a major in Statistics, Data Science or Financial Mathematics & Statistics are inherently cognate to our honours program while in borderline cases the decision of whether a major is cognate is in the hands of the Honours coordinator and the faculty);
- having a WAM of at least 65;
- *securing the agreement of a supervisor.*

In addition, the School of Mathematics and Statistics requires that the student has a total of at least 18CP (for a Data Science major), or 24CP (for Statistics major) of relevant 3XXX or 4XXX courses in which

- the average mark of Advanced level courses is at least 65;
- the average mark of Mainstream level courses is at least 75

If you have a mix of advanced and mainstream courses, where some are above and some below the thresholds, if you are not sure which of your courses are relevant, or if your average is just on the wrong side of the threshold you can seek further advice from the relevant program's honours coordinator.

1.1 Formally...

The faculty offers three main Honours pathways and it can be confusing:

- Combined Bachelor of Science/Bachelor of Advanced Studies is an option if you commenced your studies after 2018 and it allows completing Honours as an embedded pathway in the final year of the program. Requires two majors.
- Standalone Bachelor of Advanced Studies (Honours) is the same as above, including the two-majors requirement, except technically this is an appended, standalone Honours year.
- The Bachelor of Science (Honours) is a standalone (appended) Honours requiring an additional year of study. It is for students who
  - are not on track to complete two majors in the Bachelor of Science, or
  - are external students with only one major, or
  - commenced before 2018 and did not choose to transfer to the new curriculum version of their degree.

  Note that forthcoming changes will remove the current restrictions that excluded the newly formed Honours programs of Data Science (DS) from BSc (Honours). In other words from 2022 you will be able to to do Honours in DS under the BSc (Honours) program.

1.2 It's important to note that:

- All acceptances into Honours (including in cases where the schools requirements are not met) 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.
- The Faculty of Science Honours **application deadline** (for Honours commencement in Semester 1, 2022) is 31 January 2022 and for Semester 2, 2022 it is 30 June 2022.

2 Structure of Honours

An honours year in the School of Mathematics and Statistics involves four 6CP courses (worth 50% of the final mark) and a project (worth 50%).

2.1 The honours project (50%)

The honours project centres around an essay/thesis consisting of roughly 60 pages[1] 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 scientific/statistical/mathematical content and its exposition, including the written english. The thesis is due at the end of your second semester, specifically at 5pm on Monday of week 13.

Toward the end of the second semester (Friday weeks 9-10), 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.

2.1.1 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.

2.2 Course work (50%)

The honours program in *statistics* specifies a couple of core courses as well as which combination of courses can be taken. The list of available courses can be found online, however please carefully read through the list of constraints outlined in the *stats* degrees structure document.

The newly created honours program in *data science* specifies different combinations of courses that can be taken including many courses offered by the school of IT. A list of courses that will be offered in 2022 is available online. However students should carefully consult the *data science*

[1] This page number is a very rough guideline and should not be taken as binding.

degree structure document which outlines the combinations of courses that can be taken for credit. Keep in mind that unfortunately some of the course codes do not match the more updated codes online, however the course names should match.

3 Important course work information for all students

3.1 Selecting your courses

Please make sure you **select your courses after consulting the Honours supervisor and the Honours coordinator!**

3.2 AMSI courses

Students are welcomed to check the courses offered in January at the AMSI Summer School and also courses available via the Advanced Collaborative Environment (ACE). These courses can possibly be taken for credit (by enrolling in the unit AMSI4001), but this can only be done in consultation the student's supervisor and with the approvals of the specific honours coordinator as well as the School's Honours coordinator, Prof. Laurentiu Paunescu.

4 Program Administration

The Data Science Honours coordinator is

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

The Statistics as well as the Data Science Honours coordinator is

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

The Co-director of Teaching (Statistics & Data Science) is

Dr. Ellis Patrick,
Carslaw Building, Room 816, Phone 0402 159 424,
Email: ellis.patrick@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

5 Potential Supervisors and their Research Interests

See the individual staff member webpages for more detail about their research and their contact information.

5.1 The Statistics Group

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.

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

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

Doctor Ellis Patrick
Applied Statistics, Bioinformatics, Machine learning, Image analysis, Focus on Method Development for High-dimensional Biomedical Assays including High-Parameter Imaging Cytometry Data.

Associate Professor Shelton Peiris

Time Series Analysis, Estimating Functions and Applications, Statistics in Finance, Financial Econometrics, Hybrid ARIMA modelling and Applications.

Doctor Michael Stewart

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

Doctor Garth Tarr

Applied statistics, Robust Methods, Model Selection, Data Visualisation, Biometrics.

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.

Professor Jean Yang

Statistical Bioinformatics, applied Statistics, Analysis of multi-omics data, biomedical data science, single-cell data analytics, statistical learning in precision medicine.

Doctor Pengyi Yang

Machine learning, Deep learning, Statistical modelling, Single-cell omics, Multi-omics, Systems stem cell biology.

5.2 Members of the Pure and Applied Groups with Interest in Data Science

Professor Eduardo G. Altmann

complex networks, dynamical-systems modelling of social media, natural language processing, topic modelling.

Professor Georg Gottwald

Dynamical systems methods in data science, machine learning in climate science, neural networks, data assimilation.

Doctor Jonathan Spreer

Computational topology and geometry, Parameterised complexity theory, Mathematical software.

Professor Stephan Tillmann

Computational topology, Geometric structures on manifolds.

6 Honours courses in Statistics and Data Science

The following honours topics are expected to be on offer in 2022.

1. **STAT4021: Stochastic Processes and Applications (semester 1)**

   A stochastic process is a mathematical model of time-dependent random phenomena and is employed in numerous fields of application, including economics, finance, insurance, physics, biology, chemistry and computer science. In this unit you will rigorously establish the basic properties and limit theory of discrete-time Markov chains and branching processes and then, building on this foundation, derive key results for the Poisson process and continuous-time Markov chains, stopping times and martingales. You will learn about various illustrative examples throughout the unit to demonstrate how stochastic processes can be applied in modeling and analysing problems of practical interest, such as queuing, inventory, population, financial asset price dynamics and image processing. By completing this unit, you will develop a solid mathematical foundation in stochastic processes which will become the platform for further studies in advanced areas such as stochastic analysis, stochastic differential equations, stochastic control and financial mathematics.

2. **STAT4022: Linear and Mixed Models (semester 1)**

   Classical linear models are widely used in science, business, economics and technology. This unit will introduce the fundamental concepts of analysis of data from both observational studies and experimental designs using linear methods, together with concepts of collection of data and design of experiments. You will first consider linear models and regression methods with diagnostics for checking appropriateness of models, looking briefly at robust regression methods. Then you will consider the design and analysis of experiments considering notions of replication, randomization and ideas of factorial designs. Throughout the course you will use the R statistical package to give analyses and graphical displays. This unit includes material in STAT3022, but has an additional component on the mathematical techniques underlying applied linear models together with proofs of distribution theory based on vector space methods.

3. **STAT4023: Theory and Methods of Statistical Inference (semester 2)**

   In today's data-rich world, more and more people from diverse fields need to perform statistical analyses, and indeed there are more and more tools to do this becoming available. It is relatively easy to "point and click" and obtain some statistical analysis of your data. But how do you know if any particular analysis is indeed appropriate? Is there another procedure or workflow which would be more suitable? Is there such a thing as a "best possible" approach in a given situation? All of these questions (and more) are addressed in this unit. You will study the foundational core of modern statistical inference, including classical and cutting-edge theory and methods of mathematical statistics with a particular focus on various notions of optimality. The first part of the unit covers aspects of distribution theory which are applied in the second part which deals with optimal procedures in estimation and testing. The framework of statistical decision theory is used to unify many of the concepts that are introduced in this unit. You will rigorously prove key results and apply these to real-world problems in laboratory sessions. By completing this unit, you will develop the necessary skills to confidently choose the best statistical analysis to use in many situations.

4. **STAT4025: Time series (semester 1)**

   This unit will study basic concepts and methods of time series analysis applicable in many real world problems applicable in numerous fields, including economics, finance, insurance, physics, ecology, chemistry, computer science and engineering. This unit will investigate the basic methods of modelling and analyzing of time series data (ie. Data containing serially dependence structure). This can be achieved through learning standard time series procedures on identification of components, autocorrelations, partial autocorrelations and their sampling properties. After setting up these basics, students will learn the theory of stationary univariate time series models including ARMA, ARIMA and SARIMA and their properties. Then the identification, estimation, diagnostic model checking, decision making and forecasting methods based on these models will be developed with applications. The spectral theory of time series, estimation of spectra using periodogram and consistent estimation of spectra using lag-windows will be studied in detail. Further, the methods of analyzing long memory and time series and heteroscedastic time series models including ARCH, GARCH, ACD, SCD and SV models from financial econometrics and the analysis of vector ARIMA models will be developed with applications. By completing this unit, students will develop the essential basis for further studies, such as financial econometrics and financial time series. The skills gain through this unit of study will form a strong foundation to work in a financial industry or in a related research organization.

5. **STAT4026: Statistical consulting (semester 1)**

   In our ever-changing world, we are facing a new data-driven era where the capability to efficiently combine and analyse large data collections is essential for informed decision making in business and government, and for scientific research. Statistics and data analytics consulting provide an important framework for many individuals to seek assistant with statistics and data-driven problems. This unit of study will provide students with an opportunity to gain real-life experience in statistical consulting or work with collaborative (interdisciplinary) research. In this unit, you will have an opportunity to have practical experience in a consultation setting with real clients. You will also apply your statistical knowledge in a diverse collection of consulting projects while learning project and time management skills. In this unit you will need to identify and place the client's problem into an analytical framework, provide a solution within a given time frame and communicate your findings back to the client. All such skills are highly valued by employers. This unit will foster the expertise needed to work in a statistical consulting firm or data analytical team which will be essential for data-driven professional and research pathways in the future.

6. **STAT4027: Advanced Statistical Modelling (semester 2)**

   Applied Statistics fundamentally brings statistical learning to the wider world. Some data sets are complex due to the nature of their responses or predictors or have high dimensionality. These types of data pose theoretical, methodological and computational challenges that require knowledge of advanced modelling techniques, estimation methodologies and model selection skills. In this unit you will investigate contemporary model building, estimation and selection approaches for linear and generalised linear regression models. You will learn about two scenarios in model building: when an extensive search of the model space is possible; and when the dimension is large and either stepwise algorithms or regularisation

techniques have to be employed to identify good models. These particular data analysis skills have been foundational in developing modern ideas about science, medicine, economics and society and in the development of new technology and should be in the toolkit of all applied statisticians. This unit will provide you with a strong foundation of critical thinking about statistical modelling and technology and give you the opportunity to engage with applications of these methods across a wide scope of applications and for research or further study.

7. **STAT4028: Probability and Mathematical Statistics (semester 1)**

   Probability Theory lays the theoretical foundations that underpin the models we use when analysing phenomena that involve chance. This unit introduces the students to modern probability theory and applies it to problems in mathematical statistics. You will be introduced to the fundamental concept of a measure as a generalisation of the notion of length and Lebesgue integration which is a generalisation of the Riemann integral. This theory provides a powerful unifying structure that bring together both the theory of discrete random variables and the theory of continuous random variables that were introduce to earlier in your studies. You will see how measure theory is used to put other important probabilistic ideas into a rigorous mathematical framework. These include various notions of convergence of random variables, 0-1 laws, and the characteristic function. You will then synthesise all these concepts to establish the Central Limit Theorem and also verify important results in Mathematical Statistics. These involve exponential families, efficient estimation, large-sample testing and Bayesian methods. Finally you will verify important convergence properties of the expectation-maximisation (EM) algorithm. By doing this unit you will become familiar with many of the theoretical building blocks that are required for any in-depth study in probability or mathematical statistics.

8. **STAT4528: Probability and Martingale Theory (semester 1)**

   Probability Theory lays the theoretical foundations that underpin the models we use when analysing phenomena that involve chance. This unit introduces the students to modern probability theory (based on measure theory) that was developed by Andrey Kolmogorov. You will be introduced to the fundamental concept of a measure as a generalisation of the notion of length and Lebesgue integration which is a generalisation of the Riemann integral. This theory provides a powerful unifying structure that brings together both the theory of discrete random variables and the theory of continuous random variables that were introduced earlier in your studies. You will see how measure theory is used to put other important probabilistic ideas into a rigorous mathematical framework. These include various notions of convergence of random variables, 0-1 laws, conditional expectation, and the characteristic function. You will then synthesise all these concepts to establish the Central Limit Theorem and to thoroughly study discrete-time martingales. Originally used to model betting strategies, martingales are a powerful generalisation of random walks that allow us to prove fundamental results such as the Strong Law of Large Numbers or analyse problems such as the gambler's ruin. By doing this unit you will become familiar with many of the theoretical building blocks that are required for any in-depth study in probability, stochastic systems or financial mathematics.

9. **STAT5610: Advanced Inference (semester 2)**

The great power of the discipline of Statistics is the possibility to make inferences concerning a large population based on optimally learning from increasingly large and complex data. Critical to successful inference is a deep understanding of the theory when the number of samples and the number of observed features is large and require complex statistical methods to be analysed correctly. In this unit you will learn how to integrate concepts from a diverse suite of specialities in mathematics and statistics such as optimisation, functional approximations and complex analysis to make inferences for highly complicated data. In particular, this unit explores advanced topics in statistical methodology examining both theoretical foundations and details of implementation to applications. The unit is made up of 3 distinct modules. These include (but are not restricted to) Asymptotic theory for statistics and econometrics, Theory and algorithms for statistical learning with big data, and Introduction to optimal semiparametric optimality.

10. **DATA5441: Networks and High-dimensional Inference (semester 1)**

In our interconnected world, networks are an increasingly important representation of datasets and systems. This unit will investigate how this network approach to problems can be pursued through the combination of mathematical models and datasets. You will learn different mathematical models of networks and understand how these models explain non-intuitive phenomena, such as the small world phenomenon (short paths between nodes despite clustering), the friendship paradox (our friends typically have more friends than we have), and the sudden appearance of epidemic-like processes spreading through networks. You will learn computational techniques needed to infer information about the mathematical models from data and, finally, you will learn how to combine mathematical models, computational techniques, and real-world data to draw conclusions about problems. More generally, network data is a paradigm for high-dimensional interdependent data, the typical problem in data science. By doing this unit you will develop computational and mathematical skills of wide applicability in studies of networks, data science, complex systems, and statistical physics.

7 Project

7.1 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.

Lists of suggested project topics for both statistics as well as data science Honours are provided in Section 7.2 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?

7.2 Proposed project topics

The projects are divided into two categories: Data Science and Statistics. If you are a Data Science student but you prefer a project that is listed under Statistics or the other way around then feel free to talk the relevant supervisor about it. For additional projects see the Section 13 at the end of this document.

7.2.1 Proposed project topics in Data Science

1. **Complex networks and social-media data**
   Supervisor: Prof. Eduardo G. Altmann

   Project description: Please contact me if you are interested in projects combining data analysis and mechanistic models of complex networks and social-media data (time series and natural language processing).

2. **Scaling laws in urban data**
   Supervisor: Prof. Eduardo G. Altmann

   Project description: Data science research often involves finding statistical regularities universally valid in different scenarios. An example is the proposal of non-linear scaling laws in the relationship between the population of cities and data of socio-economical activities in these cities (e.g., patent production or CO2 emission). These observations have triggered the proposal of a variety of models and statistical methods for data anlaysis [1,2]. The goal of this project is to investigate the existence of such scaling laws in datasets from Australia and the development of inference methods that are able to estimate multiple parameters of generative models that take into account the spatial nature of the datasess.

   1 Bettencourt Lus M. A., Lobo Jos, Helbing Dirk, Kuhnert C., and West Geoffrey B. Growth, innovation, scaling, and the pace of life in cities. Proceedings of the National Academy of Sciences, 104(17):73017306, 4 2007.

   2 E. G. Altmann, "Spatial interactions in urban scaling laws", [PLOS ONE 15, e0243390 (2020)]

3. **Predictive models for auto-insurance using big telematics data**
   Supervisor: A/Prof. Jennifer Chan

   Project description: This project will apply some machine learning techniques to select informative variables to predict insurance claim frequency using telematics data. Traditional auto-insurance policies use drivers' demographic information such as age, gender, address, social economic status to calculate auto-insurance premiums. However these variables do not really reflect drivers' driving risky. While more and more automobiles have installed telematics which monitor driving behaviour, big but noisy telematics data are now available to inform better auto-insurance premium calculation. As predictive variables are expected to differ between safe and risky drivers, this project considers Poisson mixture model for claim frequencies so that each mixture group, for safe or risky drivers, can select a different set of driving behaviour variables. Some machine learning techniques such as lasso and random forest will be used to select informative variables that minimise certain loss functions such as area under curve. Some procedures such as resampling and parameter tuning will also be considered to avoid overfitting.

4. **Projects will be offered by discussion (for Honours in Data Science)**
   Supervisor: Prof. Georg Gottwald

   Project Description: Please contact Georg directly.

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. **Data-intensive science to understand the molecular aetiology of disease.**
   Supervisor: Dr. Ellis Patrick

   Project description: Biotechnological advances have made it possible to monitor the expression levels of thousands of genes and proteins simultaneously promising exciting, groundbreaking discoveries in complex diseases. This project will focus on the application and/or development of statistical and machine learning methodology to analyse a high-dimensional biomedical experiment. My group works on projects spanning multiple diseases including melanoma, acute myeloid leukemia, Alzheimer's disease, multiple sclerosis and HIV. We also work with various high-throughput technologies including single-cell RNA-Seq, SWATH-MS, flow cytometry, CyTOF, CODEX imaging and imaging mass cytometry.

7. **Constructing network-based biomarkers which are critical for disease prognosis.**
   Supervisor: Dr. Ellis Patrick and Prof Jean Yang

   Project description: With the advancements of single-cell sequencing technology, we are now able to explore cell identity at a resolution that was previously out of reach. From a statistical viewpoint, this data-type is very interesting as it has unusual data-structure which means there is an incredible amount of statistical methods development that needs to be completed to make full use of this amazing technology. We propose a project below that explores state-of-the-art clustering and classification approaches that, while generalizable to other contexts, will be applied to tangible and translationally relevant biomedical questions. Many classical approaches in classification are primarily based on single features that exhibit effect size difference between classes. Recently, we have demonstrated that approaches which use network-based features can be used to classify alternate subsets of patients as compared to those that use single-features. Building on our previous experience identifying network-based biomarkers (classifiers of disease) we will instead use cell-type specific networks generated from single-cell sequencing data. This process will allow us to construct network biomarkers that are specific for a cell-type of interest, are capable of assigning a score to a single individual

and can be integrated with classification approaches such as DLDA, SVM, LASSO and Random Forests. Bootstrap and resampling will be used to ensure stability and robustness of identified features.

8. **ARFIMA-ANN Hybrid Model for Time Series Forecasting**
Supervisor: A/Prof. Shelton Peiris

Project description: Autoregressive Fractionally Integrated Moving Average (ARFIMA) has been successfully applied in modelling and forecasting economic time series with long memory. It is known that Artificial Neural Network (ANN) approach can be used to capture additional complex nonlinear economic relationships with many unknown of patterns. This project proposes a hybrid model, which is distinctive in integrating the advantages of ARFIMA and ANN in modelling and the analysis of linear and nonlinear components of a time series with long memory.

9. **Comparing classifiers on publicly available datasets**
Supervisor: Dr Michael Stewart

Project description: Simple mixture models have been used as models for test statistics and p-values in large-scale multiple testing problems. Higher criticism was originally developed in Donoho and Jin (2004) as a tool for such problems, and was shown to work well at detecting certain types of mixtures. It has since been adapted as a tool for feature selection, functioning as a thresholding device to decide which p-values correspond to (potentially) useful features.

Dettling (2004) compared various popular classification methods on some widely-used publicly available datasets. Donoho and Jin (2008) extended this comparison to use a simple classification method based on higher cricism thresholding (see also Donoho (2017) for discussion) which showed that despite its simplicity it worked very well or even better than other much more complicated popular methods.

The purpose of this project is to develop similar classification methods based on other mixture detection methods and compare their performance to that of higher criticism-based and other classifiers on the same, and possibly other publicly available datasets. It will involve some theoretical work and also substantial computing.

References

M. Dettling. Bagboosting for tumor classification with gene expression data *Bioinformatics*, 20(18):3583–3593, 2004.

D. Donoho and J. Jin. Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.*, 32(3):962–994, 2004.

D. Donoho and J. Jin. Higher criticism thresholding: optimal feature selection when useful features are rare and weak. *Proc. Natl Acad. Sci. USA*, 105:14790–14795, 2008.

D. Donoho. 50 years of data science. *J. Comput. Graph. Statist.*, 26(4):745–766, 2017.

10. **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.

11. **Stable feature selection in high dimensional models**
Supervisor: Dr. Garth Tarr

Project description: Modern feature selection methods can be applied in situations where the number of variables is much greater than the number of observations. An important consideration is the stability of the set of selected features. This project will investigate feature selection stability in high dimensional regression models and consider ways of visualising and presenting this information to researchers to better inform their model selection decisions.

12. **Topic: Topological data analysis (for Honours in Data Science)**
Supervisor: Dr. Jonathan Spreer or Prof. Stephan Tillmann

Project description: The main motivation of topological data analysis is to study the shape of data in a rigorous and meaningful way. The theoretical underpinnings come from pure mathematics, in particular algebraic topology, resulting in the main tool called persistent homology. Possible projects include: mathematical foundation of the main tools; the application of persistent homology to new types of data sets; quantitative analysis to obtain geometric information from topology.

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

Project description: Over the past decade, new and more powerful -omic tools have been applied to studying complex diseases 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 a statistical method that delivers better prediction toward healthy aging by identifying a risk prediction framework that is interpretable. Students have a choice of creating a risk prediction method using one of the four case studies described below:

- *Case study I - Melanoma:* The Melanoma Institute of Australia (MIA) Stage III data collection is a unique set of multi-layered omics data (Mann et al. 2013; Jayawardana et al. 2016) with measured gene, protein, and microRNA expression and linked clinical and mutation data for subjects with Stage III lymph node metastases.

- *Case study II Cardiovascular disease:* This case study is in collaboration with Prof Gemma Figtrees team who will provide access to the [BioHEART] dataset. This is a unique large scale multi-omics data with over 1,000 samples in lipidomics, proteomics and metabolomics and 250 in CyTOF already generated to allow evaluation of scalable

models.

- *Case study III Infectious disease COVID19:* There currently exists over 30 public COVID19 multi-omics datasets, including ten single-cell RNA-seq datasets with over 300 individuals and this work will involve a newly established collaboration with Laboratory of Data Discovery for health, $D^2 4H$) on global health protection.

- *Case study IV Parkinson disease:* This case study is in collaboration with Prof Carolyn Sue (Kolling Institute). She will provide access to nearly 200 individual microbiome data with matched diet and nutrition information (with possible matched metabolomics data) to study the longitudinal impact on Parkinson disease (PD). Additional public microbiome and omics data on PD have also been curated.

14. **Machine learning for kidney allocation**
Supervisor: Prof Jean Yang

Project description: Kidney transplantation offers improved survival and quality of life for many patients with kidney failure compared to being on dialysis. Non-invasive -omics biomarkers may predict adverse events such as acute rejection after kidney transplantation and may be preferable to existing methods because of superior accuracy, timeliness and convenience. A recent study has shown that a set of gene signatures for acute rejection derived from a single study do not appear to provide adequate prediction in an independent cohort of transplant recipients. This project aims to develop an approach to integrate multiple gene signatures sets that improve the prediction performance of these markers and ensure the biomarkers behave consistently across multiple data platforms. This project will involve the evaluation of the new approach and communicate this information via an interactive web interface. The project will contribute to the widespread use of omics signatures in clinical practice.

15. **Modelling cell types using single-cell omics data**
Supervisor: Dr. Pengyi Yang

Project description: Single-cell transcriptomic profiling using RNA-sequencing (scRNA-seq) is becoming a widespread technique for studying development, cancers, and tissue regeneration. Cell type identification and modelling is one the most critical aspects in single-cell transcriptomic data analysis. While various machine learning approach (e.g. clustering, classification) have been used for this task, most of them do not provide useful insight as what genes and their networks is governing each cell type.

This project aims to develop computational methods that can identify key genes and networks that underly cell types. We will explore various deep learning techniques such as autoencoder that may use such information to predict cell types as well as infer cell trajectories during development. This will be a great opportunity for you to learning state-of-the-art machine learning methods and their applications to cutting-edge biomedical research with potential to achieve high impact in a range of translational research (e.g. development, cancers, and regenerative medicine).

16. **Identification of differential abundance in single-cell omics data**
Supervisor: Dr. Pengyi Yang and Dr. Ellis Patrick

Project description: Single-cell transcriptomic profiling using RNA-sequencing (scRNA-seq) is becoming a widespread technique for studying development, cancers, and tissue regeneration. Following cell type identification from single-cell RNA-sequencing (scRNA-seq) data, a key step is to identify differential abundance of cell types and differential gene expressions. There is a lack of statistical methodology for dealing this task taking into consideration of biological variabilities and experimental batch effects.

This project aims to develop computational methods that can detect cell types and gene expressions that are different within and between treatments. This will be a great opportunity for you to learning state-of-the-art statistical methods and their applications to cutting-edge omics data that has potential to translate into treatment and diagnosis for diseases.

7.2.2 Proposed project topics in Statistics

1. **Modelling covariance matrix time series using Wishart distribution**
   Supervisor: A/Prof. Jennifer Chan

   Project description: This project will investigate the modelling strategies for the time series of observed covariance matrices. Recent studies have considered Wishart and matrix-F distributions. The mean matrix of the distribution can be modelled with different persistence, cross persistence, and leverage effects. The models can be implemented in the Bayesian approach via some Bayesian softwares such as Stan (in R). We will apply the models to different stock market indices including cryptocurrencies and investigate how the variances and covariances change during the pandemic period.

2. **Impact of COVID 19 pandemic on cryptocurrency market using time series models with variance gamma distribution**
   Supervisor: A/Prof. Jennifer Chan

   Project description: This project will investigate properties of high frequency cryptocurrency returns 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 softwares such as stan (in R). We will consider long memory, stochastic volatility and leverage effect modelling and investigate how these features change during the pandemic period. Currently, there are few studies which investigate the impact of COVID 19 pandemic on the cryptocurrency market and so this study will be pioneering and interesting.

3. **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.

4. **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.

5. **Bayesian Moment Propagation**
   Supervisor: Dr John Ormerod

   Project description: Approximate Bayesian inference is a rapidly growing area in statistics and machine learning where models are described probabilistically and analytic approximations are brought to bear to perform prediction and inference in a fast but approximate way. For large and complex problems they are sometimes the only method can fit models in a computationally feasible time. These methods have been successfully applied in areas such as Bioinformatics, computer vision, neuroscience, and deep learning. One prominent approach is to use variational Bayes (VB) which assumes approximate posterior independence between model parameters to dramatically simplify model fitting. However, this independence assumption often leads to underestimating posterior variances and has led some to judge that such methods are not appropriate for inference. Recently, John Ormerod and his PnD student Weichang Yu have developed a way to correct posterior variance estimates for VB called Bayesian Moment Propagation (BMP). However almost nothing is known about BMP method other than it performs much better than VB on toy problems. The project could explore the theoretical underpinnings, explore the method on well known models, or extend these ideas to improve the speed or accuracy of these methods. A student with this project will gain skills in statistical computing, multivariate calculus, and multivariate statistics.

6. **Skewed Posterior Approximations**
   Supervisor: Dr John Ormerod

   Project description: Many approximate Bayesian inference methods assume a particular parametric form to approximate a posterior distribution to. A multivariate Gaussian approximation is a convenient density for such approaches, but ignores skewness. A step away from Gaussian approximation is to wade into a vast number of different skewed distributions. This project will aim at developing improvements to Gaussian approximations via exploration of the use of derivative matching, moment matching, delta method, nonlinear least squares, and stochastic gradient descent approaches to get accurate, fast, skewed approximations to posterior densities. A student with this project will gain skills in statistical computing, multivariate calculus, and multivariate statistics.

7. **Identifying changes in network structure to identify complex cellular interactions.**
   Supervisor: Dr. Ellis Patrick

Project Description: You will develop a novel network based hypothesis testing framework to detect if cells are collocating in high-dimensional cellular imaging data. This framework will inherently overcome some complications that arise in concordance based tests due to image noise and tissue inhomogeneity while also identifying the relationships that are most descriptive of the biology. The Pearson correlation coefficient approach (Manders et al. 1992) is the simplest and hence most widely used method for assessing cell-type colocalisation. We will generalise the Pearson correlation coefficient and Manders overlap coefficient methods for use with multiple markers by using partial correlation matrices, an approach I have applied to gene expression datasets (Patrick et al. 2017) for decomposing gene regulatory networks. By conceptualising colocalisation in terms of partial correlation matrices, you will test for colocalisation and changes in colocalisation in three ways:

- You will use a sparse graphical lasso to identify cell-type markers that are colocalised accounting for the behaviour of all other markers. Following the sparsity constraints you will use post-selective inference for Gaussian graphical models (GSell et al. 2013) to assign significance to each cell-cell interaction.
- Next, you will adapt a two-sample network inference approach typically used for brain connectivity analysis (Xia et al. 2017) to detect if colocalisation between two-cells, after accounting for the interactions between all other cells, is changing.
- Finally, two-sample network inference methods (Ghoshdastidar et al. 2018) can be adjusted to detect global changes in colocalisation between two conditions. This will produce a novel hypothesis testing framework to detect if whole systems of cells are interacting in distinct ways under different conditions.

8. **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).

9. **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.

10. **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:

- 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.

11. **Stable feature selection with marginality constraints**
Supervisor: Dr. Garth Tarr

Project description: In a model building process, it is often of interest to consider polynomial terms and interaction effects. This project will develop methods to evaluate the stability of model selection procedures that respect the marginality principle. The gold standard exhaustive search will be compared to stepwise methods and L1 regularised approaches for generalised linear models. On the application side, we have data from engineering and agriculture that can be explored.

12. **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.

13. **Threshold effects in nonlinear cointegrating regression**
Supervisor: Prof. Qiying Wang

Project description: There are extensive researches on estimation and inference theory of nonlinear regression models with nonstationary time series. However, it remains a challenging open research direction to investigate threshold effects in nonlinear cointegrating regression. In this project, we aim to develop estimation and inference theory in the threshold models with nonstationary time series. We will study various tests including Wald, Studentized t and the quasi-likelihood ratio for a threshold effect, investigate the asymptotics of related estimators and construct tests for model diagnostic checking.

14. **Weighted nonlinear cointegrating regression**
Supervisor: Prof. Qiying Wang

Project description: It is well-known that nonstandard asymptotic behaviour appears in nonlinear (linear) cointegrating regression. A fundamental issue raised in such a regression model with nonstatioanry time series is that the limiting distribution of least squares (LS) often depends on various nuisance parameters and/or such a limit result is cumbersome to be used in the relevant asymptotic inferences. In this project, we aim to develop new estimation theory in nonlinear cointegrating regression so that the limit distribution of suggested estimator is standard normal. This project involves deep classical probability knowledge. The interest in theoretical work is essential.

15. **Mini-batch Gibbs sampling for large-scale inference**
Supervisor: Dr. Rachel Wang

Project description: Large-scale datasets have given rise to complex models with a large number of parameters and intricate dependency structure. As a result, developing scalable algorithms with theoretical guarantees has become one of the central goals of modern day machine learning. Existing algorithms can be roughly divided into two classes: those that are based on optimisation and those that perform Bayesian inference. Since their inception, Markov chain Monte Carlo (MCMC) algorithms have been the main workhorse of Bayesian computation. However, compared to their counterparts in stochastic optimisation, standard MCMC methods do not meet the scalability requirement. Remedies have been proposed for the Metropolis-Hasting algorithm and involve making use of mini-batches of data, reminiscent of stochastic gradient descent. On the other hand, similar development for Gibbs sampling, another important class of MCMC methods, remain very nascent with the exception of [1]. This project will involve analysing the theoretical properties of a new mini-batch Gibbs algorithm and benchmarking its performance on standard models. Further applications can include structural estimation in graphical models and segmentation problems in image processing.

- De Sa, Christopher, Vincent Chen, and Wing Wong. "Minibatch Gibbs Sampling on Large Graphical Models." ICML (2018).

16. **Single cell data, network features and precision medicine**
Supervisor: Prof Jean Yang
Project description: This recent single-cell innovation generates thousands or even millions of cells in a single experiment amounting to a data revolution in single-cell biology. It poses unique statistical problems in scale and complexity. This project will develop an approach using single-cell data to identify phenotype-guided network features for different sub-populations. Interpreting complex single-cell data from a highly heterogeneous cell population remains a challenge as most existing single-cell approaches focus on cell type identification that cannot easily or directly link with specific disease outcomes. This project will initially examine methods that will use aggregate (bulk) measurement to identify cell subpopulations from single-cell data that most highly correlate with a given outcome and later expand it to identify cell interaction networks that are driven by a given outcome. The potential extension will be to examine or develop network classifiers to predict outcomes accurately and identify informative network features that involved joint analysis of multiple networks and network classification. In particular, we will identify meaningful and predictive network features by developing network classifiers that respect network structure, without reducing networks to global summary measures or treating them as a vector of edge weights.

8 Assessment

8.1 The honours grade

The student's honours grade is based on the average mark achieved by each student, over the 4 courses and the project. Courses account for 50% of the assessment and the project for the remaining 50%.

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.

8.2 The coursework mark

Students are required to attend 4 courses of 6CP during the academic year and the coursework mark is a simple average of the courses they took.

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%.

8.3 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 average 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?

8.4 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.

9 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.

10 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 internet, 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 School's social events.
- Class representative at School meetings.

11 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.

12 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.

13 Additional proposed project topics

1. **Predictability of epidemic-spreading models**
   Supervisor: Prof. Eduardo G. Altmann

   Project Description: The aim of this project is to quantify in which extent the spreading of an epidemic can be forecasted in advance. Prediction of the spreading is limited due to the combination of random fluctuations, unknown information, and the non-linear character of the underlying dynamics. The focus of this project will be on predicting the peak of an infection [1,2]. It will involve analytical and numerical investigations of ODE models and data analysis of time series of number of infections in different geographical areas.

   1 M. Castro, S. Ares, J. A. Cuesta, S. Manrubia, The turning point and end of an expanding epidemic cannot be precisely forecast. Proc. Natl. Acad. Sci. U.S.A. 117, 2619026196 (2020).

   2 Predicting an epidemic trajectory is difficult, Claus O. Wilke and Carl T. Bergstrom PNAS November 17, 2020 117 (46) 28549-28551;

2. **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.

3. **Modelling Single Cell Data with Variational Bayes**
   Supervisors: Dr. John Ormerod & Prof. Jean Yang

   Single-cell RNA sequencing (RNA-seq) data promises further biological insights that cannot be uncovered with individual datasets. Currently, for the most part simple models are entertained to model this data: Zero-inflated Poisson, normal mixture models, and latent factor models have been used. In this project we would aim to augment these models in order to either (a) cluster cells with similar gene expression profiles, or (b) perform multiclass classification using single cell data from multiple individuals. To achieve this we would use Bayesian modelling techniques, and fit the resulting model using variational Bayes. A student with this project will gain skills in statistical computing, Bayesian modelling, Bioinformatics, and approximate Bayesian inference.

4. **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.

5. **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

6. Testing for nonlinear cointergation
Supervisor: 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.

7. Nonlinear cointegrating regression with latent variables
Supervisor: Prof. Qiying Wang

Project description: Using the estimation theory currently developed in nonlinear regressionwith 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, thetraded Volatility index (VIX), and other derivatives

8. **Trans-omic data integration using statistical models**
Supervisor: Dr. Pengyi Yang

Project description: A major initiative in our group is to integrate trans-omics datasets generated by stateoftheart mass spectrometer (MS) and next generation sequencer (NGS) from various cell systems. We have now profiled various stem/progenitor cell differentiation processes using a combination of MS and NGS and have generated large-scale trans-omics datasets in these cell systems (see https://doi.org/10.1016/j.cels.2019.03.012). These data provide exciting research direction where we hypothesise that data integration across multiple omic layers is the key for comprehensive understanding of the underlying biological systems.

The aim of this project is to develop computational methods for integrating multiple omic data. Specifically, you will be learning and using unsupervised (e.g. clustering, PCA) and supervised (e.g. classification) machine learning techniques for integrating and making sense trans-omics data that capturing the dynamics of stem and progenitor cell differentiation. Knowledge discovered from this project will translate into exciting biological finding and shed light on complex diseases.

9. **Deep learning for reconstructing signalling networks**
   Supervisor: Dr. Pengyi Yang

   Project description: Signalling, such as protein phosphorylation, is a major mechanism for cells to pass extracellular signals to transcriptional and translational instructions in response to cellular micro-environment. The reconstruction of signalling networks is crucial for understanding how this layer of regulation is orchestrated. Using state-of-the-art mass spectrometry, we have profiled the global phosphoproteomes in pluripotent and unipotent stem/progenitor cells.

   This project aims to develop deep learning methods that are capable of extracting dynamic information embedded in the phosphoproteome data for predicting novel substrates of kinases and subsequently reconstruct the signalling networks. We have previously explored the traditional learning approaches (https://doi.org/10.1093/bioinformatics/btv550). The use of deep learning techniques will alleviate the difficulty in data feature engineering and allowing diverse source of information to be incorporated. You will learn from our top postgraduates (e.g. Thomas Geddes) on how to develop deep learning models using a combination of programming techniques including TensorFlow, PyTorch, and Keras. For taking this project, you will need to have experience with at least one programming language and understand the basics of machine learning.