Honours in Statistics (old) (new) and

Honours in Data Science

Detailed Guide for the 2020 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:

- For *old corriculum* students they include:
  - qualifying for the pass degree with a relevant major
  - having a SCIWAM of at least 65
  - securing the agreement of a supervisor
- For *new corriculum* students they include:
  - qualifying for the pass degree with two majors one of which should be cognate to the proposed honours stream a major which provides a suitable background for the honours stream; note that a major in Statistics, Data Science or Financial Mathematics and 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

All acceptances into Honours in Statistics or Data Science 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, 2020) is 30 November 2019.

2 Structure of Honours *(for old curriculum students doing Honours in statistics)*

An honours year in Statistics involves four 6CP courses (worth 60% of the final mark) and a project (worth 40%).

2.1 The honours project (40%)

The honours project centres around an essay/thesis consisting of 50-60 pages written on a particular topic from your chosen area. It need not contain original research (although it might) but it should clearly demonstrate that you have understood and mastered the material. The assessment of the honours thesis is based on the mathematical/statistical/scientific 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.

2.1.1 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 (60%)

There is some scope to mix and match courses from the three areas (subject to the approval of your supervisor and the Honours coordinator). Courses may require pre-requisites from Senior units of study: see the appropriate detailed guides for listings of these.

Full-time students will normally attend two 6CP lecture courses each Semester, for a total of four courses. 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 every stats Honours student needs to take either STAT4028 or STAT4528. A tentative list of the stats honour-level course offering is available in Section 7.

Subject to the approval of your supervisor and the Honours coordinator the following courses can also be taken for credit toward your required coursework:

- 4XXX or 5XXX 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)

3 Structure of Honours *(new curriculum students or Data Science Honours)*

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

3.1 The honours project (50%)

Please refer to the description above for the old honours program.

3.2 Course work (50%)

The new 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 tables attached at the end of this document. Unfortunately, the course codes do not always match the more updated codes online and in Section 7 below, however the course names should match.

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 2020 is available online. However students should carefully consult the table attached at the end of this 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.

4 Important course work information for all students

4.1 Selecting your courses

Regardless of whether you are a new or an old curriculum student make sure you **select your courses after consulting the Honours supervisor and the Honours coordinator!**

4.2 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). Note however that due to changes to our honours program these courses will not be counted toward the Honours 2020 course work requirements.

5 Program Administration

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 director of the Statistics teaching program is

Dr. Ellis Patrick,
Carslaw Building, Room 827, 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

6 Potential Supervisors and their Research Interests

6.1 The Statistics Group

Doctor Lamiae Azizi

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

Associate Professor Jennifer Chan

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

Professor Sally Cripps

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

Doctor Munir Hiabu

Nonparametric statistics, kernel smoothing, counting processes, structured models, age-period-cohort models, applications in actuarial science.

Doctor Ray Kawai

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

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

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

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, Time Dependent Categorical Data.

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, Statistical learning for biomedical data, Single cell data analytics.

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

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

Associate 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.

Associate Professor Stephan Tillmann

Computational topology,
Geometric structures on manifolds.

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

7 Honours courses in Statistics and Data Science

The following honours topics are expected to be on offer in 2020.

1. **STAT4021: Stochastic Processes and Applications**

   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**

   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**

   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**

   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**

   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**

   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**

   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**

   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**

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**

    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.

11. **DATA5710: Applied Statistics for Complex Data**

    With explosions in availability of computing power and facilities for gathering data in recent times, a key skill of any graduate is the ability to work with increasingly complex datasets. There may include, for example, data sets with multiple levels of observations gathered from diverse sources using a variety of methods. Being able to apply computational skills to implement appropriate software, as well as bringing to bear statistical expertise in the design of the accompanying algorithms are both vital when facing the challenge of analysing complicated data. This unit is made up of three distinct modules, each focusing on a different aspect of applications of statistical methods to complex data. These include (but are not restricted to) the development of a data product that interrogate large and complicated data structures; using sophisticated statistical methods to improve computational efficiency for large data sets or computationally intensive statistical methods; and the analysis of categorical ordinal data. Across all modules you will develop expertise in areas of statistical methodology, statistical analysis as well as computational statistics. Additional modules may be delivered, depending on the areas of expertise of available staff and distinguished visitors.

8 Project

8.1 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 8.2 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?

8.2 Proposed project topics

Most of the following projects are offered by members of the statistics group and the large majority of those will be appropriate for Honours in statistics as we as in data science. In addition, some projects are offered specifically for the data science honours by other members of the school. For additional projects see the Section 14 at the end of this document.

1. **Extracting information from written documents using stochastic block models (for Honours in Data Science)**
   Supervisor: A/ Prof. Eduardo G. Altmann

   Project description: One of the main challenges in Data Science is how to extract useful information from unstructured texts. Topic models are a popular machine-learning approach that infers the latent topical structure of a collection of documents. We have recently shown how a family of random graph models, known as Stochastic Block Models, can be adapted to perform topic modelling [see reference below]. The goal of this project is to incorporate into these models additional information that is usually available on the collection of written documents. For instance, documents are often classified by labels (e.g., keywords or scientific areas) or are linked with each other forming a graph (e.g., scientific papers that cite each other or webpages with hyperlinks). We will investigate how these additional information, together with the text of the documents, can be used to infer latent topical structures and to cluster the documents in groups.

   Reference: M. Gerlach, T. P. Peixoto, and E. G. Altmann, A network approach to topic models, Sci. Adv. 4, eaaq1360 (2018)

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

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

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

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

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

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

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

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

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

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

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

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

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

   Project Description: Please contact Georg directly.

8. **Least squares estimation for time-dependent coefficients in Cox's regression models**
   Supervisor: Dr. Munir Hiabu

   Cox's proportional hazard model is the most widely used model in biostatistics. Cai & Sun (2003) consider an extension with time-dependent coefficients, hence relaxing the proportionality assumption. This allows, e.g., treatment effects to vary over time. The model for the counting process intensity is

$$\lambda(t|\mathbf{x}) = \lambda_0(t) \exp\left\{ \sum_{j=1}^{p} a_j(t) x_j \right\}.$$

   The Cai & Sun (2003) estimator of $(a_j(\cdot))$ is based on the partial local likelihood idea. In this project we want to investigate whether the following hypothesis they make about their estimator is true:

   > *"One might expect a local weighted least-squares method and a local likelihood method to be asymptotically equivalent for the Cox model with time-dependent coefficients, provided that the local weighted least-squares methods use the functional structure of the model."*

Chances are that there are especially differences when the covariates $X_j, (j = 1, \dots, p)$, are correlated. The main task in this project is to develop and implement (in R) a least squares estimator and compare it to existing estimators.

- Cai, Z., & Sun, Y. (2003). Local linear estimation for timedependent coefficients in Cox's regression models. *Scandinavian Journal of Statistics*, 30(1), 93-111.

9. **Granular Reserving**
Supervisor: Dr. Munir Hiabu

In recent years there is an exploding interest in granular reserving, i.e., the prediction of outstanding liabilities (in non-life insurance) based on individual data. Starting with the seminal work of Arjas and Norberg based on counting process theory, more recently there are also many machine learning techniques emerging on that subject. Interestingly in many cases, as can for instance also be seen in the most recent Individual Claim Development with Machine Learning report of the ASTIN working party (presented on the International Congress of Actuaries (ICA) 2018, Berlin), those novel techniques still struggle to significantly outperform the simple chain-ladder technique which based on aggregated claims amounts data. The aim of this project is to review and understand existing reserving techniques and possibly develop some new ideas.

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

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

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

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

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

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

12. **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.

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

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.

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

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.

16. **One in a million - Characterising a single cell's impact on disease with state-of-the-art high-throughput biomedical technologies.**
    Supervisor: Dr. Ellis Patrick

    Project description: Over the last 20 years, researchers have used technologies such as microarray, RNA-sequencing and mass spectrometry to search for molecular drivers of disease. In order to have enough power to observe a link between a disease and a gene or protein, bulk tissues of hundreds thousands of cells needed to be assayed. The University has invested millions of dollars purchasing various state-of-the-art technologies that can facilitate the molecular profiling of single cells. The biological and medical ramifications of these technologies are unparralled and have made it possible to investigate the impact of a single cell's impact on disease. As these technologies are new, the exciting scientific hypotheses that can be explored with them, as well as the analytic approaches needed to answer these, are still being actively developed. In this project we will apply or develop analytical techniques to characterise /single cells using technologies such as single-cell RNA-Seq, high-parameter flow-cytometry, CyTOF and hyperion imaging.

17. **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.

18. **Vector Autoregressive Fractionally Integrated Moving Average (VARFIMA) Pro-**

cesses 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).

19. **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.

20. **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.

21. **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.

22. **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.

23. **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.

24. **Topic: Topological data analysis (for Honours in Data Science)**
Supervisor: A/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.

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

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

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

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

27. **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).

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

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

29. **Cell type specific network from single cell data**
    Supervisor: Prof Jean Yang and Dr. Kitty Lo

    Project description: In recent years, single cell RNA-Sequencing (scRNA-Seq) has become a key technology in understanding the variability and complexity of individual cells within tissues. However, this technology presents unique data analysis challenges due to the distinctive characteristics of the data. For example, large proportion of genes with exactly zero expression values (dropouts) as well as bimodal or multimodal distributions of the non-zero gene expression values. A potential interesting approach is to identify cell type specific network-based gene set (such as RNA splicing or transcription factors) using scRNAseq data. This project will examine different large covariance estimation techniques for construction of such networks and assess its potential impact on various diseases.

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

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

9 Assessment

9.1 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.

9.2 The coursework mark

Students are required to attend a minimum of 4 courses during the academic year. For old curriculum statistics honours students only the best 4 results which include STAT4028 or STAT4528 will be included in the overall assessment. These 4 results are weighted equally. For new curriculum students as well as data science honours students their course 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%. 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.

9.3 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?

9.4 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.

10 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.

11 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.

12 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.

13 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.

14 Additional proposed project topics by the statistics group

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

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

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

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

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

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

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

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

3. **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.*

4. **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.

5. **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.

6. **A fresh perspective - High parameter imaging and novel high throughput analytics to study HIV kinetics.**
   Supervisor: Dr. Ellis Patrick

   Project description: For the first time, imaging technologies have reached a maturity such that it is now possible to image the interaction of cells with individual HIV virions with high resolution and throughput. While the technologies have matured, the possible hypotheses that could be generated with this data are still in their infancy. In this project we can apply or develop cutting edge statistical machine learning tools to gain novel insight into the kinetics of HIV and various other diseases.

7. **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.

8. **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

9. **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.

10. **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.

11. **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 behavior of these new statistics in standard meta-analytic tests for publication bias remains questionable, yet these tests represent an important component of any meta-analysis. This project aims evaluate the behavior of variance-based effect sizes in common meta-analytic tests for publication bias, using simulated and/or real data, and if necessary to develop new tests of publication bias suitable to these statistics.

			Intensive August
Additional 4000-level COMP units to be developed for offering in 2021			

Bachelor of Advanced Studies (Honours, Data Science)

The Bachelor of Advanced Studies (Honours, Data Science) requires 48 credit points from this table including:

(i) 12 credit points of 4000-level and above Honours coursework selective units from List 1, and

(ii) 12 credit points of 4000-level and above Honours coursework selective units from List 1, List 2, List 3, or List 4 with a maximum of 6 credit points of units from List 3 or List 4 or List 5, and

(iii) 24 credit points of 4000-level Honours research project units

Honours Coursework Selective units – List 1

STAT4025 Time Series	6	**P** STAT2X11 and (MATH1X03 or MATH1907 or MATH1X23 or MATH1933) **N** STAT3925	Semester 1
STAT4026 Statistical Consulting	6	**P** At least 12cp from STAT2X11, STAT2X12, DATA2X02 and STAT3XXX **N** STAT3926	Semester 1
STAT4027 Advanced Statistical Modelling	6	**A** A three year major in statistics or equivalent including familiarity with material in DATA2X02 and STAT3X22 (Applied statistics and linear models) or equivalent. **P** STAT3x12 and STAT3x13	Semester 2
COMP5046 Natural Language Processing	6	**A** Knowledge of an OO programming language	Semester 1
COMP5328 Advanced Machine Learning	6	**A** COMP5318	Semester 2
COMP5329 Deep Learning	6	**A** COMP5318	Semester 1
COMP5338 Advanced Data Models	6	**A** This unit of study assumes foundational knowledge of relational database systems as taught in COMP5138/COMP9120 (Database Management Systems) or INFO2120/INFO2820/ISYS2120 (Database Systems 1).	Semester 2

COMP5349 Cloud Computing	6	**A** Good programming skills, especially in Java for the practical assignment, as well as proficiency in databases and SQL. The unit is expected to be taken after introductory courses in related units such as COMP5214 or COMP9103 Software Development in JAVA	Semester 1
COMP5048 Visual Analytics	6	**A** It is assumed that students will have basic knowledge of data structures, algorithms and programming skills. Note: Department permission required for enrolment	Semester 1 Semester 2
Additional 4000-level COMP units to be developed for offering in 2021			
Honours Coursework Selective units – List 2			
MATH4411 Applied Computation Mathematics	6	**A** MATH2X21 (Vector Calculus) MATH2X22 (Linear Algebra) Some familiarity with partial differential equations (MATH3978) and mathematical computing (MATH3976) is assumed.	Semester 1
MATH4412 Advanced Methods in Applied Mathematics	6	**A** MATH2X21 (Vector Calculus), MATH2X22 (Linear Algebra) or equivalent. Some familiarity with partial differential equations (MATH3978) and mathematical computing (MATH3976) is also assumed.	Semester 2
MATH4413 Applied Mathematical Modelling	6	**A** MATH2X21 and MATH3X63 or equivalent. That is, a knowledge of linear and simple nonlinear ordinary differential equations and of linear, second order partial differential equations.	Semester 1
MATH4414 Advanced Dynamical Systems	6	**A** Assumed knowledge is MATH2X21 (Vector Calculus) MATH2X22 (Linear Algebra) MATH4063 (Dynamical Systems and Applications) or equivalent. Some familiarity with partial differential equations (MATH3978) and mathematical computing (MATH3976) is also assumed.	Semester 2
MATH4061 Metric Spaces	6	**A** MATH2023 or MATH2923 or MATH2962 or MATH3068 **P** A mark of 65 or greater in 12cp from the following units (MATH3061 or MATH3066 or MATH3063 or MATH3076 or MATH3078 or MATH3962 or MATH3963 or MATH3968 or MATH3969 or MATH3971 or MATH3974 or	Semester 1

		MATH3976 or MATH3977 or MATH3978 or MATH3979) **N** MATH3961	
MATH4062 Rings fields and Galois Theory	6	**A** MATH2922 or MATH2961 **P** (MATH2922 or MATH2961) or a mark of 65 or greater in (MATH2022 or MATH2061) or 12cp from (MATH3061 or MATH3066 or MATH3063 or MATH3076 or MATH3078 or MATH3962 or MATH3963 or MATH3968 or MATH3969 or MATH3971 or MATH3974 or MATH3976 or MATH3977 or MATH3978 or MATH3979) **N** MATH3062, MATH3962	Semester 1
MATH4063 Dynamical Systems and Applications	6	**A** MATH2061 or MATH2961 or (MATH2X21 and MATH2X22) **P** (A mark of 65 or greater in 12cp of MATH2XXX units of study) or [12cp from (MATH3061 or MATH3066 or MATH3076 or MATH3078 or MATH3961 or MATH3962 or MATH3968 or MATH3969 or MATH3971 or MATH3974 or MATH3976 or MATH3977 or MATH3978 or MATH3979)]	Semester 1
MATH4068 Differential Geometry	6	**A** (MATH2921 and MATH2922) or MATH2961 **P** (A mark of 65 or greater in 12cp of MATH2XXX units of study) or [12cp from (MATH3061 or MATH3066 or MATH3063 or MATH3076 or MATH3078 or MATH3961 or MATH3962 or MATH3963 or MATH3969 or MATH3971 or MATH3974 or MATH3976 or MATH3977 or MATH3978 or MATH3979)] **N** MATH3968	Semester 2
MATH4069 Measure Theory and Fourier Analysis	6	**A** (MATH2921 and MATH2922) or MATH2961 **P** (A mark of 65 or greater in 12cp of MATH2XXX units of study) or [12cp from the following units (MATH3061 or MATH3066 or MATH3063 or MATH3076 or MATH3078 or MATH3961 or MATH3962 or MATH3963 or MATH3969 or MATH3971 or MATH3974 or MATH3976 or MATH3977 or MATH3978 or MATH3979)] **N** MATH3969	Semester 2
MATH4074 Fluid Dynamics	6	**A** (MATH2961 and MATH2965) or (MATH2921 and MATH2922) **P** (A mark of 65 or greater in 12cp of	Semester 1

		MATH2XXX units of study) or [12cp from (MATH3061 orMATH3066 or MATH3063 or MATH3076 or MATH3078 or or MATH3961 or MATH3962 or MATH3963 or MATH3969 or MATH3971 or MATH3974 or MATH3976 or MATH3977 or MATH3978 or MATH3979)] **N** MATH3974	
MATH4076 Computational Mathematics	6	**A** (MATH2X21 and MATH2X22) or (MATH2X61 and MATH2X65) **P** (A mark of 65 or greater in 12cp of MATH2XXX units of study) or [12cp from (MATH3061 or MATH3066 or MATH3063 or MATH3078 or or MATH3961 or MATH3962 or MATH3963 or MATH3969 or MATH3971 or MATH3974 or MATH3977 or MATH3978 or MATH3979)]	Semester 1
MATH4077 Lagrangian and Hamiltonian Dynamics	6	**A** 6cp of 1000 level calculus units and 3cp of 1000 level linear algebra and (MATH2X21 or MATH2X61) **P** (A mark of 65 or greater in 12cp of MATH2XXX units of study) or [12cp from (MATH3061 orMATH3066 or MATH3063 or MATH3076 or MATH3078 or MATH3961 or MATH3962 or MATH3963 or MATH3968 or MATH3969 or MATH3971 or MATH3974 or MATH3976 or MATH3978 or MATH3979)] **N** MATH3977	Semester 2
MATH4078 PDEs and Applications	6	**A** (MATH2X61 and MATH2X65) or (MATH2X21 and MATH2X22) **P** (A mark of 65 or greater in 12cp of 2000 level units) or [12cp from (MATH3061 or MATH3066 or MATH3063 or MATH3076 or MATH3961 or MATH3962 or MATH3963 or MATH3968 or MATH3969 or MATH3971 or MATH3974 or MATH3976 or MATH3977 or MATH3979)] **N** MATH3078, MATH3978	Semester 2
MATH4079 Complex Analysis	6	**A** MATH2X23 **P** (A mark of 65 or greater in 12cp of MATH2XXX units of study) or [12cp from the following units (MATH3061 or MATH3066 or MATH3063 or MATH3076 or MATH3078 or MATH3961 or MATH3962 or MATH3963 or MATH3968 or MATH3969 or MATH3971 or MATH3974 or MATH3976 or MATH3977 or MATH3978)] **N** MATH3979, MATH3964	Semester 1

STAT4022 Linear and Mixed Models	6	**A** Material in DATA2X02 or equivalent and MATH1X02 or equivalent; that is, a knowledge of applied statistics and an introductory knowledge to linear algebra, including eigenvalues and eigenvectors. **N** STAT3012, STAT3912, STAT3022, STAT3922, STAT3004, STAT3904	Semester 1
STAT4023 Theory and Methods of Statistical Inference	6	**A** STAT2X11 and (DATA2X02 or STAT2X12) or equivalent. That is, a grounding in probability theory and a good knowledge of the foundations of applied statistics. **N** STAT3013, STAT3913, STAT3023, STAT3923	Semester 2
MATH4071 Convex Analysis and Optimal Control	6	**A** MATH2X21 and MATH2X23 and STAT2X11 **P** [A mark of 65 or greater in 12cp from (MATH2070 or MATH2970 or STAT2011 or STAT2911 or MATH2021 or MATH2921 or MATH2022 or MATH2922 or MATH2023 or MATH2923 or MATH2061 or MATH2961 or MATH2065 or MATH2965 or MATH2962 or STAT2012 or STAT2912 or DATA2002 or DATA2902) or [12 cp from (MATH3075 or MATH3975 or STAT3021 or STAT3011 or STAT3911 or STAT3888 or STAT3014 or STAT3914 or MATH3063 or MATH3963 or MATH3061 or MATH3961 or MATH3962 or MATH3963 or MATH3968 or MATH3969 or MATH3971 or MATH3974 or MATH3976 or MATH3977 or MATH3978 or MATH3979)] **N** MATH3971	Semester 1
MATH4511 Arbitrage Pricing in Continuous Time	6	**A** At least 6 credit points of (2000 Advanced Mathematics or 3000 Advanced Mathematics or 4000 Mathematics units) **P** Credit average or greater in 12 credit points of 2000-level Mathematics	Semester 2
MATH4512 Stochastic Analysis	6	**A** Students should have sound knowledge of probability theory and stochastic processes from, for example, STAT2X11 and STAT3021 or equivalent.	Semester 2
MATH4513 Topics in Financial Mathematics	6	**A** Students are expected to have working knowledge of Stochastic Processes, Stochastic Calculus and mathematical methods used to price options and other financial derivatives,	Semester 2

		for example as in MATH4511 or equivalent	
MATH4311 Algebraic Topology	6	**A** (MATH2922 or MATH2961 or equivalent) and (MATH2923 or equivalent). Familiarity with abstract algebra and basic topology.	Semester 1
MATH4312 Commutative Algebra	6	**A** MATH2922 or equivalent. Familiarity with abstract algebra.	Semester 1
MATH4313 Functional Analysis	6	**A** Real Analysis (for example, MATH2X23 or equivalent), and, preferably, knowledge of Metric Spaces.	Semester 1
MATH4314 Representation Theory	6	**A** (MATH2922 or MATH2961 or equivalent). Familiarity with abstract algebra, specifically vector space theory and basic group theory.	Semester 1
MATH4315 Variational Methods	6	**A** Assumed knowledge of MATH2023 or MATH2923; MATH4061 or MATH3961; MATH3969 or MATH4069; MATH4313 or equivalent. That is, real analysis, basic functional analysis and some acquaintance with metric spaces or measure theory.	Semester 2
STAT4921 Probability and Mathematical Statistics	6	**A** STAT3X23 or equivalent: that is, a sound working and theoretical knowledge of statistical inference. **N** STAT4521	Semester 1
STAT4021 Stochastic Processes and Applications	6	**A** STAT2011 or STAT2911, and MATH1003 or MATH1903 or MATH1907 or MATH1023 or MATH1933 or equivalent. That is, thorough knowledge of basic probability and integral calculus and to have achieved at credit level or above in their studies these topics. **N** STAT3011, STAT3911, STAT3021, STAT3003, STAT3903, STAT3005, STAT3905, STAT3921.	Semester 1
STAT4027 Advanced Statistical Modelling	6	**A** A three year major in statistics or equivalent including familiarity with material in DATA2X02 and STAT3X22 (Applied statistics and linear models) or equivalent.	Semester 2
Honours Coursework Selective units – List 3			
5000-level DATA units from the School of Mathematics and Statistics			
Honours Coursework Selective units - List 4			
Other 5000-level units available in the School of Mathematics and Statistics			

Honours Coursework Selective units – List 5			
4000-level or 5000-level units at other Schools at the University			
Honours Core Research Project units			
DATA4103 Data Science Honours Project A	6	**A** Equivalent to a major in Data Science and a WAM of 65 or greater. **C** DATA4104 and DATA4105 and DATA4106	Semester 1 Semester 2
DATA4104 Data Science Honours Project B	6	**A** Equivalent to a major in Data Science and a WAM of 65 or greater. **C** DATA4103 and DATA4105 and DATA4106	Semester 1 Semester 2
DATA4105 Data Science Honours Project C	6	**A** Equivalent to a major in Data Science and a WAM of 65 or greater. **C** DATA4103 and DATA4104 and DATA4106	Semester 1 Semester 2
DATA4106 Data Science Honours Project D	6	**A** Equivalent to a major in Data Science and a WAM of 65 or greater. **C** DATA4103 and DATA4104 and DATA4105	Semester 1 Semester 2

SCIE4003 Ethics in Science	6	**A** Completion of a major **N** HSBH3004	Intensive February Intensive August

Bachelor of Advanced Studies (Honours, Statistics)

The Bachelor of Advanced Studies (Honours, Statistics) requires 48 credit points from this table including:

(i)	6 credit points of 4000-level Honours coursework selective units from List 1, and
(ii)	6 credit points of 4000-level Honours coursework selective units from List 2, and
(iii)	12 credit points of 4000-level and 5000-level Honours coursework selective units from List 2, List 3, List 4, or List 5, a maximum of 6 credit points of which may be from List 3 and a maximum of 6 credit points of which may be from List 4, and
(iv)	24 credit points of 4000-level Honours research project units

Honours Coursework Selective units – List 1

STAT4921 Probability and Mathematical Statistics	6	**A** STAT3X23 or equivalent: that is, a sound working and theoretical knowledge of statistical inference. **N** STAT4521	Semester 1
STAT4521 Probability and Martingale Theory	6	**A** STAT2X11 or equivalent and STAT3X21 or equivalent; that is, a good foundational knowledge of probability and some acquaintance with stochastic processes. **N** STAT4921	Semester 1

Honours Coursework Selective units – List 2

STAT4026 Statistical Consulting	6	**P** At least 12cp from STAT2X11, STAT2X12, DATA2X02 and STAT3XXX **N** STAT3926	Semester 1
STAT4027 Advanced Statistical Modelling	6	**A** A three year major in statistics or equivalent including familiarity with material in DATA2X02 and STAT3X22 (Applied statistics and linear models) or equivalent. **P** STAT3x12 and STAT3x13	Semester 2

Honours Coursework Selective units – List 3

A 4000- or 5000-level unit from a School other than the School of Mathematics and Statistics

Honours Coursework Selective units – List 4

A 5000-level unit available in the School of Mathematics and Statistics

Honours Coursework Selective units – List 5			
STAT4021 Stochastic Processes and Applications	6	**A** STAT2011 or STAT2911, and MATH1003 or MATH1903 or MATH1907 or MATH1023 or MATH1933 or equivalent. That is, students are expected to have a thorough knowledge of basic probability and integral calculus and to have achieved at credit level or above in their studies in these topics. **N** STAT3011, STAT3911, STAT3021, STAT3003, STAT3903, STAT3005, STAT3905, STAT3921.	Semester 1
STAT4022 Linear and Mixed Models	6	**A** Material in DATA2X02 or equivalent and MATH1X02 or equivalent; that is, a knowledge of applied statistics and an introductory knowledge to linear algebra, including eigenvalues and eigenvectors. **N** STAT3012, STAT3912, STAT3022, STAT3922, STAT3004, STAT3904.	Semester 1
STAT4023 Theory and Methods of Statistical Inference	6	**A** STAT2X11 and (DATA2X02 or STAT2X12) or equivalent. That is, a grounding in probability theory and a good knowledge of the foundations of applied statistics. **N** STAT3013, STAT3913, STAT3023, STAT3923	Semester 2
STAT4025 Time Series	6	**P** STAT2X11 and (MATH1X03 or MATH1907 or MATH1X23 or MATH1933) **N** STAT3925	Semester 1
MATH4061 Metric Spaces	6	**A** MATH2023 or MATH2923 or MATH2962 or MATH3068 **P** A mark of 65 or greater in 12cp from the following units (MATH3061 orMATH3066 or MATH3063 or MATH3076 or MATH3078 or MATH3962 or MATH3963 or MATH3968 or MATH3969 or MATH3971 or MATH3974 or MATH3976 or MATH3977 or MATH3978 or MATH3979) **N** MATH3961	Semester 1
MATH4062 Rings fields and Galois Theory	6	**A** MATH2922 or MATH2961 **P** (MATH2922 or MATH2961) or a mark of 65 or greater in (MATH2022 or MATH2061) or 12cp from (MATH3061 or MATH3066 or MATH3063 or MATH3076 or MATH3078 or MATH3962 or MATH3963 or MATH3968 or MATH3969 or MATH3971 or MATH3974 or MATH3976 or MATH3977 or MATH3978	Semester 1

		or MATH3979) **N** MATH3062, MATH3962	
MATH4063 Dynamical Systems and Applications	6	**A** MATH2061 or MATH2961 or (MATH2X21 and MATH2X22) **P** (A mark of 65 or greater in 12cp of MATH2XXX units of study) or [12cp from (MATH3061 or MATH3066 or MATH3076 or MATH3078 or MATH3961 or MATH3962 or MATH3968 or MATH3969 or MATH3971 or MATH3974 or MATH3976 or MATH3977 or MATH3978 or MATH3979)]	Semester 1
MATH4068 Differential Geometry	6	**A** (MATH2921 and MATH2922) or MATH2961 **P** (A mark of 65 or greater in 12cp of MATH2XXX units of study) or [12cp from (MATH3061 or MATH3066 or MATH3063 or MATH3076 or MATH3078 or MATH3961 or MATH3962 or MATH3963 or MATH3969 or MATH3971 or MATH3974 or MATH3976 or MATH3977 or MATH3978 or MATH3979)] **N** MATH3968	Semester 2
MATH4069 Measure Theory and Fourier Analysis	6	**A** (MATH2921 and MATH2922) or MATH2961 **P** (A mark of 65 or greater in 12cp of MATH2XXX units of study) or [12cp from the following units (MATH3061 or MATH3066 or MATH3063 or MATH3076 or MATH3078 or MATH3961 or MATH3962 or MATH3963 or MATH3969 or MATH3971 or MATH3974 or MATH3976 or MATH3977 or MATH3978 or MATH3979)] **N** MATH3969	Semester 2
MATH4071 Convex Analysis and Optimal Control Theory	6	**A** MATH2X21 and MATH2X23 and STAT2X11 **P** [A mark of 65 or greater in 12cp from (MATH2070 or MATH2970 or STAT2011 or STAT2911 or MATH2021 or MATH2921 or MATH2022 or MATH2922 or MATH2023 or MATH2923 or MATH2061 or MATH2961 or MATH2065 or MATH2965 or MATH2962 or STAT2012 or STAT2912 or DATA2002 or DATA2902) or [12 cp from (MATH3075 or MATH3975 or STAT3021 or STAT3011 or STAT3911 or STAT3888 or STAT3014 or STAT3914 or MATH3063 or MATH3963 or MATH3061 or MATH3961 or MATH3962	Semester 1

		or MATH3963 or MATH3968 or MATH3969 or MATH3971 or MATH3974 or MATH3976 or MATH3977 or MATH3978 or MATH3979)] **N** MATH3971	
MATH4074 Fluid Dynamics	6	**A** (MATH2961 and MATH2965) or (MATH2921 and MATH2922) **P** (A mark of 65 or greater in 12cp of MATH2XXX units of study) or [12cp from (MATH3061 orMATH3066 or MATH3063 or MATH3076 or MATH3078 or or MATH3961 or MATH3962 or MATH3963 or MATH3969 or MATH3971 or MATH3974 or MATH3976 or MATH3977 or MATH3978 or MATH3979)] **N** MATH3974	Semester 1
MATH4076 Computational Mathematics	6	**A** (MATH2X21 and MATH2X22) or (MATH2X61 and MATH2X65) **P** (A mark of 65 or greater in 12cp of MATH2XXX units of study) or [12cp from (MATH3061 or MATH3066 or MATH3063 or MATH3078 or or MATH3961 or MATH3962 or MATH3963 or MATH3969 or MATH3971 or MATH3974 or MATH3977 or MATH3978 or MATH3979)]	Semester 1
MATH4077 Lagrangian and Hamiltonian Dynamics	6	**A** 6cp of 1000 level calculus units and 3cp of 1000 level linear algebra and (MATH2X21 or MATH2X61) **P** (A mark of 65 or greater in 12cp of MATH2XXX units of study) or [12cp from (MATH3061 orMATH3066 or MATH3063 or MATH3076 or MATH3078 or MATH3961 or MATH3962 or MATH3963 or MATH3968 or MATH3969 or MATH3971 or MATH3974 or MATH3976 or MATH3978 or MATH3979)] **N** MATH3977	Semester 2
MATH4078 PDEs and Applications	6	**A** (MATH2X61 and MATH2X65) or (MATH2X21 and MATH2X22) **P** (A mark of 65 or greater in 12cp of 2000 level units) or [12cp from (MATH3061 or MATH3066 or MATH3063 or MATH3076 or MATH3961 or MATH3962 or MATH3963 or MATH3968 or MATH3969 or MATH3971 or MATH3974 or MATH3976 or MATH3977 or MATH3979)] **N** MATH3078, MATH3978	Semester 2
MATH4079 Complex Analysis	6	**A** MATH2X23 **P** (A mark of 65 or greater in 12cp of	Semester 1

		MATH2XXX units of study) or [12cp from the following units (MATH3061 or MATH3066 or MATH3063 or MATH3076 or MATH3078 or MATH3961 or MATH3962 or MATH3963 or MATH3968 or MATH3969 or MATH3971 or MATH3974 or MATH3976 or MATH3977 or MATH3978)] **N** MATH3979, MATH3964	
MATH4411 Applied Computation Mathematics	6	**A** MATH2X21 (Vector Calculus) MATH2X22 (Linear Algebra) Some familiarity with partial differential equations (MATH3978) and mathematical computing (MATH3976) is assumed.	Semester 1
MATH4412 Advanced Methods in Applied Mathematics	6	**A** MATH2X21 (Vector Calculus), MATH2X22 (Linear Algebra) or equivalent. Some familiarity with partial differential equations (MATH3978) and mathematical computing (MATH3976) is also assumed.	Semester 2
MATH4413 Applied Mathematical Modelling	6	**A** MATH2X21 and MATH3X63 or equivalent. That is, a knowledge of linear and simple nonlinear ordinary differential equations and of linear, second order partial differential equations.	Semester 1
MATH4414 Advanced Dynamical Systems	6	**A** Assumed knowledge is MATH2X21 (Vector Calculus) MATH2X22 (Linear Algebra) MATH4063 (Dynamical Systems and Applications) or equivalent. Some familiarity with partial differential equations (MATH3978) and mathematical computing (MATH3976) is also assumed.	Semester 2
MATH4511 Arbitrage Pricing in Continuous Time	6	**A** At least 6 credit points of (2000 Advanced Mathematics or 3000 Advanced Mathematics or 4000 Mathematics units) **P** Credit average or greater in 12 credit points of 2000-level Mathematics	Semester 2
MATH4512 Stochastic Analysis	6	**A** Students should have sound knowledge of probability theory and stochastic processes from, for example, STAT2X11 and STAT3021 or equivalent.	Semester 2
MATH4513 Topics in Financial Mathematics	6	**A** Students are expected to have working knowledge of Stochastic Processes, Stochastic Calculus and	Semester 2

		mathematical methods used to price options and other financial derivatives, for example as in MATH4511 or equivalent	
MATH4311 Algebraic Topology	6	**A** (MATH2922 or MATH2961 or equivalent) and (MATH2923 or equivalent). Familiarity with abstract algebra and basic topology.	Semester 1
MATH4312 Commutative Algebra	6	**A** MATH2922 or equivalent. Familiarity with abstract algebra.	Semester 1
MATH4313 Functional Analysis	6	**A** Real Analysis (for example, MATH2X23 or equivalent), and, preferably, knowledge of Metric Spaces.	Semester 1
MATH4314 Representation Theory	6	**A** (MATH2922 or MATH2961 or equivalent). Familiarity with abstract algebra, specifically vector space theory and basic group theory.	Semester 1
MATH4315 Variational Methods	6	**A** Assumed knowledge of MATH2023 or MATH2923; MATH4061 or MATH3961; MATH3969 or MATH4069; MATH4313 or equivalent. That is, real analysis, basic functional analysis and some acquaintance with metric spaces or measure theory.	Semester 2
Honours Core Research Project units			
STAT4103 Statistics Honours Project A	6	**C** STAT4104 and STAT4105 and STAT4106	Semester 1 Semester 2
STAT4104 Statistics Honours Project B	6	**C** STAT4103 and STAT4105 and STAT4106	Semester 1 Semester 2
STAT4105 Statistics Honours Project C	6	**C** STAT4103 and STAT4104 and STAT4106	Semester 1 Semester 2
STAT4106 Statistics Honours Project D	6	**C** STAT4103 and STAT4104 and STAT4105	Semester 1 Semester 2

Notes:

- We intend to add a specific list of units form other Schools at the University (such as Physics, Economics, etc.) to encourage breadth and reduce the number of special considerations applications to count these towards this degree