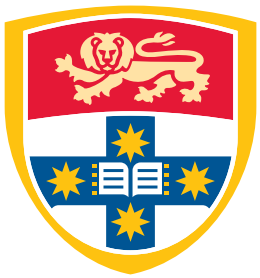
Honours in Statistics

A guide for the 2025 academic year

(updated 16/9/24)



THE UNIVERSITY OF SYDNEY

School of Mathematics and Statistics

Contents

1 Pathways to Honours

The Faculty of Science offers two main Honours pathways:

- Currently enrolled **Bachelor of Advanced Studies** students seeking to study honours have the **option** to apply for an appended disciplinary honours degree (e.g., Bachelor of Science (Honours)). Notice that this option requires two majors and is only available if you commenced a combined Bachelor of Advanced Studies degree after 2018 and before 1 July 2024. See also **Information for Bachelor of Advanced Studies students**.
- Appended **Bachelor of Science (Honours)** is a standalone Honours requiring an additional year of study. This option requires only one major and it is open for all students who satisfy the criteria below.

2 Entry requirements

Preliminary entrance into the honours program is through the Faculty of Science application portal for standalone Bachelor of Science (Honours) or, if you are enrolled into combined BSc/BAS program, then you can apply for Advanced Studies honours (look under Bachelor of Science/Bachelor of Advanced Studies) through Sydney Student in your final semester of BSc (go to Course details and apply for Advanced Studies honours). The Faculty admission requirements include that you must:

- have qualified for or be a graduate with a Bachelor of Science degree or equivalent from the University of Sydney or equivalent qualification from another tertiary institution;
- have completed a relevant major (i.e. minimum of 24 credit points of 3000-level units of study) relating to the intended Honours discipline;
- have achieved a Weighted Average Mark (WAM) of at least 65.00 or have a credit average (65.00) in 48 credit points of relevant 2000-level and 3000-level units of study (as nominated by the school); and
- satisfy any additional criteria set by the relevant Head of School or Discipline (see below).

In addition, the School of Mathematics and Statistics requires that the student **secures the agreement of a supervisor** and that the student has a total of at least 24CP of relevant 3XXX unit of studies in which

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

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

2.1 It's important to note that:

- All acceptances into Honours (including in cases where the School's requirements are not met) are ultimately at the discretion of the School. However, a student meeting all of the above criteria (or the equivalent from another institution) should be confident of acceptance.
- The Faculty of Science Honours **standard closing dates** for Honours commencement in Semester 1, 2025 is 15 January 2026, and for Semester 2, 2025 it is 25 June 2025.

3 Structure of Honours

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

3.1 The honours project (50%)

The honours project centres around an essay/thesis consisting of roughly 60 pages[1] written on a particular topic from your chosen area. It need not contain original research (although it might) but it should clearly demonstrate that you have understood and mastered the material. The assessment of the honours thesis is based on the scientific/statistical/mathematical content and its exposition, including the written english. The thesis is due at the end of your second semester, specifically at 5pm on Monday of week 13.

Toward the end of the second semester (typically 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.

3.1.1 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 Hub offers workshops for students that need help with extended written work, and a trove of online resources for improving your writing skills is also available. Make sure you make use of these resources as early as possible as writing skills develop slowly over time and with much practice.

3.2 Course work (50%)

The honours program in *statistics* specifies a couple of core courses as well as which combination of courses can be taken. The list of available courses can be found online, however please carefully read through the list of constraints outlined in the statistics Honours degrees structure document. In particular note that you have to take one course of the two in List 1 and one of the two in List 2.

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

4 Important course work information for all students

4.1 Selecting your courses

Please make sure you **select your courses after consulting your supervisor as well as your 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). These courses can possibly be taken for credit (by enrolling in the unit AMSI4001), but this can only be done in consultation the student's supervisor and with the approvals of the specific honours coordinator as well as the School's Honours coordinator, Prof. Laurentiu Paunescu.

5 Program Administration

The Statistics Honours coordinator is

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

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

A/Prof John Ormerod,
Carslaw Building, Room 815, Phone 9351 5883,
Email: john.ormerod@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 and/or Special Considerations as deemed appropriate.

6 Potential Supervisors and their Research Interests

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

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.

Dr. Shila Ghazanfar
Data science, R/Shiny, interactive data visualisation, single-cell RNA-sequencing, spatially resolved genomics data.

Dr. Clara Grazian
Bayesian statistics, mixture models, copula models, spatio-temporal data, genomic data, approximate Bayesian algorithms.

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

Doctor Linh Nghiem
Dimension Reduction, Measurement Error and Contaminated Data Modelling, Mixed Models, Graphical Models, Applied Bayesian Analysis.

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, ARMA models with Autocorrelated errors, Statistics in Finance,
Financial Econometrics, Hybrid ARIMA modelling and Applications.

Doctor Michael Stewart

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

Doctor Qiuzhuang Sun

Data-Driven Operations Research, Applied Data Science, Engineering Statistics,
Dynamic Programming, Robust Optimization.

Doctor Garth Tarr

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

Professor Qiying Wang

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

Doctor Rachel Wang

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

Professor Jean Yang

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

Associate Professor Pengyi Yang

Machine learning, Deep learning, Statistical modelling,
Single-cell omics, Multi-omics data integration, Computational Biology, Bioinformatics.

7 Honours courses in Statistics and Data Science

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

The projects are divided into two categories: Data Science and Statistics. If you are a Data Science student but you prefer a project that is listed under Statistics or the other way around then feel free to talk the relevant supervisor about it. For additional projects see the Section 14 at the end of this document.

8.2.1 Proposed project topics in Statistics

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

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

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

   Project description: This project will investigate properties of high frequency cryptocurrency returns data which often display high kurtosis. Popular heavy tail distributions like Student t and exponential power may still be inadequate to provide high enough level of kurtosis. Recent studies have considered variance gamma distribution in which the shape parameter can be made sufficiently small to provide unbounded density around the centre and heavy tails at the two ends of the distribution. As gamma variance distribution can be expressed as scale mixtures of normal, it facilitates model implementation in the Bayesian approach via some Bayesian softwares such as stan (in R). We will consider long memory, stochastic volatility and leverage effect modelling and investigate how these features change during the pandemic period. Currently, there are few studies which investigate the impact of COVID 19 pandemic on the cryptocurrency market and so this study will be pioneering and interesting.

3. **Assessing continuous changes in correlation structure for replicated single cell data.**
   Supervisor: Dr Shila Ghazanfar

   Project description: Single-cell genomics has transformed our ability to examine cell fate choice. Examining cells along a computationally ordered 'pseudotime' or across spatial coordinates offers the potential to unpick subtle changes in variability and covariation among key genes. A key challenge now is how to perform such continuous differential variation or covariation testing within a multi-sample and potentially multi-condition experiment. This project will examine the use of generalised additive models (GAMs) towards assessing changes in covariation patterns along pseudotime and spatial coordinates using a mixture of real-world and simulated data.

4. **Assessing error propagation in mosaic data integration.**
   Supervisor: Dr Shila Ghazanfar

Project description: This project aims to perform error estimation for gene expression imputation and spatial position inference from mosaic data integration of single cell spatial genomics. This proposed framework will try to extract measures of ambiguity for cells' joint integration. These noise estimates will be obtained by identifying, for each cell, the relative size of the cell's neighbourhoods in the joint network, repeatedly estimated using techniques such as bootstrapping. This will result in three key outputs, first, the estimated gene expression and associated variances for each gene for imputation of each spatial cell; second, the estimated spatial position and associated posterior distribution (represented as a field) for each non-spatial cell; and third, an ambiguity metric, represented by the relative proportion of plausible cells an either spatial or non-spatial cell is most similar to. The variance estimates will be used as feature weights for downstream analysis (e.g. differential expression) to enable propagation of error. This investigation will go towards further understanding the underlying data structure and information content of these distinct data modalities.

5. **Approximate Bayesian Computation via composite likelihood.**
   Supervisor: Dr Clara Grazian

   Project description: There have been several proposals in the literature to introduce composite likelihoods in a Bayesian setting, but usually the score function of the composite likelihood needs to be tractable. In this project, we will work with the pairwise likelihood function, which is easy to approximate either via simulation or kernel density estimation, but directly using the pairwise likelihood as estimator of the likelihood to combine with the prior distribution to obtain the posterior concentration leads to an over-concentration of the estimated posterior distribution. The first part of the project show that it is feasible to numerically approximate the summary statistics from the pairwise likelihood function. The second part of the project will be devoted to develop procedures to estimate such function in a simulation-based setting.

6. **Copula dependence techniques for inferring directionality on gene expression data.**
   Supervisor: Dr Clara Grazian

   Project description: The goal of project is to identify the directional dependence among pairs of genes through their genomic expression. The importance of this investigation is that gene regulation and expression are the foundations of the development of all cells. By understanding the dependence mechanisms that underlay gene expression, fundamental questions that have consequences in health and disease identification and progression can be addressed. The focus of this work is to identify the direction of influence just by modelling the dependence structure between gene pairs. This will be achieved by employing copulas, which are distribution functions able to isolate the marginals from the joint behaviour. Several methods based on copulas will be explored and validated on real datasets, retrieved from the DREAM5 Network Inference Challenge. The DREAM5 network consists of four networks namely in-silico, e.coli, s.aureus and s.cerevisiae. The in-silico data is a simulated network, consisting of 195 transcription factors, which produce 4012 genes interactions. The e.coli network comprises of 334 transcription factors regulating 2066 genes. The s.aureus consists of 95 transcription factors creating 518 interactions, while the s.cerevisiae network constitutes of 333 transcription factors regulating 3940 genes.

7. **ABC-MC: approximate Bayesian computation for model choice.**

Supervisor: Dr Clara Grazian

Project description: Approximate Bayesian computation has become an essential tool to handle complex models in recent years. The basic idea is to generate a proposal value for the parameter and generate a data set from the model given the proposed value. If the observed and the simulated data set and the simulated are similar in some sense, the proposed parameter value is considered to be likely to have generated the data and becomes part of the sample which will approximate the posterior distribution. This class of algorithms avoids evaluation of the likelihood by replacing it with simulation from the associated model. In the comparison between data sets, the method requires the selection of low-dimensional summary statistics which are unlikely to be sufficient in the setting of complex models. The loss of information implied in the procedure is, in general, considered acceptable for inferential problems, because it makes manageable models otherwise intractable. On the other hand, many works have demonstrated that lack of sufficiency may lead to inconsistency in problems of model choice, i.e. ABC applied to model choice may fail to recover the true model. This project will analyse possible alternative summary statistics to be used when the goal of the analysis is model choice instead of inference.

8. **Novel construction of decoys for controlling the 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 (target) database.

Unfortunately, the resulting list of optimal peptide-spectrum matches contains many incorrect, random matches. The canonical way to control the associated type I error is by controlling the false discovery rate (FDR) through target-decoy competition. Specifically, the matches to peptides in the target database are contrasted with matches to pseudo-peptides in a decoy database.

Invariably, the decoy peptides are generated by shuffling or by reversing the target peptides — methods that have limitations that have recently came to light. In this project we will look at alternative methods for constructing decoys. *No prior understanding of proteomics is required.*

9. **Competition-based approach to False Discovery Rate (FDR)**
Supervisor: A/Prof. Uri Keich

Project Description: In a ground breaking paper, Benjamini and Hochberg offered a practical novel approach to the multiple testing problem: controlling the false discovery rate (FDR). The FDR is defined as the expected value of the false discovery proportion (FDP), equivalently the proportion of falsely rejected null hypotheses among all rejected hypotheses (discoveries). Today, this is the canonical approach to type I error control in the multiple testing context when (informative) p-values can be associated with each hypothesis.

More recently, Barber and Candès (2015) offered an alternative framework for FDR control, where instead of p-values each test statistic can only be compared with a single null score. Interestingly, such competition based-approach to FDR control has been widely practiced in the mass spectrometry community for some time before Barber and Candès' work. Moreover, the mass spectrometry context continues to offer novel challenges, both theoretical and

practical in this rapidly developing area of competition-based approach to multiple testing. Interested students are welcomed to discuss potential projects with me.

10. **Sufficient dimension reduction for large datasets**
Supervisor: Dr. Linh Nghiem

Project description: In a regression setting with covariates matrix $\mathbf{X}$ and (discrete or continuous) outcome y, sufficient dimension reduction (SDR) refers to a class of dimension reduction methods that both reduce the dimensionality of $\mathbf{X}$ and ensure *no* information to predict y from $\mathbf{X}$ is lost after the reduction. Combining dimension reduction with the sufficiency principle in statistical inference, SDR has increasingly gained popularity and showed strong performance in large and big datasets. The project can range from reviewing the most popular SDR methods, checking their performance on publicly available data, to developing new SDR methods for non-standard settings, such as missing data, measurement errors, or longitudinal data.

11. **Statistical methods for contaminated data**
Supervisor: Dr. Linh Nghiem

Project description: In many applications, researchers or data practitioners do not have access to the true variables of interests; instead, they only have access to their proxies. For example, nutritional epidemiologists are interested in the relationship between long-term diets and risk diseases; nevertheless, diet data are typically only available via self–reported consumption of food during a specific time period, say 24 hour or one week. These self-report data are known to be very noisy, subject to measurement errors and personal bias, and treating them as long-term diets leads to inaccurate estimation of the effect. In a different setting, many government and industry agencies nowadays only release masked data to protect *privacy* of the participants. These masked data pose many challenges for applying appropriate statistical methods to make inference on the effects of interests. Research in this project can range from reviewing statistical methods to correct for different types of contamination, exploring tradeoffs between privacy and statistical accuracy, to developing new statistical methods for contaminated data in complex settings.

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

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

these ideas to improve the speed or accuracy of these methods. A student with this project will gain skills in statistical computing, multivariate calculus, and multivariate statistics.

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

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

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

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

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

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

Project description: This project extends the family of autoregressive fractionally integrated moving average (ARFIMA) processes to handle multivariate time series with long memory.

We consider the theory of estimation and applications of vector models in financial econometrics.

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

16. **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.

17. **Improved strong approximations for non-Donsker empirical processes**
Project Supervisor: Dr Michael Stewart

Project description: Empirical processes are appropriately normalised and centred sample averages of random functions. So long as the family of functions is not too large (in a certain sense), as the sample size increases the empirical process converges in distribution (in an appropriate sense) to a Gaussian process, a random function where any finite collection of values of the function's argument yields a multivariate normal random vector. This generalises the Central Limit Theorem. Such a well-behaved family of functions is known as a "Donsker class", following the result of Donsker (1952) who examined the empirical process based on the empirical cumulative distribution function.

In some applications, including simple mixture models, we meet empirical processes where the family of functions is too large. In such cases, the best we can do is approximate the empirical process with a certain sequence of random processes which change as the sample size increases. An example of such a "strong approximation" appears in Stewart and Robinson (2003), where a sequence of Gaussian processes over a slowly growing interval is used to approximate the (standardised) empirical moment-generating function of a normal sample.

The aim of this project is to study an improved strong approximation method which combines a Gaussian process with a Poisson process in order to capture more accurately what happens "in the tails" for certain extreme values of the function's argument. While the resulting approximation is more complicated than a Gaussian approximation, it may yield useful results in certain applications which are not available using existing methods.

References:
Monroe D. Donsker. Justification and extension of Doob's heuristic approach to the Kolmogorov-Smirnov theorems. *Ann. Math. Statistics*, 23:277–281, 1952. ISSN 0003-4851.

Michael Stewart and John Robinson. Extremes of normed empirical moment generating function processes. *Extremes*, 6(4):319–333 (2005), 2003. ISSN 1386-1999.

18. **Exploring double descent in linear models**
Project supervisor: Dr Michael Stewart

Project description: "Double descent" describes a particular phenomenon involving the test error in certain machine-learning models. Specifically, as the model complexity increases in the "underparametrised region" (where the number of parameters $p < n$, the number of observations) the test error decreases to some point and then increases again as $p \uparrow n$, but may descend again as p increases further into the "overparametrised" region where $p > n$. The purpose of this project is to study some recent work into how this effect may be apparent in standard linear models. There is potential for both theoretical and computational work in this project.

19. **Statistical modeling of recurrent event data on a network**
Supervisor: Dr. Qiuzhuang Sun
Project description: Recurrent event data refer to the repeated occurrence of an event over time, e.g., the insurance warranty claims and outbreak of diseases. Predicting the future occurrence of recurrent events is helpful in budgeting and risk management. A common method to analyse the recurrent events is to model it as a point process and estimate the intensity function of this point process. Sometimes, the occurrence of recurrent events is on a network, for instance, failures of pipelines of a water pipe network. Data analysis of recurrent events on a network is challenging due to the dependence of different segments of the network. For the above example on water pipe systems, the failures of pipelines tend to have a clustering structure. This project aims to develop a statistical model to accurately predict the occurrence time of future events on a network. We are expected to use tools such as self-exciting processes (Ertekin et al. 2015) to model the dependence structure of the recurrence data on the network. Distributed optimization algorithms like ADMM (Hallac et al. 2015) could be useful for optimization of large-scale networks. The proposed model will be used to analyse the failure dataset of a water pipe system from Scotland.

- Ertekin, Ş., Rudin, C., and McCormick, T. H. (2015), Reactive point processes: A new approach to predicting power failures in underground electrical systems. *Annals of Applied Statistics*, 9(1), 122–144.
- Hallac, D., Leskovec, J., and Boyd, S. (2015), Network Lasso: Clustering and optimization in large graphs. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 387–396).

20. **Data-driven prediction and control of glaucoma progression**
Supervisor: Dr. Qiuzhuang Sun
Project description: Periodic measurements of visual fields (VFs) are necessary in monitoring the glaucoma progression. The VF measurements, however, are well known to be subject to relatively large measurement errors. One of the reasons for the large measurement errors is that the VF measurement needs patient's concentration for a long time. As such, fatigue or unfamilarity with the measurement process can lead to the deviation of the measurements from the true underlying state. This study aims to propose a systematic modeling framework to capture such dynamics. The model needs to consider both (i) the correlation of measurements among different eyes, hemifields, and measurement locations of the same patient, and

(ii) the heterogeneity of disease progression for different patients. Based on the model, we would like to use data to make medical decisions such as the time to the next test for a new patient. The proposed model will be compared with some benchmarks in Helm et al. (2015) and Bryan et al. (2017).

- Helm, J. E., Lavieri, M. S., Van Oyen, M. P., Stein, J. D., and Musch, D. C. (2015), Dynamic forecasting and control algorithms of glaucoma progression for clinician decision support. *Operations Research*, 63(5), 979–999.
- Bryan, S. R., Eilers, P. H., Rosmalen, J. V., Rizopoulos, D., Vermeer, K. A., Lemij, H. G., and Lesaffre, E. M. (2017), Bayesian hierarchical modeling of longitudinal glaucomatous visual fields using a two-stage approach. *Statistics in Medicine*, 36(11), 1735–1753.

21. **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.

22. **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.

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

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

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

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

is standard normal. This project involves deep classical probability knowledge. The interest in theoretical work is essential.

25. **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).

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

27. **Developing deep-learning based methods for integrative analysis of single-cell and spatial omics data**
Supervisor: A/Prof. Pengyi Yang
Project description: Single-cell omics technologies enable global profiling of the molecular features (e.g. genes, proteins) in individual cells. These technologies represent a new frontier

for the discovery and characterisation of cell types in complex tissues and their changes in diseases. Recent development of spatial omics technologies adds to the molecular features captured by single-cell technologies and provides the spatial locations of cells that further enable data analysis in a cell neighbourhood with surrounding cellular environment. To this end, the development of computational methods for integrative analysis of single-cell and spatial omics data is essential for elucidating molecular mechanisms of cells and in diseases. This project aims to develop state-of-the-art and high-performing deep learning-based methods for achieving this goal.

9 Assessment

9.1 The honours grade

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

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

Grade of Honours	Faculty-Scale
First Class	80–100
Second Class, First Division	75–79
Second Class, Second Division	70–74
Third Class	65–69
Fail	0–64

For more details including on criteria for awarding a University Medal please refer to Schedule 2 of the COURSEWORK POLICY 2021 document. In addition to the guidelines in the linked document, the Faculty of Science only will only consider awarding a University Medal if the student has an Honours mark ≥ 90 and undergraduate WAM ≥ 80.

9.2 The coursework mark

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

Student performance in each honours course is assessed by a combination of assignments and examinations. The assignment component is determined by the lecturer of each course and the examination component makes up the balance to 100%.

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 average of all three marks awarded. A weighting of 50% is attached to the supervisor's original mark, while a weight of 25% is attached to each of the two marks awarded by the other examiners.

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

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

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

- Has the student shown initiative and hard work which are not superficially evident from the written report?
- Has the student coped well with a topic which is too broad or not clearly defined?

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. Students considering moving to Sydney for their studies should also be aware of the $6,000 Faculty of Science Honours Relocation Scholarships. The full list of scholarships including need-based ones is available.

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 (Australia) and here (USA).

14 Additional proposed project topics

1. **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.

2. **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.

3. **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.

4. **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

5. **Logspline-based estimation of a centre of symmetry**
Project supervisor: Dr Michael Stewart

Project description: The R package `logspline` implements a density estimation method due to Stone (1990); Kooperberg and Stone (1992); Stone *et al.* (1997) which uses cubic splines to approximate the logarithm of the density. The choice of number and location of knots is made in a data-driven manner, designed to make the estimation of the density itself as good as possible. Efficient semiparametric estimation of the centre of symmetry of a density involves first estimating the location score function (derivative of the log-density) and then solving the resultant "estimated" score equation. The logspline method can be used for this, but its performance can possibly be improved for this task by "tweaking" the algorithm used

to choose the number and location of spline knots. This project will have both a theoretical and computational component. The theory will be studied to determine how the algorithm may be adjusted to improve estimation of the derivative of the log-density under symmetry. A second aim of the project is to implement this improved algorithm and produce an efficient R package, possibly implementing the main routine in a fast low-level language, e.g. C, C++ or Fortran.

References

Charles Kooperberg and Charles J. Stone. Logspline density estimation for censored data. *Journal of Computational and Graphical Statistics*, 1(4):301–328, 1992. ISSN 10618600. URL http://www.jstor.org/stable/1390786.

Charles J. Stone. Large-sample inference for log-spline models. *Ann. Statist.*, 18(2):717–741, 1990. ISSN 0090-5364. URL https://doi.org/10.1214/aos/1176347622.

Charles J. Stone, Mark H. Hansen, Charles Kooperberg, and Young K. Truong. Polynomial splines and their tensor products in extended linear modeling. *Ann. Statist.*, 25(4):1371–1470, 1997. ISSN 0090-5364. URL https://doi.org/10.1214/aos/1031594728. With discussion and a rejoinder by the authors and Jianhua Z. Huang.

6. Testing for nonlinear cointergation
   Supervisor: Prof. Qiying Wang

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

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

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

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

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

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

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

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

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