|  |
| --- |
| UCL DEPARTMENT OF STATISTICAL SCIENCE |
| BSc PROJECTS GUIDE 2023 - 2024 |

List of Projects ii-

Project Descriptions Page 1

BSc Projects List 2023

[UCL Module Catalogue – STAT0035](https://www.ucl.ac.uk/module-catalogue/modules/project-STAT0035)

[UCL Module Catalogue – STAT0036](https://www.ucl.ac.uk/module-catalogue/modules/project-STAT0036)

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Submitting Supervisor(s)** | **Project Title** | **Weighting** |
|  | ABOURASHCHI, Dr Niloufar & RASSIAS, Dr Matina | Quantitative Methods and Applications in Finance 1 | 15/30 |
|  | ABOURASHCHI, Dr Niloufar & RASSIAS, Dr Matina | Quantitative Methods and Applications in Finance 2 | 15/30 |
|  | ABOURASHCHI, Dr Niloufar & RASSIAS, Dr Matina | Quantitative Methods and Applications in Finance 3 | 15/30 |
|  | ABOURASHCHI, Dr Niloufar & RASSIAS, Dr Matina | Quantitative Methods and Applications in Finance 4 | 15/30 |
|  | BAIO, Professor Gianluca | Bayesian Modelling of excess mortality following the COVID-19 pandemic | 30 |
|  | BAIO, Professor Gianluca | Predicting the number of bike shares in London | 30 |
|  | BARTLETT, Dr Tom | Modelling community structure in networks | 15/30 |
|  | BARTLETT, Dr Tom | Statistical learning of network structure from high dimensional data | 15/  330 |
|  | BESKOS, Dr Alexandros | Time-varying stochastic volatility modelling for financial time-series | 15/30 |
|  | BESKOS, Dr Alexandros | Hamiltonian dynamics for sampling posterior distributions | 15/30 |
|  | BRIOL, Dr Francois-Xavier | Bayesian quadrature for environmental applications | 30 |
|  | CHAKRAVARTI, Dr Purvasha | Interpretability of High-dimensional Classifiers | 15/30 |
|  | CHAKRAVARTI, Dr Purvasha | Two-Sample Testing Using High-dimensional Classifiers with an Application in High-Energy Physics | 15/30 |
|  | CHANDLER, Professor Richard | Statistical analysis of climate projections (multiple projects available) | 15/30 |
|  | CHANDLER, Professor Richard | Assessing hyper dominance in tropical forests | 30 |
|  | CHANDLER, Professor Richard | Other topics in environmental statistics | 15/30 |
|  | CIPRIANI, Dr Alessandra | Metal Microstructure through topological data analysis | 15/30 |
|  | CIPRIANI, Dr Alessandra | Curvature on graphs and a conjecture solved | 30 |
|  | COTAR, Professor Codina | Deep Hedging | 30 |
|  | COTAR, Professor Codina | Expected shortfall (ES) for different portfolios | 30 |
|  | COTAR, Professor Codina | Optimal transport to data science and statistics | 30 |
|  | DE IORIO, Professor Maria | Bayesian Shrinkage Priors | 15/30 |
|  | DELLAPORTAS, Professor Petros | Modelling wind energy | 15/30 |
|  | DONOV, Dr Alex | Applications of Copula Theory to portfolio Value-at-Risk estimation | 15/30 |
|  | DONOV, Dr Alex | Modelling the dependence structure of financial assets using vine copula | 15/30 |
|  | GRIFFIN, Professor Jim | Modelling individual athletic performance | 30 |
|  | GRIFFIN, Professor Jim | Variable Selection using the Weighted Likelihood Bootstrap | 15/30 |
|  | GRIFFIN, Professor Jim | Forecast combination | 15/30 |
|  | HERBOTS, Dr Hilde | Stochastic models in evolutionary genetics | 30 |
|  | HERNANDEZ, Dr Nicolas | Feature selection for two-sample test in High-dimensional data | 15/30 |
|  | HONNOR, Dr Tom | Analysis of spatial point pattern data 1 | 30 |
|  | HONNOR, Dr Tom | Analysis of spatial point pattern data 2 | 30 |
|  | JENDOUBI, Dr Takoua | Canonical correlation analysis (project 2) | 30 |
|  | JENDOUBI, Dr Takoua | A stable marriage algorithm | 30 |
|  | JONES, Dr Elinor | Analysis of education data | 15/30 |
|  | JONES, Dr Elinor | Analysing data from multiple studies when we can’t share the data | 30 |
|  | LEHMANN, Dr Brieuc  LEHM | Fairness considerations for conformal inferences | 15/30 |
|  | LEHMANN, Dr Brieuc | Fairness under data set shift | 15/30 |
|  | LEURENT, Mr Baptiste | Cluster-level vs. individual-level adjustment in cluster randomised trials | 30 |
|  | LEURENT, Mr Baptiste | correction for zero events in cluster randomised trials | 30 |
|  | LIVINGSTONE, Dr Samuel | Machine learning for the microbiome | 30 |
|  | LIVINGSTONE, Dr Samuel | Topics in convex optimization | 30 |
|  | LIVINGSTONE, Dr Samuel | Advanced gradient-based Markov chain Monte Carlo Methods | 30 |
|  | LO, Dr Chak Hei | Optimal Strategies for Games | 30 |
|  | LO, Dr Chak Hei | Intuitions on Extendable Options | 30 |
|  | MAIER, Dr Sebastian | Kernel-based simulation-and-regression approaches for multivariate option pricing | 15 |
|  | MAIER, Dr Sebastian | Statistical estimation of multinominal logit models using restaurant data | 15d |
|  | MANOLOPOULOU, Dr Ioanna | Statistical Choice Models | 15/30 |
|  | MANOLOPOULOU, Dr Ioanna | Modelling binary data observed with error | 15/30 |
|  | MARRA, Professor Giampiero | Improving robust estimation in flexible distributional models | 15/30 |
|  | MARRA, Professor Giampiero | Copula regression models | 15/30 |
|  | MARRA, Professor Giampiero | Flexible Parametric Regression Spline Approaches to Survival Modelling | 15/30 |
|  | NORTHROP, Dr Paul | Creating an R package | 30 |
|  | NORTHROP, Dr Paul | Inference for time series extremes | 15/30 |
|  | PAVLOU, Dr Menelaos | Sample size for the development of risk models for binary and time to event outcomes | 30 |
|  | POKERN, Dr Yvo | Heteroscedastic Error Models for Electron-Nuclear Double Resonance (ENDOR) Spectra | 30 |
|  | POKERN, Dr Yvo | Bayesian Linear Inverse Problems for De-Convolution of Spectra | 30 |
|  | POKERN, Dr Yvo | Piecewise Deterministic Samplers for Covariance Matrices | 30 |
|  | POKERN, Dr Yvo | Statistical Modelling of Real Inductor Impedances | 15/30 |
|  | POKERN, Dr Yvo | Characterizing RF attenuators and detectors | 15/30 |
|  | RUBIO ALVAREZ, Dr Javier | Practical non-identifiability of parametric models | 30 |
|  | RUBIO ALVAREZ, Dr Javier | Bayesian analysis of population growth models | 30 |
|  | SADEGHI, Dr Kayvan | Asymptotics of Discrete Determinantal Point Processes | 15/30 |
|  | SADEGHI, Dr Kayvan | Review of Structure Learning Algorithms for Causal Inference | 15/30 |
|  | SELCUK, Dr Cemil | Traffic Flow Maps | 15/30 |
|  | SELCUK, Dr Cemil | Seasonality in the Housing Market | 15/30 |
|  | SIMPSON, Dr Emma | Statistical analysis of air pollution | 30 |
|  | SIMPSON, Dr Emma | Modelling extreme sea surface temperatures | 30 |
|  | SIMPSON, Dr Emma | Conditional modelling of extreme river flow data | 30 |
|  | VAN DEN HOUT, Dr Ardo | Multi-state survival models: expected time spent in states | 15/30 |
|  | VAN DEN HOUT, Dr Ardo | Joint models for position and time in music chart | 30 |
|  | WATSON, Dr Alex | Models of Cell growth and division | 30 |
|  | WATSON, Dr Alex | Optimisation, dynamic programming and reinforcement learning | 30 |
|  | XUE, Professor Jinghao | Sparse principal component analysis and its applications | 30 |
|  | XUE, Professor Jinghao | Sparse canonical correlation analysis and its applications | 30 |
|  | XUE, Professor Jinghao | Sparse Linear discriminant analysis and its applications | 30 |
|  | XUE, Professor Jinghao | Principal component regression, ridge regression, and partial least squares | 30 |

BSc Project Descriptions for 2023

**Title:** Quantitative Methods and applications in Finance (Projects x4)

**Supervisors:** Dr Niloufar Abourashchi and Dr Matina Rassias

**Credit Value:** 15/30

**Description:** The implementation of quantitative actuarial models is a vital task to assess risk in insurance, finance and other industries and professions. This project will introduce both theoretical and practical implementation of various quantitative modelling techniques applicable to finance and insurance. Students will have the opportunity to combine quantitative disciplines such as probability/statistics and actuarial science/quantitative finance.

The aim of this project is to provide a grounding in stochastic processes and survival models and their applications. This could be suitable for students who would like to continue their studies/career in financial mathematics and/or are aiming to take any finance related professional exam such as actuarial exams.

**Title:**  Bayesian Modelling of excess mortality following the COVID-19 Pandemic

**Supervisor:** Professor Gianluca Baio

**Credit Value:** 30

**Description:** Requires skills in R and will require some learning on Bayesian modelling. The project can be split into a couple of different sub-projects that different students can take on at the same moment.

**Title:**  Predicting the number of bike shares in London

**Supervisor:** Professor Gianluca Baio

**Credit Value:** 30

**Description:** We'll use data made publicly available from TFL to estimate the relationship between the number of bikes shared around the London network on a given day, depending on weather and other characteristics to predict the capacity needed to satisfy demand at any given point. Requires R and familiarity with non-linear regression models.

**Title:** Modelling community structure in networks

**Supervisor:** Dr Tom Bartlett

**Credit value:** 15-30

**Description:** Networks and other non-Euclidean relational datasets have become important applications in modern statistics [1,2]. Much of the attention in this area has been on parametric models for network data, such as degree-based models, and community-based alternatives. One of the most widely studied of these models is the stochastic block-model (SBM), in which, under the assortative assumption, there is a greater probability of observing an edge (or interaction) between a pair of nodes (or individuals) if they are in the same block, or community. The problem of finding communities in social and biological networks has been studied for many years since Newman’s early work in this area [3], and real-life applications of this problem include identifying groups of friends in a social network and identifying functional subnetwork modules in a biological network. In this project, we will explore probabilistic models of networks such as the SBM and extensions, and we will examine real-world applications of these models.

**References**

[1] Crane H. Probabilistic foundations of statistical network analysis. CRC Press; 2018

[2] Kolaczyk ED, Csárdi G. Statistical analysis of network data with R. Springer; 2014

[3] Girvan M, Newman ME. Community structure in social and biological networks. Proceedings of the national academy of sciences. 2002;99(12):7821-6.

**Prerequisites**

Some experience of computational statistics using R, and in interest in network models and/or applications.

**Title:** Statistical learning of network structure from high-dimensional data

**Supervisor:** Dr Tom Bartlett

**Credit value:** 15-30

**Description:** Statistical network models are an effective tool for analysing and interpreting high-dimensional data, i.e., simultaneous observations of a large number of variables. There are a range of competing methods for statistical network inference using data of this type, including regression and mutual information-based methods (see for example [1, 2]). However, whilst these methods produce results that validate well at the level of network structures (i.e., aggregations of network edges), they often suffer from a high false-positive rate when detecting individual network edges. Strategies to overcome this include integrating data-sets from complementary sources, such as alternative data modalities. In this project, we will analyse a range of network inference methods, to understand their accuracy at the level of individual network edges. Following this, we will investigate how to adapt these methods appropriately for timely problems in the scientific domain.

**References**

[1] Dobra A, Hans C, Jones B, Nevins JR, Yao G, West M. Sparse graphical models for exploring gene expression data. Journal of Multivariate Analysis. 2004;90(1):196-212.

[2] Chan TE, Stumpf MP, Babtie AC. Gene regulatory network inference from single-cell data using multivariate information measures. Cell systems. 2017;5(3):251-67.

**Prerequisites**

Some knowledge and experience of computational statistics in R or Python.

**Title:** Time-varying stochastic volatility modelling for financial time-series

**Supervisor:** Dr Alexandros Beskos

**Credit Value:** 15/30

**Description:** The project will examine the literature for stochastic volatility models, both in discrete and continuous-time, and summarise earlier and current model contributions in the field. It will investigate model fitting approaches (e.g., by exploring statistical packages in R or otherwise) and model performance. The project will use datasets available at online libraries.

**Title:** Hamiltonian dynamics for sampling posterior distributions

**Supervisor:** Dr Alexandros Beskos

**Credit Value:** 15/30

**Description:** Hamiltonian equations arose in physics and model the evolution of a system where energy is preserved. Rather surprisingly, such dynamics have also found use within Bayesian inference, under the name of Hybrid Monte-Carlo (HMC). HMC can efficiently generate sampler from complex posterior distributions. The student will investigate the performance of HMC for involved hierarchical models and explore the use of ready-made statistical packages (e.g., STAN).

**Title:** Bayesian quadrature for environmental applications

**Supervisor:** Dr Francois-Xavier Briol

**Credit Value:** 30

**Description:** Bayesian quadrature is a widely applicable set of algorithms for computing expensive integrals of interest. It been applied to fields ranging from the environmental sciences to computer graphics. In this project, the student will review the state-of-the-art in this research field and perform an extensive simulation study comparing the leading set of algorithms on a test of benchmark problems.

[1] Briol, F.-X., Oates, C. J., Girolami, M., Osborne, M. A., & Sejdinovic, D. (2019). Probabilistic integration: A role in statistical computation? (with discussion). *Statistical Science*, *34*(1), 1–22.

**Prerequisite:** linear algebra and good programming ability in, for example, R, Matlab or Python.

**Title:** Interpretability of High-dimensional Classifiers

**Supervisor:** Dr Purvasha Chakravarti

**Credit Value:** 15/30

**Description:** High-dimensional classifiers are often used to make important decisions in medical science, high-energy physics, business research, and in many other areas. One example is using non-linear classification tools, such as Random Forests or Neural Networks, to classify whether a patient has a high or low risk of heart disease. While the accuracy of such classifiers can be high, understanding how a black-box classifier determines the classification of a datapoint can be challenging. Recently there have been advances in the interpretability of such classifiers via active subspace methods and Shapley values. In this project, the student will be introduced to the different interpretability methods, such as Mean Decrease Accuracy (MDA), Shapley values, and the Active Subspace technique. The methods serve different purposes as MDA and Shapley values focus on the predictive power of individual features and not on relationships between features that may have greater predictive power, while the Active Subspace technique attempts identifying relationships between features that influence the classifier. So, in the project, we will further compare the performance of the methods on different simulated data. There is also potential for these methods to be applied to real data to study if there are any potential biases in existing algorithms.

**Title:** Two-Sample Testing Using High-dimensional Classifiers with an Application in High-Energy Physics

**Supervisor:** Dr Purvasha Chakravarti

**Credit Value:** 15/30

**Description:** In the last two decades, it has become increasingly common to use classifiers for two-sample testing, i.e., to test if two samples are generated from the same distribution or not. This is because classifiers have been found to be very powerful at finding differences between samples, if they exist, especially in high-dimensional data. The idea is to assess whether there is a significant difference between the two sets of data by learning a classifier to differentiate between them and then testing if the performance of the classifier is significantly better than chance. In this project, the student will first be introduced to some state-of-the-art kernel two-sample testing methods and the two-sample tests based on the classifiers. This will include reviewing the different statistics that can be used to measure the performance of the classifiers. They will then conduct an extensive simulation analysis, comparing the classifier tests to the kernel two-sample testing methods. If the student is interested, these methods can further be applied to real-world high-energy physics datasets provided by the Large Hadron Collider at CERN.

**Title**: Statistical analysis of climate projections (multiple projects available)

**Supervisor**: Professor Richard Chandler

**Credit value**: 15/30

**Description**: there is international agreement that action is needed to limit further changes to the earth’s climate, and to mitigate the effects of changes that are already inevitable. In the first instance, such action must be informed by an understanding of how the climate may evolve under different scenarios of global social and economic development, coupled with the associated greenhouse gas emissions. The required information comes predominantly from vast archives of future climate projections, produced using a variety of different climate or “earth system” models representing the main physical, chemical, biological and geological processes that influence the climate. Statistical tools are needed to make sense of the resulting data archives, however – for example, how do you even visualise the data from more than 60 climate models, each producing daily time series at 10km spatial resolution over the whole of Europe (or the entire globe)? Similarly, if all of the models produce different projections of future climate, and all of them are known to have strengths and weaknesses, how should we combine the resulting information to obtain a “best” estimate – and, crucially for decision-making and policy development, to quantify the uncertainty in that estimate? Recent research at UCL has produced a variety of statistical tools to address these and other questions, which offer plenty of opportunities for BSc project topics (see <https://www.ucl.ac.uk/statistics/research/eurocordex-uk> for example). I am open to suggestions from students who would like to develop their own project proposals in this area, to suit their own interests and skills.

**Title**:    Assessing hyper dominance in tropical forests

**Supervisor**:   Prof Richard Chandler

**Credit value**: 30

**Description**: It is increasingly recognised that humans need to take urgent action to preserve the resources of the planet upon which we depend. Preserving biodiversity is one aspect of this, to avoid unintended consequences arising from a rapid shift in the balance of ecosystems that have evolved gradually over long time periods. A major challenge for conservation efforts and policy development is that we often do not have a good understanding of current ecosystems. In tropical forests for example, we don’t know how many tree species there are or how they are distributed – although it is generally accepted that most forests are dominated by a very small proportion of the total species. Understanding this “hyper dominance” is an important question for tropical forest ecology,  although it raises some difficult statistical issues - mainly because the data on numbers of individuals of different species are incomplete (by definition, many of the rarer species have not been sampled, which introduces a bias in the data that must be accounted for in any analysis). Recent work, in collaboration with colleagues

in UCL Geography, has proposed a new statistical methodology for addressing the problem. It is of interest to know whether the results of this methodology would change substantially under different sets of assumptions. A STAT0035 project in 2022/23 examined a part of the methodology and found that the results from this part did not change much under different assumptions. The present project aims to build on that work to examine other parts of the methodology. Excellent computational and mathematical skills will be needed.

**Title**: Other topics in environmental statistics

**Supervisor**: Professor Richard Chandler

**Credit value**: 15/30

**Description**: This project option is for students who have an interest in some aspect of the environment and who would like to apply their statistical skills in this area. Please contact Professor Chandler directly to discuss possibilities.

**Title:** Metal Microstructure through topological data analysis

**Supervisor:** Dr Alessandra Cipriani

**Credit Value:** 15/30

**Description:** Imagine you are looking for a robust, stable material like steel for a manufacturing company. Even if a bar of metal looks solid on the outside, its microscopic interior, or microstructure, is what actually determines the qualities you are looking for, like bendability, resistance to temperature and conductivity. Most metals analyzed at a tiny scale are formed by grains, also called cells. Grains can be described as polygonal shapes delimited by some lines called grain boundaries. Material scientists are keen on understanding the mathematical properties of grain cells: thanks to this modelling, they would be able to retrieve all necessary information on the mechanical features of the material.

What we will be concerned with in this thesis is the size of the grains, an important factor which for example influences the behaviour under external stress. In practice observing the microstructure of a 3D material is very costly, so scientists have to rely only on 2D cross-sections of the metal to infer its 3D volume distribution. How can this be done rigorously?

In this thesis, we will familiarize ourselves with the most common mathematical models for cell displacement such as PoissonVoronoi and VoronoiLaguerre diagrams and introduce the basic notions of topological data analysis (TDA) that are used to model microstructures. Among them, we will talk about simplicial complexes, persistence diagrams and persistence barcodes.

**References**

[1] Adler, R. J., Bobrowski, O., Borman, M. S., Subag, E., & Weinberger, S. (2010). Persistent homology for random elds and complexes. Borrowing Strength: Theory Powering Applications A Festschrift for Lawrence D. Brown. Institute of Mathematical Statistics. doi: 10.1214/10-IMSCOLL609.

[2] Chazal, F., & Michel, B. (2017). An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists. arXiv, 1710.04019. Retrieved from https://arxiv.org/abs/1710.04019v2.

**Title:** Curvature on graphs and a conjecture solved.

**Supervisor:** Dr Alessandra Cipriani

**Credit Value:** 30

**Description:** For continuum spaces curvature is a well-understood object that roughly describes how much space is "bending". It is a very useful concept not only in physics, where for example it is fundamental in the theory of relativity, but also in image processing, biology, engineering etc. Its definition cannot be clearly translated to graphs, and there have been many attempts to set a universal, coherent notion of curvature therein. In this project we will analyze different types of curvature on graphs, in particular by seeing hyperbolic graphs (like those in M.C. Escher's paintings "[Circle limit I-IV](https://eur01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fen.wikipedia.org%2Fwiki%2FCircle_Limit_III&data=05%7C01%7Ck.leport%40ucl.ac.uk%7Cb52236853cab442cd1f108db34171723%7C1faf88fea9984c5b93c9210a11d9a5c2%7C0%7C0%7C638161045989323806%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=cOfFfq1l4KubsHz1H20cHdtDvDq07I38u%2BjGRU8bUss%3D&reserved=0)"). We will also analyze a tool to compute the curvature of graphs by Stagg and Cushing.

References: [Curvature calculator](https://arxiv.org/abs/1712.03033), [curvature of graphs presentation](https://eur01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fwww.maths.dur.ac.uk%2Fusers%2Fnorbert.peyerimhoff%2Fepsrc2013%2Fworkshop%2Fjost-juergen.pdf&data=05%7C01%7Ck.leport%40ucl.ac.uk%7Cb52236853cab442cd1f108db34171723%7C1faf88fea9984c5b93c9210a11d9a5c2%7C0%7C0%7C638161045989480793%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=yPLNRywAH5%2FolKB6bDIfUiUdzWYs4XdkDkf5dkDapf0%3D&reserved=0)

Prerequisites: a course in Probability, knowledge in Stochastic processes/Markov chains is not essential but helpful.

**Title:**"Deep hedging"

**Supervisor:** Professor Codina Cotar

**Credit Value:** 30

**Description:** This thesis examines the theory of deep hedging and highlights its relevance

by overcoming the limitations of the famous Black-Scholes model. This newly introduced method for hedging derivatives by deep learning is a model independent approach, which is able to account for market frictions and trading constraints. To analyse its practicality, deep hedging is applied to real market data. As current research mostly explores deep hedging by testing it on simulated data, this research tries to verify the performance of deep hedging on real options written on the S&P

500. The deep learning strategies are compared to daily rebalanced Black-Scholes delta hedges.

Some beginner literature:

[https://www.risk.net/derivatives/7956404/jp-morgan-is-testing-quantum-deep-hedging](https://eur01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fwww.risk.net%2Fderivatives%2F7956404%2Fjp-morgan-is-testing-quantum-deep-hedging&data=05%7C01%7Ck.leport%40ucl.ac.uk%7C133a2cca1872434de85708db7bb84be2%7C1faf88fea9984c5b93c9210a11d9a5c2%7C0%7C0%7C638239803694358689%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=5fbfg6fc9NgTarXkdz7LGP1Lt9P9kZQImjRaia6DMz4%3D&reserved=0)

[https://www.risk.net/derivatives/6875321/deep-hedging-and-the-end-of-the-black-scholes-era](https://eur01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fwww.risk.net%2Fderivatives%2F6875321%2Fdeep-hedging-and-the-end-of-the-black-scholes-era&data=05%7C01%7Ck.leport%40ucl.ac.uk%7C133a2cca1872434de85708db7bb84be2%7C1faf88fea9984c5b93c9210a11d9a5c2%7C0%7C0%7C638239803694358689%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=igXqgfpiU5TPWY1%2BB3jVpJva1YrmC5ZV3fyvJ9VdFtw%3D&reserved=0)

<https://www.risk.net/cutting-edge/views/7936936/podcast-hans-buehler-on-the-data-science-behind-deep-hedging>

[http://www.deephedging.com/](https://eur01.safelinks.protection.outlook.com/?url=http%3A%2F%2Fwww.deephedging.com%2F&data=05%7C01%7Ck.leport%40ucl.ac.uk%7C133a2cca1872434de85708db7bb84be2%7C1faf88fea9984c5b93c9210a11d9a5c2%7C0%7C0%7C638239803694358689%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=2N1nXT2%2FeB0Umw4HTiqKRC2LMH7F6PgFB%2BZekfC3nFc%3D&reserved=0)

[https://www.maths.ox.ac.uk/system/files/attachments/2019%2004%2024%20Deep%20Hedging%20Frontiers%20Imperial%202.1.pdf](https://eur01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fwww.maths.ox.ac.uk%2Fsystem%2Ffiles%2Fattachments%2F2019%252004%252024%2520Deep%2520Hedging%2520Frontiers%2520Imperial%25202.1.pdf&data=05%7C01%7Ck.leport%40ucl.ac.uk%7C133a2cca1872434de85708db7bb84be2%7C1faf88fea9984c5b93c9210a11d9a5c2%7C0%7C0%7C638239803694358689%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=e0gtyY3%2FHajaveM5wjhuVmu27SUgPqnMNkOECBxmfO4%3D&reserved=0)

<https://arxiv.org/abs/1802.03042>

<https://arxiv.org/pdf/2103.16409.pdf>

**Title:** "Expected shortfall (ES) for different portfolios"

**Supervisor:** Professor Codina Cotar

**Credit Value:** 30

**Description:**  It is a well-known fact that value-at-risk2 (VaR) models do not perform well under market stress. VaR models are usually based on normal asset returns and do not work under extreme price fluctuations.

Our question is this: Is this a problem of the estimation methods, or of VaR as a risk measure? The estimation methods used for standard VaR models have problems for measuring extreme price movements. They assume that the asset returns follow a normal distribution. So they disregard the fat-tailed properties of actual returns and underestimate the likelihood of extreme price movements.

On the other hand, the concept of VaR as a risk measure has problems for measuring extreme price movements. By definition, VaR only measures the distribution quantile, and disregards extreme loss beyond the VaR level. Thus, VaR may ignore important information regarding the tails of the underlying distributions.

To alleviate the problems inherent in VaR, Artzner et al (1997, 1999) proposed the use of expected shortfall. Expected shortfall is the conditional expectation of loss given that the loss is beyond the VaR level. Thus, by definition, expected shortfall considers loss beyond the VaR level.

In this thesis, we examine whether the tail risk of VaR and expected shortfall is actually significant under market stress, and we examine whether ES performs better than ES in different conditions.

Some beginner literature

[https://www.risk.net/risk-magazine/technical-paper/1506669/var-versus-expected-shortfall](https://eur01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fwww.risk.net%2Frisk-magazine%2Ftechnical-paper%2F1506669%2Fvar-versus-expected-shortfall&data=05%7C01%7Ck.leport%40ucl.ac.uk%7C133a2cca1872434de85708db7bb84be2%7C1faf88fea9984c5b93c9210a11d9a5c2%7C0%7C0%7C638239803694358689%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=EnG5efhssvDtmnpigULsL3YBSgnuCmtwu7YuhQ5NhVM%3D&reserved=0)

[https://www.risk.net/risk-management/market-risk/2375185/hull-and-white-on-the-pros-and-cons-of-expected-shortfall](https://eur01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fwww.risk.net%2Frisk-management%2Fmarket-risk%2F2375185%2Fhull-and-white-on-the-pros-and-cons-of-expected-shortfall&data=05%7C01%7Ck.leport%40ucl.ac.uk%7C133a2cca1872434de85708db7bb84be2%7C1faf88fea9984c5b93c9210a11d9a5c2%7C0%7C0%7C638239803694358689%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=ByaMJ5vgYHxVROK1XzwdkIVf3E9%2Bk0Ov5AHSHu0qps0%3D&reserved=0)

[https://www.finalyse.com/blog/var-an-introductory-guide-in-the-context-of-frtb](https://eur01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fwww.finalyse.com%2Fblog%2Fvar-an-introductory-guide-in-the-context-of-frtb&data=05%7C01%7Ck.leport%40ucl.ac.uk%7C133a2cca1872434de85708db7bb84be2%7C1faf88fea9984c5b93c9210a11d9a5c2%7C0%7C0%7C638239803694671128%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=Rh7gw42S0f%2Bb4ch8nNyN3ZZHfRWbJsJhrziTj72zau4%3D&reserved=0)

[https://www.finextra.com/blogposting/11678/frtb-replacing-var-with-expected-shortfall-in-market-risk](https://eur01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fwww.finextra.com%2Fblogposting%2F11678%2Ffrtb-replacing-var-with-expected-shortfall-in-market-risk&data=05%7C01%7Ck.leport%40ucl.ac.uk%7C133a2cca1872434de85708db7bb84be2%7C1faf88fea9984c5b93c9210a11d9a5c2%7C0%7C0%7C638239803694671128%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=167gpgsN5OzWAhF9kokfNR1N9Mq60qsIDKRM0WuFONs%3D&reserved=0)

**Title:** Optimal transport to data science and statistics

**Supervisor:** Professor Codina Cotar

**Credit Value:** 30

**Description:**In the last 30 years, the theory of Optimal Transport has emerged as a fertile field of inquiry, and a diverse tool for exploring applications within and beyond mathematics, in such diverse fields as economics, meteorology, geometry, fluid mechanics, design problems, theoretical chemistry, and engineering. More recently, due to unexpected connections, for instance with Data Sciences and Statistical Inference, theoretical and computational aspects of Optimal Transport regained substantial attention from both theorists and practitioners.

Optimal Transport principles have been applied very recently in formulating solutions to problems in the area of statistical inference, and numerous machine learning problems such as generative learning, transfer learning,  distributionally robust optimization, and so on, with impressive results.

From a statistical point of view, transportation distances are very appealing since they quantify in a natural and meaningful way the notion of perturbations of a probability distribution (so are particularly suitable when considering mixture distributions). From a machine learning point of view, Wasserstein distances provide a very good way to compare two instances, by quantifying how much one needs to “warp” one instance to reach the other (for example, to measure similarities between images). Moreover, Optimal Transport methods have had a significant impact in the theory of nonlinear PDEs, in metric geometry, and in the development of functional inequalities.

In this project, we will explore some of the applications of optimal transport to statistics

Beginner literature:

[https://www.birs.ca/cmo-workshops/2017/17w5093/report17w5093.pdf](https://eur01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fwww.birs.ca%2Fcmo-workshops%2F2017%2F17w5093%2Freport17w5093.pdf&data=05%7C01%7Ck.leport%40ucl.ac.uk%7C133a2cca1872434de85708db7bb84be2%7C1faf88fea9984c5b93c9210a11d9a5c2%7C0%7C0%7C638239803694671128%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=IRMHCW%2Btd1oj3UVANOnVtCcNuQXLQJwD7DMeNF4u7kU%3D&reserved=0)

[https://www.stat.cmu.edu/~larry/=sml/Opt.pdf](https://eur01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fwww.stat.cmu.edu%2F~larry%2F%3Dsml%2FOpt.pdf&data=05%7C01%7Ck.leport%40ucl.ac.uk%7C133a2cca1872434de85708db7bb84be2%7C1faf88fea9984c5b93c9210a11d9a5c2%7C0%7C0%7C638239803694671128%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=WdalcJzkUNduyHkM5hrDZQ2kdHp3kWPLVom2VLJ6mkY%3D&reserved=0)

**Title:** Bayesian Shrinkage Priors

**Supervisor:** Professor Maria De Iorio

**Credit Value:** 15/30 credits

**Description:** In recent years, a rich variety of shrinkage priors have been proposed that have great promise in addressing massive regression problems. In general, these new priors are special cases of three wide classes of probability distributions: Gauss hypergeometric distribution, the compound confluent hypergeometric and the generalized hyperbolic distributions.  These project aims to provide a general overview of Bayesian shrinkage priors, by positioning them in these larger classes. This encompassing framework should prove useful in comparing the relative merits of the competing priors, considering properties and revealing close connections.

Prerequisites: Probability, Knowledge at least of R.

**Title:** Modelling wind energy

**Supervisor:** Professor Petros Dellaportas

**Credit Value:** 15/30

**Description:**Modelling data from wind turbines is very important for various reasons. Prediction of energy output in short, medium, and long terms is required for both regulatory and commercial reasons. Moreover, fault detections are a crucial aspect of wind park management.  There is a series of unanswered statistical problems that emerge from this energy application.   Depending on what the student wishes, the project may involve statistical tools such as Gaussian processes or more machine learning techniques such as deep neural networks and autoencoders.

**Title:** Applications of Copula theory to portfolio Value-at-Risk estimation.

**Supervisor:** Dr Alex Donov

**Credit:** 15/30

**Description:** Value-at-Risk (VaR) is widely used by financial institutions to manage their market risk. To correctly assess the possible losses of a portfolio over a certain horizon, an accurate measurement of the dependence structure between asset returns in the portfolio is essential. The aim of this project is to incorporate copula theory in order to describe the dependence structure of asset returns, and subsequently estimate the VaR.  The required financial data is widely available online. There will be an opportunity to develop R Shiny app.

**Prerequisites**: A good understanding of probability and statistics (STAT0005 or equivalent) and good programming skills in either MATLAB or R (STAT0023 or equivalent) are essential. Prior knowledge of copula is not necessary as you will be introduced to the basic level required for the analysis. Additional relevant modules are STAT0010 and STAT0011/38.

**Introductory reading:**

1. Meucci, Attilio, A Short, Comprehensive, Practical Guide to Copulas (2011). Available at SSRN: [https://ssrn.com/abstract=1847864](https://eur01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fssrn.com%2Fabstract%3D1847864&data=05%7C01%7Ck.leport%40ucl.ac.uk%7Cebe667ce1c864bd8e59008db61a4b81e%7C1faf88fea9984c5b93c9210a11d9a5c2%7C0%7C0%7C638211132611479577%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=0Hm%2FAkY4ABD7mV2SkQX61q5wsuHyKY%2BlEHMTReJyNTs%3D&reserved=0)

2. Jorion, Philippe, Value at Risk: The New Benchmark for Managing Financial Risk (2000). Available from the UCL Library

**Title:** Modelling the dependence structure of financial assets using vine copula.

**Supervisor:** Dr Alex Donov

**Credits:** 15/30

**Description:** The objective of this study is to apply regular vine copula to model the dependence structure of a sample of financial stocks. Vine copulas offer greater flexibility and enable the modelling of asymmetric and complex dependency structures using the rich variety of bivariate copulas. The required financial data is widely available online.  There will be an opportunity to develop R Shiny app.

**Prerequisites**: A good understanding of probability and statistics (STAT0005 or equivalent) and good programming skills in either MATLAB or R (STAT0023 or equivalent) are essential. Prior knowledge of copula is not necessary as you will be introduced to the basic level required for the analysis. Additional relevant modules are STAT0010 and STAT0011/38.

**Introductory reading:**

1.   Meucci, Attilio, A Short, Comprehensive, Practical Guide to Copulas (2011). Available at SSRN: [https://ssrn.com/abstract=1847864](https://eur01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fssrn.com%2Fabstract%3D1847864&data=05%7C01%7Ck.leport%40ucl.ac.uk%7Cebe667ce1c864bd8e59008db61a4b81e%7C1faf88fea9984c5b93c9210a11d9a5c2%7C0%7C0%7C638211132611479577%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=0Hm%2FAkY4ABD7mV2SkQX61q5wsuHyKY%2BlEHMTReJyNTs%3D&reserved=0)

2.   Aas, Kjersti, *et al*., Pair-copula constructions of multiple dependence. Available from:  [https://epub.ub.uni-muenchen.de/1855/1/paper\_487.pdf](https://eur01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fepub.ub.uni-muenchen.de%2F1855%2F1%2Fpaper_487.pdf&data=05%7C01%7Ck.leport%40ucl.ac.uk%7Cebe667ce1c864bd8e59008db61a4b81e%7C1faf88fea9984c5b93c9210a11d9a5c2%7C0%7C0%7C638211132611479577%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=IiuhgKuzEe2QOkK9vKlSL9QXlh%2Fn8hM7CUByikE1RxM%3D&reserved=0)

**Title:** Modelling individual athletic performance

**Supervisor:** Professor Jim Griffin

**Credit Value:** 30

**Description:** Understanding possible patterns of evolution in elite athletic performance is an important problem. Two directions which could be investigate in this project would be:

 identification of talented youngsters who will develop to compete at a high-level or the identification of athletes whose development is unusual and may indicate the use of performance-enhancing drugs. This project will look at addressing one of these topics using model of season's bests for a sample of athletes in sports such as track and field, swimming, or weightlifting  and using mixed effects models.

**Title:** Variable Selection using the Weighted Likelihood Bootstrap

**Supervisor:** Professor Jim Griffin

**Credit Value:** 30

**Description:** Variable selection in regression models with many possible regressors is a challenging problem. Bayesian methods have proved to be one successful method. However, these can be sensitive to the set-up of the regression model. This has lead to interest in methods where variable selection is less dependent on the choice of model. In this project, you will look at one approach, the Weighted Likelihood Bootstrap, and consider the performance of different methods within this approach at achieving robustness. The project will involve some R programming and explore methodological and practical aspects of the approach.

**Title:** Forecast combination

**Supervisor:** Professor Jim Griffin

**Credit Value:** 15/30

**Description:** Forecasts are often made about economic variables (such as the level of inflation) by experts or using models. Clearly, these forecasts can be subject to errors and this has led to interest in methods which combine the forecasts to give more accurate predictions. This project will look at these methods starting with point forecasts where a single "best" forecasts is to be made and possibly looking at density forecasts where the forecast is given as distribution over possible outcomes. The project will use R and involve some programming.

**Title:**Stochastic models in evolutionary genetics

**Supervisor:** Dr Hilde Herbots

**Credit value:**30 credits

**Description:** This project is concerned with a recently developed maximum-likelihood method which uses DNA sequence data from closely related species to make inferences about the process and timing of speciation. The method uses one pair of DNA sequences at each of a large number of different loci (different positions in the genome) and currently assumes that the relative mutation rates of these different loci are known constants, whereas in reality these relative mutation rates are typically estimated using outgroup sequences (DNA sequences from other, more distantly related species). This project will examine what effect this violation of the assumptions of the method has on the accuracy of the parameter estimates obtained, and on the results of model selection procedures. The project will involve some mathematical work and a substantial amount of computing.

Prerequisites: A good understanding of continuous-time Markov Chains (STAT0007 or equivalent) is essential, as are good programming skills in R (STAT0023 or equivalent). A good understanding of Probability and Inference (STAT0005 or equivalent) is also desirable. The 3rd year courses STAT0008 (Statistical Inference) and STAT0009 (Stochastic Systems) may also be helpful but are not required. Prior knowledge of genetics is NOT necessary.

Introductory reading:  J.Wakeley: "Coalescent Theory: An Introduction" (Roberts & Co., 2009).

**Title:** Feature selection for two-sample test in High-dimensional data

**Supervisor:** Dr Nicolás Hernández

**Suitability:** Bsc Data Science, BSc Statistics

**Credit Value:** 15/30

**Description:** The main objective of this project is to study de behaviour of the power of a Hypothesis test through variable selection in a high-dimensional context. In particular, we will investigate whether the power of a two-sample test could be boosted selecting groups of variables.

We will consider a testing procedure for testing the null hypothesis that two samples of curves observed at discrete grids (functional data) have the same underlying distribution. In this very high-dimensional context, reducing the dimension is of great relevance. To this aim, we will perform grouped variable selection using a maximum divergence criterion. The project will include an extensive simulation study and an empirical application using electrocardiogram signals.

**Pre-requisites:** Statistical Inference, Multivariate Analysis, Basic programming in R

**Title**: Analysis of spatial point pattern data 1

**Supervisor**: Dr Tom Honnor

**Credit Value**: 30 Credits

**Description**: Data doesn’t have to be limited to numbers in a spreadsheet but can include locations (points) which may be plotted to visualise by eye and analysed via a variety of theoretical and computational techniques. Interest is often in modelling the spatial locations of points (do they prefer/avoid particular areas of the space over which they are distributed) and dependencies between point locations (do they form together into clusters/spread out to avoid each other).

To give examples of the data sets and problems I’ve worked with students on in the past:

* Public health – Given the home locations of throat cancer patients and the location of an industrial incinerator is there any evidence of an increase in cancer cases around the incinerator, which could indicate that the incinerator may have a causal link to the cancers.
* Supermarkets – Combining locations of supermarkets with census information on local average incomes, population densities etc. to model which covariates are related to the spatial distribution of supermarkets and how, which could indicate the decision-making processes followed by company managers. Additionally, investigating dependency between locations of supermarkets from different retailers, which could shed light on the competition strategy of the retailer.
* Forest fires – Combining locations of forest fires with information on their causes and severities, and information on the geography of the region is there any evidence that forest fires are more/less likely in particular areas, which might be used to prioritise areas for preventative efforts or mitigation.
* Crimes – Given the locations of recorded crimes is there any evidence of hotspots, which might be used by the police to target their resources. Additionally, given the dates, times and types of crimes are there any patterns in the locations and/or timings of particular types of crimes, which might also be used by the police to target their resources.

In the past similar projects have included a balance between theory, which I don’t expect that you’ll have seen before but is in many ways an extension of things you have seen before, and practical computing in R, the vast

majority of which involves becoming familiar with the spatstat package.

Please contact me at [t.honnor@ucl.ac.uk](mailto:t.honnor@ucl.ac.uk) to discuss this topic further and see if we can come up with a particular problem and data set to suit your interests.

**Title**: Analysis of spatial point pattern data 2

**Supervisor**: Dr Tom Honnor

**Credit Value**: 30

**Description**: Please see the previous description for “Analysis of spatial point pattern data 1”. I’ve had success in the past with students working in related areas on different problems/data sets and am happy to flexible with the specific focus.

Please contact me at [t.honnor@ucl.ac.uk](mailto:t.honnor@ucl.ac.uk) to discuss this topic further and see if we can come up with a particular problem and data set to suit your interests.

**Title:** Canonical correlation analysis (2 projects).

**Supervisor:** Dr Takoua Jendoubi

**Credit Value:** 30

**Description:** Canonical correlation analysis (CCA) is a classic statistical tool for investigating complex multivariate data. Correspondingly, it has found many diverse applications, ranging from molecular biology and medicine to social science and finance. In canonical correlation analysis the aim is to find mutually orthogonal pairs of maximally correlated linear combinations of the components of two data matrices. Its standard computation is based on sample covariance matrices, which are however very sensitive to outlying observations. The obvious way for robustifying CCA is, hence, to robustly estimate the covariance matrix. Robust covariance matrix estimators can routinely be used in multivariate statistics. This project aims to study robust probabilistic models for canonical correlation analysis that can cope with atypical observations in the data. Robustness from a probabilistic point of view can be introduced by using specific distributions such as Student’s t-distribution with wide tails. During the project, we will be looking into both low and high dimensional data. High dimensionality is an additional challenge that will be addressed by means of regularization procedures by imposing specific penalty functions. The methodology can be split into two steps. The first step will be to make use of available probabilistic models in the literature to derive a robust probabilistic model for CCA [1, 2]. Then this model can be further adapted in the high dimensional setting.  The algorithms can be developed with the aim to make an R package if the student wishes to. The method will be tested on simulated data but also on real genomics data including Breast Cancer data and nutrimouse data.

Prerequisites : Programming in R, Python or Matlab

References

[1] Wilms, Ines and Christophe Croux Robust sparse canonical correlation analysis. BMC systems biology 10.1 (2016): 72.

[2] Jendoubi, T., Strimmer, K. A whitening approach to probabilistic canonical correlation analysis for omics data integration. *BMC Bioinformatics* **20,**15 (2019).

**Title:** A stable marriage algorithm

**Supervisor:** Dr Takoua Jendoubi

**Credit Value:** 30

**Description:** Stable matching algorithms have applications in a variety of real-world situations, perhaps the best known of these being in the assignment of patients to beds in a specific hospital or the assignment of medical students to hospitals or solve the college admission problem.  In a number of countries, an automated scheme accomplishes this task annually, by finding a stable matching of students to hospitals or colleges, based on the preferences of hospitals among students and vice versa.

The objective of this project is to develop an algorithm with a view to be used by UCL in similar matching problems, under specific constraints and grouping conditions. The student will be provided with a real world dataset but will also work through simulations.

Prerequisites:  Knowledge in optimization techniques is desirable, but not essential. Familiarity with either R or Matlab is essential.

Suggested Preliminary Reading: Gale, D.; Shapley, L. S. (1962). ["College Admissions and the Stability of Marriage"](https://eur01.safelinks.protection.outlook.com/?url=http%3A%2F%2Fwww.dtic.mil%2Fget-tr-doc%2Fpdf%3FAD%3DAD0251958&data=05%7C01%7Ck.leport%40ucl.ac.uk%7Cfbab85895f9348859a4a08daa5ed8114%7C1faf88fea9984c5b93c9210a11d9a5c2%7C0%7C0%7C638004736725481633%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=9ai9tkeHQjM4qhg2RKqa9MI9jzxkvLWRkqWmz8kKzmo%3D&reserved=0). [*American Mathematical Monthly*](https://eur01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fen.wikipedia.org%2Fwiki%2FAmerican_Mathematical_Monthly&data=05%7C01%7Ck.leport%40ucl.ac.uk%7Cfbab85895f9348859a4a08daa5ed8114%7C1faf88fea9984c5b93c9210a11d9a5c2%7C0%7C0%7C638004736725481633%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=bT8NcvaaANhUsuzfo4lc0zKMmqLU%2BZ7g9D8flfPVSb8%3D&reserved=0). **69** (1): 9–14. [doi](https://eur01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fen.wikipedia.org%2Fwiki%2FDoi_(identifier)&data=05%7C01%7Ck.leport%40ucl.ac.uk%7Cfbab85895f9348859a4a08daa5ed8114%7C1faf88fea9984c5b93c9210a11d9a5c2%7C0%7C0%7C638004736725481633%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=kdpSkI9xnOuA8OSybMkMLYJFyrDZQc3xlixbTToIVZE%3D&reserved=0):[10.2307/2312726](https://eur01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fdoi.org%2F10.2307%252F2312726&data=05%7C01%7Ck.leport%40ucl.ac.uk%7Cfbab85895f9348859a4a08daa5ed8114%7C1faf88fea9984c5b93c9210a11d9a5c2%7C0%7C0%7C638004736725481633%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=Eg312AsF9%2F9ZF334vES8WCLOtHgGgAq9Lvp%2BQ2FklIg%3D&reserved=0). [JSTOR](https://eur01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fen.wikipedia.org%2Fwiki%2FJSTOR_(identifier)&data=05%7C01%7Ck.leport%40ucl.ac.uk%7Cfbab85895f9348859a4a08daa5ed8114%7C1faf88fea9984c5b93c9210a11d9a5c2%7C0%7C0%7C638004736725481633%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=IXdYwNJOfGDy3AioomDhqpO9oF5s0AR6iAF44DY5XPo%3D&reserved=0) [2312726](https://eur01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fwww.jstor.org%2Fstable%2F2312726&data=05%7C01%7Ck.leport%40ucl.ac.uk%7Cfbab85895f9348859a4a08daa5ed8114%7C1faf88fea9984c5b93c9210a11d9a5c2%7C0%7C0%7C638004736725481633%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=jgvhRflnOI5WHMchV4%2FPgKeUAlN3rn2z%2B6%2FvBYSEov0%3D&reserved=0).

**Title:** Analysis of education data

**Supervisor:** Dr Elinor Jones

**Credit Value:** 15/30

**Description:** In this project, you will be expected to analyse a large education dataset. The dataset relates to historic usage of Moodle and Lecturecast. After cleaning the data, you will investigate how students use Moodle and the activities and resources therein. The research questions to be addressed will be up to you (with guidance from me).

The dataset you will use has limited number of variables but a huge number of observations. This will present challenges! You will need to be comfortable using R to carry out the data cleaning, exploratory data analysis, and eventually building suitable models depending on the research question(s) you choose.

**Title:** Analysing data from multiple studies when we can’t share the data

**Supervisor:** Dr Elinor Jones

**Credit Value:** 30

**Description:** How can we analyse data from multiple studies if we aren’t allowed to share the data between studies? This often happens in medical situations: human data are sensitive and often subject to ethico-legal requirements that prevent us from sharing data widely. Often, however, we want to borrow strength from across several studies that addressed the same research question. How do we combine information from different studies to do this without sharing data? The traditional way of doing this would be via a meta-analysis, but can we do better than this?

In this project you will investigate alternatives to meta-analysis studies. That is, method that allow us to use information from multiple sources but without actually sharing the data. This project doesn’t have a dataset to analyse. Instead, you will be expected to run simulation studies to compare the various methods (full guidance will be given if you haven’t done this before).

### **Title:** Fairness considerations for conformal inference

**Supervisor:** Dr Brieuc Lehmann

**Credit Value:** 15/30

**Description:** Conformal inference is a powerful statistical method that allows us to make predictions with quantifiable levels of uncertainty. Importantly, these uncertainty sets, also known as prediction sets, are statistically valid under minimal assumptions with respect to either the distribution of the underlying data, or the model used to study the data. The statistical or algorithmic fairness of a model refers to the quantification of biases or disparities in its performance across certain, typically protected, covariates such as ethnicity or gender. This project will explore the intersection of conformal inference and statistical fairness and provide a synthesis of recent developments in the machine learning literature. Time permitting, the project will also explore ways to incorporate fairness metrics into the evaluation of conformal inference models.

## **Title:** Fairness under Dataset Shift

**Supervisor:** Dr Brieuc Lehmann

**Credit Value:**  15/30

**Description:** Algorithmic or statistical fairness refers to the quantification of biases or disparities in the performance of statistical and machine learning models across individuals or subgroups (e.g., across gender or ethnicity). In recent years, there has been significant interest in the development of fairness metrics to assess the bias present in real-world datasets. Dataset shift occurs when the distribution of the data used to train a model differs from the distribution of the data the model will encounter in the real world. This has implications not just for the overall performance of a model, but also for the fairness characteristics of the model. The aim of this project is to explore the properties of fairness metrics under different forms of dataset shift. Time permitting, the project may also cover methods to correct or account for dataset shift.

**Title:** Cluster-level vs. individual-level adjustment in cluster randomised trials

**Supervisor:** Baptiste Leurent

**Credit Value:** 30

**Description:** Cluster randomised trial (CRT) is an experimental trial design where groups (=clusters) of individuals are randomly allocated to an intervention. For example, it could be communities, schools, or hospitals which are randomised to receive an intervention, such as a campaign to reduce smoking. The other clusters act as comparators. Statistical analysis needs to take into account of the dependencies between observations and is also commonly adjusted for the population characteristics. But there are no clear guidelines on how to best do this in CRT. For example, the adjustment could be made by adjusting by the individual characteristics, or by the average at the cluster level. The latter has been recently recognised as efficient, but there is still a lack of understanding on when this is particularly beneficial.

The aim of this project will be to explore when adjustment for the cluster-level average is more efficient than adjusting for the individual-level values in CRT. This project will involve conducting simulations, and require programming in R.

**Title:** Correction for zero events in cluster randomised trials

**Supervisor:** Baptiste Leurent

**Credit Value:** 30

**Description:** Cluster randomised trial (CRT) is an experimental trial design where groups (=clusters) of individuals are randomly allocated to an intervention. For example, it could be communities, schools, or hospitals which are randomised to receive an intervention, such as a campaign to reduce smoking. The other clusters act as comparators. The analysis of binary or rates outcomes often relies on log-transformation, which is problematic when some of the clusters have no events. One approach around this issue is to add 0.5 to all the clusters. But other approaches are possible and have not yet been compared to one another.

The aim of this project will be to conduct simulations to compare the performance on different approaches to handle clusters with zero event in CRT. We will also apply the methods to a CRT of bed-nets to reduce malaria transmission. This project will involve conducting simulations, and require programming in R.

**Title:** Machine learning for the microbiome

**Supervisor:** Dr Samuel Livingstone

**Credit Value:** 30

**Description:** The human body acts as a host for lots of tiny organisms known as microbes. Scientists believe that understanding the composition of these microbe colonies in individuals could hold the key to understanding differences in health among people. A typical dataset includes samples of microbe concentration levels for hundreds of different microbes, but typically less than one hundred patients. The problem is known as high-dimensional, as there are more features than observations, so some clever dimension reduction and supervised/unsupervised learning methods are needed to learn something interesting from the data. The student would learn about these methods and use them to analyse a microbiome data comparing healthy patients to those with irritable bowel syndrome provided by an industrial collaborator.

**Title:** Topics in convex optimization

**Supervisor:**Dr Samuel Livingstone

**Credit Value:**30

**Description:** This project can take a couple of different directions, from reading some mathematical notes and understanding proofs to exploring some different algorithms computationally by coding them up and comparing empirically on some benchmark problems.  Interested students should contact me for more details.

**Title:** Advanced gradient-based Markov chain Monte Carlo methods

**Supervisor:**Dr Samuel Livingstone

**Credit Value:**30

**Description:** Markov chain Monte Carlo methods (MCMC) are a very popular tool in Statistics, and are used even more widely than this. They are the preeminent method of fitting models using Bayesian inference, and are also very important in understanding the macroscopic behaviour of materials in statistical physics.  In this project we will explore some advanced Markov chain Monte Carlo methods in a statistical context, in particular some advanced algorithms that use gradient information about the model of interest to enhance performance.  It would suit a student interested in a mix of mathematics and computer programming.

**Title:** Optimal Strategies for Games

**Supervisor:** Dr Chak Hei Lo

**Credit Value:** 30

**Description:** More than three centuries ago we already started to develop the concept of gambler's ruin, i.e., anyone who gamble (with non-positive expectation, e.g., buying lottery or stocks) will eventually bankrupt, irrelevant to any betting strategy. However, in reality we tend not to gamble so many times and tend to stop gambling if one achieves certain goal (e.g., to be a billionaire). There can also be short term gambling with positive expectation in the market occasionally. There are subtle differences between games, gambles and investments, in particular the latest focus much on risk control.

The aim of this topic is to use various probability and statistics techniques including martingales, Optional Stopping Theorem and generating functions, to solve some cases of the general gambler's ruin problem. The problem can also be formulated in the sense of random walks. Some cases can be solved explicitly while in some cases only upper and lower bounds are known. There is also a scope of simulation for the cases without analytical solutions and considerations on practical applications on the gambler's ruin problem using various betting systems. These games in various settings exhibit stochastic behaviour and investment opportunities.

This project offers great flexibility on the object of interest to investigate in many industries. Some related examples are games in casino, bank stability, general investment decisions, insurance policies, stocks and cryptocurrencies investments.  Students with strong probability background are recommended but not necessary.

**Title:** Intuitions on Extendable Options

**Supervisor:** Dr Chak Hei Lo

**Credit Value:** 30

**Description:** While vanilla options dominate in exchange trading, exotic options are popular derivatives in the over-the-counter market. Pricing these exotic options is not an easy task in general. In most cases explicit formulae for pricing these options are not known and so Monte Carlo methods are especially useful to estimate these option prices.

In this project, we will investigate the fair price of a particular type of option called extendable options. These options have many interesting behaviours when we change various parameters. We will apply Monte Carlo pricing methods and compare with some exact pricing formula where available in Black-Scholes model. Further aims include a discussion on the advantage and limitations of the Monte Carlo methods and discovering various techniques to reduce computational speed and improve convergence rate of simulation results. Students with some understanding on pricing vanilla options in continuous-time setting and stochastic differential equations are recommended but not necessary.

**Title:** Kernel-based simulation-and-regression approaches for multivariate option pricing

**Supervisor:** Dr Sebastian Maier

**Credit Value:** 15

**Description:** Simulation-and-regression methods are widely regarded as the industry standard for the pricing of derivatives including financial options. These methods overcome the key challenge associated with solving the related optimal stopping problem, more specifically with the stochastic dynamic programming (SDP) recursion, by approximating the (generally impossible to determine) continuation function that occurs in the SDP recursion. In general, the value of these continuation functions, or conditional expectations, is estimated using both Monte Carlo sampling (i.e. simulation) and a statistical model in the form of (non)parametric regression within a standard backward recursion procedure. This project will explore the application of Kernel Ridge Regression (KRR), a widely-used non-parametric technique in the machine learning community, to solve multivariate option pricing problems, such as basket and min-max options, where the option value depends on multiple underlying assets, as opposed to standard single-factor American/Bermudan options. Considering such multi-asset options, this project will then investigate the comparative performance of KRR in terms of computational efficiency, valuation accuracy and modelling complexity.

**Prerequisites:**

Good programming skills (e.g. Matlab/Python/R) and Stochastic Methods in Finance 1 module (STAT0013).

**Reference:** Kohler, M., 2010. A review on regression-based Monte Carlo methods for pricing American options. In *Recent developments in applied probability and statistics* (pp. 37-58). Physica-Verlag HD.

**Title:** Statistical estimation of multinominal logit models using restaurant data

**Supervisor:** Dr Sebastian Maier

**Credit Value:** 15

**Description:** Modelling human behaviour using discrete choice models such as the well-known nested logit and multinominal logit (MNL) models plays an important role in a wide and diverse range of application areas, including e-commerce, marketing, transportation, energy and retailing, and are commonly used in assortment optimisation and recommender systems. To specify the parameters of such choice models, there are mainly three approaches based on either likelihood estimation – i.e. Maximum Likelihood Estimate (MLE) and expectation-maximisation (EM) – or a least-squares procedure. This project will review, study and implement these (and possibly recently proposed, newer) estimation techniques considering the widely-used and famous MNL model, and then analyse the techniques’ comparative performance in terms of numerical accuracy and computational cost, using a publicly available restaurant data-set that consists of sales transaction data. No prior knowledgeof choice models and behavioural economics is needed but some interest in the application area will be useful.

**Prerequisites:**

Good programming skills (e.g. Matlab/Python/R) as well as a good understanding of MLE (STAT0008), least-squares and optimization algorithms (STAT0025) are desirable.

**Reference:** Train, K.E., 2009. Discrete choice methods with simulation. *Cambridge university press*.

**Title:** Statistical choice models

**Supervisor:** Dr Ioanna Manolopoulou

**Credit value:** 15/30

**Description:** Choice models are models where one, or several, options are chosen out of many. For example: voting one out of several candidates; choosing multiple products out of an assortment of many; predicting which team will win in a particular sport. Many models have been proposed in the literature, typically reducing or aggregating the data to simplify the learning task. Several aspects need to be taken into account, for example the person choosing and their characteristics, the available items, the combination of items chosen together. Special cases of choice models are pairwise comparisons (eg football matches) or discrete choice (vote one out of many). Some of the vanilla types of choice models are Bradley-Terry models (pairwise comparisons), Plackett-Luce models (discrete choice models). Although these have been explored in depth, the scenario of choosing multiple out of an assortment is less developed methodologically and typically relies on assumptions about the behaviour of the choice maker. This project will focus on supervised machine learning approaches to predict such preferences, with a view to extend existing approaches to more general settings. One example might be to extend Bradley-Terry models using random forests or Bayesian Additive Regression Trees. Another will be to explore how different reduction and aggregation methods are related to each other, and how to test validity of assumptions for each. Any significant contributions will potentially be included in a python package, skpref, currently work in progress with one of my past MPhil students, Istvan Papp.

References: For an overview of choice models, see slides by

Dilan Görür <http://www.gatsby.ucl.ac.uk/teaching/courses/ml2-2008/choiceML2.pdf> Relevant background: statistical modelling, python. Suitable for students with knowledge of python.

**Title:** Modelling binary data observed with error

**Supervisor:** Dr Ioanna Manolopoulou

**Credit Value:** 15/30

**Description:** Modelling a binary outcome given a set of covariates is a well-studied problem in statistical literature. However, modelling and inference becomes much more challenging if the 0/1 labels are observed with error, especially if the error may also depend on the covariates of interest. Depending on the extent of the mislabelling, ignoring the observation error could lead to vastly inaccurate inference results and models that do not reflect the true outcomes. This project will explore a variety of methods to address this issue from the statistical literature and investigate the potential of using penalised regression methods in this context. The methods will be compared on both synthetic and real datasets in terms of the accuracy of their inference and predictive ability.

Relevant knowledge: statistical computing (STAT0023 or similar)

References:

    • Jiayi Tong, Jing Huang, Jessica Chubak, Xuan Wang, Jason H Moore, Rebecca A Hubbard, Yong Chen, An augmented estimation procedure for EHR-based association studies accounting for differential misclassification, Journal of the American Medical Informatics Association, Volume 27, Issue 2, February 2020, Pages 244–253, <https://doi.org/10.1093/jamia/>

    • Helen (Zhenzheng) Hu, Dirichlet process probit misclassification mixture model for misclassified binary data, UCL PhD thesis <https://discovery.ucl.ac.uk/id/eprint/10140643/>

# **Title:** Improving robust estimation in flexible distributional regression models

**Supervisor:** Dr Giampiero Marra

**Credit Value:** 15/30

**Description:** The validity of estimation and smoothing parameter selection for the wide class of

generalized additive models for location, scale and shape (GAMLSS) relies on the correct

specification of a likelihood function. Deviations from such assumption are known to mislead any

likelihood-based inference and can hinder penalization schemes meant to ensure some degree of

smoothness for nonlinear effects. Recently, Aeberhard et al (2021) proposed a general approach to

achieve robustness in fitting GAMLSSs by limiting the contribution of observations with low log-

likelihood values. This project will look into other options to evaluate an integral required in the log-

likelihood function, with aim of assessing the benefits gained in terms of computational stability and

efficiency. Simulation studies will be carried out and the approach applied to some real dataset.

Aeberhard WH, Cantoni E, Marra G, Radice R (2021) Robust Fitting for Generalized Additive Models for Location, Scale and Shape, *Statistics and Computing*, 31(11)

# **Title:** Copula regression models

**Supervisor:** Dr Giampiero Marra

**Credit Value:** 15/30

**Description:** Modelling two or more response variables simultaneously can be very desirable, due to

the variety of applications where modelling multiple response variables is required. The copula

approach provides a sophisticated method to produce multivariate distributions, whilst preserving

the dependence structure of the variables. This project will look into copula regression approaches

and explore the potential of various extensions and apply them to real data situations. A good

understanding of likelihood theory and good computing skills are required for this project.

Marra G, Radice R (2017), Bivariate Copula Additive Models for Location, Scale and Shape, Computational Statistics and Data Analysis, 112, 99-113.

# **Title:** Flexible Parametric Regression Spline Approaches to Survival Modelling

**Supervisor:** Dr Giampiero Marra

**Credit Value:** 15/30

**Description:** The Cox model is often used for modelling survival data even though there are some

known limitations of the method; for example, it is unable to handle non-proportional effects.

Flexible parametric models are a more general alternative, have nice theoretical properties and are

computationally more tractable. This project will look into the flexible parametric approaches

available in the literature, explore the potential of various extensions and apply them to real data

situations. A good understanding of likelihood theory and good computing skills are required for this

project.

Marra, G. & Radice, R. (2018). Joint generalized additive survival models.

Royston, P. & Parmar, M. (2002). Flexible parametric proportional-hazards and proportional odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Statistics in Medicine, 21(15), 2175-2197.

Liu, X.-R., Pawitan, Y., & Clements, M. (2016). Parametric and penalized generalized survival models. Statistical Methods in Medical Research.

**Title**: Creating an R package

**Supervisor**: Dr Paul Northrop

**Credit value**: 30

**Description**: R is a freely available language and environment for statistical computing and graphics. Packages are fundamental units of R code, including reusable R functions, help documentation that describes how to use these functions and sample data. The aim of this project is to create an R package that others, perhaps other students in the department, could use to perform particular statistical tasks. This will involve writing a set of functions to perform individual aspects of these analyses, annotating this code and writing a description of how to use each function and the package as a whole. The purpose of package will be decided by the student in consultation with me. This project would suit a student who has enjoyed learning how to use R in STAT0004 and, ideally, STAT0023, and wants to take their R programming further.

**Title**: Inference for time series extremes

**Supervisor**: Dr Paul Northrop

**Credit value**: 15/30

**Description**: In a univariate extreme value analysis the aim is to make inferences about future extreme values of a variable of interest. Typically, we need to extrapolate beyond the range of the observed data. For example, we may have a sample of wind speeds collected over the past 20 years and need to make inferences about the largest wind speed to be observed over the next 100 years. Commonly, data come in the form of a time series in which we expect some form of serial dependence, that is, the value at one time is statistically dependent on the value at another time. This causes two issues: (a) dependence in the data should be accounted for in the statistical analyses, and (b) the extent of dependence at extreme levels affects inferences about future extreme values. This project will review the theory behind the extreme value modelling of observations one variable and consider how to deal with issues (a) and (b). Simulated and real data will be used to illustrate how this theory is put into practice. The project will involve some programming in R (STAT0004, STAT0023) and likelihood theory similar to that studied in STAT0005.

**Title:** Sample size for the development of risk models for binary and time to event outcomes

**Supervisor:** Dr Menelaos Pavlou

**Credit value:** 30

**Description:** Risk models are routinely used in clinical practice to predict a health outcome in patients. For example, a risk model may be used to predict the risk of in-hospital death following a cardiac operation based on a set of patient-characteristics, the predictor variables (e.g., age, gender, blood pressure, comorbidities etc.)  For binary outcomes a logistic regression model is often while for time to event outcomes a Cox model is a common choice. Recently sample size formulae were proposed to calculate the sample size required to achieve a desired performance with respect to measures of predictive performance, such as calibration and Mean Absolute Prediction Error. In this project we aim to

1. evaluate these formulae under different scenarios such as: the outcome prevalence, the predictive strength of the model, the number of predictor variables.
2. explore under what circumstances the method of ‘shrinkage’ can be used to either achieve better performance at the recommended sizes or to safely relax the sample size requirements contrast the sample size requirements for binary and survival data.
3. make recommendations about their use in practice and suggest improvements as required.

**References**

[1] Riley, R.D., et al., Calculating the sample size required for developing a clinical prediction model. 2020. 368: p. m441.

[2] Pavlou M, Ambler G, Seaman S R, Guttmann O, Elliott P, King M et al. How to develop a more accurate risk prediction model when there are few events BMJ 2015; 351:h3868 doi:10.1136/bmj.h3868

**Prerequisites:** Knowledge of R programming. Interest in Medical Statistics and/or prediction modelling.

**Title**: Heteroscedastic Error Models for Electron-Nuclear Double Resonance (ENDOR) Spectra

**Supervisor:** Dr Yvo Pokern

**Credit Value:** **30**

**Description:** This project offers the opportunity to interact with on-going research and possibly contribute to a research publication – it is therefore particularly suitable for those who are thinking of continuing with research in statistical science.

Electron-Nuclear Double Resonance (ENDOR) spectra allow the exploration of the local environment of unpaired electrons in large enzyme molecules outperforming all other physicochemical methods of analysis in terms of specificity. The data arise from a collaboration with the Max Planck Institute for Biophysical Chemistry in Germany (see https://www.mpinat.mpg.de/bennati). The project will build on previous work (e.g. the paper in the Proceedings of the National Academy of Sciences at https://www.pnas.org/doi/10.1073/pnas.2023615118 ) including methodology and Python code arising from this collaboration.

Current statistical error models in Electron-Nuclear DOuble Resonance (ENDOR) spectra basically add an i.i.d., bivariate normally distributed error term to an expected signal. For some spectrometers, this error model does not fit the data and an extension is sought that allows one component of the covariance of the error term to evolve over time. The project will involve trying out different models for the covariance matrix, approximately computing maximum likelihood estimators for them using computational methods and assessing their fit to real data sets.

**Prerequisites:** NO background in quantum physics or chemistry is needed. Familiarity with basics of

complex numbers, linear algebra, bivariate normal distributions and MLEs and an ability or willingness to learn programming in Python are required.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Title:** Bayesian Linear Inverse Problems for De-Convolution of Spectra

**Supervisor:** Dr Yvo Pokern

**Credit Value: 30**

**Description:** A spectrum is an intensity (measured as a real value) observed for each frequency on a frequency axis. In practice, spectra are observed with additive error which may follow a normal distribution. Additionally, spectra are often observed indirectly: rather than observing them directly, we only observe a version of the spectrum convolved with a so-called 'kernel', leading to the same kind of integral that needs to be calculated when computing the pdf of the sum of two random variables whose joint pdf is known.

The measured spectrum is thus obtained from the original spectrum that (a) is convolved with the kernel and then (b) has noise added to it. Obtaining the measured spectrum from the original spectrum is known as the "forward problem".

This project aims to adopt the Bayesian viewpoint for the "inverse problem": it imposes a prior on what the original spectrum may look like, combines it with the likelihood that arises from the additive error and studies the behaviour of the Bayesian posterior distribution that describes the probable shapes of the original spectrum.

**Prerequisites:** Some linear algebra, familiarity with multivariate normal distributions and linear algebra. An ability or a willingness to learn to program in Python.

**Title:** Piecewise Deterministic Samplers for Covariance Matrices

**Supervisor:** Dr Yvo Pokern

**Credit Value: 30**

**Description:** Given i.i.d. multivariate data, it is quite straightforward to estimate the most likely covariance matrix given an assumed multivariate normal distribution: it is close to the sample covariance matrix. When comparing models that allow correlation with those that do not, one approach is to perform statistical tests. In the bivariate case, this is easy enough as there is only one correlation coefficient and hence only one test. The higher dimensional the data, the more tests arise leading to a multiple testing problem.

This project instead proposes to use Bayesian inference to estimate covariance matrices and to do so by using a Markov chain Monte Carlo method to perform this inference. One class of algorithms that has received a lot of recent interest is that of piecewise deterministic samplers where the Markov chain has some time periods of evolution according to a deterministic law (e.g. it increases by one unit each time step) and other periods of random jumps typical of the Markov chain you know.

**Prerequisites:** knowledge of Markov chains (e.g. from STAT0007), ability to program in R or Python.

**Title:** Statistical Modelling of Real Inductor Impedances

**Supervisor:** Dr Yvo Pokern

**Credit Value:** 15 or 30

**Description:** Measurements of the complex impedance of a number of inductors (realized as coils) as a function of frequency (between 3.5 and 30 MHz) are available. Simple models would suggest that the complex impedance should be given by X= i \omega L (where „i“ is the complex root of unity, \omega is the frequency and L is a real-valued parameter called the inductance which characterizes the coil). The

measurements clearly contradict these simple models and a sequence of increasingly complicated

models is available in the literature. The project aims to build an error model suitable for the vector

network analyzer used to measure these complex impedances and then estimate and assess a collection of inductor models. Since impedances are usually measured on the Smith chart, measurements can be given an interpretation in hyperbolic Riemannian geometry and it is a non-mandatory option in this project to pursue error models on the hyperbolic disk.

**Prerequisites:**

A willingness to learn some electrical engineering background (Ohm‘s law, impedances). Ability to

code in R. If the hyperbolic option is desired, prior exposure to Riemannian geometry or taking a module on Riemannian geometry concurrently is a definite advantage.

**Title:** Characterizing RF attenuators and detectors

**Supervisor:** Dr Yvo Pokern

**Credit Value:** 15 or 30

**Description:** An electric radio frequency (RF) signal is passed through a sequence of fixed and adjustable attenuators which together attenuate (weaken) the signal strength by an indicated attenuation of X dB. The strength of the resulting signal is measured through an approximately linear device, so that a measurement of the signal strength Y is available for each attenuation X. Since all components are subject to error, a statistical procedure is needed to infer the true attenuation X as well as the deviation from linearity of the indicated signal strength Y. There are two reasons to hope that the problem is identifiable: firstly, the fixed attenuator component of X is known with great precision.

Secondly, the attenuation X can be measured with good precision using direct current (DC) signals.  
A statistical evaluation of the data could start with a simple linear model relating Y to X, progress to  
a factor model based on the decomposition of X into fixed and adjustable components and  
ultimately arrive at using a Generalized Additive Model (GAM). The project would be a natural  
continuation of the "Programming for Practical Statistics" ICA and could be tackled using very similar  
methods perhaps with the addition of nonparametric estimation of the (approximately linear)  
response curve of the measurement device.

**Prerequisites:** Ability to use R (as in the module “Programming for Practical Statistics”), a willingness  
to learn about basics of RF circuits to better understand the application area.

**Title**. Practical non-identifiability of parametric models

**Supervisor:** Dr Javier Rubio Alvarez

**Credit Value:** 30

**Description**: Identifiability of parametric models is one of the most basic requirements in parametric models. Identifiability helps establish asymptotic results and facilitates the interpretation of results of data analyses based on such parametric models. However, some identifiable parametric models behave as a nonidentifiable model for specific data sets. This phenomenon is known as practical non-identifiability. The aim of this project is to present a review on the concept of practical non-identifiability and to present some examples of models which exhibit this feature.

**Prerequisites**. Knowledge of R and motivation to write and run code on your own. Knowledge of statistical inference (STAT0005 and/or STAT0008).

**Title.** Bayesian analysis of population growth models

**Supervisor:** Dr Javier Rubio Alvarez

**Credit Value:** 30

**Description.** The aim of this project is to implement Bayesian methodology for the estimation of the parameter of sigmoid population growth models (such as logistic and Gompertz). This includes investigating whether one can estimate all model parameters including the initial conditions, when they are unknown, in contrast to fixing this parameter value. Real data applications will be presented to illustrate the parameter estimation and quantification of uncertainty.

**Prerequisites**. Knowledge of R and motivation to write and run code on your own. Knowledge of statistical inference (STAT0005 and/or STAT0008). Basic knowledge of Ordinary Differential Equations.

**Title:**Asymptotics of Discrete Determinantal Point Processes

**Supervisor:** Dr Kayvan Sadeghi

**Credit Value:**15/30

**Description:**Discrete determinantal point processes (DPPs) are stochastic processes for repulsion that model in a mathematically elegant and general way negative association. Originally developed in quantum physics, DPPs arise naturally in other areas, such as combinatorics, random matrix theory, and machine learning. The main goal of this project is to ascertain how restrictive these models are asymptotically by studying their corresponding kernel matrices. This can be done through theoretical methods, but most probably through simulation studies. A good knowledge of linear algebra is required as well as programming skills in Matlab, Mathematica, or a similar programming language.

**Title:**Review of Structure Learning Algorithms for Causal Inference

**Supervisor:** Dr Kayvan Sadeghi

**Credit Value:**15/30

**Description:** There have been many different algorithms defined in the literature to learn (i.e., find) the "true causal graph" of a causal model. Our focus in this project is on score-based algorithms, which optimize an assigned score by searching through the class of graphs in order to learn the causal graph. The goal of the project is to provide a comprehensive review of these algorithms, to discuss pros and cons of each of these, and to use some real-world or simulated data sets to compare these algorithms.

**Title**: Traffic Flow Maps

**Supervisor:** Dr Cemil Selcuk

**Credit Value**: 15 credits

**Description**: A connectome is a comprehensive map showing the links between various nodes, their directions and strengths. The goal of the project is to construct a connectome for a city based on ride-hailing data. First, you will create the transition matrix across a number of zip codes. You will then embed the graph of the matrix into the city’s map and use the weights and directions across links to illustrate the amount and direction of traffic flow.

**Prerequisites**: Good programming skills, preferably MATLAB. Good understanding of probability and statistics (STAT2001 or equivalent).

**Title**: Seasonality in the Housing Market

**Supervisor:** Dr Cemil Selcuk

**Credit Value**: 15/30

**Description**: The US housing market goes through seasonal boom and bust episodes:- in summers prices rise and trade speeds up whereas in winters prices fall, it takes much longer to sell and the number of sales slides to the annual lows. The cycles are highly predictable and repetitive, seemingly defying the no-arbitrage condition. The goal of the project is two-fold: The first step involves obtaining the necessary data from publicly available sources (sale prices from the Federal Housing Financing Agency and the turnover from US Census Bureau) and then recovering and plotting the seasonal factors in each data set. The second step involves conducting a literature review in economic theory explaining the rationale behind the repetitive nature of these cycles. Familiarity with programming (R, Matlab, Python etc.) is a pre-requisite. Some knowledge of microeconomic theory would be useful, but not a pre-requisite.

Title: Statistical analysis of air pollution extremes

Supervisor: Dr Emma Simpson

Credit Value: 30

Description: Air pollution is made up of various components, and in high (or extreme) quantities, some of these can have a detrimental effect on human health. In the UK, the Department for Environment Food and Rural Affairs (Defra) provides air pollution data at over 1500 locations; modelling a subset of this will be the goal of the project.

When modelling extreme values, the focus is on the tail of a distribution. Usually, this means there is a limited amount of data corresponding to the events of interest, and we may even wish to extrapolate beyond the observations available to estimate the probability of very rare events. To model this tail behaviour, we require theoretically-justified methods, which come from the Extreme Value Theory literature.

The project will involve the application of classical univariate extreme value methods to data from a single air pollution monitoring site. There are two common approaches that can be considered: modelling maximum observations in particular time-windows using the generalised extreme value (GEV) distribution; and modelling exceedances of some high threshold using a generalised Pareto (GP) distribution. Subsequent techniques covered can depend on the interests of the student, but it may be interesting to consider temporal behaviour in the air pollution extremes and/or the dependence between extreme values across different air pollution variables.

Good programming skills in R will be useful for this project.

Title: Modelling extreme sea surface temperatures

Supervisor: Dr Emma Simpson

Credit Value: 30

Description: Sea surface temperature (SST) is a useful indicator of the health of marine habitats, such as coral reefs. Corals can adapt to some changes in SST, but if these temperatures are particularly high for a prolonged period of time, damage such as coral bleaching or even coral mortality can occur. Understanding these high temperatures is therefore important, and methods from Extreme Value Theory can be used in this context.

The project will focus on modelling SST data from the Red Sea. It will begin with an introduction to some univariate methods for modelling extremes, and the application of these to the data at single locations. It may be interesting to consider whether extremal characteristics of the univariate fits vary across the spatial domain, e.g., in relation to water depth or distance to the coast. Depending on the interests of the student, further aspects that could be studied include modelling temporal dependence in the extremes, or assessing extremal dependence between different pairs of locations, particularly with respect to how this changes with the distance between locations.

Good programming skills in R will be useful for this project.

**Title:** Conditional modelling of extreme river flow data

**Supervisor:** Dr Emma Simpson

**Credit Value:** 30

**Description:** River flow data are widely studied in hydrological applications, with large observations being linked to flooding events. Understanding the behaviour of these extreme events is important to plan for and mitigate their effect. Data for several gauging stations located across the UK river network are available from the UK National River Flow Archive; modelling extremal features of this data set will be the main aim of the project.

The project will begin with an introduction to some standard techniques for modelling univariate extremes, with these methods being applied to data from individual river gauging stations.

Multivariate features of the data can then be considered. The conditional approach of Heffernan and Tawn (2004) is a flexible method for modelling dependence in multivariate extremes. In the bivariate case, it involves conditioning on values of one variable exceeding a high threshold and modelling the behaviour of the other variable. This allows for the estimation of rare events, even extrapolating beyond those observed in the data. The bivariate conditional extremes model could be applied, for example, to data from two different gauging stations, or to data at a single station at different time lags, in order to model their joint behaviour.

Good programming skills in R will be useful for this project, as it will involve the use of existing R packages and the student coding their own functions.

**Title**: Multi-state survival models: expected time spent in states

**Supervisor**: Dr Ardo van den Hout

**Credit Value:** 15/30

**Description:** A multi-state model is an extension of the two-state survival model. Instead of having only one event time (time of death, say), there are multiple event times (times of transitions between states). An example is a model for longitudinal data for grades of cardiac allograft vasculopathy (CAV). Data for CAV are available from a follow-up study of heart-transplant patients. Four CAV states are defined: no CAV (state 1), mild/moderate CAV (state 2), severe CAV (state 3), and death (state 4).

Of interest in the project is to estimate time spent in the disease states. This implies fitting a time-continuous multi-state model to the CAV data and then used the fitted model to derive expected time spent in states.

The project starts with fitting a Markov model (see STAT007) using an R package for multi-state models. At a later stage, the project will include user-written code to deal with extended models.

**­­­­­­­­­**

**Title**: Joint Models for position and time in music chart

**Supervisor**: Dr Ardo van den Hout and Ed Burmicz (Warner Music)

**Credit Value:** 30

**Description:** Success of music tracks can be measured by the time they are in the charts and by the position of the track whilst being in the charts. Example of music charts are the ones composed by Official Charts Company (OCC, UK) and Billboard (USA).

Because time in chart is related to position in chart, it makes sense to model both outcomes jointly. When we want to model tracks in a current chart, there will be tracks for which we have longitudinal data on position but no final time in chart (as the track has not exited the chart yet). This type of data resembles data in survival analysis (with time in chart being subject to so-called right censoring).

The project will use data from OCC Top 100 singles. Of interest is to model position and time in the chart and to identify features (such as type of music, streaming history, etc.) that are correlated with a track being successful. Software: R.

**Title**: Models of cell growth and division

**Supervisor**: Dr Alex Watson

**Credit Value:** 30

**Description:** Biological cells grow slowly over time before splitting in two. The trigger for this splitting, or division, is still obscure. The three main theories are that cells divide when they reach a certain size; when they reach a certain age; or when they reach a certain multiple of their size at birth (‘adder’ model). These theories are quite hard to distinguish -- after all, cells that are old are also large.

The goal of this project is to perform simulation studies in R or Python using a flexible probabilistic model of cell growth and division, in order to find and compare certain statistics of the model, such as the population growth rate and the ‘profile’ of the population. If there is time, we will look at fitting using modern computational statistics methods.

**Prerequisites:** Familiarity with R or Python; STAT0007

**Title**: Optimisation, dynamic programming and reinforcement learning

**Supervisor**: Dr Alex Watson

**Credit Value:** 30

**Description:** Decision making under uncertainty is a skill required in every industry, and it brings in many tools from mathematics and statistics. Some of the most popular and successful approaches for this are phrasing in terms of dynamic programming problems and reinforcement learning problems. The theory for these is mathematically interesting and elegant, but the practitioner is often left with a difficult optimisation problem, for instance, to maximise some function written in an implicit form, or which is costly to compute. For this reason, the study of such problems goes hand in hand with the study of optimisation techniques.

The is a flexible project which will allow the student to concentrate on problems that interest them drawn from across a spectrum of potential application areas. It will involve a mixture of theoretical work and computational practice. Please note: because of the focus on theoretical understanding and methods with a sound mathematical basis, it is unlikely that you will be able to tackle modern machine learning problems in this project, nor will we not approach any of the popular computational frameworks for this.

**Prerequisites:** Familiarity with R or Python

**Title**: Sparse principal component analysis and its applications

**Supervisor**: Professor Jinghao Xue

**Credit Value**: 30

**Description**: Sparse principal component analysis is an extension of principal component analysis (PCA) to high-dimensional data. PCA is a popular multivariate statistical method for analysing a set of multivariate data. This project aims to study the methodology, algorithms, and theory of PCA and sparse PCA; and apply sparse PCA to some low-dimensional datasets and some high-dimensional datasets and visualise and analyse the results.

**Title**: Sparse canonical correlation analysis and its applications

**Supervisor**: Professor Jinghao Xue

**Credit Value**: 30

**Description**: Sparse canonical correlation analysis is an extension of canonical correlation analysis (CCA) to high-dimensional data. CCA is a popular multivariate statistical method for analysing two sets of multivariate data. This project aims to study the methodology, algorithms, and theory of CCA and sparse CCA; and apply sparse CCA to some low-dimensional datasets and some high-dimensional datasets and visualise and analyse the results.

**Title**: Sparse linear discriminant analysis and its applications

**Supervisor**: Professor Jinghao Xue

**Credit Value**: 30

**Description**: Sparse linear discriminant analysis is an extension of linear discriminant analysis (LDA) to high-dimensional data. LDA is a popular multivariate statistical method for classifying multivariate data. This project aims to study the methodology, algorithms, and theory of LDA and sparse LDA; and apply sparse LDA to some low-dimensional datasets and some high-dimensional datasets and visualise and analyse the results.

**Title**: Principal component regression, ridge regression, and partial least squares

**Supervisor**: Professor Jinghao Xue

**Credit Value**: 30

**Description**: Principal component regression (PCR), ridge regression, and partial least squares (PLS) are three popular multivariate statistical approaches to regression analysis. This project aims to study the methodology, algorithms, theory and relationship of these three approaches; and apply them to some low-dimensional datasets and some high-dimensional datasets, and visualise, analyse and compare their results.