UCL DEPARTMENT OF STATISTICAL SCIENCE

MSc PROJECTS GUIDE

Data Science 2018-2019

MSc Project Arrangements 2019

1. Students should plan to take a short break after their exams, before starting work on their projects. All supervisors are likely to be away from time to time during the period June – September, attending conferences or on holiday. Students should therefore see their supervisors ***as soon as*** their exams are over, to make mutually convenient arrangements for starting work on their projects.

2. During the project work, student and supervisor should arrange to meet regularly (about once a week, whenever possible) and should agree a suitable timetable for completing the project work and producing a written account.

3. Students are encouraged to use **LaTeX** for the written report. This is a powerful program for producing technical documents and is well worth learning – please see the STAT0034 Moodle page.

4. Students should submit ***draft*** versions of their project reports to their supervisors for comment by Friday **9th August 2019.**

5. Final reports should be typed and handed in to the Departmental Office by **4pm on Friday, 30th August 2019**. An electronic version should also be submitted in the designated area of the Moodle page of the MSc Project Course by **4pm on Friday, 30th August 2019.**

6. The hard copy and the electronic version must be identical. Late submissions will incur severe 'lateness' penalties (please see Section 3.11 of the UCL Academic Regulations) and reports submitted more than five working days late will receive a mark of zero. The project presentations will take place during the **week beginning 2nd September 2019** (precise date to be confirmed).

7. Students should arrange to be available during the week of **19th – 23rd August 2019**, in case their supervisors need to contact them with queries about their reports.

8. The length of the project report depends on the topic of the project and may vary considerably between projects. Lengths between 8,000 and 15,000 words (excluding the table of contents, the reference list, and any tables, graphs, computer programs, computer output and appendices) are generally acceptable. Typical project reports are between 10,000 and 12,000 words long. The final version of the project report should state its word count on the front page, and the absolute maximum allowed is 16,500 words. Project reports longer than 16,500 words will incur a 10 percentage point deduction in marks, subject to the provisions of Section 3.12 of the Academic Regulations). It is generally required that the amount of work done and demonstrated is large enough, and that the material is presented in a way understandable to fellow students with a comparable background (so 8,000 words may only be an appropriate length for a very theoretical or densely presented report). On the other hand, reports should not be too repetitive or contain unnecessary or irrelevant details, which may lead to downmarking even below 15,000 words.

9. Please see pages 25-27 of the Postgraduate Student Handbook for guidelines on what examiners will be looking for in a project. These guidelines can also be found on pages 48-50 of this document.

MSc Projects List 2019 (MSc Data Science)

PROJECT TITLE	SUBMITTING SUPERVISOR	ADDITIONAL SUPERVISORS	PROGRAMME SUITABILITY			PAGE NUMBER
			MSc Statistics	**MSc Data Science**	**MSc Medical Statistics**	
HIDDEN MARKOV MODELS WITH APPLICATIONS IN FINANCE AND THE COMPUTATIONAL METHODS FOR THEIR CALIBRATION	Beskos, Alexandros	N/A	Yes	Yes	No	1
UNDERSTANDING EARTHQUAKE DATA USING HIDDEN MARKOV MODELS	Beskos, Alexandros	N/A	Yes	Yes	No	1
BAYESIAN INFERENCE FOR CHANGE POINT MODELS	Beskos, Alexandros	N/A	Yes	Yes	No	2
MULTIVARIATE STOCHASTIC VOLATILITY MODELS FOR EXCHANGE RATES AND THEIR CALIBRATION USING STAN	Beskos, Alexandros	N/A	Yes	Yes	No	2
OTHER TOPICS IN ENVIRONMENTAL STATISTICS	Chandler, Richard	N/A	Yes	Yes	No	3
TYPHOON MODELLING FOR THE NORTHWEST PACIFIC	Chandler, Richard	N/A	Yes	Yes	No	3
UNCERTAINTY ESTIMATION FOR EARTHQUAKE MAGNITUDES	Chandler, Richard	N/A	No	Yes	No	3
MODELLING FIRST NAMES' DISTRIBUTIONS IN THE USA: 1940-2017	Cortina-Borja, Mario	Wade, Angie	No	Yes	No	4
ESTIMATING MODES FROM SAMPLES	Fearn, Tom	N/A	Yes	Yes	No	5
R PROGRAMS FOR SPECTROSCOPIC CALIBRATION	Fearn, Tom	N/A	Yes	Yes	No	5
ONE OR TWO DISTINCT PROJECTS ON THE ANALYSIS OF HIGH-DIMENSIONAL SPECTROSCOPIC DATA, TITLES WILL DEPEND ON EXACTLY WHAT AREA THE PROJECT COVERS	Fearn, Tom	N/A	Yes	Yes	No	6
ANALYSING DOCUMENTS ON OPEN QUESTIONS	Guillaumin, Arthur	Bernardoni, Mirko	Yes	Yes	No	6
GAUSSIAN PROCESS EMULATORS USING GPUS	Guillas, Serge	N/A	Yes	Yes	No	7
GAUSSIAN PROCESS EMULATORS USING HIERARCHICAL STRUCTURES	Guillas, Serge	N/A	Yes	Yes	No	7
BAYESIAN VARIABLE SELECTION IN LINEAR MODELS WITH INTERACTIONS	Griffin, Jim	N/A	Yes	Yes	No	8
GAS MODELS FOR ROBUST INFERENCE	Griffin, Jim	N/A	Yes	Yes	No	8
LOCAL PROJECTIONS FOR IMPULSE RESPONSE FUNCTION	Griffin, Jim	N/A	Yes	Yes	No	8

INVESTIGATION OF DEPENDENCE IN MULTITYPE POINT PATTERN DATA	Honnor, Thomas	N/A	Yes	Yes	No	9
INVESTIGATION OF INHIBITION IN POINT PATTERN DATA	Honnor, Thomas	N/A	Yes	Yes	No	9
INVESTIGATION OF POINT PATTERN DATA	Honnor, Thomas	N/A	Yes	Yes	No	10
WORKFLOW ARCHITECTURES AND UNIFIED INTERFACE DESIGN OF TOOLBOXES FOR STATISTICAL MODELLING AND MACHINE LEARNING IN THE PRESENCE OF STRUCTURE (INC. TEMPORAL AND HIERARCHICAL)	Kiraly, Franz	N/A	Yes	Yes	No	10
THEORETICAL AND EMPIRICAL STUDY OF GENERALIZATION ERROR ESTIMATES, QUANTITATIVE MODEL VALIDATION AND MODEL COMPARISON META-METHODOLOGY	Kiraly, Franz	N/A	Yes	Yes	No	11
ADVANCED MARKOV CHAIN MONTE CARLO METHODS FOR BAYESIAN COMPUTATION	Livingstone, Samuel	N/A	Yes	Yes	No	12
BAYESIAN MODELLING AND INFERENCE OF GENETIC ANCESTRY	Livingstone, Samuel	N/A	Yes	Yes	No	12
CONVERGENCE OF SAMPLING ALGORITHMS	Livingstone, Samuel	N/A	Yes	Yes	No	13
PREDICTING THE EFFECTS OF WEATHER CHANGES ON DEMAND FOR CHILDREN'S AMBULANCE SERVICES IN GREATER LONDON	Livingstone, Samuel	N/A	Yes	Yes	Yes	13
PATTERNS OF ORGAN DYSFUNCTION IN SEPSIS	Marra, Giampiero	Palmer, Edward	Yes	Yes	Yes	14
ADDRESSING MISSING DATA IN OBSERVATIONAL STUDIES WITH TIME-VARYING CONFOUNDING	Marra, Giampiero	Gomes, Manuel	Yes	Yes	Yes	16
REVISITING THE USE OF COPULA MODELLING IN COST-EFFECTIVENESS ANALYSIS	Marra, Giampiero	Gomes, Manuel	Yes	Yes	Yes	16
FORECASTING MOVEMENT IN FOOTBALL BETTING MARKETS	Marra, Giampiero	N/A	Yes	Yes	No	17
EVALUATING THE EFFICIENCY OF FOOTBALL BETTING MARKETS	Marra, Giampiero	N/A	Yes	Yes	No	18
STUDENT PROPOSED PROJECT	Marra, Giampiero	N/A	Yes	Yes	Yes	18
MIXTURE MODELING OF CELL SUBTYPES IN FLOW CYTOMETRY	Manolopoulou, Ioanna	N/A	Yes	Yes	Yes	19
MODELLING RING GALAXIES USING GAUSSIAN WELLS	Manolopoulou, Ioanna	N/A	Yes	Yes	No	19
MODELLING COUNT PROCESSES USING THE HAWKES PROCESS	Manolopoulou, Ioanna	N/A	Yes	Yes	No	20

BAYESIAN MULTI-STATE TRANSITION RATE MODELLING	Nicholas, Owen	Van Den Hout, Ardo	Yes	Yes	Yes	20
CREATING AN R PACKAGE	Northrop, Paul	N/A	Yes	Yes	No	21
USING BAYESIAN EXTREME VALUE MODELLING TO PREFORM OPEN SET CLASSIFICATION	Northrop, Paul	N/A	Yes	Yes	No	21
ANALYSIS OF AMATEUR RADIO CONTEST DATA	Pokern, Yvo	N/A	Yes	Yes	No	22
CONDITIONS FOR AND TESTING THE COMPOSITION PROPERTY	Sadeghi, Kayvan	N/A	Yes	Yes	No	23
RESTRICTIVENESS OF DISCRETE DETERMINANTAL POINT PROCESSES	Sadeghi, Kayvan	N/A	No	Yes	No	23
MODEL POLYTOPES FOR EXPONENTIAL RANDOM GRAPH MODELS	Sadeghi, Kayvan	N/A	Yes	Yes	No	24
CAUSAL NETWORKS WITH "WEAK ASSOCIATIONS"	Silva, Ricardo	N/A	Yes	Yes	Yes	24
CAUSALITY IN REINFORCEMENT LEARNING	Silva, Ricardo	N/A	Yes	Yes	Yes	25
MODELLING CAUSAL EFFECTS IN SOCIAL AND SPATIAL NETWORKS AND OTHER DEPENDENT DATA	Silva, Ricardo	N/A	Yes	Yes	Yes	25
MODELLING CAUSAL EFFECTS ON TIME-SERIES DATA WITH NATURAL EXPERIMENTS	Silva, Ricardo	N/A	Yes	Yes	Yes	25
STUDENT-LED PROJECT	Silva, Ricardo	N/A	Yes	Yes	Yes	25
LARGE SCALE ANALYSIS OF USER BEHAVIOUR WITH SPOTIFY DATA	Silva, Ricardo	N/A	Yes	Yes	No	26
SHAZAM FOR EARTHQUAKES – AN APPLICATION TO A REGIONAL EARTHQUAKE CATALOG	Stavrianaki, Katerina	N/A	Yes	Yes	No	26
DE-CLUSTERING EARTHQUAKE CATALOG DATA	Stavrianaki, Katerina	N/A	Yes	Yes	No	27
PARAMETRIC TIME-DEPENDENT MULTI-STATE SURVIVAL MODELS	Van Den Hout, Ardo	Nicholas, Owen	Yes	Yes	Yes	27
BIVARIATE DISCRETE DISTRIBUTIONS TO MODEL COGNITIVE FUNCTION	Van Den Hout, Ardo	N/A	Yes	Yes	Yes	28
GENERALISED TIME-DEPENDENT LOGISTIC MODELS FOR SURVIVAL DATA	Van Den Hout, Ardo	N/A	Yes	Yes	Yes	28
FALSE DISCOVERY CONTROL IN MULTIPLE TESTING	Wang, Tengyao	N/A	Yes	Yes	No	29
MATRIX COMPLETION	Wang, Tengyao	N/A	Yes	Yes	No	29
PRINCIPAL COMPONENT ANALYSIS IN HIGH DIMENSIONS	Wang, Tengyao	N/A	Yes	Yes	No	30

OPTIMISED AGGREGATION OF CITIZEN SCIENCE DATA FOR BIOMEDICAL IMAGE ANALYSIS	Xue, Jinghao	Jones, Martin	Yes	Yes	Yes	30
CLASSIFICATION OF PSEUDO CANCER VERSUS POLYP CANCER FROM RAMAN IMAGES	Xue, Jinghao	Thomas, Geraint	Yes	Yes	Yes	31
CLASSIFYING CLUSTERED RAMAN DATA OF GASTRO-INTESTINAL CANCERS	Xue, Jinghao	Thomas, Geraint	Yes	Yes	Yes	31
SEMI-SUPERVISED MACHINE LEARNING TO CLASSIFY RAMAN IMAGES OF OVARIAN CANCER	Xue, Jinghao	Thomas, Geraint	Yes	Yes	Yes	31
META-ANALYSIS OF KAPPA STATISTICS FOR COLON CANCER ASSESSMENT	Xue, Jinghao	Thomas, Geraint	Yes	Yes	Yes	32
REPRESENTATION-BASED CLASSIFICATION	Zhu, Rui	Xue, Jinghao	Yes	Yes	No	32

MSc Project Descriptions for 2019

.

Title: Understanding Earthquake Data using Hidden Markov Models

Supervisor: Dr Alexandros Beskos

Suitability: MSc Statistics or MSc Data Science

Description:

Several works in Environmental Sciences have used Hidden Markov models to fit data corresponding to Earthquake incidents. The underlying Markov state can be thought of as the “stress field” leading to the earthquake events; the observations are the magnitudes of the deduced earthquakes. The project will fit HMMs to earthquake data; we will look at maximum likelihood and Bayesian approaches, and try to quantify uncertainty of estimated model parameters and carry out model selection.

Prerequisites: Familiarity with R (e.g. as taught STAT0030), applied Bayesian methods (e.g. as taught in STAT0031) and Markov chains. Also, willingness to learn and apply computational techniques, using for instance STAN or OpenBUGS.

Title: Hidden Markov Models with Applications in Finance and the Computational Methods for their Calibration

Supervisor: Dr Alexandros Beskos

Suitability: MSc Statistics or MSc Data Science

Description:

Hidden Markov models (HMM) constitute an important class of statistical models used in a wide variety of applications. The project will explore recent Monte-Carlo techniques (e.g. combinations of Markov chain Monte-Carlo and Particle filters, known as Particle-MCMC) to fit such models to data. We will look at applications of HMM to financial models, e.g. multivariate stochastic-volatility models or GARCH-type ones.

Prerequisites: Familiarity with R (e.g. as taught in STAT0030), applied Bayesian methods (e.g. as taught in STAT0031) and Markov chains. Also, willingness to learn and apply computational techniques relevant to HMMs; e.g. STAN or OpenBUGS.

Title: Multivariate Stochastic Volatility Models for Exchange Rates and their Calibration using STAN

Supervisor: Dr Alexandros Beskos

Suitability: MSc Statistics or MSc Data Science

Description:

Recent works in Econometrics have looked at the joint modelling of multivariate exchange rates (or other time series) within a stochastic volatility context. Such models pose challenges with regards to their calibration to observations due to unobserved stochastic volatility and high-dimensionality. However, recent computational packages like STAN allow non-experts in advanced MCMC methodology, to apply powerful computational algorithms to calibrate such models and perform full Bayesian inference and prediction.

Prerequisites: Familiarity with R (e.g. as taught in STAT0030), applied Bayesian methods (e.g. as taught in STAT0031), Markov chains. Also, willingness to learn and apply MCMC techniques, using packages like OpenBUGS or STAN.

Title: Bayesian Inference for Change Point Models

Supervisor: Dr Alexandros Beskos

Suitability: MSc Statistics or MSc Data Science

Description:

Many applications in statistics involve learning change point models from available data. Such models pose challenges with regards to their calibration due to the existence of latent components and high-dimensionality. However, recent computational advances allow for fast computational statistics methods – in a Bayesian framework – to discover the change points. Potential applications include Biology or Finance.

Prerequisites: Familiarity with R (e.g. as taught in STAT0030), applied Bayesian methods (e.g. as taught in STAT0031), Markov chains. Also, willingness to learn and apply Monte-Carlo techniques, e.g. MCMC or Particle Filtering, using packages like OpenBUGS or STAN.

Title: Typhoon modelling for the northwest Pacific

Supervisor: Professor Richard Chandler

Suitable for: MSc Statistics, MSc Data Science

Description:

This project will use data from the China Meteorological Administration's "tropical cyclone best track data" to develop models of tropical cyclone (typhoon) formation and movement that can be used to assess the changing risk of storm damage in East Asia in a changing climate. The work will build on a other student projects and will required good computational skills (at least an "A" grade in STATG003) as well as familiarity with generalised linear models (STATG001 or STATG006) and a willingness to learn new stochastic modelling techniques.

Title: Other topics in environmental statistics

Supervisor: Professor Richard Chandler

Suitable for: MSc Statistics, MSc Data Science

Description: Students are welcome to suggest their own project topics in any environment-related application. Any student who wishes to take this option will need to consider the availability of suitable data sets.

Title: Uncertainty estimation for earthquake magnitudes

Supervisor: Professor Richard Chandler

Suitable for: MSc Data Science

Description:

When an earthquake occurs, waveforms are recorded at a set of seismograph stations. These are used to infer the characteristics (location, magnitude, depth and so forth) of the earthquake. This is an inverse problem, for which a variety of geophysics-based algorithms have been developed. However, the algorithms are based on imperfect models of the earth and measurement properties. As a result, earthquake magnitudes can never be known perfectly. Assessing the uncertainty in their estimates is important for applications such as seismic hazard and risk assessment. However, uncertainty assessment is rarely done in a statistically rigorous manner. The aim of this project is to fill this gap. The student will work with publically available software for earthquake waveform inversion (written mostly in C++), and will implement statistical methodology for calculating standard errors based on the theory of estimating functions. The project will require excellent computing skills in multiple programming languages, and a willingness to learn some new statistical methodology.

Title: Modelling first names' distributions in the USA: 1940-2017

Supervisors: Prof. Mario Cortina-Borja and Prof. Angie Wade (UCL Institute of Child Health)

Suitable for: MSc Data Science with specialisation in Statistics

Description:

According to the German sociologist Max Weber "Naming a baby is a social act" (2), determined by parental choice and their social environment. Analyses of changes in names distributions may give important insights into latent social processes.

In this project, we will work with a large population database of frequencies of first names by gender, and state and year of birth collected by the Office of the Chief Actuary, USA Social Security Administration. In contrast with other databases which, for confidentiality reasons, provide only frequencies ≥5 or ≥3, e.g. in the R libraries babynames and ukbabynames, we will look at distributions including frequencies ≥1.

Apart from constructing visualizations similar to those presented by Hoehle (4), we will test the hypothesis of rapid change in traditional values related to baby names choice. In particular we will look at the "increased competition and reduced popularity" process outlined by Tucker (7) and Bruhn et al (2), and the trends in unisex names (1). Summary statistics often used in analysis of vocabulary distributions or income inequalities, e.g. hapax legomena, Yule's *K* function (5), and Gini's coefficient as applied to surname distributions (4), will be used to describe such changes. Statistical models will be based on Zipf (3) and Sichel (6) conditional probability distributions.

References

1. Barry H; Harper AS (2014) Unisex Names for Babies Born in Pennsylvania 1990-2010. *Names* **62** 13-22.
2. Bruhn A; Huschka D; Wagner GG (2012) Naming and War in modern Germany. *Names* **60** 74-89.
3. Clauset A; Shalizi CR; Newman MEJ (2009). Power-Law distributions in empirical data. *SIAM Review,* **51**, 661–703
4. Hoehle M (2017) Rank uncertainty: Why the "most popular" baby names might not be the most popular. *Significance*,**14** 30-33.
5. McElduff F; Mateos P; Wade A; Cortina-Borja M (2008) What's in a name? The frequency and geographic distributions of UK surnames. *Significance*, **5** 189-192.
6. Stein GZ; Zucchini W; Juritz JM (1987) Parameter estimation for the Sichel distribution and its multivariate extension. *Journal of the American Statistical Association* **82** 938-944.
7. Tucker KD (2009) Increased competition and reduced popularity: US given name trends of the twentieth and early twenty-first centuries. *Names*, **57** 52-62.

Title: R programs for spectroscopic calibration

Supervisor: Professor Tom Fearn

Suitability: Statistics or Data Science (Statistics stream)

Description:

Near infrared (NIR) spectroscopy provides a rapid method for measuring the composition of intact foods and agricultural products, as well as many other materials. The measurement is indirect and needs calibration, in which prediction rules are derived that will either predict the quantitative composition of samples or classify them to one of two or more groups on the basis of a high-dimensional absorbance spectrum. Methods such as principal components regression (PCR) and partial least squares regression (PLSR) are standard, though nonlinear approaches such as artificial neural nets or Gaussian process regression have also been used.

I work mainly with Matlab, but (because it is free!) R is an attractive alternative. There are some R programs around that implement methods like PLS, but it is harder to find R programs that, for example, implement common spectral pre-treatments like derivatives or scatter correction. Another issue is the implementation of cross-validation, often used to tune the methods. When the data has a nested structure it is useful to be able to specify the blocks to be left out in a flexible manner and packages often do not allow this. The idea of the project would be to investigate what is available, compare the options when these exist, make some recommendations, and plug any obvious gaps with some nice well-documented code. You would obviously need to really like R to take this one on.

Title: Estimating modes from samples

Supervisor: Professor Tom Fearn

Suitability: Statistics or Data Science (Statistics stream)

Description:

A chemist collaborator of mine is interested in estimating the mode of the distribution that a set of data was sampled from. There is an R package called modeest that appears to have several methods for doing this. This project would involve trying some or all of these methods on both real and simulated data to see how well they work.

Title: One or two distinct projects on the analysis of high-dimensional spectroscopic data, titles will depend on exactly what area the project covers

Supervisor: Professor Tom Fearn

Suitability: Statistics or Data Science (Statistics stream)

Description:

Near infrared (NIR) spectroscopy provides a rapid method for measuring the composition of intact foods and agricultural products, as well as many other materials. The measurement is indirect and needs calibration, in which prediction rules are derived that will either predict the quantitative composition of samples or classify them to one of two or more groups on the basis of a high-dimensional absorbance spectrum. Methods such as principal components regression (PCR) and partial least squares regression (PLSR) are standard, though nonlinear approaches such as artificial neural nets, support vector machines or Gaussian process regression have also been used.

I have various NIR datasets, both quantitative (e.g. measurement of fatty acids in pig carcasses) and qualitative (classification of animal feed ingredients, classification of Iberian ham) and some ideas for methods to try on them, though of course students may have their own favourite methods they would like to try. I'd be happy to discuss options with interested students.

Title: Analysing documents on open questions

Supervisor: Arthur Guillaumin and Mirko Bernardoni

Suitability: Statistics or Data Science

Description:

Master project offer by Clifford Chance (www.cliffordchance.com) the world's pre-eminent law firm with significant depth and range of resources across five continents and member of the "Magic Circle" of leading British law firms.

Bidirectional encoder representations from transformers is a new method of pre-training language representations which obtains state-of-the-art results on a wide array of Natural Language Processing (NLP) tasks. We are interested in having a model that is able to work with legal and security documents and perform reading comprehension.

The project objectives can vary from being able to classify documents and paragraph, extract specific trained datapoints to a more complex full reading comprehension with open questions.

This project is looking to use the cutting edge research in this area like considering Stanford Question Answering Dataset (SQuAD) leader board.

Title: Gaussian Process emulators using GPUs

Supervisor: Professor Serge Guillas

Suitability: MSc Statistics and MSc Data Science

Description: Emulators replace computationally expensive simulators (e.g. climate, geophysical or engineering models). A typical emulator is given by a Gaussian Process (GP) regression using points selected using a design of experiment. Unfortunately, the number of points in the input and output spaces can be large and thus slow down the fitting of the GP. This project is about making use of very recent code that allows the acceleration of the fitting using Graphics Processing Units (GPU). Access will be given to the fastest cluster in the UK to illustrate the benefits of these architectures on some synthetic problems. The language used is python (some knowledge would be good but not necessary). Reference: https://github.com/UCL/gp_emulator

---

Title: Gaussian Process emulators using hierarchical structures

Supervisor: Professor Serge Guillas

Suitability: MSc Statistics and MSc Data Science

Description:

Emulators replace computationally expensive simulators (e.g. climate, geophysical or engineering models). A typical emulator is given by a Gaussian Process (GP) regression using points selected using a design of experiment. Unfortunately, the number of points in the input and output spaces can be large and thus slow down the fitting of the GP. This project is about making use of very recent code that allows the acceleration of the fitting using hierarchical structures of the covariance function. The project involves an assessment of the skills of these methods on some examples. Reference: https://github.com/jiechenjiechen/RLCM

Title: GAS models for robust inference

Supervisor: Professor Jim Griffin

Suitability: MSc Statistics and MSc Data Science

Description: Generalized Autoregressive (GAS) models are a method for constructing time series models with autoregressive structure for the parameters of a wide-range of probability models (such as the variance of a t distributions or the mean of a Poisson distribution). This is achieved by building an autoregressive structure on the score function and has been shown to have good statistical properties. In this project, you will implement GAS models for the variance of heavy tailed distribution and apply them to suitable data. The project will have opportunities for both studying the underlying theory, implementation of computational methods and applications to data.

Title: Bayesian variable selection in linear models with interactions

Supervisor: Professor Jim Griffin

Suitability: MSc Statistics and MSc Data Science

Description: Variable selection methods aim to find a subset of some available variables which are useful for predicting a response. Bayesian methods are attractive as they attach a posterior probability to all possible subsets rather than choosing a single subset. In this project, you will look at modern algorithms for Bayesian variable selection in linear models and consider how they can be extended to linear models with interactions. Inference in these models often assumes strong heredity (that an interaction can only be included if both associated main effects are included). This dependence makes the challenge of finding the posterior distribution more challenging. This project offers the opportunity to consider the computational challenges in Bayesian inference in these models.

Title: Local projections for impulse response function

Supervisor: Professor Jim Griffin

Suitability: MSc Statistics and MSc Data Science

Description: Macro-econometricians are often faced with the challenge of predicting the effect of a change of one variable on other variables. For example, a central bank predicting the effect of a change of the interest rate on inflation, growth and unemployment. This effect measured over time is often called the impulse response function. In this project, you will look at a regression based approach called local projections for impulse response functions. This project offers the opportunity to look at the usefulness of modern regression techniques in estimating the impulse response function.

Title: Investigation of inhibition in point pattern data

Supervisor: Dr Thomas Honnor

Suitability: Suitable for MSc Statistics and MSc Data Science

Description: The foundational point process model is the Poisson point processes, for which point locations are independently distributed. An increased level of complexity is added via the specification of pairwise interaction point processes, through which inhibition is modelled by reduced contributions to the likelihood for nearby points. The simplest such model is the hardcore process for which points cannot be separated by distances less than a specified minimum. Data such as the publicly available locations of trees and mobile phone masts might be modelled by such processes, with biological and economic arguments suggesting that observation locations should be inhibitive.

Students should expect to read up on the theory of point processes, gaining an understanding of Poisson point processes and extending this to pairwise interaction point processes. Real data expected to display inhibitive behaviour will then be modelled using these processes to make inference on the form and extent of the inhibition and how this might differ between data sets. Data analysis will require skilled use of R and an understanding of stochastic processes will also be useful.

Title: Investigation of dependence in multitype point pattern data

Supervisor: Dr Thomas Honnor

Suitability: Suitable for MSc Statistics and MSc Data Science

Description:

Multitype point patterns consist of point locations, each of which is assigned to one of a discrete collection of categories. The traditional real world example is the distribution of two light detecting cells within the eye - each cell has a location and may detect the presence of light or the absence of light. A question of interest is then whether the classes of points may be considered as realisations of independent point processes and the form of any dependency. Data such as the location of supermarkets categorised by company might be investigated as a multitype point process, with economic arguments suggesting a preference for similar high-population areas or different areas with reduced levels of competition.

Students should expect to read up on the theory of point processes, gaining an understanding of nonparametric methods for the summary of basic point patterns and extending this to multitype point processes. Real data for which points may be categorised will then be investigated and the existence and form of dependence determined. Data analysis will require skilled use of R and an understanding of stochastic processes will also be useful.

Title: Investigation of point pattern data

Supervisor: Dr Thomas Honnor

Suitability: Suitable for MSc Statistics and MSc Data Science

Description:

The theory of point processes can be applied to a wide range of real world scenarios in order to better understand how events are spatially distributed. Extensions to observations of multiple types of events, processes in which the location of events is dependent and to comparison of event distributions are all possible and within the scope of a potential MSc project. This project is proposed without reference to a specific data set – students might have a particular application/data set of their own in mind and should feel free to contact me to discuss in more detail the directions we might take.

Students should expect to read up on the theory of point processes, gaining an understanding of nonparametric methods for the summary of basic point patterns and extending this to multitype point processes. Real data for which points may be categorised will then be investigated and the existence and form of dependence determined. Data analysis will require skilled use of R and an understanding of stochastic processes will also be useful.

Title: Workflow architectures and unified interface design of toolboxes for statistical modelling and machine learning in the presence of structure (inc. temporal and hierarchical)

Supervisor: Dr Franz Kiraly

Suitability: MSc Statistics and MSc Data Science

Description:

Despite high importance of structured modelling/learning tasks such as time series forecasting or hierarchical modelling, toolbox support is currently poor – possibly due to the required cross-over expertise in software engineering, statistical modelling, and machine learning.

The types of modelling tasks which are well supported all more or less assume "tabular" data, i.e., data which fits inside an excel table or data frame, most notably the supervised prediction task which is covered by prominent workflow/toolbox packages such as python/sklearn and R/mlr.

Multiple projects are available in extending toolbox support to new tasks or new modelling aspects:

- temporally structured supervised and unsupervised learning tasks such as: time series forecasting, panel data prediction, anomaly detection, and change-point detection

- probabilistic modelling including Bayesian and composite frequentist prediction, density estimation, extreme value analysis and predictive tail distribution modelling
- risk modelling, combined risk/severity modelling, time-to-event modelling
- model tuning, model selection, model comparison and model performance quantification
- workflow orchestration, heterogeneous data source integration, pipelining, benchmarking
- parallelization and integration with high-performance computing intrastructure

Depending on interest and the type of task addressed, there are use case datasets which may serve as a framework test case, from the energy, finance, and health domains.

The project is in collaboration with the Alan Turing Institute, internship opportunities over the summer are also available (separately). Examples of past projects are:

The skpro toolbox for probabilistic supervised modelling:

https://github.com/alan-turing-institute/skpro

The xpandas data container interface for hierarchical data types and feature extraction

https://github.com/alan-turing-institute/xpandas

Students should have a solid background in statistical/ML modelling and experience of software development in R, python, or Julia. They will optimally (but not necessarily) have attended the 2018 iteration of the STATG019 module.

Title: Theoretical and empirical study of generalization error estimates, quantitative model validation and model comparison meta-methodology

Supervisor: Dr Franz Kiraly

Suitability: MSc Statistics and MSc Data Science

Description:

Meta-methodology for external, domain-agnostic validation of models is crucial for success control, especially in times where there are thousands of modelling strategies to choose from, but many of them black-boxes without intrinsic guarantees or model selection quantifiers. Equally, theory for comparing modelling strategies based on widely different theoretical assumptions on equal footing is crucial in any practical situation where the “best choice” is unclear.

In this context, there are a number of topics are available, which may be studied anywhere on a spectrum between theoretical/math-heavy and empirical/experiment-heavy approaches, such as:

- comparative study of model-intrinsic vs model-extrinsic performance guarantees
- model-agnostic estimates of the generalization error for prediction strategies

- hypothesis tests for comparison of prediction strategies and prediction functionals
- quantitative, model-agnostic validation of unsupervised modelling strategies
- central limit theorems and guarantees for model comparison in a sequential setting
- principled methodology for automated, model-agnostic hyper-parameter tuning

Candidates should have a solid background in the theoretical foundations of statistics and probability, and/or experience in conducting simulation studies in R or python. They will optimally (but not necessarily) have attended the 2018 iteration of the STAT3019 module.

Title: Advanced Markov chain Monte Carlo methods for Bayesian computation

Supervisor: Dr Samuel Livingstone

Suitability: MSc Statistics and MSc Data Science

Description: Markov chain Monte Carlo (MCMC) is a powerful method for fitting Bayesian models, and a thriving area of Statistics/machine learning research. There are several directions which can be taken in this project depending on the student, and potentially more than one project available depending on interest. The project would consist of learning about, implementing and improving on various MCMC methods that are used in different settings. The project will require a basic understanding of Markov chains and some enthusiasm for either mathematics, programming or a combination of both (depending on the student).

Title: Bayesian modelling and inference of genetic ancestry

Supervisor: Dr Samuel Livingstone and Dr Ioanna Manolopoulou

Suitability: MSc Statistics and MSc Data Science

Description: Inferring how far back in time and by which mechanism two organisms have a common ancestor is a key question in evolutionary biology (in particular phylogenetics), and there are various novel statistical and machine learning approaches for trying to answer these questions using genetic data and appropriate probabilistic models. The student would learn about some of the basic models used, algorithms for fitting these to data and compare some different approaches commonly used in practice. There is also scope to test out some new ideas. The project will have a good mix of interesting modelling and computation/programming tasks, so would suit a student who is interested in these things, as well as working with real scientific data and answering important questions related to it.

Pre-requisites: Programming in R

References: Hein, J., Schierup, M. H., and Wiuf, C. Gene Genealogies, Variation and Evolution – A Primer in Coalescent Theory. Oxford University Press, 2005.

Title: Convergence of sampling algorithms

Supervisor: Dr Samuel Livingstone

Suitability: MSc Statistics and MSc Data Science

Description:

This project would be for those interested in understanding the mathematics used to quantify and compare the performance of algorithms used in fitting models to data, which are currently very popular in probabilistic machine learning and statistics (particularly in more complex models in which some regularisation is needed). We would focus on sampling methods (most likely Markov chain Monte Carlo), and look at recent developments in convergence using various different metrics on the space of probability distributions, including the Wasserstein distance, total variation, maximum mean discrepancy and possibly some others. This would suit a student with a passion for the mathematical side of the subject.

Title: Predicting the effects of weather changes on demand for children's ambulance services in Greater London

Supervisor: Dr Samuel Livingstone

Suitability: All Programmes

Description:

The clinicians associated with the children's acute transport service (CATS) at Great Ormond Street hospital believe that there are strong short term fluctuations in demand for the service, particularly during the winter months, depending on weather. They have asked us to investigate this and provided 12 years of daily demand data. The student would try to capture this effect by designing and fitting several different models, using the flexible generalized additive models framework (GAMs). There is scope for the work to be published in a journal if all goes to plan. No real pre-requisites are needed, but a good understanding of basic statistical models and a willingness to get their hands data with some real data and experimentation would very advantageous. Dr Christina Pagel from the Clinical Operational Research Unit at UCL would also be involved in this project.

Title: Patterns of Organ Dysfunction in Sepsis

Supervisors: Dr. Giampiero Marra and Dr. Edward Palmer

Suitability: All Programmes

Description:

Abstract

Sepsis is a life threatening condition that kills millions worldwide. It is a very heterogeneous condition; patients present over different timescales, with different degrees of severity, and respond differently to treatment. Major research efforts are currently underway to identify patterns of physiology that might explain this heterogeneity, and thus inform clinical trial design. We have a large data resource containing over 40,000 intensive care episodes from the UK. We are interested in modelling patterns of physiology in septic patients using this data resource. Informative censoring is a major component of the data, and students will need to address this area. The project will be clinically supervised by Dr. Ed Palmer (intensive care clinician) and statistically supervised by Dr. Giampiero Marra. Academic publication is to be encouraged as an output from this project.

Introduction

Sepsis is a heterogenous syndrome of life threatening organ dysfunction caused by infection. Despite myriad potential therapeutic vectors that have shown promise in basic science research, none have been found to be efficacious when translated to humans. The mainstay of current therapies thus remain broad spectrum antibiotics and supportive care; vasoactive drugs, fluid and technology designed to augment or temporarily replace failing organs.

Describing and understanding organ function, and thus dysfunction, is complex. Current popular methods to describe organ dysfunction are based upon outdated expert consensus opinion. Data driven descriptions do exist, however they have not been widely adopted, and it is unclear if they perform any better than the expert led descriptions. At sub day resolution, all current approaches to this problem begin to breakdown.

We are looking for a motivated individual, to apply statistical learning techniques to this problem. The goal is to develop a parsimonious scoring system that is an accurate description of a patients' organ function.

Data

The Critical Care Health Informatics Collaboration (CC-HIC) is a large multi centre research project, aggregating high fidelity longitudinal data on critical care patients from 12 intensive care units across five biomedical research centres in the UK (UCL, Cambridge, Oxford, Guy's and St. Thomas' and

Imperial). 263 variables are available to study including: demographics, acute illness severity scores, high resolution bedside monitoring, drug infusions, microbiology, organ support and outcomes. At present, there are nearly half-a-billion data points inside CC-HIC.

Outputs

The expectation is that the student will devise and validate a new organ dysfunction metric, using data driven techniques. Publication and conference presentation would be strongly encouraged and supported.

Tasks

- Mini-literature review of the topic
- Devise a suitable approach for modelling organ dysfunction in a single organ system (cardiovascular or respiratory)

Prerequisites

The student is expected to have a good working knowledge of R. Python is unlikely to be available during the study period. SPSS and SAS are available, though they are not preferred.

Ethics and Governance

Once the project title has been formalised and agreed with the candidate, it will need to be registered for review by the CC-HIC scientific advisory board (Dr. Palmer will arrange). Otherwise ethics review is already in place.

Support

This project will have the advantage of close supervision and collaboration from clinical world leaders in the field of sepsis research. Proposed direct supervision structure would be:

- Dr. Edward Palmer: Primary supervisor, clinical / data resource
- Dr. Giampiero Marra: Primary supervisor, statistics

Title: Revisiting the use of copula modelling in cost-effectiveness analysis

Supervisors: Giampiero Marra and Manuel Gomes

Suitability: All Programmes

Description: By definition, cost-effectiveness studies are typically interested in making joint inferences about the effectiveness and costs of alternatives interventions being compared. This typically involves multivariate modelling or alternative ways of capturing the joint nature of costs and outcomes. This goes beyond the technical adjustment for the correlation between costs and effects. For example, it addresses the need to make joint hypothesis testing about some model coefficients (e.g. subgroup effects). Another unique feature of cost-effectiveness data is their distributional form. For example, costs are typically highly skewed or semi-continuous and outcomes such as health-related quality of life measures tend to be left-skewed or multimodal.

Parametric joint models beyond the bivariate normal case are considerably complex to implement and often require the use of MCMC methods within a Bayesian framework. A practical alternative approach is to use copulas which can be used to construct a multivariate distribution by making use of the univariate cumulative distribution functions. A key strength of this approach is its flexibility to combine different types of marginal distributions, and it can model more complex dependences between the cost-effectiveness endpoints. The aim of this project is to revisit the potential of copula modelling in cost-effectiveness analysis compared to conventional methods for producing joint inferences. The copula approach will be illustrated in a CEA of the REFLUX study evaluating laparoscopic surgery for the management of patients with reflux disease. Some familiarity with R is desirable.

Title: Addressing missing data in observational studies with time-varying confounding

Supervisors: Giampiero Marra and Manuel Gomes

Suitability: All Programmes

Description:

There is now increasing interest in using large observational studies to inform estimates of treatment effects to help agencies like NICE make recommendations about which health interventions to provide. The major concern with relying on such longitudinal, routinely-collected data is (time-varying) confounding by indication. This is a recurrent issue because patient's progression typically influences future treatments and outcomes, but it is also affected by previous treatments. A related problem is that typically these routine data sources, collected in response to clinical need, have incomplete information on the outcomes of interest and potential confounders, which can magnify the biases and raise additional challenges for tackling the confounding.

Missing data raises new challenges for the use of standard methods for addressing time-varying confounding such as inverse probability weighting (IPW)-based marginal structural models (MSMs) or G-estimation, not least because the patterns of missing data tend to be non-monotone. This project will consider alternative approaches more suitable for tackling non-monotone missing data, such as multiple imputation (MI). The project will explore combining MI with MSMs and compare that with using conventional IPW censoring weights. The methods will be illustrated in a case-study estimating the effectiveness of biological drugs for treating patients with rheumatoid arthritis, using data from the US National Data Bank for Rheumatic Diseases.

Title: Forecasting movement in football betting markets

Supervisors: Giampiero Marra

Suitability: MSc Statistics and MSc Data Science

Description:

In sports gambling markets, typically the amount of money that it is possible to bet increases as you approach kick-off. At the same time, however, the market typically becomes more efficient, leading to a volume-price tradeoff. A common problem is whether to place a low volume but generously priced bet now, or wait for higher volumes at a potentially worse price.

Smartodds, a private consultancy that provides services for professional gamblers, can provide a rich dataset of pre-match market prices for the main betting lines for supremacy, total goals and match outcome markets in football matches (e.g., respectively, Manchester City to win by more than 2 goals; more than 3 goals scored in the match; Manchester City to win) from a collection of top-flight and second tier European leagues. For each match we can provide data on how the price of these betting lines evolved in the minutes, hours and days before kick-off, from the opening price (when the market opens) to the closing price (when the match starts).

The project will try to answer the following question: If the currently available odds are X, then how far from X might the odds be when the match actually starts?

Questions that may help to frame this problem are, for example: How does this depend on how long there is left before the match starts? Does it depend on the league or market type? Are big, early moves in a particular direction predictive of later market moves?

Students could do this via any method that appealed, but one option could be to model odds moves as a random walk and try to assess how, for example, time left until the match starts, the specific league and market type affect the size of possible moves. For example, higher profile leagues typically attract more betting action and the market prices may therefore be more volatile.

The student should be able to code in R, or python.

Title: Evaluating the efficiency of football betting markets

Supervisors: Giampiero Marra

Suitability: MSc Statistics and MSc Data Science

Description:

Two facts: (1) it is possible to gamble on almost any football league and (2) betting markets tend to get more efficient over time. But is the market equally efficient in every league, how quickly is it improving and is the market a better predictor at the end of a season than at the beginning? Smartodds, a private consultancy that provides services for professional gamblers, can provide win and score probabilities implied by the closing market betting odds for the outcome of football matches across various global football leagues since 2010. For example, these will tell you what probability the market assigned to a team winning, drawing or losing a game (1X2), by how many goals they were expected to win on average (supremacy) and how many goals you would expect to see scored in a game (total goals).

This project will use regression techniques (e.g. General Additive or Linear Mixed models) or machine learning to try to determine whether the quality of market predictions for these leagues has changed over time, whether it varies across leagues and whether it varies over the course of a season. Quality of market predictions can be assessed by looking at, for example, the Brier score for 1X2 win probabilities, and RMSE for supremacy and total goals predictions, but there are many ways that this could be done depending on the student's interest and expertise.

The student should be able to code in Python or R.

Title: Student proposed project

Supervisors: Giampiero Marra

Suitability: All Programmes

Description:

A project proposed by the student in the area of survival analysis, copula regression modelling, distributional regression, or penalised spline regression.

Title: Mixture modeling of cell subtypes in flow cytometry

Supervisor: Ioanna Manolopoulou

Suitability: All Programmes

Description:

In biotechnology, flow cytometry allows efficient measurement of various protein levels on the surface of cells; a typical dataset contains tens to hundreds of thousand cell measurements. In traditional flow cytometry, cells were separated into subtypes by manual inspection (also called gating) of the distribution of measurements. Gaussian mixture modelling allows us to automatically identify the structure and location of potential cell subtypes by associating each cell subtype to a Gaussian density. This project will investigate different approaches for fitting a mixture model to a flow cytometry dataset and compare the results.

Pre-requisites: Programming in R

References: C. Chan, F. Feng, J. Ottinger, D. Foster, M. West, T.B. Kepler (2008) Statistical mixture modeling for cell subtype identification in flow cytometry. Cytometry A. 73(8): 693-701
G. McLachlan, D. Peel (2000) Finite Mixture Models. Wiley.

Title: Modelling ring galaxies using Gaussian wells

Supervisor: Ioanna Manolopoulou

Suitability: Suitable for MSc Statistics and MSc Data Science

Description:

The aim of this project is to construct a model for ring galaxies, so that they can be automatically identified and characterised. The project will use the concept of a 2-dimensional 'Gaussian well', where one Gaussian distribution has a hole in the middle due to a smaller Gaussian being subtracted, resulting in a ring-like structure. This concept of a Gaussian well will be implemented on imaging data from a ring galaxy, where we use thresholding to split the image into 'on' and 'off' pixels, the 'on' treated as observations.

Pre-requisites: Programming in R

References: Ioanna Manolopoulou, Thomas B. Kepler, and Daniel M. Merl. "Mixtures of Gaussian wells: Theory, computation, and application." Computational statistics & data analysis 56.12 (2012): 3809-3820.

Title: Modelling count processes using the Hawkes process

Supervisor: Ioanna Manolopoulou

Suitability: Suitable for MSc Statistics and MSc Data Science

Description:

The standard Hawkes process is constructed from a homogeneous Poisson process where each event produces offspring according to some exciting function. Generalisations of the Hawkes process allow for a self-exciting process of an inhomogeneous Poisson process, with different forms of exciting functions, as well as multivariate versions of the process. This project will explore various aspects of the Hawkes process, investigate the properties of its different flavours and assess its fit to a variety of datasets.

Pre-requisites: Programming in R. General knowledge of the Poisson process.

References: Fierro, R., Leiva, V., & Møller, J. (2015). The Hawkes Process with Different Exciting Functions and its Asymptotic Behavior. Journal of Applied Probability, 52(1), 37-54.

Hawkes, Alan G., and David Oakes. "A cluster process representation of a self-exciting process." Journal of Applied Probability 11.03 (1974): 493-503.

Møller, Jesper, and Jakob G. Rasmussen. "Perfect simulation of Hawkes processes." *Advances in applied probability* 37.03 (2005): 629-646.

Title: Bayesian Multi-State Transition Rate Modelling

Supervisor: Owen Nicholas and Ardo Van Den Hout

Suitability: All Programmes

Description:

Progression of an individual through a sequence of states, for example categories of health (no disease, stage 1 disease, stage 2 disease, … dead), or employment (employed, unemployed, retired), or education etc, can be described using continuous time markov multi-state models. These are really useful for understanding rates of transition between states, and for simulating trajectories of states.

When it comes to parametric Bayesian analysis of transition rates from data there are many challenges, both numerical and statistical, including efficient exponentiation of matrices and effective methods for sampling from the posterior, as well as the potential for heterogeneous transition rates between individuals, and choice of prior. This project will develop Bayesian Markov chain Monte Carlo approaches to sampling transition rate matrices, focusing on data from lung transplant patients. R, Python or Matlab are suggested as suitable packages for coding, in order to master the technical aspects of the project.

Title: Creating an R package

Supervisor: Dr Paul Northrop

Suitability: Suitable for MSc Statistics and MSc Data Science

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 creating a set of functions to perform individual aspects of these analyses, annotating this code, providing a description of how to use each function, giving some illustrative examples and writing a vignette. A vignette is a guide to the package that gives the users an overview of the problem that the package tackles and how it does this. The purpose of the package requires some careful thought and will need to be discussed with me prior to selecting the project. The package should not replicate very closely the purposes of existing R packages. One possibility is to write code to make it easy for fellow students to perform analyses that you have seen in Statistics modules that you have taken. Obviously this project would suit a student with an aptitude for programming, who has enjoyed learning how to use R in STAT0030 and wants to take their R programming further.

Title: Using Bayesian extreme value modelling to preform open set classification

Supervisor: Dr Paul Northrop

Suitability: Suitable for MSc Statistics and MSc Data Science

Description: In statistical classification we seek to identify to which class a new object belongs, based on a training set that contains quantitative information about objects (features) and their known class labels. In open set classification we allow the possibility of classifying a new object as originating from an unknown class, that is, a class that it is not represented in the data. In a recent paper (https://arxiv.org/abs/1808.09902) Vignotto and Engelke perform this task using models that originate from extreme value theory. The idea is to quantify how distant the features of a new observation are from observations in the training set. Vignotto and Engelke use a frequentist approach to this problem. The aim of this project is to develop a Bayesian approach and to compare its performance with the frequentist approach. Students who took Topic 1 of STAT0017 will be familiar with the extreme value models involved in this project. However, this project is also open to, and suitable for, other students.

Title: Analysis of Amateur Radio Contest Data

Suitability: MSc Statistics, MSc Data Science

Supervisor: Dr Yvo Pokern

Description:

Radio amateurs regularly conduct “contests”, a type of operation where each participating radio amateur attempts to make as many radio contacts with other participating radio amateurs all over the world as possible within a given time span (e.g. 2 hours, 24 hours or 48 hours). In each contact, a small amount of information is exchanged: the call-signs identifying the two radio amateurs involved in the contact, the strength of the received signal, the radio frequency of operation and a running identifier (i.e. number of contact). This information is collected by each participating radio amateur in a contest log, usually through use of standard software that also captures the date and time on which the contact took place. This log is then submitted electronically to the contest organizers at the end of the contest. A scoring system is used to award points for each contact and determine the winner of the contest. Unfortunately, the data submitted by participating radio amateurs do not always match-up: sometimes, only one of the two parties in a radio contact acknowledges the contact took place, sometimes the information (frequency, identifier, signal strength) does not match, so that some data quality issues may need to be addressed.

The project aims to study the frequency with which contacts are made – it is likely that contacts become more frequent when the reflecting and refracting properties of the ionosphere improve propagation of the frequencies used. Data is available through on-line interfaces that could be queried using scripting; a significant component in this direction may make the project suitable for the MSc in Data Science (with statistical focus). The project has a significant exploratory component so that other directions may be pursued by the student as supported by the data available.

Prerequisites:

1. a willingness to learn about amateur radio contest operation and basics of shortwave radio propagation
2. general facility with computers to enable extraction of data, format conversion etc.
3. ability to operate a standard statistical software package to analyse the data, e.g. R
4. facility with scripting (e.g. bash/python or similar) to enable web content collection if undertaken in the framework of the MSc in Data Science

Title: Conditions for and Testing the Composition Property

Suitability: MSc Statistics and MSc Data Science

Supervisor: Kayvan Sadeghi

Description:

It is well-known that if random variables A and B and random variables A and D are independent, A is not necessarily independent of (B,D). A generalization of this property when conditioning on an additional variable C is called the \emph{composition} property. The goal of the project is to investigate which distributions satisfy this property. This can be done through theoretical methods (notably using algebraic statistics or the theory of exponential family models), but also through simulation studies. Another approach is to devise algorithms that test whether composition is satisfied given data.

Title: Restrictiveness of Discrete Determinantal Point Processes

Suitability: MSc Data Science

Supervisor: Kayvan Sadeghi

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, probability and algebra. The main goal of this project is to ascertain how restrictive these models are by studying their corresponding \emph{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: Model Polytopes for Exponential Random Graph models

Suitability: MSc Statistics and MSc Data Science

Supervisor: Kayvan Sadeghi

Description:

Studying the geometry of random network models provides a systematic
framework for parameter estimation in network models as well as understanding asymptotic behavior of these models. In particular, finding the corners of model polytope is an important task for maximum likelihood estimation. The goal of the project is to understand and review the literature on the model polytope for networks. In particular, we want to show that, for non-exchangeable network models in exponential family form, the corners of the model polytope are the graphs that are uniquely realizable by the corresponding sufficient statistics. One example
of this statement is beta models where the corners of the model polytope are graphs that are uniquely realizable by the degree sequence.
In addition, what can we say, based on the understanding acquired for exchangeable network models? Some basic knowledge of convex geometry and exponential family models are required.

Title: Causal networks with "weak associations"

Supervisor: Dr Ricardo Silva

Suitability: All Programmes

Description:

Causal networks are representations of cause-effect relationships that under some conditions allow for the estimation of causal effects using observational data. Such structures however may be hard to elicit from background knowledge only. Machine learning algorithms exist that allow for estimating partial structures, but they can be unreliable if associations in the data are weak. We will investigate methods robust to weak associations. See http://auai.org/uai2017/proceedings/papers/229.pdf for an example of (specialized) reading to give some context, but I don't expect students to fully understand this paper at this stage – this is more to show some of the motivation.

Title: Causality in reinforcement learning

Supervisor: Dr Ricardo Silva

Suitability: All Programmes

Description:

The field of reinforcement learning consists of methods for learning to plan a sequence of actions out of choosing promising candidates given data collected at any stage of the estimation process. This typically requires large sample sizes and constant interaction with the environment. In this project we will explore ways of leveraging observational data to aid reinforcement learning. See https://arxiv.org/abs/1812.10576 for an example.

Title: Modelling causal effects in social and spatial networks and other dependent data

Supervisor: Dr Ricardo Silva

Suitability: All Programmes

Description:

See https://qaps.princeton.edu/sites/default/files/q-aps/files/spatial-kriging-2016-04-01.pdf for background

---

Title: Modelling causal effects on time-series data with natural experiments

Supervisor: Dr Ricardo Silva

Suitability: All Programmes

Description:

See https://ai.google/research/pubs/pub41854 for an example of problem and methodology.

Title: Student-led Project

Supervisor: Dr Ricardo Silva

Suitability: All Programmes

Description:

I'm open to student-led projects in areas such as deep learning on graphs, variational autoencoders and other problems related to efficient and effective algorithms for approximate inference in complex probabilistic models.

Title: Large Scale Analysis of User Behaviour with Spotify Data

Supervisor: Dr Ricardo Silva

Suitability: MSc Statistics and MSc Data Science

Description:

Several opportunities for developing user behaviour models using public Spotify and related data. Description of interesting datasets and problems of interest are listed below. Students are encouraged to come up with their own variants of interesting problems. Please notice that solid programming skills are required.

The Music Streaming Sessions Dataset

https://arxiv.org/pdf/1901.09851.pdf

https://www.crowdai.org/challenges/spotify-sequential-skip-prediction-challenge

Spotify Million Playlist dataset

https://recsys-challenge.spotify.com/readme

Last.fm music events dataset

http://www.cp.jku.at/datasets/LFM-1b/

WSDM 2018 music recommendation dataset

https://www.kaggle.com/c/kkbox-music-recommendation-challenge/data

WSDM 2018 Churn Prediction dataset

https://www.kaggle.com/c/kkbox-churn-prediction-challenge/data

---

Title: Shazam for earthquakes – An application to a regional earthquake catalog

Supervisor: Dr Katerina Stavrianaki

Suitability: MSc Statistics and MSc Data Science

Description:

In the recent years statistical seismology is experiencing rapid growth in the quantity of data. Earthquake detection—identification of seismic events in continuous data— is a fundamental operation for observational seismology.

Missing aftershocks, particularly after large earthquakes, can lead to erroneous estimation of the parameters in statistical models and consequently bias forecasts and seismic hazard assessments. The student will use the Fingerprint And Similarity Thresholding (FAST), an earthquake-detecting algorithm based on song- matching app Shazam, developed by a group of seismologists in Stanford University to identify missing earthquakes from an earthquake catalog of their choice.

The aim of this project is to identify missing earthquakes from a regional earthquake catalog using an algorithm that detects earthquakes by processing the waveforms recorded in the seismic networks.

The FAST algorithm will be given to the student, however the algorithm only runs on Linux systems. Knowledge of Python is also required.

Reference: Yoon, C. E., O'Reilly, O., Bergen, K. J., & Beroza, G. C. (2015). Earthquake detection through computationally efficient similarity search. Science advances, 1(11), e1501057.

Title: De-clustering earthquake catalog data

Supervisor: Dr Katerina Stavrianaki

Suitability: MSc Statistics and MSc Data Science

Description:

The Epidemic Type Aftershock Sequence (ETAS) model is the most widely used statistical model in seismology. It's a Hawkes Type self exciting stochastic model and the corresponding process that describes earthquake occurrence is equivalent to a branching process where each point can be generated either as a background event or triggered by a specific previous earthquake. This allows us to identify clusters of earthquakes and this identification is useful for seismic hazard. The aim of this project is to use machine learning clustering methods and compare the results with the outcomes of the ETAS model. R programming language skills are necessary.

Title: Parametric time-dependent multi-state survival models

Supervisor: Dr Ardo Van Den Hout and Owen Nicholas

Suitability: All programmes

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 are flexible parametric models that can describe transition hazards that increase at first, and decrease at a later time.

The project starts with exploring an R package for basic multi-state survival models, and coding the corresponding likelihood function. The next step is the extension to more flexible models.

Title: Bivariate discrete distributions to model cognitive function

Supervisor: Dr Ardo Van Den Hout

Suitability: All programmes

Description:

Longitudinal data are available for cognitive function in the older population. Using two cognitive tests, cognitive function is measured repeatedly on a bivariate discrete scale. The project is about using a bivariate discrete distribution to describe change of cognitive function over time within individuals.

The project starts with data defined by one measurement per individual. The binomial distribution and extension thereof will be applied. Next, a random-effects model will be specified for the repeated measurements.

Software: R.

Title: Generalised time-dependent logistic models for survival data

Supervisor: Dr Ardo Van Den Hout

Suitability: All programmes

Description:

The generalised time dependent logistic family comprises a wide range of models for time-to-event data. This project will investigate these models and compare them with standard models such as the Weibull and the Gompertz.

The data for this project are from a longitudinal study of bronchiolitis obliterans syndrome from lung transplant recipients. A standard survival model can be defined for death as the event. In addition, a three-state survival model can be defined consisting of two living states (presence and absence of the syndrome) and a third absorbing dead state. For both these models, the generalised time dependent logistic family will be investigated.

The project starts with the standard survival model and the corresponding likelihood function. The likelihood function can be maximised using a general-purpose optimiser. The next step is the extension to the three-state model. Software: R.

Title: False discovery control in multiple testing

Supervisor: Dr Tengyao Wang

Suitability: Suitable for MSc Statistics and MSc Data Science

Description:

The landscape of statistical hypothesis testing has changed significantly in recent years. Simultaneous testing of thousands of hypotheses is now commonplace in a wide range of application areas. Since the seminal work of Benjamini and Hochberg (1995), many procedures have been suggested to control false discovery rate in the context of multiple testing. In this project, the candidate is expected to survey the literature and compare the performance of existing methods (either empirically or theoretically). A more ambitious candidate can look at possible modifications of existing methods to achieve better performance in some specific settings.

Prerequisites: none.

Reference

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc., Ser. B*, **57**, 289--300.

Title: Matrix completion

Supervisor: Dr Tengyao Wang

Suitability: Suitable for MSc Statistics and MSc Data Science

Description:

Missing data is the rule rather than the exception in the era of Big Data. For example, in the famous Netflix user movie rating dataset, only a very small proportion of all possible ratings are observed. To be able to predict missing ratings from observable ones is highly valuable and Netflix once provided a million-dollar cash prize to find the best missing value prediction algorithm. A key idea in many of the best performing methods is to perform low rank matrix completion by solving a convex optimisation problem. The goal of this project is to give an overview of the low rank matrix completion method from a statistical perspective. Then, the candidate can either look at the performance guarantees of the matrix completion method or apply the algorithm to a real world dataset chosen by the candidate.

Prerequisites: linear algebra (eigendecomposition, matrix norms) and a good command of one programming language (e.g. Matlab, python).

Title: Principal component analysis in high dimensions

Supervisor: Dr Tengyao Wang

Suitability: Suitable for MSc Statistics and MSc Data Science

Description:

Principal component analysis (PCA), which involves projecting a sample of multivariate data onto the space spanned by the leading eigenvector of the sample covariance matrix, is one of the oldest and most widely-used dimension reduction techniques in statistics. However, the work of Johnstone and Lu (2009) and Paul (2007) shows that PCA breaks down in the high-dimensional setting that are frequently encountered in many modern application areas. Several modifications of the classical PCA has since been proposed to address this issue under various structural assumptions. This project will compare different proposed approaches for high-dimensional PCA and apply a specific method to a high-dimensional dataset chosen by the candidate.

Prerequisites: Linear algebra (eigendecomposition, matrix norms) and a good command of one programming language (e.g. R, Matlab, python).

References: Johnstone, I. M. and Lu, A. Y. (2009) On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.*, **104**, 682--693.

Paul, D. (2007) Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica.*, **17**, 1617--1642

Title: Optimised aggregation of citizen science data for biomedical image analysis

Supervisor: Dr Jinghao Xue; Dr Martin Jones (The Francis Crick Institute)

Suitability: All programmes

Description:

The use of crowdsourced “citizen science” analysis has proved to be a valuable tool in the analysis of large amounts of data, particularly in tasks where human visual processing still outperforms existing computational methods. Following on from successful projects in other fields of research, such as Galaxy Zoo [1], our project Etch a Cell obtains image segmentations from thousands of non-expert volunteers for our volume electron microscopy data [2]. A critical step is the aggregation of these data, where annotations from multiple users are combined to create a final high-quality annotation for each image. This project aims to develop a robust and optimised method for performing this aggregation, to help provide training data for downstream machine learning.

[1] Lintott et al.,Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey(2008) MNRAS doi: 10.1111/j.1365-2966.2008.13689.x

[2] Peddie & Collinson,Exploring the third dimension: volume electron microscopy comes of age(2014) Micron doi: 10.1016/j.micron.2014.01.009

Title: Semi-supervised machine learning to classify Raman images of ovarian cancer

Supervisor: Dr Jinghao Xue; Prof Geraint Thomas (Cell and Developmental Biology, UCL)

Suitability: All programmes

Description: A hyperspectral Raman dataset from ovarian cancer patients requires classification into one of two types of cancer. However, as with many biomedical problems, it only has a few labelled Raman images (9 in each group). This project aims to develop a semi-supervised machine learning method to classify the data in this dataset.

Title: Classifying clustered Raman data of gastro-intestinal cancers

Supervisor: Dr Jinghao Xue; Prof Geraint Thomas (Cell and Developmental Biology, UCL)

Suitability: All programmes

Description:

The consistency between Raman spectrometers has yet to established, contributing to slow clinical adoption of the technique. In the SMART dataset, three different Raman spectrometers at different centres were used to classify Gastro-Intestinal (GI) cancers into 5 groups, creating three similar datasets with a hierarchical structure. This project aims to develop a classification method taking into account this structure for this dataset.

Title: Classification of pseudo cancer versus polyp cancer from Raman images

Supervisor: Dr Jinghao Xue; Prof Geraint Thomas (Cell and Developmental Biology, UCL)

Suitability: All programmes

Description:

Description: Epithelial misplacement is a benign process which can occur as colon samples are handled. They are often misdiagnosed as adenocarcinomas as they share many visual features (infiltration into the submucosa). Successful classification via Raman spectroscopy could reduce the number of false positives. This project aims to develop a method to classify a dataset of 36 Raman images into polyp cancer or pseudo cancer.

Title: Meta-analysis of kappa statistics for colon cancer assessment

Supervisor: Dr Jinghao Xue; Prof Geraint Thomas (Cell and Developmental Biology, UCL)

Suitability: All programmes

Description:

Description: The inter-rater reliability and the intra-rater reliability of pathologists assessing cancer have long been noted. However, the data have yet to be collected in a meta-analysis to confirm this statistical effect. Focusing on colon cancer, this project aims to conduct a systematic literature review and meta-analysis of derived kappa statistics, taking into account statistical heterogeneity between the studies.

Title: Representation-based classification

Supervisor: Rui Zhu & Jinghao Xue

Suitability: MSc Statistics, MSc Data Science

Description:

Representation-based classification methods have been successfully applied to various fields, such as face recognition and hyperspectral image classification. Famous representation-based classification methods include sparse representation-based classification (SRC) [1] and collaborative representation-based classification (CRC) [2]. SRC and CRC first represent a test instance by using all training instances and then classify it to the class with the smallest reconstruction error. A recent study proposes to select the k-nearest classes (KNC) to represent the test instance, rather than using all training instances [3]. This project aims to study and improve representation-based classification methods, as well as compare them on real-world data.

References

[1] Wright, John, Allen Y. Yang, Arvind Ganesh, S. Shankar Sastry, and Yi Ma. "Robust face recognition via sparse representation." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, no. 2 (2009): 210-227.

[2] Zhang, Lei, Meng Yang, and Xiangchu Feng. "Sparse representation or collaborative representation: Which helps face recognition?" In *2011 International Conference on Computer Vision*, pp. 471-478. IEEE, 2011.

[3] Zheng, Chengyong, and Ningning Wang. "Collaborative representation with k-nearest classes for classification." *Pattern Recognition Letters* 117 (2019): 30-36

The below is taken from pages 25-27 in the Postgraduate Student Handbook

RESEARCH PROJECT

Guidelines for preparation and submission

Students should plan to take a short break after their written examinations, before starting work on their projects. All supervisors are likely to be away from time to time during the period June-September, attending conferences or on holiday. Students should therefore see their supervisors as soon as their examinations are over, to make mutually convenient arrangements for starting work on their projects.

Over the course of the project, student and supervisor should arrange to meet regularly (about once a week, whenever possible) and should agree a suitable timetable for completing the work and producing a written account. The supervisor should advise the student to start to write up the work, and to ask for the supervisor's feedback on their writing, early in this period.

Supervisors will provide feedback on an entire draft of the project dissertation on at least one occasion, providing it is available in at least three weeks before the deadline for submission. Any request for feedback after this deadline is at the discretion of the supervisor. Supervisors should provide feedback within two weeks.

Final (word-processed) dissertations should be handed in to the Teaching & Learning Office by 16:00 on the advertised date (this is normally at the start of September). Late submissions will incur severe “lateness” penalties (see “Late Submission Penalties” section on page 30). Furthermore, an electronic version of the dissertation should be submitted via Moodle on the same day (the MSc Tutor will circulate more detailed instructions nearer to the date).

The length of a project dissertation will depend on the topic of the project and may vary considerably. Lengths between 8,000 and 15,000 words (excluding computer programs, tables, graphs, formulae and other output) are generally acceptable. Typical projects are between 10,000 and 12,000 words long.

Each dissertation should include a table of contents, an introduction, a conclusion or discussion section, and a list of references. The reference list should include all references that have been used to support the work reported in the project; and these references should be cited in the text of the dissertation as appropriate to indicate where they have been used, following accepted conventions for citation. The pages should be clearly numbered and should have a left-hand margin of at least 2cm. Examiners attach *considerable* importance to accuracy, clarity and overall quality of presentation.

In addition to the project dissertation, each student will be required to give a presentation on their research. The time normally allocated to each presentation is 15 minutes excluding questions. Students are expected to attend and actively participate in the oral presentations by other students. Presentations normally take place in early September; students therefore need to ensure that they are available in the Department at this time.

Specific dates for the arrangements referred to in the third and fourth paragraphs above will be provided separately. *Please ensure that you are aware of them.*

Guidelines for assessment

Project dissertations are read independently by two examiners, one of whom is normally the candidate's project supervisor. Each examiner provides a brief written assessment. A selection of dissertations are also read by a visiting examiner. The final mark is agreed by the whole exam board, which includes the visiting examiner. The final mark should be interpreted in accordance with the guidance notes on page 15.

Examiners will satisfy themselves that the dissertation is the work of the candidate, and will take into account the following points:

- the difficulty and novelty of the project;
- the amount of new methodology/ application knowledge that the student was required to learn;
- the degree of direction required from the project supervisor;
- the student's progress throughout the project.

Subject to these overall criteria, examiners will consider both the content of the dissertation and its presentation, with a higher priority being attached to content. Aspects considered will usually include the following:

- *Content*: amount of work done; extent to which understanding has been demonstrated; quality and accuracy of reasoning, validity of interpretation, relevance of conclusions; critical appraisal, discussion of limitations and suggestions for further work; clarity of objectives; quality of literature review; quality of data organisation and collection (if applicable); quality of programming or use of software (if applicable).

- *Presentation*: layout of dissertation and care in its presentation; structure of the dissertation; use of appropriate judgement in selecting material; clarity of expression, readability and coherence; correctness of grammar and spelling; adequacy of diagrams, graphs and tables (if applicable); quality of presentation of mathematical material (if applicable).

A mark less than 50 will be awarded if the material, though correct, is judged to be wholly reproduced in a purely technical manner.

For a mark over 85, it is expected that the student, in addition to having submitted a well-presented dissertation demonstrating a good understanding of the material and a comparatively high amount of work, will also have shown some initiative rather than simply following instructions. Marks of 90 or more may be appropriate where in addition the technical or conceptual difficulty of the material is very high, or where some of the work could be considered original research on the part of the student.

The length of project dissertation will depend on the topic of the project and may vary considerably. Lengths between 8,000 and 15,000 words (excluding computer programs, tables, graphs, formulae and other output) are generally acceptable. Typical projects are between 10,000 and 12,000 words long. Over-length dissertations will be penalised (see page 30). It is generally required that the amount of work done and demonstrated is high enough, and that the material is presented in a way understandable to fellow students with a comparable background (so 8,000 words may only be an appropriate length for a very theoretical or densely presented dissertation). On the other hand, dissertations should not be too repetitive or contain unnecessary or irrelevant details, which may lead to downmarking.

Although the word counts given above exclude appendices, tables and program listings, these items will also be penalised if they are excessive.

Each project presentation will be assessed by two examiners. Normally, neither of the examiners will be the candidate's supervisor. The examiners make independent notes on the presentation prior to discussing and agreeing a mark. Aspects considered will usually include the following:

- *Content*: was the presentation interesting? Did it focus on the important aspects of the work and flow logically? Was there sufficient detail to be intelligible to statistically literate listeners who do not have an in-depth knowledge of the specific topic? Were there clear aims and conclusions?
- *Presentation skills*: was the verbal presentation confident and clearly audible with varied inflexion? Did the presentation engage with the audience? Were visual aids clear, well produced and well used? Were questions handled appropriately? Was the amount of material appropriate for the time allowed?