UCL DEPARTMENT OF STATISTICAL SCIENCE

MSc PROJECTS GUIDE

Medical Statistics 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 Medical Statistics)

PROJECT TITLE	SUBMITTING SUPERVISOR	ADDITIONAL SUPERVISORS	PROGRAMME SUITABILITY			PAGE NUMBER
			MSc Statistics	**MSc Data Science**	**MSc Medical Statistics**	
AN INVESTIGATION OF STOPPING RULES IN RANDOMISED TRIALS	Ambler, Gareth	N/A	Yes	No	Yes	1
AN INVESTIGATION OF THE EFFICIENT USE OF BASELINE MEASUREMENTS OF OUTCOME IN RANDOMISED TRIALS	Ambler, Gareth	Barber, Julie	No	No	Yes	1
ADDRESSING MISSING NOT-AT-RANDOM DATA IN COST-EFFECTIVENESS ANALYSIS	Baio, Gianluca	Gomes, Manuel	Yes	No	Yes	2
EXPLORING COST-EFFECTIVENESS OF THE INTRODUCTION OF THE HCM RISK-SCD RISK MODEL IN CLINICAL PRACTICE	Baio, Gianluca	Omar, Rumana	Yes	No	Yes	2
DEVELOPMENT OF AN INTERACTIVE R/SHINY TOOL FOR NMA OF TRIAL-LEVEL DATA	Baio, Gianluca	Berardi, Andrea	Yes	No	Yes	3
PREDICTING THE EFFECTS OF WEATHER CHANGES ON DEMAND FOR CHILDREN'S AMBULANCE SERVICES IN GREATER LONDON	Livingstone, Samuel	N/A	Yes	Yes	Yes	4
PATTERNS OF ORGAN DYSFUNCTION IN SEPSIS	Marra, Giampiero	Palmer, Edward	Yes	Yes	Yes	5
ADDRESSING MISSING DATA IN OBSERVATIONAL STUDIES WITH TIME-VARYING CONFOUNDING	Marra, Giampiero	Gomes, Manuel	Yes	Yes	Yes	7
REVISITING THE USE OF COPULA MODELLING IN COST-EFFECTIVENESS ANALYSIS	Marra, Giampiero	Gomes, Manuel	Yes	Yes	Yes	7
MIXTURE MODELING OF CELL SUBTYPES IN FLOW CYTOMETRY	Manolopoulou, Ioanna	N/A	Yes	Yes	Yes	8
STUDENT PROPOSED PROJECT	Marra, Giampiero	N/A	Yes	Yes	Yes	8
BAYESIAN MULTI-STATE TRANSITION RATE MODELLING	Nicholas, Owen	Van Den Hout, Ardo	Yes	Yes	Yes	9
TIME-TO-EVENT MODELLING WITH NON-PROPORTIONAL HAZARDS	O'Keeffe, Aidan	N/A	Yes	No	Yes	9
INVESTIGATING THE USE OF PERMUTATION TESTING IN RANDOMISED EXPERIMENTS	O'Keeffe, Aidan	N/A	Yes	No	Yes	10

COMPARING CLINICAL PERFORMANCE AND DETECTING OUTLIERS	Pavlou, Menelaos	Ambler, Gareth	Yes	No	Yes	10
MODELS FOR HUMAN HIP AND KNEE JOINT MOVEMENT	Pokern, Yvo	N/A	Yes	No	Yes	10
CAUSAL NETWORKS WITH "WEAK ASSOCIATIONS"	Silva, Ricardo	N/A	Yes	Yes	Yes	11
CAUSALITY IN REINFORCEMENT LEARNING	Silva, Ricardo	N/A	Yes	Yes	Yes	11
MODELLING CAUSAL EFFECTS IN SOCIAL AND SPATIAL NETWORKS AND OTHER DEPENDENT DATA	Silva, Ricardo	N/A	Yes	Yes	Yes	12
MODELLING CAUSAL EFFECTS ON TIME-SERIES DATA WITH NATURAL EXPERIMENTS	Silva, Ricardo	N/A	Yes	Yes	Yes	12
STUDENT-LED PROJECT	Silva, Ricardo	N/A	Yes	Yes	Yes	12
PARAMETRIC TIME-DEPENDENT MULTI-STATE SURVIVAL MODELS	Van Den Hout, Ardo	Nicholas, Owen	Yes	Yes	Yes	12
BIVARIATE DISCRETE DISTRIBUTIONS TO MODEL COGNITIVE FUNCTION	Van Den Hout, Ardo	N/A	Yes	Yes	Yes	13
GENERALISED TIME-DEPENDENT LOGISTIC MODELS FOR SURVIVAL DATA	Van Den Hout, Ardo	N/A	Yes	Yes	Yes	13
CLASSIFYING CLUSTERED RAMAN DATA OF GASTRO-INTESTINAL CANCERS	Xue, Jinghao	Thomas, Geraint	Yes	Yes	Yes	14
OPTIMISED AGGREGATION OF CITIZEN SCIENCE DATA FOR BIOMEDICAL IMAGE ANALYSIS	Xue, Jinghao	Jones, Martin	Yes	Yes	Yes	14
SEMI-SUPERVISED MACHINE LEARNING TO CLASSIFY RAMAN IMAGES OF OVARIAN CANCER	Xue, Jinghao	Thomas, Geraint	Yes	Yes	Yes	14
CLASSIFICATION OF PSEUDO CANCER VERSUS POLYP CANCER FROM RAMAN IMAGES	Xue, Jinghao	Thomas, Geraint	Yes	Yes	Yes	15
META-ANALYSIS OF KAPPA STATISTICS FOR COLON CANCER ASSESSMENT	Xue, Jinghao	Thomas, Geraint	Yes	Yes	Yes	15

MSc Project Descriptions for 2019

Title: An Investigation of Stopping rules in Randomised Trials

Supervisor: Dr Gareth Ambler

Suitability: MSc Statistics (Dependent on chosen modules) and MSc Medical Statistics

Description:

Interim analyses are often performed in randomised trials to investigate both safety and efficacy while the trial is still running. These analyses are performed in conjunction with statistical stopping rules to ensure that the Type I error is kept under control. In this project, simulation (and analytical methods, where feasible) will be used to investigate the characteristics and performance of several well-known stopping rules including those suggested by O'Brien and Fleming, and Pocock. In addition, the Lan-de Mets 'alpha spending' approach and stopping rules for futility will be investigated. In addition, the use of stopping rules in (multi-arm) multi-stage phase II studies may also be explored.

Title: An investigation of the efficient use of baseline measurements of outcome in randomised trials

Supervisor: Dr Gareth Ambler and Dr Julie Barber

Suitability: MSc Medical Statistics only

Description:

In randomised trials where the continuous primary outcome is measured at both baseline and follow up, an efficient analysis would adjust the treatment effect for the baseline measurement. Usually ANCOVA would be used to make this adjustment. In this project the student will investigate an alternative use of the baseline data in estimating the treatment effect, considering how this could be incorporated as part of the dependent variable using a repeated measurements model. Through use of simulation, potential efficiency gains in such an approach will be examined, especially considering the case of missing outcome data.

Title: Exploring cost-effectiveness of the introduction of the HCM Risk-SCD risk model in clinical practice

Supervisors: Gianluca Baio and Rumana Omar

Suitability: MSc Statistics and MSc Medical Statistics

Description:

HCM is a common inherited heart muscle disorder and may affect as many as 1 in 200 people. It is a leading cause of sudden cardiac death (SCD) in young adults. Patients at high risk of SCD need to be identified so they can be offered life saving treatment with an implantable cardioverter defibrillator (ICD). The HCM Risk-SCD risk model is used to estimate a patient's 5-year risk of SCD, and patients with a predicted risk of ≥ 6% receive an ICD implantation. It has been calculated that for every 13 high-risk patients who receive an ICD using this threshold, 1 patient could potentially be saved from SCD. The HCM Risk-SCD calculator can be used to avoid unnecessary ICD implants in low risk patients. The objective of this project is to develop a cost-effectiveness model to estimate the value-for-money of the introduction of the risk model. The project will involve working with clinicians and experts to create a model to describe patients pathway with and without the intervention; a targeted literature review to obtain information on the main parameters of such model; and a cost-effectiveness analysis.

Requirements: Knowledge of R (essential); Bayesian modelling (essential); health economic evaluation (desirable).

Title: Addressing missing not-at-random data in cost-effectiveness analysis

Supervisors: Gianluca Baio and Manuel Gomes

Suitability: MSc Statistics and MSc Medical Statistics

Description:

Missing data are a common issue in cost-effectiveness analysis (CEA) and are often addressed assuming the data are “missing at random” (MAR). However, this assumption is often questionable and sensitivity analyses are required to assess the implications of departures from MAR. Reference-based imputation provides an attractive approach for conducting such sensitivity analyses, because missing data assumptions are framed in an accessible way by making reference to specific sub-groups in the sample. For example, a plausible not-at-random mechanism in a placebo-controlled trial would be to assume that participants in the experimental arm who drop out stop taking their treatment, and have similar outcomes to those in the placebo arm. Recent work has considered the reference-based imputation approach in CEA by assuming the cost-effectiveness endpoints were (jointly) normally distributed. However, this is unlikely to be plausible as the outcomes of prime interest in CEA, such as costs and patient-reported outcomes, often exhibit large departures from normality (e.g. highly skewed and multimodal). This will lead to misleading inferences because the reference-based approach typically relies on plausible distributional assumptions about the data. Building on existing Bayesian approaches for handling MNAR data, this project will consider how to extend the reference-based approach to address non-normal MNAR outcomes in CEA. The methods will be applied to the IMPROVE study that evaluates an endovascular strategy compared to open surgery for patients with a ruptured aneurysm.

Requirements: Knowledge of R (essential); Bayesian modelling (essential); health economic evaluation (desirable).

Title: Development of an interactive R/Shiny tool for NMA of trial-level data

Supervisors: Gianluca Baio and Andrea Berardi

Suitability: MSc Statistics and MSc Medical Statistics

Description:

The project will involve the development of an R/Shiny application for conduction meta-analyses of trial-level data. Disease area is to be confirmed but one option is multiple sclerosis looking at the outcomes of confirmed disability progression and annualized relapse rate. The analysis will ideally consider both Frequentist and Bayesian NMA approaches, however the Frequentist approach will serve as the starting point. The Bayesian component will require linkage to an external program such as JAGS or WinBUGS for Gibbs sampling and will therefore involve greater complexity. The aim will be to develop a flexible platform wherein the user can select trials for inclusion in the network provided the network remains connected. The tool will also allow the user to test early efficacy assumptions for a new product.

Development of the tool will be in R/RStudio with use of the Shiny package to develop the user interface.

Candidate suitability: We are looking for a hard-working, self-motivated individual with a keen interest in health economics and data analytics to join the team for their summer placement.

The ideal candidate will have experience of programming in R and a strong understanding of the principles of NMA. Experience of JAGS/WinBUGS is desirable. Technical support will be provided.

Requirements:R/RStudio (Essential), Microsoft Office (Essential), JAGS and/or WinBUGS (Preferable)

This is a full-time placement based in our London office, covering a period of 3 months.

About the placement location – PAREXEL International, Euston

The Health Economics Modelling Unit (HEMU) is a Business Unit within PAREXEL Access Consulting. HEMU is a dedicated unit with 35 staff based in the UK, US and Sweden, whose purpose is to provide health economics consultancy services to international pharmaceutical and medical devices companies. The primary activities are cost-effectiveness, budget-impact modelling and data analysis.

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: 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: 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: Time-to-event modelling with non-proportional hazards

Supervisor: Dr Aidan O'Keeffe

Suitability: MSc Statistics or MSc Medical Statistics

Description:

Where an outcome of interest is the time until a pre-specified event occurs, inference is often made through estimation of a hazard. The hazard is defined as the instantaneous probability of the event occurring at a given time and, when the aim is to compare hazards for two or more groups, an assumption of proportional hazards is often made. This assumption can be quite restrictive and may often be violated in real data. This project will explore approaches to modelling time-to-event data where hazards may be non-proportional.

Title: Investigating the use of permutation testing in randomised experiments

Supervisor: Dr. Aidan O'Keeffe

Suitability: MSc Statistics or MSc Medical Statistics

Description:

When analysing data from randomised experiments, such as a parallel group randomised controlled trial, a test statistic with a known distribution under a null hypothesis (usually a null hypothesis of no treatment effect) is formulated and used to determine whether or not the null hypothesis is rejected. This approach can rely on several assumptions which may be difficult to verify in some scenarios. An alternative to a distribution-based test is a permutation test, where the observed experimental data are re-sampled and then a sampling distribution constructed from all possible samples. Such methods rely on fewer assumptions and can provide more control over Type I and Type II error rates. This project will investigate permutation testing in detail and explore its use in randomised experiments. Some example datasets will be provided.

Title: Comparing Clinical Performance and Detecting Outliers

Supervisor: Dr Menelaos Pavlou and Dr Gareth Ambler

Suitability: MSc Statistics or MSc Medical Statistics

Description:

Routinely collected data are often used to compare hospital performance (e.g. with respect to in-hospital mortality) and to identify hospitals with unusual (poor) results. Several statistical methods have been suggested for this purpose including approaches that use funnel plots and random effects modelling. Some methods are also able to incorporate risk-adjustment (ideally using a validated risk prediction model) to take into account differences in patient case-mix. This project will review and implement some of these methods using both real surgical data and simulated data. Of particular interest is comparing the performance of these methods in different scenarios and quantifying the numbers of true/false outliers detected.

Title: Models for Human Hip and Knee Joint Movement

Suitability: MSc Statistics, MSc Medical Statistics

Supervisor: Dr Yvo Pokern

Description: In the study of human movement patterns (important for problems in medicine, ergonomics and sports science), interest is in inferring the movement of the underlying bones from the movement of so-called "skin markers" - reflective patches stuck on the skin. There is precious little data available where both the underlying bone movement and the movement of the skin markers have been recorded simultaneously. Some of this data (which is also available for this project) has been analysed using straightforward ordinary least squares (OLS) estimation which is open to substantial statistical improvement. An overview of the data and these OLS estimates is available in

“A hip joint kinematics driven model for the generation of realistic thigh soft tissue artefacts”, V. Camomilla et al. , Journal of Biomechanics 46 (2013) 625-630.

The aim of this project is to apply more sophisticated models and estimation methods, e.g. employing generalized least squares with variable selection or perhaps a principal component analysis to look for common movement of several skin markers.

Prerequisites: Good skills in R for data analysis. A willingness to work with data (already published and in the scientific literature) recorded in experiments on human corpses.

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: 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: 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.

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?