MSc in Statistics 2020 Projects

January 26, 2020

This document contains the list of projects and the abstract for each of the project. When you input your list of preferred projects please use the Filenames codes in the following table to refer to each project.

IMPROVING TRAINING DATASETS OF ASSISTIVE TECHNOLOGIES FOR MOTOR NEURON DISEASE

In partnership with the Motor Neuron Disease (MND) Association, Rolls-Royce's AI hub is leading a project to improve spoken communication for people living with MND as part of its "AI for Good" initiative.

A fatal disease affecting 5,000 people in the UK alone, MND causes rapid loss of muscle movement, thus eventually leading to an almost total loss of communication. While assistive technologies for this exist, the long delays associated with spelling out each word causes frustration, to the point where many people begin limiting their social interactions.

The main aim of this initiative is to enhance current assistive technologies by utilising NLP and deep learning techniques to reduce delays and improve suggestions, while adapting to the user's personality.

In order to achieve this aim it is necessary to have a huge set of annotated conversational data, which is hard to come by. One approach which the team is experimenting with is to synthesise such dataset from publicly available data such as YouTube. However they encountered two problems with this approach.

The first problem is separating different speakers in audio recording, which is known in the literature as the problem of "speaker diarisation".

The second problem is related to what is known as "named-entity recognition", which is a subtask of information extraction that seeks to locate and classify named entities mentioned in unstructured text.

There are also a few other related problems (all concerning the improvement of MND assistive technologies) for which the team are exploring the use of data mining techniques and which might lead to some interesting research.

I intend to supervise (jointly with a co-supervisor from Rolls-Royce's AI hub) up to two student on this project each attempting at least one of the highlighted problems above. Note that no knowledge of the medical or biological aspects of MND is needed for this research.

Supervisor Ahmad Abu-Khazneh, and External Co-Supervisor from Rolls-Royce's AI hub.

Prerequisite Python, Data Science I, Data Science II.

Modelling time dependencies in behavioural biometrics data

Supervisor: Niall Adams

Biometrics, referring to measurements and observations of the human body, is often used for identity verification in security systems. Familiar examples relate to fingerprints and iris scans. *Behavioural* biometrics, referring to *how* humans do things, is becoming popular in security systems as an alternative verification mechanism. For example, the speed and cadence of typing a password has been found effective, separate from the password itself. With most security systems there is a trade-off between security and ease-of-use, leading to vendors' desire to avoid false positives (incorrectly rejecting a valid access request).

A potential problem with behavioural biometric data is the potential for behaviour to change over time, potentially compromising its value for verification by leading to an increased rate of undesirable false positives.

Callsign Inc. is a company that produces biometric verification systems for large companies. The company offering includes behavioural biometrics related to keyboard and device usage and has amassed a large collection of data from such mechanisms. This project will investigate the different types of time-dependent behaviour present in certain biometric datasets, and has wide scope for investigating various different approaches suitable for modelling the different type of time-dependencies, that could range from data normalisation to use of machine learning models that explicitly incorporate dynamic time evolution.

Prerequisites
Exploratory data analysis, modelling, machine learning

Industrial Partner: Callsign Inc (`https://www.callsign.com`)

Note: The student may be required to sign a non-disclosure agreement to have access to these data sets

Food security and climate change: how do levels of fishing impact the global environment?

Supervisors: Niall Adams, Mark Briers

Approximately 3.5 billion people in the world rely on seafood as their primary source of protein. The impact of fishing the techniques used, the catch size, the (environmentally) perverse incentives, pollution, corruption, etc – are impacting the worlds oceans at a phenomenal rate. Coupled with climate-change induced changes in ocean temperatures, and the potential for land-derived protein sources to decrease, particularly in developing nations, it is imperative that improved fishing policies are developed in order to meet the needs of the human population, whilst restoring the oceans to being a balanced and sustainable ecosystem. In order to do so, we need a better understanding of the impact of fishing on the global environment.

In collaboration with the University of British Columbia, this project will utilise large volumes of data from the Sea Around Us (`http://www.seaaroundus.org/`) project. We will explore global fishing catch data for >2000 species from over 50 years to derive new insights into fishing patterns, causal relationships, and environmental consequences and trends, given current fishing and economic practices. Such insights could help to shift marine and fisheries policy on a global scale.

Prerequisites

Python programming, Big Data, exploratory data analysis, modelling, causal inference

Industrial Partner:

- Professor Daniel Pauly, University of British Columbia
- Dr Deng Palomares, University of British Columbia
- Professor Dirk Zeller, University of Western Australia

Evaluating recent offline changepoint detection algorithms

Supervisor: Dean Bodenham

Changepoint detection procedures attempt to determine the time(s) at which the distributional properties of a time series changes. Such procedures find application in areas as diverse as manufacturing, fraud detection, astronomy, finance, computer networks and medicine. More precisely, suppose the observations $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ are realisations of the random variables $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, where for distributions $F_1, F_2, \ldots, F_m$,

$$
\begin{aligned}
&X_1, X_2, \ldots, X_{\tau_1} \sim F_1, \\
&X_{\tau_1+1}, X_{\tau_1+2}, \ldots, X_{\tau_2} \sim F_2 \neq F_1, \\
&\vdots \\
&X_{\tau_{m-1}+1}, X_{\tau_{m-1}+2}, \ldots, X_{\tau_m} \sim F_m \neq F_{m-1}.
\end{aligned}
$$

In other words, the sequence is split into m contiguous regions, where the random variables in each region follow a particular distribution, and adjacent regions have different distributions. However, the changepoints $(\tau_1, \tau_2, \ldots, \tau_m)$ are unknown. Changepoint detection algorithms attempt to find a vector $(\widehat{\tau}_1, \widehat{\tau}_2, \ldots, \widehat{\tau}_{\widehat{m}})$ that is "as close to the truth as possible". There are more general models for changepoint detection, but for the moment we assume the data follow the model above.

There have been several recent proposals for detecting changepoints in the offline univariate setting [6, 4, 3, 7], where a retrospective analysis is performed on the whole dataset. However, it is not clear how these methods perform in general, for example when the assumptions of the methods are not satisfied. While performance assessment for online (sequential) changepoint detection has been well-studied [8, 5, 2, 1], the offline case has received less attention. Furthermore, for modern applications with a large number of observations (large n), it is imperative that these offline methods be as computationally efficient as possible, although some methods avoid discussion of computational complexity.

This project will entail:

(i) Reviewing the current state-of-the-art in offline univariate changepoint detection,

(ii) Comparing existing and perhaps proposing new performance metrics for offline changepoint detection algorithms,

(iii) Analysing a selection of recently proposed methods in a variety of simulation settings,

(iv) Applying the most robust methods to real-world data from areas such as finance or medicine.

Provided there is time, there is also the potential for a student to propose a novel changepoint detection algorithm, or an extension of an established algorithm which may address shortcomings discovered in (iii).

Prerequisite skills: Basic programming skills in R

References

[1] D. A. Bodenham and N. M. Adams. Continuous monitoring for changepoints in data streams using adaptive estimation. *Statistics and Computing*, 27(5):1257–1270, 2017.

[2] S. E. Fraker, W. H. Woodall, and S. Mousavi. Performance metrics for surveillance schemes. *Quality Engineering*, 20(4):451–464, 2008.

[3] K. Frick, A. Munk, and H. Sieling. Multiscale change point inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(3):495–580, 2014.

[4] P. Fryzlewicz. Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, 42(6):2243–2281, 2014.

[5] D. Kifer, S. Ben-David, and J. Gehrke. Detecting change in data streams. In *Proceedings of the Thirtieth International Conference on Very Large Data Bases-Volume 30*, pages 180–191. VLDB Endowment, 2004.

[6] R. Killick, P. Fearnhead, and I. A. Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012.

[7] R. Maidstone, Toby H., G. Rigaill, and P. Fearnhead. On optimal multiple changepoint algorithms for large data. *Statistics and Computing*, 27(2):519–533, 2017.

[8] E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954.

Evaluating outlier detection procedures

Supervisor: Dean Bodenham

Given a sample of observed data, an outlier can be described as "an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism" [6]. Determining outliers (or anomalies) is of practical importance in areas as diverse as cybersecurity and medicine [9]. Recent surveys of outlier methods include [8, 3, 1, 5].

For univariate data, many outlier methods assume the probability distribution underlying the generation process is known [2], although there are a few methods, such as Tukey's method used in the boxplot [10, 7], which make no distributional assumptions. For multivariate data, defining outliers is far more difficult, even in the two-dimensional setting [2]. Although there are few nonparametric outlier detection methods for the multivariate setting, one popular method from the data mining literature is DBSCAN, which identifies outliers as a byproduct of its clustering algorithm [4].

An outline of this project will be to:

(i) Review the current state-of-the-art in outlier detection (univariate and multivariate),

(ii) Review performance metrics for outlier detection procedures,

(iii) Analyse a selection of popular methods in a variety of simulation settings,

(iv) Apply the most robust methods to real-world data from areas such as finance or medicine.

Provided there is enough time, there is also scope for the student to propose and evaluate a novel outlier detection procedure.

Prerequisite skills: Basic programming skills in R

References

[1] C. C. Aggarwal. Outlier analysis. In *Data mining*, pages 237–263. Springer, 2015.

[2] V. Barnett and T. Lewis. *Outliers in statistical data*. Wiley, 1974.

[3] I. Ben-Gal. Outlier detection. In *Data mining and knowledge discovery handbook*, pages 131–146. Springer, 2005.

[4] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.

[5] M. Gupta, J. Gao, C. C. Aggarwal, and J. Han. Outlier detection for temporal data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 26(9):2250–2267, 2013.

[6] D. M. Hawkins. *Identification of outliers*, volume 11. Springer, 1980.

[7] D. C. Hoaglin, B. Iglewicz, and J. W. Tukey. Performance of some resistant rules for outlier labeling. *Journal of the American Statistical Association*, 81(396):991–999, 1986.

[8] V. Hodge and J. Austin. A survey of outlier detection methodologies. *Artificial intelligence review*, 22 (2):85–126, 2004.

[9] J. Laurikkala, M. Juhola, and E. Kentala. Informal identification of outliers in medical data. In *Fifth International Workshop on Intelligent Data Analysis in Medicine and Pharmacology IDAMAP-2000 Berlin, 22 August. Organized as a workshop of the Fourteenth European Conference on Artificial Intelligence ECAI-2000*, volume 1, pages 20–24, 2000.

[10] J. W. Tukey. *Exploratory data analysis*, volume 2. Reading, Mass., 1977.

Evaluating modern multivariate two sample testing procedures

Supervisor: Dean Bodenham

Two sample testing procedures attempt to provide a statistical score to indicate if two datasets were sampled from different distributions. Mathematically, suppose that the observations $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_m)$ are realisations of the random variables $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ and $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_m)$, respectively, where $\mathbf{X}$ follows distribution F_X and $\mathbf{Y}$ follows distribution F_Y, and it is not known whether $F_X = F_Y$ or $F_X \neq F_Y$. A two sample testing procedure could be described as a function $g : \mathbb{R}^n \times \mathbb{R}^m \to [0, 1]$, where one assumes the null hypothesis $H_0 : F_X = F_Y$ to be true, and then a score $g(x, y)$ that is close to 0 indicates that the null hypothesis could be false. Note that most two sample tests do not operate on the (raw) data, but rather on summary statistics.

Perhaps the most famous example of a two-sample test is Student's t test and its variations. However, the t test (i) assumes the random samples $\mathbf{X}$ and $\mathbf{Y}$ follows normal distributions, and (ii) is only defined for univariate random variables. Other tests, such as the nonparametric Kolmogorov-Smirnov test [6, 10] and Hotelling's multivariate generalisation [5], have been proposed which separately overcome these restrictions, but until recently there were few methods which were both nonparametric and multivariate.

Then, about fifteen years ago, a collection of similar two sample testing methods [1, 11, 4] appeared which were both nonparametric and multivariate with good performance in high dimensions. Later, these tests were shown to be closely related [9]. However, despite claims to perform well in high dimensions p, recent work [7] has cast doubt on the performance metrics used for these analyses. Furthermore, other methods using different metrics have recently been proposed [8, 2, 3].

This project will:

(i) Review the current state-of-the-art in nonparametric multivariate two-sample testing,

(ii) Review performance metrics for evaluating two-sample testing procedures,

(iii) Analyse a selection of recently proposed methods in a variety of simulation settings,

(iv) Apply the most robust methods to real-world data from areas such as finance or medicine.

Prerequisite skills: Basic programming skills in R

References

[1] L. Baringhaus and C. Franz. On a new multivariate two-sample test. *Journal of Multivariate Analysis*, 88(1):190–206, 2004.

[2] H. Chen and J. H. Friedman. A new graph-based two-sample test for multivariate and object data. *Journal of the American Statistical Association*, 112(517):397–409, 2017.

[3] H. Chen, X. Chen, and Y. Su. A weighted edge-count two-sample test for multivariate and object data. *Journal of the American Statistical Association*, 113(523):1146–1155, 2018.

[4] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

[5] H. Hotelling. The generalization of student's ratio. *Annals of Mathematical Statistics*, 2(3):360–278, 1931.

[6] A. Kolmogorov. Sulla determinazione empirica di una lgge di distribuzione. *Inst. Ital. Attuari, Giorn.*, 4:83–91, 1933.

[7] A. Ramdas, S. J. Reddi, B. Póczos, A. Singh, and L. Wasserman. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[8] A. Ramdas, N. Trillos, and M. Cuturi. On wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47, 2017.

[9] D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5):2263–2291, 2013.

[10] N. Smirnov. Table for estimating the goodness of fit of empirical distributions. *The annals of mathematical statistics*, 19(2):279–281, 1948.

[11] G. J. Székely and M. L. Rizzo. Testing for equal distributions in high dimension. *InterStat*, 5(16.10): 1249–1272, 2004.

Identifiability of Hawkes Processes

Supervisor: Ed Cohen

Hawkes processes are a class of stochastic point process that model self-exciting and mutually exciting event data. That is to say, the occurrence of an event can trigger further events of the same type and/or events of a different type. Their flexibility and real-world relevancy has resulted in a host of applications. For example, in the case of financial data, this allows propagation of stock crashes and surges to be modelled. In social media, they are used for modelling 'twitter cascades'. Further examples include the modelling of earthquake events, neuron firing times, and terrorist activity.

A vital aspect of model fitting and parameter estimation is one of identifiability. Identifiability can roughly be taken to mean that there exists is a unique maximum likelihood estimator for any arbitrary dataset. When models become unidentifiable, there are multiple parameter values that are equally likely to have given rise to the data. This can cause significant problems as the fitted model can diverge significantly from the true data generating process, rendering inference and prediction useless.

Understanding when Hawkes processes become unidentifiable has yet to be fully explored, particularly in the multivariate (mutually exciting) setting and when event observations are aggregated into bins to form a time series of counts, as is common with many datasets.

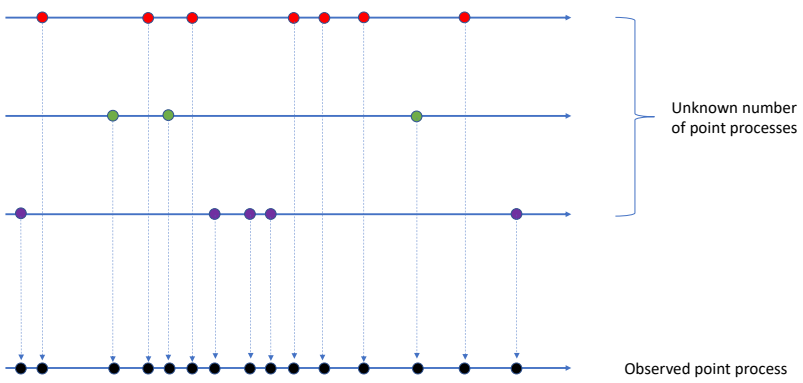
This project will investigate under what circumstances Hawkes processes become unidentifiable. In doing so you will make a vital contribution to an ongoing project in applying multivariate Hawkes processes to computer network data for cyber-security purposes.

Suitable streams: General, Theory and Methods, Data Science.

Model selection for point processes

Supervisor: Ed Cohen

Point processes (sometimes referred to as event processes) are a class of stochastic process that are used to model event data (e.g. earthquake events, crime events, arrival times of customers in a queue, the times at which neurons fire). It is quite common for the observed event data to in fact arise from the superposition of an unknown number of point processes. The immediate question then arises: given a dataset of observed event times, how many process generated it and to which process should we assign each event? Methodology for tackling this problem would have immediate impact in cyber-security where an observed sequence of events could have been generated by an several actors, both human and machine, and knowing how many is essential for characterising network behaviour and spotting abnormalities.



In this project you will explore and develop methodology for this problem. You will begin by adapting standard model selection methods (e.g. AIC and BIC) for this particular application and assessing their performance. Focus will be on Hawkes processes, a class of point process that models self-exciting behaviour (the occurrence of an event can trigger further events).

You will then look to develop this for event data that is aggregated into time series of counts. Such datasets are extremely common and handling data of this type poses many open problems. You will look to embed your model selection method alongside a MCEM (Monte-Carlo Expectation-Maximisation) parameter estimation method recently developed by this group. In doing so, you will be making a valued contribution to an ongoing project that seeks to implement point process models in a cyber-security setting.

Suitable streams: General, Theory and Methods, Data Science.

Learning point process parameters on the stream

Supervisors: Ed Cohen and Niall Adams

Point processes (sometimes referred to as event processes) are a class of stochastic process that are used to model event data (e.g. earthquake events, crime events, arrival times of customers in a queue, the times at which neurons fire). With the recent explosion in data collection and storage, hugely rich event datasets are now extremely common and hence methodologies for estimation and prediction are in high demand.

A particular challenge is to learn the parameters of a point process model on the stream. That is to say, being able to update the point process model in real time as new events are seen, particularly for point processes which exhibit significant non-stationarity. Being able to do so will allow the detection of changes in the process' structure and behaviour. Such methods could have significant impact in areas such as cyber-security where abnormal changes could indicate intruders on a network. Furthermore, real-time updating of models will provide more precise predictions for future events, something that could be of huge importance to those developing and implementing trading strategies in finance.

In this project, you will explore and develop methodology for updating parameter estimates for simple point process models in a computationally fast and efficient manor such that they can be implemented in real time. You will examine existing methods for Poisson processes and then look to extend to Hawkes processes, a class of self-exciting point process that have far reaching applications in areas such as finance, criminology, seismology, neuroscience and social media. Particular focus will be given to cyber-security, and you will make a valued contribution to an ongoing project that is implementing Hawkes processes for intruder detection.

Suitable streams: General, Theory and Methods, Data Science.

Statistical Analysis of Manifold-Valued Random Fields

Andrew Duncan: wwwf.imperial.ac.uk/~aduncan/

It is quite common that one wishes to analyse quantities which have some intrinsic non-Euclidean geometry attached to them. For example, observations of the direction of a windsock at an airport are naturally represented as points on a unit circle. Covariance matrices lies in the set of symmetric positive definite matrices which assigned various geometries [1,2]. Similarly, Principal Component Analysis (PCA) decompositions can be associated with the Stiefel manifold [3]. More complex examples arise in shape analysis and diffusion tensor imaging.

In many cases, the data under consideration is spatially distributed but it is not trivial to take into account spatial dependence in the analysis because of the non-linear geometry of the manifold [4,5]. In this project we aim to study random fields (i.e. a collection of correlated random variable indexed spatially) for which each point is a manifold-valued random variable. We shall primarily focus on the matrix-valued random fields and investigate methods for *kriging,* i.e. predicting intermediate values of the random field based on a finite set of measurements, as well as other associated problems.

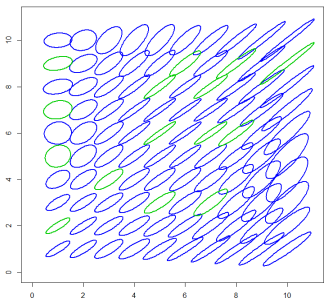


Figure 1: Predicted covariance field given the observations in green

As an application of this methodology, we could consider a problem in aerospace engineering. In the modern turbomachinery design process, blade shapes are parametrized by a set of design variables useful from aerodynamic, structural and manufacturing perspectives. Typically, the number of such parameters can range from 20 to 250. This high-dimensional design space makes parametric studies—e.g., uncertainty quantification, optimization and sensitivity analysis—challenging and computationally prohibitive, motivating the use of dimension reduction strategies. Very expensive numerical experiments using computational fluid dynamics solvers can be used to quantify the influence of the blade geometry on the engine behaviour (specifically efficiency) under various environmental conditions. In practice, engineers do not expect that all of the 250 parameters have an equal influence on the overall efficiency, and rather expect that efficiency will depend nonlinearly on a handful of these design parameters. By performing a spatially dependent PCA we aim to learn the key drivers of optimum behaviour for engine blades.

Pre-requisite modules: Bayesian Methods.
General Background: Confidence with linear algebra, and geometry will be a plus.
Stream Suitability: Theory and Methods, General.

References

[1] Yuan, Y., Zhu, H., Lin, W. and Marron, J.S. (2012) “Local polynomial regression for symmetric positive definite matrices”, J. Roy. Stat. Soc. B, 74:697–719.

[2] Shi, X., Zhu, H., Ibrahim, J.G., Liang, F., Lieberman, J. and Styner, M. (2012) “Intrinsic Regression Models for Medial Representation of Subcortical Structures”, J. Am. Stat. Assoc., 107:12–23
[3] Mahony, R. E., Helmke, U., & Moore, J. B. (1996). Gradient algorithms for principal component analysis. The ANZIAM Journal, 37(4), 430-450.
[4] Pigoli, D., Aston, J.A.D., Dryden, I.L. and Secchi, P. (2014) “Distances and Inference for Covariance Operators”, Biometrika, 101:409–422.
[5] Pigoli, D. and Secchi P. (2012) “Estimation of the mean for spatially dependent data belonging to a Riemannian manifold”, Electron. J. Stat., 6:1926–1942.

Principal Component Analysis in Very High Dimensions

Andrew Duncan: wwwf.imperial.ac.uk/~aduncan/

Principal Component Analysis (PCA) is used throughout science and engineering as a means of summarising data which is measured on many variables in terms of a smaller number of derived variables. The method originated with Pearson in 1901 and Hotelling in 1933 [1]. It's routine use in analysis boomed with the advent of electronic computers: an early classic in meteorology is the use by Lorenz to summarize air pressure data from $p = 64$ stations across the U.S.

In many settings, it's no longer uncommon that the number of variables collected is commensurate to (or even larger than) the same size of the data. In this "high-dimensional" setting, under certain assumptions on the covariance structure of the data, the statistical properties of PCA exhibit phenomena that might appear unintuitive [2].

In this project we seek to first survey the state-of-the-art on sparse PCA methods in the high dimensional setting. This will require us delving into the literature on limiting behaviour of eigenvalues of covariance operators [3], and the associated distributions and concentration phenomena and phase transitions [5]. This will require us to foray into random matrix theory and related topics. We may also investigate connections to what is known as "functional" PCA [4] in which the observation vectors have some intrinsic spatial or temporal smoothness and study the influence of these assumptions in the high dimensional regime.

There are several possible motivating examples which might be considered in this work, but one very interesting possibility will be investigating extensions the above methodology to novel decompositions of data streams with bounded variation based on rough path theory.

Pre-requisite modules: Probability for Statistics
General Background: Confidence with linear algebra will be a plus.
Stream Suitability: Theory and Methods.

References

[1] Jolliffe, I. (2011). *Principal component analysis* (pp. 1094-1096). Springer Berlin Heidelberg.
[2] Johnstone, I. M., & Paul, D. (2018). PCA in high dimensions: An orientation. *Proceedings of the IEEE*, *106*(8), 1277-1292.
[3] Johnstone IM, "On the distribution of the largest eigenvalue in principal components analysis," Annals of Statistics, vol. 29, pp. 295–327, 2001
[4] Horváth L and Kokoszka P, Inference for functional data with applications, ser Springer Series in Statistics. Springer, New York, 2012.
[5] Silverstein JW and Bai ZD, "On the empirical distribution of eigenvalues of a class of large dimensional random matrices," Journal of Multivariate Analysis, vol. 54, pp. 175–192, 1995

Hawkes Processes with Inhibition

Andrew Duncan: wwwf.imperial.ac.uk/~aduncan/

The Hawkes process is a point process model which is typically used to model 'self-exciting' processes, such as earthquakes [1,2,3], gang warfare [4] or simply likes on a twitter post [5]. The process models a sequence of event arrivals of some type over time, where each arrival excites the process in the sense that the chance of a subsequent arrival is increased for some time period after the initial arrival. Hawkes processes also possess multivariate variants where streams of events of different classes are able to excite each other in a nonlinear fashion. The process of generating, fitting and testing the goodness of fit of single and multivariate Hawkes processes is well known.

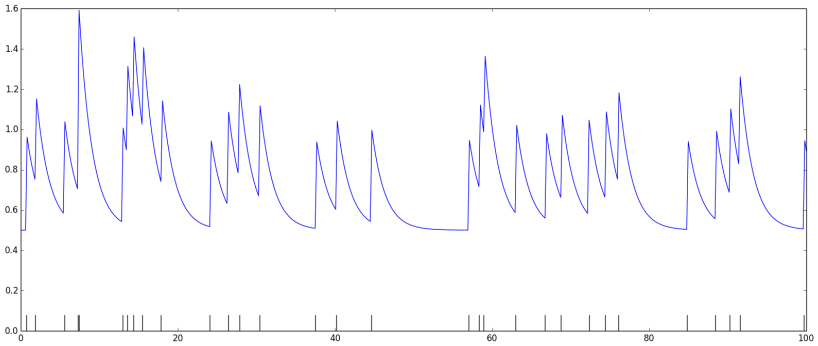


Figure 1: Realisation of a one-dimensional Hawkes Process

In this project we seek to understand better what happens when we allow not only excitation, but also inhibitory relationships in the Hawkes process. This is where the particular arrival of an event can inhibit the propensity for another type of event to occur. While this model has been considered in the statistics and stochastic process literature before, there are interesting challenges in simulation and model fitting which are swept under the carpet. Our goal will be to shed some light on these challenges and generalise the methods for fitting Hawkes processes to accommodate inhibitions in a correct manner.

There are several possible applications for this methodology which we can consider. This problem is motivating by wanting to develop a spatio-temporal model of riots and protests based on the ACLED dataset [6]. The Armed Conflict Location and Event Dataset (ACLED) collects political violence data for the world's most unstable states from 1997 onwards. Data on the date, location, groups involved and type of violence for the entire African continent have been coded and released for analysis. The types of events collected include civil war events, battles, militia attacks, and violence against civilians, riots and protests. This dataset represents the most comprehensive disaggregated collection of political violence events presently available for all states in Africa.

In previous works it has been demonstrated that riots and protests possess a self-excitatory nature. It has been postulated riots and protests which result in military intervention have a suppressive effect on future protests while those involving only police intervention are excitatory. We seek to test this conjecture through the use of an appropriately developed Hawkes process model.

Pre-requisite modules: Bayesian Methods would help.
Stream Suitability: All – scope is wide enough that it can be adapted to student's interests.

References

1. Hawkes AG, Adamopoulos L (1973) Cluster models for earthquakes—regional comparisons. Bull Int Statist Inst 45:454–461
2. Ogata Y (1988) Statistical models for earthquake occurrences and residual analysis for point processes. J Am Stat Assoc 83(401):9–27
3. Ogata Y (1998) Space-time point-process models for earthquake occurrences. Ann Inst Stat Math 50(2):379–402
4. Mohler G, Short M, Brantingham P, Schoenberg F, Tita G (2011) Self-exciting point process modeling of crime. J Am Stat Assoc 106(493):100–108
5. Kobayashi, R., & Lambiotte, R. (2016, March). Tideh: Time-dependent hawkes process for predicting retweet dynamics. In Tenth International AAAI Conference on Web and Social Media.
6. Raleigh, C., Linke, A., Hegre, H., & Karlsen, J. (2010). Introducing ACLED: an armed conflict location and event dataset: special data feature. Journal of peace research, 47(5), 651-660.

Title: Understanding the missing heritability of complex diseases
Supervisors: Dr Marina Evangelou
External Supervisor: Dr Zhana Kuncheva (C4X Discovery)

Genome-wide association studies have been successful over the years in identifying genetic variants (single nucleotide polymorphisms - SNPs) that are associated with different diseases. Unfortunately, these associations haven't been enough to explain the missing heritability of the studied diseases. In these projects we will investigate possible other sources of the unexplained heritability by looking at gene level associations, investigating SNP-SNP level interactions and gene-gene level interactions.

Project 1: We will conduct experiments for identifying the best approach for collapsing information from the SNP-level to the gene-level. We will investigate approaches that take into account the covariance structure that exists amongst the SNPs. We will look into multivariate regression models, and bootstrapping approaches for efficiently obtaining the significance of each gene.

Project 2: We will investigate different statistical models that test for SNP-SNP interactions and subsequently summarise these interactions into gene-gene level interactions. Similarly, to project 1, we will develop models that will test for the significance of each pairwise interaction term.

Prerequisite: Statistical Genetics and Bioinformatics module

Evaluation of Machine Learning approaches to investigate antiviral immune response patterns and their role in respiratory diseases.

Supervisors: Dr Sarah Filippi

External partners: Dr Sara Fontanella, Prof. Adnan Custovic (NHLI)

Objective:

Asthma is the most common chronic disease in childhood, and severe exacerbations remain one of the commonest reasons for childhood hospital admission in the developed world. Several studies showed that viral and bacterial infections are major causes of respiratory morbidity/mortality. However, the immune mechanisms governing the relationships between susceptibility to virus infection, lower respiratory tract infections, asthma pathogenesis are poorly understood.

The aim of this project is to apply different statistical learning models to investigate the association of antiviral immune responses with asthma and other clinical outcomes, and, eventually, propose the most appropriate modelling strategy for this task. The data refer to peripheral blood mononuclear cell responses (28 cytokines) to live respiratory viruses and bacteria measured in children participating in a population-based birth cohort study.

Data: 28×15 continuous variables and one or more binary response variables.

Specific Aims:

- Identify appropriate statistical machine learning solutions to evaluate the relationship between antiviral immune responses with asthma and other clinical outcomes.
- Compare results from different machine learning techniques, evaluate validity of final results and identify optimal model.

Pre-requisites modules: None

Stream suitability: Biostatistics, Applied statistics, Data Science, General.

Hierarchical Bayesian model to estimate prevalence in categories of the joint distribution of BMI and height

Supervisor: Dr Sarah Filippi

Extrernal partner: Prof Majid Ezzati (School of Public Health)

Being taller is associated with prolonged longevity as well as lower risk of cardiovascular diseases. Although height is one of the most heritable human traits, cross population differences are believed to be related to non-genetic, environmental factors. Of these, nutrition and infectious diseases during foetal period, childhood and adolescence are determinants of height during adulthood. Body Mass Index (BMI) at underweight, overweight, and obesity ranges can also lead to adverse health outcomes such as cardiovascular and kidney diseases, type 2 diabetes, some cancers, and musculoskeletal disorders. Knowledge on height and BMI can help understand the health impacts of nutrition and environment on Non Communicable Diseases.

The aim of this project is to model the joint distribution of height and BMI which are differentially affected by poor and healthy nutrition and therefore vary both geographically and in time. The focus of the project is to construct and infer a Bayesian hierarchical model of the joint distribution of height and BMI using data collated worldwide by the Non Communicable Diseases Risk factor Collaboration. The inference procedure of the model will involve the implementation of computational statistic algorithms such as Markov Chain Monte Carlo.

Data. The database of population-based health surveys is collated by the Non Communicable Diseases Risk factor Collaboration (NCD-RisC), a worldwide network of health researchers and practitioners that systematically collects data sources and report trends in NCD risk factors worldwide over time. The majority of the data sources held by the NCD-RisC group came via the collaboration with the World Health Organization (WHO) and health scientists around the world. Currently, NCD-RisC holds the largest database worldwide for NCD, accounting for over 3000 thousand surveys and almost 150 million people.

Pre-requisites modules: None

Stream suitability: Biostatistics, Applied statistics, Data Science, General.

Utilising machine learning to distinguishing macrophage populations through imaging data for tissue engineering

Supervisors: Dr Sarah Filippi

External partners: Dr. Julia Sero (University of Bath)

The use of biomaterials for tissue engineering, in which materials can be engineered to replace or repair damaged tissues, is a rapidly growing and exciting field and holds promise to revolutionise medical treatment. However, the challenge of safety and understanding how immune cells react to these materials, when introduced to the body, remains and proves to be a bottleneck to commercialisation of these novel materials.

Working with Dr. Julia Sero (University of Bath), this project aims to address better understanding of such issues. Immune cells, depending upon the cues displayed by biomaterials, can mature into different subsets which play critical roles in tissue regeneration. Macrophages, a type of white blood cell of the immune system, that engulfs and digests cellular debris and foreign substances, are first-line defenders against infection by clearing debris from the evolving wound. These cells can exist in various states, termed M1 and M2, through a process called ‘polarisation’.

The effect of various conditions on macrophage polarisation can be assessed by quantitative image analysis. Statistics can be used to characterize the cell populations by analysing complex and large datasets. The focus of this project will be to use statistical machine learning techniques to compare macrophage populations exposed to different conditions and to characterise the contributions of the various factors to cellular phenotypes and gain insight into how immune cells respond to biomaterials. The developed statistical machine learning approach could then be applied to other biomedical problems involving the distinctions between two cellular populations such as healthy vs cancerous cell tissues.

Pre-requisites modules: None

Stream suitability: Biostatistics, Applied statistics, General.

“Hawkes features”: discretised approximations to self-exciting point processes

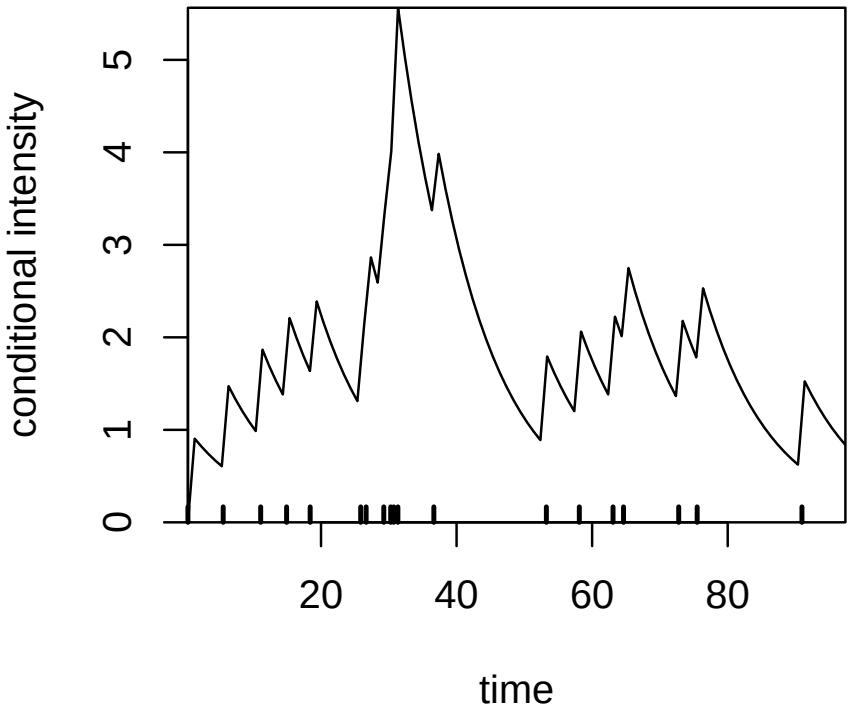
Supervisors: Seth Flaxman (Maths), Swapnil Mishra & Juliette Unwin (Medicine)

The Hawkes process is a self-exciting Poisson process, which has been used in a variety of domains from modeling earthquakes to financial data to crime (Mohler et al., 2011; Loeffler and Flaxman, 2017). While standard formulations are computationally more attractive than, for example, the log-Gaussian Cox Process, likelihood calculations are naively $\mathcal{O}(n^2)$ limiting its routine use in large data settings, especially with Bayesian inference where likelihoods must be recalculated many times. Spatiotemporal applications are also rare, with much of the literature focused on temporal data.

Given events at space-time locations $\{s_1, s_2, \ldots, s_n\}$ the Hawkes process intensity is:

$$\lambda(s) = \mu(s) + \sum_{i \,:\, s_i < s} k(s, s_i) \tag{1}$$

where $\mu(s)$ gives the underlying intensity from which events are spontaneously generated and the second term is the “self-exciting” term in which parent events trigger child events. The notation $s_i < s$ means that the event s_i occurred before the event s. $k(\cdot, \cdot)$ is a kernel function which controls how the decay in space and time of the self-exciting component of the process. A visualization is below:



In real data settings, a number of challenges arise in the use of the Hawkes process, beyond the $\mathcal{O}(n^2)$ complexity. The Hawkes framework is one of a simple point process, so there cannot be two observations with the exact same time label $t_i = t_j$. However, timestamps are in practice often rounded to the nearest hour or day. Another challenge is that of learning the underlying intensity function $\mu(s)$ from data. Inspired by “Hawkes features” (**?**Mohler and

Porter, 2018; Flaxman et al., 2019) and Diggle's computational grid (Diggle et al., 2013) we discretise to times $0, 1, \ldots, N$ days (or hours or weeks).

We calculate $\lambda(0), \lambda(1), \ldots \lambda(N)$. Now we replace $\lambda(t)$ with a piecewise constant function: $\gamma(t) = \lambda(\lfloor t \rfloor)$. Finally, we approximate the likelihood:

$$\mathcal{L} = \prod_{j=1}^{n} \lambda(t_j) \exp\left(-\int_{t=0}^{N} \lambda(t) dt\right) \tag{2}$$

$$\approx \prod_{i=0}^{N} \lambda(i)^{k_i} \exp\left(-\sum_{i=0}^{N} \lambda(i)\right) \tag{3}$$

where k_i = # of events on day i.

This formulation may yield significant advantages: for datasets with many events with the same time label, these events are aggregated into a single count yielding computational savings.

Using real datasets—either malaria cases or crime reports or both—the student will:

- investigate the various proposals for Hawkes features that have been made in the literature
- formalize the approach described above
- implement (in R or python) the existing and proposed approaches and compare them

Depending on interest, the student may focus on:

- Theoretical aspects of point processes and these approximations, including aspects of the understudied non-linear Hawkes process
- Methodological and computational advances and challenges: how should these models be used, how do they compare to the full model, how can they be implemented in frequentist or Bayesian settings?
- Data science challenges: set up a workflow to efficiently fit these models to crime data from a 10-20 cities, then perform a meta-analysis of these findings

References

P. J. Diggle, P. Moraga, B. Rowlingson, B. M. Taylor, et al. Spatial and spatio-temporal log-gaussian cox processes: extending the geostatistical paradigm. *Statistical Science*, 28(4): 542–563, 2013.

S. Flaxman, M. Chirico, P. Pereira, C. Loeffler, et al. Scalable high-resolution forecasting of sparse spatiotemporal events with kernel methods: A winning solution to the nij "real-time crime forecasting challenge". *The Annals of Applied Statistics*, 13(4):2564–2585, 2019.

S. John and J. Hensman. Large-scale cox process inference using variational fourier features. In *International Conference on Machine Learning*, pages 2367–2375, 2018.

C. Loeffler and S. Flaxman. Is gun violence contagious? a spatiotemporal test. *Journal of Quantitative Criminology*, pages 1–19, 2017.

G. Mohler and M. D. Porter. Rotational grid, pai-maximizing crime forecasts. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 11(5):227–236, 2018.

G. O. Mohler, M. B. Short, P. J. Brantingham, F. P. Schoenberg, and G. E. Tita. Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106(493), 2011.

A. Reinhart. A review of self-exciting spatio-temporal point processes and their applications. *Statistical Science*, 33(3):299–318, 2018.

Statistical machine learning methods for computer vision to accelerate low-cost microscopy

Supervisors: Seth Flaxman (Maths), Paul French (Physics)

In this project, the student will investigate deep learning and statistical machine learning methods, especially Convolution Neural Networks, to analyze imaging data for microscopy applications. The student will collaborate closely with Prof Paul French and his group in the Department of Physics. This group has a number of ongoing projects and has recently begun to investigate the use of deep learning methods (Davis et al., 2019). In addition, the student will take a data science and Design of Experiments perspective, focused on the entire pipeline from acquiring raw data to training models to interpreting results and deploying a trained model.

One possible project will be to develop statistical machine learning methods for an affordable instrument for rapid immunohistochemistry (IHC) to identify and map key proteins associated with a range of diseases, particularly head and neck cancer. In North East India there is a particular unmet need to diagnose virus-driven cancers including head and neck cancer and lymphoma. The aim is to develop low-cost, locally sustainable instruments to read out and quantify the IHC signals from a minimal panels of antibodies. Conventionally, this is realised by sequential measurements, applying and removing a few labelled antibodies at a time, using expensive instrumentation. We propose to develop a low-cost hyperspectral imaging module for the openScopes microscope platform that will enable the simultaneous readout of 5-10 different labelled antibodies in a single scan using spectral unmixing of 5-10 dyes excited with up to three excitation wavelengths. The dyes will have known excitation/emission spectra but unknown concentrations and there may be cross-talk. This will be undertaken in partnership with the IIT Guwahati and the Dr. B. Borooah Cancer Institute, Guwahati, where the openScopes histopathology instruments will be deployed.

The student should be comfortable with R or python, and prepared to learn a deep learning framework (e.g. TensorFlow or pytorch).

References

S. P. X. Davis, S. Kumar, Y. Alexandrov, A. Bhargava, G. da Silva Xavier, G. Rutter, P. Frankel, E. Sahai, S. Flaxman, P. French, and J. McGinty. Convolutional neural networks for reconstruction of undersampled optical projection tomography data applied to in vivo imaging of zebrafish. *Under review at Journal of Biophotonics*, 2019.

Efficient Bayesian nonparametric inference with variational autoencoders (VAEs)

Supervisors: Seth Flaxman (Maths), Samir Bhatt (Medicine)

Many machine learning methods consist of specifying a class of functions and then searching for the best function or set of functions in that class. Function classes can range from simple linear bases to more complex spaces such as reproducing kernel Hilbert space (RKHS). A critical ingredient in the application of these methods is a computationally efficient optimization or sampling scheme. Stochastic processes are mathematically elegant and very flexible and powerful models in theory, as they provide flexible priors over function classes and can encode a wide range of interesting assumptions. In practice, however, efficient optimization/sampling schemes are difficult to construct for stochastic processes, especially when these models are embedded in larger models and in large data and high-dimensional settings.

We have proposed a novel approach to inference in Gaussian processes and related models using a variational autoencoder (VAE) (Kingma and Welling, 2014). For popular tasks, such as spatial interpolation, our approach achieves state-of-the-art in terms of accuracy. In this project, the student will investigate other stochastic processes and Bayesian nonparametric models, e.g. Dirichlet processes, Poisson processes (especially Cox processes), Hawkes processes, and deep Gaussian processes. For some of these, extending our paradigm should be straightforward, and the student will focus on formalizing this extension and comparing to existing inference methods. For others, the extension may prove more theoretically challenging.

The student will use R and python. Knowledge of Bayesian statistics is important. Some knowledge of stochastic processes and statistical machine learning would be useful.

References

D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. *ICLR*, 2014.

Time-dependent Covariates in Complicated Event-Histories: Landmarking vs Joint Modelling

Supervisor: Professor Axel Gandy

In many areas, one is interested in the occurrence of events, for example the onset of a certain disease, or the default of a loan. In this case models from survival analysis can be used. Traditionally, only the value of covariates at baseline, i.e. at time 0 is being used in the model.

Often, updates information about individuals is becoming available over time - and one would like to include this into prediction models. There are two fundamentally different approaches available for this: landmarking (Van Houwelingen 2007) and joint modelling (Henderson, Diggle, and Dobson 2000).

The purpose of this project is to investigate and compare these two approaches.

They should also be compared to recent machine-learning based approaches such as Lee, Yoon, and Van Der Schaar (2019).

A student for this project needs to have good programming skills. A knowledge of survival analysis and machine learning would be beneficial, but is not essential.

References

Henderson, Robin, Peter Diggle, and Angela Dobson. 2000. "Joint Modelling of Longitudinal Measurements and Event Time Data." *Biostatistics* 1 (4). Oxford University Press: 465–80.

Lee, Changhee, Jinsung Yoon, and Mihaela Van Der Schaar. 2019. "Dynamic-Deephit: A Deep Learning Approach for Dynamic Survival Analysis with Competing Risks Based on Longitudinal Data." *IEEE Transactions on Biomedical Engineering*. IEEE.

Van Houwelingen, Hans C. 2007. "Dynamic Prediction by Landmarking in Event History Analysis." *Scandinavian Journal of Statistics* 34 (1): 70–85.

Model Agnostic Meta Learning for Survival Analysis

Supervisor: Professor Axel Gandy

Traditionally, the analyst/statistician is prescribing the method for learning from data, including the model choice. Classical approaches for fitting are least squares methods, method of moments and maximum likelihood methods, approaches for model choice involve information criteria (such as AIC/BIC) and cross-validation techniues. Machine learning techniques such as (deep) neural network fall under this category as well, as the analyst decides on the geometry of the network, and on how it is trained.

Meta-learning is the idea that one uses a (possibly large) number of data sets to "learn how to learn". The hope is that by learning what works in these different (but similar) data sets, one can then learn quicker on new data sets.

Recently, a model agnostic meta learning approach has been suggested in the meta learning literature (Finn, Abbeel, and Levine 2017).

The purpose of this project is to investigate how this (or other approaches) can be used for meta-learning in survival analysis.

The dominating model in survival analysis is the proportional hazards model (or Cox model). A large number of alternatives have been proposed, but none could replace the Cox model. Recently, neural network approaches for survival analysis have been proposed (Lee et al. 2018).

The project would start by assembling a set of public available survival analysis data sets. It would then use these data sets to investigate meta-learning using some of the above models.

A student for this project needs to have good programming skills. A knowledge of machine learning and survival analysis would be beneficial, but is not essential.

References

Finn, Chelsea, Pieter Abbeel, and Sergey Levine. 2017. "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks." In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, 1126–35. ICML'17. Sydney, NSW, Australia: JMLR.org.

Lee, Changhee, William R Zame, Jinsung Yoon, and Mihaela van der Schaar. 2018. "Deephit: A Deep Learning Approach to Survival Analysis with Competing Risks." In *Thirty-Second Aaai Conference on Artificial Intelligence*.

p-value Buckets - multiple testing, dependent samples

Supervisor: Professor Axel Gandy, Imperial

2020

Software packages usually report the results of statistical tests using p-values. Whilst p-values should be used cautiously, they are a very widely used part of the statistical toolkit.

Users often interpret p-values by comparing them to standard thresholds, e.g. 0.1%, 1% and 5%, which is sometimes reinforced by a star rating (***, **, *).

This project will consider statistical tests whose p-value p is not available explicitly, but can be approximated by some Monte Carlo samples, e.g. by bootstrapping, permutation approaches or by MCMC samplers. The standard implementation of such tests usually draws a fixed number of samples to approximate p. However, the probability that the exact and the approximated p-value lie on different sides of a threshold (the resampling risk) can be high, particularly for p-values close to a threshold. There are methods that uniformly bound the resampling risk, but the price to pay for this is an unbounded procedure - with no given upper bound on the maximal runtime (Gandy, 2009).

Gandy et al. (2019) introduced a method to overcome this for situations in which the Monte Carlo samples are i.i.d. The paper considers so-called p-value buckets, a finite set of given (overlapping) intervals which cover [0,1]. The suggested algorithms have finite runtime for overlapping p-value buckets and have a uniform bound on the resampling risk. The paper also proposes an extended star rating system with overlapping buckets:

Bucket	$[0, 0.1\%]$	$(0.1\%, 1\%]$	$(1\%, 5\%]$	$(5\%, 1]$
Code	***	**	*	
Bucket	$(0.05\%, 0.2\%]$	$(0.8\%, 1.2\%]$	$(4.5\%, 5.5\%]$	
Code	**~	*~	~	

For example, if the code "* " gets reported then this can be interpreted as the test being significant at the 5% level, but also potentially significant at the 1% level (but more computation effort would be needed to determine this).

This project aims to extend the ideas of p-value buckets to other situations beyond individual tests with Monte Carlo samples that are i.i.d., for example:

1. Situations where the sampling under the null is done with **dependent samples** (e.g. tests for independence in contingency tables using the MCMC sampler developed in Scott and Gandy (2019)).

2. **Multiple testing** using e.g. the FDR correction. Gandy and Hahn (2016) suggests a method for multiple tests, but it is not a finite-time procedure.

References

Gandy, A. (2009). Sequential implementation of Monte Carlo tests with uniformly bounded resampling risk. *Journal of the American Statistical Association 104*(488), 1504–1511.

Gandy, A. and G. Hahn (2016). A framework for Monte Carlo based multiple testing. *Scandinavian Journal of Statistics 43*(4), 1046–1063.

Gandy, A., G. Hahn, and D. Ding (2019). Implementing Monte Carlo Tests with P-value Buckets. *Scandinavian Journal of Statistics*. Accepted for publication. https://arxiv.org/abs/1703.09305.

Scott, J. and A. Gandy (2019). State-dependent kernel selection for conditional sampling of graphs. *Journal of Computational and Graphical Statistics*. Accepted for publication. arXiv:1809.06758 [stat.ME].

Identifying interlopers in observations of Milky Way dwarf galaxies using Gaia

Supervisors: Alex Geringer-Sameth, David van Dyk

Dwarf galaxies are small galaxies which have been found very close to (and sometimes within) the Milky Way. They are simple systems composed of populations of old stars orbiting within a clump of dark matter. As such, they are extremely important laboratories for understanding galaxy formation and the nature of dark matter. In particular, properties of a dwarf galaxy's dark matter component are related to the dispersion of the velocities of the orbiting stars. However, it is not easy to tell which stars are members of a dwarf galaxy and which are unrelated Milky Way "interlopers" that just happen to lie along the line of sight between Earth and the dwarf. Accidentally including an interloper in the dwarf sample will artificially inflate the dwarf's velocity dispersion and bias estimates of its dark matter content.

Previous analysis relied on one-dimensional line-of-sight velocity and chemical information provided by spectroscopy (observations of the frequencies of light emitted by individual stars). Differences between the distribution describing the dwarf stars and the distribution describing the interlopers allowed separation along various dimensions. We will improve upon this method by incorporating public data from the European Space Agency's Gaia spacecraft. Gaia measures not only the position of each star but also its parallax (distance), and two-dimensional proper motion (motion across the sky, rather than towards or away from us). The spectroscopic observations have been taken only for a particular, arbitrary subset of stars. First, we will "crossmatch" the spectroscopic catalog with the Gaia catalog to identify stars present in both data sets. Then we will model the full sample of stars as a mixture with missing data (as some stars are present in only one catalog or the other). While the stars associated with the dwarf will likely have simple, physics-motivated distributions, the interloper distribution will have to be empirically motivated from the data itself, perhaps non-parametrically. The end goal will be to assign membership probabilities to each star according to how likely it is to be a dwarf star or an interloper. We may well identify some misclassifications committed by previous analyses.

Prerequisites: Students should have a good knowledge of data analysis with Python (including numpy). The following modules may be helpful: Bayesian Methods, Multivariate Analysis, Data Science I.

Gamma-ray astronomy in the presence of poor background modeling

Supervisors: Alex Geringer-Sameth, David van Dyk

All analyses of astronomical sources require a model of the "background" in the vicinity of the source. The background comprises all the light generated by "uninteresting" sources, e.g., instrumental noise, unrelated astrophysical processes, and the presence of faint nearby astronomical objects. In gamma-ray astronomy, background models are generated from a recipes motivated by physics. However, due to complicated and unknown physical processes and imperfect observations, the resulting background model is only an approximation, and often a bad one. Its use can lead to false detections, wrong detection significances, and poorly estimated source properties.

Recently, Algeri [2019] proposed a new framework to compare a given background model to data, performing a hypothesis test to determine whether it is inadequate. If it is, a "correction function" is derived to improve it. The amended background model can then be used to search for the presence of an additional signal (i.e. an astrophysical object of interest) in a unified framework. The method was developed for one-dimensional data. In this project, we will adapt the algorithm to the 2D case and use it to assess a widely used model for the Milky Way gamma-ray background. We will investigate the possibility of transforming the two-dimensional sky into a one-dimensional vector through a spiral transformation. We will then apply Algeri's method to compare a background model to all-sky data from NASA's Fermi Gamma-ray Space Telescope. We may focus on particular areas of the sky where the Fermi background model is likely to break down. Visualizing the correction function will help us understand in precisely what way it does so.

Prerequisites: Students should have a good knowledge of data analysis with Python or R (Algeri's method, `LPBkg`, is available in both languages: `http://salgeri.umn.edu/my-research`).

Reference:

Sara Algeri. Detecting new signals under background mismodelling. *arXiv e-prints*, art. arXiv:1906.06615, Jun 2019.

Model-free extraction and analysis of the isotropic gamma-ray background

Supervisors: Alex Geringer-Sameth, David van Dyk

There are a great diversity of astrophysical processes that produce gamma rays (extremely high frequency light). Our pictures of the sky therefore represent messy superpositions of lots of localized and diffuse gamma-ray sources. Disentangling the observed sky map into its various components is essential for performing searches for exotic physics signals and for analyzing individual astrophysical objects. One particularly important component is the *isotropic gamma-ray background*: the cumulative light emitted by extremely distant objects, which is (statistically) the same in every direction.

We will develop new methods of inferring the properties of the isotropic background from all-sky gamma-ray maps produced using public data from NASA's Fermi Gamma-ray Space Telescope. A simple model for the distribution of gamma-ray energies will be fit independently at every location in the sky, yielding a sample. The statistics of this sample will be used to measure the isotropic component using, for example, mixture modeling or mode estimation. Samples corresponding to "pure isotropic emission" can then be mapped back onto the sky to create innovate sky maps that will hopefully reveal interesting structure. At every step, we will perform the analysis in a data-driven way, without relying on preexisting (and often inadequate) models for the structure of various emission components. Our methods will be useful for future observations where such models are lacking entirely.

Prerequisites: Students should have a good knowledge of data analysis with Python (including numpy), numerical optimization for obtaining maximum likelihood estimates, and be able to efficiently work with large data sets.

High-dimensional Genomic Data

Supervisor: Chris Hallsworth

Description

Large scale genetic variation data are now readily available, and can be used to make inferences about recent human evolution and the basis of common diseases. The framework for inference is often the *generalized linear model*. Such models are widely used across applied statistics, e.g. logistic regression for classification and Poisson models for count data. In the classical setting, n independent observations are used to determine the relationship between p covariates and the response variable, with $n >> p$. Here, estimation of GLM parameters performs well - estimators are roughly normally distributed, so that well-calibrated confidence intervals are easy to obtain. However, modern genomic methods give access to vast amounts of information on each subject: hence, p becomes large relative to n. In this regime, estimators of the regression coefficients can become biased. This project will study the resulting bias, and its onset as a function of n/p. Traditional and more modern attempts to correct this bias will be explored by simulation, with the results applied to genomic data.

Prerequisites: Good coding skills in R (or Python) are essential.

Stream suitability: General, Applied Statistics, Theory and Methods,

References

Candès, Emmanuel J., and Sur, Pragya. The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. (2018).

Sur, Pragya, and Candès, Emmanuel J. A Modern Maximum-Likelihood Theory for High-dimensional Logistic Regression. (2018).

Firth, David. Bias reduction of maximum likelihood estimates. Biometrika 80.1 (1993): 27-38.

McCullagh, Peter, and Nelder, John. Generalized linear models. Vol. 37. CRC press, (1989).

Neural Networks for Multi-Label Classification

Supervisor: Chris Hallsworth

Description

Deep neural networks are highly flexible statistical models that can be used to solve classification problems. In this project, deep neural networks will be used to analyze human genetic variation data, to infer broad-scale patterns within human population structure (3,4). This is a hierarchical multi-label classification problem.

Neural networks are trained using iterative procedures such as gradient descent. For deep neural networks, the optimization procedure is an interesting dynamical process in its own right. Saxe at al. (2) studied the dynamics of deep neural networks. They gave an explicit solutions for properties of interest in a simple hierarchical multi-label classification problem.

This project will centre on:

1. Studying the training dynamics of neural networks for multi-label classification of hierarchical data.

2. Understanding of transient illusory correlations (see (1)) that arise in training, and their rate of decay.

3. Application of deep learning methods to classification problems using human genetic variation data to study recent human evolution.

Prerequisites: Good coding skills in R (or Python) are essential.

Stream suitability: General, Applied Statistics, Theory and Methods, Biostatistics

References

(1) Saxe, Andrew M., James L. McClelland, and Surya Ganguli. A mathematical theory of semantic development in deep neural networks. Proceedings of the National Academy of Sciences 116.23 (2019): 11537-11546.

(2) Saxe, Andrew M., James L. McClellans, and Surya Ganguli. Learning hierarchical categories in deep neural networks. Proceedings of the Annual Meeting of the Cognitive Science Society. Vol. 35. No. 35. 2013.

(3) McVean, Gil. A genealogical interpretation of principal components analysis. PLoS genetics 5.10 (2009): e1000686.

(4) Patterson, Nick, Alkes L. Price, and David Reich. Population structure and eigen- analysis. PLoS genetics 2.12 (2006): e190.

Identifying Phylogenetic Covariance

Supervisor: Chris Hallsworth

Description

Unsupervised learning methods, such as principal component analysis and autoencoders, are widely used to identify hidden structure in data. Such methods typically work by finding simple representations of the variability in a dataset. This project will study how unsupervised methods perform for phylogenetic data, where the correlation takes a particularly structured form. In phylogenetics, the patterns of mutations observed in the DNA of different species are used to infer how the species are related in the evolutionary past. Once evolutionary relationships are understood, *covariance analysis* of mutation patterns can be used to understand commonalities of structure between proteins and predict their function.

The project's broad aim is to understand what signals can be recovered from phylogenetic data by PCA or an autoencoder, and learn how to distinguish patterns of covariance due to phylogenetic effects from those that suggest common biological function. This will initially involve detailed simulation, before moving on to the analysis of genetic sequence data from organisms of different species.

Prerequisites: Good coding skills in R (or Python) are essential.

Stream suitability: General, Applied Statistics, Theory and Methods, Biostatistics

References

Colwell, Lucy J., et al. Feynman-Hellmann theorem and signal identification from sample covariance matrices. Physical Review X 4.3 (2014): 031032.

Qin, Chongli, and Lucy J. Colwell. Power law tails in phylogenetic systems. Proceedings of the National Academy of Sciences (2018): 201711913.

Goodfellow, Ian, et al. Deep learning. Vol. 1. Cambridge: MIT press, 2016. Chapters 13 and 14.

Aggregating scores from correlated cyber-security analytics

Nick Heard

http://wwwf.imperial.ac.uk/~naheard

Statistical anomaly detection of cyber-security threats relies on the aggregation of evidence obtained from a range of different data science analytics, or alternatively from the same analytic calculated sequentially over time, since the value from one single analytic will not typically have sufficient discriminative power. Each analytic provides a score of surprise, which can be converted into a p-value using the distribution function for that score. Hence there is much interest in cyber-security communities on combining p-values into a single measure of surprise.

Let $p_1, p_2, \ldots p_n$ be $n \geq 1$ p-values obtained from continuous test statistics, such that under the null hypothesis that the models underlying the analytics are correct, each $p_i \sim \text{Uniform}(0,1)$. If the tests can be reasonably assumed to be independent, then these n p-values can be very simply combined in many different ways (Heard and Rubin-Delanchy, 2018) to yield an overall score of significance. For example, the method of Fisher (1929) uses the relationship $-2\sum_{i=1}^{n} \log p_i \sim \chi^2_{2n}$.

However, assumptions of independence between different test statistics, or for the same statistic calculated repeatedly over time, are often difficult to justify and can lead to erroneous conclusions. This project is concerned with learning dependency structures between analytics and proposing suitably adjusted p-value combination methods for cyber problems. Within this framework, there will be challenges in Bayesian inference and low dimensional estimation of potentially large covariance matrices.

Prerequisite courses None.

Collaborators Microsoft – Windows Defender Advanced Threat Protection.

References

R A Fisher. *Statistical methods for research workers*. Oliver & Boyd, Edinburgh, 1929.

N A Heard and P Rubin-Delanchy. Choosing between methods of combining p-values. *Biometrika*, 105(1):239–246, 2018.

Combining discrete p-values for cyber-security analytics

Nick Heard

https://www.imperial.ac.uk/people/n.heard

Statistical anomaly detection of cyber-security threats relies on the aggregation of evidence obtained from a range of different data science analytics, or alternatively from the same analytic calculated sequentially over time, since the value from one single analytic will not typically have sufficient discriminative power. Each analytic provides a score of surprise, which can be converted into a p-value using the distribution function for that score. Hence there is much interest in cyber-security communities on combining p-values into a single measure of surprise.

In cyber security applications, most of the data which can be collected are inherently discrete in nature; for example, the IP addresses participating in a network connection, the number of bytes transferred, the protocol used, and so on. Analytics derived from these quantities will necessarily be discrete, and therefore *stochastically larger* than Uniform$(0, 1)$ random variables. Hence standard methods for combining p-values (Heard and Rubin-Delanchy, 2018) cannot be directly applied to discrete p-values without some sort of correction (Rubin-Delanchy et al., 2019).

In this project, discrete p-values will be treated as *interval censored* observations of underlying (but unobserved) continuous values. This uncertainty in individual p-values can be propagated through Monte Carlo simulations, and various techniques will be considered for eventually consolidating this uncertainty into a decision. The project will be motivated by real computer network data.

Prerequisite courses None.

Collaborators Microsoft – Windows Defender Advanced Threat Protection.

References

N A Heard and P Rubin-Delanchy. Choosing between methods of combining p-values. *Biometrika*, 105(1):239–246, 2018.

P Rubin-Delanchy, N A Heard, and D J Lawson. Meta-analysis of mid-p-values: Some new results based on the convex order. *Journal of the American Statistical Association*, 114 (527):1105–1112, 2019.

Identifying malware families and suspicious internet domains

Nick Heard

http://wwwf.imperial.ac.uk/~naheard

Most *zero-day* malware is not truly new; rather, it is derivative from one or more existing pieces of malware, adapted in some way (such as a minor source code modification, or a different compression) in order to bypass anti-virus detectors looking for known malware *signatures*. For this reason, the task of identifying whether new software might be malicious can be aided by identifying a potential malware *family* from which the executable may have arisen. Given the uncertainties surrounding these assessments, learning malware *phylogenies* (family trees) can be viewed as a task in statistical inference (Bolton and Heard, 2018).

This project will look at various aspects of new data gathered on a large number of malware samples by Palo Alto Networks (Amit et al., 2018). The data give different pieces of information on up to $\sim$4.6 million pieces of software which have been classified as benign, malicious or potentially malicious.

Information sources provided include the internet domains which each malware sample connects to, along with some results from scanning those domains to see which services they run; and histograms of 4-grams observed in the byte-code for each of the $\sim$4.6 million files.

Some questions: Are there good metrics for summarising or comparing byte code frequency counts between files to identify malware or malware families? Is there interesting network structure in the malwares (through shared domains they connect to) or in their internet domains (through having common malware executables connecting to them)?

Prerequisite courses *M5MS15 Big Data* could be useful, but not essential.

Data collaborators Palo Alto Networks – Data Science.

References

I Amit, J Matherly, W Hewlett, Z Xu, Y Meshi, and Y Weinberger. Machine learning in cyber-security - problems, challenges and data sets, 2018.

A D Bolton and N A Heard. Malware family discovery using reversible jump mcmc sampling of regimes. *Journal of the American Statistical Association*, 113(524):1490–1502, 2018.

Statistical inference for the Limit Order Book

Nikolas Kantas

This project is aimed to investigate systematic tools for scalable and distributed statistical inference for Limit Order Book models. The idea is to perform inference for systems with interacting agents that can exhibit strategic behavior. We will study the effect of non-stationary and partially observable data on optimal trading models with limit order books under different assumptions. We will compare fully and partially observed data, (dark pools) using real data from the LOBSTER dataset, as well as the ABIDES platform and perform numerical investigations for to assess estimation accuracy and prediction error. If time permits we can look into extensions that involve distributed reinforcement for each agent strategies.

- Courses:
  - Introduction to Statistical Finance , Advanced Statistical Finance (essential)
  - Time series (desired)
- Very good computing skills will be advantageous. The project will involve involve numerical work on case studies using standard programming interfaces, such as R, Julia, Python etc.

External Partners: This project is based on wider collaborations with J.P. Morgan A.I. group.

References:

1. José Penalva, Sebastian Jaimungal, and Álvaro Cartea (2015) Algorithmic and High-Frequency Trading
2. David Byrd, Maria Hybinette, Tucker Hybinette Balch (2019) ABIDES: Towards High-Fidelity Market Simulation for AI Research, arXiv:1904.12066

A hierarchical Bayesian model for evolutionary adaptation and selection acting on bacterial genomes

Background

The ratio of non-synonymous to synonymous DNA transition frequencies (dN/dS or ω) is of fundamental interest in evolutionary biology. Non-synonymous transitions change protein primary structure, whereas synonymous transitions do not. Variations in this ratio can be used to determine highly conserved biological function (where ω is low), and find regions of rapid evolution (where ω is high). The latter are of particular interest in pathogenic bacteria, as these can be markers of antibiotic resistance, or can identify proteins interacting with the host immune system during disease.

Last year, collections of whole-genome datasets with over 30,000 samples have become available for bacterial pathogens (Gladstone et al., 2019). Though this amount of data could greatly increase the power of this inference, current methods for estimating ω rely on multiple sequence alignment and inference of population ancestry (Kosakovsky Pond & Frost, 2005), which do not scale to these population sizes or to the whole-genome level.

In the age of bacterial big data, methods which use counts of codon changes instead of ancestral state reconstruction are promising alternatives (Wilson & McVean, 2006). Counting state changes in a population is rapid, and the following statistical inference is then independent of sample size.

Project

The purpose of this project will be to reimplement a new method, Genomegamap (Wilson & The CRyPTIC Consortium, 2019), using standard MCMC inference tools, extend the model, and compare its performance on two large datasets. Depending on your interests, you will have the option towards the end of the project to either focus your efforts on improvements to the statistical model, work on its application to real data, or develop supervised machine learning tools to relate ω to other features of sequence evolution.

You will work with the Bacterial Evolutionary Epidemiology Group, who routinely develop evolutionary models and bioinformatic approaches to apply them to large genomic datasets. You will have access to computational resources specifically designed for large scale sequence analysis.

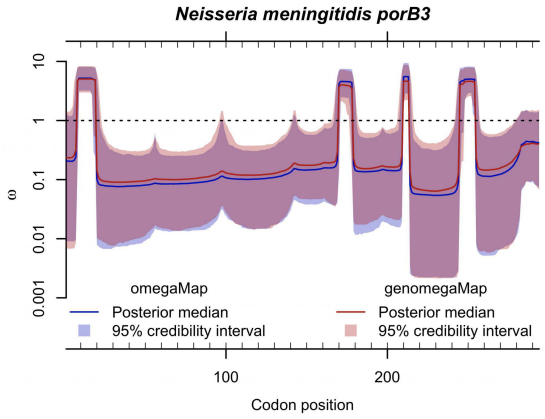


Figure 1: Application of two codon based methods to determine ω in an outer-membrane protein. Figure from Wilson and The CRyPTIC Consortium (2019).

Aims

1. Re-implement an MCMC sampler for the genomegamap likelihood, which has been previously described. We will aim to use PyMC3 (Salvatier, Wiecki, & Fonnesbeck, 2016) or STAN (Carpenter et al., 2017) to program the model. We are particularly interested in advances which may be possible using the automatic differentiation features available in deep learning libraries (Baydin, Pearlmutter, Radul, & Siskind, 2018).
2. Extend the model to make it fully hierarchical, sharing hyperparameters across all proteins in the genome.
3. Implement model selection between parameters for each sliding window versus independently for each codon.
4. Apply the model to a dataset of 10,000 *Mycobacterium tuberculosis* genomes, verifying that results can be replicated, and looking at the effect of hyperprior choice versus the fixed parameters used previously.

This will form the foundation of the project. Depending on your interests, you can then choose to work on one or more of the following problems:

- Generate input data from 35,000 *Streptococcus pneumoniae* genomes, and calculate ω for every codon. This species undergoes frequent recombination, unlike the previous example, and the power gains of the model are expected to be even higher.
- Compare some of your ω estimates from this Bayesian approach to existing maximum-likelihood packages which reconstruct ancestral state, where tractable. Consistency of ω and computational resource use are of particular interest.
- Using supervised machine learning methods such as the PyTorch library (Chen, Cofer, Zhou, & Troyanskaya, 2019), relate changes in your genome-wide ω estimates to other sequence features, such as phenotype associations or biological annotations.
- Use efficient approximate Bayesian computation methods fit evolutionary parameters (Gutmann & Corander, 2016). Your ω values will be compared against those in simulations with varying levels of selection.

Supervisors: John Lees (primary, Department of Infectious Disease Epidemiology),
Oliver Ratmann (secondary), Nicholas Croucher (Department of Infectious Disease Epidemiology).

Stream suitability: General, Applied Statistics, Biostatistics, Data Science.

Related MSc courses: No specific options are required.

References

Baydin, A. G., Pearlmutter, B. A., Radul, A. A., & Siskind, J. M. (2018). Automatic differentiation in machine learning: A survey. *J. Mach. Learn. Res. 18*(153), 1–43.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, Articles*, *76*(1), 1–32.

Chen, K. M., Cofer, E. M., Zhou, J., & Troyanskaya, O. G. (2019). Selene: A PyTorch-based deep learning library for sequence data. *Nat. Methods*, *16*(4), 315–318.

Gladstone, R. A., Lo, S. W., Lees, J. A., Croucher, N. J., van Tonder, A. J., Corander, J., ... Global Pneumococcal Sequencing Consortium. (2019). International genomic definition of pneumococcal lineages, to contextualise disease, antibiotic resistance and vaccine impact. *EBioMedicine*, *43*, 338–346.

Gutmann, M. U., & Corander, J. (2016). Bayesian optimization for likelihood-free inference of simulator-based statistical models. *J. Mach. Learn. Res. 17*(1), 4256–4302.

Kosakovsky Pond, S. L., & Frost, S. D. W. (2005). Not so different after all: A comparison of methods for detecting amino acid sites under selection. *Mol. Biol. Evol. 22*(5), 1208–1222.

Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in python using PyMC3. *PeerJ Comput. Sci. 2*, e55.

Wilson, D. J., & McVean, G. (2006). Estimating diversifying selection and functional constraint in the presence of recombination. *Genetics*, *172*(3), 1411–1425.

Wilson, D. J., & The CRyPTIC Consortium. (2019). *GenomegaMap: Within-species genome-wide dN/dS estimation from over 10,000 genomes*.

Online controlled experiments: theory and practice

Emma J. McCoy

The project focuses on the statistics behind the design of online controlled experiments (OCE). OCEs are essentially randomised controlled trials adapted for use in the Internet setting. They are a popular method used by organisations to measure the impact of new products and guide strategic decision making. Often practitioners are interested in the average treatment effect (ATE) of a variant, be it the colour of a button or a recommendation algorithm; and there is a drive to make OCEs return unbiased results more quickly.

You will have the chance to work with the wealth of data from a leading fashion e-tail company, including a clickstream with hundreds of millions of new records per day. The project will also provide the opportunity to develop skills in Python programming (and packages like pandas, numpy, and sklearn), Apache Spark, Databricks, and Azure cloud computing, skills that are sought after in the data science industry.

In the project you will begin by familiarising yourself with the basics of controlled experiments and (null hypothesis) statistical testing [1], and apply the theory to design and extract insights from experiments that can be run in an e-tail setting. You will then explore deeper in one of more of the following topics depending on your interest and progress:

Heterogeneous treatment effect detection

Traditional OCEs usually only consider the ATE for the whole treatment group, yet for organisations with a diverse user base will benefit from the insight on how different sub-groups react to the treatment. We are interested in how existing methods (e.g. [2, 3, 4]) allow us to do so without running afoul of the multiple testing problem, and which is the best method to use for the given setting.

Variance reduction

Kohavi et al. [5] note that ordinarily, successful experiments in technology companies only improve metrics by a fraction of a percent and Xie and Aurisset [6] describe the need to detect small effects as huge customer bases often translate them into substantial gains in revenue and profit. These observations motivate research into increasing the sensitivity of online tests by reducing measurement noise / variance. This in turn decreases the number of samples and/or the duration required to run an experiment and hence to obtain insights.

In recent years, methods often rely on decomposing the variance in a metric and attempting to eliminate unnecessary components using various statistical and machine learning techniques. CUPED is a control variate method that measures a metric that is modified by a linear function of the covariates [7]. The idea is to eliminate the effect of correlated covariates on the the metric under measurement. A similar idea is used by Poyarkov et al. [8] who subtract the predicted value of a metric using boosted decision trees as predictors. In the same spirit, stratified sampling is used to eliminate variance between strata. Stratified sampling can be applied pre or post experiment and while the theoretical bounds are better for pre-experiment stratification, post-stratification is often preferable for large scale online A/B tests [6].

We are interested in if the methods proposed in the existing literature is applicable to e-tail experiments and yield tangible gain by shortened time-to-insight, and if there are any further opportunities to further reduce the variance of a metric.

Bayesian testing

The Bayesian perspective is the new, cool kid in the block for evaluating OCEs [9]. It allows one to attach a probability to whether a variant is winning by e.g. calculating the Bayes factor [10]. To begin with, we are interested in if Bayesian testing offers similar test "power" as compared to frequentist approaches, and if the computation cost is practical.

Anecdotal experience from experimenters indicates that it is difficult to specify a genuine prior, i.e. a particular prior that is free from the experimenters' biases and resolves to a high confidence when there is indeed an (or lack of) impact. Deng [11] proposes leveraging past experiment data to learn an objective prior using an empirical Bayes approach. We are interested in how many past experiments we need in order to obtain a useful learnt prior, and what can be done in the case where we do not have sufficient number of past experiments.

Moreover, Johnson and Rossell [12] observe that experimenters often "include" the null hypothesis in a Bayesian A/B test's alternate hypothesis, by assigning a non-negligible probability on null hypothesis parameters in their alternate prior. Such tests favour the alternate hypothesis, and converges slowly when the null hypothesis is indeed the favoured case. They propose the use of non-local priors, which can be described as drilling hole(s) on the probability density near the null hypothesis parameter(s), when constructing the alternate prior. We are interested in the reduction in number of samples required using such method, and more importantly, whether we can further improve the prior construction by also incorporating past data.

Miscellaneous Information

Prerequisite skills There are no hard prerequisites, though previous exposure to programming in a compiled (e.g. C/C++/Java) or interpreted language (e.g. Python/R) will help.

Useful courses You may find the courses Data Science I: Data (M5MS17), Data Science II: Science (M5MS18), Big Data (M5MS15), and Machine Learning (M5MS10) useful for context.

External Partners This project will also be supervised by Bryan Liu, PhD Student at StatML CDT (Imperial/Oxford) and Machine Learning Scientist at ASOS.com.

References

[1] R. Kohavi, R. Longbotham, D. Sommerfield, and R. M. Henne, "Controlled experiments on the web: survey and practical guide," *Data Mining and Knowledge Discovery*, vol. 18, pp. 140–181, Feb 2009.

[2] "Bonferroni correction." https://en.wikipedia.org/wiki/Bonferroni_correction.

[3] W. Duivesteijn, T. Farzami, T. Putman, E. Peer, H. J. P. Weerts, J. N. Adegeest, G. Foks, and M. Pechenizkiy, "Have it both ways—from A/B testing to A&B testing with exceptional model mining," in *Machine Learning and Knowledge Discovery in Databases* (Y. Altun, K. Das, T. Mielikäinen, D. Malerba, J. Stefanowski, J. Read, M. Žitnik, M. Ceci, and S. Džeroski, eds.), (Cham), pp. 114–126, Springer International Publishing, 2017.

[4] Y. Xie, N. Chen, and X. Shi, "False discovery rate controlled heterogeneous treatment effect detection for online controlled experiments," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, (New York, NY, USA), pp. 876–885, ACM, 2018.

[5] R. Kohavi, A. Deng, R. Longbotham, and Y. Xu, "Seven rules of thumb for web site experimenters," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, (New York, NY, USA), pp. 1857–1866, ACM, 2014.

[6] H. Xie and J. Aurisset, "Improving the sensitivity of online controlled experiments: Case studies at netflix," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, (New York, NY, USA), pp. 645–654, ACM, 2016.

[7] A. Deng, Y. Xu, R. Kohavi, and T. Walker, "Improving the sensitivity of online controlled experiments by utilizing pre-experiment data," in *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, (New York, NY, USA), pp. 123–132, ACM, 2013.

[8] A. Poyarkov, A. Drutsa, A. Khalyavin, G. Gusev, and P. Serdyukov, "Boosted decision tree regression adjustment for variance reduction in online controlled experiments," in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, (New York, NY, USA), pp. 235–244, ACM, 2016.

[9] A. Deng, J. Lu, and S. Chen, "Continuous monitoring of A/B tests without pain: Optional stopping in bayesian testing," in *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 243–252, IEEE, 2016.

[10] R. E. Kass and A. E. Raftery, "Bayes factors," *Journal of the american statistical association*, vol. 90, no. 430, pp. 773–795, 1995.

[11] A. Deng, "Objective bayesian two sample hypothesis testing for online controlled experiments," in *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, (New York, NY, USA), pp. 923–928, ACM, 2015.

[12] V. E. Johnson and D. Rossell, "On the use of non-local prior densities in bayesian hypothesis tests," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 72, no. 2, pp. 143–170, 2010.

Bayesian hypothesis testing

Supervisor: Daniel Mortlock (Mathematics & Physics)

Summary
Bayesian inference provides a self-consistent method of model comparison, provided that i) there are at least two models under consideration and ii) all the models have fully-specified and proper parameter priors. Unfortunately, this is not always the case in real world situations. A particularly important example is when there is data which is (apparently) at odds with an accepted/dominant theory for which there is no good alternative.

(One famous examples is the precession of Mercury's orbit around the Sun: this is at odds with Newtonian mechanics, and was eventually explained by Einstein's General Relativity; but at the time there was no good alternative theory. How reasonable would it have been to have rejected Newtonian mechanics without an alternative?)

A hypothesis test is needed in such situations, but the conventional approach to this is to use a p-value, a quantity which cannot be related to the desired posterior probability that the null hypothesis is correct. In particular, p-values are almost always too "punitive" (e.g., Berger, J. O. & Delmpady, M., 1987, *Statistical Science*, vol. 2, p. 317).

The aim of this project is to make progress towards a Bayesian hypothesis test by: i) obtaining lower bounds on the posterior value of the null hypothesis under extremal assumptions; ii) demanding any proposed formalism satisfy Bayesian model comparison results in limiting/asymptotic circumstances; and iii) applying techniques based on training sets to obtain plausible ranges of posterior probabilities. Extensive use will be made of simulations, although analytical results can be obtained for a number of simple illustrative problems that will form a key part of the arguments made here. There will be the opportunity to apply these ideas to any subject of interest to the student, but this is primarily a project about the fundamentals of statistical inference.

Prerequisites
M5MS06 Bayesian Methods

Stream suitability
Theory and Methods; General; Applied Statistics

External partners
None

The first quasars and super-massive black holes

Supervisor: Daniel Mortlock (Mathematics & Physics)

Summary
Quasars - the glowing disks of gas around the enormous black holes at the centres of some galaxies - are some of the brightest known astronomical objects and can be seen almost to the edge of the observable Universe. One of the most important unanswered questions in astrophysics at present is how the super-massive black holes at the centre of the earliest (i.e., most distant) quasars formed so rapidly; answering this question requires knowledge of the quasar luminosity function (i.e., the relative numbers of quasars of different luminosities).

Astronomical surveys to find quasars have complicated statistical selection properties and, especially at the greatest distances, the numbers of known quasars are very low. Nonetheless, these small samples can be used to measure the quasar luminosity function, although Bayesian methods are required for a full analysis as the data-sets are inhomogeneous and the posterior distributions in the model parameters are highly degenerate. This project involves Monte Carlo simulation of the measurement process, analysis of both simulated and real samples of quasars, interpretation of the outputs and comparison against the previously published results obtained using other, less principled methods.

As such, this project will be both statistically challenging and be an important contribution to this field of astronomy.

Prerequisites
M5MS06 Bayesian Methods

Stream suitability
Applied Statistics; General; Theory and Methods

External partners
None

The expansion rate of the Universe

Supervisor: Daniel Mortlock (Mathematics & Physics)

Summary
The most important unanswered question in modern cosmology is how fast the Universe is expanding. The expansion rate is characterised by the Hubble constant, H_0, which is about 70 km/s/Mpc. Different measurements give values for H_0 that contradict each other: one explanation is new physics (*a la* dark matter or dark energy); but before jumping to such dramatic conclusions it is necessary to rule out systematic and modelling errors in the various data-sets.

The most direct measurements of H_0 come from the local "distance ladder", in which a key "rung" is the Large Magellanic Cloud (LMC), a dwarf galaxy orbiting the Milky Way. The claimed precision of the local H_0 measurement relies on being able to know the distances to various different types of stars in the LMC accurately; but the LMC has a complicated internal structure which must be taken into account and included in the overall error budget. This project will be based around developing a Bayesian hierarchical model (BHM) to describe the relevant stellar populations in the LMC and then sampling the resultant posterior distributions to explore how robust the local measurements of H_0 are to assumptions about the structure of this particular dwarf galaxy.

Prerequisites
M5MS06 Bayesian Methods

Stream suitability
Applied Statistics; General; Theory and Methods

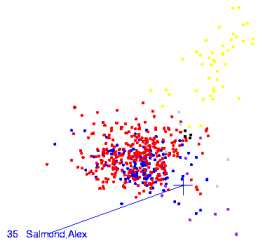
External partners
None

Web-scraped time-changing Belief Maps for British Members of Parliament

Guy Nason

It can be quite tricky to discern what individual UK parliamentary politicians think about a range of issues. Most parliamentary votes are whipped, which means that the parliamentary parties instruct Members of Parliament (MPs) how to vote on specific issues (and woe betide Members that disobey). You would have thought that one might discover the views of individual MPs during the build-up to elections but, again, the party machines exercise control over most policies, apart from those that only have local impact. One way of discovering MPs' opinions is to ask them, but they do not always tell the truth or answer the question. There is, however, a source of data that covers most politicians and reflects their genuinely held beliefs on a variety of issues. That source is the Early Day Motion database that can be accessed at the Early Day Motion website[1] and described as "Early Day Motions (EDMs) are formal motions submitted for debate in the House of Commons. However, very few are actually debated. EDMs allow MPs to draw attention to an event or cause. MPs register their support by signing individual motions." The power of Early Day Motions is that they are not whipped and are reflect the true thoughts and beliefs of MPs. They are also spontaneous and each such Motion gives a snapshot of who MPs communicate with, particularly across parliamentary boundaries. A single Motion does not, on its own, provide a lot of information. However, there can be thousands of Motions per year and it is the combined information contained within these Motions that enables us to build up a comprehensive picture of MPs' opinion.

The EDM data has been analysed a few times. One approach, that I adopted in the mid-2000s [2, 3] used multidimensional scaling. For this, we needed to calculate the distance from every MP to every other MP. This can be done easily from the lists of the Motions, who signs them and a concept known as Jaccard's distance. Some MPs sign no motions, usually Government Ministers, Whips and Parliamentary Private Secretaries, but some regular MPs too. These MPs are discarded from our analyses (However, in time local analyses, where the status of MPs changes they can sometimes be included, for example, before an MP becomes a Minister). We can present an approximate solution in two dimensions which gives a map of all MPs in space, see figure. The current project entails scraping EDM data from the website, analysing it via a variety of methods and presenting the results. A starting point might be to repeat earlier analyses on recent EDM data, but a good project would also investigate a wider range of analysis methodologies, such as those in social networking or community detection and definitely utilize modern methods of visualisation to present the results in an attractive and informative manner.



Skills required: the project will be highly computational, requiring data wrangling skills, a strong working knowledge of R and a willingness to embrace web-scraping software; experience of Unix-based data processing utilities such as `python, awk, grep, sed` will be advantageous. An aim of this project will be to explore the space of statistical and machine learning methods and decide that are useful for this data, so a good all-round knowledge of statistical science will be advantageous.

References

[1] The EDM website: `edm.parliament.uk`; [2] My earlier EDM analysis website `http://wwwf.imperial.ac.uk/~gnason/edmshow.html`; [3] Bailey, D. & Nason, G.P. (2008) Cohesion of Major Political Parties. *British Politics*, **3**, 390-417.

https://doi.org/10.1057/bp.2008.10

Deseasonalisation of Time Series using Wavelet Packets

Guy Nason

Modelling of seasonality and deseasonalisation is a topic of major practical importance carried out in many areas where time series are studied. Seasonal behaviour in a time series is the repetition of events, or cycles, that are typically aligned with key calendar attributes. For example, in time series of electricity load one might observe daily, weekly, monthly, quarterly and annual seasonality as well as week-day/weekend patterns. Seasonal patterns often appear in a wide range of official statistics such as the numbers of people in employment, tax receipts, inflation and many others. Often, the aim is to estimate and remove the seasonal pattern. The estimate can then be used to help forecast future values of the series and/or characterise the seasonal pattern to communicate to users. The residuals, obtained after the seasonal pattern has been removed, can then be further studied for signs of other (residual) seasonal pattern, trend or other stochastic structure. Current methods to model seasonal patterns rely on seasonal ARIMA models or spectral analysis. An interesting idea, due to Rebecca Killick in 2017, is to use wavelets as an alternative to Fourier (spectral) methods, to model seasonality. Wavelets partition the time-frequency plane differently to Fourier. Wavelets can focus more data onto lower frequencies that are typically of more interest to seasonal modellers and waste considerably fewer resources on higher frequencies that are not of interest.

This project will review the Killick methodology and will investigate the possibility of using wavelet packets as a wavelet replacement. Wavelet packets are libraries of oscillatory functions, which contain wavelets as a special case. Packet libraries have the ability to flexibly partition the time-frequency plane, including a tiling similar to that achieved by Fourier functions. We will also investigate if the computational frequency peak detector can be replaced, or at least augmented, by a more mathematical approach. Other opportunities include proving theoretical results about the consistency of peak frequency estimation in the wavelet and packet setup and/or considering other basis libraries (such as local cosine).

External Partner: The project will be run jointly with Duncan Elliott, Principal Methodologist, Head of the Time Series Analysis Division at the Office for National Statistics in Newport. As such this project could involve occasional trips to Newport, meeting in London and, possibly, Skype meetings. In this way, the student could gain experience of a lively operational statistical environment that is engaged in solving important real-life issues, that feed into the national debate by providing important information on the UK's economy, society and population.

Skills required: The project will suit a student who has a strong grasp of time series and spectral analysis methods and broad computational, theoretical and applied skills. The project will be involve substantial and careful R computation. Knowledge of R package construction would be useful, but not essential. The ECTS Module on *Time Series* is a prerequisite and *Non-parametric smoothing and wavelets* (the wavelets bit) will be extremely useful.

References

[1] Nason, G.P. (2008) *Wavelet Methods in Statistics with R*. Springer, New York.

[2] Percival, D.B. and Walden, A.T. (2000) *Wavelet Methods for Time Series Analysis*. Cambridge University Press, Cambridge.

Estimation and characterisation of greenhouse gas emissions from UK cities using satellite remote sensing

Guy Nason

Reducing emissions of greenhouse gases is critically important given climate change. To achieve this, we need to monitor and quantify greenhouse gas emissions accurately, so that mitigation resources are used wisely. Fortunately, given access to measurements from a growing set of deployed greenhouse gas measuring technologies, over different spatial and temporal time-scales and at different resolutions, accurate quantification is becoming feasible. Satellite remote sensing provides one source of measurements of mulitple greenhouse gases. For example, the TROPOspheric Monitoring Instrument (TROPOMI) is the satellite instrument on board the Copernicus Sentinel-5 Precursor satellite. TROPOMI [1] provides, amongst other things, measurements of methane concentrations which can be used to quantify large sources of greenhouse gas emissions such as UK cities. Measurements of column concentrations of different gases in space & time from TROPOMI are freely available [2].

The objective of the planned MSc project is to seek to address one of more of the following challenges using data from TROPOMI: By characterising emission plumes in space and time, can we quantify emissions of different greenhouse gases per annum from different UK cities. Which cities are the worst polluters? How do city emissions vary in time, for different time-scales? Is it possible to estimate a simple statistical description for the distribution of spatial plume profile for a city? What is the exposure of a remote location to emissions from a given city? Crudely, ?how far away? do emissions from a city extend? Do prevailing wind conditions, topography etc influence the distribution of plume shape? Do image enhancement techniques such as deblurring provide improved representations of plumes? Some gases are more difficult to measure by satellite. By exploiting dependence between emissions of different gases, can we improve the estimation of hard-to-measure gases? Specifically, can we use measurements of nitrogen dioxide (NO2) to provide improved estimates of emissions of methane (CH4)? To what extent does knowledge of wind conditions assist the quantification of large sources of emissions? Would it be feasible to perform probabilistic inversion to estimate gas source characteristics?

External partner: The project will be jointly supervised by Philip Jonathan (Shell's Chief Statistician) and David Randell both active statisticians at Shell. During the project, there will be opportunities to visit Shell offices, and to meet experts in methane monitoring and satellite remote sensing.

Skills required: The project will be suitable for someone interested in statistical modelling of our physical environment. Since the project will involve exploratory analysis of TROPOMI data, visualisation and model-building, it is likely that the project will be more suitable for someone with a preference for applied and computational statistics.

References

[1] `http://www.tropomi.eu/data-products/methane`

[2] `https://scihub.copernicus.eu/`

Deutsche Bank projects for MSc in Statistics

DB projects, Summer 2020

Deutsche Bank offers an exciting internship in which the intern will face the challenge of applying theoretical concepts to real world data. Projects are available at multiple desks and they will all have focus on both practical implementations of models and exploration of theoretical concepts. The intern will be able to carry out his project in Deutsche Bank's office where he will have excellent computation facilities and the opportunity to interact with the desk's quant team on a daily basis.

Assessment will take place in late January/early February and you do not need to submit a preference towards a specific project before this.

Projects include:

FX alpha research

Prediction of short horizon price movements by using both classical and machine learning approaches. Besides creating models, your work will involve development of quantitative data cleaning approaches and performance measures. A high frequency dataset with relevant quote and trade data will be provided.

Cheyette PDE for Bermudan swaptions

Collapsing the 2D Cheyette PDE resolution to 1D and analysis of impact on Bermudan swaption products.

Static and dynamic hedging for IR/FX and hybrid products

Analysis and comparison of static hedging and dynamic hedging techniques for IR/FX and Hybrids products.

Optimal trading for bonds market making

Apply statistical and machine learning techniques to optimize the trading behaviour and inventory management in bonds from a market maker perspective. **Improvement of rates futures execution**

Using limit order book data from futures contracts on European Government Bonds, build a model of the trading dynamics to inform decision making about the placement of our own order orders. In particular develop an understanding of the risk vs return trade-off for aggressive versus passive orders, to support the development of a new trading algorithm. Your study will need to be informed by the matching engine specification and communication latency. You will be provided with access to high quality order book and trade data from the EUREX futures exchange.

Project presentation and contact details

There will be a session for presenting this project and the team from Deutsche Bank on Wendesday the 5th of February, 6-7pm, HXLY 340.

To declare interest for the project and register your attendance for the project presentation session for this project please contact Laurids Nielsen, laurids-a.nielsen@db.com with subject Summer placement for MSc students in Statistics and attach an updated CV.

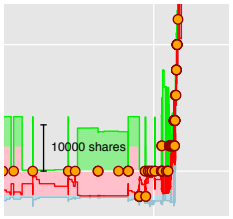
Industrial Supervisor: Laurids Nielsen, laurids-a.nielsen@db.com

Imperial Co-Supervisor: Mikko Pakkanen

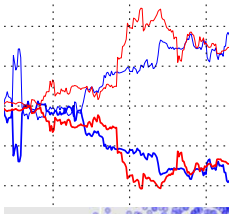
Projects in Statistical Finance

Supervisor: Mikko Pakkanen

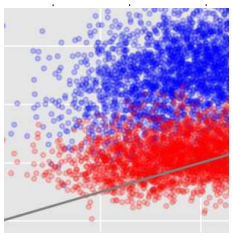
I will supervise one or two projects that showcase state-of-the-art methods of statistical finance, applying sophisticated statistical models and/or machine learning methods to large financial data sets or to challenging problems in computational finance. The exact topic and scope will be tailored according to the preferences of the selected student(s), but will be based on the following indicative research areas and questions:



Modelling of high-frequency financial data
⋆ Modelling of limit order book events using point processes
⋆ High-frequency price prediction using machine learning
⋆ Modelling of intra-day patterns in liquidity and/or trading volume
These projects can be done using high-quality tick-by-tick data from Nasdaq.



Volatility modelling
⋆ Modelling and prediction of realised volatility
⋆ Functional time series analysis and prediction of implied volatility smiles



Machine learning in computational finance
⋆ Pricing and hedging of derivatives by unsupervised deep learning
⋆ Robustness of deep learning in financial applications
⋆ Reinforcement learning in algorithmic trading

Prerequisites: The modules MATH97109 Algorithmic Trading and Machine Learning, MATH97131 Machine Learning, MATH97132 Introduction to Statistical Finance, MATH97133 Advanced Statistical Finance are in general useful but none of them is strictly necessary. Good programming skills (R or Python) are essential.

External partners: none

Questions?

Feel free to ask me questions about the projects after or between my Statistical Finance lectures or contact me via:

Email: `m.pakkanen@imperial.ac.uk`

Office hour: Fridays, 2:15-3:00pm, Room 801, Level 8, Weeks Building (16–18 Princes Gardens).

Statistical analysis of cryptocurrencies

Supervisor: Ioanna Papatsouma

Cryptocurrencies '*use strong cryptography to secure financial transactions, control the creation of additional units, and verify the transfer of assets*'. Launched in 2009, Bitcoin (BTC) is the world's largest cryptocurrency by market cap, followed by Ethereum (ETH), XRP, Bitcoin Cash (BCH) and 3279 more cryptocurrencies. Their meteoric rise has made it hard to ignore them and study their properties from a statistical perspective.

The aim of this project is to study the statistical properties of the returns of cryptocurrencies, identify the best fitting distributions, estimate the parameters of the distributions, build statistical models that can be used for prediction purposes and show the representation of principal components. All the above will be implemented using statistical methods, such as goodness-of-fit tests, hypothesis testing, maximum likelihood, regression analysis, principal component analysis etc. The project outcome can be used for investment purposes.

Prerequisite skills: programming skills in MATLAB, R etc. The particular choice will depend on the student's skills.

References:

1. M. Briere, K. Oosterlinck and A. Szafarz. Virtual currency, tangible return: Portfolio diversification with Bitcoins. *Journal of Asset Management* 16: 365–373, 2015.

2. J. A. Núñez, M. I. Contreras-Valdez and C. A. Franco-Ruiz. Statistical analysis of bitcoin during explosive behavior periods. *PLoS ONE* 14(3): e0213919, 2019.

3. S. McNally. Predicting the price of Bitcoin using Machine Learning. In: *26th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP)*, 339–343. IEEE, 2018.

4. D. Yermack. Is Bitcoin a Real Currency? An Economic Appraisal. In: Kuo Chuen David LEE editors. *Handbook of Digital Currency: Bitcoin, Innovation, Financial Instruments, and Big Data*, 31–43, London Wall: Elsevier, 2015.

A dual approach to linear discriminant analysis

Supervisor: Ioanna Papatsouma

Linear discriminant analysis (LDA) has its roots in an approach known as Fisher's linear discriminant analysis, a method used in statistics, pattern recognition and machine learning to find a linear combination of features that characterises or separates two or more classes of objects or events. It focuses simultaneously on the maximization of the between-class variance and the minimization of the within-class variance and it has been used to solve various real-life problems, such as emotion recognition by classifying the emotions or face recognition by reducing the number of pixel values.

The aim of this project is to study the LDA as a dimensionality reduction algorithm while retaining as much information as possible and as a classifier algorithm, apply both algorithms to real-life datasets and assess their performance. If time permits, we will be looking at possible techniques to make the analysis more robust.

Prerequisite skills: programming skills in MATLAB, R etc. The particular choice will depend on the student's skills.

References:

1. R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

2. G. J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognitio*n. Wiley Interscience, 2004.

3. T. T. Sajobi, L. M. Lix, B. M. Dansu, W. Laverty and L. Li. Robust descriptive discriminant analysis for repeated measures data. *Computational Statistics and Data Analysis*, 56: 2782–2794, 2012.

4. S. S. S. Yahaya, Y. F. Lim, H. Ali and Z. Omar. Robust Linear Discriminant Analysis. *Journal of Mathematics and Statistics*, 12(4): 312–316, 2016.

Using support vector machines for data classification

Supervisor: Ioanna Papatsouma

With the rapid development of information technology in this age, scientists have a wealth of data to classify, study and analyse. Support vector machines (SVMs) are a family of models used in statistics to classify data. They aim at finding decision boundaries that separate the data into classes and can be used to solve various real-life classification problems, such as text categorization or face recognition.

The SVM algorithm logic is based mainly on the construction of a hyperplane or a set of hyperplanes that separate a linearly-separable dataset into classes. In case of a non-linearly separable dataset, a kernel function is used in order to project data points into a higher dimension in which they become linearly separable.

The aim of this project is to study the SVMs from a statistical perspective and perform classifications on real-life datasets available at the UCI Machine Learning Repository. For datasets consisting of more than two classes, we will be looking at multi-class classification techniques, such as one-against-all or one-against-one.

Prerequisite skills: programming skills in MATLAB, R etc. The particular choice will depend on the student's skills.

References:

1. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

2. N. Cristianini and J. Shawe-Taylor. *An introduction to Support Vector Machines and other kernel-based learning methods.* Cambridge University Press, New York, NY, USA, 2000.

3. T. Joachims. Text categorization with Support Vector Machines: Learning with many relevant features. *ECML'98: Proceedings of the 10th European Conference on Machine Learning*, 137–142, 1998.

4. T. Hastie, R. Tibshirani and J. Friedman. *The Elements of Statistical Learning Data Mining, Inference, and Prediction (2nd edition).* Springer Series in Statistics, 2009.

5. V. N. Vapnik. *The Nature of Statistical Learning Theory*, Springer, New York, 1996.

A novel machine learning technique to characterise bird song as a model for social learning and cultural evolution

Background. Songs of oscine songbirds have long been studied as a model for social learning, speech, and cultural evolution. For example, recently, we demonstrated that some bird songs have likely persisted for many hundreds of years, suggesting that non-human cultural traditions can match those of humans [1]. At the core of this work are statistical procedures that decompose songs into lower-level units to quantify the similarity, variation, and evolution of songs.

Project. You will join a team of field biologists, evolutionary modellers and statisticians to develop a better statistical approach for decomposing songs into their lower-level units. Various algorithms and machine learning approaches have previously been evaluated for other animal models, and so there is a well-defined starting point for this work [3, 4, 2]. However, for song birds, the task is complicated by the fact that the primary cue (gap between vocalisations) is only mostly reliable. Other sources of information, such as patterns of certain repetitions, are also needed. Reasonably large corpora of human-annotated songs from different species are available for training and testing, and further scores/data can be generated by the team that you will be working with.

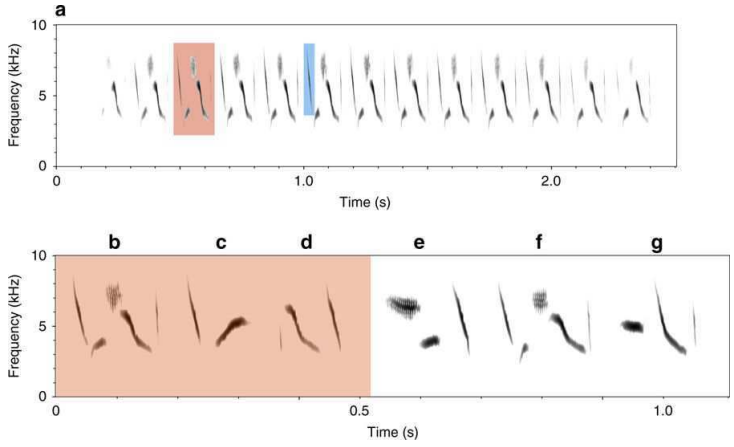


Figure 1: **Swamp sparrow song structure.** (**a**) Spectrogram of a swamp sparrow song sung by an individual from the Hudson Valley population. Swamp sparrow songs consist of one syllable (red) repeated 10 or more times. Each syllable consists of 2-5 elements or notes. (**b-g**) Examples of different syllable types. (**b-d**) Show the syllables that make up the repertoire of the same individual whose song is shown in (**a**). (**e-g**) Show three syllable types sung by other males in the population, illustrating both the considerable diversity in syllable structure found within a population, but also how, as a consequence of vocal learning, different individuals also sometimes share the same syllable-type (**b, f**)

Supervisors: Oliver Ratmann (primary), Robert Lachlan, Dept of Psychology, Queen Mary University of London (secondary)

Stream suitability: General, Applied Statistics, Biostatistics

Recommended MSc courses: Bayesian Data Analysis, Machine Learning

References

[1] Robert F Lachlan, Oliver Ratmann, and Stephen Nowicki. Cultural conformity generates extremely stable traditions in bird song. *Nature Communications*, 9(1):2417, 2018.

[2] Oisin Mac Aodha, Rory Gibb, Kate E Barlow, Ella Browning, Michael Firman, Robin Freeman,

Briana Harder, Libby Kinsey, Gary R Mead, Stuart E Newson, et al. Bat detective—deep learning tools for bat acoustic signal detection. *PLoS Computational Biology*, 14(3):e1005995, 2018.

[3] Stuart Parsons and Gareth Jones. Acoustic identification of twelve species of echolocating bat by discriminant function analysis and artificial neural networks. *Journal of Experimental Biology*, 203(17):2641–2656, 2000.

[4] Vassilios Stathopoulos, Veronica Zamora-Gutierrez, Kate E Jones, and Mark Girolami. Bat echolocation call identification for biodiversity monitoring: A probabilistic approach. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(1):165–183, 2018.

61

Bayesian viral phylogenetic source attribution: a new approach that combines phylogenetic data with estimates of infection times

Background. A key part of viral phylogenetics analyses is to reconstruct who infected whom from viral sequence data (Cohen, 2015; Oster et al., 2018). This approach of using genetic data in this way is now becoming highly influential in epidemiological investigations and policy making, because: (1) the cost of sequencing has fallen; (2) viral RNA data provides more objective information about how transmission occurred than to questionnaire answers; and (3) large amounts of data are available for analysis, especially for understanding the spread of HIV.

Project. You will formulate a Bayesian hierarchical model to describe the relationship between infected individuals and their phylogenetically identified probable source cases. The work will build on the hierarchical model proposed by (Ratmann et al., 2016), and aim to extend this model by integrating estimates of infection times (Pantazis et al., 2005; Stirrup and Dunn, 2018). You will implement a custom MCMC sampler. Each source case has low probability to be the actual transmitter, however considering many such source-recipient cases enables inference of population-level transmission patterns. This project is for someone with a strong interest in statistical modelling, integration of various data sources, and applied analysis.

The project is linked to an ongoing research project on estimating the sources of HIV infection in Greece in the last 5 years. In particular, during the economic crisis, a large outbreak of injection drug users occurred, which caught international attention (Paraskevis et al., 2018; Sypsa et al., 2017). Individual-level data on infected individuals have been pre-processed by the project partners in Greece, offering a distinct opportunity to perform an original data analysis.

Supervisors: Oliver Ratmann (primary), Dimitrios Paraskevis (Medical School of the National and Kapodistrian University of Athens), Nikos Pantazis (Medical School of the National and Kapodistrian University of Athens)

Stream suitability: General, Applied Statistics, Biostatistics

Recommended MSc courses: Bayesian Data Analysis

References

Cohen, J. (2015). Hiv family trees reveal viral spread. *Science*, 348(6240):1188–1189.

Oster, A. M., France, A. M., and Mermin, J. (2018). Molecular Epidemiology and the Transformation of HIV Prevention. *JAMA*, 319(16):1657–1658.

Pantazis, N., Touloumi, G., Walker, A. S., and Babiker, A. G. (2005). Bivariate modelling of longitudinal measurements of two human immunodeficiency type 1 disease progression markers in the presence of informative drop-outs. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(2):405–423.

Paraskevis, D., Nikolopoulos, G. K., Sypsa, V., Psichogiou, M., Pantavou, K., Kostaki, E., Karamitros, T., Paraskeva, D., Schneider, J., Malliori, M., et al. (2018). Molecular investigation of hiv-1 cross-group transmissions during an outbreak among people who inject drugs (2011–2014) in athens, greece. *Infection, Genetics and Evolution*, 62:11–16.

Ratmann, O., van Sighem, A., Bezemer, D., Gavryushkina, A., Jurriaans, S., Wensing, A., de Wolf, F., Reiss, P., Fraser, C., and (2016). Sources of hiv infection among men having sex with men and implications for prevention. *Science Translational Medicine*, 8(320):320ra2–320ra2.

Stirrup, O. T. and Dunn, D. T. (2018). Estimation of delay to diagnosis and incidence in hiv using indirect evidence of infection dates. *BMC medical research methodology*, 18(1):65.

Sypsa, V., Psichogiou, M., Paraskevis, D., Nikolopoulos, G., Tsiara, C., Paraskeva, D., Micha, K., Malliori, M., Pharris, A., Wiessing, L., Donoghoe, M., Friedman, S., Jarlais, D. D., Daikos, G., and

Hatzakis, A. (2017). Rapid Decline in HIV Incidence Among Persons Who Inject Drugs During a Fast-Track Combination Prevention Program After an HIV Outbreak in Athens. *The Journal of Infectious Diseases*, 215(10):1496–1505.

A non-parametric Bayesian approach to infer age-specific social contact rates

Background. In the last fifteen years, data from contact surveys have been increasingly used to describe the heterogeneity in the transmission of infection diseases due to social mixing (Mossong et al., 2008). There does not exist a widely accepted method to generate smoothed estimates of contact rates, and properly quantify uncertainty resulting from such surveys. There is also no method available to effectively integrate these data into infectious disease transmission analyses (Baguelin et al., 2013) and (Birrell et al., 2011).

Project. You will join a team of epidemiologists, mathematical modellers and statisticians to develop a non-parametric statistical approach for inferring social contact rates using Gaussian-process based smoothing, or other kernel-based machine learning approaches (Diggle and Ribeiro, 2007; Ton et al., 2018; Genton, 2002). The statistical framework will build on the Bayesian hierarchial model proposed by van de Kassteele et al. (2017). Here, you will investigate alternative classes of non-parametric prior distributions, and investigate if/when the model can be efficiently implemented in Stan (Carpenter et al., 2017). This will introduce you to common techniques that underlie statistical machine learning, as well as widely used inference techniques.

Building on very recent work (Chatzilena et al., 2019), we hypothesise that the above approach could provide a general framework for estimating disease spread from contact data and disease count data. If progress is going well, we would encourage you to investigate this hypothesis (see Figure 1 and (Baguelin et al., 2013)), which would link to state-of-the-art research.

Supervisors: Oliver Ratmann (primary), Marc Baguelin (secondary, Dept of Infectious Disease Epidemiology), Nikos Demiris (external, Athens University of Economics and Business, Greece)

Stream suitability: General, Applied Statistics, Biostatistics

Recommended MSc courses: Bayesian Data Analysis, Machine Learning

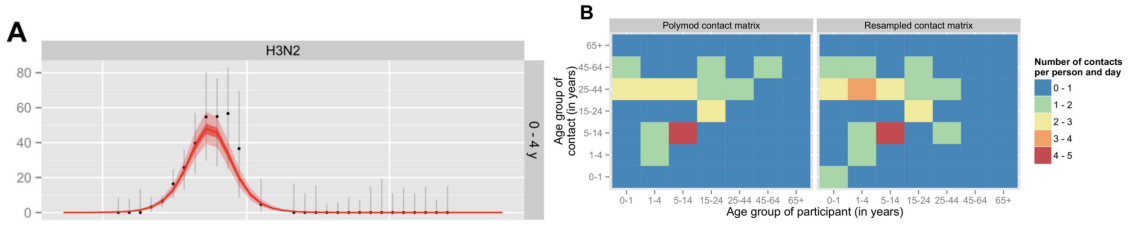


Figure 1: **Inference of a transmission matrix using information from contact surveys.** (**A**) A transmission model is used to reconstruct disease dynamics in particular age groups. The model outcome is in red and the data points from epidemiological surveillance are in black. (**A**) Intensity of contacts given by POLYMOD matrix (B - left) and intensity of contact inferred from bootstrap algorithm and used to reconstruct transmission in the epidemiological model (B - right). All figures taken from (Baguelin et al., 2013).

References

Baguelin, M., Flasche, S., Camacho, A., Demiris, N., Miller, E., and Edmunds, W. (2013). Assessing Optimal Target Populations for Influenza Vaccination Programmes: An Evidence Synthesis and Modelling Study. *PLoS Medicine*, 10(10).

Birrell, P. J., Ketsetzis, G., Gay, N. J., Cooper, B. S., Presanis, A. M., Harris, R. J., Charlett, A., Zhang, X.-S., White, P. J., Pebody, R. G., and De Angelis, D. (2011). Bayesian modeling to unmask and

predict influenza A/H1N1pdm dynamics in London. *Proceedings of the National Academy of Sciences of the United States of America*, 108(45):18238–43.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, Articles*, 76(1):1–32.

Chatzilena, A., van Leeuwen, E., Ratmann, O., Baguelin, M., and Demiris, N. (2019). Contemporary statistical inference for infectious disease models using stan. *Epidemics*, 29:100367.

Diggle, P. and Ribeiro, P. (2007). *Model-based Geostatistics.* Springer Series in Statistics. Springer.

Genton, M. G. (2002). Classes of kernels for machine learning: A statistics perspective. *J. Mach. Learn. Res.*, 2:299–312.

Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., Mikolajczyk, R., Massari, M., Salmaso, S., Tomba, G. S., Wallinga, J., Heijne, J., Sadkowska-Todys, M., Rosinska, M., and Edmunds, W. J. (2008). Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS medicine*, 5(3):e74.

Ton, J.-F., Flaxman, S., Sejdinovic, D., and Bhatt, S. (2018). Spatial mapping with gaussian processes and nonstationary fourier features. *Spatial Statistics*, 28:59 – 78. One world, one health.

van de Kassteele, J., van Eijkeren, J., and Wallinga, J. (2017). Efficient estimation of age-specific social contact rates between men and women. *Ann. Appl. Stat.*, 11(1):320–339.

Bayesian causal inference

Kolyan Ray

Inferring the causal effect of a treatment is an important problem in many applications, such as healthcare, education and economics. While carefully designed experiments are the gold standard for measuring causal effects, these are often impractical due to ethical, financial or time-constraints. An alternative is to use observational data which, while typically easier to obtain, requires careful analysis.

It is known that naively using Bayesian methods can yield poor results in causal inference problems [1]. Recently, it was shown that suitably correcting the prior can somewhat fix this problem in both theory [2] and practice [3].

The goal of this project is to first review the (Bayesian) literature for estimating average treatment effects in causal models. The idea in [2] has only been empirically investigated in the simplest setting [3], where conjugate computations are possible, and a next step would be to investigate this further. Possible extensions are to problems with binary or categorical outcomes or the use of hierarchical priors, where Markov chain Monte Carlo methods become necessary. Another question not considered in [3] is hyperparameter selection, where significant gains are possible.

[1] J. Robins & Y. Ritov. Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in Medicine* (1997).
[2] K. Ray & A.W. van der Vaart. Semiparametric Bayesian causal inference. *Annals of Statistics*, to appear.
[3] K. Ray & B. Szabo. Debiased Bayesian inference for average treatment effects. *Neural Information Processing Systems* (2019).

The Bernstein-von Mises Theorem

Kolyan Ray

The Bernstein-von Mises (BvM) theorem states that in many situations of interest, the Bayesian posterior distribution is approximately Gaussian if the number of observations is large enough. Furthermore, the mean and covariance of this Gaussian distribution are such that inference based on the posterior distribution asymptotically coincides with inference based on the frequentist maximum likelihood estimator. The BvM theorem thus provides a theoretical frequentist justification for the use of Bayesian methods.

The purpose of this project is to first review the BvM theorem for standard parametric models, where modern proofs rely on the theory of locally asymptotically normal models (models whose log-likelihood admits a certain quadratic expansion), see [1]. Beyond that, there are plenty of possible extensions: e.g. generalizations to nonparametric or semiparametric models [2,3], more complex priors or inverse problems [this is an active area of research in theoretical statistics].

Useful courses: Bayesian Methods.

Some knowledge of measure theory is helpful but not necessary.

[1] A.W. van der Vaart. Asymptotic Statistics, Cambridge University Press (1998).
[2] I. Castillo & J. Rousseau. A Bernstein–von Mises theorem for smooth functionals in semiparametric models. *Annals of Statistics* (2015).
[3] I. Castillo & R. Nickl. Nonparametric Bernstein-von Mises Theorems in Gaussian White Noise. *Annals of Statistics* (2013).

Predicting intra-daily residual demand

Supervisor: Almut Veraart

Project description

An ever increasing amount of energy is produced from renewable resources and the key challenge is to satisfy our increasing energy demand in a reliable way while at the same time reducing carbon emissions. In this project we will focus on predicting the so-called residual demand, which is the difference between the electricity demand and the wind and solar energy production based on intra-daily data.

One of the key challenges in this project is to deal with the various seasonalities arising in the wind energy production, solar energy production and electricity demand.

Moreover, we need to model potential heteroscedasticity in the data.

Initially the project will focus on point forecasts, but these will later be extended to probabilistic forecasts.

The student will then compare the forecast performance of various models and evaluate which model performs best on intra-daily time scales.

Additional information

Prerequisite skills A good background in time series analysis as well as good programming skills are required. The particular choice of the programming language will depend on the student's skills and preferences.

Essential courses Core courses, Time Series.

External partners/co-supervisors None.

Stream suitability General, Applied, Theory and Method, Data Science.

Predicting stock-price moves for micro-cap equities using news analytics

Supervisors: Almut Veraart, Neel Bhagodia (3C Capital)

Project description

Publicly listed equities typically form part of an investor's portfolio, as asset holdings are engineered to meet their risk/return profile. While larger cap equities (such as Google and Amazon) attract extensive coverage and analysis, lesser covered micro-cap equities present opportunities for higher returns with higher risk attached, as price volatility is a key reason. This project aims to research which factors can be used to reliably predict an upcoming movement in stock price for micro-cap equities, allowing investors to make more informed decisions on the riskier portion of their investment holdings.

Data: We are planning to use a dataset of historical intra-day micro-cap equity stock price data and a dataset containing intra-day news announcements.

Disclaimer: We are still in the process of sourcing the relevant data and the project might need to be modified in case there are any problems with the data sources.

Additional information

Prerequisite skills A good background in time series analysis as well as good programming skills are required. The particular choice of the programming language will depend on the student's skills and preferences.

Essential courses Core courses, Time Series (this does not need to be the Imperial time series course, but could be an equivalent course you took during undergraduate studies), Introduction to Statistical Finance, Advanced Statistical Finance.

External partners/co-supervisors Neel Bhagodia (3C Capital)

Stream suitability General, Applied, Data Science, Statistical Finance.

Spatio-temporal volatility assessment of wind energy production in the Nord Pool Market

Supervisor: Almut Veraart

Project description

Renewable energy production is highly volatile. In this project we aim to quantify the volatility of wind energy production in the Nord Pool market.

We have hourly wind energy production data available for various areas in Scandinavia. We will use tools such as realised variance and realised covariance, which were initially developed in financial econometrics to measure (and potentially predict) the (stochastic) volatility and covariance of financial assets, to assess the dependencies in the wind energy production volatility across various regions.

This project is suitable for students who would like to study continuous-time stochastic processes (in particular Itô semimartingales and related processes) through guided reading and then explore statistical techniques for estimating stochastic processes (and not just parameters) in a consistent way.

Additional information

Prerequisite skills A good background in time series analysis, continuous-time stochastic processes as well as good programming skills are required. The particular choice of the programming language will depend on the student's skills and preferences.

Essential courses Core courses, Time Series, Introduction to Statistical Finance, Advanced Statistical Finance.

External partners/co-supervisors None.

Stream suitability General, Applied, Theory and Methods, Data Science, Statistical Finance.

Adversarial training for audio source separation

K. N. Webster

Deep learning models have greatly improved the performance of audio source separation models, and there is an emerging trend towards end-to-end learning for this task, dispensing with traditional STFT preprocessing of the audio signal. However, most of the developments have still focused on supervised training of neural networks, and require a considerable amount of training data. In a recent work [1], the authors proposed a fully unsupervised deep learning autoencoder model for source separation, that requires no pre-training. This initial work serves as a proof of concept for this architecture.

The aim of this project is to extend and improve the previous work by providing a level of guidance to the source separation by the means of an adversarial training-inspired architecture. The discriminator in this case will be trained on a dataset of audio that is typical of the type of source that is to be extracted. No mixture and separation data examples will be provided to the network during training, so that the architecture can still make use of more readily abundant audio data. The proposed architecture has similarities to [2].

Prerequisite skills: the project will involve involve significant work using Python and a deep learning framework, either Tensorflow or PyTorch. Experience in working with deep learning models is essential for this project.

Essential Courses: Machine Learning

External Partners/Co-supervisors: None

References

[1] A. Bergner and K. N. Webster, "Unsupervised single channel source separation autoencoders", *Proceedings of the 5th Workshop on Intelligent Music Production 2019, Birmingham, UK* (2019).

[2] N. Mor, L. Wolf, A. Polyak, and Y. Taigman. "A universal music translation network", *Proceedings of the International Conference on Learning Representations* (2019).

Alternatives to Proportional Hazards Regression

Supervisor: David Whitney

2019-2020

The proportional hazards (PH) regression model stipulates that the conditional hazard function h_x of the distribution of T given $X = x$ satisfies

$$h_x(t) = h_0(t) \exp(\theta x) \text{ for all } t > 0 \ ,$$

where h_0 is an unspecified baseline hazard and θ is the scalar regression coefficient of interest. Under proportional hazards, the regression coefficient θ_0 is the (constant) hazard ratio value. Instead of complete observations, it is common to observe possibly right-censored event times. When the censoring variable is conditionally independent of T given X, and the PH model indeed holds, the maximizer of the partial likelihood is known to be a consistent estimator of the true regression coefficient.

When instead the PH model is used in settings where its assumptions fail, one may hope that the resulting estimator perhaps represents an average of the time-varying hazard ratio (on a logarithmic scale). It has been shown that this is indeed approximately true, though the limit in probability θ_* of the maximum partial likelihood estimator (MPLE) depends not only on the conditional time-to-event and marginal covariate distributions but also on the conditional censoring distribution [Struthers and Kalbfleisch, 1986] in a complicated manner.

A natural summary of a time-varying hazard ratio when $X = 0, 1$ is a binary covariate is

$$\theta_* := \int \log \left\{ \frac{h_1(t)}{h_0(t)} \right\} \nu(dt) \ ,$$

where h_x is the true hazard function corresponding to $X = x$, and ν represents a weight function, possibly dependent on components of the data-generating distribution. A flexible estimator of θ_* was proposed by Whitney et al. [2019]. This project will focus on evaluating the proposed estimator relative to the PH model estimator and competitors in a variety of simulation settings.

A motivated, independent student could also work on extending the parameter to settings in which X is a continuous covariate, or other dependent variables W are available.

Previous experience with survival data analysis would be helpful, though not necessary to be successful in this project.

References

Cyntha A Struthers and John D Kalbfleisch. Misspecified proportional hazard models. *Biometrika*, 73(2): 363–369, 1986.

David Whitney, Ali Shojaie, and Marco Carone. Comment: Models as (deliberate) approximations. *Statist. Sci.*, 34(4):591–598, 11 2019. doi: 10.1214/19-STS747. URL `https://doi.org/10.1214/19-STS747`.

Doubly Robust Inference for the Average Treatment Effect

Supervisor: David Whitney

2019-2020

The average treatment effect (ATE) is frequently used to quantify the population-wide impact of a new intervention, hereafter the "treatment." Commonly used augmented inverse probability of treatment weighted (AIPTW) and targeted maximum likelihood (TML) estimators are semiparametric efficient under weak assumptions [van der Laan and Rubin, 2006].

Key to establishing the asymptotic distribution of these estimators for the ATE is the consistent estimation of two nuisance regression functions, the propensity score and outcome regression. The AIPTW and TML estimators are doubly robust estimators: they are consistent if at least one of the nuisance regressions is estimated consistently, but not asymptotically normal unless both regressions are estimated consistently (at a sufficiently fast rate).

Pioneering work by van der Laan [2014] and Benkeser et al. [2017] demonstrates that asymptotic normality for TML estimators of the ATE can be preserved, provided one of the nuisance regressions is estimated consistently.

Project 1: Empirical Likelihood

Empirical likelihood (EL) methods were developed by Owen [1988] as a nonparametric analogue to the likelihood ratio test. Since then, EL has been studied in a variety of settings. Recent work by Bravo et al. [2019] shows that (under regularity conditions) EL methods remain valid when machine learning estimators of nuisance parameters are used as a first step in constructing an estimator of the target parameter. Bravo et al. [2019] prove that their methods apply to estimators of the ATE.

This project focuses on comparing doubly robust estimation and inference for the ATE based on EL to the TML estimators seen so far in the literature. There are several potential directions:

- Qin and Lawless [1995] show that EL can be used for parameter estimation, allowing for additional (known) moment constraints. By imposing moment constraints analogous to those in Benkeser et al. [2017], a maximum EL estimator for the ATE can be defined that may exhibit the same asymptotic properties as the TML estimators. Demonstrating that the proposed estimator of the ATE has the desired robustness properties is thus of interest. This will be likely be achieved through a combination of theoretical and numerical results.

- While the proposed EL estimator is expected to be asymptotically equivalent to the TML estimator, the bias of the estimators and the coverage of associated confidence intervals could be compared at small and moderate sample sizes through simulation studies.

- It may also be explored whether elements of both the TML estimator and EL approach to confidence interval and hypothesis test construction can be combined to achieve better overall performance than using either framework in isolation.

Project 2: High-dimensional Data

There has been a recent flurry of interest in utilizing doubly robust estimators in high-dimensional data settings. Work by authors such as Dukes et al. [2018], Tan [2018], Ning et al. [2018], Smucler et al. [2019], and Bradic et al. [2019] study a variant of doubly robust inference that arises when there is a large number of covariates. In their work, the authors study several approaches to inference for parameters (such as the ATE) that are doubly robust to misspecification of the sparsity of the nuisance estimators. However, the specific estimators considered by these authors typically make strong assumptions (e.g., linearity) about the functional form of the nuisance parameters.

This project will focus on comparing doubly robust inference developed in the high-dimensional contexts above with the TML approach of Benkeser et al. [2017] through numerical (simulation) studies.

References

David Benkeser, Marco Carone, Mark J van der Laan, and Peter B Gilbert. Doubly robust nonparametric inference on the average treatment effect. *Biometrika*, 104(4):863–880, 2017.

Jelena Bradic, Stefan Wager, and Yinchu Zhu. Sparsity double robust inference of average treatment effects. *arXiv preprint arXiv:1905.00744*, 2019.

Francesco Bravo, Juan Carlos Escanciano, and Ingrid van Keilegom. Two-step semiparametric empirical likelihood inference. *Annals of Statistics*, 2019.

Oliver Dukes, Vahe Avagyan, and Stijn Vansteelandt. High-dimensional doubly robust tests for regression parameters. *arXiv preprint arXiv:1805.06714*, 2018.

Yang Ning, Sida Peng, and Kosuke Imai. Robust estimation of causal effects via high-dimensional covariate balancing propensity score. *arXiv preprint arXiv:1812.08683*, 2018.

Art B Owen. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2): 237–249, 1988.

Jing Qin and Jerry Lawless. Estimating equations, empirical likelihood and constraints on parameters. *Canadian Journal of Statistics*, 23(2):145–159, 1995.

Ezequiel Smucler, Andrea Rotnitzky, and James M Robins. A unifying approach for doubly-robust ℓ_1 regularized estimation of causal contrasts. *arXiv preprint arXiv:1904.03737*, 2019.

Zhiqiang Tan. Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data. *arXiv preprint arXiv:1801.09817*, 2018.

Mark J van der Laan. Targeted estimation of nuisance parameters to obtain valid statistical inference. *The International Journal of Biostatistics*, 10(1):29–57, 2014.

Mark J van der Laan and Daniel Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.

Probability matching in Bayesian selective inference

Supervisor: Alastair Young

Abstract:
In Bayesian inference an important theoretical notion is the identification of a *probability matching prior*: a prior which, if used in a standard Bayesian analysis, yields posterior confidence limits for the parameter of interest with, to a high degree of accuracy, the correct frequentist interpretation.

Bayesian *selective inference* concerns construction of a posterior distribution for a parameter of interest, recognising that the inference is being performed only because some selection event E on the sample data y has been observed. This may be performed by treating the selection as a *data truncation*, and applying the appropriate truncated likelihood (the density of $y|E$) in the Bayesian calculation.

This project will investigate the construction of probability matching priors for Bayesian selective inference. What is the (asymptotic) theoretical form of a matching prior? What are the observed frequentist properties of confidence limits calculated from the relevant selective posterior distribution? Does the selection have any significant effect on the appropriate form of matching prior or repeated sampling properties, compared to ignoring the selection? Does performing selection on a randomised version of sample data, instead of the sample data itself, make specification of a matching prior easier?

Prerequisites: Fundamentals of Statistical Inference, Advanced Statistical Theory. Willingness to study theory and tackle numerical validation simulations in R.

External Partners: None.

Small Sample Inference from Big Data

Supervisor: Alastair Young

Abstract:
The era of 'big data' does not reduce the need for effective methods of statistical inference with very small data samples. For instance, particle physics experiments such as those conducted in the Large Hadron Collider produce huge volumes of data, but these are reduced to just a few data values, from which it is hoped that a signal will be detected in the presence of background noise. This project will have as its primary aim the analysis of simulation-based ('bootstrap') methods of inference with small data sample size, as alternatives to more analytic approaches. A starting point may be examination of probability models for signal detection in high energy physics discussed by Davison and Sartori [2008, Statistical Science, 23, 354-64]. The project may evolve in various directions, according to interests: (i) comparisons when the interest parameter is *vector valued* ; (ii) analysis of small data samples from other fields; (iii) adaptations of simulation-based methods of inference to cope with restricted parameters [see, for example, Fraser, Reid and Wong, 2004, Phys. Rev. D, 69, 033002]; (iv) detailed evaluations of objective Bayes approaches to inference with small samples.

Prerequisites: Fundamentals of Statistical Inference, Advanced Statistical Theory.

External Partners: None.

Sparse Estimation of Many Normal Means

Supervisor: Alastair Young

Abstract:

The many normal means problem serves as a template for parameter estimation in high-dimensional or 'large-scale' inference [see the book length treatment 'Large-scale Inference' by Efron, 2010]. Of particular interest in many fields, such as genomics, is the sparse case, where many or most of the true means are zero. In these circumstances, there are advantages to use of a (soft) threshold estimator, which allows estimation of some, but not necessarily all, of the true means as zero. But this process requires specification of a threshold parameter. Two methods for choosing the threshold parameter are: fixing it according to some global rule (depending only on the dimension); data-driven choice, by minimising Stein's unbiased risk estimate (SURE). There is some evidence [see, for example, Donoho and Johnstone, 1995, JASA, 90, 1200-1224] that which approach is best depends on the underlying (unknown) degree of sparsity. This project will develop adaptive ways of specification of the threshold estimator, with the aim of developing estimators which are reliable and efficient irrespective of the degree of sparsity.

Prerequisites: Fundamentals of Statistical Inference. Willingness to study theory and tackle numerical validation simulations in R.

External Partners: None.