L-Università ta' Malta
Faculty of Science

Department of
Statistics &
Operations Research

Statistics and Operations Research

Final Year Dissertation Presentations
B.Sc. (Hons.) SCI Yr IV

10/07/2023

Programme of Dissertation Presentations

Class of 2023...

Estimation Methods of Lévy Processes – A Comparative Study

by

Amy Grech

Abstract

Lévy processes became increasingly popular in the 20th century due to their ability to cater for jumps and thus have been applied to various fields, such as finance. In this dissertation, the theory of Lévy processes will be presented, along with some of their properties. Then, we will discuss different parameter estimation methods which are divided into three categories: the parametric, the semi-parametric and the non-parametric approach. This dissertation focuses on methods which fall under the parametric approach, mainly the Maximum Empirical Likelihood Estimation developed by Qin and Lawless (1994), the Integrated Squared Error Estimate proposed by Heathcote (1977) and the Generalized Method of Moments proposed by Hansen (1982). These methods are then applied to a simulation study and local financial data, and then compared.

A Multi-Neighbourhood Local Search Algorithm for the University Course Timetabling Problem

by

Martina Grima

Abstract

The University Course Timetabling Problem (UCTP) is an example of a real-world educational timetabling problem that deals with the allocation of a set of courses and lecturers to a set of rooms and timeslots subject to certain restrictions. The objective and constraints of the problem will vary according to the specific requirements of different institutions. The construction of course timetables is a challenging and time-consuming task and, in fact, the UCTP is classified as an NP-Hard optimization problem.

In this dissertation, we tackle a UCTP relating to real life data obtained from the Faculty of Science at the University of Malta. An original mathematical model is formulated specifically for this data and different problem instances are solved. Since the UCTP is an NP-Hard optimization problem, exact methods will take an exponentially longer time to converge to a globally optimal solution as the size of the problem increases. As a result, heuristic solution methods are a common alternative since they provide practical solutions in a timely manner. Both the Branch and Cut (B&C) exact method and a heuristic algorithm are applied to our problem and compared. The heuristic algorithm consists of two phases; an initial feasible solution construction phase, in which a greedy algorithm is used, and a solution quality improvement phase, in which Multi-Neighbourhood Local Search (MNLS) is performed. The results show that for small instances, the proposed MNLS algorithm can be considered a good alternative to B&C. For larger instances, it still finds relatively good solutions but running time increases significantly.

Prediction of Jellyfish Reports in Maltese Coastal Locations with Weighted Decision Trees

by

Owen Theuma Fontebrera

Abstract

Forecasting the presence of jellyfish in many beaches across the world is a daunting task due to various reasons. Although jellyfish have a positive impact on the ecosystem, their interaction with human activities can lead to significant problems. The central theme of this dissertation is to explore the application of weighted decision trees for classifying the reasons for jellyfish presence in coastal areas surrounding the Maltese islands. The dissertation will first delve into the theory behind cluster analysis, particularly weighted K-means clustering, which will be used to cluster bay locations around the Maltese Islands. Moreover, the dissertation will also explore the theory behind classification trees, particularly the weighted classification and regression tree algorithm. The weighted decision tree approach will be used to classify jellyfish presence using multiple oceanographic variables. Different weighted classification tree models will be evaluated based on their performance. This dissertation further aims to uncover interesting features of the relationship between jellyfish and oceanographic variables.

Modelling the Expected Insurance Claims using the Tweedie compound Poisson Model

by

Giada Fiorini

Abstract

A common problem in the insurance industry is that of modelling how much a policy holder will cost an insurance company. An actuary views the characteristics of a policy holder to determine the risk profile involved. In this research the Tweedie compound Poisson model is used as an approach to model insurance claims, which is a challenging task due to the rare event of a policy holder making a claim, resulting in an excess number of zeros in the data. This was firstly done by showing that the Tweedie family belongs to the exponential dispersion family, so that the model can be fit as a Tweedie generalised linear model. The framework studied follows from the widely utilised regression model in insurance claims, which is the compound Poisson model with gamma claim sizes. By reviewing the family of Tweedie's distribution, a re-parametrization of the compound Poisson model called Tweedie compound Poisson model was obtained and used to model the claim sizes. The profile likelihood estimation method, for one of the parameters, was used. Once this estimate was obtained, the generalised linear model framework using the maximum likelihood estimation followed to obtain the parameter of interest which is the expected total claim size. The importance of accurately modelling insurance claims and the literature on the topic is discussed in the introduction. Followed by the framework described and finally, an application of the model is applied to a dataset obtained from the R package 'insuranceData'.

Long Short-Term Memory Networks for Forecasting Sea Level and Seiche Occurrences in Malta

by

Nicole Borg

Abstract

As recently as 30th June 2022, a phenomenon which affected several Maltese coasts and harbours was noticed by the general public and affected the activities that the local citizens would normally carry out on a particular day. The phenomenon which was observed was referred to as a seiche. Seiches are standing waves that form in enclosed or partially enclosed bodies of water that can pose a threat to the safety of marine infrastructure and harbours. The ability to predict seiches can help prevent the damage and mitigate the risks associated with these natural phenomena. This dissertation presents a novel approach for seiche prediction in Malta using Long Short-Term Memory (LSTM) neural network models.

The proposed LSTM models are trained using time series data obtained from two tide gauge stations installed in Marsaxlokk and Portomaso by the Oceanography Malta Research Group (OMRG), University of Malta. Due to the presence of missing data, this dissertation also explores the use of gap filling methods to obtain complete time series datasets which could then be used for the LSTM model training and testing. Two LSTM neural network models are presented, where in the first instance the LSTM model is trained on the Portomaso dataset and tested on the Marsaxlokk dataset, following calibration. The other LSTM model is trained on the first 50% of the Marsaxlokk dataset and tested on the remaining 50% of the dataset. In both cases, we either consider rolling variance of sea level as a predictor, or we do not, to determine if this additional feature which needs extracting from the data improves the predictive ability of the model. We show that the LSTM models' ability to take into account long-term dependencies is useful for predicting sea levels. In particular, the seiches that occurred in 28th November, 2021 and in 30th June, 2022, were recognised by the LSTM models, including the high sea level fluctuations that occur with these phenomena.

An Application of Stochastic Dynamic Programming to Hybrid Electric Vehicles

by

Mikela Mejlaq

Abstract

Reducing carbon emissions is a goal which is currently trying to be achieved globally in order to combat the climate crisis. In light of this, electric mobility such as the use of Hybrid Electric Vehicles (HEVs) is prominent at the moment. When thinking specifically about HEVs and the way in which these help this global cause, a natural question arises: “Is there any way that the fuel consumed when using an HEV can be reduced or kept to a minimum?”.

This dissertation aims to answer this question, with the aid of Stochastic Dynamic Programming techniques. A Markov Decision Process (MDP) is created to model as accurately as possible the vehicle dynamics. The components which were chosen to be represented were based on similar studies already conducted by other researchers. The main idea is that given a state of charge of the vehicle’s battery, power demand of the driver, speed of the vehicle and whether the Internal Combustion Engine (ICE) is on or off, a specific value of ICE power must be reached in order to minimize fuel consumption as required. Real-life data corresponding to a Toyota Camry 2010 HEV was used to solve different scenarios of the MDP through two SDP algorithms, namely the Policy and Value Iteration algorithms. Similar results were observed for both algorithms. The results show that there are indeed improvements that can be made when following the rules obtained, rather than choosing randomly.

Dimension Reduction and Bayesian Asymmetric Volatility Modelling of Portfolios with Many Assets

by

Matthew Camilleri

Abstract

The impact of technology on financial markets led to a global domino effect, leaving corporations scrambling for a more reliable way of measuring their risks. This study proposes a Bayesian approach for estimating and forecasting Asymmetric GARCH models with Student's t-distributed errors for a hedge fund portfolio with many assets, where the dimension of the dataset is reduced using dynamic principal component analysis (DPCA). MCMC methods simulate the posterior distribution of the model parameters and latent variables, from which risk measures such as value at risk (VaR) and expected shortfall (ES) are estimated. Asymmetry in GARCH models is an important inclusion in modelling financial volatility due to their capability of capturing the leverage effect, scenarios where negative shocks have a stronger impact on volatility than positive shocks. Two models are applied in implementation: the GJR-GARCH model and the TGARCH model, which both cater to asymmetry. The performance of the proposed method is evaluated using real data as a test period. The forecasts are run for 4, 11, and 36 dynamic principal components to assess the importance of dimension reduction vis-a-vis the predictive ability of the modelling approach. The predictive ability of the models is assessed, and models are compared by applying cross-validation. The results show that catering for asymmetric characteristics does not provide the improved predictive ability to perform markedly compared to the standard Bayesian symmetric GARCH model. Furthermore, computational efficiency is also an issue. It is therefore concluded that the extra cost needed in estimating the added parameters required to capture any asymmetric effects was not found to be worthwhile.

Promoting Protective Factors in Mental Health: A Meta-Analysis

by

Joeline Camilleri

Abstract

Meta-analysis is a research method utilized for synthesizing research results in various fields. The word 'meta-analysis' was introduced by Gene Glass to denote "the statistical analysis of a large collection of analytic results from individual studies for the purpose of integrating the findings" (Glass 1976). The primary motivation for carrying out a meta-analysis is to improve estimates due to an increase in the number of observations and improved statistical power. This is achieved by synthesizing effect sizes from several studies into a single metric, making it a strong tool for drawing inferences from already existing literature. To carry out a successful meta-analysis, a comprehensive procedure is necessary. This involves careful selection of studies and appropriate selection of effect sizes and models.

The application of the meta-analysis investigates the impact of protective factors, namely resilience and social emotional learning (SEL) in enhancing prosocial behaviour and reducing internalizing/externalising problems. In the application, random effects models were employed because they assume that studies were conducted on distinct populations where the restricted maximum likelihood was the estimation method employed. In the analysis, forest plots, Galbraith plots and funnel plots are generated to investigate heterogeneity between the studies, while the Egger's test and the Trim and Fill analysis are carried out to investigate publication bias. Moreover, meta-regression models are fitted to investigate how resilience affects the connection between social-emotional learning and mental health. The study finds that resilience and high levels of social emotional learning abilities are crucial to alleviate internalizing/externalizing problems, implying partial mediation. On the other hand, resilience has no mediating effect on the positive relationship between social emotional learning and prosocial behaviour.

Bayesian Parameter Estimation of the Hurst Index of Fractional Brownian Motion

by

David Paul Dimech

Abstract

One of the many generalizations of Brownian Motion, Fractional Brownian Motion is very popular due to being able to account for a wide range of phenomena in various different fields ranging from finance when modelling stock data to hydrology in water turbidity analysis. Brownian Motion is unsuitable for modelling these due to the assumption of independence of increments, an assumption relaxed by Fractional Brownian Motion given it allows dependence of increments. This dependence is affected through the Hurst Index H, the parameter associated with the process. For values of H between 0 and 0.5, both excluded, negative autocorrelation between the increments is enforced while a positive one is obtained for values between 0.5 and 1, both excluded. As H approaches 0.5, the paths of the process will resemble those given by Brownian Motion and if equality holds, the process reduces to a Brownian Motion Process. In this thesis, the theory behind Fractional Brownian Motion as well as the Bayesian framework in the context of estimating H shall be discussed. Given the subjectivity involved in determining a prior, sensitivity analysis shall be performed as to analyze the effect of different priors on the posterior. Following this, the Hurst Index of real data shall be estimated. The data considered shall be genetic data relating to Covid-19 and cardiology data relating to heart rate variation.

Rooting Out Deception: The Application of Tree-Based Learners for Motor Insurance Fraud Detection

by

Lorin Grima

Abstract

This dissertation investigates motor insurance fraud detection in the Spanish market by implementing and comparing tree-based methods, renowned for their performance and interpretability. The study begins with basic Decision Trees and progresses to tree-based ensemble methods, including Random Forests, Gradient Boosting machines, and Newton-based boosting techniques such as LightGBM, XGBoost, and CatBoost. A significant challenge in motor insurance fraud detection is addressing the class imbalance. To address this issue, the dissertation evaluates cost-sensitive learning approaches and resampling techniques to optimize model performance. The analysis concludes that a cost-sensitive LightGBM model is the most effective method for this scenario, achieving a balanced accuracy of 81% and successfully identifying 83% of fraudulent cases. The findings of this study provide valuable insights into the effectiveness of tree-based methods in detecting motor insurance fraud and highlight the potential of LightGBM in efficiently identifying fraudulent cases. By presenting a rigorous comparison of different techniques and addressing the class imbalance issue, this research contributes to the ongoing development of interpretable solutions for combating insurance fraud.

A Bayesian Approach to Modelling Individual Attacking Ability in Football Using Euro2020 Data

by

Sarah Micallef

Abstract

Bayesian hierarchical models have been widely used to model team data in sports and predict final match scores. In the literature, individual player statistics have been given less attention. By focusing on Euro2020 data, more specifically data on players in the attacking position, Bayesian hierarchical models are used in this dissertation to model individual player ability in football. Two models one with a Poisson likelihood and another with a Negative Binomial likelihood are applied to the data in order to provide multiple solutions to this problem. This variety allows us to choose the best fitting model through the use of a Bayesian version of the Root Mean Square Error. The best fitting model is then used to analyse the results. All models are implemented in JAGS through R software using the package *rjags*. Due to the fact that we do not have a mathematical formulation of the posterior distribution, MCMC chains are simulated for each parameter. A k-means cluster analysis is then implemented to group the attackers according to the simulated values from the posterior distribution of each individual parameter. The clusters are then analysed using means from the raw data to be able to compare the simulated data with the raw data and therefore to show that the chosen models provided a good estimation for player ability. Finally, the use of variational inference is suggested as an alternative to MCMC sampling in future studies.