Department of Statistical Sciences (DoSS)
2024 Summer Undergraduate Research

The Department of Statistical Sciences is seeking applicants interested in conducting a summer research project under faculty supervision. Projects will take place for a 16 week period from May-Aug, 2024. The following research awards are being offered this summer: NSERC USRAs ($7,500), UTEA ($7,500), DoSS SURAs ($2000) and the Black and Indigenous Student Undergraduate Student Research Awards ($7,500). In addition, Students also interested in doing a STA496H1/STA497H1 Independent Reading course during the Summer term should fill out this form as well.

All students who are currently enrolled in a Department of Statistical Sciences (DoSS) Specialist or Major program and who have completed STA302H1 with at least a B+ by the time of application are eligible to be considered for one of these awards. You will be asked to rank your top 3 projects. Students will be considered based on their academic record, experience and their statement of research interest. Students who are shortlisted for specific projects may be invited for an interview with the prospective supervisor.

Projects

Aspects of Robust Regression Analysis

Supervised by: Nancy Reid

This project will consider comparison of Bayesian and frequentist methods on robust regression. It will involve a mix of theoretical analysis and simulations, initially in fixed-dimension regression models, and then in high-dimensional regression with regularization. Comparing Bayesian and frequentist approaches is of interest both for the foundations of inference and for the practical assessment of the reliability of Bayesian approaches; the latter is closely related to the asymptotic theory of likelihood-based inference. Robust regression methods are an important technique for ensuring that statistical conclusions remain valid even when the model used for inference differs from the model generating the data.

Extending Spatial Models for the Study of Shark Aggregations

Supervised by: Vianey Leos Barajas

Ecological statistics is a rapidly growing interdisciplinary field. Statisticians work in interdisciplinary settings with ecologists to develop novel statistical models for common ecological data structures. Within this field lies the study of animal movement. Typical types of data collected for the study of animal movement are GPS, i.e. positional data, and accelerometer data. However, as technology evolves so does the type of data that can be collected. For the study of shark aggregations, researchers are now collecting data using drones to capture aerial footage. From this we can extract locations of sharks and also collect environmental variables that may drive their aggregation patterns. This type of spatial data structure is relatively new and spatial models have not yet been developed to model repeated aggregation observations.

Investigating Dependencies in Population Projection Models

Supervised by: Monica Alexander & Radu Craiu

Probabilistic projections for human populations are commonly obtained through the use of cohort component models, where the components of population change (fertility, mortality, and migration) are themselves estimated using time series models. Existing approaches assume the components of population change are independent, however, this is not the case in most populations.

This project will investigate the degree and nature of dependency in components of population change, and sensitivities to population projections to different assumptions about dependence. The student will use copulas to model dependency in UN population data across a wide range of countries and carry out a simulation study to assess the impact of different assumptions and models on resulting projections.

Genetic Heritability of "Coupling" Between Brain Structures and Functions

Supervised by: Jun Young Park

Technological advances in brain magnetic resonance imaging (MRI) allowed researchers to use non-invasive methods to understand brain structure and functions and develop novel research questions. Among them, individual differences in “coupling” across measures of brain structure and function may underlie differential risk for neuropsychiatric disorders, and research in this area has gained significant attention in neuroscience. While several approaches have emerged for quantifying intermodal coupling at the individual level and testing its existence at the group level, it has yet to be determined whether these intermodal coupling are regulated by genetic factors (i.e., “heritable”). Understanding the genetic underpinnings of coupling, if they exist, would provide invaluable biomarkers for brain-phenotype associations.

Several methodological issues must be considered to evaluate its possibility carefully, which is the goal of the summer research. These include (i) high-dimensionality of brain MRI data, (ii) low signal-to-noise ratio, and (iii) a relatively small number of samples, all of which would lead to an underpowered study. Therefore, we will study how multivariate (spatial-extent) modelling and inference would help overcome the limitations. During the summer project, students will gain experience in (i) exploratory data analysis of real brain imaging data, (ii) methods development, and (iii) software implementation. Depending on the research progress, the resulting research outputs would be submitted to a peer-reviewed journal for publication (although it is typically expected to take over three months).

It is expected that students meet 1-2 times each week with me to discuss progress and challenges. Those are welcome to contact me (junjy.park@utoronto.ca) if they want to discuss more details of the summer research project.

Deep Reinforcement Learning Algorithms for Portfolio Optimization and Risk Management

Supervised by: Xiaofei Shi

This project aims to develop novel deep reinforcement learning algorithms and compare with existing ones for portfolio optimization problem and risk management. In particular, with risk preferences such as expected shortfall and value-at-risk, closed-form solution are very limited and efficient numerical algorithms are in need.

How to minimize risks and maximize profits in a financial market are essential tasks for financial institutions such as investment banks, hedge funds, insurance and reinsurance companies. These problems are usually formulated mathematically as portfolio optimization and/or risk management problems. Since the financial market is usually a complicated system and the optimization problems are intrinsically high-dimensional, we cannot expect simple closed-form solution. Deep reinforcement learning algorithms, due to their capabilities to overcome curse-of-dimensionalities, may offer a class of numerical solutions to these portfolio optimization and risk management problems.

The project will involve a mix of theory and development of numerical algorithms, to fully utilize the power of deep reinforcement learning as a efficient numerical tool.

Preparing a Dataset of Biochemical Knowledge Bases for Large Language Model Training

Supervised by: Christopher Maddison

Pretraining on very large-scale data is one of the key factors in the state-of-the-art (SOTA) performance of large language models (LLMs) on many natural language processing tasks. However, when it comes to their performance on biochemical tasks, these general purpose models still lag far behind specialized predictors.

In this project, our aim is to develop a large-scale dataset of biochemical data, together with science texts. There are a number of design decisions that need to be made, including the sourcing of the data, the impact of tokenization, how to manage links, and data ordering. We will evaluate the quality of our data, as well as the impact of these design consideration, by evaluating the performance of models pre-trained on our data, compared to SOTA LLMs.

Is the Universe "Broken"? Or is it just the way we look at it?: Testing the Robustness of Bayesian Inference Frameworks in Cosmological Analyses

Supervised by: Joshua Speagle & Tanveer Karim

Modern astronomical surveys are collecting data on hundreds of millions of galaxies to measure the properties of the Universe at the largest (i.e. cosmological) scales. One of the main goals of

these efforts is to finally uncover the true nature of the many mysterious components that make up our Universe, including Dark Energy, Dark Matter, and neutrinos (the smallest particles found in nature). To constrain various physics models, astronomers need to simultaneously infer the properties of multiple parameters of interest as well as a large number of nuisance parameters. This is often done under a Bayesian framework and relies on several (strong) assumptions to make claims of discovery or to test which model of cosmology best explains the data. Although many of these assumptions seem well-motivated, the extent to which these assumptions can be safely trusted is unclear.

In this project, the student will explore two interconnected research areas. The first involves developing methods to better understand how observed discrepancies between a few parameters of interest from different datasets generalize to high-dimensional spaces. The second involves exploring the robustness of various model comparison strategies when the desired parameters are close to the edge of the parameter space; this latter problem is highly relevant to the problem of estimating the sum of neutrino masses.

Active Learning Strategies for the Euclid Space Telescope

Supervised by: Joshua Speagle & Michael Walmsley

Euclid is a $500M USD space telescope that has just (Feb 14!) started operations and aims to capture the first images of hundreds of millions of galaxies at a time when the Universe was only a few billion years old. UofT researchers are providing deep learning models to measure the appearance of these galaxies (e.g. counting spiral arms) based on images of galaxies from other telescopes labelled by 100k+ volunteers. Given the massive increase in data volume in Euclid, future volunteers will only be able to provide high-quality characterizations for a tiny fraction of these galaxies. Which galaxies should these be?

This project will explore various active learning strategies to identify which will work best for Euclid and under what conditions (supercomputer access and state-of-the-art models will be provided). It will also explore the potential consequences of these strategies on expected model performance, uncertainty quantification, and robustness to domain shifts and rare events. These efforts may also involve collecting labels on new Euclid galaxies -- galaxies which the student would likely be the first person to see.

Using Spatial Modeling to Examine Patterns of Brain Metastasis

Supervised by: Meredith Franklin

The spatial distributions of brain metastasis are hypothesized to vary according to primary cancer subtype, but an understanding of these patterns remains poorly understood despite having major implications for treatment. Through this project we hope to elucidate the topographic patterns of brain metastases for 5 different primary cancers (melanoma, lung, breast, renal, and colorectal), which may be indicative of the abilities of various cancers to adapt to regional neural microenvironments, facilitate colonization, and establish metastasis. Our findings could be used as a predictive diagnostic tool and for therapeutic treatments to disrupt growth of brain metastasis on the basis of anatomical region.

To test our hypothesis that brain metastases have different spatial patterns depending on the primary cancer type, we will leverage 3D coordinates of brain metastases derived from stereostatic radiosurgery procedures in over 2100 patients. With these data we will explore two types of spatial models: one where the X, Y, Z spatial coordinates of the metastases are compared between the 5 different primary cancer types, and another where we compare the spatial coordinates of the metastases from each cancer type separately to spatially random processes on a sphere. Both approaches will use flexible generalized additive models. However, in the latter approach, methods will be developed to generate random spatial Poisson point processes in three dimensions.

Machine Learning for Real-Time Satellite Remote Sensing Observations of Aerosols

Supervised by: Meredith Franklin

Exposure to particulate matter (PM) air pollution has been associated with a myriad of adverse health outcomes, yet the relative toxicity of PM mixtures with different sizes, shapes, and chemical compositions is poorly understood. This research will help future satellite missions to be better equipped to understand aerosol particle type and its role on human health.

Using hourly data collected over the past 2 years by multiple co-located instruments at several locations in California and New York, we will explore how to predict PM properties differentiated by size and chemical composition from aerosol optical depth properties (as measured through remote sensing). Given the high dimensionality of the measured aerosol parameters, we will leverage machine learning techniques such as XGBoost with SHAP to understand what variables are important in predicting PM. Furthermore, we will incorporate temporal information to explicitly model autocorrelations in the time series data.

This work is in collaboration with the NASA and the Jet Propulsion Laboratory.

Student Eligibility:

1. Must be an undergraduate student currently enrolled in a Department of Statistical Sciences Specialist or Major program.
2. Currently registered full-time or part-time student at the time of application.
3. Must have completed STA302 with at least a B+ grade.

How to submit your application:

Please fill out and submit the following application form:
Application for DoSS Summer Undergraduate Research Awards 2024

If you have any questions regarding these awards, please contact ug.statistics@utstat.utoronto.ca

Completed applications are due by 11:59PM EST Thursday, March 14, 2024