Undergraduate Research Program in Statistics (URPS)

Overview

- Is this program for me?
- The application process
- The projects
- What to expect if I join a project
- Do I get course credit?

  Can I use the project for an honors thesis?

  For the Data Science major capstone requirement?
- Other opportunities for undergraduate research in statistics
- Other questions

Project Descriptions

Network Community Detection Using Higher-Order Structures

Description: Many real-world networks commonly exhibit an abundance of subgraphs or higher-order structures, such as triangles and by-fans, surpassing what is typically observed in randomly generated networks. However, statistical models accounting for this phenomenon are limited, especially when community structure is of interest. This limitation is coupled with a lack of community detection methods that leverage subgraphs or higher-order structures. To address these gaps, we have developed a new community detection method that effectively incorporates these higher-order structures within a network. In this project, the undergraduate students are expected to implement the new algorithm, conduct simulation studies, and compare the performance of our proposed method with some existing methods that disregard higher-order structures. Attention to details, familiarity with linear algebra, and proficiency in programming languages, particularly R and Python, are essential.

Supervisors: Ji Zhu (faculty)

A Latent Space Model for Hypergraphs with Diversity and Heterogeneous Popularity

Description: While relations among individuals make an important part of data with scientific and business interests, existing statistical modeling of relational data has mainly been focusing on dyadic relations, i.e., those between two individuals. This project addresses the less studied, though commonly encountered, polyadic relations that can involve more than two individuals. In particular, we have developed a new latent space model for hypergraphs using determinantal point processes, which is driven by the diversity within hyperedges and each node's popularity. The undergraduate students in this project are expected to implement an algorithm for estimating the model parameters, conduct simulation studies, and evaluate the performance of the proposed algorithm. Attention to details, familiarity with linear algebra, and proficiency in programming languages, particularly R and Python, are required.

Supervisors: Ji Zhu (faculty)

Personalized Prediction Using Mobile Health Data

Increased usage of smartphones and wearable devices has led to a proliferation of mobile health technologies that seek to personalize healthcare. Prior studies have shown that data collected from wearable devices and apps can be used to predict clinical mental health outcomes such as depression with high performance. This project will explore machine learning methods that can be used for personalized, early prediction of clinical outcomes related to mental health. The student will have the opportunity to work with a real mobile health data set and build a variety of machine learning prediction models, including random forests, neural networks, and recurrent neural networks. The student should have extensive experience with machine learning methods (such as those taught in EECS 445 or an equivalent course) and should be proficient in Python coding with PyTorch or TensorFlow. Depending on the student's interests, the project could also explore cluster-based prediction methods or imputation methods for missing data in more detail.

Prerequisites: EECS 445 or equivalent experience with machine learning, Proficient in PyTorch or TensorFlow

Supervisors: Ambuj Tewari (faculty), Luke Francisco (PhD student)

Enhancing Education Policy Evaluations by Modeling Student-Level Outcomes

Program and policy evaluations in the field of education benefit greatly from modeling the relationship between student-level outcomes and additional factors such as prior student achievement, student demographics, and school-level characteristics. Correct specification of such a model improves the precision of impact estimates in randomized trials and reduces bias in observational studies, giving researchers a clearer picture of program efficacy. We seek an undergraduate researcher to demonstrate the benefits of applying this procedure, known as covariance adjustment, to a policy recently implemented across a range of school districts in Texas. This researcher would fit a host of models in R to data we currently have available--and possibly additional datasets that have a theoretical basis for improving model fit--to show education researchers how they may improve existing covariance adjustment models to achieve better inference for program impacts. One can expect to learn modeling techniques not yet covered in core undergraduate courses, enhance their ability to assess the fit of complex regressions, and inform future evaluations in education research. This project requires completion of STATS 306 or 406, STATS 413, and a readiness to build on the knowledge gained in those courses.

Supervisors: Ben Hansen (faculty), Josh Wasserman (PhD student)

Frechet Bounds for Partially Observed Distributions

A variety of applications ranging from mathematical finance to biomedicine present the feature that one is interested in learning about the distribution of a random vector W (or certain functionals thereof), but that samples from the joint distribution of W are unavailable. As a simple example, one can consider the triplet W = (Y,X,Z) where Y is the true label, X is the set of features, and Z is a `weak' or 'proxy' label for Y. Consider a situation where one has samples from (Y,Z) and (X,Z) but none from (Y,X). One is interested in learning E(g(X,Y)] for some function g: e.g. $g(x,y) = 1(y \neq \psi(x))$, for some pre-trained classifier ψ. However, the joint distribution is not fully identified by the observed marginals in which case the problem becomes one of quantifying the range of E(g(X,Y)) (infimum and supremum, for example) over the class of joint distributions compatible with the observed marginals, and also ascribing confidence bounds to them.

This class of problems can be formulated in terms of optimal transport problems and extensions thereof. The proposed project will develop a variety of algorithms for computing estimates of these limits and investigate the asymptotic distributions of such estimates. Yue Yu, an extremely strong final year undergraduate student, will be working primarily on this project. He is adept at both the theoretical and computational tools needed. Subha Maity will be his graduate student supervisor. I will be the faculty supervisor.

Supervisors: Mouli Banerjee (faculty)

Expectile factor analysis for circadian heart rate analysis

This project will consider expectile analogs of classical factor analysis, and use them to understand inter-day and inter-individual variation in circadian heart rates. Expectile factor analysis identifies a low dimensional representation of a collection of high dimensional trajectories that reflect well-defined characteristics of the circadian heart rate probability distribution. One such characteristic is the mean, which coincides with traditional factor analysis, but other expectiles capture the tails of the distribution. Arguably outer expectiles are especially relevant for physiological parameters such as heart rate. Using data from wearable devices such as smart watches, we can assess whether distinct factor structures exist for the outer expectiles compared to the mean, and assess how these factor structures contribute to inter-day and inter-individual variation.

Supervisors: Kerby Shedden (faculty), Osafo Agyare (PhD student)

Causal Effect Estimation for Clustered and Blocked RCTs

Randomized controlled trials (RCTs) are used to evaluate treatment effects. When individuals are grouped, such as students in classrooms or schools, clustered RCTs are commonly conducted. The availability of baseline covariates, related to outcomes, prompts the use of blocking to mitigate covariate imbalance between treatment groups. We are interested in evaluating causal effect estimators in clustered and blocked RCTs. We propose a two-step procedure augmenting the Hajek estimator with a preliminary fit of a covariance adjustment model. A related method discussed in recent literature combines elements of Hajek estimation with least squares covariance adjustment in a single step. Several hypotheses about the advantages of the two-step procedure over the competing method remain to be explored.

We will compare these methods in various settings through simulation studies and data analysis to identify scenarios where one method dominates the other. The tentative plan involves exploring data from an RCT evaluating the effect of Minecraft Education on spatial thinking among 5th and 6th graders. This RCT randomized classrooms to intervention and blocked classrooms on grade. For this project, basic knowledge of R is required.

Supervisors: Prof. Ben Hansen (faculty), Xinhe Wang (PhD student)

Selective Inference for Time-Varying Effect Moderation

The scientific community is increasingly interested in developing data analysis techniques that can improve mobile health interventions. A key aspect of this effort involves assessing the impact of time-varying causal effect moderators. Effect modification, a scenario where the impact of treatment on outcomes varies based on other covariates, plays a significant role in decision-making processes.

When there are hundreds or thousands of covariates, it becomes necessary to use observed data to select a simpler model for effect modification and make valid statistical inferences. To achieve this, the Lasso method is used to select a model with lower complexity for effect modification. The selected model is much more interpretable compared to a full model consisting of all covariates. To ensure that our models have valid post-selection inferences, we construct a valid pivot that is asymptotically distributed as a uniform random variable.

In this project, the students will start by studying basic example models such as the weighted centered least squares (WCLS) models, and then move on to evaluating the use of selective inference in other models with time-varying moderators and applying it to real datasets. We expect the student to be familiar with Regression (Stats 413 or Stats 415 or equivalent) and Python.

Supervisors: Snigdha Panigrahi (faculty), Soham Bakshi (PhD Student)

Post-Selection Inference for Smoothed Quantile Regression

Quantile regression is a powerful technique for estimating the conditional quantiles of a response variable, which provides robust estimates for heavy-tailed responses or outliers without assuming a specific parametric distribution. However, modeling conditional quantiles in high-dimensional data and studying post-selection inferences for the quantile effects of selected covariates has computational and efficiency challenges.

The computational challenges arise because of the non-differentiable quantile loss function, while the efficiency challenges arise due to the data discarded during model selection. To address these challenges, we have developed a new approach for post-selection inference after modeling conditional quantiles with smoothed quantile regression. Our approach is fast to compute, and it no longer discards data during model selection, circumventing the disadvantages of a non-differentiable loss.

This project is an opportunity for students to learn about quantile regression and perform a wide range of numerical experiments to validate this approach. They will also explore popular model selection methods such as LASSO, SCAD, and MCP. Familiarity with regression (Stats 413 or Stats 415 or equivalent) and Python is expected.

Supervisors: Snigdha Panigrahi (faculty), Yumeng Wang (PhD Student)

Selective Inference for Gaussian Graphical Models

Abstract. Extensive work has been done on estimating an undirected graphical model from p jointly normally distributed random variables. The nodes of the graph represent these variables, while the edges between the nodes capture their conditional dependence relationships. These dependence relationships are characterized by the nonzero entries of the inverse covariance matrix, also known as the precision matrix. This model is known as the Gaussian Graphical Model (GGM).

A popular and fast approach for estimating the nonzero entries of the precision matrix in the GGM is to estimate neighborhoods of each node in the graph through multiple regression. The nodewise regression coefficients are then combined to estimate the edge structure in the graph. However, selective inference is required when making inferences in the GGM whose edge structure is data dependent, as the same dataset cannot be used twice for estimating the edge structure and inferring for the associated parameters.

In this project, students will gain knowledge about the Gaussian graphical model (GGM), selective inference techniques, and develop test-beds to evaluate novel selective inference methodologies for the GGM across various simulation setups. Prerequisites are familiarity with Python and an upper-level course covering linear regression and lasso model selection (Stats 413 or Stats 415 or equivalent).

Supervisors: Snigdha Panigrahi (faculty), Yiling Huang (PhD student)

Spatial Modeling of the SARS-CoV-2 Epidemic in Michigan

The mainstream of statistical epidemiology focuses on modeling the spread of disease in a homogeneous population over time, but it can be difficult to reason about varying transmission risk factors within a population using this approach. Spatial and hierarchical analysis has opened the door to asking more fine-grained questions about the extent and drivers of spatial variation in infection outcomes. In the proposed project, we are seeking a student to collaborate with us while we consider the problem of spatiotemporal modeling of the initial spread of SARS-CoV-2 in Michigan. Specifically, we hope to dig deeper into the ways in which residential and occupation segregation by class, race, and other social factors informed the dynamics of infection in the early period of the SARS-CoV-2 pandemic (i.e. March-Dec 2020). The proposed project would involve utilizing data including death certificates, infection records, and cellphone mobility data to identify geographic patterning of infection and highlight potential causes. We are looking for an undergraduate researcher interested in public health and spatial statistics to help with data analysis, model implementation, and visualization. In future work, we hope to use machine learning techniques (e.g. convolutional neural networks) to learn spatial correlations from large-scale health data and perform Bayesian inference at scale.

Supervisors: Prof. Jon Zelner (faculty, UM Epidemiology), Prayag Chatha (PhD student)

Mapping dark matter using astronomical images

Dark matter between Earth and distant galaxies acts as a “gravitational lens,” distorting the appearance of these galaxies (see figure). By analyzing images of galaxies, we can learn about the spatial distribution of dark matter throughout the universe. In this project, we take a probabilistic approach to mapping dark matter: given the images (observed random variables), we infer the distribution of dark matter (latent random variables) under a scientifically plausible generative model. To perform inference, we use a new technique called neural posterior estimation, which involves simulating many astronomical images with various dark matter distributions, and then training a convolutional neural network to predict the location of dark matter for each image. Undergraduate researchers will help develop the image simulator (making use of existing software), implement several varieties of neural posterior estimation, and apply them to real astronomical images. No prior knowledge of astronomy is expected. Familiarity with Bayesian statistics is helpful but not essential. Strong computational skills are required.

Supervisors: Jeffrey Regier (faculty), Tim White (PhD student)

Estimating the distances to galaxies in astronomical images

In astronomical images, the color of a galaxy indicates its distance from Earth; the more distant a galaxy is, the faster it is moving away from Earth, and therefore the redder it appears (see figure). The apparent colors of galaxies also vary due to factors other than distance, such as mass and molecular composition, and measurement error. Hence, there is ambiguity about galaxies' distances. In this project, we take a probabilistic approach to estimating galaxies' distances: given an image and a generative statistical model, we infer the distribution of distance for each imaged galaxy. To perform inference, we use a new technique called neural posterior estimation, which involves simulating many images of galaxies at various distances, and then training a convnet to predict the distance to each. Our work will focus on a particularly difficult case, which existing approaches are not equipped to handle: galaxies that overlap visually with other galaxies. Undergraduate researchers will help design and implement the image simulator, implement neural posterior estimation (making use of existing software), and apply the method to real astronomical images. No prior knowledge of astronomy is expected. Familiarity with Bayesian statistics is helpful but not essential. Strong computational skills are required.

Supervisors: Jeffrey Regier (faculty), Declan McNamara (PhD student)

Finding galaxy clusters in astronomical images

Galaxy clusters, which are made up of hundreds or thousands of nearby galaxies, are the largest structures in the universe. Yet finding them in astronomical images is challenging: galaxies often appear close together in 2d images without being part of a common cluster. In this project, we take a probabilistic approach to finding and characterizing galaxy clusters. To perform posterior inference, we'll use a new technique called neural posterior estimation, which involves simulating images with and without galaxy clusters, and training a convolutional neural network (CNN) to predict whether each galaxy is part of a cluster and if so, to predict the cluster's properties. Undergraduate researchers will 1) use Python to implement and compare several galaxy cluster simulators, 2) use PyTorch to train a neural network to predict the locations and masses of galaxy clusters in simulated images, and 3) apply the trained neural network to real astronomical images. No prior knowledge of astronomy is expected. Familiarity with Bayesian statistics is helpful but not essential. Strong computational skills are required.

Supervisors: Jeffrey Regier (faculty), Gabriel Alfonso Patrón Herrera (PhD student)

Statistical Modeling for Intensive Care Data

The aim is to uncover the complex yet crucial relationships among symptoms and procedures in critical care units using statistical embedding approaches. The primary data source for this study is the MIMIC-III (Medical Information Mart for Intensive Care) data, which contains comprehensive clinical information of patients admitted to critical care units at a large tertiary care hospital, such as vital signs, medications, laboratory measurements, observations and notes charted by care providers. This project would explore different embedding methods to generate reliable and interpretable numerical representation for each diagnosis and procedure code. These representations will not only serve as distinctive features of the symptoms and procedures, but also highlight interrelations among them. Potential applications include predicting patient symptoms, tracing causes and consequences of human disease and death, aiding the planning of service and contributing to health services research, among others. Students participating in this project will be expected to read assigned literature, perform exploratory data analysis, try out different embedding approaches, and compile coding reports. Attention to details and proficiency in programming languages, particularly R and Python, are mandatory. Familiarity with algebra and statistical learning is preferred.

Supervisors: Gongjun Xu, Ji Zhu (faculty), Shihao Wu (Ph.D. student)

Statistical Sketches of the Sun

Solar Flares are large bursts of electromagnetic radiation that are released from the Sun. The extreme solar flares can disrupt telecommunications on the Earth and in fact destroy the electronic equipment of satellites in space. The goal of the project is to study electromagnetic images of the Sun obtained from the Solar Dynamics Observatory (SDO) instruments https://sdo.gsfc.nasa.gov/mission/. The idea is to build informative covariates that can help predict extreme solar flares. Using the python package sunpy, we will download images of the Sun in the electromagnetic spectrum and implement 2d wavelet transform methods as well as various random sketching techniques such as Gaussian, sum- and max-stable sketches. The goal is to find lower-dimensional representations of these large data sets that retain features, which can help predict solar flares. The project will involve some coding in python and R shiny, learning and implementing statistical methods and algorithms.

Supervisors: Stilian Stoev (faculty), Victor Verma (PhD student)

Mapping broadband availability in Michigan

The goal of the project is to combine different types of network traffic measurement data with spatial and demographic covariates in order to produce a map of broadband availability in the State of Michigan. We have access to limited but highly accurate network quality-of-service (QoS) statistics. At the same time, there is a larger set of freely available Ookla speed-test measurement data, which are observational in their nature and involve intricate biases and confounding. The first stage of the project will involve exploratory data analysis that will study the relationship between the Ookla data and the accurate QoS measurements, when available. The second stage of the project will involve building a model that can be used to estimate broadband availability as a function of covariates in areas where no accurate QoS measurements are available. The project will involve building R shiny apps, designing and implementing statistical algorithms and collaboration with researchers from Merit Network.

Supervisors: Stilian Stoev (faculty), Moritz Korte-Stapff (PhD student)

Investigating language models' abilities to encode geographical information

Pre-trained language models have been shown to encode geographical information, such as the latitudes and longitudes of places. Nevertheless, the precise mechanisms behind the acquisition and storage of this geographical knowledge within these extensive models remain relatively unexplored. This project's primary objective is to formulate methodologies that facilitate a deeper understanding of this intriguing phenomenon. An initial approach could involve reducing data contamination by focusing on locations that are highly likely to be present in the models' training data. Subsequently, the project may progress to the identification of a functional circuit capable of learning and retaining this geographical information.

Prerequisites: Basic knowledge of machine learning (at the level of STATS 415), Python (at the level of STATS 206), and a deep learning library (at the level of STATS 315). Basic knowledge of natural language processing (NLP) is a plus.

Supervisors: Yixin Wang (faculty), Kevin Wibisono (PhD student)

The Psychology of ChatGPT

The field of psychology has long sought to better understand how humans think, often by employing ingenious experiments to explore cognitive biases, susceptibility to logical fallicies, the stability of preferences, moral reasoning, etc. Many of these classical psychology experiments can be naturally adapted and applied to ChatGPT, to potentially better understand "how ChatGPT thinks." The goal of this project is to do just that. The primary statistical content of the project will lie in the design and analysis of the experiments -- and especially the design. (No knowledge of large language models or NLP is needed.)

Supervisors: Johann Gagnon-Bartsch (faculty), Jaylin Lowe (PhD student)

Testing for network differences in neuroimaging data

This project is on applying methodology and software recently developed in our group to a neuroimaging dataset from collaborators in Psychiatry. The methodology is for testing for statistically significant differences between two groups of networks on any given part of the network, which gets power from implicitly using the information in the entire network even when testing for differences in a part of it. The neuroimaging dataset was collected to compare the effects of a particular drug tested on three groups of patients (those with schizophrenia, bipolar disorder, and healthy controls). The scientific hypothesis, supported by pilot data, is that while the particular region of the brain known to be affected by this drug is strongly and similarly affected in all patient groups, there are differences between groups in how this region then interacts with other parts of the brain.

Prerequisites: Familiarity with R (at the level of Stats 306), familiarity with statistical testing (ideally but not necessarily at the level of Stat 426).

Supervisors: Liza Levina (faculty)