Undergraduate Research Program in Statistics (URPS)

About URPS

- Overview of URPS
- The application process
- FAQs
    - Course credit
    - Can I use the project for an honors thesis?
    - For the Data Science major capstone requirement?
- Other questions?

Project Descriptions

Dynamic pricing for multiple firms under non-linear demand

Description: The problem of studying the dynamics of interaction between multiple firms competing in a market where products must be sold and each firm seeks to optimize its revenue is of canonical importance. Unlike recent works, all of which consider known underlying demand functions, our objective is to address cases where demand-price information is unknown a priori for every firm. We aim to estimate the demand function non-parametrically using shape constrained methods, where the (natural) constraint is that the demand for firm i decreases with respect to its own posted price and increases with respect to other firms' prices. After computing these estimates, we plan to analyze regret, identify sufficient assumptions to achieve the Nash equilibrium, study convergence rates, and apply our method to real-world datasets. The student is required to run simulations that validate theoretical analysis as well as delve into real data applications, and develop a broad understanding of dynamic pricing models.

Prerequisites: Background in Calculus, Linear Algebra, coding in python (or R) and some knowledge in Statistical Inference.

Faculty supervisor: Moulinath Banerjee PhD student supervisor: Daniele Bracale

Investigating how small transformers learn arithmetic

Description: Advanced large language models can implement complex reasoning tasks like coding and solving math problems. However, the mechanisms behind this emergent capability remain unexplored. This project's primary goal is to study a relatively simpler problem: how transformers learn arithmetic operations such as addition, multiplication. The first stage of the project will involve building a small transformer model and pretraining it from scratch. The second stage will involve some visualization of the pretrained model. The third stage may involve reverse engineering the model and using techniques like causal probing to study what algorithms has the model learned.

Prerequisites: 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.

Faculty supervisor: Yixin Wang PhD student supervisor: Zhiwei Xu

Martingale Posterior for Uncertianty Quantification

Description: In recent years, Generative AI has emerged as a transformative technology with applications across diverse fields, from creative content generation to scientific simulations. However, in high-stakes scenarios such as hazardous weather forecasting, medical diagnosis, and autonomous vehicles, predictions based on black-box algorithms can significantly impact downstream decision-making, raising concerns about safety and reliability. This project focuses on developing uncertainty quantification methods for generative models through a sequential predictive approach known as the Martingale posterior. Unlike traditional Bayesian methods, the Martingale posterior does not require a specified likelihood or prior; instead, it constructs a posterior predictive probability that adapts flexibly to new data. By offering uncertainty estimates that are both interpretable and theoretically grounded, this approach aims to enhance trust and accountability in AI-driven decisions, supporting the safer deployment of generative models in critical applications. In this project, the student will develop a deep understanding of the Martingale posterior and implement the algorithm in PyTorch. They will compare it with existing uncertainty quantification methods, such as conformal prediction and Bayesian posterior approaches to fundamental statistical models, and potentially extend the comparison to more complex generative models like diffusion models.

Prerequisites: Experience with Pytorch Familiar with measure theoretic probabilities - has completed STATS 412, STATS/MATH 425, ideally also STATS 426 and STATS 430 Familiar with Bayesian statistics

Faculty supervisor: Yixin Wang PhD student supervisor: Yidan Xu

Different questions in the interaction between impression and reality in sport statistics

Description: We will investigate various questions in the interaction between impression and reality in sport statistics. For example: Are the perceived players in baseball actually better? (The best prediction of a single hitter performance in the 2nd half of season is the average at bat statistics of all other players) . Is there a momentum? Is there correlation between two halves of a game (there are good reasons to believe that there is none).

Faculty supervisor: Ya'acov Ritov

Deep learning for Intensive Care Data

Description: The aim is to uncover the complex yet crucial relationships among symptoms and procedures in critical care units using deep learning methods. 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 deep learning architectures to generate numerical representations for each diagnosis and procedure code. These representations will not only serve as distinctive features of the symptoms and procedures, highlighting interrelations among them, but also are valuable for a wide range of downstream tasks. 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 deep learning architectures, and compile coding reports. Attention to details and proficiency in programming languages, particularly Python, are mandatory. Familiarity with linear algebra and statistical learning is preferred.

Prerequisites: proficiency in programming languages, particularly Python

Faculty supervisor: Ji Zhu, Gongjun Xu PhD student supervisor: Shihao Wu

Journal ranking analysis based on citation exchanges

Description: The aim is to conduct a comprehensive ranking analysis based on citation exchanges (1) for journals within a certain domain, such as statistics, and (2) for journals across multiple domains. Journal rankings play a crucial role in academic evaluations, as they influence funding decisions, publication strategies, and tenure assessments. The most commonly used metric, the journal impact factor (JIF), measures the average number of citations received by a certain journal. Although popular and straightforward, the rankings based on the journal impact factors do not always align with the views of domain experts. One reason is that the journal impact factor fails to distinguish journals from different domains as it treats all citations equally. Additionally, journal impact factors ignore mutual citation patterns between journals, which could provide valuable insights. A more comprehensive and fair ranking system is needed to better represent a journal's impact and relevance. Students participating in this project will be expected to read assigned literature, collect citation data on the web, implement various algorithms, and compile coding reports. Attention to details and proficiency in programming languages, particularly R or Python, are mandatory. Familiarity with linear algebra and statistical learning is preferred.

Prerequisites: proficiency in programming languages, particularly R or Python

Faculty supervisor: Ji Zhu, Gongjun Xu PhD student supervisor: Shihao Wu

Building Analytical Tools for Wearable Heart Rate Data

Description: The project will focus on leveraging wearable device data, specifically heart rate measurements from devices such as Fitbit and Apple Watch, as well as night-specific data from sleep studies. Students will explore the transferability of metrics developed for continuous glucose monitoring (CGM) data, previously summarized in the R package iglu, to heart rate data. The project involves identifying publicly available heart rate datasets, testing the functionality of iglu with this new data type, and ultimately developing a new R package tailored for heart rate analysis. Ideal candidates should have an interest in digital health, proficiency in data wrangling and R, and strong teamwork skills. Students will gain hands-on experience working with wearable data and using R, Git/GitHub, and potentially Python, while actively contributing to a collaborative lab environment focused on digital health analytics.

Prerequisites: R proficiency, preferably knowledge of Git/Github and potential python knowledge.

Faculty supervisor: Irina Gaynanova

The guided intermediate resampling filter

Description: The guided intermediate resampling filter (GIRF) is a recently proposed particle filtering method for statistical inference on partially observed spatiotemporal dynamic systems. The most widely used implementation is in the R package spatPomp. However, the spatPomp implementation underperforms the results in the original paper on a benchmark spatiotemporal data analysis of measles transmission. The task for this project is to improve the spatPomp implementation to close this gap. Further, a new guide function has been implemented in spatPomp::girf which is worth further investigation. You are well suited for this project if you are interested in statistical inference for stochastic dynamic systems, and you are willing to learn how to delve deep into a complex R package.

Prerequisites: Willingness to learn about advanced statistical computing and stochastic processes.

Faculty supervisor: Edward Ionides

Vector embeddings for irregular longitudinal data

Description: Longitudinal data collected at irregular time points are very common, especially in research involving human subjects. While this can be treated as a missing data problem, an alternative is to embed the observed longitudinal data into a vector space so that standard methods for multivariate data analysis can be applied. This project will consider approaches based on U statistics for constructing such embeddings, and evaluate the methods using data from several longitudinal studies of humans. Time permitting we will extend the approach to the setting of mediation analysis.

Prerequisites: Linear algebra (214, 217 preferred) is essential. Strong preference for completed or at least concurrent 413 and 415.

Faculty supervisor: Kerby Shedden PhD student supervisor: Ben Osafo Agyare

Understanding ecological response to global warming through data analysis

Description: As global temperatures continue to rise and droughts become more frequent due to climate change, understanding the effects of these shifts on plant life is increasingly critical. This project will focus on examining the impact of temperature treatments on trees in Minnesota with the possibility of expanding to drought treatments. The specific emphasis of this work will be quantifying the causal effects of multilevel temperature interventions on a functional outcome in a controlled experimental setting. By employing a rigorous methodological framework, informed by the Causal Inference with a Functional Outcome paper, this study aims to leverage advanced statistical techniques to identify and analyze the intricate relationships between climate variables and tree responses. The paper's theoretical insights and code will serve as a foundational guide, shaping our analytical approach. By leveraging these resources, we aim to produce robust findings on the causal effects of temperature treatments, contributing to a deeper understanding of ecological responses to climate change.

Prerequisites: Coding experience particularly in R and a willingness to get a crash course on related ecology

Faculty supervisor: Yang Chen PhD student supervisor: Ashlan Simpson

Misspecification Behaviors of Mixture Modeling

Description: Mixture modeling has emerged as a crucial statistical method for data clustering and pattern recognition, but they are often misspecified in practice, including the choice of kernel densities used to model the data. Under sufficient regularity, when the Kullback-Leibler (KL) minimizer is unique, the Maximum Likelihood Estimator (MLE) of a misspecified mixture model converges to the true parameters. This project focuses on studying the bias behavior of MLEs in a specific misspecification scenario: the data is generated from a finite mixture of Student-t distributions, but the model is misspecified as a finite mixture of normal distributions. Due to the intractable nature of the objective functions for MLEs in mixture models, the Expectation-Maximization (EM) framework is utilized for parameter estimation. Our prior work involves simulation studies to investigate the bias in estimating mixture locations and mixing weights under varying settings. Building on this foundation, the undergraduate student will extend these investigations by conducting additional simulation studies and theoretical analyses to uncover deeper insights into the impact of misspecification on model parameters.

Prerequisites: Strong background in probability and statistical theory, proficiency in programming with Python or R.

Faculty supervisor: Yang Chen PhD student supervisor: Jiuqian Shang

Forecasting Solar Flares with Deep Time Series Models

Description: A solar flare is an abrupt, large-scale release of electromagnetic radiation from the Sun. Forecasting strong solar flares is of great interest because they and associated space weather phenomena can inflict serious damage on communications and electrical infrastructure. In this project, the aim is to use cutting-edge time series models based on neural networks to forecast flares. These models are largely based on attention; they include foundation models pre-trained on a vast number of time series and custom architectures designed for time series forecasting. Flares are defined in terms of the solar X-ray flux time series. Many previous studies have found that the so-called SHARP parameter vector time series are useful for flare forecasting. These SHARP time series are derived summary statistics based on high-resolution images of the Sun in several frequency bands that are thought to be predictive of the solar flaring activity. The objective will be to obtain flare forecasts by using the SHARP parameter time series to predict the X-ray flux time series. If some method works reasonably well, then an interpretability method could be applied to identify patterns that presage flares. Students working on this project will write code in Python to implement the forecasting methods and evaluate their performance.

Prerequisites: No prior knowledge of astronomy is required, but strong computational skills are needed; ideally, students will have taken STATS/DATA SCI 315.

Faculty supervisor: Stilian Stoev PhD student supervisor: Victor Verma

Prognostic Modelling for Education Data with Mixed Models

Description: This project focuses on building models of student outcomes—such as test scores, attendance, and graduation rates—using mixed models and generalized linear mixed models. Students will work with a private, Texas-based dataset and apply statistical methods in R to fit and analyze these models. This modeling will be used to help Texas education administrators in a large state evaluate their programs and policies by comparing affected schools and students to unaffected but otherwise similar schools and students.

We need to model in terms of school and student characteristics various aspects of student success in their progression through school: from on-time promotion to the next grade and performance on elementary and middle school yearly achievement tests, on to chronic absenteeism and timely progression through high school end of course exams. This presents a rich opportunity to gain experience building models for real-world data, gain mastery over various mixed modeling techniques, and contribute to an active partnership with the Texas Education Agency.

The models produced by the student will form the first part of a new methodology being developed for causal matching using multi-level data, and there will be regular opportunities to discuss Causal Inference or education research for those students with relevant research interests.

Prerequisites: Proficient in modeling with R Experience with analysis of linear regression models

Preferred: Experience using GitHub for version control Experience with random effects, mixed effects, or generalized linear models Interest and eagerness to discuss education research, causal inference, or linear models

Faculty supervisor: Ben Hansen PhD student supervisor: Julian Bernado

Privacy-Protected Causal Inference

Description: Many large datasets that could potentially be used to inform public policy cannot be released to researchers due to data privacy considerations. State education databases with valuable information on student performance are one example. Databases of health and medical records are another. The overall goal of this project is to examine various methods for producing privacy-protected versions of such databases, and how these privacy-preserving methods interact with downstream causal inference analyses. Differential privacy (DP) is the most widely-accepted mathematical framework for defining what it means for one's data to be private. Undergraduate researchers will 1) use R or Python to run simulations on existing DP analysis methods to compare their performance; 2) run simulations on a novel DP analysis method designed to be more straightforward to implement in practice; and/or 3) apply the novel DP analysis methods developed to causal inference, comparing to alternate privacy-protecting procedures such as synthetic data generation.

Faculty supervisor: Johann Gagnon-Bartsch PhD student supervisor: Yizhou Gu, Sam Rosenberg