WN 2026 MRPS Project Descriptions

Statistical significance of clustering for complex data

Description: Clustering is widely used in biomedical research for identifying meaningful subgroups. However, most existing clustering algorithms do not account for the statistical uncertainty inherent in the resulting clusters, which can lead to spurious findings due to natural sampling variation. To address this issue, the Statistical Significance of Clustering (SigClust) method was developed to formally assess the significance of clusters in high-dimensional data. In this project, we will further extend SigClust for complex data such as single cell RNA-sequencing data. Extensive numerical comparisons with other existing methods will be performed.

Prerequisites: Students need have taken probability and statistics inference courses as well as knowledge in statistical machine learning tools such as clustering, and dimension reduction methods.

Faculty supervisor: Yufeng Liu PhD student supervisor: Not UM student

Inferring cosmological parameters from mass maps

Description: The standard model of Big Bang cosmology is characterized by a small set of parameters that describe the universe's composition, geometry, and expansion history. Several of these parameters can be estimated by analyzing mass maps, three-dimensional arrays that provide a noisy approximation of the distribution of matter — both visible and dark — across the sky. In this project, we take a probabilistic approach to inferring cosmological parameters from mass maps. Given a mass map produced by a realistic astrophysical simulator, we train a deep neural network to approximate the posterior distribution of cosmological parameters that could have generated it. Project participants will (1) use existing software to simulate mass maps for many different cosmological parameter combinations, (2) design and train the neural network in PyTorch using several different loss functions, and (3) apply the trained network to a collection of mass maps inferred from real astronomical images.

Prerequisites: No prior knowledge of astrophysics is expected. Familiarity with Bayesian statistics is helpful but not essential. Strong computational skills are required, and experience with PyTorch is preferred.

Faculty supervisor: Jeffrey Regier PhD student supervisor: Tim White

Energy distance dimension reduction for multilevel wearable data

Description: We will use engression and energy distance dimension reduction to understand the relationship between heartrate and step count measurements, obtained at high temporal resolution using wearables (e.g. Fitbit). Heart rate is largely a proxy for a person's 24-hour activity cycle whereas step count reveals bouts of walking at varying degrees of intensity. The data are multilevel, dyadic, and longitudinal, as they consist of paired measurements of a cancer patient and their caregiver, recorded daily for up to three months following the initiation of stem cell transplant therapy. Energy distance methods have not been developed for this non-iid setting, so this project involves methodologic work as well as efforts to align the analysis with meaningful scientific aims. Some of the statistics of interest are U-statistics which are computationally expensive to work with, so students will need to implement algorithms efficiently using a tool such as jax, rcpp, cython, julia, or another appropriate solution for such tasks.

Prerequisites:

Faculty supervisor: Kerby Shedden

Implementing Minimal TRMs for ARC-Style Puzzle Solving

Description: Tiny Recursive Model (TRM) is a novel architecture that uses recursive self-attention blocks to tackle puzzle-solving tasks such as those in the ARC-AGI benchmark. Despite using tiny models trained on small dataset, TRM outperforms large language models on difficult puzzles. While promising, the approach is still not well understood. In this project, we will implement a minimal TRM from scratch and train it on synthetic tasks to investigate its capabilities and limitations. Familiarity with self-attention and Transformers is preferred. Proficiency in Python and prior experience with a deep learning framework (e.g., PyTorch, JAX) are required.

Prerequisites: PyTorch/JAX

Faculty supervisor: Yixin Wang PhD student supervisor: Zhiwei Xu

Black Box Inverse Solver for Complex Scientific Systems

Description: Inverse problems—the challenge of determining underlying causes from observed effects—are fundamental to many scientific applications, including understanding underground geological formations, analyzing experimental physics data, and discovering new materials. Traditionally, solving these problems requires extensive manual effort: researchers must develop custom mathematical algorithms, optimization methods, and tuning procedures for each specific problem. Even minor changes in the governing equations, measurement setup, or data quality often require building an entirely new solution from scratch.

This project investigates whether we can instead learn how to solve inverse problems automatically using machine learning, rather than hand-crafting solutions each time. Our central approach uses advanced machine learning techniques—including meta-learning (learning how to learn) and foundation models (large-scale pre-trained systems)—to develop a general-purpose inverse solver that can adapt to new problems with minimal customization. This "black-box inverse solver" would work across different physical systems, automatically handle uncertainty in predictions, and require little problem-specific tuning.

Prerequisites: Proficient in Python; have math backgrounds and preferably familiar with PDE.

Faculty supervisor: Yixin Wang PhD student supervisor: Yidan Xu

Neural Posterior Estimation for Galaxy Cluster Detection

Description: Galaxy clusters, the largest gravitationally bound structures in the Universe, are crucial probes for understanding cosmology and dark matter. Current methods for detecting and cataloging these cosmic giants are limited by a lack of real-world, labeled ground-truth data, forcing scientists to rely on inaccurate algorithmic approaches rather than modern machine learning. In this project, we instead use neural posterior estimation, a cutting-edge simulation-based inference technique. We will first develop a high-fidelity simulator to generate labeled astronomical images. Then, using advanced computer vision tools such as diffusion models and flow matching, we will train a robust deep learning cluster detector to solve the inverse problem—mapping observed pixels back to the positions of galaxy clusters. This trained model will finally be applied to the real-world images from the Dark Energy Survey (DES) to create a superior galaxy cluster catalog.

Prerequisites: Strong computational skills and prior experience in computer vision (e.g., EECS 442) are essential. Experience with deep learning frameworks like PyTorch or TensorFlow is also essential. No background in astronomy or Bayesian statistics is required or expected.

Faculty supervisor: Jeffrey Regier PhD student supervisor: Gabriel A. Patron

Protein language models for understanding biological sequence data

Description: Protein language models (PLMs) are an emerging class of machine-learning methods that adapt ideas from large language models (LLMs) to study biological sequences. Rather than modeling words or sentences, PLMs learn biophysical and evolutionary constraints that shape amino-acid sequence diversity. Recent work shows that these models recover information about protein structure, function, stability, evolutionary conservation, and even mutational effects using sequence data alone. This project investigates how PLMs encode biological information and how their predictions can be interpreted and evaluated. Mentees will help fit PLMs to real biological datasets and study what these models learn about protein families. Responsibilities may include preparing protein-sequence datasets; running and modifying PLM training pipelines; visualizing embeddings and attention patterns; evaluating models on downstream prediction tasks (such as mutational-effect prediction); and comparing PLM outputs to known biological or structural annotations; examining how architecture and training data affect model generalization through simulation studies.

Prerequisites: This project is well-suited for students interested in the intersection of statistics, machine learning, and computational biology. Prior experience with Python is required. Familiarity with deep-learning frameworks (such as PyTorch or JAX), and/or knowledge of basic biology, is helpful but not necessary.

Faculty supervisor: Jonathan Terhorst PhD student supervisor: Hanbin Lee

Flexible MIRT Model Fitting via Deep Learning

Description: This project aims to develop flexible and robust methods for fitting multidimensional item response theory (MIRT) models using modern deep learning techniques. MIRT is a widely used modeling framework in fields such as educational measurement, psychology, and behavioral science, providing an important statistical foundation for representing individuals' latent traits—such as abilities, skills, or psychological attributes—and understanding how these traits shape their responses to test items or survey questions. Despite their popularity, the conventional model fitting method relies critically on some assumptions that can be restrictive in modern data analysis. This project aims to overcome this limitation via deep learning methods to learn more flexible distributions for latent traits and item response functions. This approach has the potential to greatly expand the applicability of MIRT models to real-world settings where classical model fitting methods often fall short. Students participating in this project will review relevant literature, explore various deep learning architectures, conduct simulation studies, and prepare written reports. Attention to detail and proficiency in programming languages, particularly Python, are essential.

Prerequisites: Python

Faculty supervisor: Gongjun Xu PhD student supervisor: Chengyu Cui

Latent Attribute Estimation in Cognitive Diagnosis

Description: This project aims to investigate efficient machine learning tools for recovering latent attributes in cognitive diagnosis models (CDMs). CDMs play a central role in educational and psychological assessment by modeling how individuals' mastery profiles over discrete attributes give rise to their responses to testing items. Despite their popularity, CDMs often fall short in analyzing large-scale datasets, as the computational burden of model fitting grows rapidly with assessment size and the dimensionality of the attribute space. This project will explore a variational approximation approach to improve the scalability and efficiency of CDM estimation. The project aims to provide new model-fitting strategies for CDMs suitable for large-scale assessment settings while ensuring accurate classification of individuals into mastery profiles. The project has the potential to facilitate various downstream tasks such as targeted instruction, adaptive testing, and personalized feedback. Students participating in this project will review relevant literature, conduct simulation studies, and prepare written reports. Attention to detail and proficiency in programming languages (Python and R) are essential.

Prerequisites: Python and R

Faculty supervisor: Gongjun Xu PhD student supervisor: Chengyu Cui

Assessing Educational Testing Fairness

Description: Educational assessments play a crucial role in evaluating student ability and providing meaningful comparisons across examinees. A major challenge in ensuring fairness in such assessments is Differential Item Functioning (DIF), which occurs when an item favors one group of examinees over another, even after adjusting for their underlying ability. Therefore, developing general and reliable statistical methods to detect DIF is essential for ensuring that assessments are fair and interpretable. Standard DIF detection methods, however, often rely on strong modeling assumptions and may overlook the influence of unmeasured factors, leading to potentially biased DIF conclusions. To address this need, we aim to develop general modern statistical methods that relax traditional assumptions and improve the detection of DIF, as well as methods to evaluate the robustness of the conclusions. Together, these aims contribute to a more flexible and robust method for understanding and evaluating item fairness in practical assessment applications.

Prerequisites: R and Python

Faculty supervisor: Gongjun Xu PhD student supervisor: Mengqi Lin

Evaluating prior effect in Bayesian hierarchical models

Description: In Bayesian modeling, it is common to encounter paramaters posterior predictive distributions are based on low effective sample size. In such situations, the posterior becomes too sensitive to prior specification. Sometimes the model is too complicated to explicitely evaluate the prior effect. This project aims to quantify the impact of prior on the posterior in terms of equivalent sample size. The main advantage of this measure is that it is conditional on the data at hand but nonparameteric and distribution free. Starting from validating the method for simple Gaussian prior, we can extend this to Gaussian mixture model. We plan to evaluate its performance with parameters in a high-dimensional or hierarchical setting.

Prerequisites: Course: 451 or 551 preferred, Software: R or Python

Faculty supervisor: Yang Chen PhD student supervisor: Soham Das

Phylodynamic inference: virus spillover

Description: Biologists now commonly seek to extract information on disease transmission from virus genomes sampled from infected individuals: this is the subject of phylodynamics. Statistics plays a critical role in these investigations, enabling rigorous confrontation between hypothesis and data. Hypotheses typically take the form of stochastic mechanistic models, and bringing such models into contact with data in a statistically efficient manner remains a cutting-edge challenge. This project will consider a case study arising from the multi-species dynamics of Middle-East Respiratory Syndrome (MERS) in camels and humans. The students will implement and compare existing statistical methods, looking for insights into the statistical methodologies and the scientific system.

Prerequisites:

Faculty supervisor: Aaron King and Edward Ionides

LLMs to assist inference for nonlinear dynamic systems

Description: Partially observed nonlinear stochastic dynamic systems are central to diverse scientific fields, including ecology, economics, endocrinology, entomology, epidemiology and exoplanetology. Likelihood-based inference offers the potential for statistically efficient use of noisy and incomplete data, but the computational methodology required to do this is somewhat complicated. Some research groups have therefore tried to use artificial intelligence (AI) techniques to provide alternatives to likelihood-based inference. This project takes an alternative approach, by using AI techniques to support the deployment of likelihood-based inference. Large language models (LLMs) are already widely used to support coding and data analysis, but can struggle to promote best practices for emerging scientific topics and methodologies that are not well represented in their training data. A suitably primed LLM may be required, and this project will investigate that hypothesis. Students will learn about inference for nonlinear systems and the role of LLMs in supporting data analysis.

Prerequisites:

Faculty supervisor: Edward Ionides PhD student supervisor: Aaron Abkemeier

Flow Matching for Redshift Estimation

Description: The Rubin Observatory in Chile has begun surveying the night sky in unprecedented detail. By measuring light from billions of distant galaxies, this survey offers a powerful new way to study dark matter and cosmology. However, the observatory measures each galaxy through only a few optical filters (specific wavelength ranges) rather than capturing detailed spectra. With such limited data, inferring a galaxy's redshift—a key quantity that indicates its distance—is difficult. Without accurate redshift estimates, cosmologists cannot effectively use these observations to study the large-scale structure of the Universe.

In this project, we will apply flow matching, a modern generative machine learning method, to redshift estimation. The goal is to build a model that not only predicts a galaxy's redshift but also outputs a full conditional probability distribution, capturing the uncertainty inherent in the data. Participants will train these models using real and simulated survey data, gaining practical experience with probabilistic deep learning.

Prerequisites: Strong computational skills and prior experience with deep learning frameworks like PyTorch or TensorFlow are essential. No background in astronomy or Bayesian statistics is required or expected.

Faculty supervisor: Jeffrey Regier PhD student supervisor: Jackson Loper (postdoc)

Overcoming the winner’s curse

Description: Decision-makers routinely select top-performing options—such as advertisements, recommendations, variables, or policies—based on estimated effects, choosing those that rank "highest" according to a given criterion. This gives rise to the winner’s curse phenomenon: because the estimated effects guiding the top-performing options are uncertain, and decision-makers are more likely to select an option when its performance has been overestimated. As a result, in future experiments, the selected options or the winners tend to systematically underperform relative to their initial estimates.

In this project, we focus on a flexible methodology that generates uncertainty estimates for the top-performing options while explicitly accounting for the over-optimism from the selection of winners. The undergraduate student will be responsible for evaluating the potential of this methodology on real data applications and for creating test beds to assess its performance. We expect the student to be familiar with Python. The MAS student will be in charge of developing parts of the methodology as well as conducting experiments to evaluate the proposed methods.

Prerequisites:

Faculty supervisor: Snigdha Panigrahi PhD student supervisor: Soham Bakshi

Valid Inference After Hyperparameter Tuning

Description: Hyperparameter tuning is an essential part of most machine learning workflows, allowing data scientists to improve the predictive performance of their models. In practice, hyperparameters are rarely set in advance without looking at the data. Instead, it is common to test different hyperparameter values, using tools like cross-validation, before selecting a "best" one. However, the selection of hyperparameters creates a major gap between theory and practice: classical inference tools do not account for the data-dependent nature of tuned hyperparameters, often leading to grossly invalid results.

In this project, we investigate principled inferential approaches that account for the tuning of various hyperparameters during predictive modeling, such as selecting the regularization parameter in penalized regression or determining the amount of synthetic data to integrate with gold-standard datasets. The undergraduate student will be responsible for evaluating the potential of this methodology on real data applications and for creating test beds to run simulated experiments. The MAS student will be in charge of developing parts of the methodology as well as conducting experiments to evaluate the proposed methods.

Prerequisites: Familiarity with Python and upper level regression course , like 413

Faculty supervisor: Snigdha Panigrahi

Agentic AI

Description: The project focuses on developing real-time, uncertainty-aware orchestration for multi-agent AI systems through two main directions: anomaly detection across agents and adaptive re-routing of sub-tasks. The first aims to identify and contain failures early by modeling how uncertainty and errors propagate and correlate across interconnected agents, preventing systemic breakdowns. The second develops dynamic routing policies that assign sub-tasks to the most suitable agents based on context and uncertainty, optimizing performance and cost. Together, these efforts seek to make multi-agent systems more robust, efficient, and adaptable, with potential extensions toward dynamic agent creation and general orchestration frameworks.

Prerequisites: Coding - the project involves coding with LLMs

Faculty supervisor: Raed Al Kontar PhD student supervisor: Qiyuan Chen

Making MCMC Plug-and-Play Diffusion Priors Work in Practice

Description: This project will investigate MCMC-based plug-and-play diffusion priors (PnPDP) for solving inverse problems. The idea is to take a pretrained diffusion model and use it inside an iterative algorithm to reconstruct a clean image from noisy or blurred measurements (e.g., MRI or CT).Inverse problems like denoising or deblurring are often ill-posed, meaning many different images could explain the same measurement. PnPDP methods address this by using a diffusion model as a prior for what realistic images look like. The MCMC-based variants (such as PnP-DM and MCG-diff) go one step further: given sufficient computation, they can in principle approximate the full posterior distribution rather than just output a single reconstruction.However, these samplers are sensitive to hyperparameters such as the particle count, and there is currently little systematic guidance on how to choose them. This project aims to (1) empirically study the parameter sensitivity of these methods—starting from toy problems with known ground-truth posteriors, then moving to realistic image (2) derive practical tuning guidelines for practitioners if possible .

Prerequisites: Proficiency in Python and basic knowledge of machine learning, probability and statistics inference (the level of STATS 425,426) are required. Solid programming training (the level of EECS280/281) and experience in MCMC are preferred.

Faculty supervisor: Liyue Shen PhD student supervisor: Xiaoyu Qiu

Optimizing Score Thresholds in Score-Explained Heterogeneous Treatment Effect Models

Description: We examine settings where treatment assignment is determined by some score threshold. Examples include economic aid programs and school admission, where 'treatment' is given to those meeting thresholds for income or exam scores. Traditional approaches to these problems focus on observations close to the score threshold, excluding a significant portion of the sample and only study treatment effect at the threshold. We use a model that considers most observations, and aim to find a threshold that would optimize utility for the organization (or utility for the individual). We also consider a 'performative' model, where agents can respond to a published threshold (such as extra preparation for an exam to meet a threshold if they're close), which leads to a change to the optimal cut-off point. The student will develop a theoretical understanding of such models, run simulations that motivate and validate theoretical analysis, and delve into real data applications.

Prerequisites: Multivariate Calculus, Linear Algebra, Strong Coding Background (Python), some knowledge in Probability and Statistical Inference

Faculty supervisor: Moulinath Banerjee, Ya'acov Ritov PhD student supervisor: Daniel Zou

Nash Equilibrium in Closed Markets

Description: In financial markets, trading means buying or selling assets like stocks or currencies with the goal of making money from short-term changes in prices. A trader might buy an asset if they think its price will go up and sell it later for a profit. In this project, we want to:

1) Study the equilibrium structure of a simple market, where traders interact only once and only with each other, and each trader seeks to maximize their pay-off. The equilibrium is not the standard Nash Equilibrium because, since the traders interact only with each other, the strategy sets depend on the other traders' actions, making the problem non-trivial. Indeed, let p_i be the i'th trader's selling price and b_i their buying price, and v_i denote their valuation of the product. Trader i buys from trader j at price p_j whenever v_i > p_j, and trader j buys from trader i at price b_j whenever v_i < b_j. Since the market is closed, the total number of sold items is equal to the total number of items bought. Any Nash equilibrium {(b_i^,*p_i*^)} has to satisfy this constraint, which implicitly enforces an interdependence of the strategies across traders.

2) Work through small examples (2/3 traders) that show how such an equilibrium can form.

3) Introduce statistical learning of the traders' best responses dynamically and design an algorithm converging to equilibria.

Prerequisites: Basics of Multivariate Calculus, Linear Algebra, Probability and Statistical inference. Coding required: R or Python.

Faculty supervisor: Moulinath Banerjee PhD student supervisor: Daniele Bracale

Link Prediction for Incomplete Network Data Using Generative Models

Description: Link prediction is a fundamental problem in network analysis, aiming to determine whether links exist between pairs of nodes based on observed information. Most existing link-prediction methods require at least partial observation of connections for every node. However, in real-world networks (e.g., social networks), there are often nodes for which no link information is available, such as newly added nodes, commonly referred to as the "cold-start" problem. Link prediction for these nodes becomes feasible if node features or covariates are available, as these features can reveal connectivity tendencies. In this project, our goal is to develop generative models for incomplete networks with node features, with a particular focus on predicting links for nodes that only have feature information. Participating students will be expected to read assigned literature, assist with method implementation, and conduct simulation studies and real-world data analysis.

Prerequisites: 1) Proficiency in programming languages, particularly Python. 2) Familiar with probability and statistical learning.

Faculty supervisor: Ji Zhu PhD student supervisor: Xuanyu Chen

Individualized treatment rule selection with many treatments

Description: With advances in medical research and real-world clinical practice, patients often face a wide range of treatment options for the same disease. This diversity presents significant challenges for developing optimal individualized treatment rules (ITRs), especially when multiple treatment choices and various outcomes are involved. In this project, we will study strategies to handle problems with many treatments and identify accurate and robust ITRs for such problems.

Prerequisites: Solid math and stat foundations and good programming skills.

Faculty supervisor: Yufeng Liu PhD student supervisor: Not a UM student. My Ph.D. student at UNC

Latent Space Zero-Inflated Poisson Model for Count-Weighted Networks

Description: Latent space models have been shown to be powerful tools for capturing and understanding structural characteristics in networks. Although many latent space models have been developed for binary networks, a wide range of real-world networks are inherently count-weighted, where edges represent count-valued interactions. Statistical models specifically designed for such networks remain limited. This project focuses on sparse count-weighted networks and aims to develop a latent space zero-inflated Poisson model to characterize both the existence and the strength of connections. We plan to explore estimation strategies: a projected gradient descent algorithm and an EM-based approach. Participating students will be expected to read assigned literature, assist with method implementation, and conduct simulation studies as well as real-world data analysis.

Prerequisites: 1) Proficiency in programming languages, particularly Python. 2) Familiar with probability and statistical learning.

Faculty supervisor: Ji Zhu