REU 2021 Projects

These are brief ideas. Actual projects may vary a little bit.

Jianping Sun

A Computational Efficient Method for Constructing Hierarchical Trees

Hierarchical trees are essential in genomic research, because they provide major tools for scientists to study evolution history and detect disease associated genomes. Fast developed sequencing technology, such as the next generation sequencing, has enabled researchers to obtain whole genome DNA sequences at a relatively low price. Hence there is an urgent need to develop statistical methods that can construct hierarchical trees from long DNA sequences by taking various biological complexities into account. In this project, we will develop a novel method for constructing hierarchical trees while accounting for significant evolution factors, such as mutation and recombination, simultaneously. A computational efficient algorithm will also be designed to accomplish the proposed methodology and make it practical when the length of sequence is large.

John Stufken & Rakhi Singh

Subdata Selection Methods

Data is everywhere and there is a huge amount of it. There are opportunities, more than ever, to answer relevant questions empirically by using the humongous amount of available data. For example, Walmart handles more than 1 million customer transactions every hour, which are imported into databases estimated to contain more than 2.5 petabytes (2560 terabytes) of data. As another example, a transportation company like Uber has tremendous amounts of consumer preference data that they use to predict supply, demand, location of drivers, and fares that are set for every trip. Analyzing data of this size, if feasible at all, requires gigantic computational resources and the development of novel methods. But even for smaller sized big data, depending on the available computational platform and how often an analysis or exploration needs to be performed, the computational burden can be considerable. For that reason, methods have been developed to conduct an analysis based on only some of the data, referred to as subdata. The research question that we are interested in answering is which subdata is a good subdata when (a) prediction is the goal, and (b) a good representation of the population is the goal. In this project, we expect to come up with new methods of subdata selection while using big data techniques for analyzing the data from different subdata selection methods. Even though the statistical theory can be (should be) built around the procedures used, it is extremely hard, and we will try to answer most of the questions computationally. It is an ideal project for someone who has some knowledge of statistics and programming, has an interest in learning big data analysis techniques, and is passionate about coding (in R or Python – if you want to have access to some pre-existing codes).
Untruthful Responding in Randomized Response Models

Randomized response technique (RRT) models are important survey tools when dealing with potentially sensitive questions with legal or social implications. These models allow respondents to provide scrambled (noise-added) responses which are later unscrambled at an aggregate level but not at an individual level. Some of the respondents may not trust the privacy protection provided by RRT models and may still provide untruthful responses. While there is no way to identify and fix the untruthful responses, one can measure the level of untruthfulness and account for it in the final estimates. The 2020 REU group considered this problem in the binary response setting. We plan to carry that study forward and consider a similar problem for quantitative response situations. Untruthful responding is part of the larger problem of Measurement Error in statistical estimation.

Topological Data Analysis and Firn

Project:
Topological data analysis (TDA) is a rising field at the intersection of Mathematics, Statistics, and Machine Learning. Techniques from this field have proven successful in analyzing a variety of scientific problems and datasets. The main driving force in TDA is the development of persistent homology, which studies the intrinsic shape of data. The main goal of the project is to direct TDA tools at understanding microstructure and fluid flow in porous media. The main application of this project is the Firn, a type of ice core data. The project is interdisciplinary. Students will join a team of mathematician, statistician, and climatologist to work on the gas age-ice age problem in the climate science.

Ranked Data Analysis of RCV elections

Ranked choice voting (RCV) is seeing a renewed surge of interest in the United States. For example, New York City will use ranked choice voting for local offices starting this year. While political scientists have long been in the business of modeling elections, applications of statistical models to RCV election data are still rare. This presents an exciting opportunity for new research at the intersection of statistics and political science. In this project, we will be investigating how to model RCV elections in which different voter groups are competing to elect their preferred representatives. This is especially relevant to minority representation in local government, an area where many believe RCV can have a strong positive impact in the near future.
Xiaoli Gao
Robust Classification of high-dimensional data with fuzzy group information

In cancer research and genetic studies, it is important to identify potential genomic biomarkers out of tens of thousands genetic features which are influential to certain phenotype. In this project, we will develop a robust penalized logistic regression model for simultaneous feature selection and classification with fuzzy group information in high-dimensional data settings. The proposed approach will be applied to the gene expression data on pediatric acute myeloid leukemia (AML) prognosis. Ideally, students participating in the project should have background in linear algebra and linear regression, with programming experiences in R, Matlab or Python. It will be helpful for students considering this project to browse the references in Tibshirani (1996), Yuan and Lin (2006), Zhu and Hastie (2007), Gao (2016), Gao and Feng (2018).

Scott Richter
Combined Tests for Experiments with Matched-Pairs and Independent Samples Data

In many applications a matched-pairs design is used to control for variability between subjects and allow for more precise treatment comparisons. However, in many instances, missing data occur due to the inability to obtain one of the measurements. In these situations, a mixture of complete and incomplete pairs of data will be available.

Several approaches have been proposed to incorporate information from incomplete pairs. However, there is no procedure that is superior under all conditions, and the effect of several factors on the performance of these methods is not well understood. In this project we will study one or more of the following questions/topics.

1. What is the effect of unequal variances on the performance of the tests?

The effect of unequal variance on the performance and properties of the proposed tests will be studied, and ways to mitigate negative effects of unequal variance investigated.

2. Develop a robust test to assess the assumption of equal variance.

Ways to develop an effective test of equal variance will be investigated, and the properties and performance of the test studied.

3. Can a combined test statistic improve the power of the nonparametric tests?

Ways to develop a test based on a combination of the test statistics will be considered and the properties and performance compared to those of the individual statistics.