Description of Research Projects

a. Data Confidentiality and Related Issues

Mentor: Sat Gupta, Professor of Statistics
Co-Mentor: Somya Mohanty, Assistant Professor, Computer Science
Assistant: A Statistics PhD Student

Maintaining confidentiality of data is a very important issue. Primitive tools such as stripping the name and id etc. have become obsolete. One of the ways to do this is to inject noise in the data in such a way that aggregate level properties are not disturbed much even though record level properties shift. It is important for researchers, particularly younger researchers, to recognize that the data they are working with is not necessarily the real data. We would like our student researchers to understand the basic Randomized Response Techniques (RRT) including those given by Warner (1965, 1971). They will also learn other data scrambling techniques such as changing aggregation level (releasing data at higher level of aggregation) and scrambling some variables but not all (partial scrambling). In the specific research project for the REU program, we will explore the problem of **Untruthful Responding in RRT Models.** This is one of the sources of data contamination.

A student wanting to pursue this project should have taken a course on probability and statistics and should have some familiarity with SAS or R. It will be helpful for students considering this project to browse the references Warner (1965, 1971), Gupta et al. (2002, 2010, and 2013).

a. Robust Integrative Data Analysis for Data Contamination and Heteroskedasticity

Mentor: Xiaoli Gao, Associate Professor of Statistics
Co-Mentor: Jianping Sun, Assistant Professor of Statistics
Assistant: A Statistics PhD Student

In this project, we will work on simultaneous outlier detection and feature selection with fuzzy group information. In real applications, the data can be irregular due to the contamination from outliers or leverage points and the existence of heteroskedasticity. This phenomenon becomes more common in high-dimensional settings when a large number of predictors are collected. Because of the co-existence of high-dimensionality and data contamination, simultaneous outlier detection and variable selection become important issues (She & Owen 2011, Hampel et al. 1986). Recently, Co-PI Gao and her coauthors have developed some robust high-dimensional methods dealing with this issue. This method has been successfully applied in genetics such as copy number variation (Gao, 2015; Gao and Fang, 2016, Gao and Feng, 2018). In their work, a high-dimensional penalized weighted regression model (PWR) is considered for simultaneous outlier detection, robust regression and variable selection.

Ideally, students participating in the project should have background in linear algebra and linear regression, with programming experiences in R or Matlab. It will be helpful for students considering this project to browse the references Tibshirani (1996), Yuan and Lin (2006), and Gao and Fang (2016).

a. A Big-Data Approach Towards Food Safety Using the Twitter Human Geo-Sensors

Mentor: Somya Mohanty, Assistant Professor, Computer Science
Co-Mentor: Sat Gupta, Professor of Statistics

According to a Centers for Disease Control and Prevention report, each year, approximately 48 Million people in the United States contract foodborne illness (Gould, 2013). Traditional approaches for tracking of foodborne illnesses rely upon the analysis of medical data to detect or anticipate the outbreak of diseases (Chen, 2010). However, the ubiquity of social media, has led to a repository of self-reported symptoms of illness. Real-time tracking of such self-reports can act as a *precursor* to the outbreak of such incidents. Such

an approach could lead to the timely deployment of prevention measures, attenuating the effect of outbreaks towards a more resilient food safety architecture. The overarching goal of this project is to apply an infodemiological (Eysenbach, 2009; Chew, 2010; Newkirk et al., 2012) approach towards conducting syndromic surveillance of foodborne illnesses using *human geo-sensor* (social-media user with location information) data from Twitter.

To that end, the research proposes to: *Aim 1*) Investigate the viability and develop predictive models using Twitter data as a resource for identifying foodborne illnesses; and *Aim 2*) Validate the models by triangulating results from machine-learning / traditional statistical analysis, with existing disease data sources. Twitter generates 600 million messages/tweets each day from its approximately 250 million active users; almost 60% of tweets emerging from mobile devices with built in geo-location services (Olanoff, 2015). Using Twitter public API data-access points, the research will collect geo-coded data for the region of contiguous United States.

Students participating in the project are required to have basic probability and statistics knowledge, with programming capability in Python or R.

d. A Perturbation Multivariate Approach in Family-Based Genome Wide Association Study
Mentor: Jianping Sun, Assistant Professor of Statistics
Co-Mentor: Xiaoli Gao, Associate Professor of Statistics

The advantage of high-dimensional genomic sequencing technologies leads to new investigation in the role of rare and common genetic variation underlying the complex human diseases and traits (UK10K Consortium, 2015). Due to the low power associated with separate analyses of rare events, novel methodologies for simultaneous analysis of genetic variation in small genomic regions have been developed recently (Wu et al., 2011; Lee et al., 2012; and Sun et al., 2013), including multivariate methods on joint analysis multiple traits in order to improve power to detect disease associated genetic variants (Maity et al., 2012 and Sun et al., 2016). However, these multivariate approaches are limited to the case of independent individuals and do not handle family-based designs, because of the data complexity brought by unknown correlation structures from both multiple traits and related individuals.

Recently, a series of multivariate test statistics have been developed, while allowing also for correlations between individuals in families (Sun et al., 2018). However, to explicitly derive the distribution and analytically compute the *p*-values, these test statistics are constructed based on Copula model, and hence have restrict model assumptions and are less practical in application. Aroused by resampling-based perturbation method, which has been used to joint analysis SNP and gene expression data (Huang et al., 2014), we propose to use perturbation approach to calculate *p*-values empirically for the test statistics proposed in Sun et al. (2018), so that to relax Copula model assumptions, avoid huge computation burden brought by classical re-sampling method such as permutation, and make the test statistics more practical and computationally affordable for family-based genome wide association study with multiple traits.

A student wanting to pursue this project should have taken a course on probability and statistics and should have some familiarity with R and Unix environment. It will be helpful for students considering this project to browse the references Huang et al. (2014), Wu et al. (2011), and Sun et al. (2013, 2016, 2018).

e. Combined Tests for Experiments with Matched-Pairs and Independent Samples Data
Mentor: Scott Richter, Professor of Statistics
Co-Mentor: Xiaoli Gao, Associate Professor of Statistics
Assistant: A Statistics PhD Student

A matched-pairs design is often used to allow for more precise treatment comparisons. However, missing data may occur for several reasons. A clinical trial to compare two methods of eye laser surgery (Dubnicka et al, 2002), patients may have one eye assigned to a new method and the other to the current method, but some patients may have only one eye eligible for study. Similarly some subjects are measured at two time points with an intervention in between, may be lost to follow-up. In these situations, a mixture of complete and incomplete pairs of data occurs.

Ignoring unpaired data and analyzing only the complete pairs may introduce bias due to systematic missing data as well as a loss of power Ideally, the information from observations from both treatments among the unpaired data could be combined with that of the complete pairs.

Parametric and nonparametric approaches have been proposed to incorporate information from in complete pairs (See Dubnicka et al (2002), Eisnporn and Habtzghi (2013)). However, the effect of several factors on the performance of these methods is not well understood, nor is their relative performance. In this project we will study one or more of the following questions:

What is the effect of sample size and correlation between paired observations on the tests? Is trying to incorporate information from incomplete pairs always beneficial, or does the benefit depend on the relative sample sizes, whether incomplete pair sample sizes are equal, or on the correlation between complete pairs? Can alternative weighting schemes improve power?

We will design a simulation to study the power properties of several proposed methods to investigate these questions. We will also investigate new ways to weight the paired and two-sample test statistics to improve power.

Interested students should have completed introductory courses in both mathematical statistics and applied statistical methods and should have familiarity with a programming language.

f. Subdata Selection Methods
Mentor: John Stufken, Bank of America Excellence Professor
Assistant: A PhD student or a Post Doc

Because of technological advances it is, in some areas of application, easy to collect and store enormous amounts of data. Sizes of hundreds of gigabytes, terabytes, and even petabytes are no longer uncommon. 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. Questions that arise immediately are (1) what size should the subdata have to ensure a reliable analysis; and (2) for a given size, how should the subdata be selected? The first of the above questions has not received lots of attention. For the second question, several methods have been proposed in the context of linear regression.

In this project, we will explore various ways for subdata selection and assess their optimality. Developing theory to answer these questions will be very challenging and may not be possible in many cases. Much additional knowledge and insights will have to come from simulations that are easily accessible to undergraduate students with some knowledge in statistics and programming.