Statistics and Data Science Seminar Series for Fall 2023

September 15

Speaker: Suvra Pal
Title: Support Vector Machine-Based Cure Rate Models
Abstract: In this talk, I will present a new promotion time cure model (PCM) that uses the support vector machine (SVM) to model the probability of cure. The proposed model inherits the features of the SVM and provides flexibility in capturing non-linearity in the data. Furthermore, the new model can incorporate potentially large number of covariates. For the estimation of model parameters, I will discuss the steps of an expectation maximization algorithm where I will make use of the sequential minimal optimization technique together with the Platt scaling method. Next, I will present the results of a detailed simulation study and show that the proposed model outperforms the existing logistic regression-based and spline regression-based PCM models, specifically when the true classification boundary is non-linear. I will also show that the proposed model’s ability to capture complex classification boundaries can improve the estimation results related to the survival distribution of the uncured. Finally, I will analyze a data from leukemia cancer study and show that the proposed model results in improved predictive accuracy.

September 22

Speaker: Mohamad Kazem Shirani Faradonbeh
Title: Learning from Multiple Multivariate Time-Series
Abstract: Autoregressive models are among the most popular ones for analyzing temporally dependent data. However, little is known about learning from multiple time-series trajectories, especially when it comes to multi-dimensional and non-stationary data. We study this problem and propose a joint estimation method for learning transition matrices of multiple vector autoregressive models that are unknown linear combinations of some unknown basis matrices. The setting is technically challenging due to high dimensionality of the parameter space as well as the compound nature of the uncertainty. Still, our theoretical analysis shows that the proposed joint estimator has an optimal sample-complexity and excels individual learning methods. Furthermore, applications to data-driven stabilization of dynamical systems through exogenously designed input experiments will be discussed.

September 29

Speaker: Noirrit Chandra
Title: Bayesian Nonparametric Common Atoms Regression for Generating Synthetic Controls in Clinical Trials
Abstract: The availability of electronic health records (EHR) has opened opportunities to supplement increasingly expensive and difficult to carry out randomized controlled trials (RCT) with evidence from readily available real world data. In this paper, we use EHR data to construct synthetic control arms for treatment-only single arm trials. We propose a novel nonparametric Bayesian common atoms mixture model that allows us to find equivalent population strata in the EHR and the treatment arm and then resample the EHR data to create equivalent patient populations under both the single arm trial and the resampled EHR. Resampling is implemented via a density-free importance sampling scheme. Using the synthetic control arm, inference for the treatment effect can then be carried out using any method available for RCTs. Alternatively the proposed nonparametric Bayesian model allows straightforward model-based inference. In simulation experiments, the proposed method exhibits higher power than alternative methods in detecting treatment effects for complicated response functions. We apply the method to supplement single arm treatment-only glioblastoma studies with a synthetic control arm based on historical trials.
October 13

Speaker: Jungsik Noh
Title: Granger-causality Inference Framework to Study the Actin Cytoskeleton Regulatory System from Live Cell Fluorescence Imaging Data

Abstract: Many cell regulatory systems implicate nonlinearity and redundancy among components. The regulatory network governing actin cytoskeleton structures at the cell edge is prototypical of such a system, containing tens of actin-nucleating and -modulating molecules with functional overlap and feedback loops. Due to instantaneous and long-term compensation, phenotyping the system response to perturbation provides limited information on the targeted component’s roles in the unperturbed system. Accordingly, how individual actin regulators contribute to actin cytoskeleton dynamics remains ambiguous. Here, we present a perturbation-free reconstruction of cause-effect relations among actin regulators by applying Granger-causal inference to constitutive image fluctuations that indicate regulator recruitment as a proxy for activity. Using multivariate time series representing the spatiotemporal molecular and cell morphological dynamics of the system, we identify spatially confined causal relationships between the local activities of actin and its regulators and the corresponding local cell edge motion.

October 20 – Special event: ASA Travel Course

Speaker: Babette Brumback
Title: Fundamentals of Causal Inference: With R

Abstract: One of the primary motivations for clinical trials and observational studies of humans is to infer cause and effect. Disentangling causation from confounding is of utmost importance. Fundamentals of Causal Inference: With R explains and relates different methods of confounding adjustment in terms of potential outcomes and graphical models, including standardization, doubly robust estimation, difference-in-differences estimation, and instrumental variables estimation. Several real data examples, simulation studies, and analyses using R motivate the methods throughout. The course assumes familiarity with basic statistics and probability, regression, and R. The course will be taught with a blend of lecture and worked examples. More details here: https://www.amstat-nt.org/events/short-courses

October 27

Speaker: Monnie McGee
Title: Hypothesis Testing for Multiple Groups of Compositional Data

Abstract: Let \( x_1, x_2, \ldots, x_k \) be \( k \) \( n \)-dimensional vectors such that \( \sum_{j=1}^{k} x_j = 1 \). These data are typically presented as proportions or percentages of a whole; therefore, \( \min x_j \geq 0 \). For example, suppose we examine daily activity data for thousands of people, as is possible with current fitness apps, where we obtain the time spent sleeping, eating, exercising, and working. In this case, there would be \( k = 4 \) components. Clearly, once one knows the time spent sleeping, eating, and exercising, then the time spent working is determined; therefore, the components are not independent. Suppose further that we want to compare the relative proportions of activity for different populations. Such data is frequently analyzed incorrectly by comparing the populations via a t-test (or ANOVA for multiple populations) on one component of the vector at a time.

Instead of modeling each component separately, the components can be measured together using a Dirichlet distribution or a nested Dirichlet distribution. For a composition made up of \( k \) variables, the appropriate Dirichlet distribution will have \( k \) parameters, \( \alpha = (\alpha_1, \ldots, \alpha_k) \). The \( \alpha_j \) can be thought of as counts from a prior or a current study, depending on the context, and \( A = \sum_{j=1}^{k} \alpha_j \) is known as the precision. There are two main drawbacks when using the Dirichlet distribution to model compositional data. One drawback is that components with the same mean must also have the same variance. Another limitation is that the covariance between any two components is non-positive. In practice, compositional datasets do not follow these constraints.

The nested Dirichlet distribution (NDD) is a more general form of the Dirichlet distribution that relaxes the constraints that variables with the same mean must have the same variance. It also allows for the covariance between variables to be nonnegative. The correlation structure between variables is determined by how the variables are nested. We illustrate the structure of the NDD as we derive a likelihood ratio test for \( G \geq 2 \) groups, where each
group comes from a NDD and all groups have the same nesting structure. Importantly, the overall test for equal means does not assume that the precision values for the groups are the same.

As an example, we consider a data set tracking the performance of 1975 unique major league baseball players from the years 2000 to 2010. We separately grouped at-bat outcomes for three separate years (2000, 2005, and 2010) into six components \((k = 6)\): home runs, singles, doubles, triples, outs, and other. The “other” category groups the outcomes interference, hit by pitch, and base on balls. We also split the batters into three groups based on age. The ages represented young batters \((\text{age} \leq 28)\), experienced batters \((\text{age} > 35)\), middle aged batters \((28 < \text{age} \leq 35)\). We find that the composition of the at-bat outcomes is remarkably similar from year to year, but the composition of outcomes for the age groups are not. We conclude with implications for this finding and future directions for research.

**November 3**

**Speaker:** MinJae Lee  
**Title:** Statistical methods for handling various measurement issues, and their applications in biomedical and cancer prevention research  
**Abstract:** Assessment of longitudinal biological factors or intervention effects associated with disease prevention/progression is critical for biomedical and clinical research. However, in many patient-based studies, we encounter challenges in analyzing longitudinal data that are prone to various types of measurement issues. For example, biomarker data, which provide insight into the treatment/intervention effects, are often subject to left-censoring due to measurement threshold, and/or missing due to incomplete follow-up visits. In many longitudinal studies various outcome data are often collected through patients’ self-reporting, which could be missing or inaccurate/untenable due to various sources of errors. These issues make assessments even more complicated especially when data are collected from heterogeneous and diverse at-risk populations. Although existing statistical methods can deal with specific types of measurement issues, inappropriate handling can lead to biased/incorrect results when they involve complex sources of errors. In this talk, I will introduce statistical methods that can address these challenges and their applications in various biomedical and cancer prevention studies.

**November 10**

**Speaker:** Karabi Nandy  
**Title:** The Development of a Suicidal Risk Scale through Psychometric Analyses  
**Abstract:** The Concise Health Risk Tracking Self-Report (CHRT-SR) is a self-reported survey instrument that was created to measure the risk of suicidal behavior. In this talk, I will present its psychometric properties in a representative sample of adolescent outpatients (<18 years of age). This will be done through a series of steps: first, by evaluating the scale’s factor structure using multigroup confirmatory factor analysis; next by testing measurement invariance across age and gender; then, assessing the item response theory and classical test theory characteristics of the scale; and finally, assessing the scale’s concurrent validity (both cross-sectional and as a change measure over time) by anchoring it against an independent suicide measure. Reliability and validity assessments based on these various methods revealed that the CHRT-SR has excellent psychometric properties that is sensitive to changes in suicidality over time in this adolescent sample. I will also present our study of its performance in other populations across the age spectrum and across various clinical settings.

**November 17**

**Speaker:** David Kahle  
**Title:** Variety Distributions and Applications  
**Abstract:** Nonlinear systems of polynomial equations arise naturally in many applied settings. The solution sets to these systems over the reals, called real varieties, are often positive dimensional spaces that in general may be very complicated yet have very nice local behavior almost everywhere. In this work we communicate recent progress towards a Monte Carlo framework for exploring such real solution sets. After describing how to construct
probability distributions whose mass focuses on a variety of interest, we describe how Hamiltonian Monte Carlo methods can be used to sample points near the variety that may then be moved to the variety using endgames. We conclude by showcasing trial experiments and applications.