Abstracts

Friday, October 13, 2023

Morning Session

Title: Modeling Spatio-Temporal Dynamics Under Data Sparsity
Speaker: Ali Arab, Department of Mathematics and Statistics, Georgetown University
Ph.D., University of Missouri - 2007

Abstract: Modeling the dynamics of spatio-temporal processes is often challenging and this is exacerbated under data sparsity (often the case in early stages of a process). For example, modeling the dynamics of an infectious disease at early stages is very challenging due to data sparsity (as well as potential lack of knowledge regarding the disease dynamics itself); this is an important issue for modeling emerging and re-emerging epidemics. Moreover, data sparsity may also result in inefficient inference and ineffective prediction for such processes. This is a common issue in modeling rare or emerging ecological, environmental, epidemiological, and social processes that are new or uncommon in specific areas, specific time periods, or those conditions that are hard to detect. Consequently, due to the urgency of modeling these processes in many situations (e.g., in a crisis situation), often one limited predictor data to use either because of lack of knowledge about the process or the need for fine resolution predictor data. For example, modeling the dynamics of conflict- or climate-driven human migration may be quite complex to model (in particular, when a crisis occurs and there are abrupt migration in/out flows). Classic models that are commonly used in these areas often fall short of modeling such events and are unable to provide reliable inference and reasonable or accurate forecasts. Also, the factors that are linked with migration processes are often related to long term migration and/or only available at spatially and temporally aggregated level. In this paper, we discuss strategies for dealing with some of the statistical issues of modeling dynamics under data sparsity including: utilizing blended data (i.e., both conventional and organic data sources), considering a mechanistic science-based modeling framework to model the dynamics of a spatio-temporal based on zero-modified hierarchical modeling approaches, and implementing a variable selection method to assist with inducing sparsity on predictor variables when dealing with massive datasets (as a result of using organic data). We will provide examples based on real data.

Title: Trend Filtering with Adaptive Bayesian Changepoint Analysis for Count Time Series
Speaker: Toryn Schafer, Department of Statistics, Texas A&M University
Ph.D., University of Missouri - 2020

Abstract: Model development for sequential count-valued data characterized by small counts and non-stationarities is essential for broader applicability and appropriate inference in the scientific community. Specifically, we introduce global-local shrinkage priors into a Bayesian dynamic generalized linear model to adaptively estimate both changepoints and a smooth trend for count time series. We utilize a parsimonious state-space approach to identify a dynamic signal with local parameters to track smoothness of the local mean at each time-step. This setup provides a flexible framework to detect unspecified changepoints in complex series, such as those with large interruptions in local trends. We detail the extension of our approach to time-varying parameter estimation within dynamic Negative Binomial regression analysis to identify structural breaks. Finally, we illustrate our algorithm with empirical examples in social sciences.

Title: The Link Between Health Insurance Coverage and Citizenship Among Immigrants: Bayesian Regression Modeling of Categorical Survey Data Observed with Measurement Error

Speaker: Paul Parker, Department of Statistics, University of California, Santa Cruz
Ph.D., University of Missouri - 2021

Abstract: Social scientists are interested in studying the impact that citizenship status has on health insurance coverage among immigrants in the United States. This can be done using data from the Survey of Income and Program Participation (SIPP), however, two primary challenges emerge. First, statistical models must account for the survey design in some fashion to reduce the risk of bias due to informative sampling. Second, it has been observed that survey respondents misreport citizenship status at nontrivial rates. This too can induce bias within a statistical model. Thus, we propose the use of a weighted pseudo-likelihood mixture of categorical distributions, where the mixture component is determined by the latent true response variable, in order to model the misreported data. We illustrate through an empirical simulation study that this approach can mitigate the two sources of bias attributable to the sample design and misreporting. Importantly, our misreporting model can be further used as a component in a deeper hierarchical model. With this in mind, we conduct an analysis of the relationship between health insurance coverage and citizenship status using data from the SIPP.

Title: Anopow for Replicated Nonstationary Time Series in Experiments

Speaker: Yu (Ryan) Yue, Department of Information Systems and Statistics, The City University of New York
Ph.D., University of Missouri - 2008

Abstract: We propose a novel analysis of power (ANOPOW) model for analyzing replicated nonstationary time series commonly encountered in experimental studies. Based on a locally stationary ANOPOW Cramér spectral representation, the proposed model can be used to compare the second-order time- varying frequency patterns among different groups of time series and to estimate group effects as functions of both time and frequency. Formulated in a Bayesian framework, independent two-dimensional second-order random walk (RW2D) priors are assumed on each of the time-varying functional effects for flexible and adaptive smoothing. A piecewise stationary approximation of the nonstationary time series is used to obtain localized estimates of time-varying spectra. Posterior distributions of the time-varying functional group effects are then obtained via integrated nested Laplace approximations (INLA) at a low computational cost. The large-sample distribution of local periodograms can be appropriately utilized to improve estimation accuracy since INLA allows modeling of data with various types of distributions. The usefulness of the proposed model is illustrated through two real data applications: analyses of seismic signals and pupil diameter time series in children with attention deficit hyperactivity disorder. Simulation studies and Supplementary Materials for this article are also available.

Keynote Presentation: Recent Developments for Binary Markov Random Fields

Speaker: Mark S. Kaiser, Department of Statistics, Iowa State University
Ph.D., University of Missouri - 1990

Abstract: Aside from the Gaussian Markov random field (CAR) model, Markov random fields formulated with conditional binary distributions are perhaps the most common of such models in

applications. In this talk we will provide an overview of three recent research directions for binary Markov random field models. In what are called Local Structure Graph Models, binary random fields are used to model the presence or absence of edges in a network. This type of model is illustrated with an analysis of international alliance formation between 1946 and 2000. In another direction, the problems of degeneracy and negative dependence are combined to illustrate the notion of conflicts in binary fields on regular lattices. A variety of limiting degeneracy patterns are produced by minimizing the number of conflicts in models with both positive and negative dependencies. Finally, we briefly examine a model that includes spatial dependence in a binary regression with a functional covariate process and illustrate the use of that model in an analysis of COVID vaccination rates in the Midwestern United States.

Afternoon Session

Title: A Density-Tilted Method for Data Integration of Multiple Heterogeneous Datasets with Structural Missingness
Speaker: Xiaokang Liu, Department of Statistics, University of Missouri

Abstract: In multicenter biomedical research, integrating data from multiple decentralized sites provides more accurate evidence and more generalizable findings. However, sharing individual-level data across sites is often prohibited due to privacy concerns. Many distributed algorithms, which fit a centralized model by only communicating aggregated information across sites, have been proposed to overcome the challenge of data sharing. A major challenge when applying existing distributed algorithms to real-world data is that the validity of the algorithm relies on the assumption that data across sites are independently and identically distributed, which is often violated in practice. In biomedical applications, data distributions across clinical sites can be substantially heterogeneous. Furthermore, the set of covariates available at each site can be different due to different data collection protocols. We propose a distributed inference framework for data integration in the presence of both distribution heterogeneity and data structural heterogeneity. By modeling heterogeneous and structurally missing data using density-tilted generalized method of moments, we develop a general distributed algorithm that is communication-efficient, privacy-preserving, and heterogeneity-aware. We establish the asymptotic properties of our estimator and demonstrate the validity of our methods in finite sample settings via simulation studies. We further apply our methods to identify risk factors for Alzheimer's disease using data from three Alzheimer's Disease Research Centers.

Title: Multivariate Cluster Point Process Model
Speaker: Suman Majumder, Department of Statistics, University of Missouri

Abstract: A common challenge in spatial statistics is to quantify the spatial distributions of clusters of objects. Frequently used approaches treat central object of each cluster as latent, but often cells of one or more types cluster around cells of another type. Quantifying these spatial relationships in biofilms may provide clues to disease pathogenesis. Even when clustering arrangements are not strictly parent-offspring relationships, treating the central object as a parent can enable use of parent-offspring clustering frameworks. We propose a novel multivariate spatial point process model to quantify multi-cellular arrangements with parent-offspring statistical approaches. We use the proposed model to analyze data from a human dental plaque biofilm image containing spatial locations of Streptococcus, Porphyromonas, Corynebacterium, and Pasteurellaceae, among other species and investigate any possible relationships between them. The proposed multivariate cluster point process (MCPP) model

departs from commonly used approaches in that it exploits the locations of the central parent object in clusters. It also accounts for possibly multilayered, multivariate parent-offspring clustering. In simulated datasets, the MCPP outperforms the classical Neyman-Scott process model. Applied to the motivating biofilm data, we quantified the simultaneous clustering of Streptococcus and Porphyromonas around Corynebacterium and of Pasteurellaceae around Streptococcus.

Title: A Novel Extreme Value Autoencoder Framework for Probabilistic Model Emulation and Calibration

Speaker: Likun Zhang, Department of Statistics, University of Missouri

Abstract: Large physics-based simulation models are crucial for understanding complex problems related to energy and the environment. These models are typically quite computationally expensive and there are numerous computational and uncertainty quantification (UQ) challenges when using these models in the context of calibration, inverse problems, UQ for forward simulations, and model parameterization. Surrogate model emulators have proven to be useful in recent years to facilitate UQ in these contexts, particularly when combined with Bayesian inference. However, traditional methods for model emulation such as Gaussian processes, polynomial chaos expansions, and more recently, neural networks and generative models do not naturally accommodate extreme values, which are increasingly relevant for many complex processes such as environmental impacts due to climate change and anomaly detection. Many statistical methods have been developed to flexibly model the simultaneous occurrences of extremal events, but most of them assume that the dependence structure of concurrent extremes is time invariant, which is unrealistic for physical processes that exhibit diffusive dynamics at short-time scales. We propose to develop a novel probabilistic statistical framework to explicitly accommodate concurrent and dependent extremes within a conditional variational autoencoder (CVAE) engine for enabling fast and efficient parameter estimation and uncertainty quantification. We also propose a new validation framework that is tailored to assess skill in fitting extreme behavior in model outputs. Our approach addresses, for the first time, the need to have efficient surrogate emulators of expensive simulation models that can accurately characterize, in a rigorous probabilistic manner, extreme values that are dependent in space and time and across processes.

Keynote Presentation: Analyzing Statistical Simulations: Providing Useful Information to Practitioners

Speaker: Doug Steinley, Department of Psychological Sciences, University of Missouri
Bachelor of Arts, University of Missouri - 2000

Abstract: Commonly, computational simulations are used to investigate properties of new statistical methods and/or to compare new and existing methods to each other. The wide-spread practice is to report performance of the investigated methods with average performance, both at the aggregate level and at the individual condition level of the simulation. Here, it is argued that most simulations can be viewed as a highly controlled experiment (in fact, in many ways, they are the idealized experiment) and the resultant data should be analyzed as such. With a focus on comparing competing methods for accomplishing the same task -- the example herein is rooted in cluster analysis -- it is argued that a three-tiered approach to analysis is taken: (a) methods focused on experimental design (such as factorial ANOVA), (b) accepted reporting of effect sizes, and (c) various statistical models for ranking.

Saturday, October 14, 2023

Morning Session

Keynote Presentation: Interpretable Sentiment Analysis Using the Attention-based Multiple Instance Classification Model

Speaker: Jing Cao, Department of Statistics and Data Science, Southern Methodist University
Ph.D., University of Missouri - 2005

Abstract: Sentiment analysis (SA) is widely used for analyzing text data to identify the underlying opinion or emotion expressed in a document. Neural network methods in natural language processing have produced state-of-the-art results in SA. However, many of those methods are “black-box” algorithms which don't provide interpretable results. In this paper, we aim to develop a word-level context-based method with accessibility and interpretability. Specifically, we propose an attention-based multiple instance classification model (AMIC). AMIC uses word embedding to transform text to data, employs a multiple instance classification structure, and incorporates the self-attention mechanism to include context information from document. The accessibility comes from the fact that AMIC has a transparent model structure, and it is easy to implement. The interpretability stems from AMIC's capability of providing word-level context-based sentiment metrics. In addition, AMIC can be used to construct domain-specific sentiment dictionary without requiring prior information on seed words or a pre-trained list of sentiment words. We demonstrate the performance of AMIC using a large online wine review dataset.

Title: Design Issues of Cluster Randomized Trials Which Evaluate Multiple Primary Endpoints

Speaker: Song Zhang, School of Public Health, University of Texas Southwestern Medical Center
Ph.D., University of Missouri - 2005

Abstract: Cluster randomized trials (CRTs) are widely used in different areas of medicine and public health. Recently, with increasing complexity of medical therapies and technological advances in monitoring multiple outcomes, many clinical trials attempt to evaluate multiple co-primary endpoints. In this study we present a power analysis method for CRTs with K>=2 binary co-primary endpoints. It is developed based on the GEE (generalized estimating equation) approach, and three types of correlations are considered: inter-subject correlation within each endpoint, intra-subject correlation across endpoints, and inter-subject correlation across endpoints. A closed-form joint distribution of the K test statistics is derived, which facilitates the evaluation of power and type I error for arbitrarily constructed hypotheses. We further present a theorem that characterizes the relationship between various correlations and testing power. We assess the performance of the proposed power analysis method based on extensive simulation studies. An application example to a real clinical trial is presented.

Title: scRAA: The Development of a Robust and Automatic Annotation Procedure for Single-Cell RNA Sequencing Data

Speaker: Dongyan Yan, Eli Lilly and Company
Ph.D., University of Missouri - 2019

Abstract: A critical task in single-cell RNA sequencing (scRNA-Seq) data analysis is to identify cell types from heterogeneous tissues. While the majority of classification methods demonstrated high performance in scRNA-Seq annotation problems, a robust and accurate solution is desired to generate reliable outcomes for downstream analyses, for instance, marker genes identification, differentially expressed genes, and pathway analysis. It is hard to establish a universally good metric. Thus, a universally good classification method for all kinds of scenarios does not exist. In addition, reference and query data in cell classification are usually from different experimental batches, and failure to consider batch effects may result in misleading conclusions. To overcome this bottleneck, we propose a robust ensemble approach to classify cells and utilize a batch correction method between reference and query data. We simulated four scenarios that comprise simple to complex batch effect and account for varying cell-type proportions. We further tested our approach on both lung and pancreas data. We found improved prediction accuracy and robust performance across simulation scenarios and real data. The incorporation of batch effect correction between reference and query, and the ensemble approach improve cell-type prediction accuracy while maintaining robustness. We demonstrated these through simulated and real scRNA-Seq data.

Title: Shared Consciousness through Bayer's Digital Twin in Breeding Crop Science

Speaker: Adam Gold, Bayer
Master of Arts, University of Missouri - 2012

Abstract: Bayer Crop Science has delivered industry leading seed and trait solutions for farmers for over 25 years. Precision Breeding through Data Science is leading the next wave of best-in-class innovation with the way we design, develop, and commercialize products. Our abundance of information in our digital pipeline is an asset giving us the competitive advantage in the market. Coordinating the global positioning of the next generation of seed in crop trials with complex weather, soil, and topographic environmental factors, with diverse sustainable management systems, creates a need for our innovative geospatial decision engine. Our global testing team is relying on us to make the best unified recommendations on how to develop the next generation of seed and trait solutions. High resolution market data spanning billions of acres is served as a dynamic, vivid, interactive model. Experts, scientists, and leaders can have peace of mind our data science is bringing a shared consciousness with our prescriptive testing program.

Title: Power Calculation for Cross-Sectional Stepped Wedge Cluster Randomized Trials with a Time-to-Event Endpoint

Speaker: Mary Ryan, Population Health Sciences, Biostatistics & Medical Informatics, University of Wisconsin – Madison
Bachelor of Science, University of Missouri - 2016

Abstract: Stepped wedge cluster randomized trials (SW-CRTs) are a form of randomized trial whereby clusters are progressively transitioned from the control to the intervention condition, and the timing of transition is randomized for each cluster. SW-CRTs are a common choice in implementation science, as all clusters will receive intervention by the end of the study and potentially limited study resources are not strained by implementing the intervention in many clusters simultaneously. An important task at the design stage is to ensure that the planned stepped wedge trial has sufficient power to observe clinically meaningful effects. While methods for determining study power have been developed for SW-CRTs with continuous and binary outcomes, limited methods for determining study power are available for stepped wedge

designs with censored time-to-event outcomes. In this presentation, I propose a stratified marginal Cox model to account for confounding by time in SW-CRTs, and derive the explicit expression of the robust sandwich variance to facilitate design calculations without the need for computationally intensive simulations. Power formulas based on both the Wald and robust score tests are analytically developed and compared via simulations, and demonstrate different finite-sample behaviors. Finally, I illustrate our methods using the context of a recently completed SW-CRT testing the effect of a new electronic reminder system on time to catheter removal in hospital settings.

Title: Partial Quantile Tensor Regression
Speaker: Dayu Sun, Department of Biostatistics and Health Data Science, Indiana University
Ph.D., University of Missouri - 2020

Abstract: Tensors, characterized as multidimensional arrays, are frequently encountered in modern scientific studies. Quantile regression has the unique capacity to explore how a tensor covariate influences different segments of the response distribution. In this work, we propose a partial quantile tensor regression (PQTR) framework, which novelly applies the core principle of the partial least squares technique to achieve effective dimension reduction for quantile regression with a tensor covariate. The proposed PQTR algorithm is computationally efficient and scalable to a large size tensor co-variate. Moreover, we uncover an appealing latent variable model representation for the new PQTR algorithm, justifying a simple population interpretation of the resulting estimator. We further investigate the connection of the PQTR procedure with an envelope quantile tensor regression (EQTR) model, which defines a general set of sparsity conditions tailored to quantile tensor regression. We prove the root-n consistency of the PQTR estimator under the EQTR model, and demonstrate its superior finite-sample performance compared to benchmark methods through simulation studies. We demonstrate the practical utility of the proposed method via an application to a neuroimaging study of post traumatic stress disorder (PTSD). Results derived from the proposed method are more neurobiologically meaningful and interpretable as compared to those from existing methods.

Afternoon Session

Title: Navigating the Financial Banking Industry: Statistical Methods, Regulatory Influences, and Climate Risk Integration
Speaker: Joel Miller, Rabo AgriFinance
Master of Arts, University of Missouri - 2005

Abstract: In today's dynamic financial banking industry, three critical factors are shaping its trajectory: statistical techniques, regulatory bodies like the Federal Reserve (Fed) and the European Central Bank (ECB), and the imperative to incorporate climate risk into risk management frameworks.

Statistical methods have become the backbone of decision-making within financial institutions. Advanced analytics, machine learning, and big data analytics are being harnessed to model complex financial instruments, optimize portfolios, and assess credit risks. These techniques enable banks to make data-driven decisions, minimize losses, and enhance profitability. Simultaneously, the influence of regulatory bodies such as the Fed and ECB is undeniable. These institutions play a pivotal role in shaping financial policies, ensuring stability, and preventing systemic risks. The regulatory landscape has evolved significantly since the global financial crisis of 2008. Stricter capital adequacy requirements, stress testing, and increased

transparency have become the norm. Furthermore, the financial sector is experiencing a growing push to incorporate climate risk into risk management frameworks. Climate change poses unique challenges, including physical risks from extreme weather events and transition risks associated with shifting to a low-carbon economy. Banks are recognizing the need to assess their exposure to climate-related risks and opportunities. In conclusion, successfully navigating this complex landscape requires a proactive approach that leverages advanced analytics, embraces regulatory changes, and incorporates climate risk into risk management practices. Failure to address these critical aspects could leave financial institutions vulnerable to a rapidly changing and increasingly challenging environment.

Title: Bayesian Estimation of Mixture Item Response Models
Speaker: Yanyan Sheng, Committee on Quantitative Methods, University of Chicago
Ph.D., University of Missouri - 2005

Abstract: Mixture item response theory (MixIRT) allows the presence of several latent classes that are qualitatively different but within which a conventional IRT model holds. The increase in the popularity of MixIRT models calls for an efficient estimation approach under the fully Bayesian framework via the use of Markov chain Monte Carlo (MCMC) techniques. Recent attention focuses on the no-U-turn sampler (NUTS), a non-random walk MCMC algorithm that converges to high dimensional target distributions relatively more quickly than conventional random walk MCMC algorithms. The focus of this talk is on applying NUTS to a mixture IRT model, and further to examine its performance in estimating model parameters and comparing it with conventional IRT models when factors such as sample size, test length, and number of latent classes are manipulated. The results indicate that overall, NUTS performs well in parameter recovery, and that fully Bayesian model selection methods (predictive information criteria) are suggested to be used together with their effective number of parameters in choosing the correct number of latent classes. Findings from this investigation provides empirical evidence on the performance of NUTS in fitting a specific form of the MixIRT model and suggest that researchers and practitioners in educational and psychological measurement should benefit from using NUTS for fitting complex IRT models.

Title: Semiparametric Regression of Mixed Panel Count Data with Informative Observation Processes
Speaker: Yang Li, Biostatistics & Health Data Science, Indiana University
Ph.D., University of Missouri - 2013

Abstract: In health and clinical research, mixed panel count data arise when study subjects are surveyed to report recurrent event occurrences between discrete observation times. Limited research has been established to analyze such a mixed complex data type, especially when the discrete observation times are potentially correlated with the underlying recurrent event process. In this talk, we consider regression analysis of mixed panel count data with informative observation processes. For the problem, a two-step estimation approach is proposed. An EM algorithm is developed for implementation, and the proposed estimators are shown to be consistent and asymptotically normal. An extensive simulation study is performed to assess the performance of the proposed approach and indicates that it works well in practical situations. An application to a set of real data is provided.

Title: Towards Efficient and Scalable Bayesian Inference: the Power of Variational Bayes and Quantum Computing

Speaker: Guohui Wu, Amgen
Ph.D., University of Missouri - 2014

Abstract: Bayesian inference via Markov chain Monte Carlo (MCMC) can be computationally demanding as the size of data and complexity of models grow, primarily due to the difficulties in drawing random samples from an (oftentimes) intractable target distribution and large number of MCMC iterations that are required. Faster alternatives to MCMC are variational Bayes methods which approximate a target distribution by solving an optimization problem that minimizes the Kullback-Leibler divergence from the target distribution to its variational Bayes approximation density. In this talk, three commonly used VB algorithms will be introduced and discussed: the mean field variational Bayes (MFVB), the fixed form variational Bayes (FFVB), and the integrated non-factorized variational Bayes (INFVB). Among these three algorithms, the MFVB algorithm assumes posterior independence and is often found to underestimate posterior variances, whereas the FFVB algorithm doesn't assume posterior independence but fixed parametric forms for VB approximation densities and often requires more involved stochastic optimization. Compared to the FFVB algorithm, the INFVB algorithm can also capture posterior dependence but it's more computationally appealing. By discretizing a subset of model parameters using a grid with finite grid points, the INFVB algorithm allows for parallelization. Yet there is a trade-off between controlling the dimension of model parameters to be discretized and maintaining computational tractability. Quantum computing can further improve computational efficiency of the INFVB algorithm and address this trade-off. Examples will be provided to demonstrate the effectiveness of variational Bayes methods and quantum computing.

Keynote Presentation: Running On Empty: Recharge Dynamics from Animal Movement Data

Speaker: Mevin Hooten, Department of Statistics & Data Science, University of Texas at Austin
Ph.D., University of Missouri - 2006

Abstract: Vital rates such as survival and recruitment have always been important in the study of population and community ecology. At the individual level for wild animals, physiological processes such as energetics are critical in understanding biomechanics and movement ecology and also scale up to influence food webs. Commonly used statistical models for telemetry data lack an explicit, mechanistic connection to physiological dynamics. I describe a framework for modeling telemetry data that includes an aggregated physiological process associated with decision making and movement in heterogeneous environments. This framework accommodates a wide range of movement and physiological process specifications. As an example, I present a model formulation for animal trajectories in continuous time that provides direct inference about gains and losses associated with physiological processes based on movement. The approach can also be extended to accommodate auxiliary data when available. I demonstrate the model to infer wild felid and ungulate recharge dynamics.