Featured Talks

Understanding the origin and spread of COVID-19

Liang Liu
Department of Statistics
University of Georgia
Collaborators/co-authors: Jonathan Arnold, Justine Bahl, Pengsheng Ji

Phylogenetic trees are fundamental tools for understanding the origin and spread of COVID-19. Using coalescent theory, we reconstructed a species tree from 11 genes of human, bat, and pangolin beta coronaviruses. Each gene tree reflects different histories of selection, gene flow, recombination and lineage sorting as the genes move across species boundaries. To resolve these discordances a species tree was reconstructed from the 11 gene trees using coalescent methods. The shallow species tree provides evidence of recent gene flow events between bat and pangolin beta coronaviruses predating the zoonotic transfer to humans. The species tree was also used to reconstruct the ancestral sequence of Human-SARS CoV-2, which was 2 nucleotides different from the Wuhan (WH01) sequence. The time to most recent common ancestor (tMRCA) was estimated to be Dec 8, 2019 with a bat (RaTG13) origin. The species tree is a product of evolutionary factors, providing evidence of repeated zoonotic transfers between bat and pangolin as a reservoir for future zoonotic transfers to humans. In addition, a transmission map was consructed from the species tree to illustrate the global spread of COVID-19.

Unsupervised learning in data integration studies using JIVE with Gaussian mixtures

Benjamin B. Risk
Department of Biostatistics and Bioinformatics
Rollins School of Public Health, Emory University
Collaborators/co-authors: Ganzhong (Gavin) Tian, Raphiel Murden, James Lah, John Hanfelt

A common goal in data integration studies is to identify subgroups. JIVE (joint and individual variation explained) has been proposed as a method to extract shared (joint) and unique (individual) information from each dataset, and cluster analysis is applied after extraction of joint and individual scores. We present a probabilistic JIVE model with mixture of Gaussians (JIVE-mix) that enables joint probabilistic clustering of subjects with multiple data sources. Our simulations demonstrate improvement over existing approaches. We apply our method to MRI brain imaging and CSF biomarker measurements in the Alzheimer's Disease Neuroimaging Initiative, which reveals interesting clusters that suggest distinct pathologies.

Invertible graph neural networks

Yao Xie
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
Collaborators/co-authors: Chen Xu, Xiuyuan Cheng

Inverse problem on graph data has various applications in many fields of study and applications, which considers the problem to infer the input data X given outcome labels Y, where both X and Y are defined on each node on a graph. The problem can be viewed as reversing the node prediction problem on a graph, and since the mapping from Y to X is one-to-many, it can be formulated as a conditional generative task. Such a task is important for multiple modern applications, such as inferring the predictive set (going beyond point prediction), data-driven Bayesian posterior modeling, graph prediction and design problems in protein networks, and spatio-temporal prediction for graph neural networks. We propose a model of invertible graph neural networks to address the problem, where an invertible normalizing flow network is used to construct a one-to-one mapping from X to an intermediate feature H, and then a classification network is used to map H to Y. The expressiveness of graph convolution layers is analyzed in the context of the problem and supported by experiments. In computation, we introduce Wasserstein-2 regularization in the training of the flow network. We will also discuss new designs of the invertible flow network based on Wasserstein gradient flow. This is joint work with Chen Xu at Georgia Institute of Technology and Xiuyuan Cheng at Duke University.

Technical Sessions

Feature selection: the Markov boundary approach

Anwesha Bhattacharyya
Wells Fargo N.A.
Collaborators/co-authors: Yaqun Wang, Joel Vaughan, Vijay Nair

Machine learning models offer the promise of increased predictive performance by being able to incorporate information from large numbers of features. However, some of the features are typically highly correlated which can lead to model instability, lack of generalizability, and challenges in interpretation. Ideally, one would like to use causality principles select the important features, but this is a very challenging problem with observational data. This presentation deals with a related approach for feature selection called Markov blanket. We describe the approach and outline some common problems associated with identifying a Markov blanket in structured data. We also propose a forward backward framework to tackle the challenges and demonstrate the results on simulated and real datasets.

Online Bayesian phylodynamic inference

Mandev S. Gill
Department of Statistics
University of Georgia
Collaborators/co-authors: Philippe Lemey, Marc A. Suchard, Andrew Rambaut, Guy Baele

Phylodynamic inference provides a framework to reconstruct evolutionary and epidemiological dynamics of rapidly evolving pathogens. Importantly, phylodynamic analyses can provide insights into unobserved events and processes that shape epidemic dynamics that are not obtainable through any other methods. Advances in sequencing technology enable real-time genomic surveillance as an outbreak unfolds, but widely-used Bayesian phylogenetic inference packages are not designed to accommodate the resulting continuous stream of new data. We introduce a framework for "online" Bayesian phylodynamic inference that can efficiently incorporate newly available data into existing analyses. We analyze data from the West African Ebola virus epidemic and demonstrate a considerable reduction in time required to obtain updated posterior inferences at different time points of the epidemic.

On the testing of statistical software

Ryan Lekivetz
JMP
Collaborators/co-authors: Joseph Morgan

Testing statistical software is an extremely difficult task. For many statistical packages, the development and testing are done by the same individual, who may not have formal training in software testing techniques and have limited time for testing. This makes it imperative that the adopted testing approach is both efficient and effective and, at the same time, it should be based on principles that are readily understood by the developer. As it turns out, the construction of test cases can be thought of as a designed experiment (DOE). This talk discusses how familiar DOE principles can be applied to testing statistical software.

Penalized weighted proportional hazards model for robust variable selection and outlier detection

Bin Luo
Department of Biostatistics and Bioinformatics
Duke University
Collaborators/co-authors: Xiaoli Gao, Susan Halabi

Identifying exceptional responders or non-responders is an area of increased research interest in precision medicine as these patients may have different biological or molecular features and therefore may respond differently to therapies. Our motivation stems from a real example from a clinical trial where we are interested in characterizing exceptional prostate cancer responders. We investigate the outlier detection and robust regression problem in the sparse proportional hazards model for censored survival outcomes. The main idea is to model the irregularity of each observation by assigning an individual weight to the hazard function. By applying a LASSO-type penalty on both the model parameters and the log transformation of the weight vector, our proposed method is able to perform variable selection and outlier detection simultaneously. The optimization problem can be transformed to a typical penalized maximum partial likelihood problem and thus it is easy to implement. We further extend the proposed method to deal with the potential outlier masking problem caused by censored outcomes. The performance of the proposed estimator is demonstrated with extensive simulation studies and real data analyses in low-dimensional and high-dimensional settings.

Gaussian process subspace prediction for model reduction

Simon Mak
Department of Statistical Science
Duke University
Collaborators/co-authors: Ruda Zhang, David Dunson

Subspace-valued functions arise in a wide range of problems, including parametric reduced order modeling (PROM). In PROM, each design parameter is typically associated with a subspace response, which is used for Petrov-Galerkin projections of large system matrices. Previous efforts to approximate such functions use deterministic interpolation methods on manifolds, which are inflexible and yield no uncertainty quantification. To tackle this, we propose a novel Bayesian nonparametric model for subspace prediction: the Gaussian Process Subspace regression (GPS) model. This model is extrinsic and intrinsic at the same time: with multivariate Gaussian distributions on the Euclidean space, it induces a joint probability model on the Grassmann manifold, the set of fixed-dimensional subspaces. The GPS adopts a simple yet general correlation structure, and a principled approach for model selection. Its predictive distribution admits an analytical form, which allows for efficient subspace prediction over the parameter space. We provide a suite of numerical simulations and applications which demonstrates the effectiveness of the proposed GPS model over existing subspace interpolation approaches.

Canonical joint and individual variation explained

Raphiel J. Murden
Department of Biostatistics and Bioinformatics
Rollins School of Public Health, Emory University
Collaborators/co-authors: Zhengwu Zhang, Ying Guo, Benjamin Risk

Joint and Individual Variation Explained (JIVE) is a model that decomposes multiple datasets obtained on the same subjects into shared structure, structure unique to each dataset, and noise. JIVE is an important tool for multimodal data integration in neuroimaging. The two most common algorithms are R.JIVE, an iterative approach, and AJIVE, which uses principal angle analysis. The joint structure in JIVE is defined by shared subspaces, but interpreting these subspaces can be challenging. In this paper, we reinterpret AJIVE as a canonical correlation analysis of principal component scores. This reformulation, which we call CJIVE, 1) provides an intuitive view of AJIVE; 2) uses a permutation test for the number of joint components; 3) can be used to predict subject scores for out-of-sample observations; and 4) is computationally fast. We conduct simulation studies that show CJIVE and AJIVE are accurate when the total signal ranks are correctly specified but, generally inaccurate when the total ranks are too large. CJIVE and AJIVE can still extract joint signal even when the joint signal variance is relatively small. JIVE methods are applied to integrate functional connectivity (resting-state fMRI) and structural connectivity (diffusion MRI) from the Human Connectome Project. Surprisingly, the edges with largest loadings in the joint component in functional connectivity do not coincide with the same edges in the structural connectivity, indicating more complex patterns than assumed in spatial priors. Using

these loadings, we accurately predict joint subject scores in new participants. We also find joint scores are associated with fluid intelligence, highlighting the potential for JIVE to reveal important shared structure.

Causal and counterfactual views of missing data models

Razieh Nabi
Department of Biostatistics and Bioinformatics
Rollins School of Public Health, Emory University
Collaborators/co-authors: Rohit Bhattacharya, Ilya Shpitser, James Robins

It is often said that the fundamental problem of causal inference is a missing data problem – the comparison of responses to two hypothetical treatment assignments is made difficult because for every experimental unit only one potential response is observed. In this talk, we consider the implications of the converse view: that missing data problems are a form of causal inference. We make explicit how the missing data problem of recovering the complete data law from the observed data law can be viewed as identification of a joint distribution over counterfactual variables corresponding to values had we (possibly contrary to fact) been able to observe them. Drawing analogies with causal inference, we show how identification assumptions in missing data can be encoded in terms of graphical models defined over counterfactual and observed variables. We note interesting similarities and differences between missing data and causal inference theories. The validity of identification and estimation results using such techniques rely on the assumptions encoded by the graph holding true. Thus, we also provide new insights on the testable implications of a few common classes of missing data models, and design goodness-of-fit tests around them. For relevant papers see: (i) Full Law Identification In Graphical Models Of Missing Data: Completeness Results (ICML 2020), (ii) Identification In Missing Data Models Represented By Directed Acyclic Graphs (UAI 2019), and (iii) On Testability and Goodness of Fit Tests in Missing Data Models (Preprint 2022).

Statistical methods in risk-stratified disease prevention with applications in cancer and health disparities

Parichoy Pal Choudhury
Departments of Surveillance and Health Equity Science and Population Science
American Cancer Society, Atlanta

Risk-stratified disease prevention involves tailoring of health decisions about screening and prevention based on the individualized risk predictions. This requires a comprehensive understanding of the risk factors, including genetic variants, biomarkers, lifestyle/behavioral and environmental factors leading to the development of a model for predicting absolute risk of a disease of interest. Absolute risk model development requires information on relative risks of the risk factors, population-based age-specific disease incidence rates and competing event rates and population

distributions of the risk factors. Such a model needs to be validated ideally in independent prospective cohorts before clinical applications. In this talk, I will describe a software tool for implementing absolute risk estimation of a disease integrating multiple data sources leveraging the best information available for each of the input parameters and standardized approaches for risk model validation. I will describe a major recent application of this tool in the development and validation of a comprehensive risk prediction model for breast cancer and its biologically heterogeneous subtypes based on estrogen receptor status. Model validation in two-phase study settings often involve scenarios where expensive biomarkers (e.g., polygenic risk score or PRS) are measured in smaller subsample of a prospective cohort, where subjects may be selected using complex sampling designs. I will describe a simple method for improving precisions of model validation statistics (e.g., AUC) using the partial risk factors from the full cohort and complete risk factors from the subsample. I will show an application in breast cancer risk prediction with questionnaire-based risk factors and PRS. I will also present initial findings from a recent study that investigates the contributions of access to care (e.g., health insurance coverage) in explaining racial disparities in stage of diagnosis of multiple cancers detectable by screening or clinical symptoms.

Image-based feedback control using tensor analysis

Kamran Paynabar
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
Collaborators/co-authors: Zhen Zhong, Jianjun Jan Shi

In manufacturing systems, many quality measurements are in the form of images, including overlay measurements in the semiconductor manufacturing and dimensional deformation profiles of fuselages in an aircraft assembly process. To reduce the process variability and ensure on-target quality, process control strategies should be deployed, in which the high-dimensional image output is controlled by one or more input variables. To design an effective control strategy, the process model off-line should be first estimated via relationship exploration between the image output and inputs. Next, the control law is formulated by minimizing the control objective function online. The main challenges of achieving such a control strategy include (i) the high dimensional output of a regression model, (ii) the integrated analysis of both the spatial structure of image outputs and the temporal structure of the image sequence, and (iii) non-i.i.d. noises. To address these challenges, we propose a novel tensor-based process control approach by incorporating the tensor time series and regression techniques. Based on the process model, we can then obtain the control law by minimizing a control objective function. Although our proposed approach is motivated by the 2D image case, it can be extended to higher-order tensors such as point clouds. Simulation and case studies show that our proposed method is more effective than benchmarks in terms of relative mean square error.

Optimal transport-based transfer learning for smart manufacturing

Rui Xie
Department of Statistics and Data Science
University of Central Florida
Collaborators/co-authors: Dazhong Wu

Various machine learning-based predictive modeling approaches to tool wear prediction have been introduced over the past few years. However, predicting tool wear under different operating conditions (e.g., depth of cut, feed rate, and workpiece material) with small datasets remains a challenge due to complex tool wear mechanisms. To address this issue, an optimal transport (OT)-based transfer learning algorithm is developed to transfer knowledge on tool wear from one operating condition to another. The OT-based transfer learning model has been demonstrated on a small dataset collected under different operating conditions. Experimental results have shown that the OT-based transfer learning method significantly improved tool wear prediction accuracy.

Big-data infectious disease estimation in COVID-19

Shihao Yang
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
Collaborators/co-authors: S. Ma, S. Er, S. Zhu, A. Bukharin, L. Xie, M. Santillana, S. C. Kou, P. Keskinocak, T. Zhao, Y. Xie

For epidemic control and prevention, timely insights into potential hot spots are invaluable. As an alternative to traditional epidemic surveillance, big data from the Internet could provide important information about the current epidemic trends after proper spatial-temporal modeling. This talk will present a few data-driven statistic/machine learning approaches for infectious disease prediction, with special focus on COVID-19.

Interactions among acute respiratory viruses in urban China, 2009 – 2019

Yang Yang
Department of Statistics
University of Georgia
Collaborators/co-authors: Zachary J. Madewell, Natalie E. Dean, Ira M. Longini, Li-Qun Fang

Background: A viral infection can modify the risk to subsequent viral infections via cross-protective immunity, increased immunopathology, or disease-driven behavioural change. There is limited understanding of virus-virus interactions due to lack of long-term population-level data.

Methods: Our study leverages passive surveillance data of ten human acute respiratory viruses from Beijing, Chongqing, Guangzhou, and Shanghai collected during 2009-2019: influenza A and B viruses (IAV and IBV); respiratory syncytial virus A and B (RSV-A and RSV-B); human parainfluenza virus (HPIV), adenovirus (HAdV), metapneumovirus (HMPV), coronavirus (HCoV), bocavirus (HBoV), and rhinovirus (HRV). We used a Bayesian hierarchical model to evaluate correlations in monthly prevalence of test-positive samples between virus pairs, accounting for sparse testing and autocorrelation.

Results: There were 101,643 lab-tested patients of whom 33,650 tested positive for any acute respiratory virus and 4,113 were co-infected with more than one virus. HPIV/HRV and HPIV/HCoV were positively correlated in all cities in unadjusted analyses. After adjusting for intrinsic seasonality, long-term trends and multiple comparisons, we found strong evidence for positive correlations between HPIV/HRV in all four cities and HBoV/HRV and HBoV/HMPV in three cities. Results for children revealed positive associations of HPIV/HRV, IBV/RSV-A, RSV-A/HCoV, RSV-B/HPIV, RSV-B/HMPV, RSV-B/HRV, HPIV/HMPV, HPIV/HCoV, HPIV/HBoV, HAdV/HBoV, and HBoV/HRV, and negative associations of IAV/HAdV.

Interpretation: There were strong interactions among common respiratory viruses in highly populated urban settings, particularly among children. Such interactions necessitate more studies on joint surveillance and prevention strategies for more effective control of these viruses.

Locally optimal design for A/B tests in the presence of covariates and network dependence

Qiong Zhang
School of Mathematical and Statistical Sciences
Clemson University
Collaborators/co-authors: Lulu Kang

A/B test, a simple type of controlled experiment, refers to the statistical procedure of experimenting to compare two treatments applied to test subjects. In this talk, we assume that the test subjects of the experiments, are connected on an undirected network, and the responses of two connected users are correlated. We include the treatment assignment, covariate features, and network connection in a conditional autoregressive model and propose a design criterion that measures the variance of the estimated treatment effect. A hybrid optimization approach is proposed to obtain the optimal design based on this cirterion. Through synthetic and real social network examples, we demonstrate the value of including network dependence in designing A/B experiments and validate that the proposed locally optimal design is robust to the choices of parameters.

Quickest detection of the change of community via stochastic block models

Ruizhi Zhang
Department of Statistics
University of Georgia
Collaborators/co-authors: Fei Sha

Community detection is a fundamental problem in network analysis and has important applications in sensor networks and social networks. In many cases, the community structure of the network may change at some unknown time and thus it is desirable to come up with efficient monitoring procedures that can detect the change as quickly as possible. In this work, we use the Erdős-Rényi model and the bisection stochastic block model (SBM) to model the pre-change and post-change distributions of the network, respectively. That is, initially, we assume there is no community in the network. However, at some unknown time, a change occurs, and two communities are formed in the network. We then propose an efficient monitoring procedure by using the number of k-cycles in the graph. The asymptotic detection properties of our proposed procedure are derived when all parameters are known. A generalized likelihood ratio (GLR) type detection procedure and an adaptive CUSUM type detection procedure are constructed to address the problem when parameters are unknown.

Industry Session

What is it like to work at JMP/SAS?

Ryan Lekivetz
JMP, a SAS company

For more than 30 years, we've been making JMP statistical discovery software tailored to the needs of scientists and engineers. John Sall, SAS co-founder and Executive Vice President, is also the creator of JMP. What started out as Sall's passion project has grown – by leaps and bounds – into a family of statistical software products that are used worldwide in nearly every industry.

We say that great software in the right hands can change the world. We say it because we've seen it. We've seen scientists and engineers use JMP to speed new drugs to market, to design better products and processes, to figure out how to restore ecosystems. You get the idea. Advancements are made when brilliant people use JMP statistical discovery software to see what they've not seen before.

You've probably used a JMP or SAS product at some point in your graduate degree. Have you ever wondered what it's like to work for SAS? I'll give you my own perspective as a statistician working at SAS and discuss the diversity of jobs in the organization for individuals with quantitative backgrounds.

Data science at State Farm

Megan Lutz
State Farm

At State Farm, we celebrated our 100th anniversary in 2022, and over that time we have seen our analytics function grow and expand. State Farm has 150 analytics professionals, including 80 data scientists. Advanced analytics supports the enterprise both in customer-facing analytics and as in-house consultants for teams as diverse as Human Resources and Claims. State Farm is the number one automobile and homeowner insurance company in the US and has been for over 60 years. We use data science, machine learning, and AI to identify new solutions and maintain our competitive advantage. Our Modeling and Analytics Graduate Network (MAGNet) programs, located in Athens, GA and Champaign, IL, are proven pipelines for developing new, full-time data scientists. We will discuss the MAGNet program and full time analytics work at State Farm during this presentation.

Natural language processing in banking

Rahul Singh
Wells Fargo N.A.

The use of Natural Language Processing (NLP) is increasingly becoming popular in banking and finance. In this presentation, I will provide an overview of research in our team, encompassing applications to text and sentiment classification, chatbots, conversational AI, , named entity recognition, and topic modeling. Another important component of our research is developing novel diagnostic techniques to assess model weakness and to provide explainability for model decisions. If time permits, other topics including model robustness, knowledge distillation, paraphrasing, model explainability will also be covered.

Posters

Exploration of integrating gene expression and spatial information based on multiple graph-based deep learning methods

Jiazhang Cai
Department of Statistics
University of Georgia
Collaborators/co-authors: Huimin Cheng, Guocheng Yuan, Ping Ma, Wenxuan Zhong

With the development of sequencing technology, more information is available which can provide us with a deep understanding of the cell's mechanism. Recently, spatial information can be observed together with gene expression information in many experiments, e.g., Multiplexed Error- Robust Fluorescence ISH (MERFISH) and sequential fluorescence ISH (seqFISH). Since the actual location of cells does not have a direct relationship with cell types, how to exploit spatial information has become a challenge. In this poster, we introduce a new method that takes advantage of multiple graph-based deep learning methods, including spaGCN and SpaceFlow, through WNN. Compared with using gene expression information only, the proposed method can largely recover the structure of the tissue and perform better in cell clustering. It also provides better insight for the subsequent analysis, such as the trajectory inference.

Evaluating cell type specific marker genes' characteristics from population-level single-cell RNA-seq

Luxiao Chen
Department of Biostatistics and Bioinformatics
Rollins School of Public Health, Emory University
Collaborators/co-authors: Hao Wu

Single cell RNA-seq (scRNA-seq), a technique helping scientists study samples with extremely high resolutions, provides gene expression profile of each cell in a sample. In recent years, a rapid increasing research interest to apply this technique at population level and its decreasing expense facilitate appearing of a bunches of datasets containing multiple samples measured by scRNA-seq. Our real data exploration suggests that in scRNA-seq data, cell type specific marker (cs-marker) genes may not consistently appear across all samples. However, cs-marker genes applied in analyses like cell typing or deconvolution of bulk samples, which rely on scRNA-seq data as reference, are expected to be consistent across samples. Motivated by this observation, we first applied a statistical model to identify cs-marker genes that consistently appear in historical population-level scRNA-seq data. We then designed a strategy to incorporate these consistent cs-marker genes identified from historical data into analyses like cell-typing or deconvolution of bulk samples. Extensive data analyses demonstrate that applying consistent cs-marker genes in analyses can help to improve accuracy of cell typing or bulk sample deconvolution.

Informative node selection for graph and scalar association study

Yongkai Chen
Department of Statistics
University of Georgia
Collaborators/co-authors: Ping Ma, Wenxuan Zhong

In many studies, we often collect replicated samples of a graph and its attribute. Current methods for studying the association between the graph and its attribute suffer from high model complexity, low interpretability, and high computational cost. To address this issue, we consider a sparse dependence between the graph and its attribute. In particular, we define a novel concept of the *informative node* to characterize this sparse dependence. To identify the informative nodes, we develop a parsimonious yet representative subgraph extraction method using the Wasserstein distance. Since our objective function is non-convex, we implement a highly efficient stepwise selection algorithm via splitting and shuffling. The proposed informative node selection method, named (INS), is statistically efficient and easy to implement. The empirical performance of INS has been carefully assessed through extensive simulations and real experiments.

Graphon convolutional network: A highly efficient learner for random graph

Huimin Cheng
Department of Statistics
University of Georgia
Collaborators/co-authors: Shushan Wu, Jun Yu, Haoran Lu, Wenxuan Zhong

Classic graph convolutional network (GCN) uses the Laplacian matrix of a given graph as the kernel matrix to train the node classifier. This approach assumes that the given graph is error-free. However, for a dense graph or a graph with a large number of nodes, we usually assume an edge between two nodes is sampled from a Bernoulli distribution whose mean is generated by a graphon function f. Under this assumption, the observed graph can be biased since some edges may not be observed. Thus, classifiers obtained by overlapping neighboring nodes around the origin node can be biased as the neighborhood defined by the graph is no longer error-free. To overcome this challenge, we propose the graphon convolutional network, which replaces the kernel matrix with a graphon estimation. As graphon is the limit of a graph, when the number of nodes n goes to infinity, it is less susceptible to encountering connectivity errors. The superiority of the proposed method is demonstrated by various synthetic and real experiments, especially those large and dense graphs.

Was there any widespread fraud in 2020 presidential election? What does Benford's Law say?

Deeya Datta
Department of Statistics
University of Georgia

Fair elections free of any interference are integral tenets of any functioning democracy, and widespread election fraud is undoubtedly a serious threat to a free republic. While instances of electoral fraud are much more prevalent in countries with illiberal democracies, the U.S has recently faced such an accusation. Although he was unable to provide any concrete evidence, the former U.S. President Donald Trump accused his opponent, Joe Biden, now president, of electoral fraud after the presidential election. Fortunately, election forensics are often successful in investigating the validity of such fraud allegations. In this paper, I applied Benford's law, a rule that should stand up to any large set of natural numbers, such as un-tampered electoral data. Using this law and basic statistical analysis of votes of U.S. counties for candidates of the two major parties, I completed a forensic analysis to investigate Mr. Trump's allegation. My comprehensive investigation does not find any evidence supporting his allegation.

Spatio-temporal point processes with deep non-stationary kernels

Zheng Dong
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
Collaborators/co-authors: Xiuyuan Cheng, Yao Xie

Deep neural networks, especially recurrent neural network (RNN) models, have become a popular tool for analyzing point process data. Despite the powerful expressiveness and memorizing ability of RNN models, they may not successfully model sophisticated non-stationary dependencies among data due to the recurrent structure. Meanwhile, another type of deep model for point process data was recently proposed, which represents the influence kernel rather than the intensity function by neural networks. This paper develops a deep non-stationary influence kernel for spatio-temporal point processes with a novel parameterization that enables us to well approximate complicated kernels in a low-rank form. A log-barrier penalty is introduced during network optimization to maintain the non-negativity of conditional intensity. Our new method can also be extended to model high-dimensional marks, and we demonstrate outstanding performance gain on real police text data. The new approach significantly reduces the model and computational complexities, and the benefits of kernel recovery and event prediction are demonstrated using synthetic and real point process data.

An efficient pseudo-likelihood estimator for DNA sequence evolution

Gregory Ellison
Department of Statistics
University of Georgia
Collaborators/co-authors: Liang Liu

The Generalized Time Reversible (GTR) model provides a probabilistic framework for DNA sequence evolution, and is useful in statistical inference of phylogenies based on DNA sequence data. However, the maximum likelihood approach to estimating GTR model parameters also depends on the topology of the phylogenetic tree, which in practice requires computation over a large search space of possible tree topologies. We consider a pseudo-likelihood estimator of the GTR model parameters, which allows for efficient estimation of the GTR model parameters by ignoring the tree topology.

OTSFAL: An active learning framework for deep neural network

Luyang Fang
Department of Statistics
University of Georgia
Collaborators/co-authors: Cheng Meng, Lin Zhao, Wenxuan Zhong, Tianming Liu, Ping Ma

The remarkable achievements of recent super-large deep neural networks rely on a large number of the labeled training dataset, which is extremely or even impossible to obtain. Therefore, an essential problem is how do we get the model to achieve desired outputs with only a limited amount of labeled data. One solution to this is active learning (AL), where a model asks an oracle to label a subset of the unlabeled pool, and then adds the newly labeled subset to the training set to train the model refresh. In this paper, we propose a novel active learning framework (OTSFAL) combining optimal transport and space-filling to select the subset such that it can best represent the distribution of the entire data pool. And therefore, we can make the best use of the information contained in the data pool and improve the current model accordingly. Our OTSFAL algorithm advances the existing AL methods in two aspects. First, OTSFAL selects a subset of data that can best represent the distribution of the whole data pool, which avoids serious sample bias and undesirable performance over the entire data pool. Second, the proposed algorithm can be applied with any learning model and different tasks, rather than being designed for a specific task. Extensive empirical evaluations on challenge image datasets such as CIFAR 10 also demonstrate that the proposed algorithm can achieve promising results.

Tailoring capture-recapture methods to estimate registry-based case counts based on error-prone diagnostic signals

Lin Ge
Department of Biostatistics and Bioinformatics
Rollins School of Public Health, Emory University
Collaborators/co-authors: Yuzi Zhang, Kevin C. Ward, Timothy L. Lash, Lance A. Waller, Robert H. Lyles

Surveillance research is of great importance for effective and efficient epidemiological monitoring of case counts and disease prevalence. Taking specific motivation from ongoing efforts to identify recurrent cases based on the Georgia Cancer Registry, we extend recently proposed "anchor stream" sampling design and estimation methodology. Our approach offers a more efficient and defensible alternative to traditional capture-recapture (CRC) methods by leveraging a relatively small random sample of participants whose recurrence status is obtained through a principled application of medical records abstraction with one or more existing signaling data streams, some of which may be, and often likely are, non-representative of the full registry population. The key extension developed here accounts for the common problem of false positive or negative diagnostic signals from one or more of the existing data streams. In particular, we show that the design only requires documentation of positive signals in the non-anchor surveillance streams, and permits valid estimation of the true case count based on an estimable positive predictive value (PPV) parameter. We borrow ideas from the multiple imputation paradigm to provide accompanying standard errors, and develop a Bayesian credible interval approach that yields favorable frequentist coverage properties. We demonstrate the benefits of the proposed methods through simulation studies, and provide a data example targeting estimation of the breast cancer recurrence case count among Metro Atlanta area patients from the Georgia Cancer Registry-based Cancer Recurrence Information and Surveillance Program (CRISP) database.

A practical revealed preference model for separating preferences and availability effects in marriage formation

Shuchi Goyal
Department of Statistics
University of California, Los Angeles
Collaborators/co-authors: Mark S. Handcock, Heide Marie Jackson, Michael S. Rendall, Fiona Y. Yeung

Many problems in demography require models for partnership formation that separate latent preferences for partners from the availability of partners. We consider a model for matchings within a bipartite population where individuals have utility for people based on known and unknown characteristics. People can form a partnership or remain unpartnered. The model represents both the availability of potential partners of different types and preferences of individuals for such people. We develop Menzel's (2015) framework to estimate preference parameters based on sample survey data on partnerships and population composition. We conduct simulation studies based

on new marriages observed in the Survey of Income and Program Participation (SIPP) to show that, for realistic population sizes, the model recovers preference parameters that are invariant under different population availabilities. We also develop bias correction of parameters for small population sizes and confidence intervals for estimates that have correct coverage. This model can be applied in family demography to understand individual preferences given different availabilities.

Sampled-boosting regression transfer for atmospheric pollution prediction

Shrey Gupta
Department of Computer Science
Emory University
Collaborators/co-authors: Jianzhao Bi, Yang Liu, Avani Wildani

The unprecedented trends in global climate change can be attributed to the rise in atmospheric pollution among other major factors. The atmospheric pollution is often measured using *Particulate Matter 2.5* levels (PM 2.5), and requires the installation of costly equipment every few kilometers for a successful prediction of PM 2.5. However, installation of such useful atmospheric pollution prediction equipment often gets neglected in the countries across the world especially under-developed regions. Hence, these regions suffer from equipment-based data deficiency. In our research, we apply knowledge transfer methodologies that utilize data from data-rich regions like the United States and Western Europe and adapt it for PM 2.5 modeling for data-scarce regions of the world. The focus of our model is generalization and uniformity across multi-modal distributions commonly observed in spatio-temporal pollution data. We achieve this by performing instance-transfer domain-adaptation with a keen focus on measuring the complexity in order to peek into the behavior of the distribution. Our incrementally novel, regression-transfer boosting methodology performs better than the competitive transfer learning methodologies 63% of the time, as well as displays consistency in its performance as opposed to the fluctuating performance of other methodologies.

Overall ranking of populations using Bayesian methods

Yiren Hou
Department of Statistics
University of Georgia
Collaborators/co-authors: Gauri Datta, Abhyuday Mandal

Methods that provide direct measure of uncertainty for the estimated overall ranking of populations are less available compared to measure of uncertainty in individual ranks. The direct measure of the uncertainty in the estimated overall ranking would involve all the populations simultaneously and their relative standing to each other, instead of considering ranking of populations based on a real-valued parameter. Motivated by the construction of a joint confidence region for an overall ranking of populations by Klein, Wright, and Wieczorek, a Bayesian method is developed and

applied to the same dataset on mean travel time to work for states in 2011 collected by the US Census Bureau. A joint credible region for the true unknown overall ranking is constructed, and it is more informative than the joint confidence region developed from the frequentist perspective. There is significant reduction in the volume of the "elliptical" credible set for the vector of means over the "rectangular" frequentist confidence set. Additionally, the Bayesian method provides probabilities of various rankings a state may have based on its mean and the probabilities that various states may occupy a particular ranking based on their means.

Constrained minimum energy designs

Chaofan Huang
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
Collaborators/co-authors: V. Roshan Joseph, Douglas M. Ray

Space-filling designs are important in computer experiments, which are critical for building a cheap surrogate model that adequately approximates an expensive computer code. Many design construction techniques in the existing literature are only applicable for rectangular bounded space, but in real world applications, the input space can often be non-rectangular because of constraints on the input variables. One solution to generate designs in a constrained space is to first generate uniformly distributed samples in the feasible region, and then use them as the candidate set to construct the designs. Sequentially Constrained Monte Carlo (SCMC) is the state-of-the-art technique for candidate generation, but it still requires large number of constraint evaluations, which is problematic especially when the constraints are expensive to evaluate. Thus, to reduce constraint evaluations and improve efficiency, we propose the Constrained Minimum Energy Design (CoMinED) that utilizes recent advances in deterministic sampling methods. Extensive simulation results on 15 benchmark problems with dimensions ranging from 2 to 13 are provided for demonstrating the improved performance of CoMinED over the existing methods.

Police text analysis: Topic modeling and spatial relative density estimation

Sarah Huestis-Mitchell
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
Collaborators/co-authors: Xiuyuan Cheng, Yao Xie

We analyze a large corpus of police incident narrative documents in understanding the spatial distribution of the topics. The motivation for doing this is that police narratives in each incident report contains very fine-grained information that is richer than the category that is manually assigned by the police. Our approach is to split the corpus into topics using two different unsupervised machine learning algorithms - Latent Dirichlet Allocation and Non-negative Matrix Factorization. We validate the performance of each learned topic model using model coherence.

Then, using a k-nearest neighbors density ratio estimation (kNN-DRE) approach that we propose, we estimate the spatial density ratio per topic and use this for data discovery and analysis of each topic, allowing for insights into the described incidents at scale. We provide a qualitative assessment of each topic and highlight some key benefits for using our kNN-DRE model for estimating spatial trends.

Jackknife empirical likelihood methods for the Cox regression model

Ali Jinnah
Department of Mathematics and Statistics, Georgia State University, USA
Collaborators/co-authors: Yichuan Zhao, Lauren Drinkard

In this paper, we propose a jackknife empirical likelihood method to draw inference for the regression parameters in Cox regression model. We develop the jackknife empirical likelihood (JEL), adjusted jackknife empirical likelihood (AJEL), mean jackknife empirical likelihood (MJEL), transformed jackknife empirical likelihood (TJEL) and transformed adjusted jackknife empirical likelihood (TAJEL) methods. We profile the set of nuisance parameters to study the parameter of interest. Extensive simulation studies show that the proposed jackknife empirical likelihood methods have better performance than the normal approximation in certain cases. We apply the proposed methods to study the Bone Marrow Transplant Patients (BMT), Larynx, and Myeloma real datasets for illustration.

Messaging fragmentation: an initial conceptualization and exploratory examination

Elise Karinshak
Department of Statistics
University of Georgia

Social media is central to online information seeking and sharing behaviors, and social media platforms significantly impact the creation and dissemination of information. Users are increasingly exposed to large amounts of brief, digestible information (as opposed to a few in-depth sources). This research proposes the Messaging Fragmentation Model (MFM), which delineates the evolution of information as it is transmitted via social media platforms. It proposes the distillation process, conceptualizing the emergence of information subsets on social media and how these subsets affect information dissemination. Key phenomena in the distillation process include: the emergence of information subsets, homogenization of descriptors, and divergence of topics of discussion. This process is illustrated through a case analysis of Twitter's manipulated media announcement; Twitter data is analyzed and discussed. This study reveals the impact of information subsets on information transmission and recommends additional analysis with the MFM as a valuable framework for information dissemination on social media.

Characterization of worldwide Mycobacterium bovis gene presence/absence to determine specific genomic signatures of across scales evolution

Noah Legall
Institute of Bioinformatics
University of Georgia
Collaborators/co-authors: Liliana Salvador

Mycobacterium bovis, a bacterial zoonotic pathogen responsible for the economically and agriculturally important livestock disease bovine tuberculosis (bTB), infects a broad mammalian host range worldwide. This characteristic has led to bidirectional transmission events between livestock and wildlife species as well as the formation of wildlife reservoirs, impacting the success of bTB control measures. Next Generation Sequencing (NGS) has transformed our ability to understand disease transmission events by tracking variant sites, however the genomic signatures related to host adaptation following spillover, alongside the role of other genomic factors in the M. bovis transmission process are understudied problems. Recently, computational and sequencing advances in tandem have made it feasible to conduct large scale comparative genomic analyses through the use of a pangenome framework, which looks to glean information from highly conserved regions (core genome) and variably present and absent regions (accessory genome) amongst a group of similar bacterial samples. By tracking the changes in gene variation amongst a large group of M. bovis, it might be possible to infer the genomic variations that coincide with adaptation at varying ecological scales (i.e. geographical, population cluster, and host species). We analyzed publicly available M. bovis datasets collected from a large amount of hosts worldwide to investigate how gene presence absence predicted an isolates membership across multiple ecological scales. We used the tool mbovpan to infer a pangenome of M. bovis, followed by using the 5-fold cross validated Random Forest (RF) models on the inferred accessory genome data to give individual genes an importance metric for each scale that was being investigated. The genes detected within these genomic regions harbor various pathogenic functions. The results of this study demonstrate how comparative genomics alongside machine learning approaches are useful to investigate further the nature of M. bovis host-pathogen interactions.

Targeting clinical equipoise via propensity score weighting

Yi Liu
Department of Statistics
North Carolina State University
Collaborators/co-authors: Roland A. Matsouaka, Yunji Zhou

Causal identification of a marginal treatment effect from observational data is an important focus in statistics. There has been a recent surge on propensity score methods for causal inference in recent literature, since the classical inverse probability weighting (IPW) method relies on some assumptions among which the "positivity" assumption can sometimes be violated. A number of questions have been posed about the goals and intent of these methods: to infer causality, what are they really estimating and what are their target populations?

We present a series of our current efforts on studying different aspects of these methods. First, we highlight some specific characteristics of the equipoise weights and corresponding estimators. We discuss three distinct potential motivations for weighting under the lack of positivity when estimating causal effects: (1) What essentially separates equipoise weights from IPW and IPW trimming? (2) How do equipoise weights target the clinical equipoise? (3) When should we expect similar results from these methods, even if the treatment effect is heterogeneous? Our work is illustrated via extensive simulation studies and data analysis. Second, we proposed some variance estimation methods for two common estimators of equipoise treatment effects under regular parametric models for treatment and outcome. Three sources of uncertainty are associated when we evaluate these their variances, i.e., when we estimate the treatment, outcome and the desired treatment effect. Third, we present our ongoing effort on a generalized estimating framework for causal effects on the treated population, namely equipoise treatment effect on the treated (EATT). We expect the new method can address the issue of lack of positivity on the control group.

Two-layer state-space model for smoothing spline with phase transition

Haoran Lu
Department of Statistics
University of Georgia
Collaborators/co-authors: Huimin Cheng, Ye Wang, Wenxuan Zhong, Ping Ma

Loan behavior modeling has been an essential but challenging issue in financial engineering. In this paper, we focus on the prediction of loan prepayment based on time series data of multiple customers. Existing approaches such as logistic regression or nonparametirc regression could only model the direct relationship between customer features and prepayment behavior, that is, the probability of prepayment is a function of features. Motivated from extracting some hidden phases of loan behavior, we propose the "smoothing spline with phase transition" (SSPT), based on a hidden Markov model (HMM) with varying transition and emission matrices modeled by smoothing splines. In contrast to existing methods, SSPT benefits from capturing the loans' unobserved phase transitions, which not only increases prediction performances but also provides more interpretability. The overall model is learned by EM-algorithm iterations, and within each iteration, smoothing splines are fitted with penalized least squares. Simulation studies demonstrate the effectiveness of SSPT, and a detailed real data analysis on loan data reveals the customers' hidden behavior patterns, which provides reliable predictions and meaningful interpretations for the financial industry.

Covariance estimators for the ROOT-SGD algorithm in online learning

Yiling Luo
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
Collaborators/co-authors: Xiaoming Huo, Yajun Mei

Online learning naturally arises in a lot of statistical and machine learning problems. The most wildly used family of methods in online learning is the stochastic first-order algorithm. Among this family of algorithms, there is a recently developed algorithm – Recursive One-Over-T SGD (ROOT-SGD). ROOT-SGD is advantageous over the previous algorithms in that it converges at a non-asymptotically faster rate, and its estimator further achieves an asymptotically normal distribution so that one can calibrate its uncertainty. However, the uncertainty measurement, in particular the covariance, in the ROOT-SGD depends on the unknown population risk function, and thus cannot be directly applied to measure the uncertainty. To fill this gap, we develop two estimators for the asymptotic covariance of ROOT-SGD. Using our covariance estimators, one will be able to conduct statistical inference for the ROOT-SGD estimator. Our first estimator is based on the idea of plug-in. For each unknown component in the asymptotic covariance, we estimate it by its empirical counterpart. The plug-in estimator converges to the true asymptotic covariance at a rate $\mathcal{O}(1/\sqrt{t})$, where t is the number of data samples. Despite its quick convergence, the plug-in estimator has the limitation that it relies on the Hessian of the stochastic loss function, which might be unavailable in some cases. Moreover, the computation of the plug-in estimator is expensive. Our second estimator is a Hessian-free estimator that overcomes the above disadvantages. The Hessian-free estimator uses the technique of random-scaling, and we show that it is an asymptotically consistent estimator for the true covariance. Observations in our numerical experiments are consistent with our theorems.

Cellcano: supervised cell type identification for single cell ATAC-seq data

Wenjing Ma
Department of Computer Science
Emory University
Collaborators/co-authors: Jiaying Lu, Hao Wu

Computational cell type identification (celltyping) is a fundamental step in single-cell omics data analysis. Supervised celltyping methods have gained increasing popularity in single-cell RNA-seq data because of the superior performance and the availability of high-quality reference datasets. Recent technological advances in profiling chromatin accessibility at single-cell resolution (scATAC-seq) have brought new insights to the understanding of epigenetic heterogeneity. With continuous accumulation of scATAC-seq datasets, supervised celltyping method specifically designed for scATAC-seq is in urgent need. In this work, we develop Cellcano, a novel computational method based on a two-round supervised learning algorithm to identify cell types from scATAC-seq data. The method alleviates the distributional shift between reference and target data and improves the prediction performance. We systematically benchmark Cellcano on 50 well-designed experiments

from various datasets and show that Cellcano is accurate, robust, and computational efficient. Cellcano is well-documented and freely available at `https://marvinquiet.github.io/Cellcano/`.

Using machine learning to predict which NFL teams will make the playoffs based on quarterback statistics

JD Miller
Department of Statistics
University of Georgia

I am attempting to determine which model or types of models, that I know are being used, are best at predicting the winner of an NFL game. I am examining existing models and building my own and examining their results over at least a 6 week period and seeing which are the most accurate. Some examples of existing models I am examining are a Guassian model built by Warner in 2010, a machine learning model built by Hamadani in 2005, FiveThirtyEight a popular website known for all things predictive, and others.

Computational modeling of endocrine disruption of gonadotrophin-dependent ovarian follicle maturation

Sarahna Moyd
Department of Biostatistics and Bioinformatics
Rollins School of Public Health, Emory University
Collaborators/co-authors: Shuo Xiao, Qiang Zhang

Selection of ovarian dominant follicles is a key step in folliculogenesis to ensure successful ovulation. The selection is a gonadotrophin-dependent process, requiring stimulation of antral follicles by follicle-stimulating hormone (FSH), which rises above a threshold in the early stage of the menstrual cycle in women. Many environmental pollutants, including microcystins, per- and polyfluorinated substances (PFAS), and a variety of endocrine disrupting chemicals (EDCs), can interfere with the follicle maturation process. To better understand and predict the dose-response relationships for the adverse outcomes of reproductive EDCs, it is helpful to construct quantitative adverse outcome pathway (qAOP) models simulating the perturbed signaling dynamics of ovarian follicles. Here we reported our effort in developing a computational model of the signal transduction and gene regulatory network that underpin follicle dominance selection and the feedback interaction between ovarian hormones and FSH. The dynamical model contains primarily the PKA and AKT pathways in the granulosa cells and the transcriptional program supporting follicle dominance and maturation to the preovulatory stage. Among the induced gene products, CYP19A1 (aromatase), insulin-like growth factors (IGF), and pregnancy-associated plasma protein (PAPPA) can synergize the activities of the signal transduction pathways, forming multiple intrafollicular positive feedback loops (FPLs). The PFLs function collectively as a bistable switch, underpinning the FSH thresholds required for the acquisition and maintenance of follicle dominance. Dominant follicles secrete inhibin and E2 into the circulation to inhibit pituitary FSH

secretion, preventing subordinate follicles from becoming dominant. By disrupting the cross-talk in signal transduction and blocking E2 signaling, the model was able to recapitulate the effects of microcystins and aromatase inhibitors in arresting antral follicles at pre-dominance stage. The model can help to understand the pathophysiology of women's ovarian diseases such as anovulation and polycysticovarian syndrome associa ted with environmental exposure.

Distance-based independence screening for canonical analysis

Yijin Ni
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
Collaborators/co-authors: Chuanping Yu, Andy Ko, Xiaoming Huo

In this paper, we propose Distance-based Independence Screening for Canonical Analysis (DISCA), a novel distance covariance-based technique for dimension reduction. DISCA is a method that simultaneously reduces the dimensions of two random variables to lower dimensional linear subspaces. DISCA is a stepwise algorithm: depending on its forward/backward version, in each step, it solves an optimization problem to identify a direction; directions from all the steps will form the final estimate. Numerically, DISCA solves a non-convex optimization problem at each stage, which can be written as a difference-of-convex (DC) optimization problem, allowing it to be solved by existing algorithms. Theoretically, consistency of the DISCA algorithm is established, and a concentration inequality that gives non-asymptotic error bounds is provided. We present exemplary cases where DISCA can perform dimension reduction while other methods cannot. In the simulation studies and the real data cases where the other state-of-the-art dimension reduction methods are applicable, we observe that DISCA performs comparably or better than most. All codes of our DISCA method can be found in GitHub `https://github.com/Yijin911/DISCA`, including an R package *DISCA*.

Efficient algorithms for learning to control bandits with unobserved contexts

Hongju Park
Department of Statistics
University of Georgia
Collaborators/co-authors: Mohamad Kazem Shirani Faradonbeh

Contextual bandits are canonical models for sequential decision-making under uncertainty in environments with time-varying components. In this setting, the expected reward of each bandit arm consists of the inner product of an unknown parameter with the context vector of that arm. The classical bandit settings heavily rely on assuming that the contexts are fully observed, while the study of the richer model of imperfectly observed contextual bandits is immature. This work considers Greedy reinforcement learning policies that take action as if the current estimates of the parameter and of the unobserved contexts coincide with the corresponding true values. We establish that the non-asymptotic worst-case regret grows poly-logarithmically with the time horizon

and the failure probability, while it scales linearly with the number of arms. Numerical analysis showcasing the above efficiency of Greedy policies is also provided.

A compressed sensing based least squares approach to semi-supervised local cluster extraction

Zhaiming Shen
Department of Mathematics
University of Georgia
Collaborators/co-authors: Ming-Jun Lai

A least squares semi-supervised local clustering algorithm based on the idea of compressed sensing is proposed to extract clusters from a graph with known adjacency matrix. The algorithm is based on a two-stage approach similar to a pioneering work under the same framework but with weaker model assumptions and less computational complexity. Our algorithm is shown to be able to find a desired cluster with high probability. The "one cluster at a time" feature of our method distinguishes it from other global clustering methods. Several numerical experiments are conducted on the synthetic data such as stochastic block model and real data such as MNIST, political blogs network, ATT and YaleB human faces data sets to demonstrate the effectiveness and efficiency of our algorithm.

A systematic view of information-based optimal subdata selection: Algorithm development, performance evaluation, and application in financial data

Difan Song
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
Collaborators/co-authors: Li He, William Li, Min Yang

With the urgent need of analyzing extraordinary amount of data, the information-based optimal subdata selection (IBOSS) approach has gained considerable attention in the recent literature due to its ability to maintain rich information of the full data. On the other hand, there lacks a systematic exploration of the framework, especially the characterization of the optimal subset when the model is more complex than first-order linear models. Motivated by a real finance case study concerning the impact of corporate attributes on firm value, we systematically explore the framework consisting of the exact steps one can follow when employing the idea of IBOSS for data reduction. In the context of the second-order models, we develop a novel algorithm of selecting an informative subdata. We also provide a thorough evaluation of the performance of the proposed algorithm from the standpoints of both predictions and variable selection, the latter of which is important for complex models but has not been given enough attention in the IBOSS field. Empirical studies including a real example demonstrate that the new algorithm adequately addresses the trade-off between the computation complexity and statistical efficiency, one of six

core research directions for theoretical data science research proposed by the US National Science Foundation. The real case study demonstrates the potential impact of the IBOSS strategy in scientific fields beyond statistics. In particular, we note that finance field, where the speed is critically important, is a promising area for applications of IBOSS.

Analysis of increase in punter and kicker statistics over thirteen years

Sloka Sudhin
Department of Statistics
University of Georgia

Over the past thirteen years, Competitive Sports Analysis (CSA) observed that each graduating class of punters and kickers had better performance stats than the last. This study aimed to assess why this occurred and how to account for this improvement in sports analytics. Literature review demonstrated that the biggest changes to performance statistics were specialized training camps (Battista, 2018; NFL Films, 2016; Murray, 1994; Neisen, n.d.); specifically, one specialized training camp, Pro Kick Australia, is well known for being a pipeline for Australian punters and kickers to enter the NCAA and the NFL (Bishara, 2018; Malchow, 2021). The study aimed to determine if the number of high school Australian punters and kickers joining the NCAA was greater or less than the number of high school American punters and kickers; results showed that there was a statistically significant difference between the number of American high school players in comparison to the number of Australian players recruited to the NCAA ($p < 0.05$). These results are inconsistent with the literature regarding professional players (Malchow, 2021; Tanier, 2015). In order to make the bell-curved, yet skewed, distributions of punter / kicker data a normal distribution, a statSCORE-subtraction algorithm was implemented that allowed CSA to create a normal distribution of statSCOREs by shifting values based on the difference between their distribution's mean and the projected mean of a normal distribution of statSCOREs. This program was deemed successful because the means of the new distributions were equal to the projected mean. Overall, this study provides college coaches with realistic expectations of the competition that their trainees are facing, allowing them to plan training in a way that benefits their players.

Data twinning

Akhil Vakayil
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
Collaborators/co-authors: V. Roshan Joseph

In this work, we develop a method named `Twinning` for partitioning a dataset into statistically similar twin sets. `Twinning` is based on `SPlit`, a recently proposed model-independent method for optimally splitting a dataset into training and testing sets. `Twinning` is orders of magnitude

faster than the `SPlit` algorithm, which makes it applicable to Big Data problems such as data compression. `Twinning` can also be used for generating multiple splits of a given dataset to aid divide-and-conquer procedures and k-fold cross validation.

Statistical inference of disease-associated genes on tissue-specific markers generates novel hypotheses on pathogenesis of complex human diseases

Boqi Wang
College of Arts and Sciences
Emory University
Collaborators/co-authors: Ammar Aleem Rashied, Steve Zhaohui Qin

Accurate identification of affected tissue of human diseases or traits is important for the derivation of disease etiology and the development of new treatment strategies. For example, which part of the brain is directly relevant to a specific psychiatric disorder can help us understand its pathogenesis. In this study, we tackle the problem using newly emerged genetics and genomics big data and develop a logistic regression-based method named LRDisTissue. The central hypothesis is that most disease-associated genes are expressed preferentially in affected organs or tissues. LRDisTissue takes advantage of newly emerged data on disease-related genes as well as tissue-specific gene expression data from Genotype-Tissue Expression (GTEx) V8 across 47 tissues. The unique feature of LRDisTissue is that it takes into account the strength of gene-disease associations. We applied LRDisTissue to a total of 3,261,324 gene-disease associations collected from DisGeNET covering 30,170 diseases and traits and 21,666 genes. Our approach has presented significantly more accurate results compared to others like linear regression and the chi-square test. Various tissue-trait combinations were revealed for 978 diseases and traits while some suggested potential explanations for disease pathogenesis. The results showed great consistency with past studeis and were proven effective by empirical plots and gene set enrichment analysis. Overall, LRDisTissue has shown great potential in uncovering novel pathogenesis mechanisms of complex diseases. In-depth analysis and experimental validation were required to fully understand these discovered tissue-trait associations and their enriched genes.

Sequential change-point detection for mutually exciting point processes over networks

Haoyun Wang
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
Collaborators/co-authors: Liyan Xie, Yao Xie, Alex Cuozzo, Simon Mak

We present a new CUSUM procedure for sequentially detecting change-point in the self and mutual exciting processes, a.k.a. Hawkes networks using discrete events data. Hawkes networks have become a popular model for statistics and machine learning due to their capability in modeling

irregularly observed data where the timing between events carries a lot of information. The problem of detecting abrupt changes in Hawkes networks arises from various applications, including neuronal imaging, sensor network, and social network monitoring. Despite this, there has not been a computationally and memory-efficient online algorithm for detecting such changes from sequential data. We present an efficient online recursive implementation of the CUSUM statistic for Hawkes processes, both decentralized and memory-efficient, and establish the theoretical properties of this new CUSUM procedure. We then show that the proposed CUSUM method achieves better performance than existing methods, including the Shewhart procedure based on count data, the generalized likelihood ratio (GLR) in the existing literature, and the standard score statistic. We demonstrate this via a simulated example and an application to population code change-detection in neuronal networks.

Sinkhorn distributionally robust optimization

Jie Wang
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
Collaborators/co-authors: Rui Gao, Yao Xie

We study distributionally robust optimization with Sinkhorn distance – a variant of Wasserstein distance based on entropic regularization. We derive its convex programming dual reformulation when the nominal distribution is a general distribution. Compared with Wasserstein DRO, it is computationally tractable for a larger class of loss functions, and its worst-case distribution is more reasonable. We propose an efficient stochastic mirror descent algorithm to solve the dual reformulation with provable convergence guarantees. Finally, we provide various numerical examples using both synthetic and real data to demonstrate its competitive performance and light computation cost.

Output space-filling design

Shangkun Wang
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
Collaborators/co-authors: V. Roshan Joseph

Space-filling designs are commonly used in computer experiments to fill the space of inputs so that the input-output relationship can be accurately estimated. However, when the relationship is highly nonlinear, a good space-filling design on the input space may produce large gaps in the output space and thereby, deteriorating the prediction performance. In this article, we propose a new experimental design method that tends to fill the space of the outputs. The method is adaptive and model-free, and therefore is expected to be robust to different kinds of modeling choices and input-output relationships. Several examples are given to show the advantages of the proposed method over the traditional space-filling designs.

Alignment of spatially resolved single-cell transcriptome using optimal transport

Zhen Wang
Department of Statistics
University of Georgia
Collaborators/co-authors: Guo-Cheng Yuan, Ping Ma, Wenxuan Zhong

Spatial studies of transcriptomes provide biologists with gene expression maps of heterogeneous and complex tissues. However, unlike single-cell RNA sequencing, spatial transcriptomics is at lower resolution and with limited sensitivity. To overcome these limitations, we present an algorithm named graphOT, a computational framework that probabilistically assigns cells to tissue locations. GraphOT can help incorporate spatial information to study tissue organization and spatial gene expression patterns. We demonstrate GraphOT on healthy mouse brain cortex with reconstructing spatial map at single-cell resolution.

Memory efficient kernel CUSUM for online change-point detection

Song Wei
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
Collaborators/co-authors: Yao Xie

We develop a non-parametric method for online change-point detection, namely memory efficient kernel CUSUM, which is based on Maximum Mean Discrepancy (MMD). Our procedure features a window-limited maximization to locate the change-point and an online recursive update of Gram matrix, which closely resembles the classic cumulative sum (CUSUM) procedure. We leverage the change-of-measure technique via exponential tilting and martingale concentration inequality to derive closed-form approximations to average run length (ARL) and expected detection delay (EDD), respectively. Our non-asymptotic analysis reveals the optimal choice of window size parameter, which is on the order of log ARL, to achieve (asymptotic) zero performance loss compared with the oracle procedure at the minimal computation and memory cost. Most importantly, this optimal choice as well as the corresponding EDD are analogous to classic results for window-limited generalized likelihood ratio (GLR) statistic and CUSUM statistic, respectively, which bridges our theoretical findings in non-parametric regime and the classic parametric results together. We conduct extensive numerical experiments to verify our theoretical findings and demonstrate our proposed method's good performance by comparing to various benchmarks.

Sex education and its future implications in physical health of American students

Ella Wileman
Department of Statistics
University of Georgia

By the 2009 United Nations Educational, Scientific, and Cultural Organization (UNESCO) guidelines, successful sex education is defined by cultural neutrality, age appropriateness, scientific accuracy, and ability to decrease sexually transmitted diseases (STDs) and teenage pregnancy (UNESCO et al., 2009; Ivanova et al., 2020 p. 8183). This definition of success has not yet been applied on a localized scale, which is necessary for nations such as the United States that have no centralized legislation for the issue (Keogh et al., 2019, pp. 119-137; Guttmacher Institute, 2022). This study details models constructed to predict negative outcomes of STDs and teen pregnancy using the curriculum, bias, and content of state-level sex education legislation as predictor variables. Demographic variables including religious makeup, unemployment, immigration, adult population proportion, income, and poverty were incorporated to increase accuracy (Sexually Transmitted Diseases, 2021). Using R Studio, two models were built for each outcome variable: a multivariate regression model and a binary logistic regression model (R Studio, 2021). The predictor variables were transformed into their most significant forms for each type of regression using bivariate analysis. The multivariate model predicted a rate for each outcome variable and the logistic models classified either a "high" or "low" rate. Demographic variables were found to be more significant to the transmission of disease than the sex education policies, excluding policy bias regarding abstinence and sexuality. Five outcome variables were studied: teen birth, chlamydia, syphilis, gonorrhea, and Human Immunodeficiency Virus (HIV). A Graphical User Interface (GUI) was built around the logistic teen birth rate model, where users can input local demographic data and view whether these indicate a high or low teen birth rate.

Targeted image poisoning

Lauren Rose Wilkes
Department of Statistics
University of Georgia
Collaborators/co-authors: Alaa Khaddaj, Saachi Jain, Aleksander Mądry

State of the art image classification tasks use advanced machine learning models that regularly achieve high accuracy. However, many implementations fail to account for the risk of adversarial examples, which are inputs specifically designed to break the model. Harmful players can take advantage of these examples to conduct data poisoning, where certain images are given a trigger during training and subsequentially get their label flipped to a target class. This causes the model to learn the trigger and when the trigger is applied during test time, the model will classify most images as the target class regardless of their original label. Previously, poisoning points has been conducted by randomly selecting images. This project will leverage datamodels, a tool that can determine the influence of the training images on each other, to select the most influential points. By poisoning the images that are most influential to other images in the model, this project will

determine if we can effectively break the model using fewer poisoned images. By successfully choosing poisons, we expose the ease and risk of data poisoning if someone has access to training data and investigate how machine learning models learn patterns.

Unsupervised anomaly detection and diagnosis in power electronic networks: Informative leverage and multivariate functional clustering approaches

Shushan Wu
Department of Statistics
University of Georgia
Collaborators/co-authors: Luyang Fang, Jinan Zhang, Stephen J. Coshatt, Feraidoon Zahiri, Alan Mantooth, Jin Ye, Wenxuan Zhong, Ping Ma, WenZhan Song

We propose a novel unsupervised anomaly detection and diagnosis algorithm in power electronic networks. Since most anomaly detection and diagnosis algorithms in the literature are based on supervised methods that can hardly be generalized to broader scenarios, we propose unsupervised algorithms. Our algorithm extracts the Time-Frequency Domain (TFD) features from the three-phase currents and three-phase voltages of the point of coupling (PCC) nodes to detect anomalies and distinguish anomaly types, cyber-attacks, and physical faults. To detect anomalies through TFD features, we propose a novel Informative Leveraging for Anomaly Detection (ILAD) algorithm. The proposed unsupervised ILAD algorithm automatically extracts noise-reduced anomalous signals, achieving more accurate anomaly detection results than other score-based methods. We apply a novel Multivariate Functional Principal Component Analysis (MFPCA) clustering method to assign anomaly types for anomaly diagnosis. Unlike the deep learning methods, the MFPCA clustering method does not need labels to train, yields more accurate results than other deep embedding based clustering approaches, and is even comparable to supervised algorithms in both offline and online experiments. To the best of our knowledge, the proposed unsupervised framework accomplishing anomaly detection and anomaly diagnosis tasks is the first of its kind in power electronic networks.

A fast algorithm for the Wasserstein-distance-based independence test

Yiling Xie
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
Collaborators/co-authors: Yiling Luo, Xiaoming Huo

We observe that computing empirical Wasserstein distance in the independence test is an optimal transport (OT) problem with a special structure. This observation inspires us to study a special type of OT problem and propose a modified Hungarian algorithm to solve it exactly. For an OT problem between marginals with m and n atoms ($m \geq n$), the computational complexity of the proposed algorithm is $\mathcal{O}(m^2 n)$. Computing the empirical Wasserstein distance in the independence

test requires solving this special type of OT problem, where we have $m = n^2$. The associate computational complexity of our algorithm is $\mathcal{O}(n^5)$, while the order of applying the classic Hungarian algorithm is $\mathcal{O}(n^6)$. Numerical experiments validate our theoretical analysis.

An alternative approach to train neural networks using monotone variational inequality

Chen Xu
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
Collaborators/co-authors: Xiuyuan Cheng, Yao Xie

Theoretical understanding of the neural network training procedures remains limited, especially in providing performance guarantees of testing performance due to the non-convex optimization problem. The current paper investigates an alternative approach to neural network training by reducing to another problem with convex structure — to solve a monotone variational inequality (MVI) — inspired by a recent work of (Juditsky & Nemirovski, 2019). We propose a practical and completely general algorithm called stochastic variational inequality (SVI), and demonstrate its applicability in training fully-connected neural networks, graph neural networks (GNN), and convolutional networks (CNN); SVI is completely general for training other networks. In special cases, we obtain performance guarantee of ℓ_2 and ℓ_∞ bounds on model recovery and prediction accuracy.

Data-driven method for combining measurements from a group of heterogeneous raters for the evaluation of a new device

Qi Yu
Department of Biostatistics and Bioinformatics
Rollins School of Public Health, Emory University
Collaborators/co-authors: Ying Cui, Jeong Hoon Jang, Amita Manatunga

With advanced technology, many computer-aided diagnostic (CAD) devices are being introduced to detect and monitor complex diseases. This work is motivated by the need to develop a statistical framework to evaluate the performance and acceptability of a newly developed CAD device. In absence of a "gold standard" test, often in practice, the evaluation of a new CAD device is achieved by comparing its measurement to the average of multiple measurements from clinicians who are regarded as the best available standard (imperfect gold standard). This approach, however, may lead to biased evaluations as these clinicians may have different experiences and accuracy levels. In this work, we propose a novel weighting strategy to combine measurements from a heterogeneous group of clinicians. Specifically, an induction method, which learns from data in an unsupervised manner, is proposed to sequentially assign higher weights to clinicians who consistently agree with others and to assign lower weight to those who mostly disagree with others, providing a fair

evaluation of a new device according to the consistent opinions among clinicians. Our method is applicable to any of the agreement measures, including CCC and TDI. We demonstrate the practical utility of the proposed method via extensive simulation studies and an application to the data from a renal study.

Design and analysis for multi-fidelity finite element simulations

Henry Shaowu Yuchi
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
Collaborators/co-authors: V. Roshan Joseph, C. F. Jeff Wu

The numerical accuracy of Finite Element Analysis (FEA) depends on the number of finite elements used in the discretization of the space, which can be varied using the mesh size. The larger the number of elements, the more accurate the results are. However, the computational cost increases with the number of elements. In current practice, the experimenter chooses a mesh size that is expected to produce a reasonably accurate result, and for which the computer simulation can be completed in a reasonable amount of time. Improvements to this approach have been proposed using multi-fidelity modeling by choosing two or three mesh sizes. However, mesh size is a continuous parameter and therefore, multi-fidelity simulations can be performed easily by choosing a different value for the mesh size for each of the simulations. In this work, we develop a method to optimally find the mesh sizes for each simulation and satisfy the same time constraints as a single or double mesh size experiment.

Novel empirical likelihood methods for the cumulative hazard ratio

Dazhi Zhao
Department of Mathematics and Statistics
Georgia State University
Collaborators/co-authors: Yichuan Zhao

In medical research, it is often very meaningful to evaluate the effect of a treatment via the cumulative hazard ratio, especially when those hazards may be nonproportional. In order to capture the cumulative treatment effect, the ratio of the treatment and specific cumulative baseline hazards is often used as a measure of the treatment effect. Pointwise and simultaneous confidence bands associated with the estimated ratio provide a global picture of how the treatment effect evolves over time. Our research project can be divided into two parts. In the first part, we construct a pointwise confidence interval for the ratio using a plug-in type empirical likelihood approach, in which we use the full likelihood method to estimate the unknown parameters in the Cox model. In the second part, we construct a pointwise confidence interval for the ratio based on the profile empirical likelihood approach, which can make full use of the information contained in the survival model, and the limiting distribution of the proposed statistic is optimized to the standard chi-square distribution. We will present the model framework for both parts and some simulation results and real data analysis.