Georgia Statistics Day 2025

October 31, 2025

Hosted by:

Department of Statistics

UNIVERSITY OF
GEORGIA

Welcome

The Department of Statistics at the University of Georgia welcomes you to Athens for the 2025 Georgia Statistics Day.

The Georgia Statistics Day (GSD) is an annual event, designed to promote interdisciplinary statistics research among the three premier academic institutions in the State of Georgia: the University of Georgia, Georgia Institute of Technology, and Emory University. The venue of the conference rotates among the three participating institutions on a three-year cycle. In 2015, the inaugural Georgia Statistics Day was held at UGA. This year, UGA is again pleased to serve as host for Georgia Statistics Day 2025. We are delighted to have Dr. Liza Levina, Vijay Nair Collegiate Professor of Statistics at the University of Michigan as the keynote speaker for our 2025 event.

As before, the format of the GSD 2025 is such that, apart from the keynote lecture, most of the research talks will be delivered by faculty from the three host institutions. A significant part of the one-day workshop will also be student poster sessions and we expect to have around 50 poster presentations. Awards will be given to outstanding student posters in several categories.

In addition to facilitating the exchange of cutting-edge research in statistics and data science, an important focus of GSD 2025 is to build collaborations with top industries, expose our students to applied research carried out in industry, and more importantly, help them build connections with industry sponsors. As in the three previous GSDs, we have reached out to leading businesses like State Farm, JMP/SAS, and others for participation in GSD 2025. Along with the poster sessions mentioned above, there will be a special Industry Session where statisticians from these companies will discuss careers in industry and job opportunities.

We appreciate your attendance and look forward to a successful conference.

Once again, welcome to Georgia Statistics Day 2025!

Contents

GSD 2025 Committees

Organizers

Abhyuday Mandal, University of Georgia

Steering Committee

Xiaoming Huo, Georgia Institute of Technology
Abhyuday Mandal, University of Georgia
Robert Krafty, Emory University

Local Organizing Committee

Abhyuday Mandal, University of Georgia
Sen Na, Georgia Institute of Technology
Tianwen Ma, Emory University

Poster Evaluation Committee

Graduate Posters

Ting Zhang, University of Georgia (Chair)
Shuyang (Ray) Bai, University of Georgia
Mandev Gill, University of Georgia
Whitney Huang, Clemson University
Yuan Ke, University of Georgia
Liang Liu, University of Georgia
Rongjie (RJ) Liu, University of Georgia
Aditya Mishra, University of Georgia
Sen Na, Georgia Institute of Technology
Subhadeep Paul, Ohio State University
Zhaohui (Steve) Qin, Emory University
Tharuvai (TN) Sriram, University of Georgia

Undergraduate Posters

Joshua Lukemire, Emory University (Chair)
William (Bill) Fisher, JMP
Mark Werner, University of Georgia

Staff Support

Ryan Robinson
Wendy Brown
Tanya Boyd

Student Volunteers

Maxwell Baxley
Cong Cheng
Angelina Garcia
Bingnan Li
Andrew Mosbo
Bella Salter
George Whittington

Sponsors

UNIVERSITY OF GEORGIA

Department of Statistics
Franklin College of Arts and Sciences
Graduate School
Morehead Honors College
Office of the Senior Vice President for Academic Affairs and Provost

jmp Statistical Discovery™ From SAS.

Georgia Tech ISyE H. Milton Stewart School of Industrial and Systems Engineering

ASA

EMORY UNIVERSITY

UPCR

NISS National Institute of Statistical Sciences Academic Affiliate Program Member

IISA International Indian Statistical Association

Georgia Statistics Day 2025 acknowledges the generous support from the following sponsors (in alphabetical order)

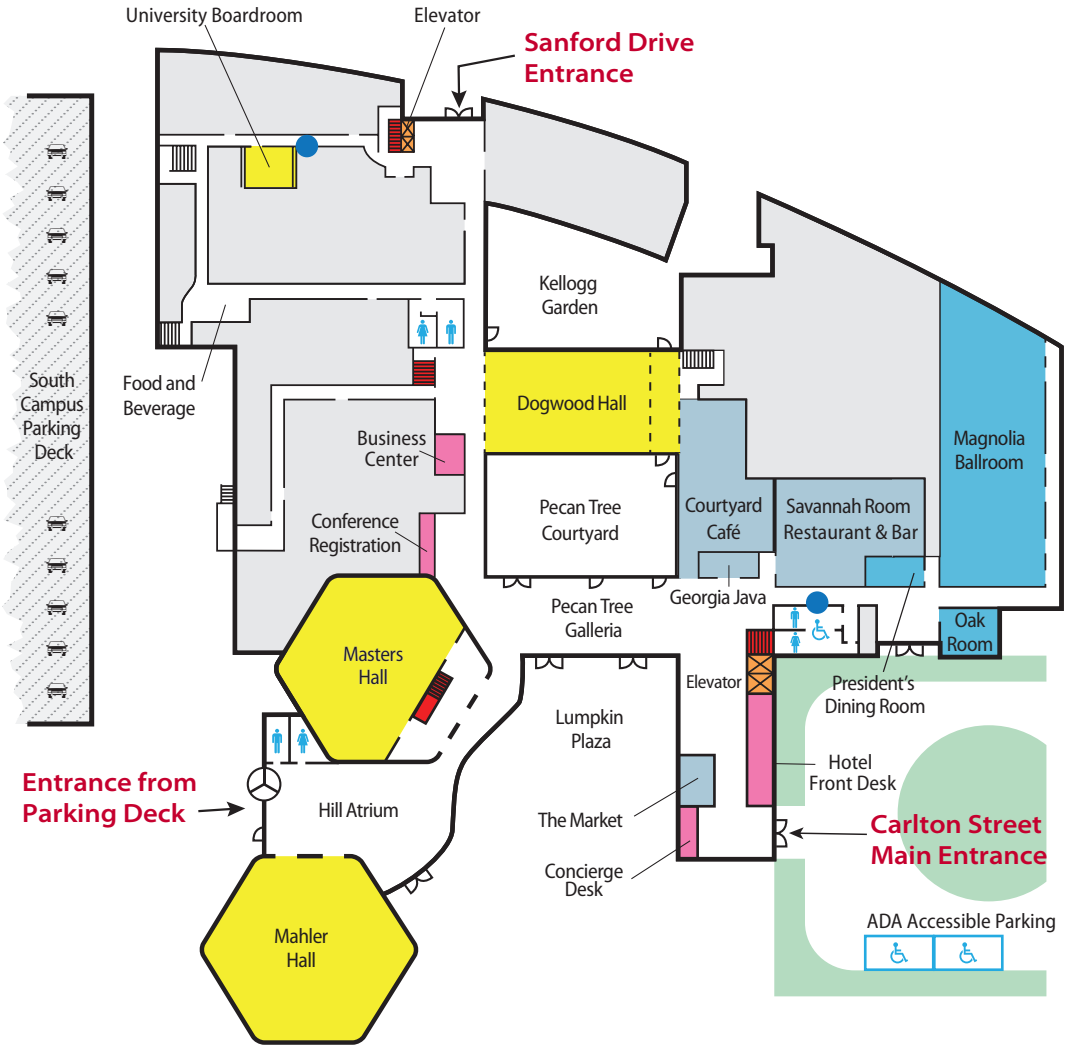
- American Statistical Association, Georgia Chapter
- Department of Biostatistics and Bioinformatics, Emory University
- Department of Statistics, University of Georgia
- Jere W. Morehead Honors College, University of Georgia
- JMP® Statistical Software
- National Institute of Statistical Sciences
- The Graduate School, University of Georgia
- Office of the Senior Vice President for Academic Affairs and Provost, University of Georgia
- The H. Milton Stewart School of Industrial & Systems Engineering, Georgia Institute of Technology
- The International Indian Statistical Association (IISA)
- Unclaimed Property Consulting & Reporting (UPCR)

Conference Venue

The venue of the conference is the University of Georgia Center for Continuing Education and Hotel

Address: 1197 S Lumpkin St, Athens, GA 30602

First floor

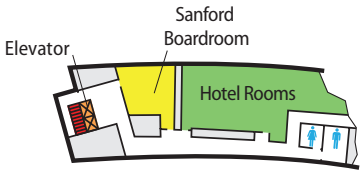


UNIVERSITY OF GEORGIA
Center for Continuing Education & Hotel

Restrooms	Meeting Space
ADA Accessible	Banquet Dining
Elevators	Restaurants
Stairs	Registration & Service Areas
Water Bottle Filling Station	Hotel Rooms

Second floor

Fifth floor



8-18/186793

Georgia Center Map

Scientific Program – At a Glance

List of Sessions

Friday, October 31, 2025

Time	Location	Session
7:30–9:00	Magnolia Balllroom	Breakfast
8:00–8:45	Connector	Registration
8:45–9:00	Master's Hall	Opening Ceremony ● Dan Hall, Professor and Interim Head, Department of Statistics, UGA
9:00–10:00	Master's Hall Session Chair: T. N. Sriram	Keynote Lecture – Liza Levina, Vijay Nair Collegiate Professor of Statistics at the University of Michigan
10:00–10:15	Kellogg Concourse	Break
10:15–11:45	Room KL Session Chair: Rongjie Liu	Parallel Session A1 ● Ting Zhang, UGA ● Limin Peng, Emory ● Xiaoming Huo, Georgia Tech
	Room R Session Chair: Liang Liu	Parallel Session A2 ● Ruizhi Zhang, UGA ● William (Bill) Fisher, JMP ● Kamran Paynabar, Georgia Tech
	Room Q Session Chair: Yuan Ke	Parallel Session A3 ● Subhadeep Paul, OSU ● Yichuan Zhao, GSU ● Shihao Yang, Georgia Tech

Friday, October 31, 2025

Time	Location	Session
11:45–1:00	Magnolia Ballroom	Lunch Break
1:00–2:30	Pecan Tree Galleria	Posters
2:30–3:30	Room KL Session Chair: Dan Hall	Industry Session ● William (Bill) Fisher, JMP ● Andrew Neely, Unclaimed Property Consulting & Reporting LLC
3:30–3:45	Kellogg Concourse	Break
3:45–5:15	Room KL Session Chair: Mandev Gill	Parallel Session B1 ● Whitney Huang, Clemson ● Yaotian Wang, Emory ● Roshan Joseph, Georgia Tech
	Room R Session Chair: Ray Bai	Parallel Session B2 ● Xiaotian Zheng, UGA ● Yijian (Eugene) Huang, Emory ● Hanwen Huang, Augusta
5:15–5:20	Break	
5:20–5:30	Room KL	Closing remarks and award ceremony

Featured Speakers

Keynote Lecture

Elizaveta (Liza) Levina

Liza Levina is the Vijay Nair Collegiate Professor of Statistics at the University of Michigan, and affiliated faculty at the Michigan Institute for Data and AI in Society and the Center for the Study of Complex Systems. She received her PhD in Statistics from UC Berkeley in 2002, and has been at the University of Michigan since, serving as the department chair from 2020 to 2025. Her research interests are in network analysis, high-dimensional statistics, statistical learning, and applications to neuroscience and imaging. Honors include being selected as a fellow of the American Statistical Association, a fellow of the Institute of Mathematical Statistics, a Web of Science Highly Cited Researcher, an IMS Medallion lecturer, and an ICM invited speaker.

Talks

Keynote Lecture

Liza Levina	Towards Interpretable and Trustworthy Network-Assisted Prediction

Invited Talks

William Fisher	Combinatorial Testing of Engineered Systems with Hard-to-Change Factors
Hanwen Huang	Statistical Inference in Classification of High-dimensional Gaussian Mixture
Whitney Huang	A Physics-Statistics Approach for Geophysical Extreme Event Risk Assessment
Yijian Huang	Learning Neyman–Pearson Classifiers for Cancer Detection
Xiaoming Huo	Asymptotic Behavior of Adversarial Training Estimator under ℓ_∞-Perturbation
Roshan Joseph	Factor Importance Ranking and Selection
Subhadeep Paul	Heterogeneous Transfer Learning with Statistical Error Bounds
Kamran Paynabar	Sequential Sampling for Optimization under Functional Uncertainty: A Robust Approach in Function Space
Limin Peng	Estimation and Prediction of Time-in-range with Inpatient Continuous Glucose Monitoring
Yaotian Wang	A Bayesian Blind Source Separation Framework for Uncovering Reliable Neural Circuitry in the Developing Connectome
Shihao Yang	Modernizing Time Series Forecasting: Reuniting Statistical Structure with Deep Learning through Transformer

Ruizhi Zhang	Optimal Bounded-Influence Procedure for Robust Sequential Change-Point Detection
Ting Zhang	Tail Spectral Density Estimation and Its Uncertainty Quantification: Another Look at Tail Dependent Time Series Analysis
Yichuan Zhao	Novel Empirical Likelihood Method for the Cumulative Hazard Ratio Under Stratified Cox Models
Xiaotian Zheng	Mixture Modeling for Temporal Point Processes with Memory

Industry Talks

William Fisher	What Is It Like to Work at JMP Statistical Discovery LLC?
Linh Le	Data Science and MAGNet at State Farm

Posters

Posters

Postdoctoral Posters

Jamshid Namdari	Localized Sparse Principal Component Analysis of Multivariate Time Series in Frequency Domain
Anik Roy	Capturing Global Heterogeneous Associations Between Tensor Outcomes and Predictors

Graduate Student Posters

Yisa Abolade	Parameter Estimation in Epidemiological Models Using the Sum of Absolute Deviations Method
Conglin Bao	Imputing Alzheimer's Disease Pathologies in Living Participants Using Deep Learning Models Trained on Decedent Data
Arghadeep Basu	Joint Credible Distribution of Ranks for Unit Level Small Area Estimation
Maxime Bouadoumou	Divide-and-Conquer Empirical Likelihood for the Fine–Gray Model in Large-Scale Competing Risks Data
Jiazhang Cai	SpaDiff: Denoising for Sequence-based Spatial Transcriptomics via Diffusion Process
Menghui Chen	Age Effect Explorer: Bulk and Single-Cell Dashboards for Human Aging Studies
Xinchen Du	Online Statistical Inference of Constrained Stochastic Optimization via Random Scaling
Lihui Duan	Prospective Association between Cognition and COVID-19 Serology: A Hurdle Analysis of Anti-N Positivity and Titer Intensity in HAALSI
Prasanjit Dubey	Necessary Optimality Conditions in Multiple Hypotheses Testing under Exact Family-Wise Error Rate Control – a Path to Fast Computing
Greg Ellison	Fast and Accurate Divergence Time Estimation Using the Pairwise Likelihood
Jingying Gao	Tail Index Estimation for Tail Adversarial Stable Time Series with an Application to High-Dimensional Tail Clustering
Joseph (Joe) Hart	Forecasting with Physics-Informed Statistical Learning
Umma Hafsah Himu	Statistical Meta-analysis to Investigate the Association of SNP (rs4680) in COMT Gene with Multiple Cancers

Youwei Hu	Identifying Quantitative EEG Biomarkers to Evaluate Alzheimer's Disease Severity Using a Standardized Preprocessing Pipeline
Yu Huang	Neural Network Models for Group Testing Data
Collin Hunt	Shape Based Analysis of Subcortical Structures in Alzheimers Disease Classification
Sihan Jia	Methodological Approaches for the Estimation of Confidence Intervals on Partial Youden Index under Verification Bias
Xiaoye Jiang	MRI2Sex: A 2D CNN Model for Sex Classification from MRI Scans
Katherine Kreuser	Dynamic Surrogate Modeling Methods for Online Optimization
Junghwan (Jay) Lee	Flow-based Conformal Prediction for Multi-dimensional Time Series
Tianqi Li	From Pixel to Transcript: Evaluating the Effective Subcellular Resolution of Visium HD Spatial Transcriptomics
Zhe Liu	Unmasking Technical Zeros in Multi-omics: Challenges and Solutions
Ziyu Liu	STAR: Shared-Tree Adaptive Representation for Video Compression
Amy Moore	Methods for Estimating VE Using Routine School Testing Data with Differential Testing Behavior
Yijin Ni	An Analytical Framework to Precisely Characterize the Accuracy-Fairness Trade-off in Fair Representation Learning
Kayode Okunola	Effect of Life Expectancy on Economy Growth for High-Income Nations
Zhengyi Ou	Effect of Three Intra-articular Injections on Patient Reported Knee Osteoarthritis Outcomes: A Retrospective Longitudinal Study
Yumo Peng	Supervised Subdata Selection of Big Data Using Weighted Support Points
Guangbin Quan	Multi-Task Learning of Cognitive Functions Using Brain Connectome Data
Ahmer Raza	Private and Efficient Surveillance Using Group Testing
Vishal Routh	Causal Discovery in Recursive Homogeneous Models with Heavy Tailed Innovations
Dinuka M Senevirathne	Comparing Logistic and Survival Analysis Modeling for Tree-Level Mortality Prediction Across Different Forest Conditions
Difan Song	Efficient Optimization of Plasma Radiation Detectors Using Imperfect Inference Models
Zeliang Sun	Trait Regression for Brain Connectomes Based on Deep Parcellation Analysis
Shirui Wang	Imputation of Missing Cancer Stage at Diagnosis Accounting for Stage-specific Survival
Tao Wang	GASDU: Gauss–Southwell Dynamic Update for Efficient LLM Fine-Tuning
Jianliang Ye	Statistical Inference of Constrained Model Estimation via Derivative-Free Stochastic Sequential Quadratic Programming
Meizhi Yu	Extracting Viral Signatures from Complex Raman Spectra
Wei-Yang Yu	Automated Analysis of Experiments Using Hierarchical Garrote
Zhuoran Yu	Region-specific Gene Co-expression Network Inference for Spatial Transcriptomics Data
Chenyang Yuan	mcDETECT: Characterizing mRNA Localization in Polarized Neuronal Compartments
Xiaotian Zhang	Spatio-Temporal Epidemic Forecasting Using GNN and Transformer Informed by Mobility and Transmission Dynamics

Undergraduate Student Posters

Deeya Datta	Empirical Bayes Approach to Overall Ranking of Populations
Rui Gong	Evaluating Machine Learning Models for Breast Tumor Classification: A Comparative Study
Yumin Min	Tumor Mutational Burden and Its Association With Immune Response: A TCGA Bioinformatics Study.
Sneha Shadija	Assessing AMF Colonization and Disease Susceptibility in 12 Sorghum Accessions Using Linear Mixed-Effects Models: Treatment vs. Control
Jacob Song	Robotics-Based Modeling of Poultry Flock Behavior for Disease Spread Simulation
Sloka Sudhin	d-QPSO: an R Package for Searching for Efficient Designs
Xiaohan Sun	Integrating Antigen Processing and Presentation Prediction with MHCflurry for Enhanced Neoantigen Prioritization
Siddarth Suresh	Optimizing Diagnostic Thresholds in Cardiovascular Risk: Youden Index Analysis with Machine Learning and Deep Learning Approaches

Abstracts

Keynote Lecture

Towards Interpretable and Trustworthy Network-Assisted Prediction

Liza Levina
Department of Statistics
University of Michigan

When training data points for a prediction algorithm are connected by a network, it creates dependency, which reduces effective sample size but also creates an opportunity to improve prediction by leveraging information from neighbors. Multiple prediction methods on networks taking advantage of this opportunity have been developed, but they are rarely interpretable or have uncertainty measures available. This talk will cover two contributions bridging this gap. One is a conformal prediction method for network-assisted regression. The other is a family of flexible network-assisted models built upon a generalization of random forests (RF+), which both achieves competitive prediction accuracy and can be interpreted through feature importance measures. Importantly, it allows one to separate the importance of node covariates in prediction from the importance of the network itself. These tools help broaden the scope and applicability of network-assisted prediction to practical applications.

Technical Sessions

Combinatorial Testing of Engineered Systems with Hard-to-Change Factors

William Fisher
JMP Statistical Discovery LLC
Collaborators/co-authors: Ryan Lekivetz and Joseph Morgan

The testing of engineered systems is a complex task which often requires a test engineer to construct a suite of test cases in order to investigate system behavior, with the primary purpose of this endeavor being to detect errors in the system. Test engineers often must strategically construct their suite of test cases due to budgetary and other operational constraints. *Combinatorial testing* is an effective and efficient method to construct such test suites. The mathematical object typically used for construction is a *covering array*, where the columns of the array correspond to the *factors* of the system under test (SUT). In some instances, there may be a subset of factors that are expensive or hard to change. In the statistical design of experiments setting, *split-plot* designs are used when hard-to-change factors are present to accomodate randomization constraints. We adopt a similar structural approach, focusing on addressing practical implementation constraints such as reducing the number of resets of hard-to-change factors, which can be costly or time-consuming. Borrowing the concept of split-plot designs from design of experiments, we extend the methodology of combinatorial testing by presenting a method for constructing covering arrays with split-plot structure. We then demonstrate the efficiency of our method in a case study involving the testing of an implementation of XGBoost.

Statistical Inference in Classification of High-Dimensional Gaussian Mixture

Hanwen Huang
Department of Biostatistics, Data Science and Epidemiology
Medical College of Georgia, Augusta University
Collaborators/co-authors: Peng Zeng

We consider the classification problem of a high-dimensional mixture of two Gaussians with general covariance matrices. Using the replica method from statistical physics, we investigate the asymptotic behavior of a broad class of regularized convex classifiers in the limit where both the sample size n and the dimension p approach infinity while their ratio $\alpha = n/p$ remains fixed. This approach contrasts with traditional large-sample theory in statistics, which examines asymptotic behavior as $n \to \infty$ with p fixed. A key advantage of this asymptotic regime is that it provides precise quantitative guidelines for designing machine learning systems when both p and n are large but finite. Our focus is on the generalization error and variable selection properties of the estimators. Specifically, based on the distributional limit of the classifier, we construct a de-biased estimator to perform variable selection through an appropriate hypothesis testing procedure. Using L_1-regularized logistic regression as an example, we conduct extensive computational

experiments to verify that our analytical findings align with numerical simulations in finite-sized systems. Additionally, we explore the influence of the covariance structure on the performance of the de-biased estimator.

A Physics-Statistics Approach for Geophysical Extreme Event Risk Assessment

Whitney Huang
School of Mathematical and Statistical Sciences
Clemson University
Collaborators/co-authors: Katherine Kreuser

Catastrophic events, such as hurricane-induced storm surges, earthquake-triggered tsunamis, and volcanic pyroclastic flows, can cause significant damage to human society and environmental systems. Accurately quantifying the risks associated with these events is essential for effective risk assessment and mitigation. However, the rarity of such events limits the availability of observational data, making purely data-driven approaches insufficient. Computer models, which incorporate physical knowledge, offer a promising alternative by generating synthetic events to augment limited data, thereby enabling more robust risk analysis.

This talk presents a statistical workflow for a physics-statistics approach to long-term extreme event risk assessment, involving the following tasks: - Estimating the computer model input distribution.
- Emulating the computer model input-output relationship via surrogate models.
- Performing forward propagation by integrating the input distribution with the input-output surrogate model to estimate the tail distribution of the output (e.g., the 1-in-100-year return level).
A case study on storm surge risk in Southwest Florida will be presented to illustrate the approach.

Learning Neyman–Pearson Classifiers for Cancer Detection

Yijian Huang
Department of Biostatistics and Bioinformatics
Rollins School of Public Health, Emory University

The Neyman–Pearson (NP) classification framework seeks to maximize class-1 accuracy while maintaining a prespecified control level on class-0 accuracy, making it an attractive paradigm for biomarker-based cancer detection. We study a learning approach that maximizes empirical utility and introduce an efficient algorithm for linear NP classification that handles non-smooth, non-convex optimization. However, the resulting classifier often fails to meet the desired control level on average. Through higher-order asymptotic analysis, we identify the source of this discrepancy and develop improved methods for classifier estimation. We further propose risk prediction methods based on the training data. Simulation studies and an application to aggressive prostate cancer detection demonstrate the effectiveness of the proposed approach.

Asymptotic Behavior of Adversarial Training Estimator under ℓ_∞-Perturbation

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

Adversarial training has been proposed to protect machine learning models against adversarial attacks. This article focuses on adversarial training under ℓ_∞-perturbation, which has recently attracted much research attention. The asymptotic behavior of the adversarial training estimator is investigated in the generalized linear model. The results imply that the asymptotic distribution of the adversarial training estimator under ℓ_∞-perturbation could put a positive probability mass at 0 when the true parameter is 0, providing a theoretical guarantee of the associated sparsity-recovery ability. Alternatively, a two-step procedure is proposed—adaptive adversarial training, which could further improve the performance of adversarial training under ℓ_∞-perturbation. Specifically, the proposed procedure could achieve asymptotic variable-selection consistency and unbiasedness. Numerical experiments are conducted to show the sparsity-recovery ability of adversarial training under ℓ_∞-perturbation and to compare the empirical performance between classic adversarial training and adaptive adversarial training. Supplementary materials for this article are available online, including a standardized description of the materials available for reproducing the work. This paper is to appear in the Journal of the American Statistical Association, Theory and Methods.

Factor Importance Ranking and Selection

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

There can be numerous factors present in a system, but only a few may affect the output. In this talk I will explain how to identify the important factors and rank them according to their importance using total Sobol' indices. I will start with a simple case where we know the input-output relationship and can simulate the inputs using experimental design techniques. Then I will move on to the more complicated case of observational data and introduce FIRST indices. FIRST is a fast and nonparametric method, which performs better than the state-of-the-art techniques for variable selection and importance ranking.

Heterogeneous Transfer Learning with Statistical Error Bounds

Subhadeep Paul
Department of Statistics
The Ohio State University
Collaborators/co-authors: Jae Ho Chang, Chenze Li, Massimiliano Russo

In the first part of the talk, we consider the problem of transferring knowledge from a source domain to a new target domain for learning a high-dimensional regression model with different features. Recently, the statistical properties of homogeneous transfer learning have been investigated. However, most homogeneous transfer and multi-task learning methods assume that the target and proxy domains have the same feature space, limiting their practical applicability. In applications, target and proxy feature spaces are frequently inherently different, for example, due to the inability to measure some variables in the target data-poor environments. Conversely, existing heterogeneous transfer learning methods do not provide statistical error guarantees, limiting their utility for scientific discovery. We propose learning the relationship between the missing and observed features through a feature map that may be linear or nonparametric through unknown functions in the source data. Then, we solve a joint penalized regression optimization problem in the target data. We develop upper bounds on the method's estimation and prediction risks, assuming that the proxy and the target domain parameters are sparsely different. Our results elucidate how estimation and prediction error depend on the complexity of the model, sample size, the extent of overlap, and the association between matched and mismatched features. In the second part of the talk, I will discuss my recent work on developing a transfer learning method for nonparametric regression problems using a random forest with distance covariance-based feature weights. The unknown source and target regression functions are assumed to differ for a small number of features. We derive an upper bound on the mean square error rate of the procedure that theoretically brings out the benefits of transfer learning in random forests.

Sequential Sampling for Optimization under Functional Uncertainty: A Robust Approach in Function Space

Kamran Paynabar
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
Collaborators/co-authors: Pouya Ahadi

Sequential sampling is widely used to identify optimal input values that achieve a de- sired response. Most existing work focuses on scalar responses and/or assumes Gaus- sian uncertainty. We introduce Functional Robust Bayesian Optimization (FRBO), a novel framework for optimizing functional responses under uncertainty, where the en- tire mapping from design inputs to responses may vary within a structured function space. FRBO assumes the true response function lies within a smooth, norm-bounded region surrounding an unknown reference function. We model this ambiguity using a Reproducing Kernel Hilbert Space (RKHS) ball and derive a robust surrogate objective that captures both interpolation error and epistemic uncertainty. This approach enables principled acquisition without relying on posterior sampling, facilitating non-parametric robustness in black-box optimization settings. FRBO accommodates both scalar and functional-valued responses and is applicable across diverse domains such as optics and materials design. Beyond providing theoretical guarantees on worst-case regret, we validate our approach through numerical studies, demonstrating that FRBO consistently outperforms existing baseline methods.

Estimation and Prediction of Time-in-range (TIR) with Inpatient Continuous Glucose Monitoring

Limin Peng
Department of Biostatistics and Bioinformatics
Rollins School of Public Health, Emory University
Collaborators/co-authors: Qi Yu, Guillermo Umpierrez

Continuous glucose monitoring (CGM) has been increasingly used in US hospitals for the care of patients with diabetes. Time-in-range (TIR), which measures the percent of time over a specified time window with glucose values within a target range, has served as a pivotal CGM-metric for assessing glycemic control. As inpatient glycemic control generally involves clinical decisions (e.g., insulin adjustments) at sequential time points, dynamic prediction of TIR is of high interest to both clinicians and patients. However, this task is prone to multi-fold complications, which include a complex missing mechanism inherent to inpatient CGM data, the boundedness constraint to TIR which precludes straightforward regression modeling that typically requires some linearity assumption, and the presence of a large number of potential predictors. To address these challenges, we propose a random forest procedure which can accommodate nonlinear effects and interactions between predictors while simultaneously performing variable selection and dynamic prediction. Through utilizing a newly proposed nonparametric estimator of TIR, our proposal properly handles the complex data missingness associated with inpatient CGM data. Results from our numerical studies demonstrate the advantages of the proposed method over benchmark approaches.

A Bayesian Blind Source Separation Framework for Uncovering Reliable Neural Circuitry in the Developing Connectome

Yaotian Wang
Department of Biostatistics and Bioinformatics
Rollins School of Public Health, Emory University
Collaborators/co-authors: Ying Guo

Numerous studies have shown that neural circuits are key to understanding brain development, yet delineating them within the brain's complex, high-dimensional connectome remains challenging. Here, we introduce a novel Bayesian blind source (BSS) separation framework that decomposes connectivity data to uncover latent neural circuits. It flexibly incorporates diverse forms of domain knowledge (e.g., spatial and functional information) about brain regions via a novel hierarchical Ising-SBM prior. It also enables estimation of group-specific latent circuits to capture heterogeneity in the circuitry that drives population differences (e.g., sex- and age-related differences). We apply it to resting-state functional connectivity data from the Lifespan Human Connectome Project in Development (HCP-D; ages 8–21 years; $n = 602$), revealing neural circuits and their sex-specific development. These circuits are successfully reproduced in independent data from the Philadelphia Neurodevelopmental Cohort (PNC; ages 8–21 years; $n = 494$), supporting their validity as genuine neural circuits during development. The findings obtained through our Bayesian BSS framework advance understanding of the developing connectome.

Modernizing Time Series Forecasting: Reuniting Statistical Structure with Deep Learning through Transformer

Shihao Yang
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
Collaborators/co-authors: Jiecheng Lu

The accelerating adoption of deep learning in time series forecasting has brought renewed attention to Transformer-based architectures. Despite their empirical success, many existing models lack interpretability and alignment with traditional statistical principles, often underperforming in multivariate or long-horizon settings. This talk attempts to bridge classical time series models with modern deep architectures, advancing both accuracy and interpretability.

We begin by revisiting linear attention in Transformers through a Vector Autoregressive (VAR) lens, revealing structural misalignments in current deep models and introducing SAMoVAR, a VAR-aligned Transformer that restores interpretability and forecasting fidelity. Building on this foundation, we propose WAVE, a novel attention mechanism integrating ARMA dynamics via implicit moving average weights, boosting the ability to capture both local and global temporal dependencies efficiently.

Addressing scalability and context-awareness, we introduce the In-Context Time Series Predictor, which reformulates time series forecasting tasks for efficient in-context learning, eliminating the need for parameter updates while maintaining competitive performance in full, few-, and zero-shot scenarios. For complex multivariate forecasting challenges, we explore two orthogonal innovations: (1) ARM, which adapts to series-specific temporal patterns using techniques like Adaptive Univariate Effect Learning, Random Dropping, and Multi-kernel Local Smoothing; and (2) CATS, which constructs Auxiliary Time Series to enrich inter-series dependencies while minimizing model complexity.

Optimal Bounded-Influence Procedure for Robust Sequential Change-Point Detection

Ruizhi Zhang
Department of Statistics
University of Georgia

We study the problem of robust sequential change-point detection under Huber's gross error model. We incorporate ideas of the influence function from the classical offline robust statistics literature and propose a new definition called the false alarm influence function to quantify the robustness of general sequential change detection procedures. Then, we propose a general family of robust detection procedures and derive their false alarm influence functions. The result shows that they all have a bounded influence function, which implies that any single arbitrary outlier will have only a limited impact on their detection performance. Furthermore, we construct the optimal robust

bounded-influence procedure in that general family with the smallest detection delay, subject to the false alarm rate constraint and the false alarm influence function constraint. It turns out the optimal procedure is based on the truncated likelihood ratio statistic and has a simple form. Finally, we demonstrate the robustness and the detection efficiency of the optimal robust bounded-influence procedure through extensive simulations.

Tail Spectral Density Estimation and Its Uncertainty Quantification: Another Look at Tail Dependent Time Series Analysis

Ting Zhang
Department of Statistics
University of Georgia
Collaborators/co-authors: Beibei Xu

We consider the estimation and uncertainty quantification of the tail spectral density, which provide a foundation for tail spectral analysis of tail dependent time series. The tail spectral density has a particular focus on serial dependence in the tail, and can reveal dependence information that is otherwise not discoverable by the traditional spectral analysis. Understanding the convergence rate of tail spectral density estimators and finding rigorous ways to quantify their statistical uncertainty, however, still stand as a somewhat open problem. In the current talk, we aim to fill this gap by providing a novel asymptotic theory on quadratic forms of tail statistics in the double asymptotic setting, based on which we develop the consistency and the long desired central limit theorem for tail spectral density estimators. The results are then used to devise a clean and effective method for constructing confidence intervals to gauge the statistical uncertainty of tail spectral density estimators, and it can be turned into a visualization tool to aid practitioners in examining the tail dependence for their data of interest. Numerical examples including data applications are presented to illustrate the developed results.

Novel Empirical Likelihood Method for the Cumulative Hazard Ratio Under Stratified Cox Models

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

Evaluating the treatment effect is a crucial topic in clinical studies. Nowadays, the ratio of cumulative hazards is often applied to accomplish this task, especially when those hazards may be nonproportional. The stratified Cox proportional hazards model, as an important extension of the classical Cox model, has the ability to flexibly handle nonproportional hazards. In this article, we propose a novel empirical likelihood method to construct the confidence interval for cumulative hazard ratio under the stratified Cox model. The large sample properties of the proposed profile

empirical likelihood ratio statistic are investigated, and the finite sample properties of the empirical likelihood-based estimators under some different situations are explored in simulation studies. The proposed method was finally applied to perform statistical analysis on a real world dataset on the survival experience of patients with heart failure.

Mixture Modeling for Temporal Point Processes with Memory

Xiaotian Zheng
Department of Statistics
University of Georgia
Collaborators/co-authors: Athanasios Kottas, Bruno Sansó

In this talk, I will present a constructive approach to building temporal point processes that incorporate dependence on their history. The dependence is modeled through the conditional density of the duration, i.e., the interval between successive event times, using a mixture of first-order conditional densities for each one of a specific number of lagged durations. Such a formulation provides a tractable and scalable class of models for different types of point events. By specifying appropriate families of distributions for the first-order conditional densities, we can obtain either self-exciting or self-regulating point processes. From the duration-processes perspective, we develop a method to specify a stationary marginal density. The resulting model, interpreted as a dependent renewal process, introduces high-order Markov dependence among identically distributed durations. Furthermore, we provide extensions to cluster point processes, which can describe duration clustering behaviors attributed to different factors. The point process models are implemented within the Bayesian framework for inference, model assessment, and prediction. The methods will be illustrated with data examples from environmetrics and finance. Thi is joint work with Athanasios Kottas and Bruno Sansó from the University of California, Santa Cruz.

Industry Session

What Is It Like to Work at JMP Statistical Discovery LLC?

William Fisher
JMP Statistical Discovery LLC

JMP is a statistical software system that is used by scientists, engineers, and researchers to aid in a range of statistical tasks such as: analyzing data, building predictive models, designing experiments, and visualization. In this talk, I will share some of my experiences working at JMP Statistical Discovery LLC as a research statistician developer, and discuss what work looks like for other roles in R&D. Along the way, I will highlight some of the projects (both research and non-research related) I had the opportunity to work on as an intern and during my recent transition to being a full-time employee. I will also discuss more broadly the work that is done in other R&D roles such as testing and documentation. Lastly, I will conclude with a discussion on what internships at JMP are like for any interested students.

Data Science at UPCR

Andrew Neely
Unclaimed Property Consulting & Reporting (UPCR)

Unclaimed Property Compliance & Reporting (UPCR) helps organizations manage unclaimed property obligations—ensuring compliance with state regulations while reuniting rightful owners with their assets.

Our work sits at the intersection of data, technology, and finance, using analytics to solve real-world compliance challenges for clients across industries.

During this session, Andrew Neely, a UGA alumnus and Data Analyst Consultant at UPCR, will share insights from his own career path—from studying Management Information Systems (MIS) at UGA to applying data analysis in a consulting environment. He'll discuss what unclaimed property is and what UPCR does, how the company leverages statistics in decision making for client and internal projects, and what career and internship opportunities are available for students interested in applying their analytical skills to business and compliance.

UPCR actively recruits and mentors UGA interns, with the goal of developing talented students into full-time team members. We look forward to continuing our partnership with UGA and supporting the next generation of analytical professionals.

Posters

Parameter Estimation in Epidemiological Models Using the Sum of Absolute Deviations Method

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

Reliable parameter estimation is vital for accurate predictions in infectious disease modeling, especially during public health emergencies. The Least Squares (LSQ) method is traditionally favored for its computational efficiency and closed-form solutions, assuming normally distributed errors. However, LSQ is highly sensitive to outliers, which can lead to biased parameter estimates when dealing with noisy data which is a common scenario in real-world epidemiological studies. To address this issue, we introduce the Sum of Absolute Deviations (SAD) as a robust estimation technique that minimizes the absolute differences between observed and predicted values. Unlike LSQ, SAD is less affected by outliers because it imposes a linear penalty on residuals, making it better suited for handling heavy-tailed error distributions epidemiological datasets. This study evaluates the performance of SAD using both simulated and real-world infectious disease data, demonstrating its advantages over LSQ in scenarios with outliers or non-normally distributed errors. By adapting concepts from signal processing, where SAD has proven effective in recovering signals from corrupted data, we apply these techniques to epidemiological modeling. Our findings indicate that SAD not only enhances the robustness of parameter estimation but also improves the accuracy of epidemic forecasts, offering a promising alternative to conventional LSQ methods. These results have significant implications for real-time epidemic tracking, where robust estimation methods are crucial for guiding timely public health interventions. This research contributes to the expanding literature on robust parameter estimation methods and provides a framework for applying SAD to epidemiological models. We believe that SAD, given its successful application in other fields like signal processing, holds considerable potential for improving the reliability of forecasts in infectious disease modeling, ultimately supporting more effective public health strategies.

Imputing Alzheimer's Disease Pathologies in Living Participants Using Deep Learning Models Trained on Decedent Data

Conglin Bao
Department of Biostatistics and Bioinformatics
Rollins School of Public Health, Emory University
Collaborators/co-authors: Jingjing Yang, Qile Dai

Accurate characterization of Alzheimer's disease (AD) pathology in living participants remains a major challenge due to the lack of direct post-mortem validation. This study develops a deep learning framework to infer longitudinal AD pathologies—including amyloid, tangles, global pathology

(gpath) and NIA-Reagan scores—by leveraging clinical and cognitive trajectories from the Rush Religious Orders Study (ROS) and Memory and Aging Project (MAP).

We first trained and evaluated multiple recurrent neural architectures (LSTM, BiLSTM, LSTM-ReLU, BiLSTM-ReLU) using data from all autopsied participants. A five-fold cross-validation procedure was employed to identify, for each pathology, the optimal model architecture, the target-column configuration, and the set of hyperparameters based on the mean out-of-fold R^2 performance.

After determining the optimal configuration for each pathology, we conducted a second five-fold cross-validation, in which each fold's model generated full-time step pathology predictions for its 20% holdout test subset. The five sets of predictions were then combined to construct a complete, cross-validated "clean imputation" of pathology trajectories for all decedents.

Subsequently, the entire decedent dataset was used to retrain the corresponding models to obtain final network weights, which were then applied to all living participants to infer their longitudinal pathology profiles across all visits.

Imputed longitudinal pathologies were further used to perform unsupervised k-means clustering on decedents, using inferred baseline and five-year pathology change features, revealing distinct progression patterns potentially corresponding to subtypes of AD dementia. The same models and clustering approach were subsequently applied to living participants to identify analogous trajectory subgroups.

This framework demonstrates the feasibility of integrating longitudinal deep learning and unsupervised clustering to bridge the gap between observed clinical profiles and unobservable neuropathological progression in living adults, offering new insight into AD heterogeneity.

Joint Credible Distribution of Ranks for Unit Level Small Area Estimation

Arghadeep Basu
Department of Statistics
University of Georgia
Collaborators/co-authors: Gauri Sankar Datta, Abhyuday Mandal and Aditya Mishra

We propose a hierarchical Bayesian small-area estimation framework that augments the classical nested-error regression with a bivariate radial-basis penalized spline on spatial coordinates under weakly informative priors, thereby flexibly capturing both local covariate relationships and broad spatial trends. We also propose a framework, where instead of considering a spatial spline component we work on a random regression model. Posterior inference yields full joint distributions for population means of small areas, which we summarize via a simultaneous Cartesian credible hyperrectangle and a double stochastic ranking matrix that assigns every entity/small area a posterior probability for each possible rank. Applied to Acid Neutralizing Capacity data across 113 eight-digit Hydrologic Unit Code watersheds in the northeastern U.S., our method identifies the

top and bottom ranking basins within New England, the Mid-Atlantic, and Great Lakes/Ohio regions. The resulting probabilistic maps and rank-probability profiles provide all stakeholders in environmental and ecological research with uncertainty-aware guidance for prioritizing acidification mitigation.

Divide-and-Conquer Empirical Likelihood for the Fine-Gray Model in Large-Scale Competing Risks Data

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

The Fine-Gray model is widely used for modeling the cumulative incidence function (CIF) in the presence of competing risks. However, standard likelihood-based inference becomes computationally expensive in large-scale survival datasets due to complex, non-monotonic risk sets. We propose a novel divide-and-conquer empirical likelihood (DAC-EL) framework for scalable and statistically valid inference in the Fine–Gray model. Our method partitions the full dataset into K subsets, fits the subdistribution hazard model on each, and aggregates the results using empirical likelihood principles. We incorporate a Hessian-preconditioned formulation that enables efficient and robust score approximation without requiring full likelihood evaluation. Through extensive simulations, we demonstrate that the proposed DAC-EL Fine–Gray estimator achieves nominal empirical coverage and short confidence intervals across varying levels of censoring and model complexity. Our method is highly parallelizable, interpretable, and directly applicable to large biomedical or industrial competing risks datasets. Applications to real-world datasets highlight its scalability and practical effectiveness.

SpaDiff: Denoising for Sequence-based Spatial Transcriptomics via Diffusion Process

Jiazhang Cai
Department of Statistics
University of Georgia
Collaborators/co-authors: Yongkai Chen, Luyang Fang, Wenxuan Zhong, Guo-Cheng Yuan, Ping Ma

Spatial transcriptomics enables transcriptome-scale analysis with spatial resolution but suffers from spot-swapping, where RNA molecules drift from their true locations, introducing noise and reducing spatial specificity. We introduce SpaDiff, a denoising method that treats spot-swapping as a diffusion process. SpaDiff simulates the displacement of RNA molecules and reverses it to restore their original spatial distribution while preserving molecular counts. Evaluations on simulated and real data demonstrate that SpaDiff enhances spatial specificity of gene expression, improves data integrity, and supports more accurate downstream analyses such as clustering and spatial domain identification.

Age Effect Explorer: Bulk and Single-Cell Dashboards for Human Aging Studies

Menghui Chen
Department of Biostatistics and Bioinformatics
Rollins School of Public Health, Emory University
Collaborators/co-authors: Mingrui Li, Ronnie Y. Li, Jie Jiang, Zhaohui S. Qin

Understanding age-related transcriptional changes is crucial for clarifying the molecular mechanisms of human aging. First of all, we developed Age Effect Explorer, an interactive R Shiny dashboard that utilizes a large number of GTEx v10 transcriptomes from 54 tissues and 981 donors. Using gene-wise linear models of genes with age, sex and technical covariates (Benjamini–Hochberg FDR control), this application identified more than 10,000 genes significantly related to age and verified known markers such as EDA2R, as well as revealing tissue-specific patterns such as ribosomes, mitochondria and immune pathways. Based on this, we now expand the platform to the single-cell level by adding another single-cell dashboard. We integrated OneK1K and PsychENCODE sc/snRNA-seq, applied comparable regression models at the cell type resolution, and used CAMERA for pathway analysis to detect coordinated age-related perturbations. These two dashboards provide integrated and accessible frameworks for exploring aging biology from tissue to cell type.

Empirical Bayes Approach to Overall Ranking of Populations

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

Reliable ranking of populations is crucial in fields like engineering, public health, and policy. Rankings based solely on point estimates of means overlook sampling variability and can lead to misleading conclusions. This paper proposes a practical and theoretically grounded empirical Bayes (EB) approach for robust overall ranking in small-sample settings, extending recent hierarchical Bayes methods. Unlike prior EB work by Laird and Louis (1989, `https://doi.org/10.3102/10769986014001029`), which focused on individual rank intervals, our method provides joint interval estimates for the entire rank vector, identifying plausible rank orderings with specified confidence and identifying the most frequent ones. Additionally, for any individual entity, one can determine the most plausible ranks it may hold, along with their estimated posterior probabilities for a given confidence level. Furthermore, we extend the Klein, Wright, and Wieczorek (2020, `https://doi.org/10.1111/rssc.12402`) frequentist approach using simultaneous EB confidence intervals of the means. We illustrate the methods using an example with a known "gold standard" rank vector, enabling comparison with competing approaches.

Online Statistical Inference of Constrained Stochastic Optimization via Random Scaling

Xinchen Du
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
Collaborators/co-authors: Wanrong Zhu, Wei Biao Wu, Sen Na

Constrained stochastic nonlinear optimization problems have attracted significant attention for their ability to model complex real-world scenarios in physics, economics, and biology. As datasets continue to grow, online inference methods have become crucial for enabling real-time decision-making without the need to store historical data. In this work, we develop an online inference procedure for constrained stochastic optimization by leveraging a method called Adaptive Inexact Stochastic Sequential Quadratic Programming (AI-SSQP). As a generalization of (sketched) Newton methods to constrained problems, AI-SSQP approximates the objective with a quadratic model and the constraints with a linear model at each step, then applies a randomized sketching solver to inexactly solve the resulting subproblem, along with an adaptive random stepsize to update the primal-dual iterates. Building on this design, we first establish the asymptotic normality guarantee of *averaged* AI-SSQP and observe that the averaged iterates exhibit better statistical efficiency than the last iterates, in terms of a smaller limiting covariance matrix. Furthermore, instead of estimating the limiting covariance matrix directly, we study a new online inference procedure called *random scaling*. Specifically, we construct a test statistic by appropriately rescaling the averaged iterates, such that the limiting distribution of the test statistic is free of any unknown parameters. Compared to existing online inference procedures, our approach offers two key advantages: (i) it enables the construction of *asymptotically valid* and *statistically efficient* confidence intervals, while existing procedures based on the last iterates are less efficient and rely on a plug-in covariance estimator that is inconsistent; and (ii) it is *matrix-free*, i.e., the computation involves only primal-dual iterates themselves without any matrix inversions, making its computational cost match that of advanced first-order methods for unconstrained problems. We validate our theoretical findings through numerical experiments on nonlinearly constrained regression problems and demonstrate the superior performance of the random scaling method over existing inference procedures.

Prospective Association between Cognition and COVID-19 Serology: A Hurdle Analysis of Anti-N Positivity and Titer Intensity in HAALSI

Lihui Duan
Department of Biostatistics and Bioinformatics
Rollins School of Public Health, Emory University
Collaborators/co-authors: Johnathan Alex Edwards

Background. Cognitive function may shape COVID-19 exposure risk and immune response, yet evidence from community cohorts is limited. We examined whether cognition is associated with (a) anti-nucleocapsid (anti-N) seropositivity and (b) antibody titer intensity among seropositives.

Methods. We assembled a primary cohort from linked Survey and COVID records ($N = 2{,}235$), then partitioned it into two subcohorts: Survey-Only (SO; $N = 977$) and Survey-Lab (SL; $N = 1{,}258$). After de-duplication and cohort definition, cognitive variables were summarized via PCA into Factor 1 (objective cognition) and Factor 2 (subjective cognitive complaints). We removed obviously derived covariates, addressed multicollinearity using VIF, and screened covariates associated with both exposure and outcome to identify potential confounders. The final adjustment set for each cohort equaled the union of the screened set and prespecified clinical covariates (age, sex, education, wealth index, country/region, BMI category, diabetes, hypertension, dyslipidemia). Outcomes were modeled with a hurdle framework: (Step 1) logistic regression for anti-N positivity ($Y > 1$); (Step 2) among positives, intensity on the log scale, comparing log-normal OLS vs. Gamma(log) by AIC; heteroskedasticity assessed with Breusch–Pagan (HC SEs used if indicated).

Results. Step 1 (positivity): In the primary cohort, positivity was 52.7% and female sex was associated with higher odds of positivity (OR ≈ 1.43, $p = 0.001$). In SL, positivity was 53.0% and female sex again showed higher odds (OR ≈ 1.69, $p = 0.0009$); wealth index Q2 vs Q1 showed lower odds (OR ≈ 0.63, $p = 0.022$); "other" region had higher odds (OR ≈ 1.39, $p = 0.038$). In SO, positivity was 52.2%; higher values of cognition Factor 2 were associated with lower odds (OR ≈ 0.82 per 1-unit increase, $p = 0.025$), and older age with lower odds (OR ≈ 0.98 per year, $p = 0.013$). **Step 2 (titer among positives):** Log-normal OLS fit best in all cohorts (AIC: SL 1851.6 vs 6084.6; SO 1278.4 vs 4128.7; Primary 3040.6 vs 9985.7). Breusch–Pagan suggested no strong heteroskedasticity in SL ($p = 0.59$) or Primary ($p = 0.383$); SO showed evidence ($p = 0.0045$). In SL, cognition Factor 2 was positively associated with titers ($\beta = 0.132$; geometric mean ratio ≈ 1.14 per unit, $p = 0.026$).

Conclusions. Cognitive function showed cohort-consistent associations with COVID serology: PCA-derived cognition Factor 2 related inversely to seropositivity (SO) yet directly to titer intensity among seropositives (SL). Female sex was robustly associated with higher odds of anti-N positivity. Findings support cognition as a correlate of both infection likelihood and antibody magnitude, warranting longitudinal studies to clarify mechanisms and temporality.

Necessary Optimality Conditions in Multiple Hypotheses Testing under Exact Family-Wise Error Rate Control – a Path to Fast Computing

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

In the multiple hypotheses testing problem, the computation of the most powerful test is numerically challenging. In the state-of-the-art, searching for the most powerful test can be done via solving a constrained optimization problem. The dual problem has been derived, and strong duality has been established in a very general sense. However, solving the dual problem is still an

open question and could result in an exhaustive search, which is not numerically efficient. In this paper, under the exact family-wise error rate control, we proved some necessary conditions for the optimizer of the dual problem. These necessary conditions can lead to coordinate-wise fast algorithms, which may result in numerically efficient algorithms to solve the dual problem, which consequently solve the primal problem for identifying the most powerful test. We prove the linear convergence of our algorithm (that is, the computational complexity of our proposed algorithm is proportional to the logarithm of the reciprocal of the target error). To the best of our knowledge, this is the first time such a fast algorithm has been proposed for finding the most powerful test with family-wise error rate control.

Fast and Accurate Divergence Time Estimation Using the Pairwise Likelihood

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

Bayesian statistical methods are commonly used in phylogenetic inference due to its ability to flexibly model complex models and datasets. With the advent of next-gen DNA sequencing technologies, however, genome scale datasets pose problems for Bayesian computation since large phylogenetic tree topology spaces are difficult to explore efficiently and computing the likelihood function on large tree topologies is a computational bottleneck. For long sequences, the pairwise composite likelihood enjoys a significant computational advantage over the full likelihood. Since the pairwise likelihood loses information relative to the full likelihood, estimates of parameter uncertainty can be overoptimistic; we correct for this by adjusting the pairwise likelihood to match the moments of the pairwise likelihood ratio test (LRT) statistic to that of the full likelihood. We examine the performance of the pairwise likelihood via simulation and a dating analysis using a large dataset of 125 bird taxa.

Tail Index Estimation for Tail Adversarial Stable Time Series with an Application to High-Dimensional Tail Clustering

Jingying Gao
Department of Statistics
University of Georgia
Collaborators/co-authors: Hanyue Cao, Yu Shao, T. N. Sriram, Weiliang Wang, Fei Wen, Ting Zhang

For stationary time series with regularly varying marginal distributions, an important problem is to estimate the associated tail index which characterizes the power-law behavior of the tail distribution. For this, various results have been developed for independent data and certain types of dependent data. In this article, we consider the problem of tail index estimation under a

recently proposed notion of serial tail dependence called the tail adversarial stability. Using the technique of adversarial innovation coupling and a martingale approximation scheme, we establish the consistency and central limit theorem of the tail index estimator for a general class of tail dependent time series. Based on the asymptotic normal distribution from the obtained central limit theorem, we further consider an application to cluster a large number of regularly varying time series based on their tail indices by using a robust mixture algorithm. The results are illustrated using numerical examples including Monte Carlo simulations and a real data analysis.

Evaluating Machine Learning Models for Breast Tumor Classification: A Comparative Study

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

Accurate classification of breast tumors as benign or malignant is essential for early diagnosis and effective treatment planning. However, the optimal machine learning approach for leveraging multi-dimensional tumor characteristics remains incompletely explored. In this study, a comprehensive breast tumor dataset covering tumor characteristics and their classifications from multiple patients was analyzed to evaluate the performance of different machine learning models. The dataset exhibits distinct linear relationships among features, slight class imbalance, and significant scale variations, representing a binary classification problem with multi-dimensional feature interactions. Exploratory data analysis using ggplot2 violin plots identified key discriminatory features, and Principal Component Analysis (PCA) was performed to validate the linear separability of the data and guide model selection. Logistic regression and random forest algorithms were implemented and compared to balance interpretability with predictive power. Model evaluation employed a multi-dimensional metric system including F1 Score, sensitivity, specificity, Log Loss for probability prediction quality, and ROC curves with AUC values to measure comprehensive discriminative capability. The results revealed that the logistic regression model achieved superior performance on this dataset, demonstrating high prediction accuracy while maintaining excellent interpretability through its linear decision boundary. Comprehensive evaluation across multiple metrics confirmed the model's robust classification ability and clinical applicability. In addition, the linear characteristics of the data were fully validated, supporting the use of simpler models over complex ensemble methods for this particular dataset. Collectively, these findings underscore logistic regression as a highly effective and interpretable approach for breast tumor classification, highlighting its value as a clinically relevant tool for supporting diagnostic decision-making and guiding personalized treatment strategies in breast cancer care.

Forecasting with Physics-Informed Statistical Learning

Joseph Hart
Department of Mathematical and Statistical Sciences

Clemson University
Collaborators/co-authors: Chris McMahan, Hudson Smith

Forecasting the trajectory of infectious disease outbreaks is challenging, particularly at the start when data is sparse and noisy. Classical statistical approaches, such as ARMA, require stationary time series and careful tuning of parameters, while modern machine learning models, including LSTMs and Transformers, often demand lots of data to achieve accurate predictions. Mechanistic models, such as compartmental models (e.g., SIR, SEIR, SIHR), describe disease dynamics through differential equations but are difficult to calibrate to real-world data and may not capture all relevant complexities.

Here, we present a physics-informed machine learning approach (PIML) that integrates epidemiological knowledge with observed case and hospitalization data. Through regularization on the loss function, we bridge the gap between purely data-driven and purely mechanistic models. We demonstrate our approach on COVID data collected in the state of SC.

Statistical Meta-analysis to Investigate the Association of SNP (rs4680) in COMT Gene with Multiple Cancers

Umma Hafsah Himu
Department of Mathematics and Statistics
Georgia State University
Collaborators/co-authors: Md. Harun-Or-Roshid, Md. Rezwane Sadik, Md. Bayazid Hossen, Saroje Kumar Sarkar, Md. Nurul Haque Mollah, Yichuan Zhao

Previous meta-analyses have investigated the association of SNP (rs4680) in COMT gene with different cancers individually as well as collectively with multiple cancers. However, in some cases, meta-analysis results with multiple cancers did not support the cancer-specific meta-analyses results. It may be happened due to fewer individual studies and the selection of inappropriate statistical models. To make a consensus decision more accurately, we aimed to determine the association of SNP (rs4680) with multiple cancers by taking more individual studies and appropriate statistical models in the extended meta-analysis. We found 115 case-control datasets by the systematic review that comprises a total of 103308 samples (case: 44,764 and control: 58,544). Then we performed a comprehensive statistical meta-analysis to investigate the association of SNP (rs4680) in COMT gene with multiple cancers. The current meta-analysis with each genetic model indicated that the SNP rs4680 is not significantly associated with the overall cancer risk (p-value ¿ 0.05). The meta-analysis under the cancer-specific sub-group suggested that rs4680 is significantly associated with bladder, prostate, and esophageal cancers, whereas the previous meta-analysis of rs4680 with multiple cancers did not mention the risk of prostate cancer. We observed that the risk of esophageal cancer increases (under A versus G, AA versus GG, GA versus GG, AA+AG versus GG models), but the risk of prostate cancer decreases (under A versus G, AA versus GG models) due to the presence of A allele of rs4680. The Trial Sequential Analysis (TSA) results

also supported the association of rs4680 with the risk of prostate and esophageal cancers. However, additional studies are needed to evaluate the association of rs4680 with the bladder cancer risk by TSA. Moreover, our analysis revealed that the presence of the A allele of SNP (rs4680) is associated with the increasing overall cancer risk in Asian populations, while it is associated with the decreasing overall cancer risk in African and mixed populations. Conversely, the G allele shows the opposite trend. The findings of this study recommended that the COMT gene may be utilized as the diagnostic and prognostic biomarker for some particular cancers, including prostate and esophageal cancers.

Identifying Quantitative EEG Biomarkers to Evaluate Alzheimer's Disease Severity Using a Standardized Preprocessing Pipeline

Youwei Hu
Department of Biostatistics and Bioinformatics
Rollins School of Public Health, Emory University
Collaborators/co-authors: Navnika Gupta, Andres Rodriguez, Allan Levey, Jim Lah, and Tianwen Ma

Quantitative EEG (QEEG) has become a promising biomarker for assisting in the diagnosis and prognosis of Alzheimer's disease (AD). This study aimed to extract spectral-domain features from EEG recordings to assess AD severity in a cohort of ICU patients through a reproducible preprocessing pipeline. Resting-state EEG data from patients with AD and mild cognitive impairment (MCI) were collected and preprocessed using a seven-step workflow: (1) Extract the first 60 minutes for long recordings, (2) Select 19 scalp electrodes and perform a bandpass filter (0.5–100 Hz), (3) Perform preliminary artifact removal with Clean Rawdata and Artifact Subspace Reconstruction (CR-ASR), (4) Conduct independent component analysis (ICA) and remove noise components via Multiple Artifact Rejection Algorithm (MARA), (6) Re-reference to the common average, and (7) Assess the effectiveness of the signal preprocessing. Spectral-domain features including absolute, relative band powers (delta, theta, alpha, and beta), slow (delta+theta), fast (alpha+beta), delta–alpha ratio (DAR), (theta+alpha)/delta, delta/(alpha+beta), and (delta+theta)/(alpha+beta) were calculated. The Wilcoxon rank-sum test was applied to compare continuous variables between groups, and the Pearson correlation was used to compare continuous variables with continuous variables. Thirty-one patients from the Emory Neurology clinic were included (6 MCI, 25 AD; median age = 63 years; median MoCA = 17; median abeta/tau index = 0.4; median DAR = 3.3). Although none of univariate comparisons of QEEG features between groups were statistically significant ($p < 0.05$), patients with AD had lower relative alpha power (Median: 0.1, IQR: 0.1-0.2) than patients with MCI (Median: 0.2, IQR: 0.1-0.3, p=0.08). In addition, MoCA scores were significantly correlated with DAR (r = –0.49, p = 0.007) and relative delta power (r = –0.56, p = 0.002), and positively correlated with relative alpha power (r = 0.41, p = 0.028). Future work will expand the cohort and incorporate advanced features from connectivity and complexity domains to explore the potential of QEEG in non-invasive diagnosis and tracking disease progression.

Neural Network Models for Group Testing Data

Yu Huang
School of Mathematical and Statistical Sciences
Clemson University
Collaborators/co-authors: Christopher Steven McMahan

Group testing is a cost-effective strategy for screening large populations for infectious diseases. Instead of testing individuals one at a time, biospecimens (e.g., blood, urine, swabs) are pooled and tested together, substantially reducing overall testing costs. In infectious disease screening programs that make use of group testing, it is often desirable to relate individual-level covariates (e.g., age, gender, symptoms) to the underlying infection status, typically through regression methods. However, this task is challenging because individual infection statuses are unobserved; i.e., they are masked by the effects of imperfect testing and potentially the pooling protocol. While regression methodologies that address these issues have been developed, existing approaches generally lack the ability to automatically detect and account for nonlinear associations and higher-order interactions between the covariates and the infection status. To address these limitations, we propose a neural network framework for group testing data that automatically detects and accounts for nonlinear relationships and high-order interactions if they are present. We evaluate the performance of our approach through extensive numerical studies and further demonstrate its practical utility using Chlamydia group testing data collected by the Iowa Public Health Laboratory.

Shape Based Analysis of Subcortical Structures for Alzheimer's Disease Classification

Collin Hunt
Department of Statistics
University of Georgia
Collaborators/co-authors: Rongjie Liu

Magnetic Resonance Imaging (MRI) has become a prominent tool in the research and diagnosis of neurological diseases due to the detailed anatomical information it provides while exposing patients to less radiation than its counterparts. Each scan produces a volumetric grid that shows tissue density within each 1mm^3 unit of the brain. This unique data type has prompted medical researchers and statisticians alike to develop novel methods of modeling and analysis. Shape analysis is a promising approach that strikes a balance between interpretability and performance when compared to traditional models trained on numeric summaries, such as volume and surface area, or black-box models trained on unprocessed data. This project develops a pipeline for processing MRI data into functional principal components that encode shape features for visualization of volume-based morphometry and the training of statistical models. Previous research has found Alzheimer's Disease to be correlated with functional and structural decay of an anatomical structure known as the Corpus Callosum, which is widely recognized for facilitating communication between the two brain hemispheres. A dataset of volumetric MRIs labeled as cognitively normal, mildly impaired,

or Alzheimer's disease will be processed through this pipeline to construct a decision-tree classifier, providing a framework for interpretable shape-based biomarkers of neurodegeneration.

Methodological Approaches for the Estimation of Confidence Intervals on Partial Youden Index under Verification Bias

Sihan Jia
Department of Mathematics and Statistics
Georgia State University
Collaborators/co-authors: Shirui Wang

The advancement of precision medicine hinges on accurately tailored diagnostic strategies yet estimating reliable confidence intervals (CIs) for the maximal partial Youden Index under verification bias presents considerable challenges, especially within critical false positive rate (FPR) ranges (e.g., (0,0.1), (0.05,0.2)) vital for specific clinical applications. While previous work established the partial Youden Index framework, and methods like Full Imputation (FI), Mean Score Imputation (MSI), Inverse Probability Weighting (IPW), and Semiparametric Efficient (SPE) address verification bias, robustly integrating these for the partial index across demanding FPRs has needed further development. This paper significantly advances this area by adapting and applying these four bias-correction techniques to estimate the partial Youden Index and its confidence interval (CIs) under verification bias. We systematically evaluate their performance with the proposed (bootstrap-based, MOVER) CI construction approaches. Extensive simulations demonstrate distinct method-specific patterns across verification proportions and FPR ranges, revealing the complexities in achieving reliable estimates. Bootstrap-based CIs exhibit greater robustness to model misspecification, a common clinical uncertainty, while analytical CIs often face undercoverage issues. A cardiovascular disease biomarker analysis corroborates these findings, showing Blood Pressure's superior discriminatory capability compared to Pulse Rate. Operating under the Missing at Random (MAR) assumption, these results offer crucial, updated guidance for CI estimation in diagnostic studies with incomplete verification, providing significant value where precise evaluation in specific FPR regions is paramount and complete verification is unfeasible. Our findings enhance the statistical foundation for diagnostic test evaluation, extending beyond previous work by comprehensively addressing the partial Youden Index with these updated verification bias correction and CI formula applications.

MRI2Sex: A 2D CNN Model for Sex Classification from MRI Scans

Xiaoye Jiang
Department of Biostatistics and Bioinformatics
Rollins School of Public Health, Emory University
Collaborators/co-authors: Zhaohui Qin

We developed **MRI2Sex**, a 2D convolutional neural network (CNN) framework for sex classification from brain MRI. To verify its validity, we first conducted a simulation study by modifying

voxel values within selected 10×10×10 brain cubes in female images using a normal distribution (mean = 2×original, SD = 1). The model accurately detected these artificial changes, confirming sensitivity to localized structural variation.

Applying MRI2Sex to real MRI data, each of the sagittal, coronal, and axial mid-slices was processed by an independent CNN. The sagittal plane achieved the best performance, with mean Matthews correlation coefficients (MCC) of 0.42–0.52 across thresholds. Predicted probabilities showed clear bimodal distributions, indicating effective class separation. Despite reduced 2D input, MRI2Sex demonstrated strong performance and biological interpretability. Future work will extend this approach to 3D CNNs and other clinical outcomes in balanced datasets.

Dynamic Surrogate Modeling Methods for Online Optimization

Katherine Kreuser
Department of Mathematical and Statistical Sciences
University of Clemson
Collaborators/co-authors: Whitney Huang

Real-time decision-making often requires optimizing an objective function that evolves over time. In many applications, the objective function is only partially observed and can be expensive to evaluate due to computational or monetary costs. This motivates the need for a data-driven dynamic surrogate modeling approach capable of curve fitting time-varying objectives under limited information. Specifically, we propose a dynamic surrogate modeling strategy based on Gaussian Process (GP) regression with autoregressive (AR) structures to enable curve fitting and dynamic modeling of evolving objectives, and it is integrated into an online optimization framework, allowing for adaptive decision-making. We illustrate the proposed strategy in a simplified autonomous vehicle application. Preliminary results indicate that our AR-GP modeling can enhance online optimization by providing accurate time-varying function reconstruction and proving improving downstream decision quality.

Flow-based Conformal Prediction for Multi-dimensional Time Series

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

Time series prediction underpins a broad range of downstream tasks across many scientific domains. Recent advances and increasing adoption of black-box machine learning models for time series prediction highlight the critical need for reliable uncertainty quantification. While conformal prediction has gained attention as a reliable uncertainty quantification method, conformal prediction for time series faces two key challenges: (1) adaptively leveraging correlations in features and

non-conformity scores, and (2) constructing prediction sets for multi-dimensional outcomes. To address these challenges jointly, we propose a novel conformal prediction method for time series using flow with classifier-free guidance. We provide coverage guarantees by establishing exact non-asymptotic marginal coverage and a finite-sample bound on conditional coverage for the proposed method. Evaluations on real-world time series datasets demonstrate that our method constructs significantly smaller prediction sets than existing conformal prediction methods while maintaining target coverage.

From Pixel to Transcript: Evaluating the Effective Subcellular Resolution of Visium HD Spatial Transcriptomics

Tianqi Li
Department of Biostatistics and Bioinformatics
Rollins School of Public Health, Emory University
Collaborators/co-authors: Jian Hu

Spatial transcriptomics has emerged as a leading frontier in genomics, with platforms presenting trade-offs between spatial resolution and transcript diversity. Among them, Visium HD achieves 2 µm resolution with whole-transcriptome profiling, whereas Xenium offers single-transcript resolution (280 nm) but limited gene coverage. In this study, we evaluate Visium HD's capability to infer RNA localization patterns at the subcellular scale using mouse brain tissue. Xenium-derived distributions of selected marker genes serve as the ground truth, providing a reference for assessing Visium HD's accuracy in recapitulating spatial patterns. We specifically analyzed neuronal and synaptic marker genes, comparing their cytoplasmic enrichment across the two platforms. Our results show that despite its high spatial resolution, Visium HD cannot fully capture the single-transcript distributions observed in Xenium. This limitation restricts its utility for studying subcellular structures such as synapses and dendrites and underscores the need for careful interpretation of Visium HD data at near-subcellular scales.

Unmasking Technical Zeros in Multi-omics: Challenges and Solutions

Zhe Liu
Department of Statistics
University of Georgia
Collaborators/co-authors: Aditya Mishra

Microbes are to be found in almost every type of environment. It is known that microbial communities can be a relevant factor in in fields such as biology, medicine and agriculture. Yet due to the nature of multi-omics sequencing technology, utilization of statistical models can be constrained by the sparse nature of microbiome data. The inflated zeros can cause bias in taxa abundance distribution and mask signals in discovery of important taxa. At the same time, it is challenging to distinguish technical zeros from biological zeros and impute the correct zeros. In this presentation

we use an existing method – mblmpute – to identify and recover technical zeros in a poultry gut microbiome dataset of farming practices. We evaluate this method by comparing the statistical analysis results from original dataset and imputed dataset. We discuss the challenges of current imputation methods for multi-omics data and the future directions to encounter such challenges.

STAR: Shared-Tree Adaptive Representation for Video Compression

Ziyu Liu
Department of Statistics
University of Georgia
Collaborators/co-authors: Xiangnan Feng, Rongjie Liu

We introduce Shared-Tree Adaptive Representation (STAR), a Bayesian framework designed for video compression. STAR efficiently handles both spatial and temporal redundancies in video compression, by applying Recursive Dyadic Partitioning (RDP) and domain adaptation: it reuses the RDP tree inferred from the previous frame and selectively updates only the regions with significant pixel changes across video frames. The RDP tree, inferred via a Bayesian model, serves as a nonlinear spatial transformation that adaptively partitions and prunes the image based on local intensity homogeneity which efficiently captures spatial correlations within each frame. To handle temporal redundancy, STAR introduces a tree-sharing strategy that aligns naturally with the principles of domain adaptation: instead of recomputing the RDP tree for every frame, it treats each frame as a new target domain and reuses the tree structure inferred from the previous frame, updating only the branches corresponding to regions with significant pixel changes. The hierarchical nature of the RDP tree enables such domain adaptation to be performed in a localized and interpretable manner, facilitating efficient adaptation to evolving video content while maintaining high compression quality. Unlike conventional codecs or deep learning-based methods, STAR requires no pre-training or external datasets, enabling fast deployment across diverse video sources. The resulting wavelet-based representation is not only parsimonious but also physically interpretable, making it also suitable for other downstream tasks. Extensive evaluations on surveillance video datasets show that STAR consistently outperforms JPEG, JPEG2000, H.264/AVC, H.265/HEVC, and deep learning-based method under high compression ratio requirements.

Tumor Mutational Burden and Its Association With Immune Response: A TCGA Bioinformatics Study

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

Tumor mutational burden (TMB) has emerged as a promising biomarker for predicting tumor immunogenicity and responsiveness to immune checkpoint inhibitors. However, the mechanistic

relationships between TMB, immune gene expression, and immune cell infiltration across different cancer types remain incompletely understood. In this study, somatic mutation and transcriptomic data from The Cancer Genome Atlas (TCGA) were analyzed across 33 tumor types to explore the association between TMB and immune response. TMB was defined as the total number of nonsynonymous mutations per megabase, and patients were stratified into high- and low-TMB groups. Differential expression analysis and gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment were performed to identify immune-related molecular changes. Immune cell infiltration levels were estimated using the CIBERSORT algorithm, and survival analyses were conducted through Kaplan–Meier and Cox regression models. The results revealed that high-TMB tumors showed increased expression of immune checkpoint molecules such as *PDCD1* (PD-1), *CD274* (PD-L1), and *CTLA4*, as well as enrichment of pathways involved in interferon signaling, antigen processing, and T cell–mediated cytotoxicity. In addition, computational deconvolution indicated elevated infiltration of $CD8^+$ T cells, NK cells, and activated dendritic cells in high-TMB samples. Survival analyses further demonstrated that elevated TMB correlated with improved overall survival in several tumor types, including melanoma, non-small cell lung cancer, and bladder cancer. Collectively, these findings underscore the strong link between TMB and immune activation, highlighting TMB as a clinically relevant biomarker for predicting immunotherapy response and guiding personalized cancer treatment strategies.

Methods for Estimating VE Using Routine School Testing Data with Differential Testing Behavior

Amy Moore
Department of Biostatistics and Bioinformatics
Rollins School of Public Health, Emory University
Collaborators/co-authors: Paige E. Harton, Allison T. Chamberlain, Elizabeth T. Rogawski McQuade, Natalie Dean

During the COVID-19 pandemic, many school systems implemented opt-in regular testing for students to track the spread of disease and detect cases early. Beyond the primary use of these testing programs as surveillance, the observational data collected from these programs can be leveraged to measure vaccine effectiveness (VE) among school-aged children. This project is specifically motivated by a data set collected by a large public school district over the course of 2 school years (2021-23). Despite the opt-in nature of the testing program permitting equal access to testing for all students, the motivating dataset shows evidence of differences in testing behavior between vaccinated and unvaccinated students, which violates the assumption of similar testing results between vaccination groups for estimating VE. To combat this issue, we explore strategies to adjust for differences in testing behavior observed over the course of the school year. We apply 2 methods for measuring VE to the observational data: a target trial emulation approach with matching of participants across vaccination groups and a test-negative design. For both methods, we compare losses to sample size due to study design, compare point estimates and confidence intervals for the estimation of VE, and consider additional sources of bias due to unmet assumptions for each adjustment strategy.

Localized Sparse Principal Component Analysis of Multivariate Time Series in Frequency Domain

Jamshid Namdari
Department of Biostatistics and Bioinformatics
Rollins School of Public Health, Emory University
Collaborators/co-authors: Robert T. Krafty, Amita Manatunga, and Fabio Ferrarelli

Principal component analysis has been a main tool in multivariate analysis for estimating a low dimensional linear subspace that explains most of the variability in the data. However, in high-dimensional regimes, naive estimates of the principal loadings are not consistent and difficult to interpret. In the context of time series, principal component analysis of spectral density matrices can provide valuable, parsimonious information about the behavior of the underlying process, particularly if the principal components are interpretable in that they are sparse in coordinates and localized in frequency bands. In this paper, we introduce a formulation and consistent estimation procedure for interpretable principal component analysis for high-dimensional time series in the frequency domain. An efficient frequency-sequential algorithm is developed to compute sparse-localized estimates of the low-dimensional principal subspaces of the signal process. The method is motivated by and used to understand neurological mechanisms from high-density resting-state EEG in a study of first episode psychosis.

An Analytical Framework to Precisely Characterize the Accuracy-Fairness Trade-off in Fair Representation Learning

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

We are developing an analytical framework to give a mathematically precise quantification of the accuracy-fairness trade-off in Fare Representation Learning (FRL). The main tool is to introduce a novel kernel-based formulation of the Equalized Odds (EO) criterion, denoted as EO_k, for FRL in supervised settings. The central objective of FRL is to mitigate discrimination regarding a sensitive attribute S while preserving prediction accuracy for the target variable Y. Our proposed criterion enables a rigorous and interpretable quantification of three core fairness objectives: independence ($\widehat{Y} \perp\!\!\!\perp S$), separation—also known as equalized odds ($\widehat{Y} \perp\!\!\!\perp S \mid Y$), and calibration ($Y \perp\!\!\!\perp S \mid \widehat{Y}$). Under both unbiased ($Y \perp\!\!\!\perp S$) and biased ($Y \not\perp\!\!\!\perp S$) conditions, we show that EO_k satisfies both independence and separation in the former, and uniquely preserves predictive accuracy while lower bounding independence and calibration in the latter, thereby offering a unified analytical characterization of the tradeoffs among these fairness criteria. We further define the empirical counterpart, $\widehat{\mathrm{EO}}_k$, a kernel-based statistic that can be computed in quadratic time, with linear-time approximations also available. A concentration inequality for $\widehat{\mathrm{EO}}_k$ is derived, providing performance guarantees and error bounds, which serve as practical certificates of fairness compliance.

Effect of Life Expectancy on Economy Growth for High-Income Nations

Kayode Okunola
Department of Mathematics and Statistics
Georgia State University
Collaborators/co-authors: Deborah Okunola, Oladimeji Adewuyi

The global age distribution has undergone substantial changes in recent years due to a rise in life expectancy. Based on projections, the global population of those aged 60 and beyond is expected to reach 2 billion by 2050, representing almost 25% of the total population. By the year 2050, it is expected that the proportion of adults aged 80 years and older will rise by 1% to 4% of the global population. Because of this trend, economic growth may be hampered. The growing reliance on elderly people results in an increase in taxation, while political pressures may cause public funding to be redirected to adult social care. If this option is made, it could be detrimental to both growth and investment. The present study uses panel data from high-income countries to determine if life expectancy is a favorable predictor of economic growth using Granger causality and panel regression. The Hausman test was used to evaluate pooled, random, and fixed effect models in order to determine which model was the most appropriate. Based on the results, the fixed effect model tends to perform better, as indicated by the p-value being less than 0.05. Furthermore, the findings convey that life expectancy has a negative impact on economic growth.

Effect of Three Intra-articular Injections on Patient Reported Knee Osteoarthritis Outcomes: A Retrospective Longitudinal Study

Zhengyi Ou
Department of Biostatistics and Bioinformatics
Rollins School of Public Health, Emory University
Collaborators/co-authors: Nicholas Hooper, Ramnik Gill, Kenneth Mautner, Tianwen Ma

Background: Intra-articular injections such as bone marrow concentrate (BMC), microfragmented adipose tissue (MFAT), and platelet-rich plasma (PRP) are increasingly used for knee osteoarthritis, but comparative effectiveness over time remains unclear.

Purpose: We aim to examine the association between longitudinal treatment outcomes, Knee Injury and Osteoarthritis Outcome Score-Junior (KOOS-JR) and Visual Analog Scale (VAS) and three injection methods, patient characteristics and disease severity.

Methods: The study utilized a national database collected from a US cloud-based software and platform company. Descriptive statistics of patient- and knee-level covariates were reported. Univariate comparison across injection methods were performed. Spaghetti plots of treatment outcomes were produced. Linear-mixed models (LMMs) were fitted to examine the association between KOOS-JR and VAS values over 18 months and injection methods, adjusting for age, sex, BMI, and arthritis severity with a random intercept for knee clustering. Missing values were reported and addressed via multiple imputation with 10 replications. We started by including main

effects only and adding interaction terms if possible. Bayesian Information Criteria was used for model selection. The final parameter estimates were computed by pooling the estimates calculated from each imputed dataset.

Results: *Patient-Level:* Mean age was 62.1 years (SD, 12.5), with MFAT patients older (65.3 years; $P < .001$) and higher BMI (29.9; $P < .001$) than BMC (61.3 years; BMI 28.3) and PRP (61.6 years; BMI 28.3).

Knee-Level: Among 3452 knees, disease severity ($P < .001$; 75.1% mild, MFAT with more moderate/severe cases) and compartment involvement ($P < .001$; BMC higher across compartments) were significant. *LMMs:* The KOOS-JR, on average, improved by 0.36 per month ($p < .001, 95\%CI : [0.31, 0.40]$), while VAS decreased by 0.06 per month($P < .001, 95\%CI : [0.05, 0.07]$). Both BMI, age, and disease severity were significant ($P < .001$). Injection methods and their interactions with time were not significant. Base models for both treatment outcomes were selected.

Conclusion: BMC, MFAT, and PRP showed comparable changes of KOOS-JR and VAS over time. Age, BMI, and disease severity were significant predictors, with no differential effects by disease severity nor injection type.

Supervised Subdata Selection of Big Data Using Weighted Support Points

Yumo Peng
Department of Statistics
University of Georgia
Collaborators/co-authors: V. Roshan Joseph, Abhyuday Mandal

The phenomenon of big data has become ubiquitous across disciplines, posing new challenges for statistical and machine learning methods due to high computational and storage costs. A practical solution is to fit models on a carefully selected subset of the data, making subdata selection a crucial step in large-scale analysis. Existing approaches range from random subsampling to optimal experimental-design-based methods, yet they often fail when different regions of the data require distinct allocation strategies, such as in imbalanced data classification. First, we propose a supervised data reduction method incorporating output information to guide subsampling by leveraging weighted support points, a model-independent method for data compression. Second, we develop an efficient optimization-based algorithm to solve for the resampling weights under a fixed sampling budget. Finally, we apply the proposed method to imbalanced classification problems and benchmark it against state-of-the-art resampling and cost-sensitive learning techniques and submodular-based coreset methods for data reduction. Results demonstrate consistent improvements in both predictive performance and computational efficiency.

Multi-task Learning of Cognitive Functions Using Brain Connectome Data

Guangbin Quan
Department of Biostatistics and Bioinformatics
Rollins School of Public Health, Emory University
Collaborators/co-authors: Xuan Kan, Yaotian Wang, Joshua Lukemire, Carl Yang, Ying Guo

Deep learning methods using functional magnetic resonance imaging (fMRI) have shown a strong potential to predict individual cognitive functions and mental disorders. However, traditional single-task learning approaches often fail to capture the shared neural mechanisms among different cognitive processes, and their clinical utility remains limited by insufficient interpretability. In this study, we developed a multi-task learning (MTL) model to jointly predict 12 cognitive functions using both resting-state and task-based functional connectivity data from the Adolescent Brain Cognitive Development (ABCD) dataset (N = 2,180). Model performance and the reliability of integrated gradients (IG)–derived predictive features were evaluated using cross-validation. We further conducted permutation tests on IG-based statistics to identify brain connections significantly contributing to cognitive predictions. The MTL model successfully captured interrelationships among correlated cognitive functions, demonstrating improved interpretability over single-task models. The IG-based predictive features exhibited high test–retest reliability across validation folds, indicating robust and consistent predictive capability. Furthermore, analysis of significant predictive connections revealed convergent brain network patterns across three cognitive domains—General Ability, Executive Function, and Learning/Memory. Together, these findings highlight the effectiveness of the MTL framework in revealing shared and domain-specific neural substrates underlying diverse cognitive functions, offering a promising step toward interpretable deep learning in neuroimaging.

Private and Efficient Surveillance Using Group Testing

Ahmer Raza
School of Mathematical and Statistical Sciences
Clemson University
Collaborators/co-authors: Christopher S. McMahan, Rafael G. L. D'Oliveira

Understanding disease prevalence within a population is essential for guiding effective public health interventions. Surveillance studies must simultaneously address concerns of statistical utility, resource efficiency, and privacy. Group testing, initially developed to reduce diagnostic costs, has recently been shown to enhance statistical utility over individual-level testing when disease prevalence is low and diagnostic tests are imperfect. In this work, we demonstrate that group testing also provides inherent privacy benefits. Specifically, we develop an optimal group testing-based surveillance strategy for prevalence estimation, providing differential privacy guarantees under two practical resource constraints: a fixed number of diagnostic tests and a fixed number of participants. We illustrate our theoretical results by designing optimal chlamydia and gonorrhea surveillance strategies for the University of Iowa State Hygienic Laboratory.

Causal Discovery in Recursive Homogeneous Models with Heavy Tailed Innovations

Vishal Routh
Department of Statistics
University of Georgia
Collaborators/co-authors: Shuyang Bai

We study causal discovery at the population level for recursive homogeneous structural causal models on directed acyclic graphs with heavy-tailed noise. Each node is generated by a common aggregation function f that combines parental inputs and idiosyncratic noise. We assume f is 1-homogeneous, coordinatewise monotone, continuous, nonnegative, and vanishes at the origin, and that the noise variables $(\varepsilon_i)_i$ are i.i.d. and regularly varying with index $\alpha > 0$. Under these assumptions, we introduce the population causal tail coefficient Γ^* and show that it encodes the exact pairwise ancestral relation: $\Gamma^*_{ij} = 1$ if and only if i is an ancestor of j. Consequently, Γ^* induces a generational (layered) decomposition of the nodes. Leveraging exact ancestry together with this generational structure, we identify—at the population level—the transitive reduction of the DAG, i.e., the minimal edge set that preserves reachability.

Capturing Global Heterogeneous Associations Between Tensor Outcomes and Predictors

Anik Roy
Department of Biostatistics and Bioinformatics
Rollins School of Public Health, Emory University
Collaborators/co-authors: Limin Peng, Ying Guo, Amita Manatunga

In mental health research, it is of interest to understand how predictors such as clinical symptoms are associated with neuroimaging phenotypes, which are often represented as tensors of large dimensions. Investigating this problem can be complicated by the large sizes of neuroimaging tensors as well as their potentially complex relationships (e.g., nonlinear) to relevant clinical variables. In this work, we propose a robust two-step procedure to address these challengs. Specifically, we propose to captures the association between a clinical variable and a neuroimaging tensor by leveraging interval quantile index (IQI), and imposing a meaningful low-rank structure to achieve a parsimonious representation. A key premise of our approach is its ability to provide a coherent global summary of potentially heterogeneous associations, thereby enabling more powerful detection of neuroimaging features linked to clinical symptoms. We demonstrate the advantages of our approach over several benchmark methods through simulation studies. Finally, we apply our method to the Grady Trauma Project to uncover non-homogeneous association between resting-state fMRI connectivity and symptoms.

Modeling Tree Mortality Dynamics Using Logistic Regression and Survival Analysis: A Statistical Framework for Short-Term Forecasting

Dinuka M. Senevirathne
Warnell School of Forestry and Natural Resources
University of Georgia
Collaborators/co-authors: Sheng-I Yang, Dehai Zhao, Song Xiao, Quang Cao, Bronson Bullock, Stephen Kinane and Richard Chandler

Modeling tree mortality, a key process that affects forest productivity and ecosystem stability, remains a statistical challenge due to its complex, nonlinear, and age-dependent effects. This study integrates logistic regression and Cox proportional hazards models to quantify short-term (2–3 year) tree-level mortality in loblolly pine (*Pinus taeda*) plantations across the southeastern United States. Model performance was evaluated using AUC and Brier score, confirming that both models accurately captured mortality patterns with high predictive accuracy (AUC = 0.6–0.9). Notably, age-based logistic and segmented Cox models showed similar prediction accuracy. The analysis identified tree size (diameter at breast height), stand density, and physiographic region as key mortality drivers, with risk increasing significantly in dense, younger stands. This statistical framework provides a robust and adaptable tool for forecasting forest resilience, offering valuable insights for sustainable forest management under diverse environmental conditions.

Assessing AMF Colonization and Disease Susceptibility in Sorghum Accessions Using Linear Mixed-Effects Models: Treatment vs. Control

Sneha Shadija
Department of Genetics, Institute of Bioinformatics
Department of Statistics
University of Georgia
Collaborators/co-authors: Jonathan Arnold, Abhyuday Mandal

Arbuscular mycorrhizal fungi (AMF) play a critical role in plant nutrient uptake and disease resistance. This study describes a greenhouse experiment conducted to further investigate the relationship between AMF colonization and disease resistance in a controlled environment. Twelve sorghum accessions were selected and grown under treatment and control conditions to test the hypothesis that AMF colonization provides protective mechanisms against foliar pathogens. The controlled environment allowed us to isolate variables and confirm findings from prior field studies, eliminating external factors that might influence outcomes. We hypothesize that treatment will enhance AMF colonization and reduce disease susceptibility, with responses varying across accessions due to genetic differences. Data will be analyzed using linear mixed-effects models to evaluate the effects of treatment, accession, and their interaction. This approach aims to elucidate the genotype-dependent impact of AMF on plant health and provide insights for targeted breeding and sustainable management strategies.

Efficient Optimization of Plasma Radiation Detectors Using Imperfect Inference Models

Difan Song
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
Collaborators/co-authors: William E. Lewis, Patrick F. Knapp, C.F. Jeff Wu, V. Roshan Joseph

The configurations of instruments fielded on an experiment affect the amount of information captured and the quality of subsequent inference. We investigate the problem of optimizing plasma x-ray radiation detectors in a magneto-inertial fusion experiment at Sandia National Laboratories. It is impossible to directly measure properties such as the temperature of the thermonuclear fusion plasma produced in these experiments because of the extreme environment and destructive nature of the experiment. Among other diagnostics, several detectors are placed with significant standoff from the fusion target to capture the x-rays emitted by the fusion plasma, which can be used to infer some of its properties. To optimize the configuration of these detectors, a high-fidelity model (HFM) is used for simulating outputs and a low-fidelity model (LFM) is used for inference. We develop methods based on A- and L-optimality criteria that are efficient to compute while explicitly accounting for the discrepancy between the HFM and the LFM. The method allows us to find detector configurations that perform similarly to or better than the configuration obtained using an existing sampling-based optimization method while decreasing computational time by a factor of 50.

Robotics-Based Modeling of Poultry Flock Behavior for Disease Spread Simulation

Jacob Song
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
Collaborators/co-authors: Nathan Damen

Avian influenza, a highly infectious respiratory disease, remains a prominent issue in the poultry industry. It can lead to severe health complications within chicken populations, leading to large-scale economic losses in the poultry industry. Since 2022, 1689 flocks of birds and 168.62 million birds have tested positive for highly pathogenic avian influenza (HPAI). Robotics offers a novel way to simulate poultry house dynamics, providing insight into possible disease spread. We present a behavior tree framework designed to replicate fundamental chicken behaviors: (1) scanning (sporadic monitoring of the environment for safety), (2) feeding (modeled by pecking order), (3) drinking (mirroring the feeding pattern except with a drinker projected), and (4) resting (herding into groups). Each behavior was implemented as a branch of the model and tested in the Georgia Tech Robotarium simulator and then on the physical GRITSBot and GRITSBot X robots. The robots successfully exhibited the four targeted behaviors, illustrating the feasibility of modeling the dynamics of chicken behavior in a swarm environment. However, several challenges emerged: barrier certificates intended to prevent collisions were inconsistent, with higher success rates in smaller swarms. Larger groups frequently experienced collisions or boundary drift, causing program termination. This study demonstrates the potential of behavior tree models in robotic swarms to approximate chicken behavior, with applications for understanding the spread of avian influenza. Future efforts will focus on synchronizing multiple behaviors into one cohesive program, improving barrier certificate reliability, and scaling models to larger populations.

d-QPSO: an R Package for Searching for Efficient Designs

Sloka Sudhin
Department of Statistics
University of Georgia
Collaborators/co-authors: Joshua Lukemire, Abhuyuday Mandal

An experimental design ξ consists of the settings of all factors for each experimental run to be conducted. The corresponding responses are typically observed under some error, and thus statistical techniques are required to quantify the uncertainty about the resulting estimates of β. This uncertainty is a function of the Fisher information matrix (FIM). In many settings, the experimental response is modeled under a generalized linear model (GLM), in which case the information matrix can be written as $I(\xi, \beta) = X^T W_\beta X$, which depends on the true underlying values of β. Due to the complexity of the FIM and W_β as a function of the unknown parameters for GLMs, finding a design that is optimal is a highly difficult numeric and combinatorial problem that lacks analytic solutions except for specific cases. Lukemire et. al. (2019) proposed $d-$QPSO, a modified variant of quantum-based particle swarm optimization targeting this problem for binary response experiments. In this work, we develop a general software package, `dqpso`, capable of finding designs for experiments with outcomes modeled under a multitude of link functions. Moreover, our approach can be used to search for exact or approximate designs, locally optimal or Bayesian designs, and designs with mixed categorical and continuous factors. The superiority of the generated optimal designs are further validated by a sensitivity function.

Evaluating MHCflurry for Robust Peptide–MHC Binding Prediction with Length- and Allele-Specific Insights

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

Accurate prediction of peptide-MHC class I binding affinity is essential for neoantigen discovery and vaccine design. We evaluated the deep learning model **MHCflurry** using 97,389 peptide-allele pairs from IEDB. Prediction accuracy was quantified with Pearson correlation ($r = 0.70$) and classification metrics, achieving AUC-ROC = 0.85 and AUC-PR = 0.81. These results demonstrate strong linear correlation and robust discriminative ability. Stratified analysis revealed best performance for 8-10 mer peptides ($r \approx 0.71$, AUC-ROC ≈ 0.86), with degraded accuracy for longer peptides (e.g., $r = –0.08$ at 15-mer). Allele-specific evaluation indicated strong results on common human HLA alleles (e.g., $r = 0.80$ for HLA-A*02:01), but weaker performance for underrepresented non-human primate alleles (e.g., $r = 0.42$ for Patr-A*0101). Overall, our study highlights both the robustness and limitations of MHCflurry. While the model performs well across common peptide lengths and alleles, it tends to overestimate binding affinity, suggesting calibration could further improve accuracy. These findings provide guidance for developing integrated prediction pipelines that extend beyond binding affinity to antigen presentation and immunogenicity modeling.

Trait Regression for Brain Connectomes Based on Deep Parcellation Analysis

Zeliang Sun
Department of Biostatistics & Epidemiology
University of Georgia
Collaborators/co-authors: Chao Huang, Rongjie Liu, Xiaohe Chen

The study of human brain connectomes offers a unique framework for linking structural networks with cognitive and behavioral traits. Traditional atlas-based anatomical parcellation analysis (APA) relies on predetermined regions of interest, which often introduces bias and reduces reproducibility. To address these limitations, we propose Deep Parcellation Analysis (DPA), a data-driven regression pipeline for predicting human traits from structural connectomes. DPA integrates both local and global geometric features of white matter tracts, including three-dimensional coordinates, curvature, torsion, arc length, and global descriptors. These features are encoded into a latent representation using a novel fiber point autoencoder (fiberPAE), which combines local multi-layer perceptrons and symmetric pooling to capture nonlinear geometric information while preserving permutation invariance. Population-level clustering on these latent embeddings yields subject-specific parcellations, from which compositional connectomes are constructed. Trait associations are then modeled through sparse regression, enabling both prediction and interpretability by identifying fiber bundles that contribute to specific outcomes.

We evaluate DPA using diffusion MRI and behavioral data from 1,065 participants in the Human Connectome Project. Predictive performance was compared with several benchmark approaches, including Principal Parcellation Analysis (PPA), PCA-based clustering (PCPA), and APA-based methods such as APA-Lasso and APA-SBL. Across seven cognitive traits, DPA demonstrated notable improvements for language-related measures, particularly receptive vocabulary and oral reading recognition, where it consistently outperformed competing frameworks. Visualization of active fiber bundles revealed biologically plausible associations involving occipital, lingual, insular, and fusiform regions, aligning with established literature on language processing.

To further assess model components, we conducted ablation experiments. Excluding curvature and torsion reduced predictive accuracy, underscoring the contribution of local geometry. Omitting global descriptors also degraded performance, confirming the added value of global features. Replacing K-means clustering with Gaussian mixture models indicated that improvements were not solely dependent on clustering choice, while substituting Ridge for LASSO regression highlighted the interpretability benefits of sparsity-inducing penalties. These studies collectively emphasize the robustness and flexibility of the DPA framework.

In summary, DPA provides a scalable, atlas-independent approach for brain parcellation and trait prediction. By leveraging deep representation learning and data-driven clustering, it advances the study of connectome-based biomarkers and offers new opportunities for uncovering the structural basis of human cognition.

Optimizing Diagnostic Thresholds in Cardiovascular Risk: Youden Index Analysis with Machine Learning and Deep Learning Approaches

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

Accurate medical diagnoses require the selection of classification thresholds that balance sensitivity and specificity, directly influencing patient outcomes. I investigated threshold optimization using the Youden index in multiple machine learning models applied to a heart disease data set. Logistic regression achieved a Youden index of 0.796 at a threshold of 0.480, while random forests reached 0.731 at 0.500. A small neural network with Monte Carlo dropout produced a Youden index of 0.789 at 0.262, with a 95% confidence interval of [0.687, 0.917], providing a probabilistic measure of uncertainty. I explored multidimensional threshold surfaces and calibration analyses, revealing interactions between predictors that affect diagnostic performance. Comparison of traditional and uncertainty-weighted thresholds highlighted areas where predictions are less reliable, illustrating the value of integrating classical biostatistics with machine learning and uncertainty modeling. This work provides a reproducible framework to improve decision-making in clinical and biological applications, bridging statistical rigor and modern predictive methods.

Imputation of Missing Cancer Stage at Diagnosis Accounting for Stage-specific Survival

Shirui Wang
Department of Mathematics and Statistics
Georgia State University
Collaborators/co-authors: Farhad Islami, Rebecca L. Siegel, Ahmedin Jemal, Parichoy Pal Choudhury

Cancer stage at diagnosis is a primary determinant of patient prognosis and guides clinical decision-making. Yet, in the US population-based cancer registries, a substantial number of cases are recorded without stage information. This missingness introduces systematic bias into cancer epidemiologic studies, undermining the accuracy of stage-specific incidence, survival, and mortality rates. Existing imputation methods based on demographic and clinical predictors do not use information from stage-specific cancer survival. We propose an imputation method using an Expectation-Maximization (EM) algorithm that leverages stage-specific survival information to estimate the distribution of localized, regional, and distant disease among patients with missing stage at diagnosis. The proposed method is evaluated through comprehensive simulation studies and demonstrates strong performance and robustness across various clinically relevant scenarios. We apply our method to address missing stage information in the Surveillance, Epidemiology, and End Results (SEER) dataset for breast (sample size = 381,735 with 4,910 missing, 1.3%) and lung (sample size = 189,328 with 3,514 missing, 2%) cancers in ages 45-64 years and

observe that missing stage at diagnosis disproportionately corresponds to advanced stage disease. Among breast cancer cases with missing stage at diagnosis, the estimated proportions of regional and distant disease (with 95% bootstrap CIs) were 0.80(0.74,0.84) and 0.20(0.16,0.26) for all races combined; 0.78(0.75,0.82) and 0.22(0.18,0.25) for non-Hispanic White; 0.85(0.78,0.91) and 0.15(0.09,0.22) for non-Hispanic Black; 0.73(0.67,0.79) and 0.27(0.21,0.33) for non-Hispanic Other; and 0.85(0.80,0.91) and 0.15(0.09,0.20) for Hispanic, respectively. Localized disease was negligible across all groups. Among lung cancer cases with missing stage at diagnosis, the estimated proportions of regional and distant disease were 0.45(0.31,0.60) and 0.52(0.41,0.64) for all races combined; 0.51(0.45,0.57) and 0.49(0.43,0.54) for non-Hispanic White; 0.62(0.38,0.73) and 0.38(0.27,0.51) for non-Hispanic Black; 0.16(0,0.63) and 0.62(0.35,0.75) for non-Hispanic Other; and 0.32(0.09,0.48) and 0.68(0.52,0.84) for Hispanic, respectively. Localized disease was estimated at 0.004(0,0.03) for all races combined, 0.22(0,0.36) for non-Hispanic Other and negligible across all other groups.

GASDU: Gauss–Southwell Dynamic Update for Efficient LLM Fine-Tuning

Tao Wang
Department of Statistics
University of Georgia
Collaborators/co-authors: Ping Ma

Parameter-efficient fine-tuning (PEFT) is crucial for adapting large language models (LLMs), yet existing methods trade off accuracy, latency, and compute: some add inference-time modules, others fix a static parameter set that can drift from evolving gradients, and dynamic variants can be costly. We propose **GA**uss–**S**outhwell **D**ynamic **U**pdate (GASDU), which performs *periodic Gauss–Southwell-k* selection: every M steps it uses the current gradients to select the k largest-magnitude coordinates and updates only those entries while reusing the mask until the next refresh. The Top-k selection is implemented in a streaming, tile-wise way to avoid materializing dense gradients, making the amortized refresh cost negligible. Theoretically, under a local Polyak–Łojasiewicz condition, we prove that GASDU enjoys a linear convergence rate scaled by a measurable gradient-retention factor and show that the factor degrades sublinearly within each refresh window. This sublinear decay implies that a moderate M can maintain a high retention factor, which in turn explains GASDU's near–full–fine-tuning behavior. Empirically, GASDU sustains high retention between refreshes at an extreme parameter budget (0.01%) and consistently outperforms strong PEFT baselines and closely tracks or exceeds full fine-tuning across diverse commonsense and arithmetic reasoning benchmarks and LLMs (LLaMA-2/3 and GPT-OSS-20B). For training efficiency, it delivers $\sim$10.6$\times$ higher token throughput and $\sim$70% lower peak memory than full fine-tuning.

Statistical Inference of Constrained Model Estimation via Derivative-Free Stochastic Sequential Quadratic Programming

Jianliang Ye
Department of Mathematics
Georgia Institute of Technology
Collaborators/co-authors: Sen Na

We propose a derivative-free stochastic sequential quadratic programming (DFSSQP) method for solving nonlinear equality-constrained stochastic optimization problems using only zero-order information. Our algorithm estimates gradients and Hessians via randomized finite differences and introduces an online debiasing technique that aggregates past iterates to reduce bias without excessive memory cost. We establish global and local convergence, asymptotic normality of the averaged iterates, and a functional central limit theorem (FCLT) for the optimization path. Notably, we extend existing FCLT results to a double-averaging setting arising from the debiasing process, addressing technical challenges that do not appear in standard single-averaging schemes. Our method also relaxes restrictive step size conditions by allowing random stabilization, improving practical applicability. We discuss a sketching-based extension to reduce complexity, making DF-SSQP suitable for large-scale derivative-free inference and optimization tasks.

Extracting Viral Signatures from Complex Raman Spectra

Meizhi Yu, Yingchuan Zhang
Department of Statistics
University of Georgia
Collaborators/co-authors: Haoran Lu, Yanjun Yang, Jiaheng Cui, Yiping Zhao, Ping Ma

Surface-enhanced Raman Spectroscopy (SERS) enables virus detection at low concentrations but yields complex mixtures of viral, buffer, and noise components. We propose a physically constrained machine learning framework to extract true viral spectra while preserving spectral realism and robustness. Each SERS signal is modeled as a concentration-dependent combination of basis spectra, with the viral component represented by a ResNet-based neural network that learns spectral patterns across wavenumbers. The training objective integrates mean-squared error with additivity, nonnegativity, and flatness penalties to ensure physical plausibility. Fourier feature mapping further enhances reconstruction of subtle and high-frequency spectral structures. Experiments on SARS-CoV-2 and pyocyanin spectra demonstrate accurate and stable recovery across varying concentrations and noise levels. Analyses using signal-to-noise ratio and Dynamic Range Deviation metrics reveal data boundaries where reliable viral spectra can be reconstructed.

Automated Analysis of Experiments Using Hierarchical Garrote

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

In this work, we propose an automatic method for the analysis of experiments that incorporates hierarchical relationships between the experimental variables. We use a modified version of the nonnegative garrote method for variable selection which can incorporate hierarchical relationships. The nonnegative garrote method requires a good initial estimate of the regression parameters for it to work well. To obtain the initial estimate, we use generalized ridge regression with the ridge parameters estimated from a Gaussian process prior placed on the underlying input-output rela-

tionship. The proposed method, called HiGarrote, is fast, easy to use, and requires no manual tuning. Analysis of several real experiments are presented to demonstrate its benefits over the existing methods.

Region-specific Gene Co-expression Network Inference for Spatial Transcriptomics Data

Zhuoran (Angela) Yu
Department of Biostatistics and Bioinformatics
Rollins School of Public Health, Emory University
Collaborators/co-authors: Chang Su, Ying Ma

Gene co-expression networks (GCN), defined by correlations of gene expression, characterize gene regulatory relationship and biological pathways. Recent advances in spatial transcriptomics technologies enabled the inference of GCNs while incorporating spatial information. Understanding how GCNs vary across spatial regions can provide regulatory insights into the biological pathways underlying spatial heterogeneity. For instance, differences in GCNs between pathological and healthy tissue regions may reveal disease-associated gene regulation. To this end, we introduce a new statistical framework that characterizes GCNs across spatial regions while accounting for spatial correlations among cells. Our method employs a moment-based regression approach to robustly estimate region-specific GCNs and rigorously test differences in GCNs across spatial regions. Simulations demonstrate that the proposed model achieves proper type I error control and high statistical power.

Characterizing mRNA Localization in Polarized Neuronal Compartments

Chenyang Yuan
Department of Biostatistics and Bioinformatics, Rollins School of Public Health
Emory University
Collaborators/co-authors: K. Patel, H. Shi, H.-L. Wang, F. Wang, R. Li, Y. Li, V. Corces, H. Shi, S. Das, J. Yu, P. Jin, B. Yao, J. Hu

Background: Neurons are highly polarized cells with distinct distal compartments—such as synapses, dendrites, and axons—that enable connectivity and communication. Thousands of mRNAs have been shown to localize to these distal sites, supporting local protein synthesis critical for their plasticity and function. The diversity and regulation of these mRNAs are closely linked to neuronal function and are implicated in various neurodegenerative diseases. However, conventional RNA sequencing methods, such as single-cell/nucleus RNA sequencing, are limited to capturing nuclear and somatic transcripts, leaving the mRNA content of distal compartments unexplored. Recent advances in in situ spatial transcriptomics (iST) allow subcellular measurement of gene expression by capturing individual mRNAs in their spatial context. This resolution creates new opportunities to study compartmentalized transcripts. As most analytical pipelines still focus on

mRNAs located in the cell body, the primary challenge lies not in the technology itself, but in the development of appropriate analytical methods.

Methods: We present mcDETECT, a machine learning framework for studying mRNA localization in polarized neuronal compartments using iST data. mcDETECT first applies density-based clustering to identify RNA granules within distal subcellular regions. It then integrates information from nearby RNA granules to reconstruct compartment-specific expression profiles for individual neurons. These profiles reveal neuronal cell states that complement conventional cell type classifications based on nuclear mRNA.

Results: We applied mcDETECT to mouse brain datasets generated using various state-of-the-art iST platforms, including Xenium 5K, MERSCOPE, and CosMx. mcDETECT successfully identified RNA granules within polarized neuronal compartments and further classified them into distinct subtypes, each associated with specific functional roles across brain regions. Leveraging compartment-specific expression profiles, mcDETECT consistently uncovered previously unrecognized neuronal states across all different brain regions. In an Alzheimer's disease (AD) mouse model, mcDETECT revealed alterations in both compartmentalized RNA patterns and neuronal states prior to observable neuronal loss, potentially serving as molecular markers for early AD diagnosis and therapeutic intervention.

Conclusion: mcDETECT is the first method to systematically characterize mRNAs for polarized neuronal compartments, offering novel molecular insights and revealing potential therapeutic targets within these distal regions for neurodegenerative diseases.

Spatio-Temporal Epidemic Forecasting Using GNN and Transformer Informed by Mobility and Transmission Dynamics

Xiaotian Zhang
Department of Statistics
University of Georgia
Collaborators/co-authors: Pengsheng Ji, Yang Yang

Epidemic forecasting is fundamental to disease preparedness and control, yet many learning-based approaches struggle to capture both mechanistic transmission dynamics and time-varying data among different regions. Graph Neural Networks (GNNs) have emerged as a natural fit for spatially structured outbreak data, offering improved handling of graph data and more precise predictions. Recent hybrid models combined compartmental dynamics with GNNs (e.g., CausalGNN, MepoGNN, Epi-cola-GNN), but most either operated on a single region at a time or neglected future, time-varying mobility that reshapes transmission pathways. Here we propose a multi-region, mobility-aware hybrid framework that integrates transmission mechanism with a Graph Convolutional Network (GCN) for evolving connectivity, a spatio-temporal GRU encoder, and a Transformer for time-varying mobility. Using U.S. COVID-19 case and mobility data, we conduct rolling-window experiments to evaluate forecast accuracy and parameters. Empirical results indicate superior multi-horizon accuracy alongside identifiable, interpretable transmission parameters and flow estimates, enabling transparent epidemiological inference.

Participants

List of Participants

Last Name	First Name	Email	Affiliation
Abolade	Yisa	yabolade1@student.gsu.edu	Georgia State University
Akinbote	Abiodun	akinboteabiodunmary@gmail	Georgia State University
Allotey	Prince	Prince.Allotey@uga.edu	University of Georgia
Bai	Shuyang	bsy9142@gmail.com	University of Georgia
Bao	Conglin	cbao25@emory.edu	Emory University
Basnet	Man	mbasnet@uga.edu	University of Georgia
Basu	Arghadeep	argharivu2504@gmail.com	University of Georgia
Baxley	Maxwell	mmb54078@uga.edu	University of Georgia
Bouadoumou	Maxime	mbouadoumou@student.gsu.edu	Georgia State University
Chen	Hua Xuan	hua.xuan.chen@emory.edu	Emory University
Chen	Menghui	mche494@emory.edu	Emory University
Cheng	Cong	cong.cheng@uga.edu	University of Georgia
Choi	Kaeum	kchoi66@emory.edu	Emory University
Choudhary	Saurav	Saurav.Choudhary@uga.edu	University of Georgia
Chowdhury	Monsur	monsur.chowdhury@uga.edu	University of Georgia
Clarke	Emma	efc35927@uga.edu	University of Georgia
Daluwatumulle	Sekha	sdaluwa@emory.edu	Emory University
Datta	Deeya	dd20557@uga.edu	University of Georgia
Du	Xinchen	xinchendu@gatech.edu	Georgia Institute of Technology
Duan	Lihui	lihui.duan@emory.edu	Emory University
Dubey	Prasanjit	pdubey31@gatech.edu	Georgia Institute of Technology

Last Name	First Name	Email	Affiliation
Ejisoby-Nwosu	Ifeyinwa	iejisob@emory.edu	Emory University
Ellison	Greg	gmhellison@gmail.com	University of Georgia
Ewusi Dadzie	Samuel	sewusida@uga.edu	University of Georgia
Fisher	William	Bill.Fisher@jmp.com	JMP Statistical Discovery LLC
Floyd	Audrey	alf56122@uga.edu	University of Georgia
Fu	Zhirui	zfu66@emory.edu	Emory University
Garcia	Angelina	ag95547@uga.edu	University of Georgia
Gill	Mandev	mandev.gill@uga.edu	University of Georgia
Gong	Rui	rgong2@student.gsu.edu	Georgia State University
Guo	Anna	anna.guo@emory.edu	Emory University
Guo	Diqing	dg46076@uga.edu	University of Georgia
Hall	Daniel	danhall@uga.edu	University of Georgia
He	Yixuan	yixuan.he@emory.edu	Emory University
Himu	Umma Hafsah	uhimu1@student.gsu.edu	Georgia State University
Hart	Joseph	jchart@clemson.edu	Clemson University
Hu	Youwei	youwei.hu@emory.edu	Emory University
Huang	Hanwen	hhuang1@augusta.edu	Augusta University
Huang	Whitney	wkhuang@clemson.edu	Clemson University
Huang	Yijian	yhuang5@emory.edu	Emory University
Huang	Yu	yhuang3@clemson.edu	Clemson University
Hunt	Collin	cch79931@uga.edu	University of Georgia
Huo	Xiaoming	huo@gatech.edu	Georgia Institute of Technology
Jia	Sihan	jocelynjia0126@gmail.com	Georgia State University
Jiang	Xiaoye	xiaoye.jiang@emory.edu	Emory University
Jones	Alyssa	realamj2006@gmail.com	University of Georgia
Kapalavai	Venugopala Nikhil	nikhil.kapalavai@gmail.com	University of Georgia
Ke	Yuan	yuan.ke@uga.edu	University of Georgia
Kim	Hanna	hk20807@uga.edu	University of Georgia
Kim	Kyurhi	kyurhi.kim@emory.edu	Emory University
Krafty	Robert	rkrafty@emory.edu	Emory University
Kreuser	Katherine	kkreuse@clemson.edu	Clemson University
Lee	Jinae	jinaelee@uga.edu	University of Georgia
Lee	Junghwan	jlee3541@gatech.edu	Georgia Institute of Technology
Lei	Yumiao	yl79386@uga.edu	University of Georgia
Levina	Liza	elevina@umich.edu	University of Michigan
Li	Bingnan	bl17467@uga.edu	University of Georgia
Li	Tianqi	tianqi.li@emory.edu	Emory University

Last Name	First Name	Email	Affiliation
Li	Zilin	zilin.li@uga.edu	University of Georgia
Liu	Liang	lliu@uga.edu	University of Georgia
Liu	Rongjie	rjliu@uga.edu	University of Georgia
Liu	Xuran	xl85299@uga.edu	University of Georgia
Liu	Yuchi	yuchi.liu@emory.edu	Emory University
Liu	Yutong	yutong.liu@emory.edu	Emory University
Liu	Zhe	zhe.liu2@uga.edu	University of Georgia
Liu	Ziqi	ziqi.liu@emory.edu	Emory University
Liu	Ziyu	zl23565@uga.edu	University of Georgia
Lu	Yueqi	yueqi.lu@uga.edu	University of Georgia
Lukemire	Joshua	joshua.lukemire@emory.edu	Emory University
Ma	Ping	pingma@uga.edu	University of Georgia
Ma	Tianwen	ma3tian1wen2@emory.edu	Emory University
Ma	Wenpu	wma4@gsu.edu	Georgia State University
Manatunga	Amita	amanatu@emory.edu	Emory University
Mandal	Abhyuday	amandal@stat.uga.edu	University of Georgia
Mandal	Dhruba	dhrubacu24@gmail.com	University of Georgia
McDonnell	Brendan	bmm34505@uga.edu	University of Georgia
McNealey	Amaya	amcnelaey3@gatech.edu	Georgia Institute of Technology
Min	Yumin	myumin1@student.gsu.edu	Georgia State University
Mishra	Aditya	aditya.mishra@uga.edu	University of Georgia
Moore	Amy	amoor53@emory.edu	Emory University
Mosbo	Andrew	andrew.mosbo@uga.edu	University of Georgia
Na	Sen	senna@gatech.edu	Georgia Institute of Technology
Namdari	Jamshid	jamshid.namdari@emory.edu	Emory University
Ni	Yijin	yni64@gatech.edu	Georgia Institute of Technology
Okunola	Kayode	okunolakayusman@gmail.com	Georgia State University
Ou	Zhengyi	zou23@emory.edu	Emory University
Paul	Subhadeep	paul.963@osu.edu	The Ohio State University
Paynabar	Kamran	kpaynabar3@gatech.edu	Georgia Institute of Technology
Peng	Limin	lpeng@emory.edu	Emory University
Peng	Yumo	yp30751@uga.edu	University of Georgia
Qian	Weijia	weijia.qian@emory.edu	Emory University
Qin	Zhaohui	zhaohui.qin@emory.edu	Emory University
Quan	Guangbin	gquan2@emory.edu	Emory University
Ray	Connor	clr52393@uga.edu	University of Georgia
Raza	Ahmer	araza@g.clemson.edu	Clemson University
Reeves	Jaxk	jaxk@uga.edu	University of Georgia
Routh	Vishal	vr12741@uga.edu	University of Georgia

Last Name	First Name	Email	Affiliation
Roy	Anik	anik.roy@emory.edu	Emory University
Ryan	Riley	rr60552@uga.edu	The Working Data.com
Salter	Bella	mis12120@uga.edu	University of Georgia
Senevirathne	Dinuka M.	dms32073@uga.edu	University of Georgia
Seymour	Lynne	seymour@uga.edu	University of Georgia
Shadija	Sneha	sas99053@uga.edu	University of Georgia
Shenvi	Neeta	nshenvi@emory.edu	Emory University
Shi	Xinyu	xinyu.shi25@uga.edu	University of Georgia
Shiau	Justine	Justine.Shiau@uga.edu	University of Georgia
Song	Difan	dfsong@gatech.edu	Georgia Institute of Technology
Song	Jacob	songjacob81@gmail.com	Georgia Institute of Technology
Sriram	Tharuvai	tn@uga.edu	University of Georgia
Sudhin	Sloka	sps61703@uga.edu	University of Georgia
Sun	Xiaohan	xsun21@student.gsu.edu	Georgia State University
Sun	Zeliang	zs58484@uga.edu	University of Georgia
Suresh	Siddarth	siddarthsuresh00@gmail.com	Georgia State University
Sweeney	Stella	sps84839@uga.edu	University of Georgia
Wan	Zhihua	robert.wan@emory.edu	Emory University
Wang	Shirui	swang58@student.gsu.edu	Georgia State University
Wang	Tao	tw95546@uga.edu	University of Georgia
Wang	Wenyi	wwan258@emory.edu	Emory University
Wang	Xiaotian	xw17904@uga.edu	University of Georgia
Wang	Yaotian	ywan964@emory.edu	Emory University
Wang	Zihang	Zihang.wang@emory.edu	Emory University
Wells	Justin	jsw02185@uga.edu	University of Georgia
Werner	Mark	mwerner@uga.edu	University of Georgia
Whittington	George	George.Whittington@uga.edu	University of Georgia
Wilkes	Forrest	roy.wilkes40@gmail.com	University of Georgia
Wu	Charles	cw93334@uga.edu	University of Georgia
Wu	Wencheng	wwu227@emory.edu	Emory University
Xie	Yujia	yxie@gatech.edu	Georgia Institute of Technology
Xing	Shi	xs51911@uga.edu	University of Georgia
Xu	Baijia	baijia.xu@emory.edu	Emory University
Yang	Shihao	shihao.yang@isye.gatech.edu	Georgia Institute of Technology
Yang	Yang	yang.yang4@uga.edu	University of Georgia
Ye	Jianliang	jye347@gatech.edu	Georgia Institute of Technology
Yeh	Chi-Kuang	cyeh@gsu.edu	Georgia State University
Yu	Meizhi	mzyu018@gmail.com	University of Georgia
Yu	Wei-Yang	wyu322@gatech.edu	Georgia Institute of Technology

Last Name	First Name	Email	Affiliation
Yu	Zhuoran	angela.yu@emory.edu	Emory University
Yuan	Chenyang	chenyang.yuan@emory.edu	Emory University
Zeng	Yixuan	yz54048@uga.edu	University of Georgia
Zhang	Ruizhi	ruizhi.zhang@uga.edu	University of Georgia
Zhang	Ting	tingzhang@uga.edu	University of Georgia
Zhang	Xiaotian	zxtiantian34@uga.edu	University of Georgia
Zhang	Yingchuan	yz54720@uga.edu	University of Georgia
Zhang	Zilong	zzhang52@student.gsu.edu	Georgia State University
Zhao	Guantao	gzhao46@gatech.edu	Georgia Institute of Technology
Zhao	Jiamin	jiamin.zhao@emory.edu	Emory University
Zhao	Yichuan	yichuan@gsu.edu	Georgia State University
Zheng	Xiaotian	xzheng@uga.edu	University of Georgia

Georgia Statistics Day 2025

October 31, 2025