Featured Talks

Understanding the origin and spread of COVID-19

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

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

Unsupervised learning in data integration studies using JIVE with Gaussian mixtures

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

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

Invertible graph neural networks

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

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

Technical Sessions

Feature selection: the Markov boundary approach

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

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

Online Bayesian phylodynamic inference

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

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

On the testing of statistical software

Ryan Lekivetz
JMP
Collaborators/co-authors: Joseph Morgan

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

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

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

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

Gaussian process subspace prediction for model reduction

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

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

Canonical joint and individual variation explained

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

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

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

Causal and counterfactual views of missing data models

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

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

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

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

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

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

Image-based feedback control using tensor analysis

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

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

Optimal transport-based transfer learning for smart manufacturing

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

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

Big-data infectious disease estimation in COVID-19

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

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

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

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

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

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

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

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

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

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

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

Quickest detection of the change of community via stochastic block models

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

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