Keynote Lecture

Beyond tech: Machine learning in science and policy

David Dunson
Arts and Sciences Professor of Statistical Science
Duke University

Applications in the tech industry have driven much of the progress in machine learning in recent years. Most tech applications are fundamentally different in structure than applications in other fields, such as science and policy making. Hence, algorithms that can be highly successful in tech can be dramatically unsuccessful in other domains.

In this talk, I will provide a brief review of the types of application areas in which widely popular machine learning algorithms (e.g., deep learning) perform well and will highlight key differences between these settings and other areas. I will argue that there can be disastrous results in naively applying off-the-shelf ML algorithms in areas including criminal justice (e.g, automating sentencing and bail), science (e.g, neuroscience), policy (e.g., regulating chemical exposures), and health decision making. Instead, one needs to carefully develop targeted methods that deal with crucial issues of selection bias, uncertainty quantification, limited training data, and complex/high-dimensional observations. As illustration, I focus in more detail on two problems: (1) removing the influence of a sensitive variable (e.g, race/ethnicity) to obtain a fair predictive algorithm (also related to causal inference and privacy); and (2) obtaining interpretable predictive models of human traits based on an individuals brain connection structure.

Featured Talks

C. F. Jeff Wu
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
Collaborators/co-authors: Simon Mak

This talk introduces a novel method for selecting main effects and a set of reparametrized predictors called conditional main effects (CMEs), which capture the conditional effect of a factor at a fixed level of another factor. CMEs represent interpretable, domain-specific phenomena for a wide range of applications in engineering, social sciences and genomics. The key challenge is in incorporating the grouped structure of CMEs within the variable selection procedure itself. We propose a new method, cmenet, which employs two principles (CME coupling and CME reduction) to effectively navigate the selection algorithm. Simulations demonstrate the improved performance of cmenet over generic variable selection methods. Applied to a gene association study on fly wing shape, cmenet not only provides more parsimonious models and improved predictive performance over existing methods, but also reveals important insights on gene activation behavior which can guide further experiments.

Asympirical analysis: A new paradigm for data science

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

Large samples have been generated routinely from various sources. Classic statistical and analytical methods are not well equipped to analyze such large samples due to expensive computational costs. In this talk, I will present an asympirical (asymptotic + empirical) analysis in large samples. The proposed method can significantly reduce computational costs in high-dimensional and large-scale data. We show the estimator based on the proposed methods achieves the optimal convergence rate. Extensive simulation studies will be presented to demonstrate numerical advantages of our method over competing methods. I will further illustrate the empirical performance of the proposed approach using two real data examples.

Dynamic correlation analysis for high-throughput expression data

Tianwei Yu
Department of Biostatistics and Bioinformatics
Rollins School of Public Health, Emory University

Dynamic correlations are pervasive in high-throughput data. Large numbers of gene pairs can change their correlation patterns in response to observed/unobserved changes in physiological states. Finding changes in correlation patterns can reveal important regulatory mechanisms. Currently there is no method that can effectively detect global dynamic correlation patterns in a dataset. Given the challenging nature of the problem, the currently available methods use genes as surrogate measurements of physiological states, which cannot faithfully represent true underlying biological signals. In this study we develop a new method that directly identifies strong latent dynamic correlation signals from the data matrix, named DCA: Dynamic Correlation Analysis. At the center of the method is a new metric for the identification of pairs of variables that are highly likely to be dynamically correlated, without knowing the underlying physiological states that govern the dynamic correlation. We validate the performance of the method with extensive simulations. We applied the method to three real datasets: a single cell RNA-seq dataset, a bulk RNA-seq dataset, and a microarray gene expression dataset. In all three datasets, the method reveals novel latent factors with clear biological meaning, bringing new insights into the data.

Technical Sessions

Penalized multivariate count models for genomic data

Deepak Ayyala
Division of Biostatistics and Data Science at the Department of Population Health Sciences
Augusta University

Genomic and metagenomic experiments yield high dimensional count data which are extremely sparse, making parameter estimation very difficult. Features which are differentially expressed under two or more conditions are hard to detect due to the high sparsity. In this talk, I will present a penalized model for multivariate count data using Dirichlet-Multinomial distribution. The l2-based penalty function is designed to perform simultaneous gene selection and cell-type detection in single-cell RNA-seq experiments. We implemented a fast Newton-Raphson algorithm by avoiding large matrix inversions. The computation cost is reduced to a quadratic rate with respect to number of variables.

Leveraged subsampling for vector autoregression

Shuyang Bai
Department of Statistics
University of Georgia
Collaborators/co-authors: Ping Ma, Zenghan Wang, Rui Xie, Wenxuan Zhong

When estimating a model for streaming multivariate time series data, the computational capacity often cannot afford using all the data available. The sample size needs to be reduced and a subsample must be selected. In the context of vector autoregression models, we propose a subsample selection method based on a leverage score. The method is shown to have a performance superior to some naive approaches.

Nonparametric doubly-robust inference for the mean outcome under a longitudinal treatment decision rule

David Benkeser
Department of Biostatistics and Bioinformatics
Rollins School of Public Health, Emory University
Collaborators/co-authors: Peter Gilbert, Marco Carone

Doubly robust estimators are a popular tool for assessing causal effects of treatments and interventions. Such estimators are consistent for the parameter of interest if one of two nuisance parameters is consistently estimated. However, recent work has demonstrated that when adaptive nuisance estimators are utilized, the notion of double robustness does not extend to inference. We study this problem in the context of estimation of the mean counterfactual outcome under a

longitudinal treatment decision rule. The main contribution of this work is a theorem, which provides conditions on nuisance estimators that are sufficient to ensure that a substitution estimator of the target parameter is doubly robust with respect to asymptotic linearity, even when considering adaptive nuisance estimators. Moreover, the analytic form of such an estimators' influence function is derived, which provides a natural pathway to closed form, doubly robust confidence intervals and hypothesis tests. We propose two concrete algorithms for constructing estimators that satisfy the conditions of our theorem and evaluate their performance via simulation. Our study demonstrates that ignoring this issue has severe repercussions for statistical inference, while our proposed estimators remedy these issues.

Interpretable Machine Learning with Applications to Banking

Linwei Hu
Advanced Technologies for Modeling
Wells Fargo

Machine Learning (ML) algorithms have become popular in recent years due to their flexibility in model fitting and increased predictive performance. Despite this, they have not been widely adopted in banking and finance due to their opaqueness and lack of explainability, a necessary requirement to gain regulatory acceptance. There is much research currently to develop techniques to understand the model behavior and develop insights into the "black box". I will provide an overview of research being conducted by our team in this area with a focus on locally interpretable models and effects.

Factor-adjusted regularized model selection

Yuan Ke
Department of Statistics
University of Georgia
Collaborators/co-authors: Jianqing Fan, Kaizheng Wang

This paper studies model selection consistency for high dimensional sparse regression when data exhibits both cross-sectional and serial dependency. Most commonly-used model selection methods fail to consistently recover the true model when the covariates are highly correlated. Motivated by econometric studies, we consider the case where covariate dependence can be reduced through factor model, and propose a consistent strategy named Factor-Adjusted Regularized Model Selection (FarmSelect). By separating the latent factors from idiosyncratic components, we transform the problem from model selection with highly correlated covariates to that with weakly correlated variables. Model selection consistency as well as optimal rates of convergence are obtained under mild conditions. Numerical studies demonstrate the nice finite sample performance in terms of both model selection and out-of-sample prediction. Moreover, our method is flexible in a sense that it pays no price for weakly correlated and uncorrelated cases. Our method is applicable to a wide range of high dimensional sparse regression problems. An R-package *FarmSelect* is also provided for implementation.

Locating targets via wireless sensor networks

Rohit Patra
Department of Statistics
University of Florida
Collaborators/co-authors: George Michailidis, Moulinath Banerjee

Wireless sensor networks (WSNs) serve as key technological infrastructure for monitoring diverse systems across space and time. Examples of their widespread applications include: precision agriculture, surveillance, animal behavior, drone tracking, and emergent disaster response and recovery. A WSN consists of hundreds or thousands of identical sensors at fixed locations where each individual sensor observes the surrounding at fixed time intervals. In this work we estimate the location of a (signal emitting) target under the assumption that magnitude of signal detected at the sensor is a strictly decreasing function of the distance between the sensor and the signal emitting target. We propose an automated root-n-consistent estimator of the location the target under under only the monotonicity assumption. Our estimator is tuning parameter free. We show that our estimator has a Gaussian limit distribution and construct asymptotic confidence region for the location target.

A non-parametric Bayesian binary regression approach with application to latent class regression analysis

Nong Shang
Centers for Disease Control and Prevention

We consider a binary regression problem for which the success probability is generated from a random field over covariate domain. While it is difficult to specify a full random field for probability value changing from 0 to 1, for many practical problems, it is suffice to know the corresponding marginal point wise distributions. Based on a new perspective on statistical conditional estimation, we developed a non-parametric approach to construct the marginal point wise distributions from data with similarity among nearby distributions being reflected through the construction process. Such approach provides an approximate yet more efficient way to conduct non-parametric Bayesian binary regression. We investigated performance of the approach and explored its application in non-parametric latent class regression analysis.

Change-point detection for a network of Hawkes processes

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

Hawkes processes has been a popular point process model for capturing mutual excitation of discrete events. In the network setting, this can capture the mutual influence between nodes,

which has a wide range of applications in neural science, social networks, and crime data analysis. In this talk, I will present a statistical change-point detection framework to detect a change in the influence using streaming discrete events. Theoretical results are provided for controlling false alarms, characterizing the trade-off between the average-run-length and the expected detection delay, as well proving an online estimation procedure is nearly optimal.

Decentralized consensus optimization with delayed and stochastic gradients on networks

Xiaojing Ye
Department of Mathematics and Statistics
Georgia State University
Collaborators/co-authors: Benjamin Sirb

Decentralized consensus optimization has extensive applications in many emerging big data, machine learning, and sensor network problems. In decentralized computing, nodes in a network privately hold parts of the objective function and need to collaboratively solve for the consensual optimal solution of the total objective, while they can only communicate with their immediate neighbors during updates. In real-world networks, it is often difficult and sometimes impossible to synchronize these nodes, and as a result they have to use stale (and stochastic) gradient information which may steer their iterates away from the optimal solution. In this talk, we focus on a decentralized consensus algorithm by taking the delays of gradients into consideration. We show that, as long as the random delays are bounded in expectation and a proper diminishing step size policy is employed, the iterates generated by this algorithm still converge to a consensual optimal solution. Convergence rates of both objective and consensus are derived. Numerical results will also be presented.

A Monte Carlo simulation study: The number of minimum events and an adequate sample size required in Firth's penalized partial likelihood approach in Cox regression models

Chao Zhang
Biostatistics and Bioinformatics Shared Resource
Winship Cancer Institute, Emory University
Collaborators/co-authors: Jeanne Kowalski

Objectives: To determine the minimum number of events required in Firth's correction Cox regression models. Study Design and Setting: We used a series of Monte Carlo simulations to detect the impact of the number of events on the accuracy and precision of the estimated of parameters of the fitted models. For accuracy of regression coefficient of the models, the signed percent relative bias was used, and estimated of model variance, empirical simulation variance and their ratio of the fitted model were used to measure the precision of the models. Results: A minimum of approximately 6-8 events for a continuous covariate tended to result in estimation of regression coefficients with relative bias of less than 5%. And with this minimum number of events, the ratio

between the model variance and empirical simulation variance approximately equals 1, and around 100 models were converged with this minimum number of events. For a binary covariate, it also required at least 6-8 events for low-prevalence group to get the same conclusion as continuous covariate. Conclusion: Firth's bias correction Cox models require at least 6-8 events for a continuous predictor and same value for low- prevalence group of a binary predictor for adequate estimation of regression coefficient, standard error, and model convergence.

Towards Understanding Nonconvex Stochastic Optimization in Machine Learning

Tuo Zhao
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Stochastic Gradient Descent-type (SGD) algorithms have been widely applied to many nonconvex optimization problems in machine learning, e.g., training deep neural networks, variational Bayesian inference and collabrative filtering. Due to current technical limit, however, establishing convergence properties of SGD for these highly complicated nonoconvex problems is generally infeasible. Therefore, we propose to analyze the algorithm through a simpler but nontrivial nonconvex problems − streaming PCA. This allows us to make progress toward understanding SGD and gaining new insights for more general problems. Specifically, by applying diffusion approximations, we cast the solution trajectories of SGD as continuous time stochastic processes. This naturally characterizes the algorithmic behaviors and provides some principles on applying SGD to nonconvex optimization problems in machine learning.

Industry Session

What it is like to work for LexisNexis

Renu Midha
LexisNexis

LexisNexis combines cutting-edge technologies, unique data and advanced scoring analytics to provide innovative products and services addressing client needs in insurance, utilities, gaming, finance, government and healthcare. We process over 10,000 data sources daily adding to a database of over 2 petabytes of information which is used by our advanced statistical analysts to develop products through statistical analysis to help customers understand their individual business, driving growth and profit for both LexisNexis and all of LexisNexis' customers.

We at LexisNexis understand the uniqueness of our business and have training tracks set up for all employees, both directly out of college and those that have years of experience from other companies. These training tracks not only give our employees the knowledge on products that LexisNexis provides, but more so allow them to be innovators and strategic thinkers when future products are developed.

At LexisNexis, "We are innovators, passionate about challenging the status quo and improving outcomes."

Career opportunities at Wells Fargo for statisticians

Anqi Zou
Wells Fargo

Wells Fargo has a large "quant" community that is engaged in statistical and mathematical modeling. We're looking to recruit talented graduating PhDs, as well as a smaller number of Master's candidates, through the Quantitative Associate Program. Wells Fargo representatives will describe the program and various roles and responsibilities available at the bank, as well as giving examples of the role of Statistics in credit risk modeling and other areas.

The Wells Fargo **Quantitative Associate program** is designed to provide qualified candidates with the opportunity to gain comprehensive professional and industry experience that prepares them to develop, implement, calibrate, and/or validate various analytical models. The presenter will discuss aspects and benefits of the rotational program and how PhD and Master's students can apply.

The 12 month rotational program consists of two track selections: Our **Capital Markets** track gives Associates the opportunity to develop and validate mathematical models for pricing and hedging complex financial instruments. They will also educate the trading desk on the strengths

and weaknesses of models and provide model analysis. Our **Credit & Operational Risk** track gives Associates the opportunity to develop, maintain and validate statistical models for loss forecasting, credit risk scorecard, capital management, stress testing, detection and prediction of suspicious activity, and estimating bank-wide operational risk regulatory capital levels.

Advanced analytics at State Farm

Jeff Stoiber and Jeremy Mulcahey
State Farm

The analytics community at State Farm is faced with a wide variety of exciting and challenging problems that are disrupting and evolving the automotive insurance industry. A recent evolution in auto insurance rating and discount plans is State Farms Drive Safe and Save program, which is a form of usage-based insurance that analyzes driver behaviors based on mobile telematics data. Actuarial Statisticians work as leaders in the analytics community to navigate the challenges of implementing evolutionary and disruptive, as well as traditional, actuarial pricing efforts in a regulated environment.

Posters

Reduce computation in jackknife empirical likelihood for comparing two Gini indices

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

We propose an alternative approach of the jackknife empirical likelihood method to reduce the computation cost associated with the construction of confidence interval for the difference of two Gini indices from paired samples. We also investigate the adjusted jackknife empirical likelihood and the bootstrapped calibrated jackknife empirical likelihood to improve coverage accuracy for small samples. We, then, avoid the maximization over a nuisance parameter used by a previous profile jackknife empirical likelihood method proposed by Wang and Zhao (2016). We establish the Wilks theorem and show through simulations that the proposed methods perform better than Wang and Zhao (2016)'s methods in terms of coverage accuracy and require less computation. A real data application proves that the proposed methods work perfectly in practice.

High-order Laplacian-based regularization achieves the optimal rate in function estimation

Shanshan Cao
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
Collaborators/co-authors: Yibiao Lu, Zhouwang Yang, Xiaoming Huo

Graph Laplacian based regularization has been widely used in learning problems to take advantage of the information on the geometry towards the marginal distribution. In this paper, we consider the high-order Laplacian regularization, which takes the form of $\mathbf{f}^T\mathbf{L}^m\mathbf{f}$ with $\mathbf{L}$ being the graph Laplacian matrix of the sample data, in the context of supervised learning, and provide the theoretical foundations in the non-parametric setting. And we call the resulting estimator the *Graph Laplacian Smoother (GLS)*. The high-order Laplacian regularization technique, which is proved to converge to the Sobolev semi-norm regularization, has been successfully used in the literature of semi-supervised learning and supervised learning problems without theoretical guarantees. In this work, it is shown that nearly all good asymptotic properties of the existing state-of-the-art approaches are inherited by the Laplacian-based smoother. Specifically, we prove that as the sample size goes to infinity, the expected Mean Square Error (MSE) is of order $O(n^{-\frac{2m}{2m+d}})$, which is the *optimal convergence rate*, where m is the order of the Sobolev semi-norm used in the regularization, d is intrinsic dimension of the domain. Besides, we propose a *generalized cross validation* (GCV) approach for choosing the penalty parameter λ, which is guaranteed to be *asymptotically optimal*.

Regularized aggregation of statistical parametric maps

Jongik Chung
Department of Statistics
University of Georgia
Collaborators/co-authors: Li-Yu Wang, Cheolwoo Park, Hosik Choi, Amanda L. Rodrigue, Jordan E. Pierce, Brett A. Clementz, Jennifer E. McDowell

Combining statistical parametric maps (SPM) from individual subjects is the goal in some types of group-level analyses of functional magnetic resonance imaging data. Brain maps are usually combined using a simple average across subjects, making them susceptible to subjects with outlying values. Furthermore, t tests are prone to false positives and false negatives when outlying values are observed. We propose a regularized unsupervised aggregation method for SPMs to find an optimal weight for aggregation, which aids in detecting and mitigating the effect of outlying subjects. We also present a bootstrap-based weighted t test using the optimal weights to construct an activation map robust to outlying subjects. We validate the performance of the proposed aggregation method and test using simulated and real data examples. Results show that the regularized aggregation approach can effectively detect outlying subjects, lower their weights, and produce robust SPMs.

Functional directed graphical models

Ana María Estrada Gómez
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
Collaborators/co-authors: Kamran Paynabar, Massimo Pacella

A directed graphical model aims to represent the probabilistic relationship between variables in a system. Learning a directed graphical model from data includes parameter learning and structure learning. Several methods have been developed for graphical models with low-dimensional variables. However, the case in which the variables are high-dimensional has not been studied thoroughly. Nowadays, in many applications, the variables are high-dimensional, and need to be treated as functional random variables. This paper proposes a novel methodology to learn the structure and predict in the functional setting. When the structure of the network is known, function-to-function regression is used to estimate the parameters of the graph. When the goal is to learn the structure, a penalized least square loss with a group Lasso penalty, for variable selection, and an L_2 penalty, to handle group selection of nodes, is defined. Cyclic coordinate accelerated proximal gradient descent algorithm is employed to minimize the loss function and learn the structure of the directed graph. Through simulations and a case study, the advantage of the proposed method is proven. This paper enlarges the methodology to learn directed graphical models with high-dimensional variables.

Batch effect correction of single-cell RNA sequencing data through sample distance matrix correction

Teng Fei
Department of Biostatistics and Bioinformatics
Rollins School of Public Health, Emory University
Collaborators/co-authors: Tianwei Yu

Motivation: Batch effects is a frequent challenge in second generation sequencing data analysis. Several batch effect removal algorithms have shown good sample clustering results, providing important sample pattern information in corrected distance matrices. However downstream analyses, such as differential expression analysis, still suffer from remaining artifacts. Method: We have previously developed a method to correct the sample distance matrix, resulting in good sample clustering results. However the method doesn't correct the original count data matrix. In the current work, utilizing the corrected distance matrix as reference distance matrix, we numerically solve a novel least squares loss function to conduct linear transformation on the raw count matrix. The resulting corrected count matrix yields Pearson correlation that approximates the sample pattern reflected by the reference distance matrix. Downstream analyses can then be performed on the corrected count matrix. Results: We compared the proposed method with popular batch effect algorithms `ComBat` and `mnnCorrect`. In simulation studies, our method achieves better DE gene detection with high true positive rates and low false positive rates. In real data analysis, our method also shows improved performance in both clustering and DE gene analyses.

A personalized threshold method via boosting for sepsis screening

Chen Feng
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
Collaborators/co-authors: Yajun Mei, Paul Griffin, Shravan Kethireddy

Sepsis is one of the biggest risks to patient safety, with a natural mortality rate between 25% and 50%. It is difficult to diagnose and no validated standard for diagnosis currently exists. A commonly used scoring criteria is the quick sequential organ failure assessment (qSOFA). It has been shown to have low sensitivity in practice, however. We developed a method to personalize thresholds in qSOFA that incorporates easy to measure patient baseline characteristics. We compared the personalized threshold method to qSOFA, five previously published methods that obtain an optimal constant threshold for a single biomarker, and to the machine learning algorithms based on logistic regression and AdaBoosting using patient data in the MIMIC-III database. The personalized threshold method achieved higher accuracy than qSOFA and the five published methods and had comparable performance to machine learning methods. Personalized thresholds, however, are much easier to adopt in practice than machine learning methods as they are computed once for a patient and used in the same way as qSOFA, whereas the machine learning methods require an update after each observation.

Collaborative spectral clustering in attributed networks

Xiaodong Jiang
Department of Statistics
University of Georgia
Collaborators/co-authors: Pengsheng Ji

We propose a novel spectral clustering algorithm for attributed networks, where each node has p-dimensional meta-covariates from various formats such as text, image, speech, etc. The connectivity matrix $W_{n\times n}$ is constructed with the adjacency matrix $A_{n\times n}$ and covariate matrix $X_{n\times p}$, and $W = (1-\alpha)A + \alpha K(X, X')$, where $\alpha \in [0, 1]$ and K is a kernel to measure the covariate similarities. We then perform the classical k-means algorithm on the element-wise ratio matrix of the first K leading eigenvector of W. Theoretical and simulation studies show the consistent performance under both Stochastic Block Model (SBM) and Degree-Corrected Block Model (DCBM), especially in unbalanced networks where most community detection algorithms fail.

Optimal generalized quadrature functional regression

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

We study the generalized functional linear model with a scalar response with quadrature form functional predictors. The response given the functional predictors is assumed to follow a distribution of an exponential family. The unknown coefficient function is estimated by a penalized likelihood approach. An example of tensor product cubic spline is given to illustrate the feasibility of the proposed method. The penalized likelihood estimator attains the optimal convergence rate in the sense of prediction. Our simulations shows a better performance against the existing approach. The method is further illustrated in the use of the image classification and gene expression explained by histone modification.

Application of EL methods on bivariate MRL and quantile correlation estimators

Ali Jinnah
Department of Mathematics and Statistics
Georgia State University

Kulkarni and Rattihalli (2002) proposed an estimator for the bivariate mean residual life (MRL) function. In this paper, we apply the empirical likelihood (EL) and adjusted empirical likelihood (AEL) methods to the MRL function. The Wilk's theorem is established under general conditions. We profile the nuisance parameter in the EL and develop EL for the univariate MRL function. Extensive simulation studies show EL methods for both bivariate and one-dimensional MRL functions perform better than the normal approximation (NA) method in terms of coverage

probabilities. AEL methods result in noticeable better coverage probability. AEL method based on F-distribution calibration results in better coverage probability for small sample sizes. Two real data sets are used to illustrate the proposed procedure. We further apply jackknife empirical likelihood (JEL) method to quantile correlation and quantile partial correlation functions.

Propensity score matching for multi-level and spatially-structured data

Behzad Kianian
Department of Biostatistics and Bioinformatics
Rollins School of Public Health, Emory University
Collaborators/co-authors: Rachel Patzer, Lance Waller

In many health studies, an individual's treatment assignment and health outcomes are determined both by where they live and where they seek care. We consider propensity score matching in an observational data setting with a binary treatment and the presence of unmeasured cluster-level and spatial confounders. We present a systematic simulation study comparing various existing approaches, such as propensity score models that include spatial coordinates, random effects, and fixed effects, and the recently developed method of distance-adjusted propensity score matching (DAPSm). We additionally propose a method that synthesizes two methods recently developed in a two-stage procedure: (1) match within the same cluster where possible; (2) for the remaining treated units that were unmatched in the first stage, use the DAPSm method so that remaining matches are spatially close despite not being in the same cluster. We present initial results on a motivating application based on kidney disease patients who have recently started dialysis. Patients are either informed of their transplant options or not; this decision and the patient's outcomes are likely impacted by individual, facility, and area-level factors.

Distributional clustering: A distribution-preserving clustering method

Arvind Krishna
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
Collaborators/co-authors: Simon Mak, V. Roshan Joseph

One of the key uses of clustering is to compact big data into a set of representative points for further analysis. However, the distribution of the cluster centers of k-means or other clustering methods does not necessarily reflect the distribution of the data. We assume that a data analyst would in general want the distribution of the reduced sample to reflect the data distribution, except in cases where a task requires the data to be sampled in a specific way. For example, in the case of pattern recogniton problems, the data distribution should be reflected in the sample. We present a clustering method called 'Distributional Clustering', where the distribution of the obtained cluster centers asymptotically approaches the data distribution. We also demonstrate the performance of distributional clustering on a real data set and compare it with other sampling methods.

Dissecting differential signals in high-throughput data from complex tissues

Ziyi Li
Department of Biostatistics and Bioinformatics
Rollins School of Public Health, Emory University
Collaborators/co-authors: Zhijin Wu, Peng Jin and Hao Wu

Samples from clinical practices are often mixtures of different cell types. The high-throughput data obtained from these samples are thus mixed signals. The cell mixture brings complications to data analysis, and will lead to biased results if not properly accounted for. We develop a method to model the high-throughput data from mixed, heterogeneous samples, and to detect differential signals. Our method allows flexible statistical inference for detecting a variety of cell-type specific changes. Extensive simulation studies and analyses of two real datasets demonstrate the favorable performance of our proposed method compared with existing ones serving similar purpose.

Transformation and additivity in Gaussian process

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

Transformation of response is a common technique used in regression analysis. However, this method is not as prevalent when modeling deterministic functions. We proposed Transformed Additive Gaussian (TAG) process, finding a transformation to improve the additivity of the deterministic function. An additive function is easier to approximate and therefore the approximation obtained using such a transformation is expected to perform better. Furthermore, we extended TAG to a more general model called transformed approximately additive Gaussian (TAAG) process, for not just approximating but interpolating data from deterministic functions. Because TAG and TAAG both possess additive structures, efficient estimation techniques can be developed for their unknown parameters. Our proposed methods address several weaknesses of commonly used Gaussian process with product kernels, such as the occasional misidentification of important input variables and lower efficiency when fitting high dimensional and big data. These advantages are illustrated through numerical examples. We also applied TAG and TAAG to a noisy dataset.

Bayesian joint modeling of multiple brain functional networks

Joshua Lukemire
Department of Biostatistics and Bioinformatics
Rollins School of Public Health, Emory University
Collaborators/co-authors: Suprateek Kundu, Giuseppe Pagnoni, Ying Guo

Brain function is organized in coordinated modes of spatio-temporal activity (functional networks) exhibiting an intrinsic baseline structure with variations under different experimental conditions.

Existing approaches for uncovering such network structures typically do not explicitly model shared and differential patterns across networks, thus potentially reducing the detection power. We develop an integrative modeling approach for jointly modeling multiple brain networks across experimental conditions. The proposed Bayesian Joint Network Learning approach develops flexible priors on the edge probabilities involving a common intrinsic baseline structure and differential effects specific to individual networks. Conditional on these edge probabilities, connection strengths are modeled under a Bayesian spike and slab prior on the off-diagonal elements of the inverse covariance matrix. The model is fit under a posterior computation scheme based on Markov chain Monte Carlo. An application of the method to fMRI Stroop task data provides unique insights into brain network alterations between cognitive conditions.

Latent scale prediction model for network valued covariates

Xin Ma
Department of Biostatistics and Bioinformatics
Rollins School of Public Health, Emory University
Collaborators/co-authors: Suprateek Kundu

Network valued data commonly arise in various areas such as neuroimaging, genetics and social sciences. Although networks contain rich information, there have been limited advances in regression approaches involving network valued covariates in literature. One of the main challenges is the high dimensionality of the networks which often results in models with an inflated number of parameters, or alternative approaches which reduce the dimension of the networks and then use the low-dimensional structure to predict the outcome of interest. The first class of approaches are difficult to fit from a computational perspective and may result in inaccurate estimates, while the second class of methods often lack interpretability in terms of the connections in the observed network. In this work, we develop a novel regression model involving network valued covariates which uses a Bayesian framework to find a node-specific low-rank representation for the network covariates through the latent space model, then use a flexible regression framework for prediction. The approach results in a dramatic reduction in the number of regression parameters and is able to maintain interpretability at the node level. The posterior computation is realized through Markov Chain Monte Carlo with an efficient Gibbs Sampler involving a data augmentation scheme. We evaluate our performance in prediction and inference in simulations and real data example seeking to predict a clinical outcome based on structural connections in the human brain.

Selecting a representative subsample via SDART

Cheng Meng
Department of Statistics
University of Georgia
Collaborators/co-authors: Jingyi Zhang, Jinyang Chen, Wenxuan Zhong, Ping Ma

Given the large number of observations generated from an unknown continuous population density function p, we propose SDART, aiming to select subsamples which can approximately best-

represent p. SDART is derived by combining several algorithmic principles, namely measure-preserving transformation, space-filling design and nearest neighbor searching. SDART has an appealing visual representation of the population, requiring essentially $O(n)$ memory and $O(n\log(n))$ time. When applied to kernel density estimation, SDART achieves faster convergence rate than state of the art sampling methods. SDART is also suitable for the measure-constrained problem, as verified on several real-world datasets.

Generalized fairness regularizers for neural networks

Yonatan Mintz
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
Collaborators/co-authors: Matt Olfat

Many predictive tasks (e.g. image recognition and natural language processing) have been revolutionized by Artificial Neural Networks (ANN). However, as these deep learning methods become more ubiquitous, there have been several examples of these models exhibiting anthropomorphic bias (e.g. making predictions correlated with race or gender for unrelated tasks) by overfitting to, and thus amplifying, existing biases in training data. We address this problem by considering a novel regularization approach for deep learning, inspired by the constrained optimization literature, that directly penalizes unwanted disparities in treatment of populations proportionally to their impact on observed prediction bias. Using this method, we can control bias at training time as opposed to in a pre- or post-processing step; this results in concurrent out-of-sample improvements in fairness and can even improve model accuracy for some data sets. Our method fits well into existing training approaches and can be easily generalized across network architectures and notions of fairness. We validate our method empirically on several real world data sets that contain implicit bias. Namely we examine the prediction of recidivism, income, and wine quality while remaining fair with respect to race, gender, and wine color, respectively.

Combining satellite imagery and numerical model simulation to estimate ambient air pollution

Nancy L. Murray
Department of Biostatistics and Bioinformatics
Rollins School of Public Health, Emory University
Co-authors: Howard H. Chang, Heather A. Holmes, Yang Liu

Ambient fine particulate matter less than 2.5 micrometers in aerodynamic diameter ($PM_{2.5}$) has been linked to various adverse health outcomes. However, the sparsity of air quality monitors greatly restricts the spatial-temporal coverage of $PM_{2.5}$ measurements, potentially limiting the accuracy of $PM_{2.5}$-related health studies. We develop a method to combine estimates for $PM_{2.5}$ using satellite-retrieved aerosol optical depth (AOD) and simulations from the Community Multiscale Air Quality (CMAQ) modeling system. While most previous methods utilize AOD or CMAQ separately, we aim to leverage advantages offered by both methods in terms of resolution and coverage

by using Bayesian model averaging. In an application of estimating daily $PM_{2.5}$ in the Southeastern United States, the ensemble approach outperforms previously developed spatial-temporal statistical models that use either AOD or bias-corrected CMAQ simulations in cross-validation (CV) analyses. The enhanced prediction performance that the ensemble technique provides at fine-scale spatial resolution, as well as the availability of prediction uncertainty, can be further used in health effect analyses of air pollution exposure.

Partially collapsed Gibbs sampling for latent Dirichlet allocation

Hongju Park
Department of Statistics
University of Georgia
Collaborators/co-authors: Taeyoung Park, Yung-Seop Lee

A latent Dirichlet allocation (LDA) model is a Bayesian hierarchical model that identifies latent topics from text corpora. Current popular inferential methods to fit the LDA model are based on variational Bayesian inference, collapsed Gibbs sampling, or a combination of these. However, these methods can suffer from large bias, particularly when text corpora consist of various clusters with different topic distributions. This paper proposes an inferential LDA method to efficiently obtain unbiased estimates under flexible modeling for text corpora by using the method of partial collapse and the Dirichlet process mixtures. The method is illustrated using a simulation study and an application to a corpus of 1300 documents from neural information processing systems (NIPS) conference articles during the period of 2000–2002 and British Broadcasting Corporation (BBC) news articles during the period of 2004–2005.

Planning and analyzing clinical trials with competing risks: Recommendations for choosing appropriate statistical methodology

J. C. Poythress
Department of Statistics
University of Georgia
Collaborators/co-authors: Misun Yu Lee, Jim Young

In the analysis of time-to-event data, competing risks occur when multiple event types are possible, and the occurrence of a competing event precludes the occurrence of the event of interest. In this situation, statistical methods that ignore competing risks (e.g., by treating competing events as censoring) can result in biased inference regarding the event of interest. We review the mechanisms that lead to bias, and describe several statistical methods that have been proposed to avoid bias by formally accounting for competing risks in the analyses of the event of interest. Through simulation, we illustrate that Gray's test should be used in lieu of the logrank test for non-parametric hypothesis testing. We also compare the two most popular models for semi-parametric modelling: the cause-specific hazards (CSH) model and Fine-Gray (F-G) model. We explain how to interpret estimates obtained from each model, and identify conditions under which the estimates of the hazard ratio and subhazard ratio differ numerically. Finally, we evaluate several model

diagnostic methods with respect to their sensitivity to detect lack-of-fit when the CSH model holds, but the F-G model is misspecified, and vice versa. Our results illustrate that adequacy of model fit can strongly impact the validity of statistical inference. We recommend analysts incorporate a model diagnostic procedure and contingency to explore other appropriate models when developing the analysis plan for trials in which competing risks are anticipated.

On the probability distributions of durations of heatwaves

Sohini Raha
Department of Statistics
North Carolina State University
Collaborators/co-authors: Sujit Ghosh

Characterization of heatwaves is becoming increasingly important in environmental research as they pose a significant threat to many human lives worldwide. Though several quantifications of the extremities of a heatwave have been proposed in literature, they are mostly improvised and there does not exist a universally accepted definition of heatwave. In this paper, we devise a probabilistic inferential framework to characterize heatwave, and come up with a definition which can capture the essence of all existing ad hoc definitions. Based on results for sums of dependent Bernoulli random variables, we derive an approximate distribution on the frequency of such durations for a stationary time series. We select a daily time series e.g. maximum ambient temperature or heat-index (based on temperature and relative humidity) and define "Duration" as the amount of days the time series stays above a chosen threshold in one up-crossing in a fixed location. We then propose a hierarchical model for the durations and validate it using two different datasets, one with Atlanta data, and the other one with 126 USCRN weather stations spread across the United States. Using the distributions of the durations, we compute the expected duration of an up-crossing corresponding to a threshold in a fixed location, and define an up-crossing to be a heatwave if the duration of that exceeds the expectation. Moreover, we demonstrate a quadratic relationship between the threshold quantiles and the expected duration which makes it easier to identify the heatwaves at any given level of the quantiles of the time series that are generally used to define extreme heatwave.

Empirical likelihood inference for the two-way partial AUC

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

Receiver operative curve (ROC) is an important tool for observing the performance of binary classifiers. Area under the ROC curve (AUC) is calculated in this regard. It has been observed that putting restrictions on both true positive rates and false positive rates while calculating AUC provides much better classification ability. Such statistic is termed as two-way partial AUC. Empirical likelihood procedure, which does not require underlying distributional assumption is

much suitable in the estimation of such statistic. In this paper, a confidence interval for the two-way partial AUC based on empirical likelihood method is constructed. The performance of our proposed confidence interval is also compared with bootstrap confidence interval using simulation studies. The procedure is applied to a practical data set.

Predicting tree timber quality proportions by using a logit GLMM model

Héctor I. Restrepo
Warnell School of Forestry and Natural Resources
University of Georgia
Collaborators/co-authors: Nicole Lazar, Bronson P. Bullock

The ultimate goal of forest modeling, from the timber management perspective, is to obtain an accurate estimation of the merchantable volume as an input for financial return calculations. Finding the proportions of timber in each of the commercial pine product classes (i.e. sawtimber, chip-n-saw, and pulpwood) is an important part in estimating merchantable volume. Since timber products have different market prices, merchantable volume must be adjusted based on product classes. Stem quality may differ as a function of forest growth factors, i.e. age, environmental conditions, genetics, planting density and management regimes. The objective of this research is to predict the timber product class proportions over time as a function of forest growth factors and early field measurements of stem quality at age six. Data from a designed research trial evaluating the impacts of density and management over ages 6-21 (years) were used to estimate a logit GLMM model. The random component of the model is associated with the repeated measures of the units (plots). Proportions of timber product classes can be used to adjust the timber prices. Those proportions can be used to obtain a blended price of timber, resulting in a simplification of the financial calculations. Furthermore, timber product class proportions can be used to optimize financial returns by performing marginal analysis of applying certain management regimes.

Statistical inference for Rossby waves: A time series approach to persistent homology

Richard Ross
Department of Statistics
University of Georgia
Collaborators/co-authors: Lynne Seymour, Nicole Lazar, Thomas L Mote

Persistent Homology is a tool used within the framework of Topological Data Analysis that helps to summarize data based on its homological structure. In previous work, we've used this tool to generate time series which describe the relative strength of evidence for certain climatological patterns (a given number of Rossby waves). In this work, we implement inference for dependent time series data to provide concrete statistical inference. We present the results of this time series analysis, present a possible interface for climate researchers to use our work and results, and describe the next steps in developing our models and inference.

Comparison of groups based on cluster analysis of fMRI data

Arunava Samaddar
Department of Statistics
University of Georgia
Collaborators/co-authors: Cheolwoo Park, Nicole Lazar, Jennifer E. McDowell, Brooke Jackson

We aim to evaluate brain activation using functional Magnetic Resonance Imaging (fMRI) data and activation changes across time associated with practice related cognitive control during tasks. FMR images are acquired from participants engaged in 1 block design and 5 probability event related designs at two scan sessions: 1) pre-practice before any exposure to the task, and 2) post-practice, after 4 days of daily practice on either general antisaccade (generating a glance away from the cue) tasks or specific probability related event runs, which are a mixture of antisaccade and prosaccade (generating a glance towards the cue) task. The clustering technique is composed of several steps: detrending, data aggregation, wavelet transform and thresholding, the adaptive pivotal thresholding test, principal component analysis and K-medoids clustering. We use the structural similarity index to compare similarity between pre- and post- scan session images on the probability event related runs.

Empirical likelihood inference for the panel count data with informative observation process

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

Panel count data are interval-censored recurrent event data, which often arises from longitudinal studies. Each study subject is observed only at discrete time points rather than continuously. As a result, one can only know the total number of events occurred between two observation time points instead of actual time of the events. Furthermore, the observation times can be different among subjects and carry important information about the underlying recurrent process. In this paper, empirical likelihood (EL) method for panel count data with informative observation times is proposed. An empirical likelihood ratio for the vector of regression coefficients is formulated and the Wilks theorem is established. Simulation studies are carried out to show the performance of empirical likelihood and compare those with normal approximation methods. We compare the EL with existing methods using an illustrative example from bladder cancer study.

A Bayesian multidimensional trend filter

Stella C. W. Self
Department of Mathematical Sciences
Clemson University
Collaborators/co-authors: Christopher McMahan

The Bayesian multidimensional trend filter is an extension of the one-dimensional trend filtering technique popular in time series analysis. The method is suitable for data collected over 2 or more dimensions, such as spatial or spatio-temporal data. The multidimensional trend filter estimates a smooth trend function over the entire support space. The technique is computationally tractable, even for a large number of observations or a large support space. Furthermore, the computational expense of the method is determined by a user-specified level of discretezation and is nearly independent of the number of observations. An adaptive method of performing the discretezation of the support space is developed.

Dynamic regression with recurrent events

Jae Eui Soh
Department of Biostatistics and Bioinformatics
Rollins School of Public Health, Emory University
Collaborators/co-authors: Yijian Huang

Recurrent events often arise in follow-up studies where a subject may experience multiple occurrences of the same event. Examples include opportunistic infections in AIDS patients, successive pulmonary exacerbations in cystic fibrosis patients, myocardial infarctions, seizures in epileptic patients, and successive tumors in cancer studies. By the multiplicity nature of recurrent events, a dependence structure is often observed within a subject. Therefore, accommodating such intra-individual correlation is important in analyzing recurrent events. Most regression models with recurrent events tacitly assume constant effects of covariates. However, the effects may actually vary over time in many applications. In a clinical study for AIDS patients, for example, a treatment may take time to reach its full efficacy rather than right after randomization; meanwhile, the treatment effect may also erode over time as drug resistance develops, e.g., Eshleman et al (2001) and Wu et al (2005). Such circumstances call for more general models to accommodate the evolving effects of covariates. To address the time-varying effects, we propose a dynamic regression model to target the mean frequency of recurrent events at the population level. We develop an estimation procedure that fully exploits the observed data. Consistency and weak convergence of the proposed estimator are established. Simulation studies demonstrate that the proposed method works well, and two real data analyses illustrate its practicality.

Factor analysis on report citations using a combined latent and graphical model

Namjoon Suh
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
Collaborators/co-authors: Xiaoming Huo, Erice Heim, Timothy Van Slyke and Lee Seversky

We propose a combined latent and graphical model for the citation network, where either a latent model or a graphical model alone is often insufcient to capture the structure of the data. The proposed model has a latent (i.e., factor analysis) model to represents the main technological trends (aka factors), and adds a sparse graphical component that captures the remaining ad-hoc dependence. Model selection and parameter estimation are carried out simultaneously through construction of a pseudo-likelihood function and properly chosen penalty terms. The convexity of the pseudo-likelihood function allows us to develop an efcient algorithm, while the penalty terms generate a low-dimensional latent component and a sparse graphical structure.Simulation results are reported that show the new method works well in practical situations. The proposed method has been applied to a real application in HEP-Ph(high energy physics phenomenology) citation data set.

A novel hierarchical independent component modeling framework with application to longitudinal fMRI study

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

In recent years, longitudinal neuroimaging study has become increasingly popular in neuroscience research to investigate disease-related changes in brain functions, to study neurodevelopment or to evaluate treatment effects on neural processing. One of the important goals in longitudinal imaging analysis is to study temporal changes in brain functional networks and its association with subjects' clinical or demographical covariates. In neuroscience literature, one of the most commonly used tools to extract and characterize brain functional networks is independent component analysis (ICA), which separates multivariate signals into linear mixture of independent components. However, existing ICA methods are developed only for cross-section study and not suited for modelling repeatedly measured imaging data. In this paper, we proposed a novel longitudinal independent component model (L-ICA) as the first formal statistical modeling framework that extends ICA to longitudinal setting. By incorporating subject-specific random effects and visit-specific covariate effects, L-ICA is able to provide more accurate estimates of changes in brain functional networks on both the population- and individual-level, borrow information within the same subject to increase statistical power in detecting covariate effects on temporal changes in the networks, and allow for model-based prediction for changes in brain networks related to disease progression, treatment or neurodevelopment. We developed fully traceable exact EM algorithm to obtain maximum likelihood estimation of L-ICA. We further develop two approximate

EM algorithms based on voxel-specific subspace and sub-sampling techniques which greatly reduce the computation time while still retaining high accuracy. Moreover, we proposed a voxel-specific approximate inference procedure for examining covariate effects on brain network changes. Simulation results demonstrate the advantages of our proposed methods. We apply L-ICA to ADNI2 study to investigate deterioration in brain functional network in Alzheimer disease. Results from the L-ICA provides biologically insightful findings which are not revealed using the traditional methods.

First-order optimal sequential subspace change-point detection

Liyan Xie
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
Collaborators/co-authors: George V. Moustakides, Yao Xie

We consider the sequential change-point detection problem of detecting changes that are characterized by a subspace structure. Such changes are frequent in high-dimensional streaming data altering the form of the corresponding covariance matrix. In this work we present a Subspace-CUSUM procedure and demonstrate its first-order asymptotic optimality properties for the case where the subspace structure is unknown and needs to be simultaneously estimated. To achieve this goal we develop a suitable analytical methodology that includes a proper parameter optimization for the proposed detection scheme. Numerical simulations corroborate our theoretical findings.

Distance-based independence screening for canonical analysis

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

We introduce a new method named Distance-based Independence Screening for Canonical Analysis (DISCA) to reduce dimensions of two random vectors with arbitrary dimensions. DISCA is based on the distance-based independence measure, also known as the distance covariance, proposed by Székely and Rizzo in 2007. Unlike the existing canonical analysis methods, DISCA does not need the assumption that the dimension of the reduced subspaces of the two random vectors are equal. Besides, it can be applied to any types of distributions, continuous or discrete, light- or heavy-tailed. DISCA is eventually to solve a non-convex optimization problem. We reformulate it as a difference-of-convex (DC) optimization problem, and then it can be solved efficiently by adopting the alternating direction method of multipliers (ADMM) on the convex part of the DC problem. This formulation avoids the potentially numerically-intensive bootstrap method to determine the dimension of the reduced subspaces. In both the simulation studies and the real data cases, DISCA can not only handle the cases that other methods cannot do, but also perform comparably or better than other dimension reduction methods.

Decentralized computing methods for feature fusion

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

Due to the data transmission cost and data privacy, some multi-site data fusion can only be conducted through feature level. Under this situation, traditional statistical tools cannot be directly applied to the separated datasets. Decentralized computing methods have received significant attention recently. The methods tackle the multi-site problem by keeping the data in local nodes and exchanging only the estimator in each optimization step. In this paper, we consider the multi-site feature fusion problem when the data pooling is hard to process. We propose a decentralized computing method to fuse features. Theoretical results confirm the convergence and asymptotic property of the proposed method, even when there exist batch effects on different nodes. Simulation studies and real data examples indicate the proposed method dominates the existing traditional methods. We will study the communication efficiency in the future study.

Robustness and Tractability for High-Dimensional M-estimators

Ruizhi Zhang
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
Collaborators/co-authors: Yajun Mei, Jianjun Shi, Huan Xu

We investigate two important properties of M-estimator, namely robustness and tractability, in linear regression when the data are contaminated by arbitrary outliers. Specifically, robustness means the statistical property that the estimator should always be close to the true parameters *regardless of the distribution of the outliers*, whereas tractability means the computational property that the estimator can be computed efficiently even though the objective function of the M-estimator can be *non-convex*. In this article, by learning the landscape of the empirical risk, we show that under mild conditions, many M-estimators enjoy nice robustness and tractability properties simultaneously when the percentage of outliers is small. We extend our analysis to the high-dimensional setting in which the number of parameters is greater than the number of samples, and provide that when the proportion of outliers is small, many penalized M-estimators with L_1 penalty will enjoy robustness and tractability simultaneously. Our research provides a principled approach to choose tuning parameters for some families of M-estimators through balancing the tradeoff between robustness and tractability when the data are contaminated by arbitrary outliers. Simulation and case study results are presented to illustrate the usefulness of our theoretical results for M-estimators under Welsch's exponential loss.

A multi-armed bandit approach for online monitoring high-dimensional data in resource constrained environments

Wanrong Zhang
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
Collaborators/co-authors: Yajun Mei

We investigates the problem of online monitoring high-dimensional streaming data in resource constrained environments, where one has limited capacity in data acquisition, transmission or processing, and thus can only observe or utilize partial, not full, data for decision making. It is assumed that an undesired event might occur and change the distributions of some unknown components of data at some unknown time, and one wants to decide how to smartly observe a limited number of local components of data per time step so that one can still detect the undesired event as quickly as possible subject to the false alarm constraint. We propose a multi-armed bandit approach to adaptively sampling useful local components of data, and our method, termed Thompson-Sampling-Shiryaev-Roberts-Pollak (TSSRP) algorithm, is to combine the Thompson Sampling algorithm in the multi-armed bandits problem with the Shiryaev-Roberts-Pollak procedure in the sequential change-point detection literature. Our proposed TSSRP algorithm is able to balance between exploiting those observed local components that maximize the immediate detection performance and exploring new local components that might accumulate new information to improve future detection performance. In particular, our TSSRP algorithm performs similar to a random sampling algorithm in the in-control state when there are no changes, and performs similar to a greedy sampling algorithm on those affected local components in the out-of-control state after the event occurs. The usefulness of our TSSRP algorithm is validated through extensive numerical simulations and a real case study of solar flare detection.

Crime series detection from large-scale police report data

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

One of the most important problems in crime analysis is that of *crime series detection*. Technically speaking, *crime series* is a subset of *crime events* committed by the same individual or group. Generally, criminals follow a modus operandi (M.O.) that characterizes their crime series. The main scope of the project is to develop an efficient algorithm that can detect the correlation between crime incidences, using large-scale streaming police report data, both the structured (e.g., time, location) and unstructured (the so-called *free-text*).