Keynote Lecture

The Lasso: An Application to Cancer Detection and Some New Tools for Selective Inference

Robert Tibshirani

Health Research and Policy, Stanford School of Medicine and Department of Statistics
Stanford University

Collaborators/co-authors: Richard Lockhart (Simon Fraser University), Jonathan Taylor (Stanford University) and Ryan Tibshirani (Carnegie Mellon University)

First I will review the lasso method and show an example of its utility in cancer diagnosis via mass spectometry. Then I will consider the testing the significance of the terms in a fitted regression, fit via the lasso or forward stepwise regression.

I will present a novel statistical framework for this problem, one that provides p-values and confidence intervals that properly account for the inherent selection in the fitting procedure. I will give other examples of this procedure, including graphical models and PCA, and describe an R language package for its computation.

Technical Session

Symbolic Data Analysis: Are Distributions the Numbers of the Future? An Illustrative Answer

Lynne Billard
Department of Statistics
University of Georgia

Massively large data sets are routine and ubiquitous given modern computer capabilities. What is not so routine is how to analyse these data. One approach is to aggregate the data sets according to some scientific criteria. The resultant data are perforce symbolic data, i.e., lists, intervals, histograms, and so on. Applications abound, especially in the medical and social sciences. Other data sets (small or large in size) are naturally symbolic valued, such as species data, data with measurement uncertainties, confidential data, and the like.

Unlike classical data which are points in p-dimensional space, symbolic data are hypercubes or Cartesian products of distributions in p-dimensional space. We describe such data and how they arise. We look briefly at some of the differences between classical and symbolic data and their respective methodologies, through illustrations.

Simulation and Optimization Using Minimum Energy Designs

Roshan J. Vengazhiyil
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
Collaborators/co-authors: Tirthankar Dasgupta, Rui Tuo, C. F. Jeff Wu

Space-filling designs are commonly used as experimental designs in computer experiments. For example, maximin distance designs fill the experimental region uniformly while maximizing the minimum pairwise distance among the design points. In this talk, we will show that they can be modified to follow any arbitrary distribution by appropriately assigning weights to each design point. This method has a physical analogy of minimizing the total potential energy of electrically charged particles inside a box and therefore, we call the new space-filling design as minimum energy design. We will explain how this can be used in the simulation of complex probability distributions and in the global optimization of expensive black-box functions.

Evaluating Data Science Contributions in Teaching and Research

Lance A. Waller
Department of Biostatistics and Bioinformatics
Rollins School of Public Health, Emory University

The emergence, growth, and evolution of the field of data science include (or should include) meaningful and ongoing collaboration with the faculty members in the fields of statistics and biostatistics. This collaboration often yields new forms of scholarship and especially new formats and venues for teaching, research, and service that differ from traditional classroom teaching, peer-review publication, and committee assignments. From the perspective of a department chair, I outline opportunities and challenges in recognizing, valuing, documenting, and defending new scholarly contributions in a new field, particularly within the traditional academic promotion process for tenure-track faculty. Topics include MOOCs, blogging, software as scholarly output, and other emerging forms of scholarship within an evolving field in light of standard promotion evaluation systems.

Consulting Session

Training Statistics Students to Collaborate in the Academic Environment

Kim Love-Myers
Department of Statistics
University of Georgia

The University of Georgia Statistical Consulting Center (SCC) has a threefold mission: to provide collaborative research assistance to faculty, research staff, and students in all departments of the University of Georgia; to increase the quality of quantitative research performed at the University; and to provide an advantageous educational experience to students of statistics through training as statistical collaborators. This talk will focus on the third part of that mission, providing an educational experience to students of statistics. The SCC has provided assistantships for $4-9$ graduate consultants at a time each year for a number of years, in addition to providing volunteer opportunities for numerous graduate and undergraduate students with an interest in statistical consulting and collaboration. I will provide a brief history and overview of the SCC, and discuss the SCC's recent changes in training and development of these students, which has helped to place them and the SCC on a path to success.

Statistical Consulting - Experience From an GT-ISyE Faculty

Jye-Chyi (JC) Lu
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Many of us have consulting experience in our student and faculty life. Statistical consulting can be challenging but also rewarding. Statistician applies his/her problem-solving and good people skills to work with clients for solving their problems. This presentation will first summarize a few personal past projects ranging from quality improvement in manufacturing plants, product cost modeling for an electrical-mechanical-system assembly company, health-insurance fraud-detection, market-performance monitoring and so on. Then, we offer a few observations learned from dealing with clients in understanding problem background, signing contracts, jointly solving problems, providing reports and program codes, and delivering final presentations. The presentation will end with a few words about potential consulting opportunities in the (big-data) analytics field.

Industry Session

What it is like to work for LexisNexis

Hicham Elhassani
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."

What Is It Like to Work at SAS?

Ryan Lekivetz
JMP Division of SAS

SAS is the leader in business analytics software and services, and the largest independent vendor in the business intelligence market. Through innovative solutions, SAS helps customers at more than 70,000 sites improve performance and deliver value by making better decisions faster. Since 1976 SAS has been giving customers around the world "The Power to Know".

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

Wells Fargo Quantitative Associate Program for New PhDs

Misty Ritchie
Wells Fargo

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 or validate various analytical models. Dr. Misty Ritchie will discuss aspects and benefits of the rotational program and how PhD students can apply.

The 12 month rotational program consists of three distinct track selections:

Our **Capital Markets** track, sponsored by Corporate Risk and Wells Fargo Securities, 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 Risk** track, sponsored by Corporate Risk, Business Direct, Consumer Lending and Wealth, Brokerage and Retirement, gives Associates the opportunity to work with various lines of business to develop, maintain and validate statistical models for loss forecasting, credit risk scorecard, risk segmentation, capital management, and stress testing for a variety of lending products.

Our **Corporate Risk** track, sponsored by the Financial Crimes Risk Management, Operational Risk Management and Regulatory Risk Compliance Management areas of Corporate Risk, gives Associates the opportunity to develop, maintain and validate statistical models for the detection and prediction of suspicious activity, developing and maintaining statistical models used in estimating bank-wide operational risk regulatory capital levels and provide high-quality analytics to help our consumer lending businesses identify, quantify, and mitigate risks.

Advanced Analytics at State Farm

Jonathan Sauls
State Farm

From our founding father, G. J. Mecherle, who in 1922 observed that rural drivers deserved lower rates due to reduced risk, to State Farm becoming the first commercial SAS customer in 1972, to today's Advanced Analytics team of over 50 statisticians, analytics has played an important role in making State Farm the number one automobile, homeowner, and life insurance company in the US. The Advanced Analytics team uses a wide array of data science and predictive modeling methods to transform data into actionable understanding and competitive advantage, and we are invested in the development of tomorrow's analysts through our **M**odeling and **A**nalytics **G**raduate **N**etwork programs at the University of Illinois and the University of Georgia. We look forward to sharing more information with you about the exciting work that we are doing in analytics at State Farm.

Posters

Model calibration with censored data

Fang Cao
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
Collaborators/co-authors: Shan Ba, William Brenneman, V. Roshan Joseph

The purpose of model calibration is to make inference about the unknown input parameters, known as calibration parameters, of a computer model. The classical Kennedy-O'Hagan approach is widely used for model calibration, which can account for the inadequacy of the computer model while simultaneously estimating the calibration parameters. In many applications, the phenomenon of censoring occurs when the exact outcome of the physical experiment is not observed, but is only known to fall within certain region. In such cases, the Kennedy-O'Hagan approach cannot be used directly, and we propose a method to incorporate the censoring information when performing model calibration. The method is applied to study the compression phenomenon of liquid, and the results show significant improvements over the traditional methods.

Small-Area Estimation of Unmet Need for Preventive Dental Care for Children in Georgia

Shanshan Cao
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
Collaborators/co-authors: Monica Gentili, Nicoleta Serban, Paul Griffin, Susan Griiffin

The implementation of the Affordable Care Act (ACA) has led to a significant increase in the number of children receiving some form of dental benefit which is projected to result in further increase in utilization of dental care services for children. It is thus important to understand how well the existing supply of dental services will meet the resulting increase in need. In this study, we introduce a modeling framework to derive estimates of unmet need defined as lack of accessibility and availability of pediatric preventive dental care. We pilot our study for the population of children in Georgia. Our measurement models are based on optimization models that match need with supply of service under a series of user and provider system constraints. On the provider side, an example of constraint is the maximum caseload of providers with different levels of training, e.g., dentist and dental hygienist. On the user side, an example of constraint is the reduced acceptability of patients with public insurance. The results of the optimization model are used to measure accessibility, availability and unmet need of preventive dental care for different population groups (overall population of children, publicly-insured children, and privately-insured children) at the census tract levels and for different geographic areas (rural vs.urban areas). We compare our estimates of unmet need to those derived using the Human Resources and Service Administration (HRSA) shortage criteria. The results of the analysis are useful to provide guidance as to which policies can best address this unmet need.

Poisson Matrix Completion

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

We extend the theory of matrix completion to the case where we make Poisson observations for a subset of entries of a low-rank matrix. We consider the (now) usual matrix recovery formulation through maximum likelihood with proper constraints on the matrix M of size d_1-by-d_2, and establish theoretical upper and lower bounds on the recovery error. Our bounds are nearly optimal up to a factor on the order of $\mathcal{O}(\log(d_1 d_2))$. These bounds are obtained by adapting the arguments used for one-bit matrix completion (although these two problems are different in nature) and the adaptation requires new techniques exploiting properties of the Poisson likelihood function and tackling the difficulties posed by the locally sub-Gaussian characteristic of the Poisson distribution. Our results highlight a few important distinctions of Poisson matrix completion compared to the prior work in matrix completion including having to impose a minimum signal-to-noise requirement on each observed entry. We also develop an efficient iterative algorithm and demonstrate its good performance in recovering solar flare images.

Statistical Approaches for Exploring Brain Connectivity with Multi-Modal Neuroimaging Data

Phebe B. Kemmer
Department of Biostatistics and Bioinformatics
Rollins School of Public Health, Emory University
Collaborators/co-authors: Ying Guo, DuBois Bowman

By combining various types of neuroimaging data, multimodal imaging analyses enable us to study the relationship between brain structure and function, and investigate the connectivity disruption pathways that characterize certain brain diseases. We develop a novel measure to quantify the strength of structural connectivity (sSC) underlying functional networks identified from fMRI using data-driven methods such as independent component analysis (ICA). The sSC statistic can be defined on both the voxel- or region-level using tractography procedures from diffusion tensor imaging (DTI) data. We provide a framework to conduct statistical inference for sSC, which overcomes many computational challenges due to spatial correlations within the data and the estimation of a large variance-covariance matrix. We present simulation results, to assess the performance of our measure, and illustrate the application of this multimodal analysis using an fMRI and DTI dataset of 20 healthy controls and 20 patients with major depressive disorder. We find that the reliability of functional networks estimated by ICA is informed by sSC.

M- Statistic For Kernel Change-Point Detection

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

Detecting the emergence of an abrupt change-point is a classic problem in statistics and machine learning. Kernel-based nonparametric statistics have been proposed for this task which make fewer assumptions on the distributions than traditional parametric approach. However, none of the existing kernel statistics has provided a computationally efficient way to characterize the extremal behavior of the statistic. Such characterization is crucial for setting the detection threshold, to control the significance level in the offline case as well as the average run length in the online case. In this paper we propose two related computationally efficient M-statistics for kernel-based change-point detection when the amount of background data is large. A novel theoretical result of the paper is the characterization of the tail probability of these statistics using a new technique based on change-of-measure. Such characterization provides us accurate detection thresholds for both offline and online cases in computationally efficient manner, without the need to resort to the more expensive simulations such as bootstrapping. We show that our methods perform well in both synthetic and real world data.

K-regression Clustering for Interval-Valued Data

Fei Liu
Department of Statistics
University of Georgia
Collaborators/co-authors: Lynne Billard

Symbolic data records are becoming a more powerful instrument to deal with large size data sets. Interval-valued data are a special type of symbolic data, for which each observation is a vector of intervals. The typical K-means methods for interval-valued data suppose the data separate to spherical clusters. It usually cannot converge to the correct clusters if the data are not clustering spherically. We propose a K-regression based clustering method for interval-valued data to recover a more complicated data structure. Assuming the response and predictor variables follow K different linear relationships, the data are initially split into K groups randomly. Then, we apply the new developed "symbolic variation" least squares to estimate the parameters of the K symbolic regressions. A data point is then relocated to its closest group in terms of its symbolic distance to the regression lines. This two-step dynamic clustering algorithm continues until the clusters are stable. Further, a plot of symbolic regression R-square versus the number of cluster K is used to determine the optimal number of clusters. The simulation shows that our method performs better than the K-means method.

Scalable SUM-Shrinkage Schemes for Distributed Monitoring Large-Scale Data Streams

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

Motivated by biosurveillance and censoring sensor networks, we investigate the problem of distributed monitoring large-scale data steams where an undesired event may occur at some unknown time and affect only a few unknown data streams. We propose to develop scalable global monitoring schemes by parallel running local detection procedures and by combining these local procedures together to make a global decision based on SUM-shrinkage techniques. The SUM-shrinkage technique is to apply shrinkage methods to combine the local detection statistics of the local detection procedures together. By using the SUM-shrinkage technique, we are able to filter out those unchanging local data streams and to make a global decision based on those likely affected data streams.

Our approach is illustrated in two concrete examples: one is the nonhomogeneous case when the pre-change and post-change local distributions are given, and the other is the homogeneous case of monitoring a large number of independent $N(0, 1)$ data streams where the means of some data streams might shift to unknown positive or negative values. Numerical simulation studies demonstrate that the usefulness of the proposed schemes.

Weighted Leverage Score for High Dimensional Variable Screening

Yiwen Liu
Department of Statistics
University of Georgia
Collaborators/co-authors: Wenxuan Zhong, Peng Zeng

With the rapid development of science and technology, a large amount of high dimensional data has occurred in areas such as genomics, finance, image processing and Internet search. How to extract useful information from massive data becomes the key issue nowadays. In spite of the urgent need in statistical tools to deal with such data, there are limited methods that can fully address the high dimensional problem.

Motivated by sliced inverse regression, we here propose a novel feature screening method named weighted leverage score (WLS). WLS screening procedure does not impose a specific form of relationship between the response variable and the predictors, and it can identify all relevant predictors consistently. As showed in our theoretical analysis, WLS not only possesses consistency in selection, but also has competitive performance in empirical studies and real data application.

Using particle swarm optimization to identify optimal designs for generalized linear models with binary response

Joshua Lukemire
Department of Biostatistics and Bioinformatics
Rollins School of Public Health, Emory University
Collaborators/co-authors: Abhyuday Mandal, Weng Kee Wong

Identifying optimal designs for experiments in which the response is modeled by a generalized linear model is a difficult task due to the dependence of the information matrix on the model parameters. Theoretical results for such design problems are often unavailable, resulting in the use of computational tools to construct designs. This research explores the use of one such computational tool, Particle Swarm Optimization (PSO), to identify D-optimal designs for generalized linear models taking a binary response. We provide comparisons of PSO to other popular design algorithms for all discrete and all continuous factor experiments. Finally, we provide results for experiments with a mixture of discrete and continuous factors.

A Bayesian Approach for Envelope Model

Subhadip Pal
Department of Biostatistics and Bioinformatics
Rollins School of Public Health, Emory University
Collaborators/co-authors: Zhihua Su, Kshitij Khare

The envelope model is a new paradigm to address estimation and prediction in multivariate analysis. Using sufficient dimension reduction techniques, it has the potential to achieve substantial efficiency gains compared to standard models. This model was first introduced in Cook (2010) for multivariate linear regression, and has since been adapted to many other contexts. However, a Bayesian approach for analyzing envelope models has not yet been investigated in the literature. In this paper, we develop a comprehensive Bayesian framework for estimation and model selection in envelope models in the context of multivariate linear regression. Our framework has the following attractive features. Firstly, we use the matrix Bingham distribution to construct a prior on the orthogonal basis matrix of the envelope subspace. This prior respects the manifold structure of the envelope model, and can directly incorporate prior information about the envelope subspace through the specification of hyperparamaters. This feature has potential applications in the broader Bayesian sufficient dimension reduction area. Secondly, sampling from the resulting posterior distribution can be achieved by using a block Gibbs sampler with standard associated conditionals. This in turn facilitates computationally efficient estimation and model selection. Thirdly, unlike the current frequentist approach, our approach can accommodate situations where the sample size is smaller than the number of responses. Lastly, the Bayesian approach inherently offers comprehensive uncertainty characterization through the posterior distribution. We illustrate the utility of our approach on simulated and real datasets.

Tracking Concept Drift Using a Constrained Penalized Regression Combiner

Li-Yu Wang
Department of Statistics
University of Georgia
Collaborators/co-authors: Cheolwoo Park, Kyupil Yeon, Hosik Choi

The objective of this work is to develop a predictive model when data batches are collected in a sequential manner. With streaming data, information is constantly being updated and a major statistical challenge for these types of data is that the underlying distribution and the true input-output dependency might change over time, a phenomenon known as concept drift. The concept drift phenomenon makes the learning process complicated because a predictive model constructed on the past data is no longer consistent with new examples. In order to effectively track concept drift, we propose new novel model-combining methods using constrained and penalized regression that possesses a grouping property. The new learning methods enable us to select data batches as a group that are relevant to the current one, reduce the effects of irrelevant batches, and adaptively reflect the degree of concept drift emerging in data streams. We study theoretical properties of the proposed methods and finite sample performance using simulated and real examples. The analytical and empirical results indicate that the proposed methods can effectively adapt to various types of concept drift and show superior performance over existing methods.

A Metagenomics Method for Simultaneously Identifying Microbial Species and Estimating their Abundance in Multiple Samples

Xin Xing
Department of Statistics
University of Georgia
Collaborators/co-authors: Jun S. Liu, Wenxuan Zhong

Metagenomics refers to the study of a collection of genomes, typically microbial genomes, presenting in environmental samples, such as samples from the gastrointestinal tract of a human patient or samples of soil from a particular ecological origin. By sequencing bulk DNA that is directly extracted from environmental samples, one can bypass the difficulties arising in cell cultivation. Moreover, we can easily identify novel microbial species and study their distribution variation along different samples. However, these advantages cannot really benefit the biological researchers before we have a high-resolution and reference-free Metagenomic tool, because most existing methods that focus on basic taxonomy ranks either need to align short reads to a reference genome or can only produce very rough estimates. It is very challenging to identify species that are not well studied. The method we introduce in this article can overcome this difficulty and provide a reliable detection of microbial species. Our method leverages the matrix factorization method to simultaneously estimating known and unknown species and their proportions in a microbial colony. We demonstrate our method in both simulation and a real biological study.

Quantifying and Understanding Adherence to Recommended Care Practices for Pediatric Asthma Care

Yuchen Richard Zheng
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
Collaborators/co-authors: Ross Hilton, Nicoleta Serban, Anne Fitzpatrick, James Bost

The disadvantaged populations, specifically Medicaid-insured children, have historically utilized the care system disparately, without following evidenced-based care practices. In this study, we adopt sequential clustering to quantify and understand adherence to basic recommended care for pediatric asthma in the Medicaid system. The model output is used to visualize transitions between providers, quantify cost-savings of interventions for improving adherence and evaluate the impact of geographic access on patient-level utilization using multinomial regression. We find more similarities than dissimilarities in pediatric asthma healthcare utilization for GA and NC Medicaid systems. It is most efficient for policy-makers to focus on interventions raising the levels of adherence for patients visiting ED regularly. The potential costs associated with such interventions can be offset by the cost-savings in the Medicaid payments primarily for Georgia. One contributor to prevalent ED utilization is lower geographic access to asthma care. Targeting communities and not individuals for interventions will be most effective in improving adherence.