Program & Abstracts

ISBIS 2017
June 7-9

IBM T. J. Watson Research Center in Yorktown Heights, NY

Sponsored by

IBM Research

THE CLIMATE CORPORATION

facebook

NESS | New England Statistical Society

New England Statistics Symposium since 1987 & New England Statistical Society since 2017

Google

Novel Hierarchical Bayesian Approaches for Business and Government Applications, Auditorium

New Statistical Methods for Stochastic Volatility and Functional Data in Finance and Engineering , Room 20-043

Analytics Promoting Social Good: Money Access, Housing and Health, Room CR3

Full program can be found online at www.isbis2017.org/program.

Program & Abstracts

Wednesday June 7th, 2017

8:45-9:00

Welcome Address, Auditorium

Dario Gil, Vice President, Science and Solutions, IBM Research

9:00-10:00

Keynote Address, Auditorium

"Riding Technology Waves: Perspectives and Opportunities for Leveraging Data" - Brenda L. Dietrich, IBM Fellow and VP

Abstract: This talk begins with a fly-by of five decades of information technology beginning with its use to automate business processes and extending to its current role in consumer self-service and in intermediating social processes. The resulting "data exhaust" together with the availability of low cost computing capacity spawned the age of analytics, the rise of big data and the birth of cognitive computing. The past, current and potential role of analytic methods in these technology waves will be discussed.

10:30-12:00

ASMBI Invited Session, Room CR3

Org/Chair: Emmanuel Yashchin, IBM Research

"Graphics to Facilitate Informative Discussion and Team Decision-Making" - Christine Anderson-Cook, Los Alamos National Laboratory

Abstract: Everyone knows the expression "A picture is worth a thousand words." However, in many cases, the goal for the right graphical summary is not only to convey information for solving the right problem, but also to encourage and guide discussion and to help focus a team on making a carefully considered, defensible, data-driven decision. The aims of graphics differ if we are trying to communicate the merits of a single choice versus outline several contending alternatives for further comparison and discussion, each with their own strengths and weaknesses. They also serve different purposes at various stages of decision-making. Often the role of statisticians is not to provide a single answer, but to provide rich information and summaries in a manageable and compact form to enable productive discussion among their team. Through a series of diverse examples, we present principles and strategies for examining multiple objectives, framing trade-offs between alternatives, and examining the impact of subjective priorities and uncertainty on the final decision.

Discussants:

Jennifer van Mullecom, Virginia Tech
Tom Donnelly, SAS

Statistical Models for Complex Data, Room 20-043

Org/Chair: Veronika Rockova, University of Chicago

”An Interaction Analysis of Social Media and Traditional Platform Effects in the Consumer Purchasing Funnel” - Daniel Zantedeschi, The Ohio State University

Abstract: We advance a pragmatic empirical strategy aiming at measuring synergistic effects of online platforms for targeted advertising in a relationship between a firm and a multichannel advertising agency. Gauging meaningful interaction effects between activities on different platforms and within different parts of the purchasing funnel is very challenging. This is due to (a) the presence potential endogeneity biases where the most active users end up being targeted more frequently on different platforms and (b) “rare outcomes” indicating that ultimate conversion rates are negligible. We tackle these issues by a combination of cutting-edge, but yet established, tools in the epidemiology and machine learning literature comprising (a) case-control design to match retrospectively users showing a similar level of activity and (b) post-regularized choice models, proved to be effective even in the presence of rare outcomes. Our empirical analysis finds that segmenting customers based on the similarity of their browsing activities mitigates “path-to-purchase” heterogeneity and offers more accurate associational measures related to platform effects. Second, targeting across platforms is positively associated with ultimate conversion for the lower funnel, but there is no measurable synergistic effect for the upper funnel. Also, we find that main effect of social media is positively related to the ultimate conversions for users in the early stages but has no incremental impact when consumers move down to the lower funnel.

”Bayesian Causal Forests” - Richard Hahn, University of Chicago

Abstract: In this talk I will describe a semi-parametric Bayesian regression model for estimating heterogeneous treatment effects from observational data. Standard nonlinear regression models, which may work quite well for prediction, can yield badly biased estimates of treatment effects when fitted to data with strong confounding. The new Bayesian causal forest model is able to eliminate this adverse bias by jointly modeling the treatment and the response conditional on control variables.

”Sparse Autoregressive Processes for Dynamic Variable Selection” - Veronika Rockova, University of Chicago

Abstract: We consider the problem of dynamic variable selection in time series regression models, where the set of active predictors is allowed to evolve over time. To capture time-varying variable selection uncertainty, we introduce new dynamic shrinkage priors for the time series of regression coefficients. These priors are characterized by two main ingredients: smooth parameter evolutions as well as intermittent zeroes for modeling predictive breaks. This is achieved with a multiple shrinkage formulation by switching between two shrinkage targets: either zero or the vicinity of the previous value. More formally, our proposed Autoregressive Spike-and-Slab Process (ASSP) priors are constructed as mixtures of two processes: a spike process for the irrelevant coefficients and a slab autoregressive process for the active coefficients. The mixing weights are themselves time-varying and depend on a lagged value of the series. A remarkable feature of ASSP is that their stationary distribution is fully known, where the marginals are characterized by benchmark spike-and-slab priors. This property guarantees marginal stability and equilibrium between selection and smoothing. The practical appeal is the availability fast MAP estimation algorithms, a strategy

pursued here. By turning our priors into penalty functions, we formalize the notion of dynamic shrinkage, borrowing adaptively from both future and past information. We characterize dynamic selection thresholds for MAP smoothing and implement a one-step-late EM algorithm for efficient calculations. Illustrations show that ASSP succeed far better at finding signal relative to classical autoregressive state-space models, when many redundant predictors are present.

Recent Advances in Spatio-Temporal Statistical Modeling, Room 20-001

Org/Chair: Huijing Jiang, IBM Research

"Dynamic Multiscale Spatiotemporal Models for Poisson Data" - Marco Ferreira, Virginia Tech

Abstract: We propose a new class of dynamic multiscale models for Poisson spatiotemporal processes. Specifically, we use a multiscale spatial Poisson factorization to decompose the Poisson process at each time point into spatiotemporal multiscale coefficients. We then connect these spatiotemporal multiscale coefficients through time with a novel Dirichlet evolution. Further, we propose a simulation-based full Bayesian posterior analysis. In particular, we develop filtering equations for updating of information forward in time and smoothing equations for integration of information backward in time, and use these equations to develop a forward filter backward sampler for the spatiotemporal multiscale coefficients. Because the multiscale coefficients are conditionally independent a posteriori, our full Bayesian posterior analysis is scalable, computationally efficient, and highly parallelizable. Moreover, the Dirichlet evolution of each spatiotemporal multiscale coefficient is parametrized by a discount factor that encodes the relevance of the temporal evolution of the spatiotemporal multiscale coefficient. Therefore, the analysis of discount factors provides a powerful way to identify regions with distinctive spatiotemporal dynamics. Finally, we illustrate the usefulness of our multiscale spatiotemporal Poisson methodology with two applications. The first application examines mortality ratios in the state of Missouri, and the second application considers tornado reports in the American Midwest.

"Spatially Fused Time-Varying Lattice Models for Agricultural Management Zoning" - Rodrigue Ngueyep Tzoumpe, IBM Research

Abstract: In many applications where both predictors and responses are collected across geographical regions over time, the impact of the predictors to responses are often not static but time-varying. Moreover, the time-varying impact of the predictors may vary across different regions. To identify nearby regions where these time-varying impact behave similarly, we propose a spatially fused time-varying lattice model. We model time-varying impact of spatio-temporal predictors via a spatial lattice model with time-varying coefficients. Furthermore, we utilize fusion penalty to allow nearby regions to share same time-varying coefficients. The model parameters can be efficiently estimated via ADMM algorithm. One motivation application of our method is to identify agriculture management zones where the time-varying impact of environment attributes (e.g., growing degree days, heat stress, precipitation) on the crop yield is similar. Once these zones are identified, same planting policy could be implemented within these zones.

“Spatio-Temporal Data Science at The Climate Corporation” - Maria Terres, The Climate Corporation

Abstract: Recent advances in precision agriculture have enabled farmers to better optimize management decisions on their operations through the use of data science techniques which are often frequently associated with tech companies like Apple, Google, and Facebook. The Climate Corporation is at the forefront of the development of these digital tools. Our data scientists have the opportunity to work on a variety of statistical challenges, all of which are inherently spatio-temporal due to the non-homogenous nature of farmers’ fields and the impacts that in-season weather events can have on crops. The resulting algorithms are implemented in Climate FieldView™ Pro where farmers can visualize spatial maps of their fields’ health in-season, create spatial prescriptions for seeding rates, monitor the spatio-temporal dynamics of the nitrogen in their soil, and more.

Statistical Applications in The Financial Services and Media Industries, Auditorium

Org/Chair: Aliza Heching, IBM Research

”Quantitative Strategies Using Sentiment Classification of Financial News Using Statistical Techniques in Machine Learning” - Arun Verma, Bloomberg

Abstract: The high volume and time sensitivity/dependence of news and social media stories necessitates automated processing to extract actionable information, while the unstructured nature of textual information presents challenges that are comfortably addressed by machine-learning techniques. We have applied a novel machine learning technique combining 3 separate support vector machines . In this talk we examine these scores, focusing on using news and social sentiment information in trading strategies that can achieve good risk-adjusted returns.

”Multi-Factor Models: Risk and Attribution” - Samer Takriti, Viking Global Investors LP

Abstract: For decades, quantitative portfolio construction has relied on fundamental multi-factor risk models to predict portfolio volatility and attribute sources of risk and return. These models are built by regressing market returns on a set of "intuitive" fundamental factors. This talk describes the typical construction methodology of risk models and discusses some of the cautions which must be exercised in their use and interpretation.

“Applications of Machine and Deep Learning in Media” - Rahel Jhirad, Hearst

Abstract: TBA

2:00-3:30

Data-Driven Learning in Industrial Settings, Room 20-043

Org/Chair: Yada Zhu, IBM Research

"Towards Cognitive Product Data Cleaning in e-Commerce" - Brian Quanz, IBM Research

Abstract: As part of managing, planning for, and carrying out e-commerce operations like order fulfillment, retailers and many of their e-commerce engines depend on product item weights and dimensions. Missing or incorrect item data can lead to poor, costly decisions, for example, leading to inaccurate shipping cost estimates resulting in more costly fulfillment decisions, or showing incorrect information to customers. However, retailers often struggle to get accurate and complete item weight and dimensions, with many items and high item turn-over, and as a result many item characteristic values are missing or inaccurate. In this talk I will present results showing the impact of missing item weight and dimension data on fulfillment cost. I will describe a comprehensive cognitive, data-driven approach to address the missing and inaccurate data in an adaptive manner, as well as present experimental results showing the cost improvement of an initial data-driven approach which has been implemented and is currently being used in a production e-commerce order fulfillment system.

"Moderate Deviations in Cloud Computing via Stein-Chen Method" - Yingdong Lu, IBM Research

Abstract: Estimating the fulfillment probability (or equivalently loss probability) is a crucial problem in capacity planning for cloud computing. We developed a novel method for this estimation through new development on the Stein-Chen method for moderate deviations of linear combinations of Poisson distributions. While this method has been shown to be numerically efficient even for large scale networks, this new study provide theoretical error bounds. This is a joint work with Yue Tan and Cathy Xia from Ohio State University.

"An Efficient Graph Algorithm for Customer Return Prediction in e-Commerce Industry" - Yada Zhu, IBM Research

Abstract: With the booming of e-commerce in the last decade, return cost becomes a key challenge faced by e-retailers. To improve customer engagement and experience, more and more retailers offer hassle-free returns. As online sales are growing, e-commerce return costs are also going up. However, as far as we know, there is no systematic framework to analyze customer return behaviors for return cost reduction. To fill in this gap, we develop a return affinity model that influences customer search results based on their historical purchase and return behaviors. The proposed algorithm is evaluated on real world online retail data and demonstrates potential for large scale implementation.

Statistics in the IT Industries, Auditorium

Org/Chair: David Banks, Duke University

"Towards Scalable Automatic Machine Learning" - Erin LeDell, H2O

Abstract: In recent years, the demand for machine learning experts has outpaced the supply, despite the surge in people entering the field. To address this gap, there have been big strides in the development of user-friendly machine learning software that can be used by non-experts. Open source machine learning tools such as scikit-learn and H2O have lowered the barrier to entry to the field by offering user-friendly interfaces to advanced machine learning algorithms. H2O in particular, has introduced a fully-featured web-based GUI that allows users to train and evaluate machine learning models on large datasets, all without writing a single line of code.

Although these tools have made it easier for non-experts to experiment with machine learning, there is still a fair bit of knowledge and background in data science that is required to produce high-performing, production-ready machine learning models. Deep Neural Networks in particular are notoriously difficult for a non-expert to tune properly. In order for AI software to truly be accessible to non-experts, such systems must be able to automatically perform proper data pre-processing steps and return a good model.

H2O.ai, creator of H2O, an open source platform for scalable, distributed machine learning, has recently developed a distributed Automatic Machine Learning system called H2O AutoML. We will present our methodology for automating the machine learning workflow, which includes feature pre-processing and automatic training of many models within a user-specified time-limit. Stacked ensembles are also automatically trained using a subset of the individual models to produce a highly predictive final model.

"Random Forests, Decision Trees, and Categorical Predictors: The 'Absent Levels' Problem" - Timothy Au, Google

Abstract: The "absent levels" problem for random forests and other decision tree based models occurs when a tree node splits on a categorical predictor where some of the variable's unordered categorical levels are not present in the node's subset of the training data, but still present in out-of-bag or test observations that reach the node. Although an inherent issue for decision based tree models, the absent levels problem has been overlooked in both the literature and in popular software implementations of these models. By using Leo Breiman's random forests FORTRAN code and the randomForest R package as motivating examples, we show how absent levels can dramatically and systematically bias a model.

"Scalable Bayesian Modeling and Monitoring of Dynamic Network Flow Data with Online Advertising Applications" - Xi Chen, LinkedIn

Abstract: Traffic flow count data in networks arise in many applications, such as automobile or aviation transportation, certain directed social network contexts, and Internet studies. Using an example of Internet browser traffic flow through site-segments of an international news website aiming to optimize online advertising strategies, we present Bayesian analyses of two linked classes of models which, in tandem, allow fast, scalable and interpretable Bayesian inference. We first develop flexible state-space models for streaming count data, able to adaptively characterize and quantify network dynamics efficiently in real-time. We then use these models as emulators of more

structured, time-varying gravity models that allow formal dissection of network dynamics. This yields interpretable inferences on traffic flow characteristics, and on dynamics in interactions among network nodes. Bayesian monitoring theory defines a strategy for sequential model assessment and adaptation in cases when network flow data deviates from model-based predictions. Exploratory and sequential monitoring analyses of evolving traffic on a network of web site-segments in e-commerce demonstrate the utility of this coupled Bayesian emulation approach to analysis of streaming network count data.

Data Science for Decision Support, Room CR3

Org/Chair: Kay See Tan, Memorial Sloan Kettering Cancer Center

”Using analytics to measure and increase energy savings” - Nancy Hersh, Independent

Abstract: Climate change is not simply a supply-side issue but a demand-side challenge as well. Learn how Opower uses behavioral science and data science to reduce the energy consumption of over 5 million individuals worldwide. Understand the techniques and approaches employed to precisely measure the impact of those behavior changes.

“Automated External Feature Sensitivity Scoring Tool” - Karina Kervin, IBM Research

Abstract: The goal in this work was to quickly identify which time series from a matrix of N time series are most likely to be impacted by external factors while maintaining interpretability of results. The recent increase in use of data collection technologies, such as IoT devices, means important time series forecasts can occur at increasingly finer levels of analysis. While this increase in data can lead to more detailed insights, it can also dramatically increase the time to find those insights. Many statistical methods, such as neural networks, sacrifice interpretability for accuracy in their ability to handle large, complex data sets. A hybrid approach is proposed to provide this interpretability, combining the output from baseline SARIMAX models with external data, and feeding this into Lasso models. The Lasso models generate a score that describes the improvement of modeling accuracy when adding these external features to the modeled time series. A secondary output is the subset of external features that impacted each time series model. This tool was tested on retail sales data for selected beverages, successfully identifying impactful weather patterns on sales data that would not have been considered in a purely manual analysis.

"Interfacing With Decision Makers Across Warby Parker" - Maxwell Shron, Warby Parker

Abstract: Warby Parker produces and sells affordable eyewear using its own e-commerce site, physical retail locations, marketing channels and supply chain. Because so many of these functions are fulfilled in-house there are many opportunities to apply statistical methods to improve operating efficiency, from selecting sites for new stores to optimizing marketing spend to designing a network of high-throughput optical labs. This talk will review some of these applications at a high level, discussing elements of the statistical methodology but largely emphasizing the techniques used by the data science team to ensure that the output of our models closely match the needs of decision makers.

4:00-5:30

Impact Measurement Methodologies, Auditorium

Org/Chair: Marianna Dizik, Google Inc.

”Bayesian methods in media mix modeling” - Yuxue Jin, Google Inc.

Abstract: Media mix models are used by advertisers to measure the effectiveness of their advertising and provide insight in making future budget allocation decisions. Advertising usually has lag effects and diminishing returns, which are hard to capture using linear regression. We propose a media mix model with flexible functional forms to model the carryover and shape effects of advertising. The model is estimated using a Bayesian approach in order to make use of prior knowledge accumulated in previous or related media mix models. We illustrate how to calculate attribution metrics such as Return on Ad Spend (ROAS), marginal Return on Ad Spend (mROAS) and optimal media mix on simulated data sets, as well as on data from a shampoo advertiser.

”Longitudinal approach to measure treatment impact” - Qing Wu, Google Inc.

Abstract: When randomized experiments are not possible and a large number of heterogeneous members can adopt treatment over a long time period (such as Google advertisers adopt new product features), estimating treatment impact is not a trivial task. Instead of using covariates to match treatment candidates to non-treatment candidates, we utilize the histories from both groups to predict the counterfactuals and estimate the differences between the predictions and actuals. Methods such as “nearest neighbour”, “elastic net” and “Bayesian seasonal structural time series” are explored. A further placebo test is used to adjust potential prediction bias.

”Impact Measurement using SEM” - Yongjian Kang, Google Inc.

Abstract: Google’s main business is focused on providing search advertising (adwords) services to small-medium businesses and enterprises alike - at a high level, we have developed various revenue programs targeting advertisers by how much they spend. One key objective for us is to evaluate the incremental revenue driven by our sales teams. We have traditionally set aside a share of advertisers as “control group”, and by comparing advertisers that our sales teams pitch to similar advertisers in the control group, we can estimate the incremental revenue they were responsible for. However, not all of our sales teams have a control group, and so far we haven’t been able to measure their incrementality. Recently, our team has been working to develop a new method that can do this. It uses a statistical approach called structural equation modeling (SEM) to estimate unobserved variables (in the case, what advertisers would have spent without us). Thus, by recovering the 'organic spend', we could obtain our impact estimation.

Applications of Data-Driven Decision Making, Room 20-043

Org/Chair: Tahir Ekin, Texas State University

"Data-Intensive Time and Motion Studies for Manufacturing Operations" - Francis Mendez, Texas State University

Abstract: The objective of this study is to implement motion and biometric sensing technology to collect data from simulated manufacturing processes and apply time and motion studies to improve human based manufacturing processes. A sample of workers, of various physical characteristics, will perform a series of pre-designed manufacturing-like movements. These movements have been cataloged, from a series of basic and simple movements to movements that are more complex in nature and more difficult to capture. The movements will be captured by infrared optical reflector technology capable of detecting fine movements (i.e., to the hundredth of a millimeter) in three dimensions. The statistical analysis of the data requires repeated observations for each subject. Therefore, the movements are carefully designed, cataloged and the data segmented in a form suitable for statistical analysis. The data will be analyzed in order to identify characteristic, common, as well as unusual patterns of movements. One of the objectives to this study is to use the results of the analysis to develop methods to improve the training of workers who perform these and other manufacturing processes. The researchers believe that it will be possible to identify factors that can be used to improve worker training to enhance safety and efficiency. To help identify these patterns, the researchers will use standard statistical techniques together with data-intensive methods such as data mining and machine learning techniques.

"High-Frequency Trading in Risk-Averse Portfolio Optimization with Higher-Order Risk Measures" - Sitki Gulten, Stockton University

Abstract: This study examines the application of risk-averse optimization techniques to high-frequency trading (HFT) in real-time portfolio management. First, I develop efficient clustering methods for scenario tree construction. Then, I construct a two-stage stochastic programming problem with higher-order conditional measures of risk, which is used to rebalance the portfolio on a rolling horizon basis, with transaction costs included. Finally, I present an extensive simulation study on both interday and high-frequency intraday real-world data of the methodology.

"Data-Driven Pre-Screening of Claims for Medical Audits" - Tahir Ekin, Texas State University

Abstract: Three to ten percent of the annual healthcare spending is estimated to be lost to overpayments. This increases the importance of resource allocation decisions within medical audits. Unsupervised data mining can be used as a pre-screening aid in medical fraud assessment. We present a Bayesian co-clustering method which can help identify the hidden patterns among providers and medical procedures via outlier detection and similarity assessment. We illustrate the proposed method using U.S. Medicare Part B data.

Reliability, Room CR3

Org/Chair: Kassie Fronczyk, Institute for Defense Analyses

"Reliability Assessment of Multiple Component Systems Using Bayesian Hierarchical Models" - *James Gilman, North Carolina State University*

Abstract: We use a Bayesian hierarchical model to assess the reliability of the Joint Light Tactical Vehicle (JLTV), which is a family of vehicles. The proposed model effectively combines information across three phases of testing and across common vehicle components. The analysis yields estimates of failure rates for specific failure modes and vehicles as well as an overall estimate of the failure rate for the family of vehicles. We are also able to obtain estimates of how well vehicle modifications between test phases improve failure rates. In addition to using all data to improve on current assessments of reliability and reliability growth, we illustrate how to leverage the information learned from the three phases to determine appropriate specifications for subsequent testing that will demonstrate if the reliability meets a given reliability threshold.

"Economic Complexity and Globalization of Services" - Saurabh Mishra, International Finance Corporation, World Bank Group, University of Maryland College Park

Abstract: Economic complexity has emerged as an important metric to measure nations' inherent capabilities embodied in the structure of economic production. However, this metric only measures manufacturing-based capabilities. Ignoring the role of services in economic production and growth may misinform critical policy and investment decisions, especially in developing countries where services are an increasingly important ingredient of competitiveness. This paper builds the first universal matrix of global trade incorporating not only physical goods, but also services to measure the complexity of exports. A non-linear iterative algorithm that ranks the 'fitness' of countries based on the diversity and complexity of their exports is applied. Two findings relating to the economic potential of emerging and developing markets are highlighted. First, emerging markets and many middle-income countries show greater economic strength adding services trade. Second, resource rich countries improve ranking when services are also used to measure economic complexity. The evidence highlights that technology-based "modern" services increase country fitness, and will remain an important component for future growth strategies.

"Reliability Modeling Incorporating Load Share and Frailty" - Vincent Raja Anthonisamy, University of Guyana

Abstract: The stochastic behavior of lifetimes of a two component system is often influenced primarily by the system structure and by the covariates shared by the components. Any meaningful attempt to model the lifetimes must take into consideration the factors affecting their stochastic behavior. In particular, for a load share system, we describe a reliability model incorporating both the load share dependence and the effect of observed and unobserved covariates. The model includes a bivariate Weibull (Lu (1989)) to characterize load share, a positive stable distribution to describe frailty, and incorporates effects of observed covariates. We investigate various interesting reliability properties of this model including cross ratio functions and conditional survivor functions. We implement profile maximum likelihood estimation of the model parameters and discuss model adequacy and selection. We illustrate our approach using a simulation study. For a real data situation, we demonstrate the superiority of the proposed model which incorporates both load share

and frailty effects over competing models that incorporate just one of these effects. An attractive and computationally simple cross-validation technique is introduced to reconfirm the claim. We conclude with a summary and discussion.

Thursday June 8th, 2017

9:00-10:00

Keynote Address, Auditorium

"Bayesian Model Choice: Past, Present, Future" - Merlise A. Clyde, Professor of Statistical Science, Duke University

Abstract: Significant advances have been made in Bayesian model selection and model averaging over the last 30 plus years in theory, computation and applications. While the Bayesian paradigm for model selection and uncertainty quantification is simple to describe, a challenge for practitioners is specification of prior distributions for the parameters defined for each candidate model. In variable selection this task becomes quickly daunting, particularly in the large p small n paradigm, as the number of models grows rapidly with the number of predictors p. Because of the difficulty of subjective prior specification, there have been a number of attempts to define conventional or objective prior distributions for Bayesian model selection ranging from Zellner's g-prior or mixtures of g-priors to generalized ridge prior distributions. In addition to prior distributions on model specific parameters, prior probabilities on models play a key role in the large p paradigm. We discuss various criteria that have been deemed essential for model selection priors in the context of linear and generalized linear models and extensions. We highlight recent advances in theory, computation and software, and close with a discussion of challenges that need to be addressed.

10:30-12:00

From Samples to Target Populations, Room 20-043

Org/Chair: Chaitra H. Nagaraja, Fordham University

"Observational Methods for Health Policy Decision-Making" - Frank Yoon, IBM Watson Health

Abstract: Healthcare reform has been spurred by recent innovations in service and payment delivery, such as accountable care organizations or behavioral health integration, often tested in pilots or other limited settings. To make decisions about scaling these pilots to the national level, policymakers must know their impacts. With large administrative claims databases, the analyst can estimate those impacts using observational methods grounded in (1) good study design and (2) flexible analytics. By good study design, we will demonstrate how blocking or stratification removes bias due to confounders, and by flexible analytics, we will illustrate computationally quick approaches, including Bayesian modeling, to estimate impacts with fewer statistical assumptions. In a commercial claims database, we will describe an analytic approach to estimate impacts of a behavioral health program that could be scaled nationwide. Our approach will show how to apply (1) blocking to define subgroups and (2) the Bayesian framework to borrow strength across multiple domains to increase statistical precision.

"Designing Randomized Trials for Making Generalizations to Policy-Relevant Populations" - Elizabeth Tipton, Teachers College, Columbia University

Abstract: Randomized trials are common in education, the social sciences, and medicine. While random assignment to treatment ensures that the average treatment effect estimated is causal, studies are typically conducted on samples of convenience, making generalizations of this causal effect outside the sample difficult. This talk provides an overview of new methods for improving generalizations through improved research design. This includes defining an appropriate inference population, developing a sampling plan and recruitment strategies, and taking into account planned analyses for treatment effect heterogeneity. The talk will also briefly introduce a new webtool useful for those planning randomized trials in education research.

”An Instrumental Variable Approach to Generalizing Experimental Results” - Chaitra H. Nagaraja, Fordham University

Abstract: While assignment to treatment and control groups are randomized in experiments (e.g., clinical trials), selection into the trial itself is most often not. This makes it more difficult to generalize experimental results to the wider population. Current techniques to model trial participation generally require population-level auxiliary data to supplement the sample information. We take a different approach. Our strategy is to use an instrumental variable, such as distance from the trial site, to model these trial participation probabilities without confounding. Such a variable avoids the omission bias by being correlated with trial participation but not directly with trial outcomes or other covariates. From this foundation, we develop a statistical test to check for generalizability.

Statistics Engineering at Facebook, Room CR3

Org/Chair: Daniel Merl, Facebook

”Forecasting at Scale” - Ben Letham, Facebook

Abstract: There are a variety of challenges that come with producing a large number of forecasts across a diverse collection of time series. Our approach to forecasting "at scale" is a combination of configurable models and thorough analyst-in-the-loop performance analysis. I will talk about a forecasting approach we have developed based on a decomposable model with interpretable parameters that can be intuitively adjusted by the analyst. It is optimized for the type of business forecasting problems that we frequently encounter at Facebook: strong multiple seasonalities, missing data and outliers, large trend shifts, important holiday effects, and saturating growth. I will discuss the performance analyses that we use to help analysts use their expertise most effectively and scale up forecasting across the company. The software has been open sourced and is available in both Python and R.

”Bayesian optimization for infrastructure systems” - Brian Karrer, Facebook

Abstract: Facebook has numerous infrastructure systems supporting its operation. These systems are configured through even more numerous parameters that control their performance, as well as alter trade-offs between quantities such as memory usage and CPU consumption. Tuning these configuration parameters can be a difficult task even for experts with substantial domain knowledge. Even when this is possible, manual tuning by experts can be hard to scale and maintain.

Similar to hyperparameter tuning for machine learning, an alternative to expert tuning is to utilize Bayesian optimization. In this presentation, I discuss efforts to apply Bayesian optimization to tune parameters at Facebook through sequential experimentation, focusing on the practical challenges that have arisen, including measurement error and experiment parallelism.

”Challenges of A/B Testing at Facebook” - John Myles White, Facebook

Abstract: In this talk, I'll describe the challenges that Facebook has addressed while developing Deltoid, an internal system that automatically analyzes the results of A/B tests. Deltoid is designed to enable engineers, analysts and product managers to rapidly perform custom analyses of their A/B tests in accordance with statistical best practices. What makes Deltoid interesting to statisticians is that Deltoid's success as a system is measured in both statistical and software engineering terms. For example, Deltoid strives to report valid point and interval estimates to end-users within one minute after the user requests an analysis of their A/B test. Given that Deltoid is used to analyze datasets that range in size from kilobytes to terabytes, this total computation time constraint introduces challenges that are seldom considered in statistical research. By elaborating upon these kinds of technical challenges, I will show how Deltoid's design provides a useful case study for statisticians interested in making tradeoffs between computational complexity and statistical optimality.

Spatio-Temporal Statistics for Environmental Sciences and Climatology, Auditorium

Org/Chair: Huijing Jiang, IBM Research

”Bayesian Spatio-Temporal Factor Analysis for Prediction” - Candace Berrett, Brigham Young University

Abstract: Monitoring sea levels at a local level is vital for protecting coastal populations, ecological systems, and natural resources. To aid in monitoring these levels, the US National Oceanic and Atmospheric Administration (NOAA) collects hourly measurements of sea level using tide gauges at various locations around the US. These data provide essential information for learning about the patterns and changes in sea level at their respective locations. However, these data pose several challenges that are prevalent in many environmental research problems. First, many of the monitors have large portions of missing data; second, daily and seasonal tidal patterns vary widely from one location to another and from one year to another; and third, it is impossible to obtain measurements at all locations along the coast and therefore difficult to examine sea levels at unmonitored locations. In this talk, we present a comprehensive spatio-temporal model for accommodating each of these challenges in order to provide accurate inferences on sea level patterns at missing times and unmonitored locations. We illustrate this model using hourly tidal readings from the past 30 years at 38 gauges from the east coast of the US.

”Spectral Radiance in Climate Study” - Taps Maiti, Michigan State University

Abstract: In climate change study, the infrared spectral signatures of climate change have recently been conceptually adopted, and widely applied to identifying and attributing atmospheric composition change. We propose a Bayesian hierarchical model for spatial clustering of the high-dimensional functional data based on the effects of functional covariates. We couple the functional mixed-effects model with a generalized spatial partitioning method for: (1) identifying subregions for the high-dimensional spatio-functional data; (2) improving the computational feasibility via parallel computing over subregions or multi-level partitions; and (3) addressing the near-boundary ambiguity in model-based spatial clustering techniques. The proposed model extends the existing spatial clustering techniques to produce spatially contiguous partitions for spatio-functional data. Moreover, it improves the model exploration and captures the variability near boundaries to address uncertainty in partitioning large and high-dimensional spatial data. Dimension reduction in the vertical direction is also achieved via Bayesian wavelets. The model successfully captured the regional effects of the atmospheric and cloud properties on the spectral radiance measurements. This elaborates the importance of considering spatially contiguous partitions for identifying regional effects and small-scale variability.

”Coupled Physical and Statistical Models for Renewable Energy Integration” - Lloyd Treinish, The Weather Company & IBM Research

Abstract: Electricity customers want energy that is more reliable and cleaner as well as the ability for self-generation, especially via solar power. In addition, utilities are being required to manage the increased penetration of renewable resources, yet maintain competitive markets for generation. Renewable energy production and energy demand have significant sensitivity to local, short-term weather conditions. The resultant intermittency in generation coupled with variation in demand can lead to reliability issues with the transmission system, grid stability problems and shortfall of energy. In Vermont, there are additional challenges from an operational perspective as a result of the local geography, including complex terrain, variable weather conditions at a local scale and the diversity of electricity consumers.
The aforementioned issues lead to inherent uncertainty in the information required to improve grid reliability with further penetration of renewable generation. Since this uncertainty is poorly quantified, conservative grid management leads to curtailment of renewable power production (i.e., wind). Over the last couple of years, there has been a rapid expansion in the deployment of solar power systems, especially at the smaller scale, which has added to the complexity of this situation.

To address these and related challenges, the Vermont Weather Analytics Center (VWAC) has been developed. It employs a system of coupled models to enable improved understanding of both production and demand. While each model attempts to deliver more accurate results, the models also quantify the inherent uncertainty in each respective component. The effort starts with weather, which is addressed through physical modelling in a system we call Deep Thunder. These methods, known as numerical weather prediction, involve a finite-difference solution of a set of coupled partial differential equations that model a diversity of atmospheric processes as an initial and boundary value problem. While we apply these methods globally to support a diversity of applications, the targeted version of this model for the VWAC is run operationally every 12 hours at 1-km horizontal resolution with high vertical resolution in the lower boundary layer for regional coverage for 72 hours. To reduce and characterize the errors in the initial state, three-dimensional variational data

assimilation is performed around each initialization time using observations from the Vermont mesonet, which was developed at part of the VWAC, as well as from other sources of public and private weather measurements.

Once each execution of the weather model is completed, the results are abstracted to include key variables at the appropriate temporal and spatial resolution. The variables include both direct model output as well as diagnostic fields derived from specialized post-processing. These data then permit execution in parallel of data-driven (i.e., via statistical and machine-learning) models to predict wind and solar power, and electricity demand. All of these models operate at a granularity that enables aggregation from the 1 km computational weather grid. Hence, at the finest scale, wind power is done at the turbine level, solar at each utility-scale facility and demand at the distribution substation. In addition, the demand model predicts solar power generation for distributed systems behind the meter (e.g., rooftop deployments), aggregated to the substation level. These models use training sets, which consist of both forecasts and hindcasts of the weather model, and historical power and other data from the utilities.

We will present an overview of the deployed capabilities along with the results to date and the overall effectiveness of our particular approach. We will also discuss ongoing issues such as calibration of data and quantifying uncertainty as well as recommendations for future work.

2:00-3:30

Bayesian Analysis for Large and Complex Data, Room CR3

Org/Chair: Tamara Broderick, MIT

“Online Patient Monitoring Using Medical Time Series Data” - Barbara Engelhardt, Princeton University

Abstract: Electronic health records have a wealth of noisy data capturing patient state and medical interventions for patients across time. In this work, we develop a multivariate Gaussian process model for online monitoring of hospital patients. Using a structured kernel across clinical traits, we find we are able to estimate patient state and also interpret relationships among the different clinical covariates. Finally, we use this online monitoring framework with a reinforcement learning approach to estimating a treatment policy for removing a patient in the ICU from ventilation.

”Machine Learning for Improving Healthcare” - Katherine Heller, Duke University

Abstract: Machine learning and statistical modeling have the potential to revolutionize the way medical services and patient care operate. In this talk I will discuss dynamic models for EHR and mobile app-based data. This includes applications to chronic kidney disease, surgical complications prediction, infectious disease, and chronic disease mobile data.

“Approximate Sufficient Statistics for Scalable Bayesian Inference” - Tamara Broderick, MIT

Abstract: The use of Bayesian methods in large-scale data settings is attractive because of the rich hierarchical models, uncertainty quantification, and prior specification they provide. However, standard Bayesian inference algorithms can be computationally difficult or infeasible exactly in the settings of modern interest: large data sets and complex models. Sufficient statistics allow easy, large-scale computation for many models. But not all models have sufficient statistics readily available. We propose instead to discover and compute fast summaries of the data, which we call approximate sufficient statistics, as a pre-processing step. We provide theoretical guarantees on the size and approximation quality of our summaries. And we show that our approach can be extended to streaming and parallel settings, with minimal additional effort. We demonstrate the efficacy of our approach on a number of synthetic and real-world datasets.

Machine Learning with Mixed Data Types, Auditorium

Org/Chair: Bonnie Ray, Arena

”Variable Selection for Correlated Bivariate Mixed Outcomes Using Penalized Generalized Estimating Equations” - Elizabeth Schifano, University of Connecticut

Abstract: We propose a penalized generalized estimating equations framework to jointly model correlated bivariate binary and continuous outcomes involving multiple predictor variables. We use sparsity-inducing penalty functions to simultaneously estimate the regression coefficients and perform variable selection on the predictors, and use cross-validation to select the tuning parameters. We further propose a method for tuning parameter selection that can control a desired false discovery rate. Using simulation studies, we demonstrate that the proposed joint modeling approach performs better in terms of accuracy and variable selection than separate penalized regressions on the binary and the continuous outcomes. We demonstrate the application of the method on a medical expenditure data set.

”A Semiparametric Method for Clustering Mixed Data” - Marianthi Markatou, University at Buffalo

Abstract: Despite the existence of a large number of clustering algorithms, clustering remains a challenging problem. As large datasets become increasingly common in a number of different domains, it is often the case that clustering algorithms must be applied to heterogeneous sets of variables, creating an acute need for robust and scalable clustering methods for mixed continuous and categorical scale data. We show that current clustering methods for mixed-type data suffer from at least one of two central challenges: (1) they are unable to equitably balance the contribution of continuous and categorical variables without strong parametric assumptions; or (2) they are unable to properly handle data sets in which only a subset of variables are related to the underlying cluster structure of interest. We first develop KAMILA (KAy-means for MIxed LArge data), a clustering method that addresses (1) and in many situations (2), without requiring strong parametric assumptions. We next develop MEDEA (Multivariate Eigenvalue Decomposition Error Adjustment), a weighting system that addresses (2) even in the face of a large number of uninformative variables. We study theoretical aspects of our method and demonstrate their superiority in a series of Monte Carlo simulation studies and a set of real-world applications.

“Scalable Computation with Skinny Gibbs Sampler for High Dimensional Bayesian Models” - Naveen Naidu Narisetty, University of Illinois at Urbana-Champaign

Abstract: The Bayesian paradigm offers a flexible modeling framework for analyzing data with complex structures, but relative to penalization-based methods, it faces a harsher computational burden due to the posterior computation involved. A particularly challenging problem is to devise scalable Bayesian computational methods for high dimensional data settings. In this talk, I will introduce a new Gibbs sampling algorithm for posterior computation called "Skinny Gibbs”, which is much more scalable than the standard Gibbs samplers for large datasets. In particular, the complexity of the algorithm is only linear in the number of variables at each iteration. Our Skinny Gibbs algorithm results in the property of strong model selection consistency and is flexible to use in a variety of problems including linear and logistic regressions, and a more challenging problem of censored quantile regression where a non-convex loss function is involved. I will demonstrate statistical and computational performance of our approach through empirical studies.

Web Applications in Statistics and Machine Learning, Room 20-043

Org/Chair: Cheryl Flynn, AT&T Labs Research

”Television and Digital Advertising: Second Screen Response and Coordination with Sponsored Search” - Shawndra Hill, Microsoft Research

Abstract: I will present research that considers the potential to improve the efficiency and efficacy of broader advertising efforts through cross channel coordination between TV and digital advertising. In our research, we consider the types of devices on which search response predominantly manifests following TV advertisements, and the degree to which shifts in search activity can be used to evaluate the success of TV advertisers’ targeting efforts. We leverage data on TV advertising around Microsoft Windows 10 and an Xbox video game, in combination with large-scale proprietary search data from Microsoft Bing. Our identification strategy hinges on a combination of geographic heterogeneity in TV advertising exposure and continuous variation in the cost of TV advertisements (a proxy for TV audience size). We first demonstrate that search response peaks within three minutes of the airing of a TV advertisement, and that this manifests primarily via second-screen devices. Our estimated elasticities indicate that a 20% increase in advertising spend equates to an approximately 2.5% (3.4%) increase in search volumes for Windows 10 (the Xbox game). Second, we show that, indeed, the demographic groups targeted by TV advertisements are those most likely to respond, and we thereby demonstrate that TV ad effectiveness can be usefully measured via online search data. Third, examining sponsored search clicks in our query-level data, for queries involving brand-related keywords, we demonstrate a significant increase in rank-ordering effects in searches that take place in the minutes immediately following a TV advertisement, which implies a complementarity between TV and sponsored search advertisements. In my talk, I will also discuss the future of cross-channel advertising coordination and the many projects we have underway combining (Re)Search and TV.

“Blockchain and Ledger Analytics” - Roman Vaculin, IBM Research

Abstract: In this talk we will discuss our efforts focused on Blockchain and ledger analytics -- analytics capabilities on data inside blockchain applications. The objective of our work is to enable a set of analytics and cognitive capabilities that will allow clients running production blockchain networks to analyze blockchain data in-place maintaining blockchain’s inherent characteristics and helping preserve the security, trust and integrity of the data without transferring the data out of the blockchain through the ETL process which has the potential to expose data to security risks.

“From Which World Is Your Graph?” - Zhenming Liu, College of William and Mary

Abstract: Discovering statistical structure from links is a fundamental problem in the analysis of social networks. Choosing a misspecified model, or equivalently, an incorrect inference algorithm will result in an invalid analysis or even falsely uncover patterns that are in fact artifacts of the model. This work focuses on unifying two of the most widely used link-formation models: the stochastic blockmodel (SBM) and the small world (or latent space) model (SWM). Integrating techniques from kernel learning, spectral graph theory, and nonlinear dimensionality reduction, we develop the first statistically sound polynomial-time algorithm to discover latent patterns in sparse graphs for both models. When the network comes from an SBM, the algorithm outputs a block structure. When it is from an SWM, the algorithm outputs estimates of each node's latent position.

”Deconstructing Domain Names to Reveal Latent Topics” - Cheryl Flynn, AT&T Labs

Abstract: Measurement of the lexical properties of domain names enables many types of relatively fast, lightweight web mining analyses. These include unsupervised learning tasks such as automatic categorization and clustering of websites, as well as supervised learning tasks, such as classifying websites as malicious or benign. In this paper we explore whether these tasks can be better accomplished by identifying semantically coherent groups of words in a large set of domain names using a combination of word segmentation and topic modeling methods. By segmenting domain names to generate a large set of new domain-level features, we compare three different unsupervised learning methods for identifying topics among domain name keywords: spherical k-means clustering (SKM), Latent Dirichlet Allocation (LDA), and the Biterm Topic Model (BTM). We successfully infer semantically coherent groups of words in two independent data sets, finding that BTM topics are quantitatively the most coherent. Using the BTM, we compare inferred topics across data sets and across time periods, and we also highlight instances of homophony within the topics. Finally, we show that the BTM topics can be used as features to improve the interpretability of a supervised learning model for the detection of malicious domain names. To our knowledge this is the first large-scale empirical analysis of the co-occurrence patterns of words within domain names.

4:00-5:30

Data Tales from Industry, Auditorium

Org/Chair: Claudia Perlich, Dstillery

”Match Making for Tech Jobs” - Jon Krohn, Untapt

Abstract: Job-seekers predominantly sift through employment possibilities by manually navigating job boards or consulting with human recruitment specialists that have limited bandwidth and finite opportunities. This is akin to using the classifieds section of the newspaper or word of mouth to find a romantic partner today. Your contemporaries, meanwhile, are finding their soulmate by leveraging explicit (e.g., OkCupid, Match.com) or implicit (Tinder, The League) preference algorithmic matching to identify their soulmate from an unbounded pool of possibilities. At untapt, we have developed a correspondingly modern matchmaking experience for jobs in the technology sector. We have built an ensemble model of Bayesian regression and deep-learning neural network approaches and applied it to a data set consisting of a million software developer profiles and tens of thousands of hiring decisions to learn explicit and implicit preferences. The probabilities of job-application success output by the model enable our platform to programmatically suggest the best-suiting roles to candidates, provide instantaneous feedback to prospective job-seekers, and filter applications presented to hiring firms. In aggregate, these features culminate in interview rates that greatly surpass industry benchmarks.

”You are probabilistically here: Lessons from working with mobile device generated geodata” - Peter E Lenz, Dstillery

Abstract: Mobile devices have created an explosion in the amount of spatial data available to researchers, but come with unique quality issues. For the sake of scale we have traded any control over the methodology used to generate it. This talk will discuss the multitude of methods available to mobile device to determine your location, how a device would decide to apply each, and methods Dstillery has developed to quality control, classify, and learn from this blind-source geodata.

”Improving Police Stop Efficiency in New York City” - Ravi Shroff, New York University

Abstract: Recent studies have examined the use of stop-and-frisk—a widely employed but controversial policing tactic—in New York City. Much controversy has centered around both the high volume of stops as well as racial disparities in the distribution of stopped individuals. We investigate by analyzing 750,000 stops in New York City over five years, focusing on cases where officers suspected the stopped individual of criminal possession of a weapon (CPW). For each such CPW stop, we estimate the ex-ante probability that the detained suspect would have a weapon, and demonstrate that by conducting only the 6% ex-ante highest hit rate stops, one can both recover the majority of weapons and mitigate racial disparities. Finally, we develop stop heuristics that can be implemented as a simple scoring rule, with comparable accuracy to our full statistical model.

Statistical Methods in Medicine, Room CR3

Org/Chair: Rebecca Yates Coley, Group Health Research Institute

"Simulation Controlled Seamless Phase II/III Clinical Trials" - Lindsay Berry, Duke University

Abstract: This paper focuses on seamless phase II/III trials, an adaptive trial consisting of two stages: the first stage compares multiple treatment arms to a control and determines the appropriate arm, and the second stage undertakes a more traditional comparison of the selected arm to a control arm. The data from both stages is used in the final analysis of the treatment, with the goal of more power for arm selection and confirmation than two separate trials. This paper introduces a method of evaluating seamless trials through simulation. This method relies on simulation of clinical trials under the null hypothesis to find critical values that control type one error. Power comparisons between the Posch method, the simulation method, and two separate trials conclude that seamless trials are more powerful than separate trials, and the simulation method is slightly more powerful than the Posch method when analyzing seamless trials. Power comparisons of separate trials and the seamless trial with the same total sample size reveals the seamless trials are optimized with larger phase II portions, while maintaining superior overall power.

"Unlocking real-world oncology data using electronic health records" - Sandy Griffith, Flatiron Health

Abstract: Overall survival (OS) is an important outcome measure in oncology, but it may require large sample sizes and long follow-up time to provide meaningful insights. As a supplement to OS, outcomes based on disease progression and tumor response typically require less follow-up time and are also captured in oncology clinical trials to assess treatment efficacy. The standard methods to collect these outcomes in clinical trials, however, may not be feasible using electronic health records (EHR), but such data present an opportunity to answer research questions at a scale and recency not available from clinical trials, while also reflecting treatment patterns and populations seen in routine clinical practice. As we develop real-world versions of these outcomes, it is critical to apply a statistical framework to understand their quality and ensure reliability and validity. We illustrate such a framework with a case study evaluating the reliability and validity of real-world progression in advanced non-small cell lung cancer. We discuss statistical considerations unique to working with EHR data, and identify methodological gaps that present opportunities for future development.

"A Data Science Framework for Learning Health Systems" - Rebecca Yates Coley, Kaiser Permanente Washington Health Research Institute

Abstract: Learning health systems promise to improve medical decision-making in the era of Big Data by making up-to-date analyses of patient information and scientific knowledge available to physicians and patients in real time. Data scientists play an essential role in developing tools that enable intelligent, dynamic use of patient data to provide the "right treatment to the right patient at the right time". In this talk, we will outline a data science framework for supporting local, disease-specific learning health systems. We will discuss an example from Johns Hopkins Medicine, where a multidisciplinary team has developed and deployed a tool to assist in personalized management of prostate cancer.

Data Driven Semi-Parametric Models, 20-043

Org/Chair: Emre Barut, George Washington University

"Predicting Class Size for Fun and Profit with Semi-parametric Regression Models" - Harlan Harris, WeWork

Abstract: Educational companies that are trying to make a profit often need to make cancellation and facility decisions in the weeks before a class starts. In this talk, I will recall the challenges faced when trying to predict class size, and will describe a quasi-Bayesian approach to extrapolating enrollments in a principled, statistically honest manner. A key piece of this approach is a class of semi-parametric regression models called GAMLSS -- Generalized Additive Models for Location, Scale, and Shape. By modeling the shape of enrollment priors and timing curves as semi-parametric functions of class properties, we are able to make predictions for individual classes early enough that business decision-makers could minimize the impact on students and faculty. For nontechnical users of predictions, it is important not to be overly precise, as inevitable errors lead quickly to mistrust. The use of posterior distributions and prediction intervals, instead of point estimates, was important in the roll-out of this model.

"Bayesian Pollution Source Identification via an Inverse Physics Model" - Youngdeok Hwang, IBM Research

Abstract: Behavior of air pollution is governed by the complex dynamics, in which air quality of a site is affected by the pollutants transported from neighboring locations via physical processes. To estimate the source of observed pollution, it is crucial to take the atmospheric condition account. Traditional approach to build empirical models uses observations, but is not able to incorporate the physical knowledge. This drawback becomes particularly severe for the situations where a near-real time source estimation is needed. To that end, we propose a Bayesian method to estimate the pollution sources, by exploiting both the physical knowledge and observed data. The proposed method uses a flexible approach to utilize the large scale data from the numerical weather prediction model while incorporating the physical knowledge into the model. The method is illustrated with a real data set.

"An Efficient Method for Parameter Estimation Under Model Contamination" - Emre Barut, George Washington University

Abstract: In parameter estimation problems, maximum likelihood (ML) approaches possess a series of advantageous properties, which has led to their common use in everyday statistical applications. Unfortunately, ML based methods are also known to be very sensitive to model misspecification issues or outliers. In that spirit, we provide a new framework for parameter estimation called DAS estimator, which is given as the empirical minimizer of a second order U-statistic. When estimating parameters in the exponential family, the estimator is shown to be a solution of a quadratic convex problem that can be efficiently solved. For parameter estimation, our approach significantly improves upon MLE when outliers are present, or when the model is misspecified. Furthermore, we show how DAS estimator can be used to efficiently fit to distributions with unknown normalizing constants. Extensions of DAS estimators for regression and their implications for statistical modeling are discussed.

Friday June 9th, 2017

9:00-10:00
Keynote Address, Auditorium

"When Less is More: Adaptive data collection for economic indicators" - *Moorea Brega, Sr. Director, Data Science, Premise Data*

Abstract: The Consumer Price Index (CPI) is a statistical construct used to estimate what typical consumers pay for a standard basket of goods and services for domestic consumption. By tracking price changes over time, the CPI series can be used to quantify headline inflation trends. Inflation trends are fundamental to accurately characterising key macroeconomic aggregates (such as real Gross Domestic Product and real exchange rates) as well as important microeconomic variables (such as household living standards and purchasing power). However, collecting the quality and quantity of data needed to compute the CPI and other price indices has generally been both a statistical and operational challenge for National Statistical Offices. As a result of these challenges and resource constraints, data collection is often infrequent and data validation is minimal.

Premise is a worldwide network and predictive analytics platform that maps ground truth to drive better outcomes. We are utilizing our network of on-the-ground contributors to collect price data via our mobile app in order to produce high frequency estimates of price indices. Unlike traditional, static data collection efforts, the Premise platform leverages incoming data to understand price variability and its impact on the resulting price index in order to more effectively target future data collection. As a result, Premise can deliver more accurate and frequent estimates for the same data volume and budget than would be achieved through a more traditional static design. In this talk, I will describe the approach we employ to improve index accuracy as well as some of the operational challenges and constraints involved in crowdsourcing price data in developing countries.

10:30-12:00
Panel Discussion: Succeeding as an Early-Career Data Scientist in Today's Industry, Room 20-001

Org/Chair: Grant Weller, Savvysherpa

Panel discussants:

Mattia Ciollaro, Spreemo Health
Reka Daniel-Weiner, Dstillery
Ryan Roundy, Oracle

Novel Hierarchical Bayesian Approaches for Business and Government Applications, Auditorium

Org/Chair: Nalini Ravishanker, University of Connecticut

“Regression and Reliability Models for Predicting Customer Churning” - Sujit K. Ghosh, NCSU and SAMSI

Abstract: In today’s analytics, successful prediction of the lifetime of customer’s subscription has huge impact on business revenue sources. In the era of IoT, the fact that customers (e.g., with subscription-based services) have multiple choices among service providers, it is of huge interest to predict if and when a customer will churn (i.e. stop service/subscription) based not only on his/her own past history but also on longitudinally collected information on `similar' group of customers. This talk presents some of the best practices and tools used for predicting customer churning, and a newly developed semiparametric model attempts to identify key factors that leads to churning.

”Multivariate Spatio-Temporal Survey Fusion with Application to the American Community Survey and Local Area Unemployment Statistics” - Scott Holan, University of Missouri and US Census Bureau

Abstract: There are often multiple surveys available that estimate and report related demographic variables of interest that are referenced over space and/or time. Not all surveys produce the same information, and thus, combining these surveys typically leads to higher quality estimates. That is, not every survey has the same level of precision nor do they always provide estimates of the same variables. In addition, various surveys often produce estimates with incomplete spatio-temporal coverage. By combining surveys using a Bayesian approach, we can account for different margins of error and leverage dependencies to produce estimates of every variable considered at every spatial location and every time point. Specifically, our strategy is to use a hierarchical modelling approach, where the first stage of the model incorporates the margin of error associated with each survey. Then, in a lower stage of the hierarchical model, the multivariate spatio-temporal mixed effects model is used to incorporate multivariate spatio-temporal dependencies of the processes of interest. We adopt a fully Bayesian approach for combining surveys; that is, given all of the available surveys, the conditional distributions of the latent processes of interest are used for statistical inference. To demonstrate our proposed methodology, we jointly analyze period estimates from the US Census Bureau's American Community Survey, and estimates obtained from the Bureau of Labor Statistics Local Area Unemployment Statistics program.

This is joint work with Jonathan R. Bradley (Florida State University) and Christopher K. Wikle (University of Missouri)

“Dynamic Models for Multivariate Times Series of Counts” - Nalini Ravishanker, University of Connecticut

Abstract: Discrete-valued time series modeling is emerging as an important research area with diverse applications, as discussed in the recent CRC Handbook of Discrete-valued Time Series. Using Markov Chain Monte Carlo (MCMC) methods for Bayesian hierarchical dynamic modeling of vector time series of counts under a multivariate Poisson sampling distributional assumption may be

computationally demanding, especially in high dimensions. An alternate flexible level correlated model (LCM) framework is described in this talk. This enables us to combine different marginal count distributions and to build a hierarchical model for the vector time series of counts, while accounting for association between components of the response vector. We employ the Integrated Nested Laplace Approximation for fast approximate Bayesian modeling using the R-INLA package (r-inla.org). The approach lends itself to application in diverse areas such as ecology, marketing and transportation safety. In this talk, we describe analysis of marketing data from a large multinational pharmaceutical firm. We describe models for monthly new prescription counts that are written by physicians for the firm's focal drug and for competing drugs, as functions of physician-specific and time-varying predictors. To enhance computational speed, we first build a model for each physician, use features of the estimated trends in the time-varying parameters in order to cluster the physicians into groups, and fit aggregate models for all physicians within each cluster. Our three-stage analysis can provide useful guidance to the pharmaceutical firm on their marketing actions.

New Statistical Methods for Stochastic Volatility and Functional Data in Finance and Engineering , Room 20-043

Org/Chair: Mengyang Gu, Johns Hopkins University

”Efficient Portfolio Allocation with Sparse Volatility Estimation for High-Frequency Financial Data” - Jian Zou, Worcester Polytechnic Institute

Abstract: Traditionally, investors try to estimate short term portfolio volatility based on daily return. When tick-by-tick data are available, investors use different volatility estimators based on high-frequency data to evaluate the portfolio risk aiming at outperforming those based on low-frequency data. In this paper, we optimize block realized kernel estimator in Hautsch et al. (2015) and propose another more efficient way when we deal with the large portfolio allocation. Our research contribution focuses on the benefits of high-frequency data for portfolio allocation based on sparse volatility estimate methods. This process provides us new insights and alternatives when we want to set up a sensible investment strategy especially for risk averse investors.

”Shape-constrained Semiparametric Additive Stochastic Volatility Models” - Xinyi Xu, Ohio State University

Abstract: The Gaussian stochastic process is the most commonly used approach for modeling time series data. The Gaussianity assumption, however, is known to be insufficient or inappropriate in many problems. On the other hand, nonparametric stochastic volatility models provide great flexibility for modeling the volatility equation, but they often fail to account for useful shape information. For example, a model may not use the knowledge that the autoregressive component of the volatility equation is monotonically increasing as the lagged volatility increases. In this work, we propose a class of additive stochastic volatility models, which capture the asymmetry and heavy tails of many real-world time series data and allow for different shape constraints to improve estimation efficiency. We develop a Bayesian fitting algorithm and demonstrate model performances on simulated and empirical datasets. Unlike general nonparametric models, our model sacrifices little when the true volatility equation is linear. In nonlinear situations we improve the model fit and the ability to estimate volatilities over general, unconstrained, nonparametric models.

”Robust Gaussian Stochastic Process” - Mengyang Gu, Johns Hopkins University

Abstract: We consider the estimation of the parameters of a Gaussian Stochastic Processes (GaSP) with multiple inputs through various generalized maximum likelihood methods, mostly involving finding posterior modes as the full Bayesian analysis is typically computationally expensive. This is a difficult estimation problem, with poorly behaved likelihoods, so study of the robustness of the estimators is crucial. We demonstrate that certain parameterizations result in more robust estimators than others, and that some parameterizations which are in common use should clearly be avoided. These results are applicable to many frequently used covariance functions, e.g., power exponential, Matern, rational quadratic and spherical covariance; we also generalize the results to GaSP with a nugget parameter. Both theoretical and numerical evidence will be presented concerning the performance of the studied procedures. Examples in emulating complex computer models and estimating utility functions will be discussed.

Analytics Promoting Social Good: Money Access, Housing and Health, Room CR3

Org/Chair: Marian Farah, The Climate Corporation

”Route-based data collection to determine mobile money access in Africa” - Ankur Gupta, Premise Data

Abstract: In many developing countries, geospatial information about infrastructure such as financial institutions, availability of utilities (e.g. electricity and water), and access to health care is not readily available. This information is essential for making development decisions such as where to build new banks, power stations, and health clinics. Premise provides a real-time, scalable platform for collecting ground truth data and providing actionable insights in “data dark” locations using on-the-ground contributors in developing countries. One such project at Premise is aimed at assessing the mobile money infrastructure in Africa. Mobile money provides legal, secure, convenient, and low-cost financial access to underbanked people in Africa and has been shown to reduce poverty. As a part of this project, we utilize the Premise platform to obtain current information about locations and types of mobile money services offered within a geographical region. We use a route-based data collection methodology to determine the geographical availability of mobile money. Contributors are paid to walk along a predefined route of length 0.5 -- 1 km and identify mobile money locations in the form of shops, kiosks, and street vendors. When contributors find a mobile money location, they fill out a short survey while the Premise smartphone app records the geolocation of the mobile money vendor. In this talk, I will present some of the statistical methods we use to design and implement route-based data collection.

”Risk factors for eviction in rapidly developing cities” - Ryan Brady, Apteligent

Abstract: An unfortunate side effect of urbanization is that existing families are displaced. In San Francisco, roughly 13,000 evictions have taken place in the last 20 years (roughly 6% of the total rental stock). In this talk I will introduce the data visualizations already produced by the Anti-Eviction Mapping Project, and discuss ongoing work in linking specific geospatial data from public San Francisco data to eviction risk. Identifying these risk factors will enable activist groups to make more informed choices on how to allocate limited advocacy resources

“A Bayesian Approach for Predicting Neotropical Primate Reservoirs of Zika Virus” - Kush Varshney, IBM Research

Abstract: Emerging infectious diseases are an increasing global problem. While our responses to novel disease outbreaks are becoming more efficient and expedient, it is clear that a reactive approach to future disease threats is unsustainable, especially as disease events become more frequent and more widely distributed in an increasingly global society. One recent example is the emergence in the western hemisphere of Zika virus, spread by mosquitos and hosted by primates as reservoirs. Given the difficulty of effective mosquito eradication, combined management of mosquito removal and wild reservoir management provides the best chance of reducing Zika virus spillover into human populations. In the paleotropics (Africa), documented wildlife reservoirs include multiple primate species whose natural history and ecology are relatively well known, but the reservoirs are unknown in the neotropics (South America). To predict neotropical reservoirs, we collate and impute relevant biological and ecological data and then use a supervised learning approach with paleotropical primates as training samples. We use a Bayesian hierarchical latent factor model to obtain disease-specific risk scores for all primates, identifying high-risk undetected primates in the process. Important predictors are also identified using variable loadings in the latent factors. The model predictions have been informally validated via field testing by an independent research group.